Post-Graduate Degree Programme (CBCS)

in

ZOOLOGY

SEMESTER-III

SPECIAL PAPER

CYTOGENETICS AND MOLECULAR BIOLOGY

ZDSE(MJ)T-301

SELF LEARNING MATERIAL

DIRECTORATE OFOPEN AND DISTANCE LEARNING

UNIVERSITY OF KALYANI

KALYANI, NADIA,

W.B. INDIA

Content Writer:

Dr.Subhabrata Ghosh, Assistant Professor of Zoology, Directorate of Open and Distance Learning, University of Kalyani.

Acknowledgements:

The author thankfully acknowledges all the faculty members of Department of Zoology, University of Kalyani for their academic contribution and valuable suggestions regarding the preparation of Self Learning Material.

MAY 2023

Directorate of Open and Distance Learning, University of Kalyani.

Published by the Directorate of Open and Distance Learning,
University of Kalyani, Kalyani-741235, West Bengal.

Director's Message

Satisfying the varied needs of distance learners, overcoming the obstacle of distance and reaching the unreached students are the threefold functions catered by Open and Distance Learning (ODL) systems. The onus lies on writers, editors, production professionals and other personnel involved in the process to overcome the challenges inherent to curriculum design and production of relevant Self Learning Materials (SLMs). At the University of Kalyani a dedicated team under the able guidance of the Hon'ble Vice-Chancellor has invested its best efforts, professionally and in keeping with the demands of Post Graduate CBCS Programmes in Distance Mode to devise a self-sufficient curriculum for each course offered by the Directorate of Open and Distance Learning (DODL), University of Kalyani.

Development of printed SLMs for students admitted to the DODL within a limited time to cater to the academic requirements of the Course as per standards set by Distance Education Bureau of the University Grants Commission, New Delhi, India under Open and Distance Mode UGC Regulations, 2017 had been our endeavour. We are happy to have achieved our goal.

Utmost care and precision have been ensured in the development of the SLMs, making them useful to the learners, besides avoiding errors as far as practicable. Further suggestions from the stakeholders in this would be welcome.

During the production-process of the SLMs, the team continuously received positive stimulations and feedback from Professor (Dr.) Manas Kumar Sanyal, Hon'ble Vice-Chancellor, University of Kalyani, who kindly accorded directions, encouragements and suggestions, offered constructive criticism to develop it within proper requirements. We gracefully, acknowledge his inspiration and guidance.

Sincere gratitude is due to the respective chairpersons as well as each and every member of PGBOS (DODL), University of Kalyani. Heartfelt thank is also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani.

Their persistent and co-ordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode.

All the Self Learning Materials are self writing and collected from e-book, journals and websites.

Director
Directorate of Open and Distance Learning
University of Kalyani

Module	Unit	Content	Credit	Page No
ZDSE(MJ)T301 CytogeneticsandMolecularBiology	I	Eukaryotic chromosome organization: Packagingof DNA in eukaryoticcell;Chromatin structure; histones and non-histones; nucleosome.	2	
	II	Higher order structure of chromatin; Domains and scaffold; organization of active chromatin and assembly of Chromatin during replication.		
	III	Mapping genomes - physical maps, EST, SNPs as physical markers, radiation hybrids, FISH, optical mapping, gene maps, integration of physical and genetic maps; sequencing genomes.		
	IV	Strategies of sequencing recognition of coding and non-coding regions and annotation of genes, quality of genome- Sequence data, base calling and sequence accuracy.		
	V	Microbial genetics: organization of prokaryotic genome; single stranded DNA phages; RNA phages; cycle and gene expression in SV40 virus.		
	VI	Lytic and lysogenic phage morphogenesis in λ Phage; bacterial conjugation, transduction and transformation		

	VII	Protein folding and processing Chaperones and folding; enzymes and Protein folding., protein cleavage, glycosylation, attachment of lipids.		
	VIII	Modes of cell communications; Signaling molecules and receptors; Signal transduction and amplification; Response to signals-Gene expression, Cellular growth and metabolism		
	IX	Cell death; DNA damage and repair signaling ; Extra cellular matrix and cell signaling; Signaling crosstalk		
	X	Proteomics –Proteomes, expression analysis,Post-translational modification, 2D Electrophoresis		
	XI	Transcriptome, Transcriptome analysis, DNA microarray expression profiling, Data processing and presentation, RNA Sequencing		
	XII	Protein sequencing. Protein structure analysis,protein-protein interaction, Protein-DNA interaction		
	XIII	Signaling defects, disease and therapeutic drugs: Signaling defects in human disease-Alzheimer'sdisease, Diabetes		

		mellitus and Cysticfibrosis; Pathogens (bacteria, virus) targetof host signaling.		
	XIV	**Human disease and therapeutic drugs targeting GPCR, JAK STAT and TLR signaling pathways.**		

UNIT-I

Eukaryotic chromosome organization: Packaging of DNA in eukaryotic cell; Chromatin structure; histones and non- histones; nucleosome

UNIT-II

Higher order structure of chromatin;Domains and scaffold; organization of active chromatin and assembly of Chromatin during replication

Objective: In these two unitswe will discuss about eukaryotic chromosome organization. We will discuss about Packaging of DNA in eukaryotic cell; chromatin structure; histones and non-histones; nucleosome; higher order structure of chromatin; domains and scaffold; organization of active chromatin and assembly of chromatin during replication.

Introduction:

Eukaryotic genomes contain levels of complexity that are not encounteredin prokaryotes. In contrast to prokaryotes, most eukaryotes are diploid, having two complete sets of genes, one from each parent. Although eukaryotes haveonly about 2 to 15 times as many genes as *E. coli,* they have orders ofmagnitude more DNA. Moreover, much of this DNA does not containgenes, at least not genes encoding proteins or RNA molecules.Not only do most eukaryotes contain many times the amount of DNA in prokaryotes,but also this DNA is packaged into several chromosomes, and each chromosomeis present in two (diploids) or more (polyploids) copies. Recall that the chromosomeof*E. coli*has a contour length of 1500μm, or about 1.5 mm. Now consider that thehaploid chromosome complement, or genome, of a human contains about 1000 mmof DNA (or about 2000 mm per diploid cell). Moreover, this meter of DNA is subdividedamong 23 chromosomes of variable size and shape, with each chromosomecontaining 15 to 85 mm of DNA. In the past, geneticists had little information as tohow this DNA was arranged in the chromosomes. Is there one molecule of DNA perchromosome as in prokaryotes, or are there many? If many, how are the moleculesarranged relative to each other? How does the 85 mm (85,000 μm) of DNA in thelargest human chromosome get condensed into a mitotic metaphase structure that isabout 0.5 μm in diameter and 10 μm long?

CHEMICAL COMPOSITIONOF EUKARYOTIC CHROMOSOMES:

Interphase chromosomes are usually not visible with the light microscope. However chemical analysis, electron microscopy, and X-ray diffraction studies of isolatedchromatin (the complex of the DNA, chromosomal proteins, and other chromosomeconstituents isolated from nuclei) have provided valuable information about thestructure of eukaryotic chromosomes.When chromatin is isolated from interphase nuclei, the individual chromosomesare not recognizable. Instead, one observes an irregular aggregate of nucleoprotein.Chemical analysis of isolated chromatin shows that it consists primarily of DNA andproteins with lesser amounts of RNA (Figure 9.16). The proteins are of two majorclasses: (1) basic (positively charged at neutral pH) proteins called histones and (2) aheterogeneous, largely acidic (negatively charged at neutral pH) group of proteinscollectively referred to as non-histone chromosomal proteins.

Histones play a major structural role in chromatin. They are present in the chromatinof all eukaryotes in amounts equivalent to the amounts of DNA. This relationship suggests that an interaction occurs between histones and DNA that is conservedin eukaryotes. The histones of all plants and animals consist of five classes of proteins.

These five major histone types, called H1, H2a, H2b, H3, and H4, are present in almostall cell types. A few exceptions exist, most notably some sperm, where the histones arereplaced by another class of small basic proteins called protamines. The five histone types are present in molar ratios of approximately 1 H1:2 H2a:2H2b:2 H3:2 H4. Four of the five types of histones are specifically complexed with DNAto produce the basic structural subunits of chromatin, small (approximately 11 nm indiameter by 6.5 nm high) ellipsoidal beads called nucleosomes. The histones have beenhighly conserved during evolution—four of the five types of histone are similar in alleukaryotes.

Most of the 20 amino acids in proteins are neutral in charge; that is, they haveno charge at pH 7. However, a few are basic and a few are acidic. The histones arebasic because they contain 20 to 30 percent arginine and lysine, two positively charged amino acids. The exposed -$NH3^{+}$groups of arginine and lysineallow histones to act as polycations. The positively charged side groups on histones areimportant in their interaction with DNA, which is polyanionic because of the negativelycharged phosphate groups.

The remarkable constancy of histones H2a, H2b, H3, and H4 in all cell types ofan organism and even among widely divergent species is consistent with the idea thatthey are important in chromatin structure (DNA packaging) and are only non-specificallyinvolved in the regulation of gene expression. However, as will be discussed later, chemicalmodifications of histones can alter chromosome structure, which, in turn, can enhance ordecrease the level of expression of genes located in the modified chromatin.

In contrast, the non-histone protein fraction of chromatin consists of a largenumber of heterogeneous proteins. Moreover, the composition of the non-histonechromosomal protein fraction varies widely among different cell types of thesame organism. Thus, the non-histone chromosomal proteins probably do not playcentral roles in the packaging of DNA into chromosomes. Instead, they are likelycandidates for roles in regulating the expression of specific genes or sets of genes.

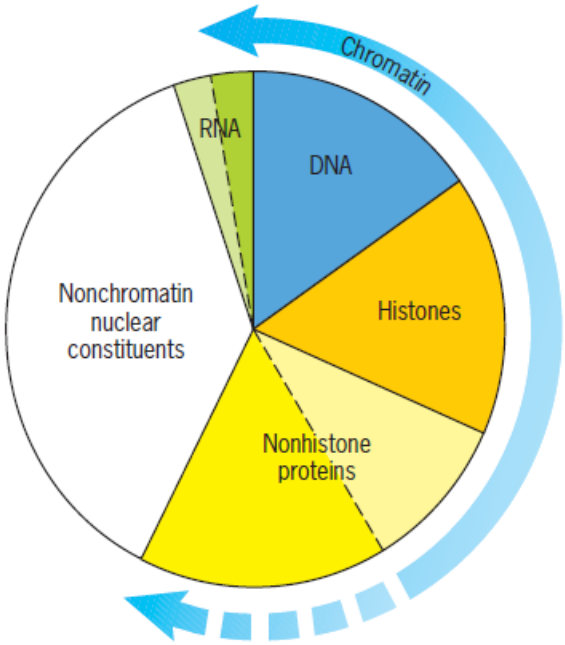


Figure : The chemical composition of chromatin as a function of the total nuclear content. The DNA and histone contents of chromatin are relatively constant, but the amount of non-histone proteins present depends on the procedure used to isolate the chromatin (dashed arrow).

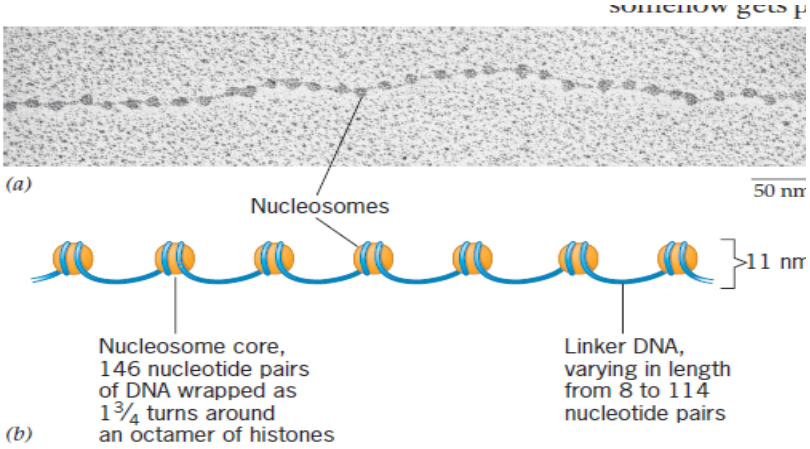
ONE LARGE DNA MOLECULE PER CHROMOSOME:

A typical eukaryotic chromosome contains 1 to 20 cm of DNA.During metaphase of meiosis and mitosis, this DNA is packaged in a chromosomewith a length of only 1 to 10 um.
Considerable evidence now indicates that each chromosome contains a single, giantmolecule of DNA that extends from one end through the centromere all the way to theother end of the chromosome. However, as we will discuss in the following section, thisgiant DNA molecule is highly condensed (coiled and folded) within the chromosome.

THREE LEVELS OF DNA PACKAGINGIN EUKARYOTIC CHROMOSOMES:

The largest chromosome in the human genome contains about 85 mm (85,000 um, or8.5 _ 107 nm) of DNA that is believed to exist as one giant molecule. This DNA moleculesomehow gets packaged into a metaphase structure that is about 0.5 um in diameter andabout 10 umin length—a condensation of almost 104-fold in lengthfrom the naked DNA molecule to the metaphase chromosome. Howdoes this condensation occur? What components of the chromosomesare involved in the packaging processes? Is there a universalpackaging scheme? Are there different levels of packaging? Clearly meiotic and mitotic chromosomes are more extensively condensedthan interphase chromosomes. What additional levels of condensationoccur in these special structures that are designed to assure theproper segregation of the genetic material during cell divisions?

Are DNA sequences of genes that are being expressed packageddifferently from those of genes that are not being expressed? Letus investigate some of the evidence that establishes the existence ofthree different levels of packaging of DNA into chromosomes.When isolated chromatin from interphase cells is examined byelectron microscopy, it is found to consist of a series of ellipsoidalbeads (about 11 nm in diameter and 6.5 nm high) joined by thinthreads (Figure 9.17*a*). Further evidence for a regular, periodic packagingof DNA has come from studies on the digestion of chromatinwith various nucleases. Partial digestion of chromatin with these nucleases yielded fragments of DNA in a set of discrete sizes that were integralmultiples of the smallest size fragment. These results are nicely explained if chromatinhas a repeating structure, supposedly the bead seen by electron microscopy(Figure 9.17*a*), within which the DNA is packaged in a nuclease-resistant form(Figure 9.17*b*). This "bead" or chromatin subunit is called the nucleosome.According to the present concept of chromatin structure, the linkers**,** or interbeadthreads of DNA, are susceptible to nuclease attack.



■ **FIGURE 9.17** Electron micrograph (*a*) and low-resolution diagram (*b*) of the beads-on-a-string nucleosome substructure of chromatin isolated from interphase nuclei. *In vivo,* the DNA linkers are probably wound between the nucleosomes forming a condensed 11-nm fiber.

Figure 9.17: Electron micrograph (a) and low-resolution diagram(b) of the beads-on-a-string nucleosome substructure of chromatinisolated from interphase nuclei. In vivo, the DNA linkers are probablywound between the nucleosomes forming a condensed 11-nm fibre.

After partial digestion of the DNA in chromatin with an endonuclease (anenzyme that cleaves DNA internally), DNA approximately 200 nucleotide pairsin length is associated with each nucleosome (produced by a cleavage in eachlinker region). After extensive nuclease digestion, a 146-nucleotide-pair-longsegment of DNA remains present in each nucleosome. This nuclease-resistantstructure is called the nucleosome core.Its structure—essentially invariant ineukaryotes—consists of a 146-nucleotide-pair length of DNA and two moleculeseach of histones H2a, H2b, H3, and H4. The histones protect the segmentof DNA in the nucleosome core from cleavage by endonucleases. Physicalstudies (X-raydiffraction and similar analyses) of nucleosome-core crystals haveshown that the DNA is wound as 1.65 turns of a superhelix around the outsideof the histone octamer (Figure 9.18*a*).
The complete chromatin subunit consists of the nucleosome core, thelinker DNA, and the associated nonhistone chromosomal proteins, all stabilizedby the binding of one molecule of histone H1 to the outside of the structure(Figure 9.18*b*).

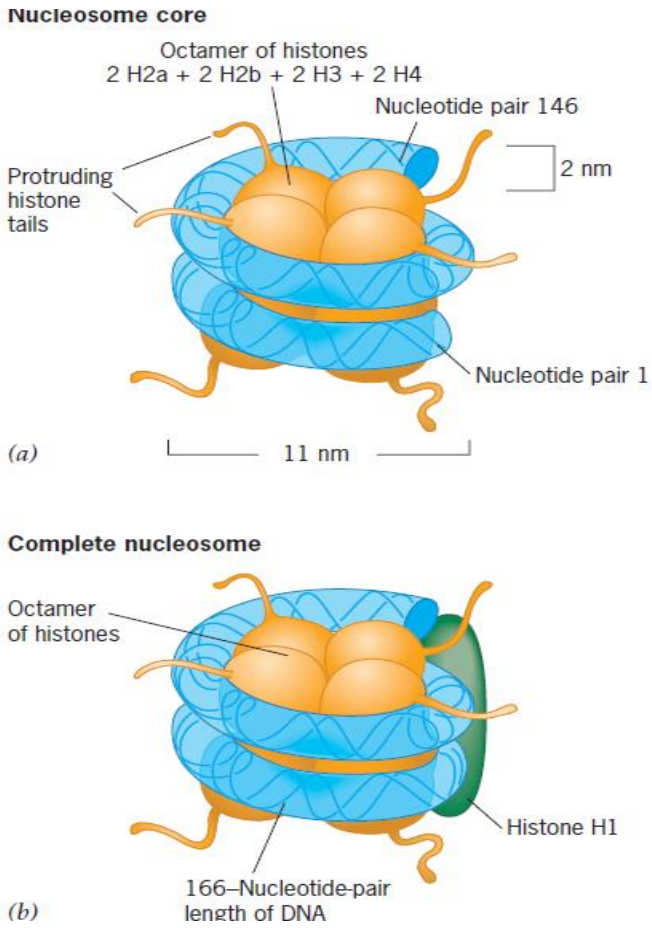


Figure 9.18 Diagrams of the gross structure of (a)the nucleosome core and (b) the complete nucleosome.The nucleosome core contains 146 nucleotide pairswound as 1.65 turns of negatively supercoiled DNAaround an octamer of histones—two molecules each ofhistones H2a, H2b, H3, and H4. The complete nucleosomecontains 166 nucleotide pairs that form almosttwosuperhelical turns of DNA around the histoneoctamer. One molecule of histone H1 is thought **tostabilize the complete nucleosome.**

The size of the linker DNA varies from species to speciesandfrom one cell type to another. Linkers as short as eight nucleotide pairs and aslong as 114 nucleotide pairs have been reported. Evidence suggests that thecomplete nucleosome (as opposed to the nucleosome core) contains two fullturns of DNA superhelix (a 166-nucleotide-pair length of DNA) on the surfaceof the histone octamer and the stabilization of this structure by the binding ofone molecule of histone H1 (Figure 9.18*b*).

The structure of the nucleosome core has been determined with resolutionto 0.28 nm by X-ray diffraction studies. The resulting high-resolution map of the nucleosome core shows the precise location of all eight histone molecules and the 146 nucleotide pairs of negatively

supercoiled DNA (Figure 9.19*a* and *b*). Some of the terminal segments of the histones pass over and between the turns ofthe DNA superhelix to add stability to the nucleosome. The interactions between the various histone molecules and between the histones and DNA are seenmost clearly in the structure of one-half of the nucleosome core (Figure 9.19*c*) which contains only 73 nucleotide pairs of supercoiled DNA.

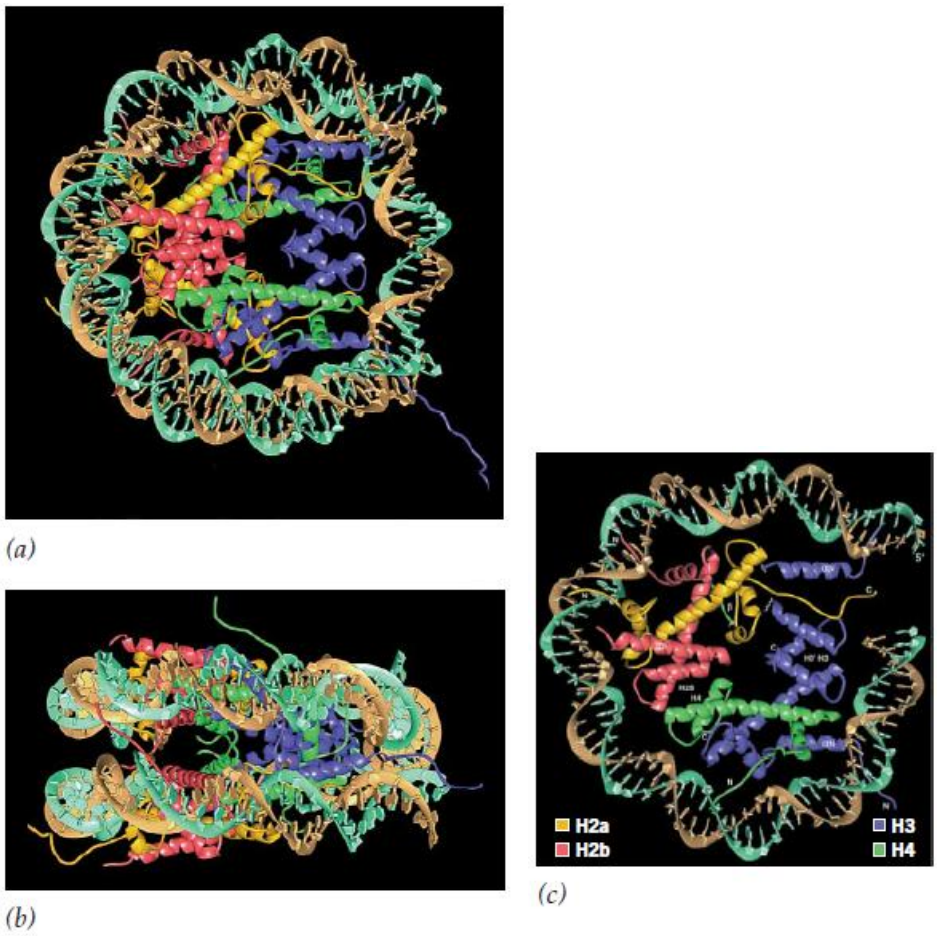


Figure 9.19: Structure of the nucleosome core basedon X-ray diffraction studies with 0.28-nm resolution. The macromolecular composition of the nucleosome core is shown looking along (a) or perpendicular to (b) the axisof the superhelix. (c) Diagram of the structure of a halfnucleosome which shows the relative positions of theDNA superhelix and the histones more clearly. Thecomplementary strands of DNA are shown in brown andgreen, and histones H2a, H2b, H3, and H4 are shown inyellow, red, blue, and green, respectively.

The basic structural component of eukaryotic chromatin is the nucleosome. The structure of nucleosomes in transcriptionally active regions of chromatin is known to differ from that of

nucleosomes in transcriptionally inactive regions. The tails of some of the histone molecules protrudefrom the nucleosome and are accessible to enzymes that add and remove chemical groups such as methyl (—CH3) and acetyl groups. The addition of these groups can change the level of expression of genes packaged in nucleosomes containing the modified histones.

Electron micrographs of isolated metaphase chromosomes show masses of tightly coiled or folded lumpy fibres (Figure 9.20). These chromatin fibres have an average diameter of 30 nm. When the structures seen by light and electronmicroscopy during earlier stages of meiosis are compared, it becomes clear thatthe light microscope simply permits one to see those regions where these 30-nmfibres are tightly packed or condensed. Indeed, when interphase chromatin is isolated using very gentle procedures, it also consists of 30-nm fibres (Figure 9.21*a*).

However, the structure of these fibres seems to be quite variable and depends on the procedures used. When observed by cryo-electron microscopy (microscopy using quickly frozen chromatin rather than fixed chromatin),the 30-nm fibres show less tightly packed "zigzag" structures(Figure 9.21*b*).

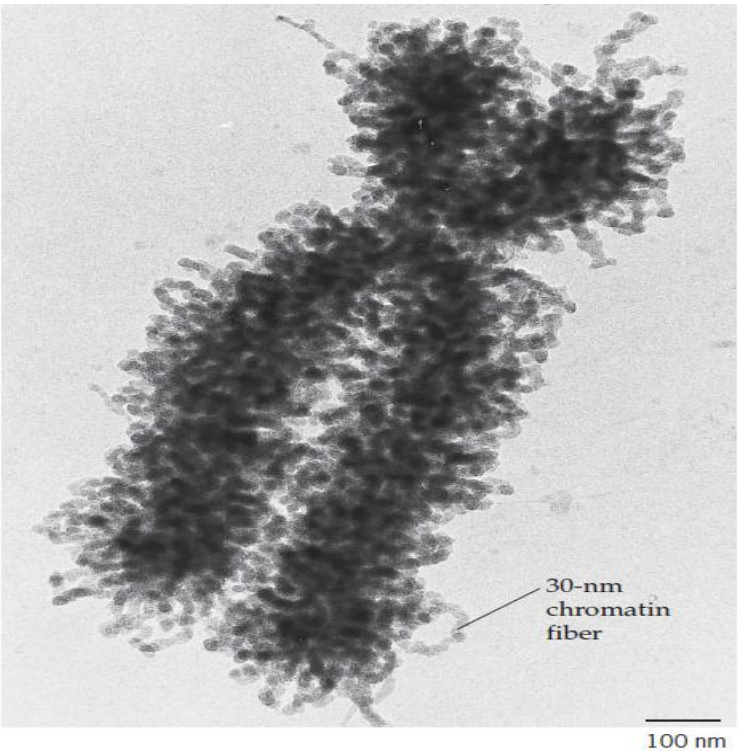


Figure 9.20 Electron micrograph of a human metaphasechromosome showing the presence of 30-nm chromatin fibres. The available evidence indicates that each chromatidcontains one large, highly coiled or folded 30-nm fibre.

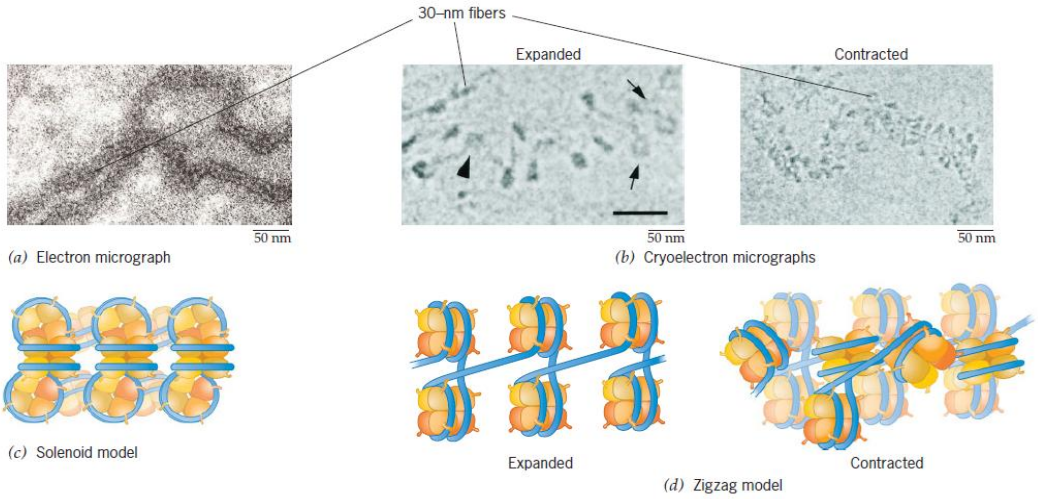


Figure 9.21 Electron micrograph (a) and cryo-electron micrographs (b) of the 30-nm chromatin fibres in eukaryotic chromosomes. The structure of 30-nm chromatin fibres seems to vary based on the procedures used to isolate and photograph them. (c) According to one popular model, the 30-nm fibre is produced by coilingthe 11-nm nucleosome fibre into a solenoid structure with six nucleosomes per turn. (d) However, when chromatin is visualized after cryopreservation (quick freezing) without fixation, it exhibits a zigzag structure whose density—expanded versus contracted—varies with ionic strength and with chemical modifications of the histone molecules.

The two most popular models are the solenoid model(Figure 9.21*c*) and the zigzag model (Figure 9.21*d*).*In vivo,* the nucleosomes clearly interact with one another to condense the 11-nm nucleosomes into 30-nm chromatin fibres. Whether these have solenoid structures or zigzag structures, or both, depending on the conditions,is still uncertain. What is certain is that chromatin structure isnot static; chromatin can expand and contract in response to chemical modifications of histone H1 and the histone tails that protrude fromthe nucleosomes.

Metaphase chromosomes are the most condensed of normaleukaryotic chromosomes. Clearly, the role of these highly condensed chromosomes is to organize and package the giant DNA molecules of eukaryotic chromosomes into structures that will facilitate their segregation to daughter nuclei without the DNA molecules of different chromosomes becoming entangled and, as a result, being broken during the anaphase separation of the daughter chromosomes. As we notedin the preceding section, the basic structural unit of the metaphase chromosome is the 30-nm chromatin fibre. However, how are these 30-nm fibres further condensed into the observed metaphase structure? Unfortunately, there is still no clear answer to this question. There is evidence that the gross structure of metaphase chromosomes is notdependent on histones. Electron micrographs of isolated metaphase

chromosomes from which the histones have been removed reveal ascaffold**,** or central core, which is surrounded by a huge pool or halo ofDNA.

This chromosome scaffold must be composed ofnonhistone chromosomal proteins. Note the absence of any apparentends of DNA molecules in the micrograph shown in Figure 9.22; this finding again supports the concept of one giant DNA molecule per chromosome.In summary, at least three levels of condensation are required to package the103 to 105 um of DNA in a eukaryotic chromosome into a metaphase structure a fewmicrons long (Figure 9.23).

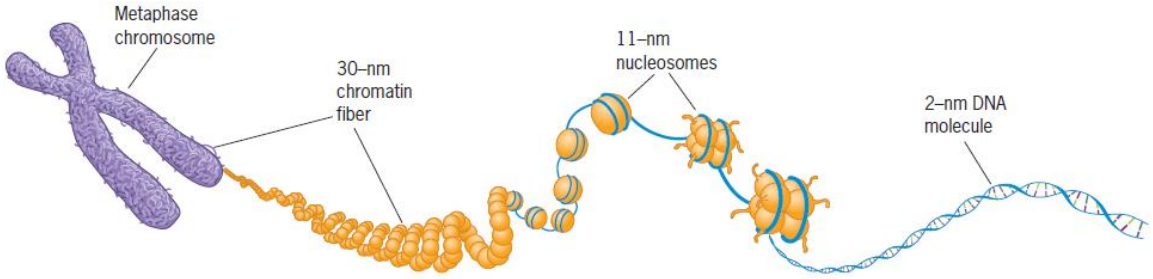


Figure 9.23. Diagram showing the different levels of DNA packaging in chromosomes. The 2-nm DNA moleculeis first condensed into 11-nm nucleosomes, which are further condensed into 30-nm chromatin fibres.The 30-nm fibres are then segregated into supercoiled domains or loops via their attachment to chromosome scaffolds composed of nonhistone chromosomal proteins.

1. The first level of condensation involves packaging DNA as a negative supercoil into nucleosomes, to produce the 11-nm-diameter interphase chromatin fibre. This clearly involves an octamer of histone molecules, two each of histones H2a, H2b,H3, and H4.

2. The second level of condensation involves an additional folding or supercoiling ofthe 11-nm nucleosome fibre, to produce the 30-nm chromatin fibre. Histone H1 is involved in this supercoiling of the 11-nm nucleosome fibre to produce the 30-nm chromatin fibre.

3. Finally, nonhistone chromosomal proteins form a scaffold that is involved in condensing the 30-nm chromatin fibre into the tightly packed metaphase chromosomes.This third level of condensation appears to involve the separation ofsegments of the giant DNA molecules present in eukaryotic chromosomes intoindependently supercoiled domains or loops. The mechanism by which this third level of condensation occurs is not known.

REPEATED DNA SEQUENCES:

The centromeres and telomeres contain DNA sequences that are repeated many times. Indeed, the chromosomes of eukaryotes contain many DNAsequences that are repeated in the haploid chromosome complement, sometimes as many as a million times. DNA containing such repeated sequences, called **repetitive DNA,** is a major component (15 to 80 percent) of eukaryotic genomes.

The first evidence for repetitive DNA came from centrifugation studies of eukaryotic DNA. When the DNA of a prokaryote, such as *E. coli*, is isolated, fragmented and centrifuged at high speeds for long periods of time in a 6*M* Cesium Chloride (CsCl) solution, the DNA will form a single band in the centrifuge tube at theposition where its density is equal to the density of the CsCl solution.For *E. coli*, this band will form at a position where the CsCl density is equal to the density of DNA containing about 50 percent A:T and 50 percent G:C base pairs.

DNA density increases with increasing G:C content. The extra hydrogen bond in a G:C base pair results in a tighter association between the bases and thus a higher density than for A:T base pairs. The centrifugation of DNAs from eukaryotes to equilibrium conditions in such CsCl solutions usually reveals the presence of one large main band of DNA and one to several small bands. These small bands of DNA are called satellite bands (from the Latin word *satelles,* meaning "an attendant" or "subordinate")and the DNAs in these bands are often referred to as satellite DNAs**.** For example, the genome of *Drosophila virilis*, a distant relative of *Drosophila melanogaster,* contains three distinct satellite DNAs, each composed of a repeating sequence of seven base pairs.

Other satellite DNAs in eukaryotes has long repetitive sequences. Much of what we know about the types of repeated DNA sequences in the chromosomes of various eukaryotic species resulted from DNA renaturation experiments. The two strands of a DNA double helix are held together by a large number of relatively weak hydrogen bonds between complementary bases. When DNA molecules in aqueous solution are heated to near 100^0C these bonds are broken and the complementary strands of DNA separate. This process is calleddenaturation.If the complementary single strands of DNA are cooled slowly underthe right conditions, the complementary base sequences will find each other andwill re-form base-paired double helices. This reformation of double helices fromthe complementary single strands of DNA is called renaturation.

If a DNA sequence is repeated many times, denaturation will yield a large numberof complementary single strands that will renature rapidly, faster than the rateof renaturation of sequences that are present only once in the genome. Indeed, therate of DNA renaturation is directly proportional to copy number (the number ofcopies of the sequence in the genome)—the higher the copy number, the faster therate and the less time required for renaturation. Mathematical analyses of the ratesof renaturation of DNA sequences in eukaryotic genomes provided strong evidencefor the presence of different classes of repeated DNA sequences, or repetitive DNA,in eukaryotic chromosomes. The recent genome

sequencing projects have providedadditional information about the different types of repetitive DNA sequences ineukaryotic genomes, and ongoing sequencing projects are providing informationabout the sequence variability that occurs in human populations. The locations of different DNA sequences inchromosomes can be determined directly by procedures similar to the renaturation experiments described here. With this procedure, called *in situ* hybridization, labelled strands of DNA form double helices with denatured DNA is still present in chromosomes.

The most highly repeated sequences in eukaryotic genomes do not encodeproteins. Indeed, they are not even transcribed. Other less repetitive sequences encode proteins, such as ribosomal proteins and the muscle proteins actin and myosin that are needed in large amounts and are each encoded by several genes.The genes that specify ribosomal RNAs are also multicopy genes because cellsneed large amounts of ribosomal RNA to produce the ribosomes required for protein synthesis.

The most prevalent of the repeated DNA sequences are transposable genetic elements**,** DNA sequences that can move from one location in a chromosome toanother or even to a different chromosome, or inactive sequences derived from transposable elements. In *D. melanogaster,* about 90 different families of transposable elements have been characterized and been given interesting namessuch as *hobo, pogo,* and *gypsy* that suggest their mobility. A much larger proportion—between 40 and 50 percent—of the human genome contains transposable elements orsequences derived from them. As much as 80 percent of the corn genome may consistof transposable genetic elements or their derivatives.

DUPLICATION OF NUCLEOSOMES AT REPLICATION FORKS:

The DNA in eukaryotic interphase chromosomes is packaged in approximately 11-nm beads called nucleosomes. Each nucleosome contains 166 nucleotide pairs of DNA wound in two turns around an octamer of histone molecules. Given the size of nucleosomes and the large size of DNA replisomes, it seems unlikely that a replication fork can move past an intact nucleosome. Yet, electron micrographs of replicating chromatin in *Drosophila* clearly show nucleosomes with approximately normal structure and spacing on both sides of replication forks (Figure 10.33*a*); that is, nucleosomes appear to have the same structure and spacing immediately behind a replication fork (post replicative DNA) as they do in front of a replication fork (prereplicative DNA). This observation suggests that nucleosomes must be disassembled to let the replisome duplicate the DNA packaged in them and then be quickly reassembled; thatis, DNA replication and nucleosome assembly must be tightly coupled. Since the mass of the histones in nucleosomes is equivalent to that of the DNA, large quantities of histones must be synthesized during each cell generation in order for the nucleosomes to duplicate. Although histone synthesis occurs throughout the cell cycle, there is a burst of histone biosynthesis during S phase that generates enough histones for

chromatin duplication. When density-transfer experiments were performed to examine the mode of nucleosome duplication, the nucleosomes on both progeny DNA molecules were found to contain both old (prereplicative) histone complexes and new (post-replicative) complexes. Thus, at the protein level, nucleosome duplication appears to occur by a dispersive mechanism.

A number of proteins are involved in the disassembly and assembly of nucleosomes during chromosome replication in eukaryotes. Two of the most important are *n*ucleosome *a*ssembly *p*rotein-*1* (Nap-1) and *c*hromatin *a*ssembly *f*actor-*1* (CAF-1). Nap-1 transports histones from their site of synthesis in the cytoplasm to the nucleus, and CAF-1 carries them to the chromosomal sites of nucleosome assembly (Figure 10.33*b*). CAF-1 delivers histones to the sites of DNA replication by binding to PNCA (*p*roliferating *c*ell *n*uclear *a*ntigen)—the clamp that tethers DNA polymerase to the DNA template (Figure 10.32). CAF-1 is an essential protein in *Drosophila,* but not in yeast where other proteins can perform some of its functions.

Many other proteins affect nucleosome structure. Some are involved in chromatin remodeling—changing nucleosome structure in ways that activate or silence the expression of the genes packaged therein. Others modify nucleosome structure by adding methyl or acetyl groups to specific histones. In addition, eukaryotes contain several minor histones with structures slightly different from the major histones, and the incorporation of these minor histones into nucleosomes can change their structure. In *Drosophila,* for example, the incorporation of histone H3.3 into nucleosomes results in high levels of transcription of the genes therein. Thus, nucleosome structure is not invariant; to the contrary, it plays an important role in modulating gene expression.

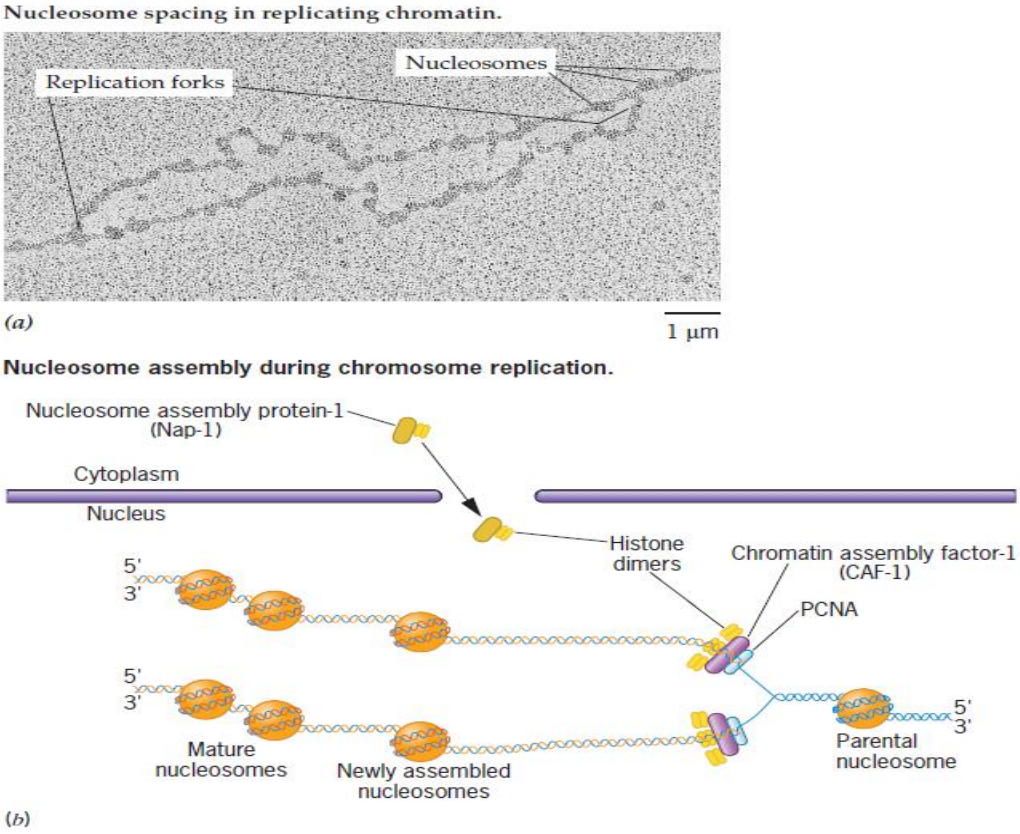


Figure 10.33 The disassembly and assembly of nucleosomes during the replicationof chromosomes in eukaryotes. (a) An electron micrograph showing nucleosomes onboth sides of two replication forks in Drosophila. Recall that DNA replication is bidirectional;thus, each branch point is a replication fork. (b) The assembly of new nucleosomesduring chromosome replication requires proteins that transport histones fromthe cytoplasm to the nucleus and that concentrate them at the site of nucleosomeassembly. PCNA - proliferating cell nuclear antigen.

Assembly of Chromatin in replication:

The duplication of a chromosome requires replication of the DNA and the reassembly of the associated proteins on each daughter DNA molecule. The latter process is tightly linked to DNA replication to ensure that the newly replicated DNA is rapidly packaged into nucleosomes. Although the replication of DNA requires the nucleosome disassembly, the DNA is rapidly repackaged into nucleosomes in an ordered series of events. The first step in the assembly of a nucleosome is the binding of an H3.H4 tetramer to the DNA. Once the tetramer is bound, two H2A.H2B dimers associate to form the final nucleosome. H1 joins this

complex last, presumably during the formation of higher-order chromatin assemblies.To duplicate a chromosome, at least half of the nucleosomes on the daughter chromosomes must be newly synthesized. The fate of the old histones is a particularly important issue given the effects that histone modification can have on the accessibility of the resulting chromatin. If the old histones were lost completely, then chromosome duplication would erase any "memory" of the previously modified nucleosomes. In contrast, if the old histones were retained on a single chromosome, that chromosome would have a distinct set of modifications relative to the other copy of the chromosome. Mixing is not entirely random,however. H3.H4 tetramers and H2A.H2B dimers are composed of either all new or all old histones. Thus, as the replication fork passes, nucleosomes are broken down into their component subassemblies. H3.H4 tetramers appear to remain bound to one of the two daughter duplexes at random and are never released from DNA into the free pool of histones. In contrast, the H2A.H2B dimers are released and enter the local pool, available for new nucleosome assembly.

The distributive inheritance of old histones during chromosome duplication provides a mechanism for the propagation of the parental pattern of histone modification. By this mechanism, old modified histones will tend to rebind one of the daughter chromosomes at a position near their previous position on the parental chromosome. The old histones have an equal probability of binding either daughter chromosome. This localized inheritance of modified histones ensures that a subset of the modified histones is located in similar positions on each daughter chromosome

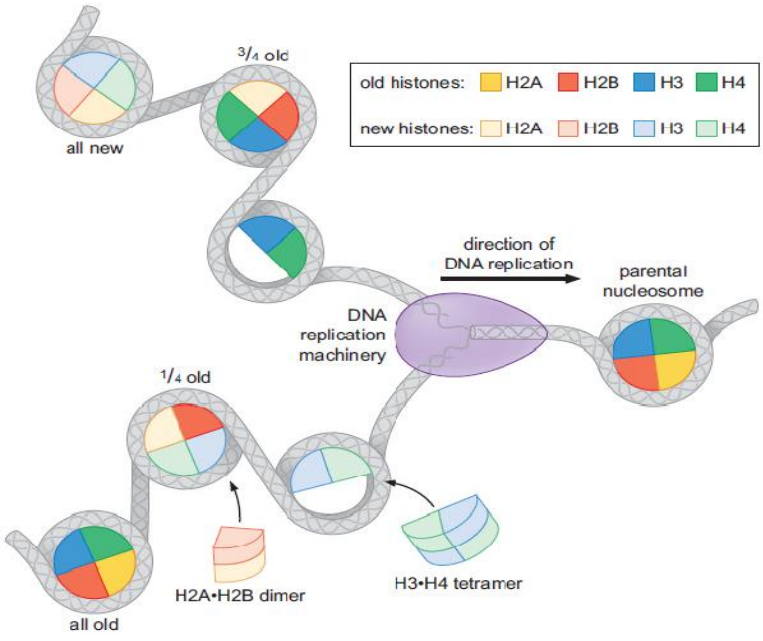


Figure: Inheritance of histone and assembly of chromatin during replication.

Probable questions:

1. Write down the chemical composition of Eukaryotic chromosome.
2. How DNA are packed into chromosomes.
3. Name histone proteins which took part in nucleosome formation.
4. What is scaffold?
5. What are the role of histone and non histone proteins in DNA packaging?
6. How duplication of nucleosomes occur at replication forks?
7. What do you know about repeated DNA sequences?

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

UNIT-III

Mapping genomes - physical maps, EST, SNPs as physical markers, radiation hybrids, FISH, optical mapping, gene maps, integration of physical and genetic maps; sequencing genomes.

UNIT-IV

High-throughput sequencing, strategies of sequencing, recognition of coding and non-coding regions and annotation of genes, quality of genome-sequence data, base calling and sequence accuracy.

Objective: In this unit you will learn physical maps, EST, SNPs as physical markers, radiation hybrids, FISH, optical mapping, gene maps, integration of physical and genetic maps; sequencing genomes: high-throughput sequencing, strategies of sequencing, recognition of coding and non-coding regions and annotation of genes, quality of genome-sequence data, base calling and sequence accuracy

Gene mappingdescribes the methods used to identify the locus of a gene and the distances between genes.The essence of all genome mapping is to place a collection of molecular markers onto their respective positions on the genome. Molecular markers come in all forms. Genes can be viewed as one special type of genetic markers in the construction of genome maps, and mapped the same way as any other markers.

Genetic and Physical Maps

The convention is to divide genome mapping methods into two categories.

- Genetic mapping is based on the use of genetic techniques to construct maps showing the positions of genes and other sequence features on a genome. Genetic techniques include cross-breeding experiments or, in the case of humans, the examination of family histories (pedigrees).
- Physical mapping uses molecular biology techniques to examine DNA molecules directly in order to construct maps showing the positions of sequence features, including genes.

Genetic Mapping

As with any type of map, a genetic map must show the positions of distinctive features. In a geographic map these markers are recognizable components of the landscape, such as rivers, roads and buildings. What markers can we use in a genetic landscape?

Genes were the first markers to be used

The first genetic maps, constructed in the early decades of the 20th century for organisms such as the fruit fly, used genes as markers. This was many years before it was understood that genes are segments of DNA molecules. Instead, genes were looked upon as abstract entities responsible for the transmission of heritable characteristics from parent to offspring. To be useful in genetic analysis, a heritable characteristic has to exist in at least two alternative forms or phenotypes, an example being tall or short stems in the pea plants originally studied by Mendel. Each phenotype is specified by a different allele of the corresponding gene. To begin with, the only genes that could be studied were those specifying phenotypes that were distinguishable by visual examination. So, for example, the first fruit-fly maps showed the positions of genes for body color, eye color, wing shape and suchlike, all of these phenotypes being visible simply by looking at the flies with a low-power microscope or the naked eye. This approach was fine in the early days but geneticists soon realized that there were only a limited number of visual phenotypes whose inheritance could be studied, and in many cases their analysis was complicated because a single phenotype could be affected by more than one gene. For example, by 1922 over 50 genes had been mapped onto the four fruit-fly chromosomes, but nine of these were for eye color; in later research, geneticists studying fruit flies had to learn to distinguish between fly eyes that were colored red, light red, vermilion, garnet, carnation, cinnabar, ruby, sepia, scarlet, pink, cardinal, claret, purple or brown. To make gene maps more comprehensive it would be necessary to find characteristics that were more distinctive and less complex than visual ones.

The answer was to use biochemistry to distinguish phenotypes. This has been particularly important with two types of organisms - microbes and humans. Microbes, such as bacteria and yeast, have very few visual characteristics so gene mapping with these organisms has to rely on biochemical phenotypes such as those listed in With humans it is possible to use visual characteristics, but since the 1920s studies of human genetic variation have been based largely on biochemical phenotypes that can be scored by blood typing. These phenotypes

include not only the standard blood groups such as the ABO series (Yamamoto et al., 1990), but also variants of blood serum proteins and of immunological proteins such as the human leukocyte antigens (the HLA system). A big advantage of these markers is that many of the relevant genes have multiple alleles. For example, the gene called *HLA-DRB1* has at least 290 alleles and *HLA-B* has over 400. This is relevant because of the way in which gene mapping is carried out with humans. Rather than setting up many breeding experiments, which is the procedure with experimental organisms such as fruit flies or mice, data on inheritance of human genes have to be gleaned by examining the phenotypes displayed by members of a single family. If all the family members have the same allele for the gene being studied then no useful information can be obtained. It is therefore necessary for the relevant marriages to have occurred, by chance, between individuals with different alleles. This is much more likely if the gene being studied has 290 rather than two alleles.

DNA markers for genetic mapping

Genes are very useful markers but they are by no means ideal. One problem, especially with larger genomes such as those of vertebrates and flowering plants, is that a map based entirely on genes is not very detailed. This would be true even if every gene could be mapped because, as we saw in Chapter 2, in most eukaryotic genomes the genes are widely spaced out with large gaps between them (see Figure). The problem is made worse by the fact that only a fraction of the total number of genes exist in allelic forms that can be distinguished conveniently. Gene maps are therefore not very comprehensive. We need other types of marker.

Mapped features that are not genes are called DNA markers. As with gene markers, a DNA marker must have at least two alleles to be useful. There are three Types of DNA sequence feature that satisfy this requirement: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), and single nucleotide polymorphisms (SNPs). Discussed underneath is the SNPs which is a comparatively modern technique used in current times:

Single nucleotide polymorphisms (SNPs)

These are positions in a genome where some individuals have one nucleotide (e.g. a G) and others have a different nucleotide (e.g. a C) (Figure). There are vast numbers of SNPs in every genome, some of which also give rise to RFLPs, but many of which do not because the sequence in which they lie is not recognized

by any restriction enzyme. In the human genome there are at least 1.42 million SNPs, only 100 000 of which result in an RFLP (SNP Group,2001).

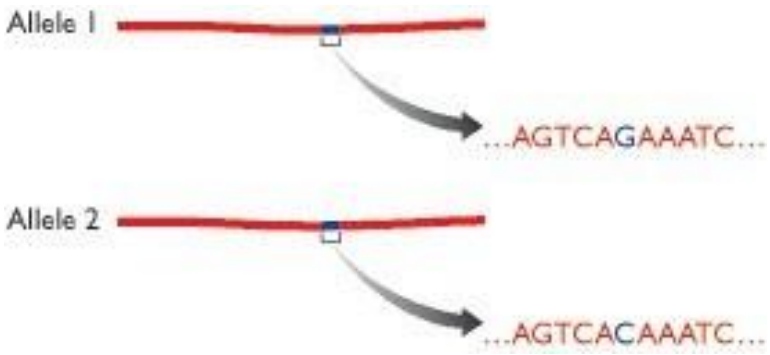
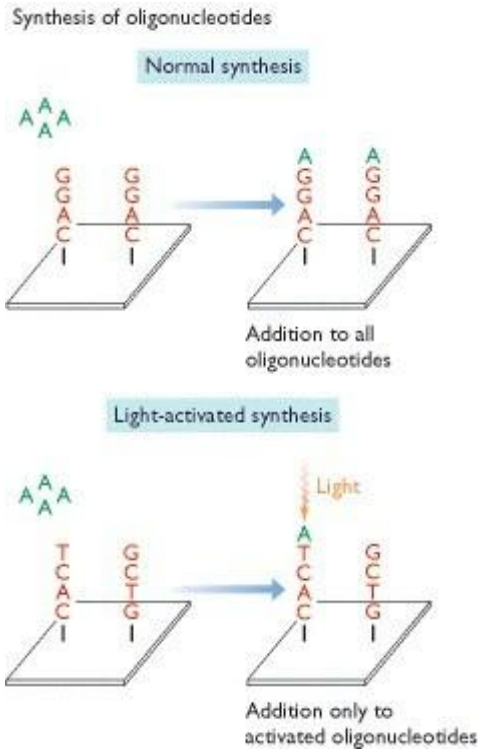


Figure : A single nucleotide polymorphism (SNP)

Although each SNP could, potentially, have four alleles (because there are four nucleotides), most exist in just two forms, so these markers suffer from the same drawback as RFLPs with regard to human genetic mapping: there is a high possibility that a SNP does not display any variability in the family that is being studied. The advantages of SNPs are their abundant numbers and the fact that they can be typed by methods that do not involve gel electrophoresis. This is important because gel electrophoresis has proved difficult to automate so any detection method that uses it will be relatively slow and labor-intensive. SNP detection is more rapid because it is based on oligonucleotide hybridization analysis. An oligonucleotide is a short single-stranded DNA molecule, usually less than 50 nucleotides in length,that is synthesized in the test tube. If the conditions are just right, then an oligonucleotide will hybridize with another DNA molecule only if the oligonucleotide forms a completely base-paired structure with the second molecule. If there is a single mismatch - a single position within the oligonucleotide that does not form a base pair - then hybridization does not occur (Figure below). Oligonucleotide hybridization can therefore discriminate between the two alleles of an SNP. Various screening strategies have been devised (Mir and Southern, 2000), including DNA chip technology and solution hybridization technique



A DNA chip is a wafer of glass or silicon, 2.0 cm^2 or less in area, carrying many different oligonucleotides in a high-density array. The DNA to be tested is labeled with a fluorescent marker and pipetted onto the surface of the chip. Hybridization is detected by examining the chip with a fluorescence microscope, the positions at which the fluorescent signal is emitted indicating which oligonucleotides have hybridized with the test DNA. Many SNPs can therefore be scored in a single experiment (Wang et al., 1998; Gerhold et al.,1999).

Solution hybridization techniques are carried out in the wells of a microtiter tray, each well containing a different oligonucleotide, and use a detection system that can discriminate between unhybridized single-stranded DNA and the double-stranded product that results when an oligonucleotide hybridizes to the test DNA. Several systems have been developed, one of which makes use of a pair of labels comprising a fluorescent dye and a compound that quenches the fluorescent signal when brought into close proximity with the dye. The dye is attached to one end of an oligonucleotide and the quenching compound to the other end. Normally there is no fluorescence because the oligonucleotide is designed in such a way that the two ends base-pair to one another, placing the quencher next to the dye (Figure 5.9). Hybridization between oligonucleotide and test DNA disrupts this base pairing, moving the quencher away from the dye and enabling the fluorescent signal to be generated (Tyagi et al.,1998).

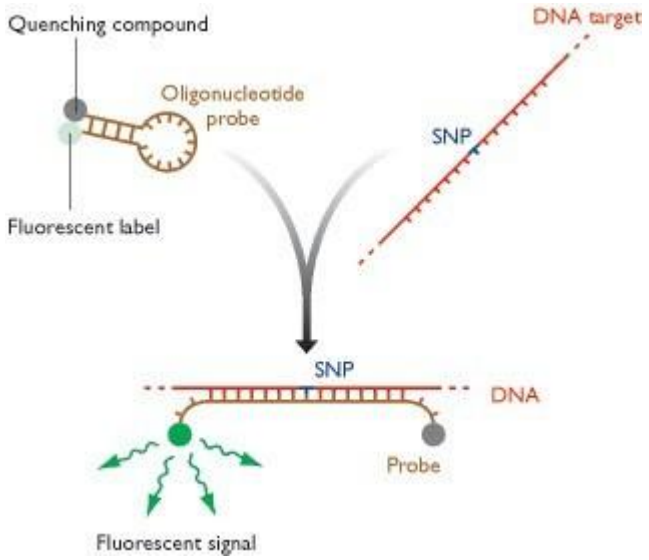


Figure: One way of detecting an SNP by solution hybridization.

Linkage analysis is the basis of genetic mapping

Now that we have assembled a set of markers with which to construct a genetic map we can move on to look at the mapping techniques themselves. These techniques are all based on genetic linkage, which in turn derives from the seminal discoveries in genetics made in the mid 19th century by Gregor Mendel.

Working out a genetic map from recombination frequencies

Once Morgan had understood how partial linkage could be explained by crossing-over during meiosis he was able to devise a way of mapping the relative positions of genes on a chromosome. In fact the most important work was done not by Morgan himself, but by an undergraduate in his laboratory, Arthur Sturtevant (Sturtevant, 1913). Sturtevant assumed that crossing-over was a random event, there being an equalchance of it occurring at any position along a pair of lined-up chromatids. If this assumption is correct then two genes that are close together will be separated by crossovers less frequently than two genes that are more distant from one another. Furthermore, the frequency with which the genes are unlinked by crossovers will be directly proportional to how far apart they are on their chromosome. The recombination frequency is therefore a measure of the distance between two genes. If you work out the

recombination frequencies for different pairs of genes, you can construct a map of their relative positions on the chromosome.

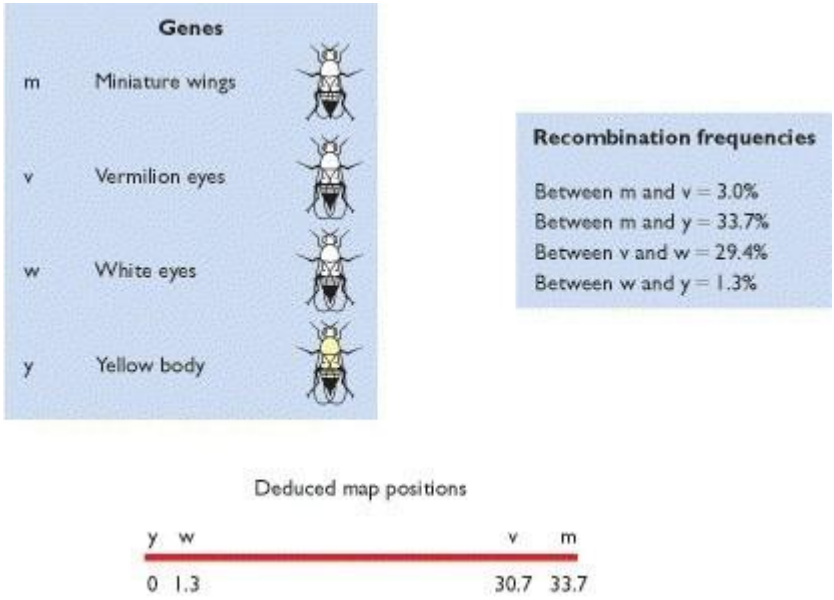


Figure: Working out a genetic map from recombination frequencies. The example is taken from the original experiments carried out with fruit flies by Arthur Sturtevant. All four genes are on the X chromosome of the fruit fly.

It turns out that Sturtevant's assumption about the randomness of crossovers was not entirely justified. Comparisons between genetic maps and the actual positions of genes on DNA molecules, as revealed by physical mapping and DNA sequencing, have shown that some regions of chromosomes, called recombination hotspots, are more likely to be involved in crossovers than others. This means that a genetic map distance does not necessarily indicate the physical distance between two markers (see Figure). Also, we now realize that a single chromatid can participate in more than one crossover at the same time, but that there are limitations on how close together these crossovers can be, leading to more inaccuracies in the mapping procedure. Despite these qualifications, linkage analysis usually makes correct deductions about gene order, and distance estimates are sufficiently accurate to generate genetic maps that are of value as frameworks for genome sequencing projects.

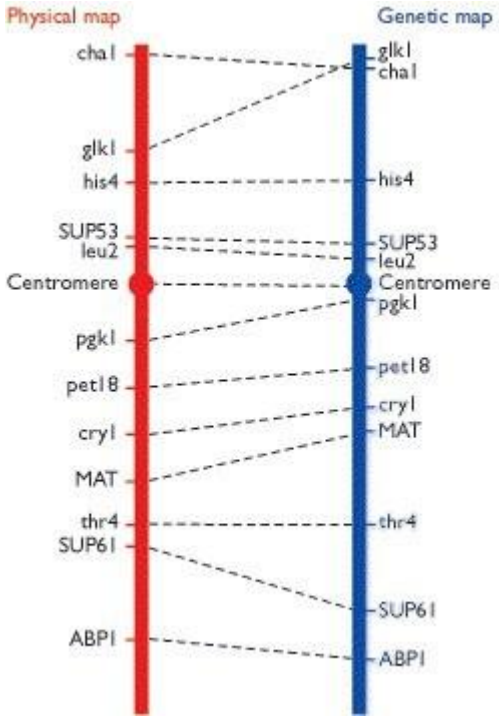


Figure: Comparison between the genetic and physical maps of *Saccharomyces cerevisiae* chromosome III. The comparison shows the discrepancies between the genetic and physical maps, the latter determined by DNA sequencing.

Physical Mapping

A map generated by genetic techniques is rarely sufficient for directing the sequencing phase of a genome project. This is for two reasons:

- **The resolution of a genetic map depends on the number of crossovers that have been scored .**This is not a major problem for microorganisms because these can be obtained in huge numbers, enabling many crossovers to be studied, resulting in a highly detailed genetic map in which the markers are just a few kb apart. For example, when the *Escherichia coli* genome sequencing project began in 1990, the latest genetic map for this organism comprised over 1400 markers, an average of one per 3.3 kb. This was sufficiently detailed to direct the sequencing program without the need for extensive physical mapping. Similarly, the *Saccharomyces cerevisiae* project was supported

by a fine-scale genetic map (approximately 1150 genetic markers, on average one per 10 kb). The problem with humans and most other eukaryotes is that it is simply not possible to obtain large numbers of progeny, so relatively few meiosis can be studied and the resolving power of linkage analysis is restricted. This means that genes that are several tens of kb apart may appear at the same position on the genetic map.

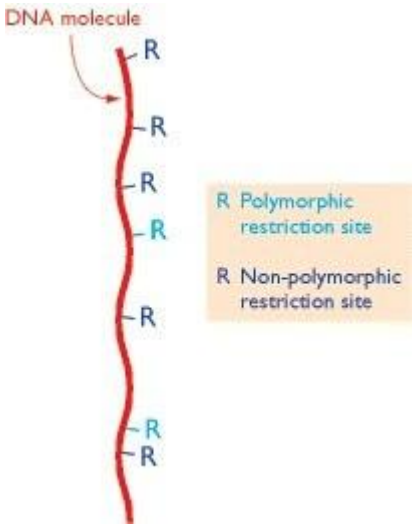
- **Genetic maps have limited accuracy .**Sturtevant's assumption that crossovers occur at random along chromosomes is only partly correct because the presence of recombination hotspots means that crossovers are more likely to occur at some points rather than at others. The effect that this can have on the accuracy of a genetic map was illustrated in 1992 when the complete sequence for S*. cerevisiae* chromosome III was published (Oliver et al., 1992), enabling the first direct comparison to be made between a genetic map and the actual positions of markers as shown by DNA sequencing (Figure). There were considerable discrepancies, even to the extent that one pair of genes had been ordered incorrectly by genetic analysis. Bear in mind that *S. cerevisiae* is one of the two eukaryotes (fruit fly is the second) whose genomes have been subjected to intensive genetic mapping. If the yeast genetic map is inaccurate then how precise are the genetic maps of organisms subjected to less detailed analysis?

These two limitations of genetic mapping mean that for most eukaryotes a genetic map must be checked and supplemented by alternative mapping procedures before large-scale DNA sequencing begins. A plethora of physical mapping techniques has been developed to address this problem, the most important being:

- Restriction mapping, which locates the relative positions on a DNA molecule of the recognition sequences for restriction endonucleases;
- **Fluorescent *in situ* hybridization (**FISH**)**, in which marker locations are mapped by hybridizing a probe containing the marker to intact chromosomes;
- Sequence tagged site **(**STS**) mapping**, in which the positions of short sequences are mapped by PCR and/or hybridization analysis of genome fragments.

Restriction mapping

Genetic mapping using RFLPs as DNA markers can locate the positions of polymorphic restriction sites within a genome, but very few of the restriction sites in a genome are polymorphic, so many sites are not mapped by this technique (Figure below). Could we increase the marker density on a genome map by using an alternative method to locate the positions of some of the non-polymorphic restriction sites? This is what restriction mapping achieves, although in practice the technique has limitations which mean that it is applicable only to relatively small DNA molecules. We will look first at the technique and then consider its relevance to genome mapping.



The basic methodology for restriction mapping:

The simplest way to construct a restriction map is to compare the fragment sizes produced when a DNA molecule is digested with two different restriction enzymes that recognize different target sequences. An example using the restriction enzymes *Eco*RI and *Bam*HI is shown in Figure. First, the DNA molecule is digested with just one of the enzymes and the sizes of the resulting fragments are measured by agarose gel electrophoresis. Next, the molecule is digested with the second enzyme and the resulting fragments again sized in an agarose gel. The results so far enable the number of restriction sites for each enzyme to be worked out, but do not allow their relative positions to be determined. Additional information is therefore obtained by cutting the DNA molecule with both enzymes together. In the example shown in Figure below this double restriction enables three of the sites to be mapped. However, a

problem arises with the larger *Eco*RI fragment because this contains two *Bam*HI sites and there are two alternative possibilities for the map location of the outer one of these. The problem is solved by going back to the original DNA molecule and treating it again with *Bam*HI on its own, but this time preventing the digestion from going to completion by, for example, incubating the reaction for only a short time or using a suboptimal incubation temperature. This is called a partial restriction and leads to a more complex set of products, the complete restriction products now being supplemented with partially restricted fragments that still contain one or more uncut *Bam*HI sites. In the example shown in Figure, the size of one of the partial restriction fragments is diagnostic and the correct map can be identified.

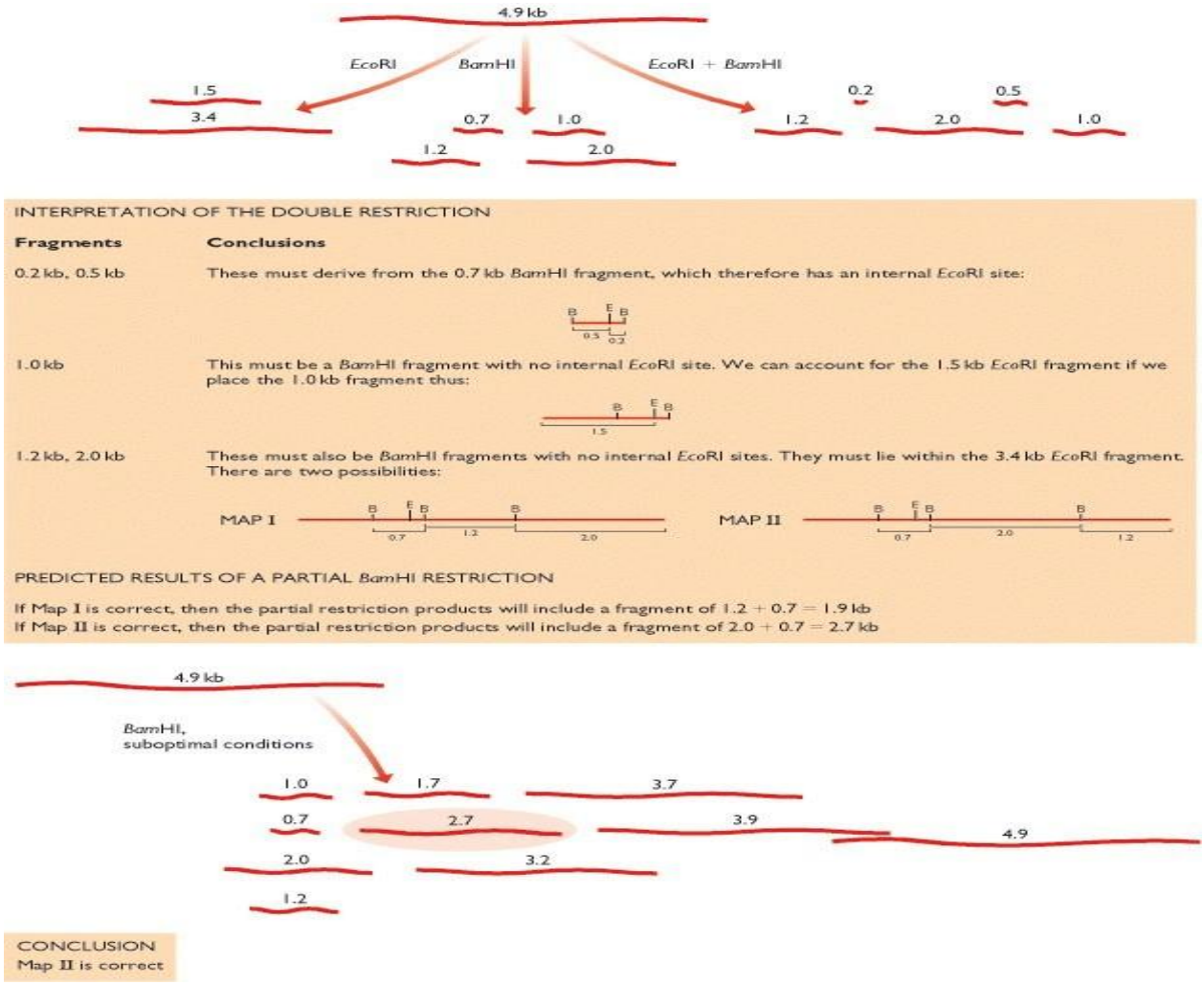


Figure : Restriction mapping. The objective is to map the EcoRI (E) and BamHI (B) sites in a linear DNA molecule of 4.9 kb. The results of single and double restrictions are shown at the top. The sizes of the fragments given after double restriction enable two alternative maps to be

constructed, as explained in the central panel, the unresolved issue being the position of one of the three BamHI sites. The two maps are tested by a partial BamHI restriction (bottom), which shows that Map II is the correct one.

A partial restriction usually gives the information needed to complete a map, but if there are many restriction sites then this type of analysis becomes unwieldy, simply because there are so many different fragments to consider. An alternative strategy is simpler because it enables the majority of the fragments to be ignored. This is achieved by attaching a radioactive or other type of marker to each end of the starting DNA molecule before carrying out the partial digestion. The result is that many of the partial restriction products become 'invisible' because they do not contain an end-fragment and so do not show up when the agarose gel is screened for labeled products. The sizes of the partial restriction products that are visible enable unmapped sites to be positioned relative to the ends of the startingmolecule.

The scale of restriction mapping is limited by the sizes of the restriction fragments

Restriction maps are easy to generate if there are relatively few cut sites for the enzymes being used. However, as the number of cut sites increases, so also do the numbers of single, double and partial restriction products whose sizes must be determined and compared in order for the map to be constructed. Computer analysis can be brought into play but problems still eventually arise. A stage will be reached when a digest contains so many fragments that individual bands merge on the agarose gel, increasing the chances of one or more fragments being measured incorrectly or missed out entirely. If several fragments have similar sizes then even if they can all be identified, it may not be possible to assemble them into an unambiguous map.

Restriction mapping is therefore more applicable to small rather than large molecules, with the upper limit for the technique depending on the frequency of the restriction sites in the molecule being mapped. In practice, if a DNA molecule is less than 50 kb in length it is usually possible to construct an unambiguous restriction map for a selection of enzymes with six-nucleotide recognition sequences. Fifty kb is of course way below the minimum size for bacterial or eukaryotic chromosomes, although it does cover a few viral and organelle genomes, and whole- genome restriction maps have indeed been important in directing sequencing projects with these small molecules. Restriction maps are equally useful after bacterial or eukaryotic genomic DNA has been cloned, if the cloned fragments are less than 50 kb, because a detailed restriction map can

then be built up as a preliminary to sequencing the cloned region. This is an important application of restriction mapping in sequencing projects with large genomes, but is there any possibility of using restriction analysis for the more general mapping of entire genomes larger than 50 kb?

The answer is a qualified 'yes', because the limitations of restriction mapping can be eased slightly by choosing enzymes expected to have infrequent cut sites in the target DNA molecule. These 'rare cutters' fall into two categories:

- **Enzymes with seven- or eight-nucleotide recognition sequences.** A few restriction enzymes cut at seven- or eight-nucleotide recognition sequences. Examples are SapI (5′- GCTCTTC-3′) and SgfI (5′-GCGATCGC-3′). The seven-nucleotide enzymes would be expected, on average, to cut a DNA molecule with a GC content of 50% once every 4^7 = 16 384 bp. The eight-nucleotide enzymes should cut once every 4^8 = 65 536 bp. These figures compare with 4^6 = 4096 bp for six-nucleotide enzymes such as BamHI and EcoRI. Seven- and eight-nucleotide cutters are often used in restriction mapping of large molecules but the approach is not as useful as it might be simply because not many of these enzymes are known.
- **Enzymes whose recognition sequences contain motifs that are rare in the target** DNA **.** Genomic DNA molecules do not have random sequences and some are significantly deficient in certain motifs. For example, the sequence 5′-CG-3′ is rare in human DNA because human cells possess an enzyme that adds a methyl group to carbon 5 of the C nucleotide in this sequence. The resulting 5-methylcytosine is unstable and tends to undergo deamination to give thymine (Figure below). The consequence is that during human evolution many of the 5′-CG-3′ sequences that were originally in our genome have become converted to 5′-TG-3′. Restriction enzymes that recognize a site containing 5′-CG-3′ therefore cut human DNA relatively infrequently. Examples are SmaI (5′-CCCGGG-3′), which cuts human DNA on average once every 78 kb, and BssHII (5′- GCGCGC-3′) which cuts once every 390 kb. Note that NotI, an eight-nucleotide cutter, also targets 5′-CG-3′ sequences (recognition sequence 5′-GCGGCCGC-3′) and cuts human DNA very rarely - approximately once every 10Mb.

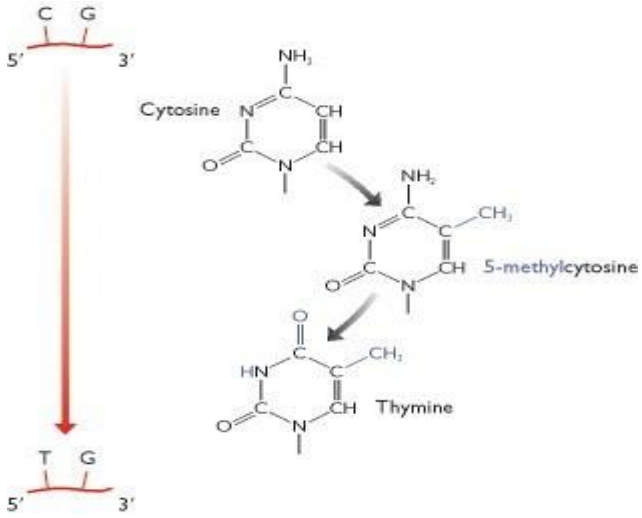


Figure: The sequence 5′-CG-3′ is rare in human DNA because of methylation of the C, followed by deamination to give T.

The potential of restriction mapping is therefore increased by using rare cutters. It is still not possible to construct restriction maps of the genomes of animals and plants, but it is feasible to use the technique with large cloned fragments, and the smaller DNA molecules of prokaryotes and lower eukaryotes such as yeast and fungi.

If a rare cutter is used then it may be necessary to employ a special type of agarose gel electrophoresis to study the resulting restriction fragments. This is because the relationship between the length of a DNA molecule and its migration rate in an electrophoresis gel is not linear, the resolution decreasing as the molecules get longer (Figure A). This means that it is not possible to separate molecules more than about 50 kb in length because all of these longer molecules run as a single slowly migrating band in a standard agarose gel. To separate them it is necessary to replace the linear electric field used in conventional gel electrophoresis with a more complex field. An example is provided by orthogonal field alternation gel electrophoresis (OFAGE), in which the electric field alternates between two pairs of electrodes, each positioned at an angle of 45° to the length of the gel (Figure B). The DNA molecules still move down through the gel, but each change in the field forces the molecules to realign. Shorter molecules realign more quickly than longer ones and so migrate more rapidly through the gel. The overall result is that molecules much longer than those separated by conventional gel electrophoresis can be resolved. Related techniques include CHEF (contour clamped homogeneous electric fields) and FIGE (field inversion gel electrophoresis).

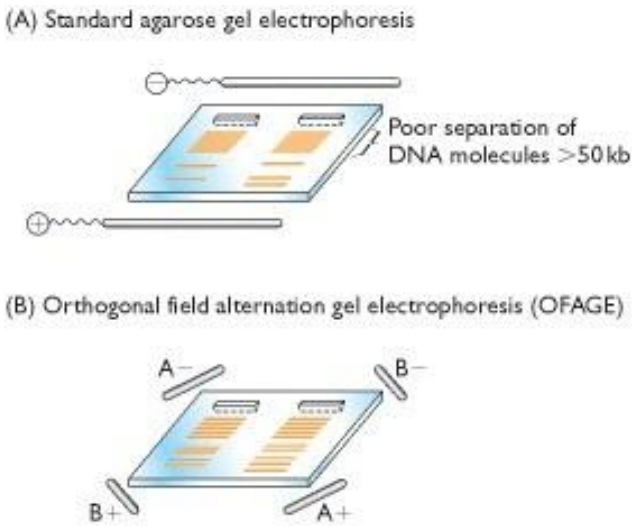


Figure : Conventional and non-conventional agarose gel electrophoresis. (A) In standard agarose gel electrophoresis the electrodes are placed at either end of the gel and the DNA molecules migrate directly towards the positive electrode. Molecules longer than about 50 kb cannot be separated from one another in this way. (B) In OFAGE, the electrodes are placed at the corners of the gel, with the field pulsing between the A pair and the B pair.

Optical mapping:

It is also possible to use methods other than electrophoresis to map restriction sites in DNA molecules. With the technique called optical mapping (Schwartz et al., 1993), restriction sites are directly located by looking at the cut DNA molecules with a microscope. The DNA must first be attached to a glass slide in such a way that the individual molecules become stretched out, rather than clumped together in a mass. There are two ways of doing this: gel stretching and molecular combing. To prepare gel-stretched DNA fibres (Schwartz et al., 1993), chromosomal DNA is suspended in molten agarose and placed on a microscope slide. As the gel cools and solidifies, the DNA molecules become extended (Figure A). To utilize gel stretching in optical mapping, the microscope slide onto which the molten agarose is placed is first coated with a restriction enzyme. The enzyme is inactive at this stage because there are no magnesium ions, which the enzyme needs in order to function. Once the gel has solidified it is washed with a solution containing magnesium chloride, which activates the restriction enzyme. A fluorescent dye is added, such as DAPI (4,6-diamino-2-phenylindole dihydrochloride), which stains the DNA so that the fibres can be seen when the slide is examined with a high-power fluorescence

microscope. The restriction sites in the extended molecules gradually become gaps as the degree of fibre extension is reduced by the natural springiness of the DNA, enabling the relative positions of the cuts to be recorded.

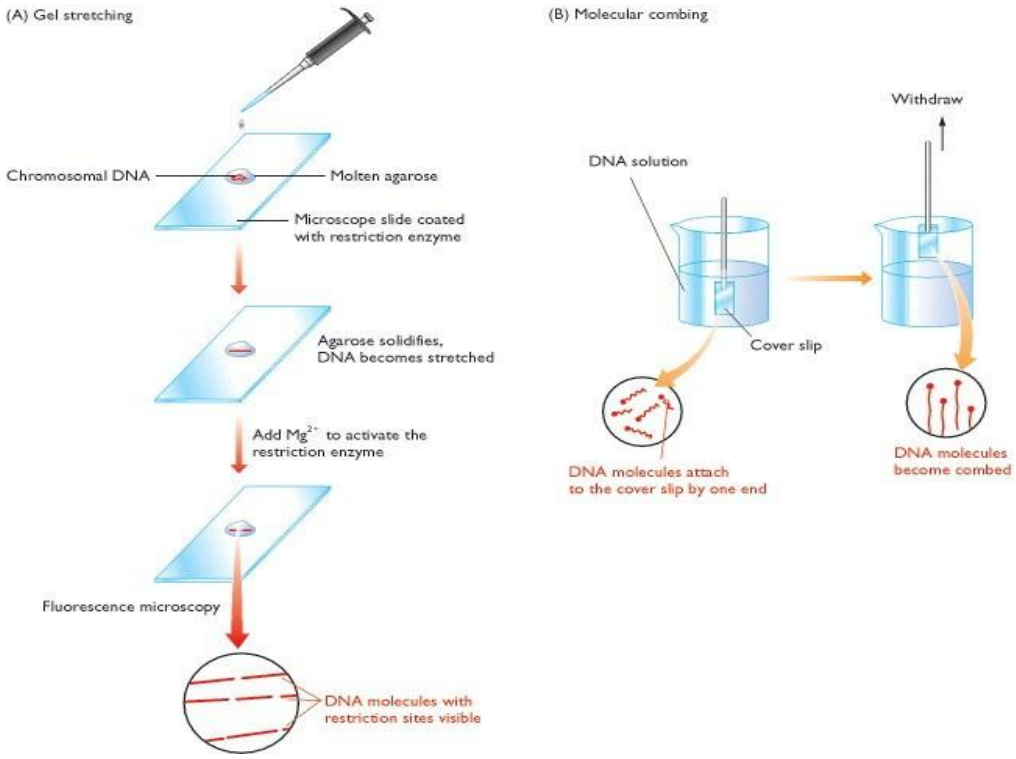


Figure: Gel stretching and molecular combing. (A) To carry out gel stretching, molten agarose containing chromosomal DNA molecules is pipetted onto a microscope slide coated with a restriction enzyme. As the gel solidifies, the DNA molecules become stretched.

In molecular combing (Michalet et al., 1997), the DNA fibres are prepared by dipping a silicone- coated cover slip into a solution of DNA, leaving it for 5 minutes (during which time the DNA molecules attach to the cover slip by their ends), and then removing the slip at a constant speed of 0.3 mm s^{-1} (Figure B). The force required to pull the DNA molecules through the meniscus causes them to line up. Once in the air, the surface of the cover slip dries, retaining the DNA molecules as an array of parallel fibres.

Optical mapping was first applied to large DNA fragments cloned in YAC and BAC vectors. More recently, the feasibility of using this technique with genomic DNA has been proven with studies of a 1-Mbchromosome of the malaria parasite Plasmodium falciparum (Jing et al., 1999), and the two chromosomes and single mega plasmid of the bacterium *Deinococcus radiodurans* (Lin et al., 1999)

Fluorescent in situ hybridization (FISH)

The optical mapping method described above provides a link to the second type of physical mapping procedure that we will consider - FISH (Heiskanen et al., 1996). As in optical mapping, FISH enables the position of a marker on a chromosome or extended DNA molecule to be directly visualized. In optical mapping the marker is a restriction site and it is visualized as a gap in an extended DNA fibre. In FISH, the marker is a DNA sequence that is visualized by hybridization with a fluorescent probe.

In situ hybridization with radioactive or fluorescent probes

In situ hybridization is a version of hybridization analysis in which an intact chromosome is examined by probing it with a labeled DNA molecule. The position on the chromosome at which hybridization occurs provides information about the map location of the DNA sequence used as the probe (Figure). For the method to work, the DNA in the chromosome must be made single stranded ('denatured') by breaking the base pairs that hold the double helix together. Only then will the chromosomal DNA be able to hybridize with the probe. The standard method for denaturing chromosomal DNA without destroying the morphology of the chromosome is to dry the preparation onto a glass microscope slide and then treat with formamide.

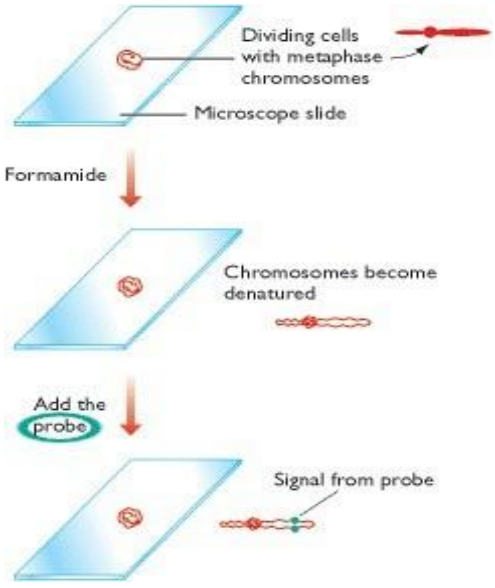


Figure :Fluorescent in situ hybridization. A sample of dividing cells is dried onto a microscope slide and treated with formamide so that the chromosomes become denatured but do not lose their characteristic metaphase morphologies. The position at which the probe hybridizes to the chromosomal DNA is visualized by detecting the fluorescent signal emitted by the labeled DNA.

In the early versions of in situ hybridization the probe was radioactively labeled but this procedure was unsatisfactory because it is difficult to achieve both sensitivity and resolution with a radioactive label, two critical requirements for successful in situ hybridization. Sensitivity requires that the radioactive label has a high emission energy (an example of such a radiolabel is ^{32}P), but if the radiolabel has a high emission energy then it scatters its signal and so gives poor resolution. High resolution is possible if a radiolabel with low emission energy, such as ^{3}H, is used, but these have such low sensitivity that lengthy exposures are needed, leading to a high background and difficulties in discerning the genuine signal.

These problems were solved in the late 1980s by the development of non-radioactive fluorescent DNA labels. These labels combine high sensitivity with high resolution and are ideal for in situ hybridization. Fluoro-labels with different colored emissions have been designed, making it possible to hybridize a number of different probes to a single chromosome and distinguish their individual hybridization signals, thus enabling the relative positions of the probe sequences to be mapped. To maximize sensitivity, the probes must be labeled as heavily as possible, which in the past has meant that they must be quite lengthy DNA molecules - usually cloned DNA fragments of at least 40 kb. This requirement is less important now that techniques for achieving heavy labeling with shorter molecules have been developed. As far as the construction of a physical map is concerned, a cloned DNA fragment can be looked upon as simply another type of marker, although in practice the use of clones as markers adds a second dimension because the cloned DNA is the material from which the DNA sequence is determined. Mapping the positions of clones therefore provides a direct link between a genome map and its DNA sequence.

If the probe is a long fragment of DNA then one potential problem, at least with higher eukaryotes, is that it is likely to contain examples of repetitive DNA sequences and so may hybridize to many chromosomal positions, not just the specific point to which it is perfectly matched. To reduce this non-specific hybridization, the probe, before use, is mixed with unlabeled DNA from the organism being studied. This DNA can simply be total nuclear DNA (i.e. representing the entire genome) but it is better if a fraction enriched for repeat

sequences is used. The idea is that the unlabeled DNA hybridizes to the repetitive DNA sequences in the probe, blocking these so that the subsequent *in situ* hybridization is driven wholly by the unique sequences (Lichter et al., 1990). Non-specific hybridization is therefore reduced or eliminated entirely (Figure).

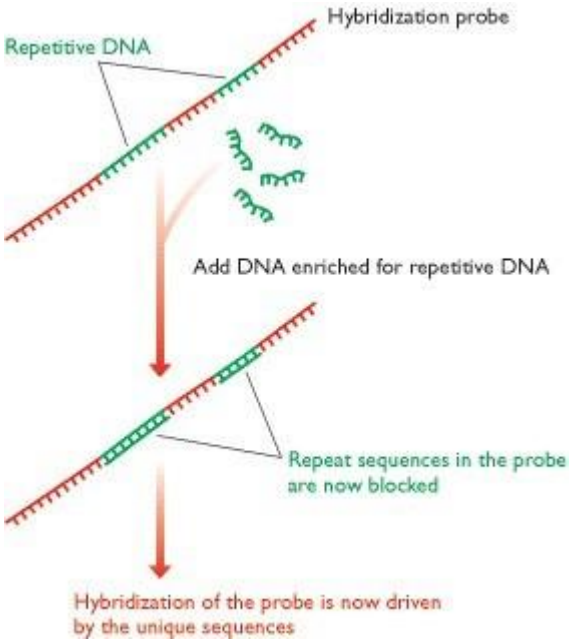


Figure :A method for blocking repetitive DNA sequences in a hybridization probe. In this example the probe molecule contains two genome-wide repeat sequences (shown in green). If these sequences are not blocked then the probe will hybridize to any copies of these genome-wide repeats in the target DNA. To block the repeat sequences, the probe is prehybridized with a DNA fraction enriched for repetitiv

FISH in action

FISH was originally used with metaphase chromosomes . These chromosomes, prepared from nuclei that are undergoing division, are highly condensed and each chromosome in a set takes up a recognizable appearance, characterized by the position of its centromere and the banding pattern that emerges after the chromosome preparation is stained (see Figure). With metaphase chromosomes, a fluorescent signal obtained by FISH is mapped by measuring its position relative to the end of the short arm of the chromosome (the FLpter value). A disadvantage is that the highly condensed nature of metaphase chromosomes means that only low-resolution mapping is possible, two markers having to be at least 1 Mb apart to be resolved as separate hybridization signals (Trask et al., 1991). This degree of resolution is insufficient for the construction of useful chromosome maps, and the main application of metaphase FISH has been in determining the chromosome on which a new marker is located, and providing a rough idea of its map position, as a preliminary to finer scale mapping by other methods.

For several years these 'other methods' did not involve any form of FISH, but since 1995 a range of higher resolution FISH techniques has been developed. With these techniques, higher resolution is achieved by changing the nature of the chromosomal preparation being studied. If metaphase chromosomes are too condensed for fine-scale mapping then we must use chromosomes that are more extended. There are two ways of doing this (Heiskanen et al., 1996):

- **Mechanically stretched chromosomes** can be obtained by modifying the preparative method used to isolate chromosomes from metaphase nuclei. The inclusion of a centrifugation step generates shear forces which can result in the chromosomes becoming stretched to up to 20 times their normal length. Individual chromosomes are still recognizable and FISH signals can be mapped in the same way as with normal metaphase chromosomes. The resolution is significantly improved and markers that are 200–300kb apart can be distinguished.
- **Non-metaphase chromosomes** can be used because it is only during metaphase that chromosomes are highly condensed: at other stages of the cell cycle the chromosomes are naturally unpacked. Attempts have been made to use prophase nuclei because in these
- the chromosomes are still sufficiently condensed for individual ones to be identified. In practice, however, these preparations provide no advantage over mechanically stretched chromosomes. Interphase chromosomes are more useful because this stage of the cell cycle (between nuclear divisions) is when the chromosomes

are most unpacked. Resolution down to 25 kb is possible, but chromosome morphology is lost so there are no external reference points against which to map the position of the probe. This technique is therefore used after preliminary map information has been obtained, usually as a means of determining the order of a series of markers in a small region of a chromosome.

Interphase chromosomes contain the most unpacked of all cellular DNA molecules. To improve the resolution of FISH to better than 25 kb it is therefore necessary to abandon intact chromosomes and instead use purified DNA. This approach, called fibre-FISH, makes use of DNA prepared by gel stretching or molecular combing and can distinguish markers that are less than 10 kb apart.

Sequence tagged site (STS)mapping

To generate a detailed physical map of a large genome we need, ideally, a high-resolution mapping procedure that is rapid and not technically demanding. Neither of the two techniques that we have considered so far - restriction mapping and FISH - meets these requirements. Restriction mapping is rapid, easy, and provides detailed information, but it cannot be applied to large genomes. FISH can be applied to large genomes, and modified versions such as fibre-FISH can give high-resolution data, but FISH is difficult to carry out and data accumulation is slow, map positions for no more than three or four markers being obtained in a single experiment. If detailed physical maps are to become a reality then we need a more powerful technique.

At present the most powerful physical mapping technique, and the one that has been responsible for generation of the most detailed maps of large genomes, is STS mapping. A sequence tagged site or **STS** is simply a short DNA sequence, generally between 100 and 500 bp in length, that is easily recognizable and occurs only once in the chromosome or genome being studied. To map a set of STSs, a collection of overlapping DNA fragments from a single chromosome or from the entire genome is needed. In the example shown in Figure, a fragment collection has been prepared from a single chromosome, with each point along the chromosome represented on average five times in the collection. The data from which the map will be derived are obtained by determining which fragments contain which STSs. This can be done by hybridization analysis but PCR is generally used because it is quicker and has proven to be more amenable to automation. The chances of two STSs being present on the same fragment will, of course, depend on how close together they are in the genome. If they are very close then there is a good chance that they will always be on the same fragment; if they are further apart then sometimes they will be on the same fragment and sometimes they will not (Figure below). The data can therefore be used to calculate the distance between two markers, in a manner analogous to the way in which map distances are determined by linkage analysis. Remember that in linkage analysis a map distance is

calculated from the frequency at which crossovers occur between two markers. STS mapping is essentially the same, except that each map distance is based on the frequency at which *breaks* occur between two markers.

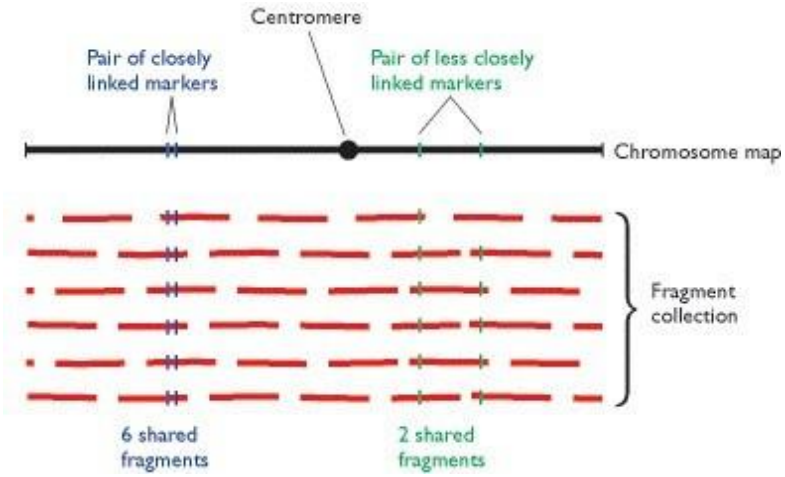


Figure : A fragment collection suitable for STS mapping. The fragments span the entire length of a chromosome, with each point on the chromosome present in an average of five fragments. The two blue markers are close together on the chromosome map and there is a high probability that they will be found on the same fragment. The two green markers are more distant from one another and so are less likely to be found on the same fragment

Any unique DNA sequence can be used as an STS

To qualify as an STS, a DNA sequence must satisfy two criteria. The first is that its sequence must be known, so that a PCR assay can be set up to test for the presence or absence of the STS on different DNA fragments. The second requirement is that the STS must have a unique location in the chromosome being studied, or in the genome as a whole if the DNA fragment set covers the entire genome. If the STS sequence occurs at more than one position then the mapping data will be ambiguous. Care must therefore be taken to ensure that STSs do not include sequences found in repetitive DNA.

These are easy criteria to satisfy and STSs can be obtained in many ways, the most common sources being **expressed sequence tags (ESTs)**, SSLPs, and **random genomic sequences**.

Expressed sequence tags *(ESTs):* These are short sequences obtained by analysis of cDNA clones (Marra et al., 1998). Complementary DNA is prepared by converting an mRNA preparation into double-stranded DNA (Figure). Because the mRNA in a cell is derived from protein-coding genes, cDNAs and the ESTs obtained from them represent the genes that were being expressed in the cell from which the mRNA was prepared. ESTs are looked upon

as a rapid means of gaining access to the sequences of important genes, and they are valuable even if their sequences are incomplete. An EST can also be used as an STS, assuming that it comes from a unique gene and not from a member of a gene family in which all the genes have the same or very similar sequences.

SSLPs: In earlier sections, we examined the use of microsatellites and other SSLPs in genetic mapping. SSLPs can also be used as STSs in physical mapping. SSLPs that are polymorphic and have already been mapped by linkage analysis are particularly valuable as they provide a direct connection between the genetic and physical maps.

Random genomic sequences: These are obtained by sequencing random pieces of cloned genomic DNA, or simply by downloading sequences that have been deposited in the databases.

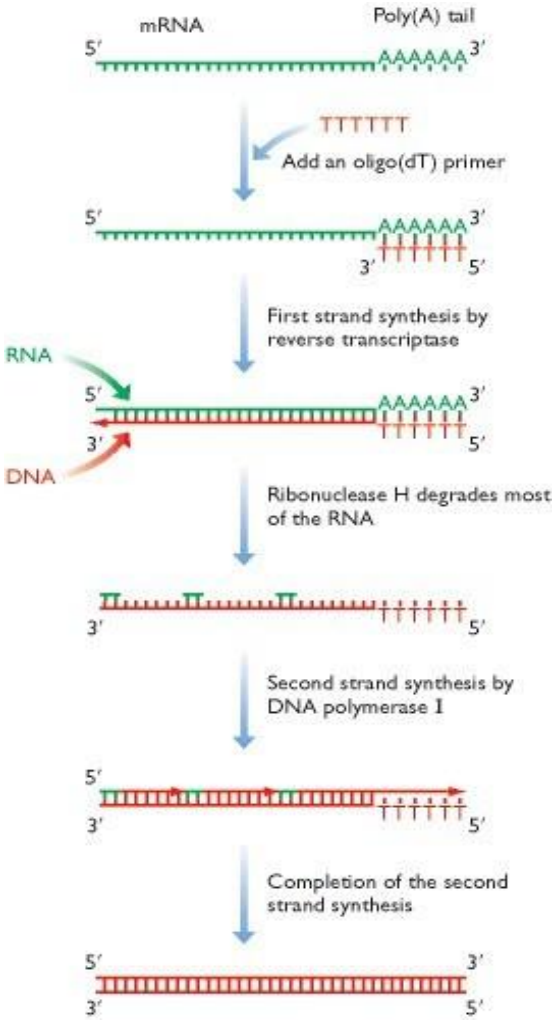


Figure : Most eukaryotic mRNAs have a poly(A) tail at their 3′ end . This series of A nucleotides is used as the priming site for the first stage of cDNA synthesis, carried out by reverse transcriptase - a DNA polymerase that copies an RNA template. The primer is a short synthetic DNA

oligonucleotide, typically 20 nucleotides in length, made up entirely of Ts (an 'oligo(dT)' primer). When the first strand synthesis has been completed, the preparation is treated with ribonuclease H, which specifically degrades the RNA component of an RNA-DNA hybrid. Under the conditions used, the enzyme does not degrade all of the RNA, instead leaving short segments that prime the second DNA strand synthesis reaction, this one catalyzed by DNA polymerase-I.

Fragments of DNA for STS mapping

The second component of an STS mapping procedure is the collection of DNA fragments spanning the chromosome orgenome being studied. This collection is sometimes called the mapping reagent and at present there are two ways in which it can be assembled: as a clone library and as a panel of radiation hybrids. We will consider radiation hybrids first.

A radiation hybrid is a rodent cell that contains fragments of chromosomes from a second organism (McCarthy, 1996). The technology was first developed in the 1970s when it was discovered that exposure of human cells to X-ray doses of 3000–8000 rads causes the chromosomes to break up randomly into fragments, larger X-ray doses producing smaller fragments (Figure below). This treatment is of course lethal for the human cells, but the chromosome fragments can be propagated if the irradiated cells are subsequently fused with non-irradiated hamster or other rodent cells. Fusion is stimulated either chemically with polyethylene glycol or by exposure to Sendai virus. Not all of the hamster cells take up chromosome fragments so a means of identifying the hybrids is needed. The routine selection process is to use a hamster cell line that is unable to make either thymidine kinase (TK) or hypoxanthine phosphoribosyl transferase (HPRT), deficiencies in either of these two enzymes being lethal when the cells are grown in a medium containing a mixture of hypoxanthine, aminopterin and thymidine (HAT medium). After fusion, the cells are placed in HAT medium. Those that grow are hybrid hamster cells that have acquired human DNA fragments that include genes for the human TK and HPRT enzymes, which are synthesized inside the hybrids, enabling these cells to grow in the selective medium. The treatment results in hybrid cells that contain arandomselection of human DNA fragments inserted into the hamster chromosomes. Typically the fragments are 5–10 Mb in size, with each cell containing fragments equivalent to 15–35% of the human genome. The collection of cells is called a radiation hybrid panel and can be used as a mapping reagent in STS mapping, provided that the PCR assay used to identify the STS does not amplify the equivalent region of DNA from the hamster genome.

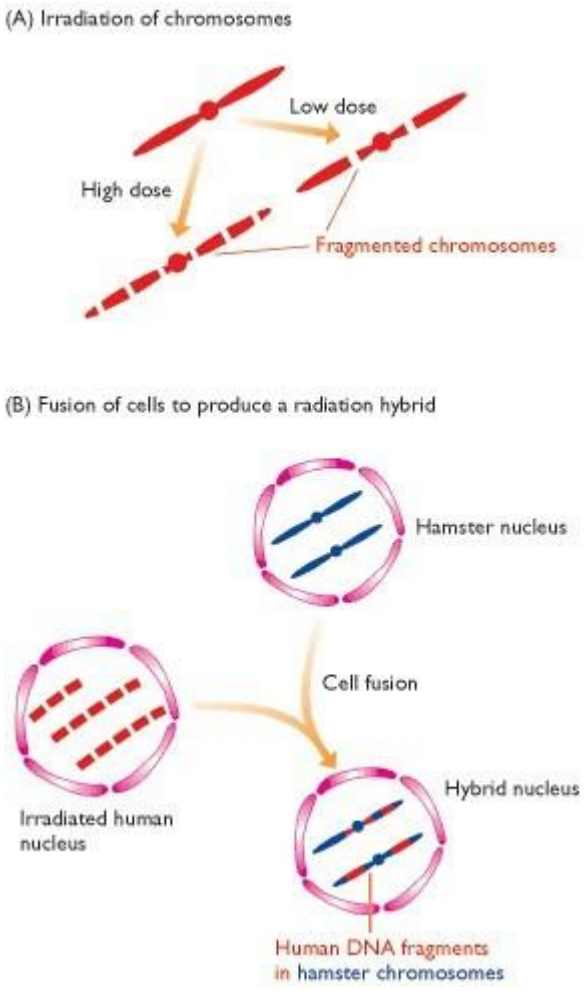


Figure: Radiation hybrids. (A) The result of irradiation of human cells: the chromosomes break into fragments, smaller fragments generated by higher X-ray doses. In (B), a radiation hybrid is produced by fusing an irradiated human cell with an untreated hamster cell. For clarity, only the nuclei are shown.

A second type of radiation hybrid panel, containing DNA from just one human chromosome, can be constructed if the cell line that is irradiated is not a human one but a second type of rodent hybrid. Cytogeneticists have developed a number of rodent cell lines in which a single human chromosome is stably propagated in the rodent nucleus. If a cell line of this type is irradiated and fused with hamster cells, then the hybrid hamster cells obtained after selection will contain either human or mouse chromosome fragments, or a mixture of both. The ones containing human DNA can be identified by probing with a human-specific genome-wide repeat sequence, such as the short interspersed nuclear element (SINE) called Alu, which has a copy number of just over 1 million and so occurs on average once every 4 kb in the human genome. Only cells containing human DNA will hybridize to Alu probes, enabling the

uninteresting mouse hybrids to be discarded and STS mapping to be directed at the cells containing human chromosome fragments.

Radiation hybrid mapping of the human genome was initially carried out with chromosome-specific rather than whole-genome panels because it was thought that fewer hybrids would be needed to map a single chromosome than would be needed to map the entire genome. It turns out that a high-resolution map of a single human chromosome requires a panel of 100–200 hybrids, which is about the most that can be handled conveniently in a PCR screening program. But whole-genome and single-chromosome panels are constructed differently, the former involving irradiation of just human DNA, and the latter requiring irradiation of a mouse cell containing much mouse DNA and relatively little human DNA. This means that the human DNA content per hybrid is much lower in a single-chromosome panel than in a whole-genome panel. It transpires that detailed mapping of the entire human genome is possible with fewer than 100 whole-genome radiation hybrids, so whole-genome mapping is no more difficult than single- chromosome mapping. Once this was realized, whole-genome radiation hybrids became a central component of the mapping phase of the Human Genome Project. Whole-genome libraries are also being used for STS mapping of other mammalian genomes and for those of the zebra fish and the chicken (McCarthy,1996).

A clone library can also be used as the mapping reagent for STS analysis

A preliminary to the sequencing phase of a genome project is to break the genome or isolated chromosomes into fragments and to clone each one in a high-capacity vector, one able to handle large fragments of DNA . This results in a clone library, a collection of DNA fragments, which, in this case, have an average size of several hundred kb. As well as supporting the sequencing work, this type of clone library can also be used as a mapping reagent in STS analysis.

As with radiation hybrid panels, a clone library can be prepared from genomic DNA, and so represents the entire genome, or a chromosome-specific library can be made if the starting DNA comes from just one type of chromosome. The latter is possible because individual chromosomes can be separated by flow cytometry. To carry out this technique, dividing cells (ones with condensed chromosomes) are carefully broken open so that a mixture of intact chromosomes is obtained. The chromosomes are then stained with a fluorescent dye. The amount of dye that a chromosome bind depends on its size, so larger chromosomes bind more dye and fluoresce more brightly than smaller ones. The chromosome preparation is diluted and passed through a fine aperture, producing a stream of droplets, each one containing a single chromosome. The droplets pass through a detector that measures the amount of fluorescence, and hence identifies which droplets contain the particular chromosome being sought. An electric charge is applied to these drops, and no others (Figure), enabling the droplets containing the desired chromosome to be deflected and separated

from the rest. What if two different chromosomes have similar sizes, as is the case with human chromosomes 21 and 22? These can usually be separated if the dye that is used is not one that binds non-specifically to DNA, but instead has a preference for AT- or GC-rich regions. Examples of such dyes are Hoechst 33258 and chromomycin A_3, respectively. Two chromosomes that are the same size rarely have identical GC contents, and so can be distinguished by the amounts of AT- or GC-specific dye that they bind.

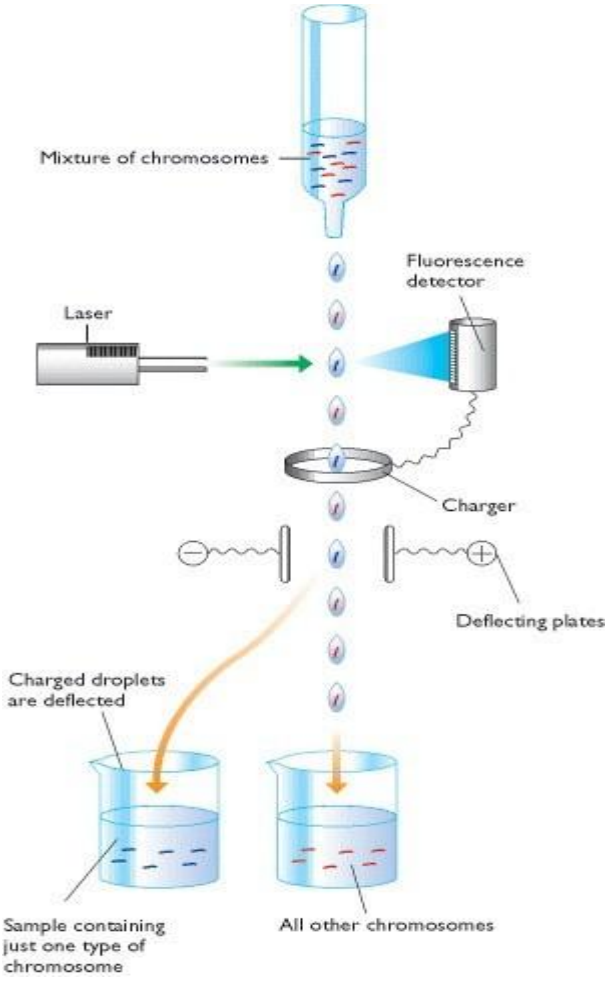


Figure : Separating chromosomes by flow cytometry A mixture of fluorescently stained chromosomes is passed through a small aperture so that each drop that emerges contains just one chromosome. The fluorescence detector identifies the signal from drops containing the correct chromosome and applies an electric charge to these drops. When the drops reach the electric plates, the charged ones are deflected into a separate beaker. All other drops fall straight through the deflecting plates and are collected in the waste beaker.

Compared with radiation hybrid panels, clone libraries have one important advantage for STS mapping. This is the fact that the individual clones can subsequently provide the DNA that is actually sequenced. The data resulting from STS analysis, from which the physical map is generated, can equally well be used to determine which clones contain overlapping DNA fragments, enabling a clone contig to be built up (Figure); for other methods for assembling clone contigs see). This assembly of overlapping clones can be used as the base material for a lengthy, continuous DNA sequence, and the STS data can later be used to anchor this sequence precisely onto the physical map. If the STSs also include SSLPs that have been mapped by genetic linkage analysis then the DNA sequence, physical map and genetic map can all be integrated.

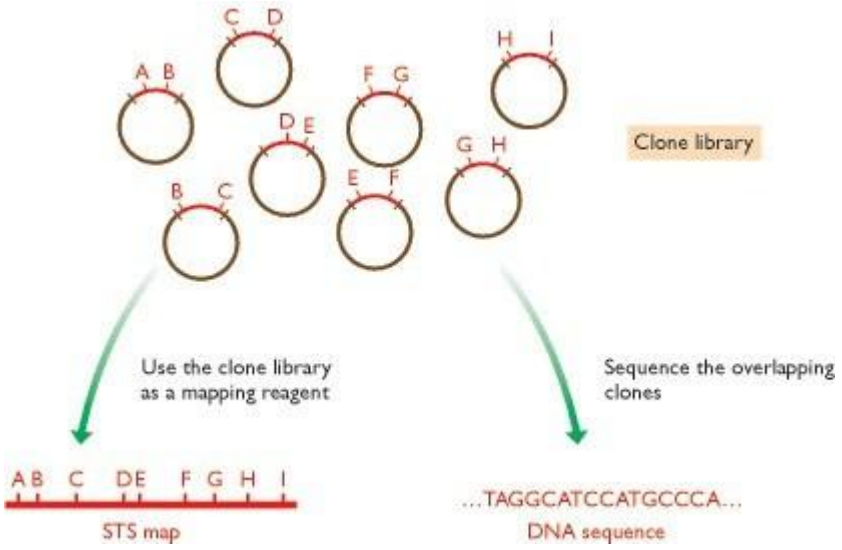


Figure The value of clone libraries in genome projects. The small clone library shown in this example contains sufficient information for an STS map to be constructed, and can also be used as the source of the DNA that will besequenced.

Genome Sequencing:

The objective of a genome project is the complete DNA sequence for the organism being studied, ideally integrated with the genetic and/or physical maps of the genome so that genes and other interesting features can be located within the DNA sequence. This chapter describes the techniques and research strategies that are used during the sequencing phase of a genome project, when this ultimate objective is being directly addressed. Techniques for sequencing DNA are clearly of central importance in this context and we will begin the chapter with a detailed examination of sequencing methodology. This methodology is of little value however, unless the short sequences that result from individual sequencing experiments can be linked together in the correct order to give the master sequences of the

chromosomes that make up the genome. The middle part of this chapter describes the strategies used to ensure that the master sequences are assembled correctly. Finally, we will review the way in which mapping and sequencing were used to produce the two draft human genome sequences that were published in February 2001.

The Methodology for DNA Sequencing

Rapid and efficient methods for DNA sequencing were first devised in the mid-1970s. Two different procedures were published at almost the same time:

- The chain termination method (Sanger et al., 1977), in which the sequence of a single- stranded DNA molecule is determined by enzymatic synthesis of complementary polynucleotide chains, these chains terminating at specific nucleotide positions;
- The **chemical degradation method** (Maxam and Gilbert, 1977), in which the sequence of a double-stranded DNA molecule is determined by treatment with chemicals that cut the molecule at specific nucleotide positions.

Both methods were equally popular to begin with but the chain termination procedure has gained ascendancy in recent years, particularly for genome sequencing. This is partly because the chemicals used in the chemical degradation method are toxic and therefore hazardous to the health of the researchers doing the sequencing experiments, but mainly because it has been easier to automate chain termination sequencing. As we will see later in this chapter, a genome project involves a huge number of individual sequencing experiments and it would take many years to perform all these by hand. Automated sequencing techniques are therefore essential if the project is to be completed in a reasonable time-span.

Chain termination DNA sequencing

Chain termination DNA sequencing is based on the principle that single-stranded DNA molecules that differ in length by just a single nucleotide can be separated from one another by polyacrylamide gel electrophoresis. This means that it is possible to resolve a family of molecules, representing all lengths from 10 to 1500 nucleotides, into a series of bands (Figure).

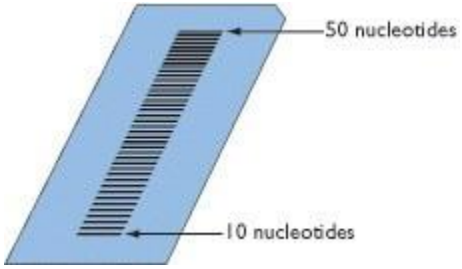


Figure: The banding pattern is produced after separation of single- stranded DNA molecules by denaturing polyacrylamide gel electrophoresis. The molecules are labeled with a radioactive marker and the bands visualized by autoradiography. The bands gradually get closer together towards the top of the ladder. In practice, molecules up to about 1500 nucleotides in length can be separated if the electrophoresis is continued for long enough. Chain termination sequencing in outline

The starting material for a chain termination sequencing experiment is a preparation of identical single-stranded DNA molecules. The first step is to anneal a short oligonucleotide to the same position on each molecule, this oligonucleotide subsequently acting as the primer for synthesis of a new DNA strand that is complementary to the template (Figure). The strand synthesis reaction, which is catalyzed by a DNA polymerase enzyme and requires the four deoxyribonucleotide triphosphates (dNTPs - dATP, dCTP, dGTP and dTTP) as substrates, would normally continue until several thousand nucleotides had been polymerized. This does not occur in a chain termination sequencing experiment because, as well as the four dNTPs, a small amount ofa dideoxynucleotide (e.g.ddATP) is added to the reaction. The polymerase enzyme does not discriminate between dNTPs and ddNTPs, so the dideoxynucleotide can be incorporated into the growing chain, but it then blocks further elongation because it lacks the 3′-hydroxyl group needed to form a connection with the next nucleotide (Figure B).

If ddATP is present, chain termination occurs at positions opposite thymidines in the template DNA. Because dATP is also present the strand synthesis does not always terminate at the first T in the template; in fact it may continue until several hundred nucleotides have been polymerized before a ddATP is eventually incorporated. The result is therefore a set of new chains, all of different lengths, but each ending in ddATP. Now the polyacrylamide gel comes into play. The family of molecules generated in the presence of ddATP is loaded into one lane of the gel, and the families generated with ddCTP, ddGTP and ddTTP loaded into the three adjacent lanes. After electrophoresis, the DNA sequence can be read directly from the positions of the bands in the gel (Figure D). The band that has moved the furthest represents the smallest piece of DNA, this being the strand that terminated by incorporation of a ddNTP at the first position in the template. In the example shown in

Figure this band lies in the 'G' lane (i.e. the lane containing the molecules terminated with ddGTP), so the first nucleotide in the sequence is 'G'. The next band, corresponding to the molecule that is one nucleotide longer than the first, is in the 'A' lane, so the second nucleotide is 'A' and the sequence so far is 'GA'. Continuing up through the gel we see that the next band also lies in the 'A' lane (sequence GAA), then we move to the 'T' lane (GAAT), and so on. The sequence reading can be continued up to the region of the gel where individual bands are not separated.

Chain termination sequencing requires a single-stranded DNA template

The template for a chain termination experiment is a single-stranded version of the DNA molecule to be sequenced. There are several ways in which this can be obtained:

- **The** DNA **can be cloned in a plasmid vector :** The resulting DNA will be double stranded so cannot be used directly in sequencing. Instead, it must be converted into single-stranded DNA by denaturation with alkali or by boiling. This is a common method for obtaining template DNA for DNA sequencing, largely because cloning in a plasmid vector is such a routine technique. A shortcoming is that it can be difficult to prepare plasmid DNA that is not contaminated with small quantities of bacterial DNA and RNA, which can act as spurious templates or primers in the DNA sequencing experiment.

- **The** DNA **can be cloned in a bacteriophage M13 vector.** Vectors based on M13 bacteriophage are designed specifically for the production of single-stranded templates for DNA sequencing. M13 bacteriophage has a single-stranded DNA genome which, after infection of Escherichia coli bacteria, is converted into a double-stranded replicative form. The replicative form is copied until over 100 molecules are present in the cell, and when the cell divides the copy number in the new cells is maintained by further replication. At the same time, the infected cells continually secrete new M13 phage particles, approximately 1000 per generation, these phages containing the single-stranded version of the genome. Cloning vectors based on M13 vectors are double-stranded DNA molecules equivalent to the replicative form of the M13 genome. They can be manipulated in exactly the same way as a plasmid cloning vector. The difference is that cells that have been transfected with a recombinant M13 vector secrete phage particles containing single-stranded DNA, this DNA comprising the vector molecule plus any additional DNA that has been ligated into it. The phages therefore provide the template DNA for chain termination sequencing. The one disadvantage is that DNA fragments longer than about 3 kb suffer deletions and rearrangements when cloned in an M13 vector, so the system can only be used with short pieces of DNA.

- **The** DNA **can be cloned in a phagemid.** This is a plasmid cloning vector that contains, in addition to its plasmid origin of replication, the origin from M13 or another phage with a single-stranded DNA genome. If an E. coli cell contains both a phagemid and the replicative form of a helper phage, the latter carrying genes for the phage replication enzymes and coat proteins, then the phage origin of the phagemid becomes activated, resulting in synthesis of phage particles containing the single-stranded version of the phagemid. The double-stranded plasmid DNA is therefore converted into single-stranded template DNA for DNA sequencing. This system avoids the instabilities of M13 cloning and can be used with fragments up to 10 kb ormore.

- PCR **can be used to generate single-stranded** DNA**.** There are various ways of generating single-stranded DNA by PCR, the most effective being to modify one of the two primers so that DNA strands synthesized from this primer are easily purified. One possibility is to attach small metallic beads to the primer and then use a magnetic device to purify the resulting strands.

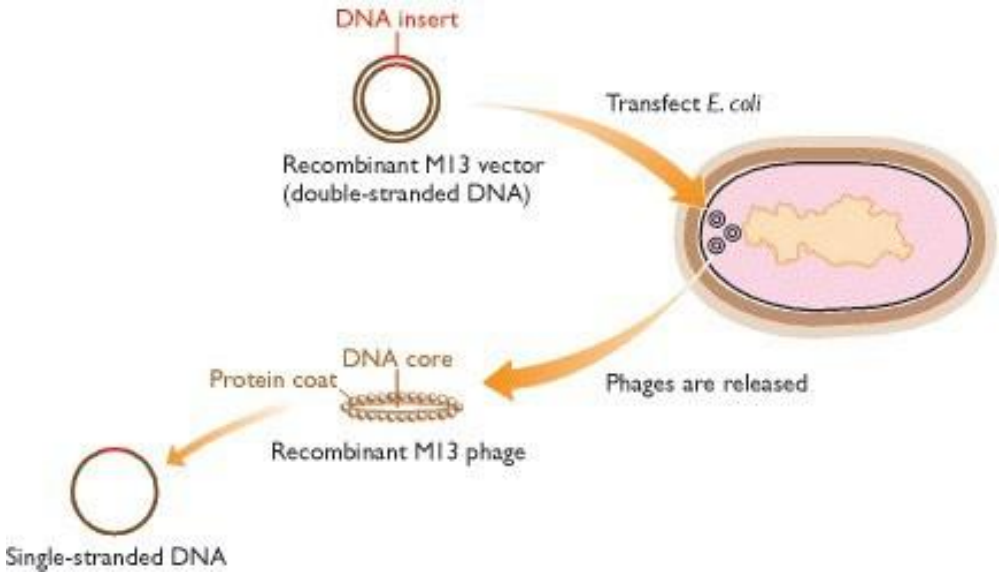


Figure. M13 vectors can be obtained in two forms: the double-stranded replicative molecule and the single-stranded version found in bacteriophage particles. The replicative form can be manipulated in the same way as a plasmid cloning vector (Section 4.2.1) with new DNA inserted by restriction followed by ligation. The recombinant vector is introduced into Escherichia coli cells by transfection. Once inside an E. coli cell, the double-stranded vector replicates and directs synthesis of single-stranded copies, which are packaged into phage particles and secreted from the cell. The phage particles can be collected from the culture medium after centrifuging to pellet the bacteria. The protein coats of the phages are removed by treating with phenol, and the single-stranded version of the recombinant vector is purified for use in DNA sequencing.

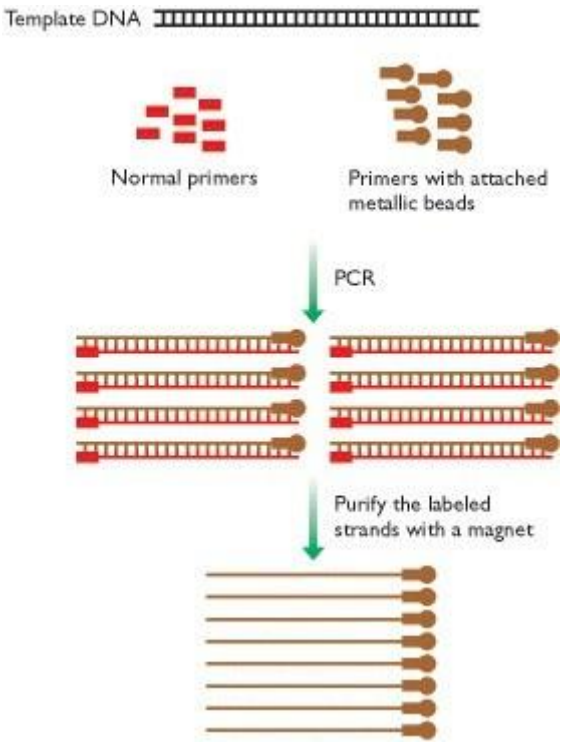


Figure :The PCR is carried out with one normal primer (shown in red), and one primer that is labelled with a metallic bead (shown in brown). After PCR, the labelled strands are purified with a magnetic device. For more details about PCR The primer determines the region of the template DNA that will be sequenced

To begin a chain termination sequencing experiment, an oligonucleotide primer is annealed onto the template DNA. The primer is needed because template-dependent DNA polymerases cannot initiate DNA synthesis on a molecule that is entirely single-stranded: there must be a short double-stranded region to provide a 3′ end onto which the enzyme can add new nucleotides.

The primer also plays the critical role of determining the region of the template molecule that will be sequenced. For most sequencing experiments a 'universal' primer is used, this being one that is complementary to the part of the vector DNA immediately adjacent to the point into which new DNA is ligated (Figure below). The same universal primer can therefore give the sequence of any piece of DNA that has been ligated into the vector. Of course if this inserted DNA is longer than 750 bp or so then only a part of its sequence will be obtained, but usually this is not a problem because the project as a whole simply requires that a large number of short sequences are generated and subsequently assembled into the contiguous master sequence. It is immaterial whether or not the short

sequences are the complete or only partial sequences of the DNA fragments used as templates. If double-stranded plasmid DNA is being used to provide the template then, if desired, more sequence can be obtained from the other end of the insert. Alternatively, it is possible to extend the sequence in one direction by synthesizing a non- universal primer, designed to anneal at a position within the insert DNA (Figure below). An experiment with this primer will provide a second short sequence that overlaps the previous one.

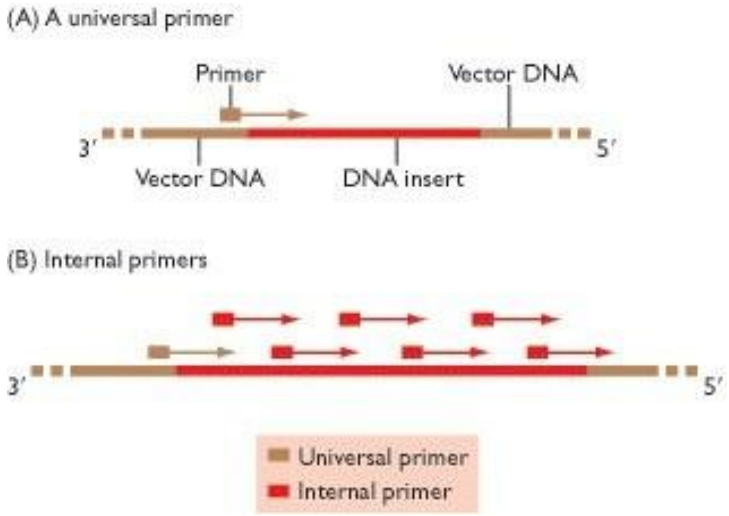


Figure : A universal primer anneals to the vector DNA, adjacent to the position at which new DNA is inserted. A single universal primer can therefore be used to sequence any DNA insert, but only provides the sequence of one end of the insert. (B) One way of obtaining a longer sequence is to carry out a series of chain termination experiments, each with a different internal primer that anneals within the DNAinsert.

Thermal cycle sequencing offers an alternative to the traditional methodology

The discovery of thermostable DNA polymerases, which led to the development of PCR, has also resulted in new methodologies for chain termination sequencing. In particular, the innovation called thermal cycle sequencing (Sears et al., 1992) has two advantages over traditionalchainterminationsequencing.First,it uses double-stranded rather than single-stranded DNA as the starting material. Second, very little template DNA is needed, so the DNA does not have to be cloned before being sequenced.

Thermal cycle sequencing is carried out in a similar way to PCR but just one primer is used and each reaction mixture includes one of the ddNTPs (Figure below). Because there is only one primer, only one of the strands of the starting molecule is copied, and the product accumulates in a linear fashion, not exponentially as is the case in a real PCR. The presence of the ddNTP in the reaction mixture causes chain termination, as in the standard methodology, and the family of resulting strands can be analyzed and the sequence read in the normal manner by polyacrylamide gel electrophoresis.

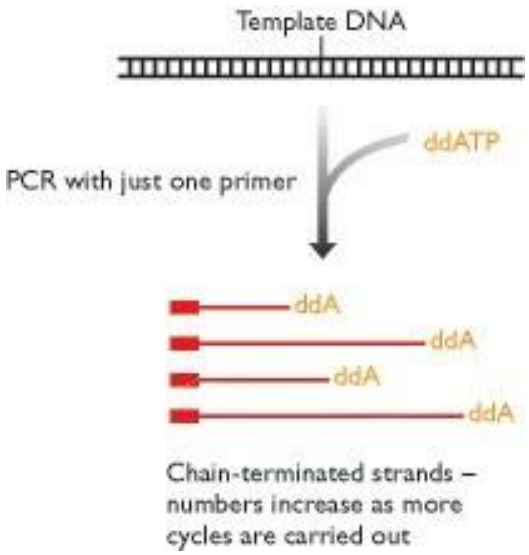


Figure: PCR is carried out with just one primer and with a dideoxynucleotide present in the reaction mixture. The result is a family of chain-terminated strands - the 'A' family in the reaction shown. These strands, along with the products of the C, G and T reactions, are electrophoresed as in the standard methodology

The standard chain termination sequencing methodology employs radioactive labels, and the banding pattern in the polyacrylamide gel is visualized by autoradiography. Usually one of the nucleotides in the sequencing reaction is labeled so that the newly synthesized strands contain radiolabels along their lengths, giving high detection sensitivity. To ensure good band resolution, ^{33}P or ^{35}S is generally used, as the emission energies of these isotopes are relatively low, in contrast to ^{32}P, which has a higher emission energy and gives poorer resolution because of signal scattering.

Previously we saw how the replacement of radioactive labels by fluorescent ones has given a new dimension to in situ hybridization techniques. Fluorolabeling has been equally important in the development of sequencing methodology, in particular because the detection system for fluorolabels has opened the way to automated sequence reading

(Prober et al., 1987). The label is attached to the ddNTPs, with a different fluorolabel used for each one (Figure below). Chains terminated with A are therefore labeled with one fluorophore, chains terminated with C are labeled with a second fluorophore, and so on. Now it is possible to carry out the four sequencing reactions - for A, C, G and T - in a single tube and to load all four families of molecules into just one lane of the polyacrylamide gel, because the fluorescent detector can discriminate between the different labels and hence determine if each band represents an A, C, G or T. The sequence can be read directly as the bands pass in front of the detector and either printed out in a form readable by eye (Figure B) or sent straight to a computer for storage. When combined with robotic devices that prepare the sequencing reactions and load the gel, the fluorescent detection system provides a major increase in throughput and avoids errors that might arise when a sequence is read by eye and then entered manually into a computer. It is only by use of these automated techniques that we can hope to generate sequence data rapidly enough to complete a genome project in a reasonable length of time.

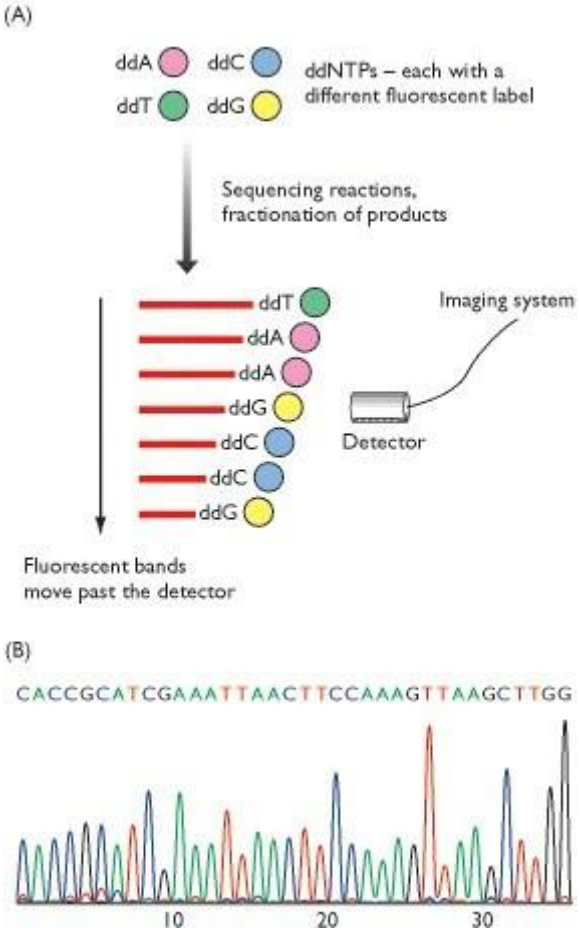


Figure (A) The chain termination reactions are carried out in a single tube, with each dideoxynucleotide labeled with a different fluorophore. In the automated sequencer, the bands in the electrophoresis gel move past a fluorescence detector, which

identifies which dideoxynucleotide is present in each band. The information is passed to the imaging system. (B) The printout from an automated sequencer. The sequence is represented by a series of peaks, one for each nucleotide position. In this example, a green peak is an 'A', blue is 'C', black is 'G', and red is 'T'.

High throughput sequencing methods

In spite of the development of automated techniques, conventional DNA sequencing suffers from the limitation that only a few hundred bp of sequence can be determined in a single experiment. In the context of the Human Genome Project, this means that each experiment provides only one five-millionth of the total genome sequence. Attempts are continually being made to modify the technology so that sequence acquisition is more rapid, a recent example being the introduction of new automated sequencers that use capillary separation rather than a polyacrylamide gel. These have 96 channels so 96 sequences can be determined in parallel, and each run takes less than 2 hours to complete, enabling up to 1000 sequences to be obtained in a single day (Mullikan and McMurray, 1999). Other systems that are being developed will increase data generation even further by enabling 384 or 1024 sequences to be run at the same time (Rogers, 1999).

There have also been attempts to make sequence acquisition more rapid by devising new sequencing methodologies. One possibility is pyrosequencing, which does not require electrophoresis or any other fragment separation procedure and so is more rapid than chain termination sequencing (Ronaghi et al., 1998). In pyrosequencing, the template is copied in a straightforward manner without added ddNTPs. As the new strand is being made, the order in which the dNTPs are incorporated is detected, so the sequence can be 'read' as the reaction proceeds. The addition of a nucleotide to the end of the growing strand is detectable because it is accompanied by release of a molecule of pyrophosphate, which can be converted by the enzyme sulfurylase into a flash of chemiluminescence. Of course, if all four dNTPs were added at once then flashes of light would be seen all the time and no useful sequence information would be obtained. Each dNTP is therefore added separately, one after the other, with a nucleotidase enzyme also present in the reaction mixture so that if a dNTP is not incorporated into the polynucleotide then it is rapidly degraded before the next dNTP is added. This procedure makes it possible to follow the order in which the dNTPs are incorporated into the growing strand. The technique sounds complicated, but it simply requires that a repetitive series of additions be made to the reaction mixture, precisely the type of procedure that is easily automated, with the possibility of many experiments being carried out in parallel.

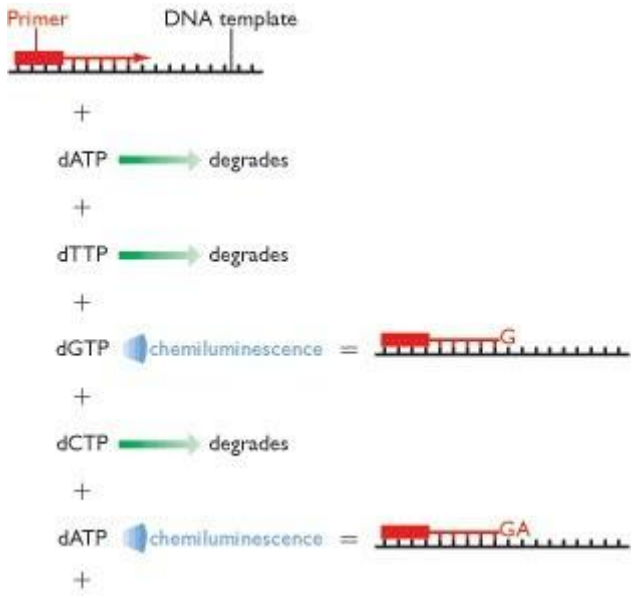


Figure Pyrosequencing. The strand synthesis reaction is carried out in the absence of dideoxynucleotides. Each dNTP is added individually, along with a nucleotidase enzyme that degrades the dNTP if it is not incorporated into the strand being synthesized. Incorporation of a nucleotide is detected by a flash of chemiluminescence induced by the pyrophosphate released from the dNTP. The order in which nucleotides are added to the growing strand can therefore be followed.

A very different approach to DNA sequencing through the use of DNA chips might one day be possible. A chip carrying an array of different oligonucleotides could be used in DNA sequencing by applying the test molecule - the one whose sequence is to be determined - to the array and detecting the positions at which it hybridizes. Hybridization to an individual oligonucleotide would indicate the presence of that particular oligonucleotide sequence in the test molecule, and comparison of all the oligonucleotides to which hybridization occurs would enable the sequence of the test molecule to be deduced (Figure). The problem with this approach is that the maximum length of the molecule that can be sequenced is given by the square root of the number of oligonucleotides in the array, so if every possible 8-mer oligonucleotide (ones containing eight nucleotides) were attached to the chip - all 65 536 of them - then the maximum length of readable sequence would be only 256 bp(Southern, 1996). Even if the chip carried all the 1 048 576 different 10-mer sequences, it could still only be used to sequence a 1 kb molecule. To sequence a 1 Mb molecule (this being the sort of advance in sequence capability that is really needed) the chip would have to carry all of the 1×10^{12} possible 20-mers. This may sound an outlandish proposition but advances in miniaturization, together with the possibility of electronic rather than visual detection of hybridization, could bring such an array within reach in the future.

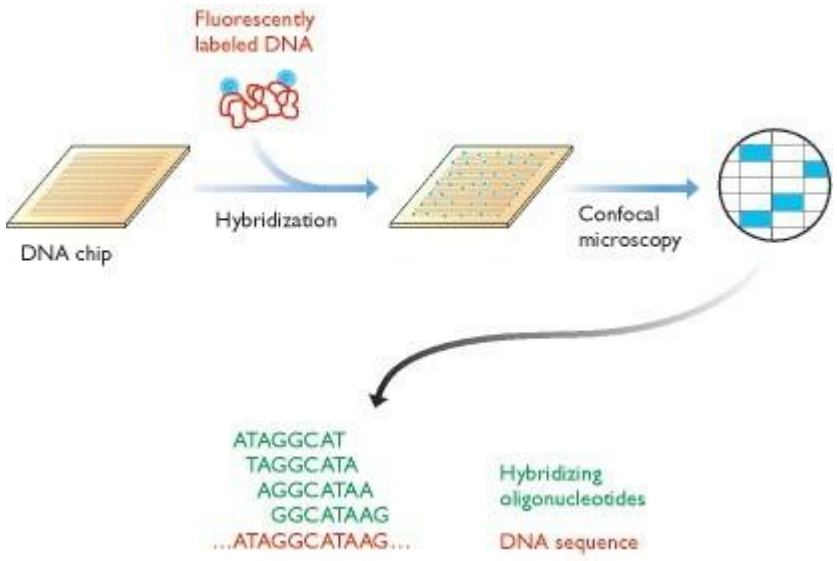


Figure The chip carries an array of every possible 8-mer oligonucleotide. The DNA to be sequenced is labeled with a fluorescent marker and applied to the chip, and the positions of hybridizing oligonucleotides determined by confocal microscopy. Each hybridizing oligonucleotide represents an 8-nucleotide sequence motif that is present in the probe DNA. The sequence of the probe DNA can therefore be deduced from the overlaps between the sequences of these hybridizing oligonucleotides.

DNA Annotation

DNA annotation or genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do. An annotation (irrespective of the context) is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it.

For DNA annotation, a previously unknown sequence representation of genetic material is enriched with information relating genomic position to intron-exon boundaries, regulatory sequences, repeats, gene names and protein products. This annotation is stored in genomic databases such as Mouse Genome Informatics, FlyBase, and Worm Base. Educational materials on some aspects of biological annotation from the 2006 Gene Ontology annotation camp and similar events are available at the Gene Ontology website.

The National Center for Biomedical Ontology develops tools for automated annotation of database records based on the textual descriptions of those records.

As a general method, dcGO has an automated procedure for statistically inferring associations between ontology terms and protein domains or combinations of domains from the existing gene/protein-level annotations.

Process:

Genome annotation consists of three main steps:.

1. identifying portions of the genome that do not code for proteins
2. identifying elements on the genome, a process called gene prediction, and
3. attaching biological information to these elements.

Automatic annotation tools try to perform all this by computer analysis, as opposed to manual annotation (a.k.a. curation) which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline.

The simplest way to perform gene annotation relies on homology based search tools, like BLAST, to search for homologous genes in specific databases, the resulting information is then used to annotate genes and genomes. However, nowadays more and more additional information is added to the annotation platform. The additional information allows manual annotators to deconvolute discrepancies between genes that are given the same annotation. Some databases use genome context information, similarity scores, experimental data, and integrations of other resources to provide genome annotations through their Subsystems approach. Other databases (e.g.Ensembl) rely on both curated data sources as well as a range of different software tools in their automated genome annotation pipeline.

Structural annotation consists of the identification of genomic elements.

- ORFs and their localization
- Gene structure
- Coding regions
- location of regulatory motifs

Functional annotation consists of attaching biological information to genomic elements.

- Biochemical function
- Biological function
- involved regulation and interactions
- expression

These steps may involve both biological experiments and in silico analysis. Proteo-genomics based approaches utilize information from expressed proteins, often derived from mass spectrometry, to improve genomics annotations.

A variety of software tools have been developed to permit scientists to view and share genome annotations. Genome annotation remains a major challenge for scientists investigating the human genome, now that the genome sequences of more than a thousand human individuals and several model organisms are largely complete. Identifying the locations of genes and other genetic control elements is often described as defining the biological "parts list" for the assembly and normal operation of an organism. Scientists are still at an early stage in the process of delineating this parts list and in understanding how all the parts "fit together".[

Genome annotation is an active area of investigation and involves a number of different organizations in the life science community which publish the results of their efforts in publicly available biological databases accessible via the web and other electronic means. Here is an alphabetical listing of on-going projects relevant to genome annotation:

- Encyclopedia of DNA elements(ENCODE)
- Entrez Gene
- Ensembl
- GENCODE
- Gene Ontology Consortium
- GeneRIF
- RefSeq
- Uniprot
- Vertebrate and Genome Annotation Project (Vega)

Base calling and sequence accuracy:

Base calling is the process by which an order of nucleotides in a template is inferred during a sequencing reaction. Next generation sequencing platforms that use fluorescently labeled reversible terminators have a unique color for each base. These are incorporated into the complementary strand of the DNA template and captured with a sensitive CCD camera. These images are processed into signals which are used to infer the order of nucleotides, also known as base calling. While sequencing platforms typically have integrated base calling software, the development of high performing base calling algorithms is an area of

ongoing research.

Base calling accuracy is typically measured by a Q score (Phred quality score), a common metric to assess the accuracy of a sequencing run. Q scores are defined as logarithmically related to base calling error probability.

$Q = - 10 \log P / \log 10$

If a sequencing run is assigned a Q score of 40, this is equal to the probability of an incorrect base call of 1 in 10,000 times, or 99.99% base calling accuracy.

Q Score 10 - Base calling accuracy 1 in 10 - Probability of incorrect base 90% Q Score 20 - Base calling accuracy 1 in 100 - Probability of incorrect base 99%
Q Score 30 - Base calling accuracy 1 in 1,000 - Probability of incorrect base 99.9%

Q Score 40 - Base calling accuracy 1 in 10,000 - Probability of incorrect base 99.99%

Q Score 50 - Base calling accuracy 1 in 100,000 - Probability of incorrect base 99.999%

A lower Q score of 10 means, there is the probability of an incorrect call in 1 of 10 bases. Lower Q scores can lead to increases in false positive variant calls and reduces the overall confidence an investigator has in their sequencing data

Probable Questions:

1. How DNA markers can be used for genetic mapping?
2. Write down the importance of SNP markers in genetic mapping.
3. Describe basic methodology of restriction mapping.
4. What is optical mapping. Describe the technique.
5. Describe FISH method. What are its significance.
6. Describe STTS mapping.
7. Describe the methodology of DNA sequencing?
8. What is high throughput DNA sequencing?
9. What do you mean by base calling?
10. What is DNA annotation?

Suggested Readings:

1. Jing JP, Lai ZW, Aston C. et al. Optical mapping of Plasmodium falciparum chromosome 2. Genome Res. (1999);9:175–181. [PMC free article][PubMed]
2. Lichter P, Tang CJ, Call K. et al. High resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. Science. (1990);247:64–69.[PubMed]
3. Lin J, Qi R, Aston C. et al. Whole-genome shotgun optical mapping of Deinococcusradiodurans. Science. (1999);285:1558–1562.[PubMed]
4. Marra MA, Hillier L, Waterston RH. Expressed sequence tags - ESTablishingbridges between genomes. Trends Genet. (1998);14:4–7.[PubMed]
5. McCarthy L. Whole genome radiation hybrid mapping. Trends Genet.(1996);12:491–493. [PubMed]
6. Oliver SG, van der Aart QJM, Agostoni-Carbone ML. et al. The complete DNA

sequence of yeast chromosome III. Nature. (1992);357:38–46.[PubMed]

7. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang Y-K. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. Science.(1993);262:110–114.[PubMed]

8. SNP Group (The International SNP Map Working Group). A map of human genome sequence variation containing 1.42 million singlenucleotide polymorphisms. Nature. (2001);409:928–933. [PubMed]

9. Sturtevant AH. The linear arrangement of six sex-linked factors in Drosophila asshown by mode of association. J. Exp. Zool.(1913);14:39–45.

10. Yamamoto F, Clausen H, White T, Marken J, Hakamori S. Molecular genetic basis of the histo-blood group ABO system. Nature. (1990);345:229–233.[PubMed]

11. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc. Natl Acad. Sci. USA. (1977);74:560–564.[PMC free article][PubMed]

12. Mullikan JC, McMurray AA. Sequencing the genome, fast. Science. (1999);283:1867– 1868.[PubMed]

13. Murray JC, Buetow KH, Weber JL. et al. A comprehensive human linkage map with centimorgan density. Science. (1994);265:2049–2054.[PubMed]

14. Prober JM, Trainor GL, Dam RJ. et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science.(1987);238:336–341. [PubMed]

15. Rogers J. Gels and genomes. Science. (1999);286:429.[PubMed]

16. Ronaghi M, Ehleen M, Nyrn P. A sequencing method based on real-time pyrophosphate. Science. (1998);281:363–365. [PubMed]

17. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain terminating inhibitors. Proc. Natl Acad. Sci. USA. (1977);74:5463–5467. [PMC freearticle] [PubMed]

18. Genomes. 2nd edition. Brown TA. Oxford: Wiley-Liss;2002.

UNIT-V

Microbial genetics: organization of prokaryotic genome; single stranded DNA phages; RNA phages; cycle and gene expression in SV40 virus.

UNIT-VI

Lytic and lysogenic phage morphogenesis; bacterial conjugation, transduction and transformation

Objective: In this unit you will learn about organization of prokaryotic genome, single stranded DNA phages, RNA phages, cycle and gene expression in SV40 virus; Lytic and lysogenic phage morphogenesis, bacterial conjugation, transduction and transformation

Genome organization in prokaryotes: Much of what is known about prokaryotic chromosome structure was derived from studies of *Escherichia coli*, a bacterium that lives in the human colon and is commonly used in laboratory cloning experiments. In the 1950s and 1960s, this bacterium became the model organism of choice for prokaryotic research when a group of scientists used phase-contrast microscopy and autoradiography to show that the essential genes of *E. coli* are encoded on a single circular chromosome packaged within the cell nucleoid.

Prokaryotic cells do not contain nuclei or other membrane-bound organelles. Most of the well-characterized prokaryotic genomes consist of double-stranded DNA organized as a single circular chromosome 0.6-10 Mb in length and one or more circular plasmid species of 2 kb-1.7 Mb. In fact, the word "prokaryote" literally means "before the nucleus." The nucleoid is simply the area of a prokaryotic cell in which the chromosomal DNA is located. This arrangement is not as simple as it sounds, however, especially considering that the *E. coli* chromosome is several orders of magnitude larger than the cell itself. So, if bacterial chromosomes are so huge, how can they fit comfortably inside a cell-much less in one small corner of the cell? The answer to this question lies in DNA packaging. Whereas eukaryotes wrap their DNA around proteins called histones to help package the DNA into smaller spaces, most prokaryotes do not have histones (with the exception of those species in the domain Archaea). Thus, one way prokaryotes compress their DNA into smaller spaces is through supercoiling (Figure 1).

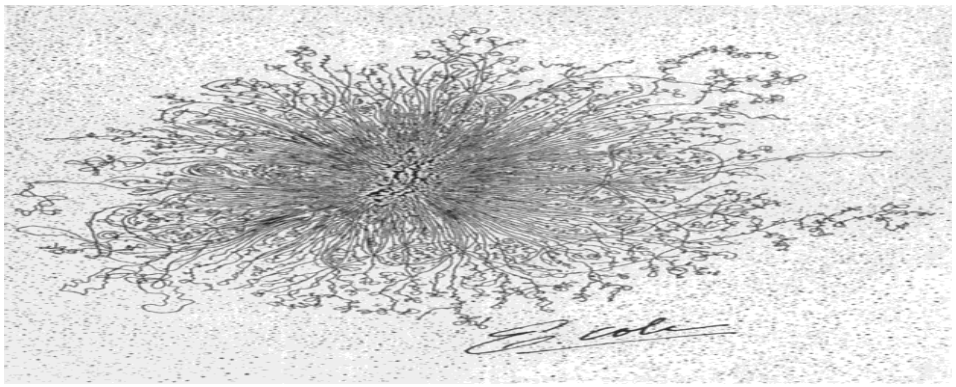


Fig. 1: Supercoiled DNA in prokaryotes (*E. coli)*

Imagine twisting a rubber band so that it forms tiny coils. Now twist it even further, so that the original coils fold over one another and form a condensed ball. When this type of twisting happens to a bacterial genome, it is known as supercoiling (Fig.2: a and b).

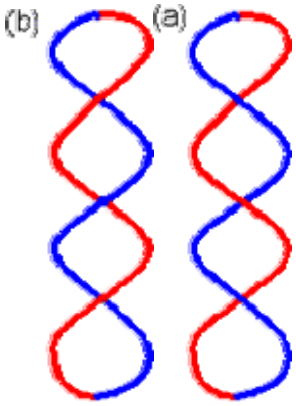


Fig.2: (a) Positive supercoils (the front segment of a DNA molecule cross over the back segment from left to right). (b) Negative supercoils.

Genomes can be negatively supercoiled, meaning that the DNA is twisted in the opposite direction of the double helix, or positively supercoiled, meaning that the DNA is twisted in the same direction as the double helix. Most bacterial genomes are negatively supercoiled during normal growth.

Proteins Involved in Supercoiling: During the 1980s and 1990s, researchers discovered that multiple proteins act together to fold and condense prokaryotic DNA. In particular, one protein called HU, which is the most abundant protein in the nucleoid, works with an enzyme called topoisomerase I to bind DNA and introduce sharp bends in the chromosome, generating the tension necessary for negative

supercoiling. Recent studies have also shown that other proteins, including integration host factor (IHF), can bind to specific sequences within the genome and introduce additional bends. The folded DNA is then organized into a variety of conformations that are supercoiled and wound around tetramers of the HU protein, much like eukaryotic chromosomes are wrapped around histones.

Once the prokaryotic genome has been condensed, DNA topoisomerase I, DNA gyrase, and other proteins help maintain the supercoils. One of these maintenance proteins, H-NS, plays an active role in transcription by modulating the expression of the genes involved in the response to environmental stimuli. Another maintenance protein, factor for inversion stimulation (FIS), is abundant during exponential growth and regulates the expression of more than 231 genes, including DNA topoisomerase-I.

Accessing Supercoiled Genes:

Supercoiling explains how chromosomes fit into a small corner of the cell, but how do the proteins involved in replication and transcription access the thousands of genes in prokaryotic chromosomes when everything is packaged together so tightly? It has been determined that prokaryotic DNA replication occurs at a rate of 1,000 nucleotides per second, and prokaryotic transcription occurs at a rate of about 40 nucleotides per second, so bacteria must have highly efficient methods of accessing their DNA strands.

Researchers have noted that the nucleoid usually appears as an irregularly shaped mass within the prokaryotic cell, but it becomes spherical when the cell is treated with chemicals to inhibit transcription or translation. Moreover, during transcription, small regions of the chromosome can be seen to project from the nucleoid into the cytoplasm (i.e., the interior of the cell), where they unwind and associate with ribosomes, thus allowing easy access by various transcriptional proteins. These projections are thought to explain the mysterious shape of nucleoids during active growth. When transcription is inhibited, however, the projections retreat into the nucleoid, forming the aforementioned spherical shape.

Because there is no nuclear membrane to separate prokaryotic DNA from the ribosomes within the cytoplasm, transcription and translation occur simultaneously in these organisms. This is strikingly different from eukaryotic chromosomes, which are confined to the membrane-bound nucleus during most of the cell cycle. In eukaryotes, transcription must be completed in the nucleus before the newly synthesized mRNA molecules can be transported to the cytoplasm to undergo translation into proteins.

Variations in Prokaryotic Genome Structure: Recently, it has become apparent that one size does not fit all when it comes to prokaryotic chromosome structure. While most prokaryotes, like *E. coli*, contain a single circular DNA molecule that makes up their entire genome, recent studies have indicated that some prokaryotes contain as many as four linear or circular chromosomes. For example, *Vibrio cholerae*, the bacteria that causes

cholera, contains two circular chromosomes. One of these chromosomes contains the genes involved in metabolism and virulence, while the other contains the remaining essential genes. An even more extreme example is provided by *Borrelia burgdorferi*, the bacterium that causes Lyme disease. This organism is transmitted through the bite of deer ticks, and it contains up to 11 copies of a single linear chromosome. Unlike *E. coli*, *Borrelia* cannot supercoil its linear chromosomes into a tight ball within the nucleoid; rather, these strands are diffused throughout the cell. Other organisms, such as *Bacillus subtilis*, form nucleoids that closely resemble those of *E. coli*, but they use different architectural proteins to do so. Furthermore, the DNA molecules of Archaea, a taxonomic domain composed of single-celled, non-bacterial prokaryotes that share many similarities with eukaryotes, can be negatively supercoiled, positively supercoiled, or not supercoiled at all. It is important to note that archaea are the only group of prokaryotes that use eukaryote like histones, rather than the architectural proteins described above, to condense their DNA molecules. The acquisition of histones by archaea is thought to have paved the way for the evolution of larger and more complex eukaryotic cells. Nobel Prize winner Arthur Kornberg used ΦX174 as a model to first prove that DNA synthesized in a test tube by purified enzymes could produce all the features of a natural virus. In 2003, it was reported by Craig Venter's group that the genome of ΦX174 was the first to be completely assembled *in vitro* from synthesized oligonucleotides. The ΦX174 virus particle has also been successfully assembled invitro.

This bacteriophage has a [+] circular single-stranded DNA genome of 5386 nucleotides encoding 11 proteins. Of these 11 genes, only 8 are essential to viral morphogenesis. The GC-content is 44% and 95% of nucleotides belong to coding genes. Infection begins when G protein binds to lipopolysaccharides on the bacterial host cell surface. H protein (or the DNA Pilot Protein) pilots the viral genome through the bacterial membrane of *E.coli*bacteria most likely via a predicted N- terminal transmembrane domain helix. Additionally, H protein induces lysis of the bacterial host at high concentrations as the predicted N-terminal transmembrane helix easily pokes holes through the bacterial wall. The DNA is ejected through a hydrophilic channel at the 5-fold vertex. It is understood that H protein resides in this area but experimental evidence has not verified its exact location. Once inside the host bacterium, replication of the [+] ssDNA genome proceeds via negative sense DNA intermediate. This is done as the phage genome supercoils and the secondary structure formed by such supercoiling attracts a primosome protein complex. This translocates once around the genome and synthesizes a [-]ssDNA from the positive original genome. [+] ssDNA genomes to package into viruses are created from this by a rolling circle mechanism. This is the mechanism by which the double stranded supercoiled genome is nicked on the positive strand by a virus-encoded A protein, also attracting a bacterial DNA polymerase (DNAP) to the site of cleavage. DNAP will use the negative strand as a template to make positive sense DNA. As it translocates around the genome it displaces the outer

strand of already-synthesised DNA, which is immediately coated by SSBP proteins. The A protein will cleave the complete genome every time it recognizes the origin sequence.
As D protein is the most abundant gene transcript, it is the most protein in the viral procapsid. Similarly, gene transcripts for F, J, and G are more abundant than for H as the stoichiometry for these structural proteins is 5:5:5:1. The primosomes are protein complexes which attach/bind the enzyme helicase on the template. Primosomes gives RNA primers for DNA synthesis to strands. Phi X is regularly used as a positive control in DNA sequencing due to its relatively small genome size in comparison to other organisms, its relatively balanced nucleotide content- about23% G, 22% C, 24% A, and 31% T, i.e., 45% G+C and 55% A+T, for its 5,386 nucleotide long sequence.

RNA phages: Bacteriophages occur abundantly in the biosphere, with different genomes, and lifestyles. Phages are classified by the International Committee on Taxonomy of Viruses (ICTV) according to morphology and nucleic acid. Nineteen families are currently recognized by the ICTV that infect bacteria and archaea. Of these, only two families have RNA genomes, and only five families are surrounded by an envelope.

Family	**Morphology**	**Nucleic acid**	**Examples**
Cystoviridae	Enveloped, spherical	Segmented dsRNA	
Leviviridae	Nonenveloped, isometric	Linear ssRNA	MS2, Q

Table 1: ICTV classification of prokaryotic (bacterial and archaeal) viruses

Bacteriophage MS2

The bacteriophage MS2 is an icosahedral, positive-sense single-stranded RNA virus that infects the bacterium Escherichia coli and other members of the Enterobacteriaceae. MS2 is a member of a family of closely related bacterial viruses that includes bacteriophage f2, bacteriophage Qβ, R17, and GA. The MS2 genome is one of the smallest known, consisting of 3569 nucleotides of single-stranded RNA. It encodes just four proteins: the maturation protein (A-protein), the lysis protein, the coat protein, and the replicase protein. The gene encoding lysis protein (lys) overlaps both the 3'-end of the upstream gene (cp) and the 5'-end of the downstream gene (rep), and was one of the first known examples of overlapping genes. The positive-stranded RNA genome serves as messenger RNA, and is translated

upon viral uncoating within the host cell. Although the four proteins are encoded by the same messenger/viral RNA, they are not all expressed at the same levels; expression of these proteins is regulated by a complex interplay between translation and RNA secondary structure.

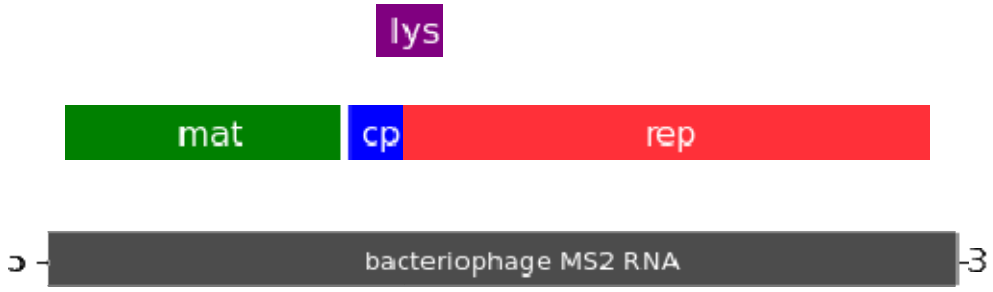


Fig.3: Location of protein-coding genes within bacteriophage MS2 RNA. Note that the lys gene overlaps segments of both the cp and rep genes. Scale is approximate.

Capsid structure: An MS2 virion (viral particle) is about 27 nm in diameter, as determined by electron microscopy. It consists of one copy of the maturation protein and 180 copies of the coat protein (organized as 90 dimers) arranged into an icosahedral shell with triangulation number T=3, protecting the genomic RNA inside. The structure of the coat protein is a five-stranded β- sheet with two α-helices and a hairpin. When the capsid is assembled, the helices and hairpin face the exterior of the particle, while the β-sheet faces the interior.

Life cycle: Once the viral RNA has entered the cell, it begins to function as a messenger RNA for the production of phage proteins. The gene for the most abundant protein, the coat protein, can be immediately translated. The translation start of the replicase gene is normally hidden within RNA secondary structure, but can be transiently opened as ribosomes pass through the coat protein gene. Replicase translation is also shut down once large amounts of coat protein have been made; coat protein dimers bind and stabilize the RNA "operator hairpin", blocking the replicase start. The start of the maturation protein gene is accessible in RNA being replicated but hidden within RNA secondary structure in the completed MS2 RNA; this ensures translation of only a very few copies of maturation protein per RNA. Finally, the lysis protein gene can only be initiated by ribosomes that have completed translation of the coat protein gene and "slip back" to the start of the lysis protein gene, at about a 5%frequency.

Replication of the plus-strand MS2 genome requires synthesis of the complementary minus strand RNA, which can then be used as a template for synthesis of a new plus strand RNA. MS2 replication has been much less well studied than replication of the highly related

bacteriophage Qβ, partly because the MS2 replicase has been difficult to isolate, but is likely to be similar.

The formation of the virion is thought to be initiated by binding of maturation protein to the MS2 RNA; in fact, the complex of maturation protein and RNA is infectious. The assembly of the icosahedral shell or capsid from coat proteins can occur in the absence of RNA; however, capsid assembly is nucleated by coat protein dimer binding to the operator hairpin, and assembly occurs at much lower concentrations of coat protein when MS2 RNA is present. Bacterial lysis and release of newly formed virions occurs when sufficient lysis protein has accumulated. Lysis protein forms pores in the cytoplasmic membrane, which leads to loss of membrane potential and breakdown of the cell wall, while the β-sheet faces the interior.

Applications:

Since 1998, the MS2 operator hairpin and coat protein have found utility in the detection of RNA in living cells. MS2 and other viral capsids are also currently under investigation as agents in drug delivery, tumor imaging, and light harvesting applications.MS-2, due to its structural similarities to noroviruses, its similar optimum proliferation conditions, and non-pathogenicity to humans, has been used as substitute for noroviruses in studies of disease transmission.

Cycle and gene expression in SV 40 virus:

In 1960, Sweet and Hilleman first described an agent, which they named SV40, induced cytopathic effects and vacuole formation in monkey cells. Since its discovery, simian virus 40 (SV40) has been one of the most intensely studied animal viruses. The molecular biology of SV40 has led to seminal discoveries in the fields of transcription, DNA replication, and oncogenic transformation. Over the last decade, provocative evidence has accumulated that suggests SV40 may be a human pathogen. Does SV40 infect humans? If so, when did this monkey polyomavirus enter the human population and where is the reservoir? What is the behavior of SV40 in human cells? Does it cause or contribute to acute or chronic disease?

SV40 was isolated from normal monkey kidney cells, stocks of the Sabin poliovirus vaccine, and an adenovirus vaccine. The last two reagents were prepared in primary kidney cell cultures derived from rhesus monkeys. Subsequent analyses found that the Salk poliovirus vaccine administered from 1955 to 1963 in the United States was also contaminated with SV40, potentially exposing an estimated 100 million people. Although poliovirus in the Salk vaccine was inactivated by formalin treatment, the conditions were insufficient to completely inactivate SV40. Soon thereafter, it was demonstrated that SV40 could infect humans and also induce tumors in experimental animals. These observations raised concerns that vaccinated people worldwide may have been inadvertently exposed to an oncogenic virus. Early epidemiological studies allayed these fears, revealing no increased incidence of cancers directly related to immunization status. However, these initial analyses were necessarily limited in that it was unknown whether (i) the virus could be transmitted, either horizontally

or vertically; (ii) vaccinated, immunocompetent individuals would be at equal risk for development of cancer with others having defective immunity or a cancer predisposition; (iii) the power of the analysis was sufficient to detect increases in rare cancers; and (iv) SV40 normally circulated in humans before development of the poliovirus vaccine. A recent review of all epidemiological data by the Institute of Medicine concluded that evidence to date was "inadequate to accept or reject a causal relationship between SV40-containing poliovirus vaccines and cancer". Criticisms included misclassification bias, lack of confidence intervals for the data, and "ecological" study design, which are unlikely to be remedied by further follow-up of the study populations.

A brief overview of the biology of SV40 is relevant to understand the concerns raised by these initial analyses. When SV40 infects its natural host, it initially undergoes a lytic replication cycle. The early viral genes encode the tumor (T) antigens: large T antigen (LT), small t antigen (ST), and 17K T antigen (also tiny T or T′). LT plays a dominant role in infection, repressing early viral gene transcription and stimulating late viral gene transcription. LT is also an initiation factor for viral DNA replication, recruiting the DNA polymerase α-primase complex to the origin of replication and acting as a helicase. Following the strategy of other DNA viruses, the SV40 early proteins dysregulate the cell cycle and impede cell apoptosis in order to maximize virus production. LT binds the members of the retinoblastoma protein family, pRb, p107, and p130, resulting in release and activation of E2F transcription factors, which stimulate expression of genes involved in S-phase progression and DNA synthesis. LT also binds p53 and inactivates its function, preventing the infected cell from undergoing apoptotic cell death. After viral DNA replication is under way, the infection enters the late phase, when viral structural proteins are synthesized and new virions are produced. Ultimately, the infected cell releases progeny virions, frequently but not always by cell lysis. The immune system is critical for controlling the initial lytic phase in vivo, quenching the initial infection to a state of persistent low-level or nonreplicating genomes (i.e., in the proximal renal tubular epithelium for SV40), with detectable lytic viral reactivation coincident only with host immune suppression.

Some data on the infectivity of SV40 in humans were obtained from volunteers and individuals receiving contaminated vaccines. However, antibody data from many surveys must be viewed with the knowledge that the human BK virus (BKV) and JC virus (JCV) (closely related human polyomavirus family members) might give an indistinguishable response in these assays due to the high degree of cross-reactivity between capsid protein antigens. Melnick and Stinebaugh found SV40 (by cytopathic effects in monkey cells) in the stools of children 3 to 4 weeks after ingestion of 100 to 1,000 PFU of SV40 with oral poliovirus vaccine. Morris et al. gave SV40 intranasally to volunteers and found subclinical infections. They were able to isolate virus 7 to 11 days after administration from 3 of 8 subjects, and they detected antibody responses of various amplitudes. Horváth and Fornosi found SV40 excreted in the stools of 10 of 35 children 1 to 2 weeks after being given

contaminated oral poliovirus vaccines. Thus, SV40 may replicate in humans after oral administration, but the efficiency and duration of the replication may be low in these immunocompetent subjects who were given small inocula.

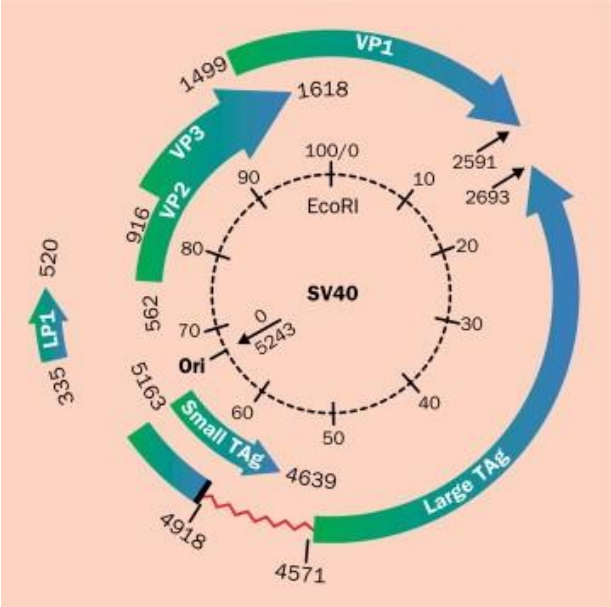


Fig 4: Genome organization of SV40 virus

The biology of SV40 in human cells was first studied in the 1960s with fibroblast cell lines or primary human fibroblast cell cultures. Whereas uninfected primary human fibroblasts can only be passaged a finite number of times before ceasing to divide and undergoing senescence, cell cultures infected with SV40 undergo a "crisis" at this same stage, followed by the outgrowth of a small number of cells that are phenotypically transformed. During the initial phase of the infection, generally the first 4 weeks, approximately 0.1% of the cells produce 500 to 1,000 virions per cell. Virus output from the culture then remains at a constant, very low level but with 100% of the cells producing virus at a rate of approximately 1 to 2 virions/cell. Once the cell culture progresses through crisis, virus production generally decreases, accompanied by a concomitant decrease in production of viral capsid proteins and an increase in the production of LT. One interpretation of these data is that the cells producing large amounts of virus are killed, but the cells that produce low levels of virus (as assayed by infectious center assays) survive. Finally, as the culture reaches its passage limit, most cells die, but those expressing a threshold level of LT overgrow the culture. Interestingly, the onset of transformation varies quite significantly in cells isolated from different individuals, ranging from 20 to almost 50 weeks in culture. Based on these early studies, human cells were termed semi permissive for SV40 growth. This nomenclature is confusing since the virus can clearly replicate in some human cell types more efficiently than in others, although the development of cytopathic effect is more rapid in African green monkey kidney cells.

SV40 is highly oncogenic in experimental animals and readily transforms rodent cells in culture. Hamsters inoculated with SV40 develop lymphomas, brain tumors, osteosarcomas, and mesotheliomas. SV40 is likely oncogenic in rodents because LT is unable to interact functionally with the rodent DNA polymerase α-primase complex. In this setting, the oncogenic functions of the T antigens are engaged but the productive cycle is not completed, resulting in uncontrolled cell division rather than cell lysis. LT is both necessary and sufficient for initiation and maintenance of transformation of rodent cells in tissue culture in most instances. Under certain conditions, however, usually involving primary cells in the absence of growth factors, ST is also required. ST functions by inhibiting the activity of the cellular phosphatase PP2A, resulting in activation of cell growth signal transduction pathways. Mice that are transgenic for LT transcriptionally regulated by tissue-specific promoters develop tumors in those tissues (for an example, see reference 105). Transgenic mice in which LT expression is regulated by the native viral promoter elements specifically develop tumors of the choroid plexus (6), the specialized epithelial structure of the brain ependymal lining that produces cerebrospinal fluid. This finding is interesting in view of the discovery of SV40 DNA in certain brain tumors, as discussed below. After the discovery of SV40's tumorigenic and cell transformation properties, a wave of studies in the 1960s and 1970s pursued the identification of viral oncogenic agents in humans. SV40 DNA was detected on rare occasions, usually in brain tumors, using relatively low-sensitivity Southern hybridization techniques, immunostaining for LT, and electron microscopy. Also during this period the distinctly human polyomaviruses BKV and JCV were identified, and the destructive brain white matter disease progressive multifocal leukoencephalopathy was attributed to JCV infection. These human viruses were also shown to induce tumors in animals and to transform rodent and human cells in culture. However, virtually all investigations failed to reveal any significant associations between human malignancy and these suspected oncogenic viruses. With the discovery of oncogenes, the emphasis in cancer research shifted from viruses to genomic mutations.

Lytic and Lysogenic cycles - phage multiplication cycle:A Lytic or virulent phages are phages which can only multiply on bacteria and kill the cell by lysis at the end of the life cycle.

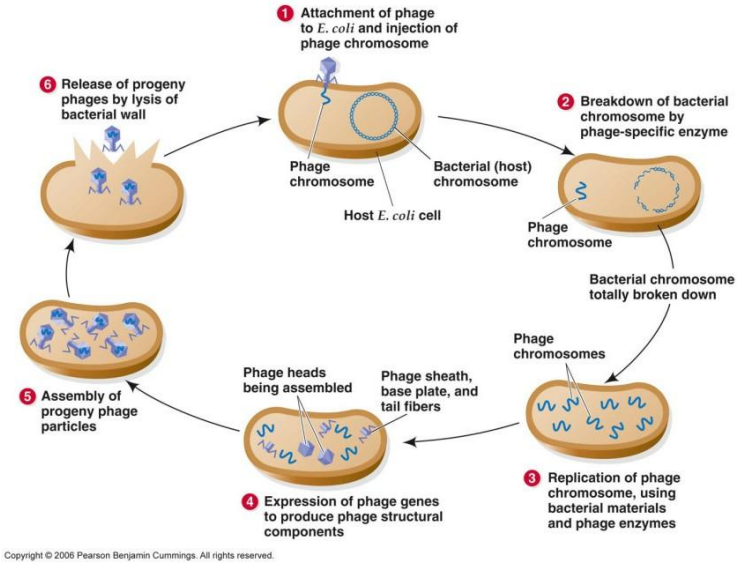


Fig 5: Lytic cycle of Phage

Steps: This cycle consists of the following steps

a. Eclipse period - During the eclipse phase, no infectious phage particles can be found either inside or outside the bacterial cell. The phage nucleic acid takes over the host biosynthetic machinery and phage specified m-RNA's and proteins are made. There is an orderly expression of phage directed macromolecular synthesis, just as one sees in animal virus infections. Early m- RNA's code for early proteins which are needed for phage DNA synthesis and for shutting off host DNA, RNA and protein biosynthesis. In some cases the early proteins actually degrade the host chromosome. After phage DNA is made late m-RNA's and late proteins are made. The late proteins are the structural proteins that comprise the phage as well as the proteins needed for lysis of the bacterial cell.

b. Intracellular Accumulation Phase - In this phase the nucleic acid and structural proteins that have been made are assembled and infectious phage particles accumulate within the cell. c. Lysis and Release Phase - After a while the bacteria begin to lyse due to the accumulation of the phage lysis protein and intracellular phage are released into the medium. The number of particles released per infected bacteria may be as high as1000.

B. Lysogenic or Temperate Phage

Lysogenic or temperate phages are those that can either multiply via the lytic cycle or enter a quiescent state in the cell. In this quiescent state most of the phage genes are not transcribed; the phage genome exists in a repressed state. The phage DNA in this repressed state is called a prophage because it is not a phage but it has the potential to produce phage. In most cases the phage DNA actually integrates into the host chromosome and is replicated along with the host chromosome and passed on to the daughter cells. The cell harboring a prophage is not adversely affected by the presence of the prophage and the lysogenic state may persist indefinitely.The cell harboring a prophage is termed a lysogen.

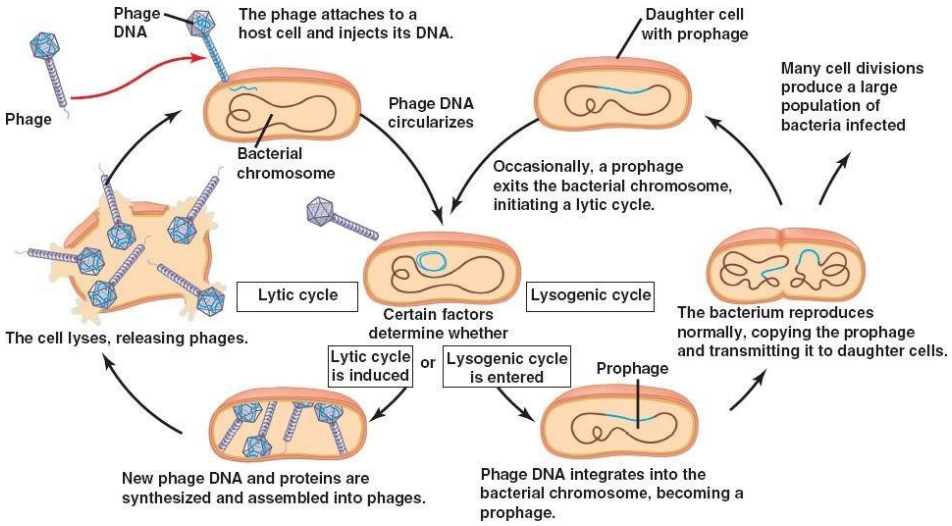


Fig 6: Lytic and Lysogenic cycle of phages

Significance of Lysogeny:

(a) Model for animal virus transformation -Lysogeny is a model system for virus transformation of animal cells.

(b) Lysogenic conversion-When a cell becomes lysogenized, occasionally extra genes carried by the phage get expressed in the cell. These genes can change the properties of the bacterial cell. This process is called lysogenicorphageconversion.Thiscanbeofsignificanceclinically.e.g.Lysogenic phages have been shown to carry genes that can modify the Salmonella O antigen, which is one of the major antigens to which the immune response is directed. Toxin production by *Corynebacterium diphtheriae* is mediated by a gene carried by a phage. Only those strains that have been converted by lysogeny are pathogenic.

Methods of Sexual Reproduction in Bacteria:

There are three methods by which sexual reproduction of Bacteria take place. They are described below

A. Conjugation:

Lederberg and Tatum (1946) discovered conjugation in E. coli and its detailed studies were made by Woolman and Jacob (1956). Conjugation, is a process by which genetic material is transferred from one bacterial cell ("donor cell" or "male cell") to another ("recipient cell" or "female cell") through a specialized intercellular connection called sex-pilus or conjugation tube.

The maleness and femaleness of bacterial cells are determined by the presence or absence of F-plasmid (also called F-factor or sex factor). F- plasmid, an extra chromosomal genetic material, is always present in the cytoplasm of donor or male cells, and the latter develop specialized cell surface appendages called F-pili or sex-pili under the control of F-plasmid. Recipient or female cells always lack F-pIasmids and, therefore, F-pili are not present on their surface.

F-plasmid or F-factor can exist in two different states:

(i) The autonomous state in which it lies free in the cytoplasm and replicate independent of the bacterial chromosome (DNA); a donor or male cell containing F- factor in autonomous state is called F^+ cell, and
(ii) The integrated state in which it is integrated (inserted) into the bacterial chromosome (DNA) and replicate along with it; a donor or male cell containing F-factor in integrated state is called Hfr cell (for high frequency recombination) or high frequency male cell. However, the recipient or female cell lacks F-factor and this is called F^- cell.

1. Conjugation between a F^+ (donor) cell and a F^- (recipient) cell:

In conjugation between a F^+ (donor) cell and a F^- (recipient) cell, it is the autonomous F-factor (F-plasmid) which is transferred, never the bacterial DNA (Fig. 29.2). When the two cells (F^+ and F^-) come close to each other, the F-pilus of the F^+ (donor) cell attaches with the F^- (recipient) cell and acts as a conjugation tube.
Simultaneously, the double-stranded circular F-factor DNA is nicked at a specific point, and begins to replicate producing a single-stranded copy of the F- factor DNA, which migrates through the tube into the cytoplasm of the F^- (recipient) cell.
It becomes double- stranded, and circulars and lies free in the cytoplasm thus rendering the recipient cell to become F^+ donor cell. In this way, mixing a population of F^+ (donor) cells with a population of F^+ (recipient) cells result in the conversion of virtually all the cells in the population becoming F^+ (donor) cells.

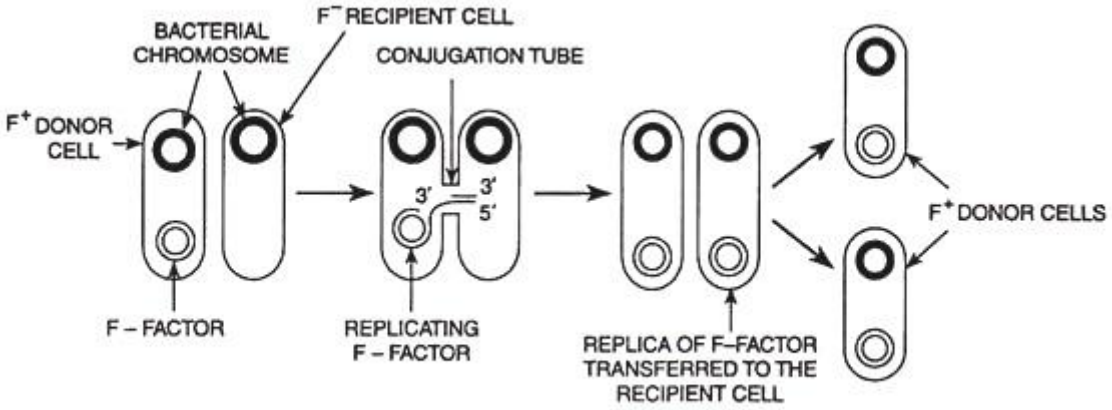


FIG. 29.2. Diagrammatic representation of conjugation between F⁺ (donor) cell and F⁻ (recipient) cell.

2. Conjugation between Hfr Donor Cells and Recipient (F^-) Cell:

The Hfr donor cells are considered to be fertile because, unlike F^+ (donor) cells, their chromosomal segments are transferred from donor to recipient cells and the F-factor remains in situ.

When the two cells (Hfr and F^-) come in contact, a conjugation tube develops between them. The circular DNA of Hfr donor cell is nicked and replication is initiated. The integrate F-factor always lies at the rear end of the DNA molecule. The replication of DNA starts towards the end near the conjugation tube and the newly synthesized single strand starts migrating through the tube into the recipient (F^-) cell.

In nature, the mating of two cells exists for a short period and gets interrupted resulting in the migration of only a portion of the donor DNA into the recipient cell. Since the F-factor lies at the rear end of the molecule, it is rarely transferred to the recipient cell.

The genes of the newly entered DNA fragment may replace the homolgous genes of the DNA of the recipient cell, resulting in a recombinant genetic material. The newly formed recombinant genetic material now possesses those male characters that have been transferred through recombination with the migrated DNA fragment (Fig. 29.3).

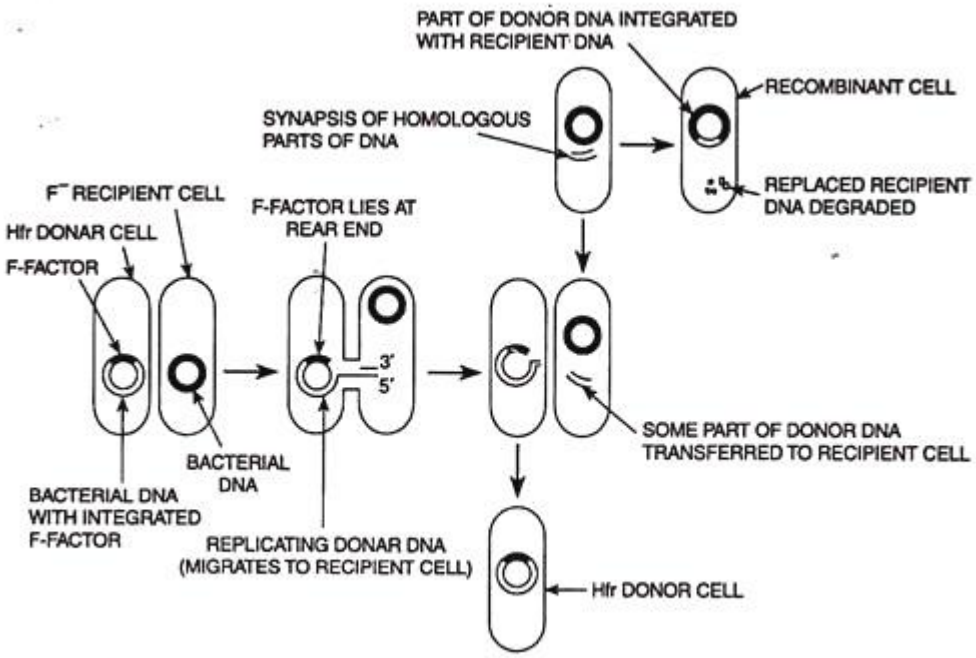


FIG. 29.3. Diagrammatic representation of conjugation between Hfr donor cell and F⁻ (recipient) cell.

3. Conjugation between F' (F-prime) Male and F⁻ (Recipient) Cell (Sex-Duction):

Existence of Hfr donor cells is not absolute. The F-factor integrated into the bacterial DNA of Hfr donor cells may dissociate and become free in the cytoplasm (Fig. 29.4).The dissociation may be occasionally anomalous during which the dissociated F-factor may bring with it some genes of the bacterial chromosome. Adelberg and Burns (1958) first identified such a modified F-factor and called it F' ("F-prime") factor; the donor cell possessing this factor is called F' (F-prime) male.

When a F' male conjugates with F⁻ (recipient) cell, the F'-factor is transferred from donor to the recipient cell, and such a recipient bacterial cell becomes heterozygous (merozygous) for that part of the bacterial chromosome, which the F'- factor had obtained during its anomalous dissociation.
Transfer of F'-factor to recipient cell apparently occurs by the same mechanism as F-factor, transfers during in F⁺ and F⁻ mating and chromosome transfer in Hfr and F⁻ cell mating. Genetic recombination of this type, mediated by F'-factor, is called sex-duction or F-duction.

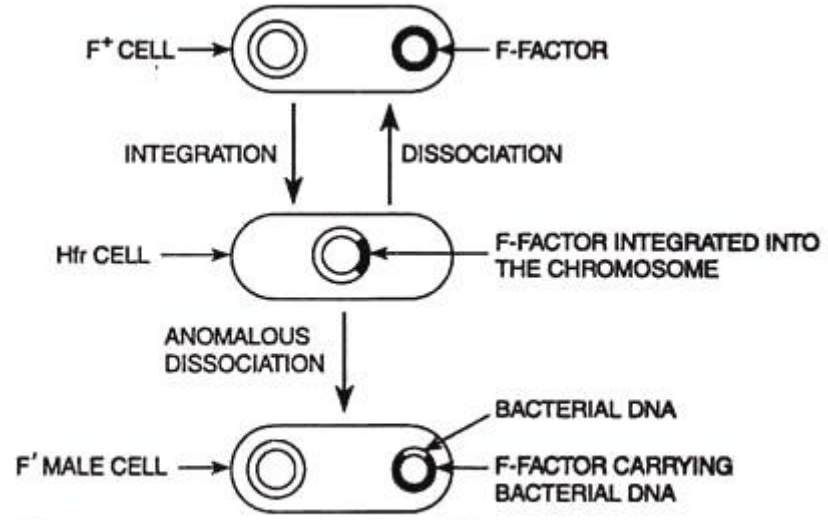


FIG. 29.4. Integration and dissociation of F-factor resulting in different types of donor (male) cells in bacteria.

B. Transformation:

This process of genetic recombination was first studied by Griffith (1928), an English bacteriologist. He took two strains of the bacterium Streptococcus pneumoniae (= Pneumococcus pneumoniae), then called Diplococcus pneumoniae. One of the two strains was virulent or pathogenic and capsulated normal; it formed smooth colonies. The other strain was non-pathogenic or avirulent and non-capsulated; it formed rough colonies on the culture medium.

He experimented on mice as summarised below:

Virulent strain (capsulated) ⟶ Injected into mice ⟶ Mice died
Avirulent strain (non-capsulated) ⟶ Injected into mice ⟶ No effect on mice
Heat-killed (dead) virulent cells ⟶ Injected into mice ⟶ No effect on mice
Avirulent cells + Heat killed virulent cells ⟶ Injected into mice ⟶ Mice died

Dead mice $\xrightarrow{\text{Isolation of bacteria}}$ Virulent strain (capsulated)

It is obvious from Griffith's experiment that the avirulent or non-pathogenic strain becomes virulent or pathogenic when mixed with heat-killed (dead) virulent strain thus causing the death of mice. Griffith named this change of avirulent into virulent strain as 'transformation'.
Griffith reasoned that there was a transfer of some factor from heat-killed (dead) virulent strain to the avirulent strain and called it "transforming principle". Further, he said that the transforming principle was the polysaccharide of capsule of heat-killed virulent strains.

The idea of polysaccharide as transforming principle came to an end in 1944 when Avery, MacLeod and McCarty showed that it is the DNA which works as transforming principle not the polysaccharide of the capsule. They proved for the first time that DNA is the genetic material in organisms.In transformation

(Fig. 29.5) a free (naked) DNA molecule is transferred from a donor to a recipient bacterial cell. The donor bacterium undergoes lysis to free the DNA molecule and the recipient bacterium must be competent to receive it.

This competence of bacterial cell is not a permanent feature; it has been demonstrated in relatively few bacterial genera and depends upon the growth phase of bacteria and the environmental conditions.When donor DNA comes into contact with the competent bacterial cell, it first binds on the cell surface and then is taken up inside the cell. In some of the cases, it is observed that the double-strand (ds) DNA enters inside the bacterial cell as such and its one strand is degraded by endonuclease enzyme therein leaving single-strand (ss) DNA whereas in others such as some species of Bacillus and Streptococcus it appears that only single-strand (ss) DNA enters the recipient bacterial cell.

An endonuclease enzyme now degrades one of the strands of dsDNA of recipient bacterial chromosome in corresponding region and this gap is filled by the donor ssDNA with the help of ligase enzyme which joins it with the DNA of the recipient bacterial chromosome. If the allelic forms of the donor and recipient genes are not identical, the donor DNA forms a heteroduplex with the recipient bacterial DNA.

When the bacterial cell containing 'heteroduplex' undergoes binary fission the heteroduplex replicates forming two 'homoduplexes'. One of these is a normal duplex which is all recipient in origin and the daughter cell containing it is like the recipient bacterial cell.The other homoduplex is a transformed duplex (hybrid genome) different from that of either the donor or the recipient bacterial genome. The daughter cell containing transformed duplex is a 'transformed cell' and contains some of the characteristics of the donor bacterial cell which are inherited progeny to progeny.

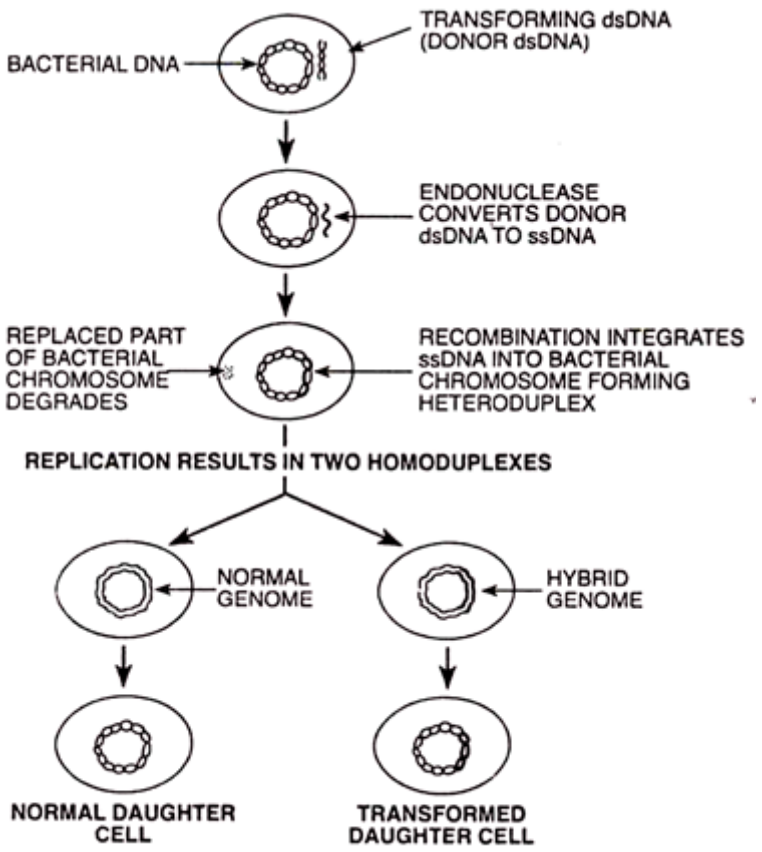


FIG. 29.5. Mechanism of transformation.

C. Transduction:

This process of genetic recombination was discovered by Zinder and Lederberg (1952) in Salmonella typhimurium during their experiments with the objective of discovering whether E. coli type of genetic exchange also existed in S. typhimurium.

In contrast to transformation, wherein free (naked) DNA is transferred, fragments of DNA are transferred from one bacterial cell to the other with the help of a viral carrier (bacteriophage) during transduction i.e., the transduction is a phage-mediated process of genetic material transfer in bacteria.

The bacteriophage acquires a portion of the bacterial DNA of the host cell in which it reproduces and then transfers this acquired DNA to another bacterial cell to which it infects. Such bacteriophage is called 'transducingphage'. Transduction is of the following two types: generalized (non- specialised) and specialized (restricted).

1. Generalized Transduction:

Transduction, which results in transfer of any bacterial gene from one bacterial cell to the other is referred to as generalized or non-specialized transduction. It is mediated by some virulent phagesand

certain temperate phages; E. coli phage P1, Salmonella phage P22, and Bacillus subtilis phages PBS1 and SP10 are such phages.

In generalized transduction (Fig. 29.6), some of the developing progeny phages, during their normal lytic- cycle may accidentally acquire pieces of bacterial DNA. Such phages, after the lysis of the host bacterial cell and their release, attach to and inject their DNA into a new recipient cell but fail to re-establish lytic-cycle therein.Once inside the recipient bacterial cell, the injected DNA may be degraded by nucleases, in which case genetic exchange does not occur. The injected DNA, however, may undergo integration resulting in homologous recombination, as a result, the transduced cell may possess new combination of genes. The transduced bacterial cell now undergoes usual binary fission and produces progeny cells containing new combination of genes.

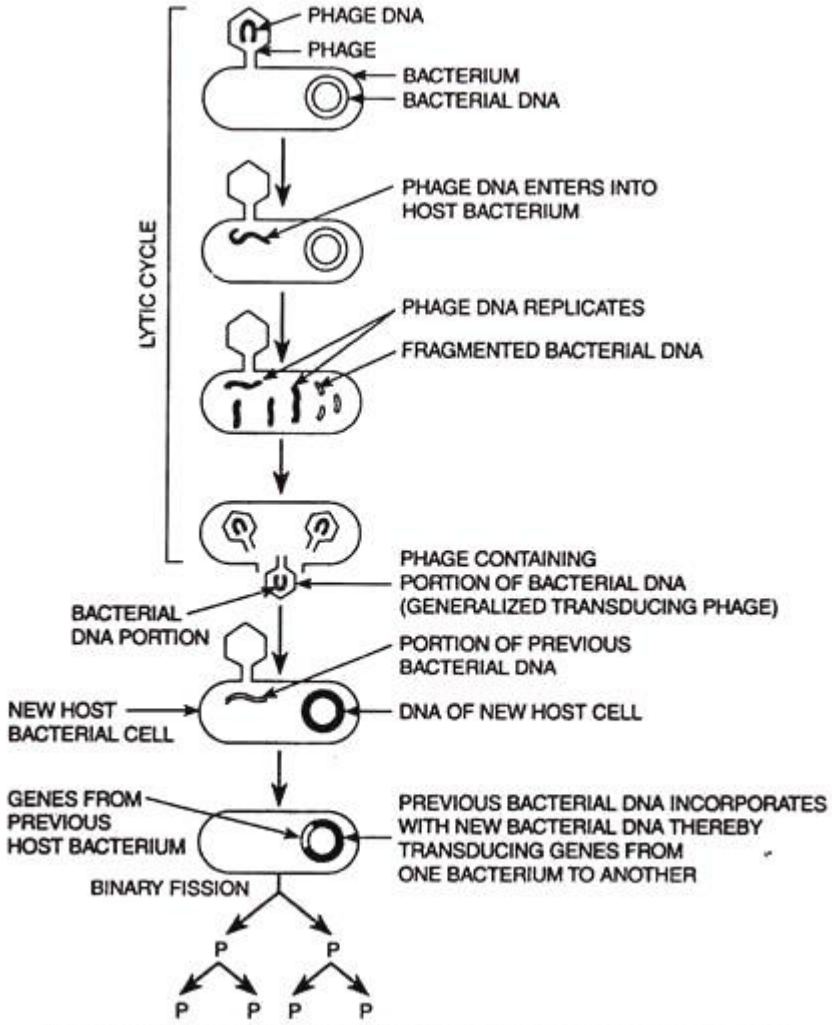


FIG. 29.6. Diagrammatic representation of generalised transduction. P = progeny.

2. Specialized Transduction:

In contrast to generalized (non-restricted) transduction, which results in transfer of any gene from donor to recipient bacterial cell, specialized (restricted) transduction is that which leads to the transfer of only specific (restricted) genes from donor to recipient cell.

Specialized transduction (Fig.29.8) is mediated by those temperate bacteriophages (e.g., lambda (λ) phage, mu (μ) phage and ϕ80 phage) that usually incorporate (integrate) their DNA into the bacterial chromosome.

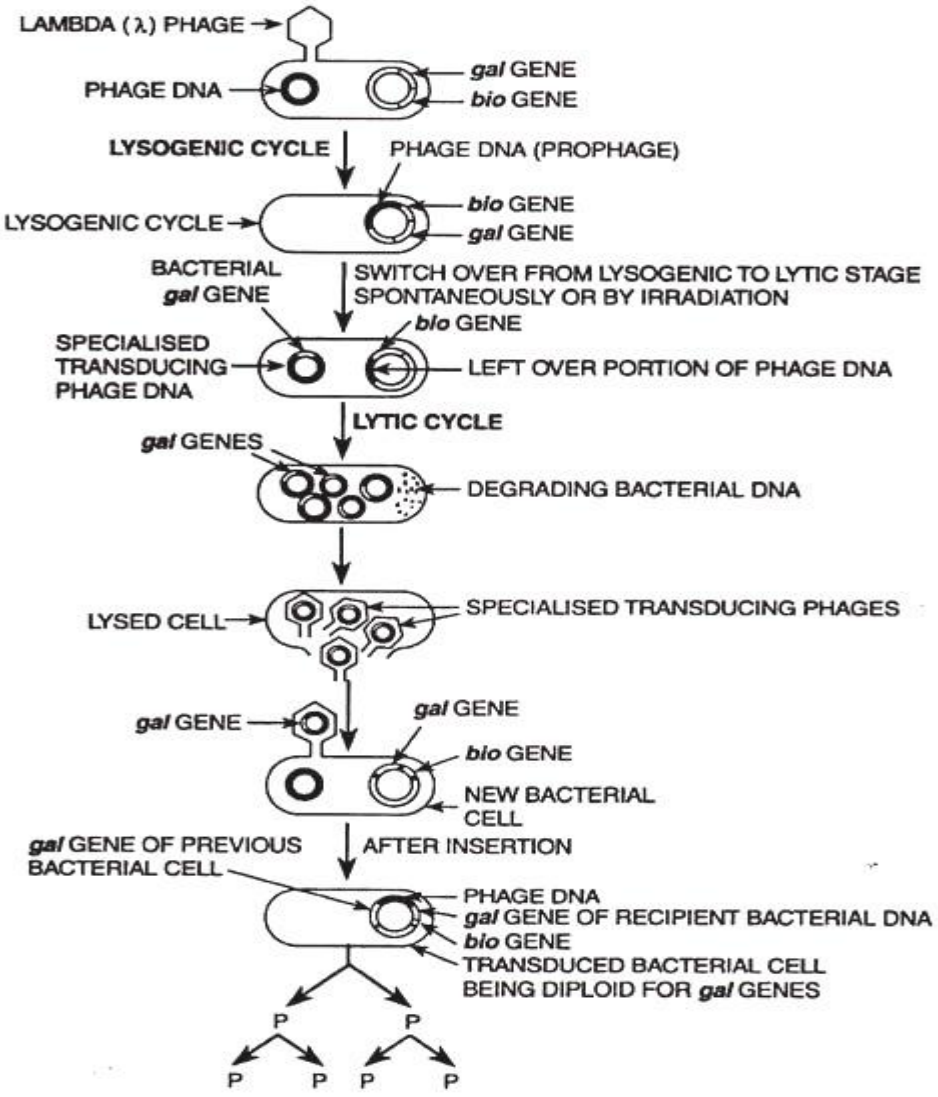


FIG. 29.8. Specialised transduction (diagrammatic). P = progeny.

The phage-DNA is called ‘prophage’ in its integrated state with the bacterial chromosome; the bacterium having a prophage is said to be lysogenic, and this phage-host-relationship is called lysogeny.

Lysogenic temperate phages spontaneously switch over from lysogenic to lytic state at a low rate (about one in 19^5 cell divisions) in nature, or they may be induced to do so by irradiation with ultraviolet light. During this transition, the prophage is usually excised precisely from the specific site of integration in its exactly original form.But occasionally, it may excise imprecisely so that it takes with it that specific portion of bacterial chromosome which lies close to the site of prophage insertion and leaves a portion of its own DNA remaining integrated within the bacterial chromosome.
Such prophage is called ‘specialized transducing principle’ and is packaged into a developing phage particle inside the host bacterial cell. Phage particle so developed is called ‘specialized transducing phage’ and is released after the host bacterial cell undergoes lysis.Only those specialized transducing phages are viable that contain an amount of greater than 73% and less than 110% of the phage-DNA. When a viable specialized-transducing-phage infects a new bacterial cell, its specialized-transducing principle that already contains specific portion of bacterial chromosome inserts into the recipient bacterial chromosome thus making the latter diploid for that specific bacterial gene (partial diploid or heterogenote or merogenote).

Since the specialized transducing phage is ‘defective’ phage as it has lost some genes during the excision, it functions in recipient bacterial cell only when the latter is already infected by another phage (termed as helper phage) that contains the missing genes and thereby complements the lost phage-functions of the specialized transducing-phage.The partial diploid contains two copies of the concerned genes, one from donor bacterium and other from recipient bacterium, and is unstable. As a result, bacterial cells containing gene of donor bacterium and those containing gene of recipient bacterium segregate at a frequency of about one in 1,000 cell divisions.
For example, lambda (λ) phage integrates between the gal genes (required for the utilization for galactose as an energy source) and the bio genes (essential for the synthesis of biotin amino acid) in the E. coli chromosome.It transduces, therefore, only gal or bio genes thus making the recipient bacterial chromosome diploid for either gal or bio genes. Similarly, phage ϕ80 integrates near the trp genes (required for the synthesis of tryptophan amino acid) and transduces them.

Probable Questions:

1. Describe genome organization in prokaryotes.
2. Describe the variations found in prokaryotic genome structure.
3. Describe the life cycle of MS2 bacteriophages.
4. Describe lytic cycle of a phage with examples.
5. Describe lysogenic cycle of a bacteriophage.
6. Describe significance of lysogeny.
7. Define conjugation. Describe the process in a bacteria.
8. What is transduction? What is the difference between simple and specialized transduction?
9. Define transformation in bacteria. Describe the process in brief with suitable diagram.
10. What are the significances of transformation in bacteria?

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

Unit-VII

Protein folding and processing: Chaperones and folding; enzymes and protein folding, protein cleavage, glycosylation, attachment of lipids.

Objective: In this unit we will discuss about different aspects of protein folding.

Introduction:

Translation completes the flow of genetic information within the cell. The sequence of nucleotides in DNA has now been converted to the sequence of amino acids in a polypeptide chain. The synthesis of a polypeptide, however, is not equivalent to the production of a functional protein. To be useful, polypeptides must fold into distinct three-dimensional conformations, and in many cases multiple polypeptide chains must assemble into a functional complex. In addition, many proteins undergo further modifications, including cleavage and the covalent attachment of carbohydrates and lipids, that are critical for the function and correct localization of proteins within the cell.

Chaperones and Protein Folding

The three-dimensional conformations of proteins result from interactions between the side chains of their constituent amino acids. The classic principle of protein folding is that all the information required for a protein to adopt the correct three-dimensional conformation is provided by its amino acid sequence. This was initially established by Christian Anfinsen's experiments demonstrating that denatured RNase can spontaneously refold in vitro to its active conformation. Protein folding thus appeared to be a self-assembly process that did not require additional cellular factors. More recent studies, however, have shown that this is not an adequate description of protein folding within the cell. The proper folding of proteins within cells is mediated by the activities of other proteins.

Proteins that facilitate the folding of other proteins are called molecular chaperones. The term "chaperone" was first used by Ron Laskey and his colleagues to describe a protein (nucleoplasmin) that is required for the assembly of nucleosomes from histones and DNA. Nucleoplasmin binds to histones and mediates their assembly into nucleosomes, but nucleoplasmin itself is not incorporated into the final nucleosome structure. Chaperones thus act as catalysts that facilitate assembly without being part of the assembled complex. Subsequent studies have extended the concept to include proteins that mediate a variety of

other assembly processes, particularly protein folding.

It is important to note that chaperones do not convey additional information required for the folding of polypeptides into their correct three-dimensional conformations; the folded conformation of a protein is determined solely by its amino acid sequence. Rather, chaperones catalyze protein folding by assisting the self-assembly process. They appear to function by binding to and stabilizing unfolded or partially folded polypeptides that are intermediates along the pathway leading to the final correctly folded state. In the absence of chaperones, unfolded or partially folded polypeptide chains would be unstable within the cell, frequently folding incorrectly or aggregating into insoluble complexes. The binding of chaperones stabilizes these unfolded polypeptides, thereby preventing incorrect folding or aggregation and allowing the polypeptide chain to fold into its correct conformation.

A good example is provided by chaperones that bind to nascent polypeptide chains that are still being translated on ribosomes, thereby preventing incorrect folding or aggregation of the amino- terminal portion of the polypeptide before synthesis of the chain is finished (Figure 1). Presumably, this interaction is particularly important for proteins in which the carboxy terminus (the last to be synthesized) is required for correct folding of the amino terminus. In such cases, chaperone binding stabilizes the amino-terminal portion in an unfolded conformation until the rest of the polypeptide chain is synthesized and the completed protein can fold correctly. Chaperones also stabilize unfolded polypeptide chains during their transport into subcellular organelles—for example, during the transfer of proteins into mitochondria from the cytosol (Figure 2). Proteins are transported across the mitochondrial membrane in partially unfolded conformations that are stabilized by chaperones in the cytosol. Chaperones within the mitochondrion then facilitate transfer of the polypeptide chain across the membrane and its subsequent folding within the organelle. In addition, chaperones are involved in the assembly of proteins that consist of multiple polypeptide chains, in the assembly of macromolecular structures (e.g., nucleoplasmin), and in the regulation of protein degradation.

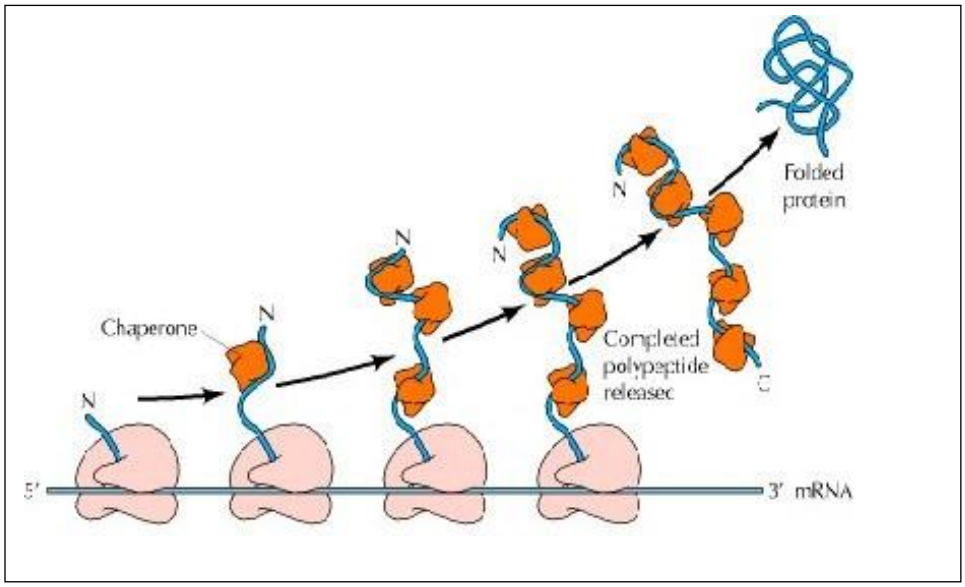


Figure1. Action of chaperones during translation. Chaperones bind to the amino (N) terminus of the growing polypeptide chain, stabilizing it in an unfolded configuration until synthesis of the polypeptide is completed. The completed protein is then released from the ribosome and is able to fold into its correct three-dimensional conformation.

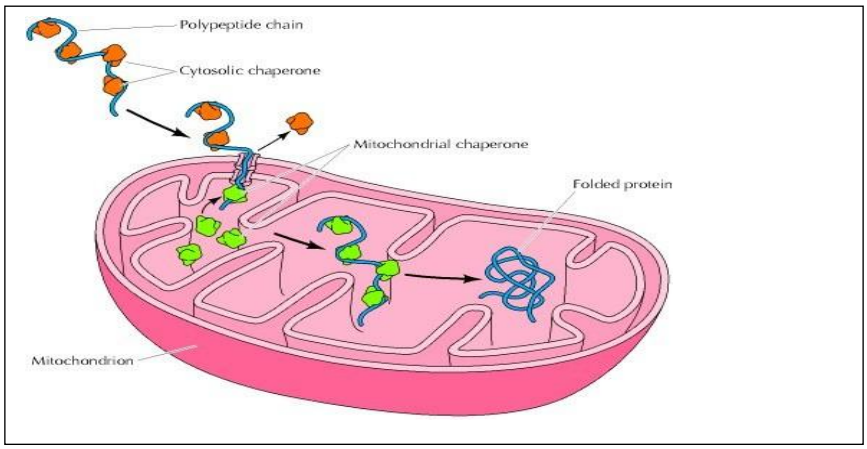


Figure 2.Action of chaperones during protein transport. A partially unfolded polypeptide is transported from the cytosol to a mitochondrion. Cytosolic chaperones stabilize the unfolded configuration. Mitochondrial chaperones facilitate transport and subsequent folding of the polypeptide chain within the organelle.

Many of the proteins now known to function as molecular chaperones were initially identified as heat-shock proteins, a group of proteins expressed in cells that have been subjected to elevated temperatures or other forms of environmental stress. The heat-shock proteins (abbreviated Hsp) are highly conserved in both prokaryotic and eukaryotic and are thought to stabilize and facilitate the refolding of proteins that have been partially denatured as a result of exposure to elevated temperature. However, many members of the heat- shock protein family are expressed and have essential cellular functions under normal growth conditions. These proteins serve as molecular chaperones, which are needed for polypeptide folding and transport under normal conditions as well as in cells subjected to environmental stress.

The Hsp70 and Hsp60 families of heat-shock proteins appear to be particularly important in the general pathways of protein folding in both prokaryotic and eukaryotic cells. The proteins of both families function by binding to unfolded regions of polypeptide chains. Members of the Hsp70 family stabilize unfolded polypeptide chains during translation as well as during the transport of polypeptides into a variety of subcellular compartments, such as mitochondria and the endoplasmic reticulum. These proteins bind to short segments (seven or eight amino acid residues) of unfolded polypeptides, maintaining the polypeptide chain in an unfolded configuration and preventing aggregation.

Members of the Hsp60 family (also called chaperonins) facilitate the folding of proteins into their native conformations. Each chaperonin consists of 14 subunits of approximately 60 kilodaltons (kd) each, arranged in two stacked rings to form a “double doughnut” structure (Figure-3). Unfolded polypeptide chains are shielded from the cytosol by being bound within the central cavity of the chaperonin cylinder. In this isolated environment protein folding can proceed while aggregation of unfolded segments of the polypeptide chain is prevented by their binding to the chaperonin. The binding of unfolded polypeptides to the chaperonin is a reversible reaction that is coupled to the hydrolysis of ATP as a source of energy. ATP hydrolysis thus drives multiple rounds of release and rebinding of unfolded regions of the polypeptide chain to the chaperonin, allowing the polypeptide to fold gradually into the correct conformation.

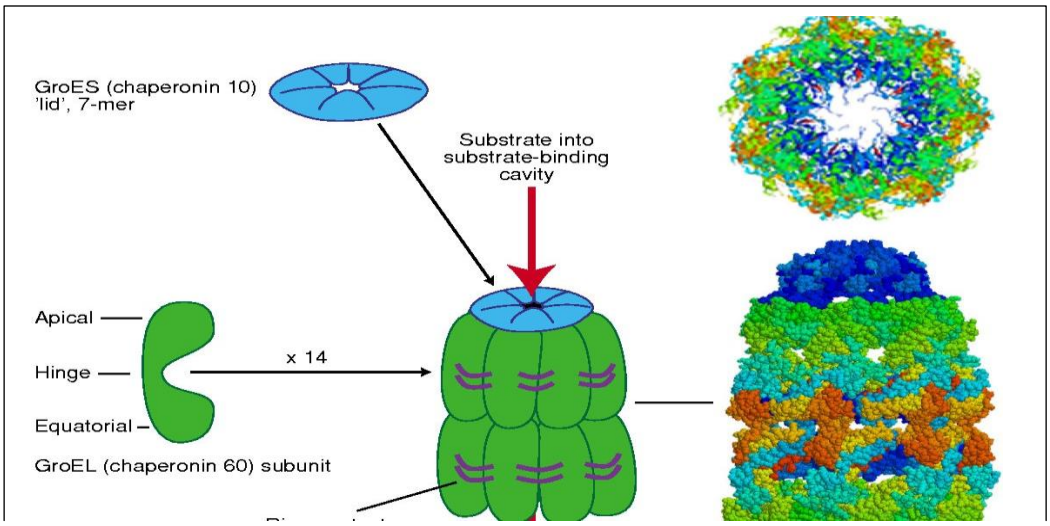


Figure 3.Structure of a chaperonin. Structure of a chaperonin. GroEL, a member of the Hsp60 family, is a porous cylinder composed of two stacked rings. Each ring consists of seven subunits. (Courtesy of Paul B. Sigler, Yale University.)

In some cases, members of the Hsp70 and Hsp60 families have been found to act together in a sequential fashion. For example, Hsp70 and Hsp60 family members act sequentially during the transport of proteins into mitochondria and during the folding of newly synthesized proteins in E. coli (Figure 4). First, an Hsp70 chaperone stabilizes nascent polypeptide chains until protein synthesis is completed. The unfolded polypeptide chain is then transferred to an Hsp60 chaperonin, within which protein folding takes place, yielding a protein correctly folded into its functional three-dimensional conformation. Members of the Hsp70 and Hsp60 families are found in the cytosol and in subcellular organelles (e.g., mitochondria) of eukaryotic cells, as well as in bacteria, so the sequential action of Hsp70 and Hsp60 appears to represent a general pathway of protein folding. An alternative pathway for the folding of some proteins in the cytosol and endoplasmic reticulum may involve the sequential actions of Hsp70 and Hsp90 family members, although the function of Hsp90 is not yet well understood.

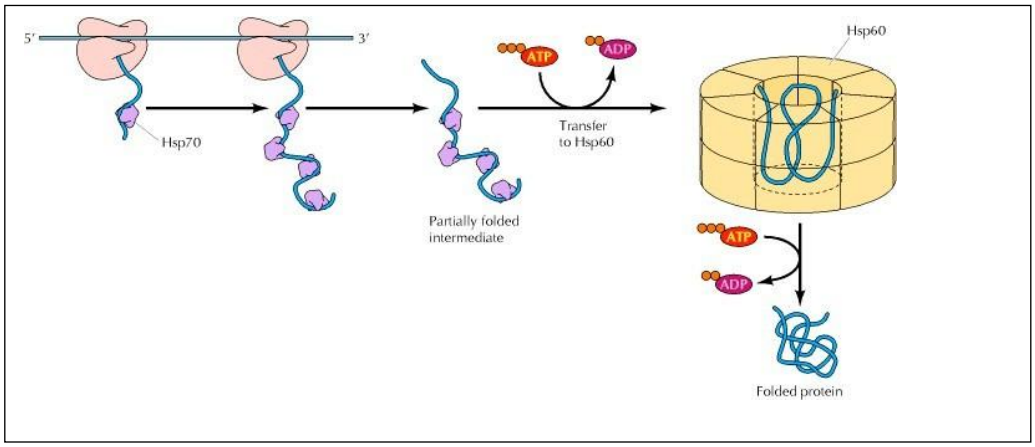


Figure 4.Sequential actions of Hsp70 and Hsp60 chaperones. Chaperones of the Hsp70 family bind to and stabilize unfolded polypeptide chains during translation. The unfolded polypeptide is then transferred to chaperones of the Hsp60 family, within which protein folding takes place. ATP hydrolysis is required for release of the unfolded polypeptide from Hsp70 as well as for folding within Hsp60.

Enzymes and Protein Folding

In addition to chaperones, which facilitate protein folding by binding to and stabilizing partially folded intermediates, cells contain at least two types of enzymes that catalyze protein folding by breaking and re-forming covalent bonds. The formation of disulfide bonds between cysteine residues is important in stabilizing the folded structures of many proteins. Protein disulfide isomerase, which was discovered by Christian Anfinsen in 1963, catalyzes the breakage and re-formation of these bonds (Figure 5). For proteins that contain multiple cysteine residues, protein disulfide isomerase (PDI) plays an important role by promoting rapid exchanges between paired disulfides, thereby allowing the protein to attain the pattern of disulfide bonds that is compatible with its stably folded conformation. Disulfide bonds are generally restricted to secreted proteins and some membrane proteins because the cytosol contains reducing agents that maintain cysteine residues in their reduced (—SH form), thereby preventing the formation of disulfide (S—S) linkages. In eukaryotic cells, disulfide bonds form in the endoplasmic reticulum, in which an oxidizing environment is maintained. Consistent with the role of disulfide bonds in stabilizing secreted proteins, the activity of PDI in the endoplasmic reticulum is correlated with the level of protein secretion in different types of cells.

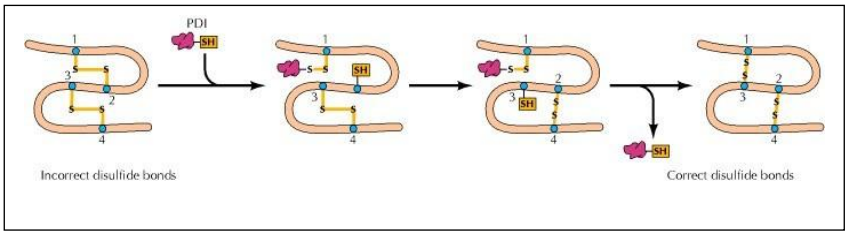


Figure 5.The action of protein disulfide isomerase. Protein disulfide isomerase (PDI) catalyzes the breakage and rejoining of disulfide bonds, resulting in exchanges between paired disulfides in a polypeptide chain. The enzyme forms a disulfide bond with a cysteine residue of the polypeptide and then exchanges its paired disulfide with another cysteine residue. In this example, PDI catalyzes the conversion of two incorrect disulfide bonds (1-2 and 3-4) to the correct pairing (1-3 and2-4).

The second enzyme that plays a role in protein folding catalyzes the isomerization of peptide bonds that involve proline residues (Figure 6). Proline is an unusual amino acid in that the equilibrium between the cis and trans conformations of peptide bonds that precede proline residues is only slightly in favor of the trans form. In contrast, peptide bonds between other amino acids are almost always in the trans form. Isomerization between the cis and trans configurations of prolyl peptide bonds, which could otherwise represent a rate-limiting step in protein folding, is catalyzed by the enzyme peptidylprolyl isomerase. This enzyme is widely distributed in both prokaryotic and eukaryotic cells and can catalyze the refolding of at least some proteins. However, its physiologically important substrates and role within cells have not yet been determined.

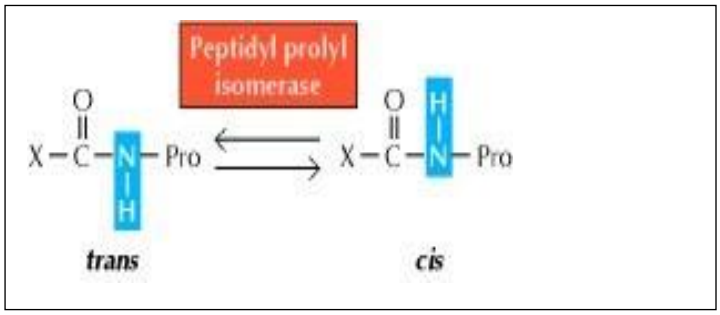


Figure-6.The action of peptidylprolyl isomerase. Peptidylprolyl isomerase catalyzes the Isomerization of peptide bonds that involve proline between the cis and transconformations.

Protein Cleavage

Cleavage of the polypeptide chain (proteolysis) is an important step in the maturation of many proteins. A simple example is removal of the initiator methionine from the amino terminus of many polypeptides, which occurs soon after the amino terminus of the growing polypeptide chain emerges from the ribosome. Additional chemical groups, such as acetyl groups or fatty acid chains (discussed shortly), are then frequently added to the amino-terminal residues.

Proteolytic modifications of the amino terminus also play a part in the translocation of many proteins across membranes, including secreted proteins in both bacteria and eukaryotes as well as proteins destined for incorporation into the plasma membrane, lysosomes, mitochondria, and chloroplasts of eukaryotic cells. These proteins are targeted for transport to their destinations by amino-terminal sequences that are removed by proteolytic cleavage as the protein crosses the membrane. For example, amino-terminal signal sequences, usually about 20 amino acids long,

target secreted proteins to the plasma membrane of bacteria or to the endoplasmic reticulum of eukaryotic cells while translation is still in progress (Figure 7). The signal sequence, which consists predominantly of hydrophobic amino acids, is inserted into the membrane as it emerges from the ribosome. The remainder of the polypeptide chain passes through a channel in the membrane as translation proceeds. The signal sequence is then cleaved by a specific membrane protease (signal peptidase), and the mature protein is released. In eukaryotic cells, the translocation of growing polypeptide chains into the endoplasmic reticulum is the first step in targeting proteins for secretion, incorporation into the plasma membrane, or incorporation into lysosomes.

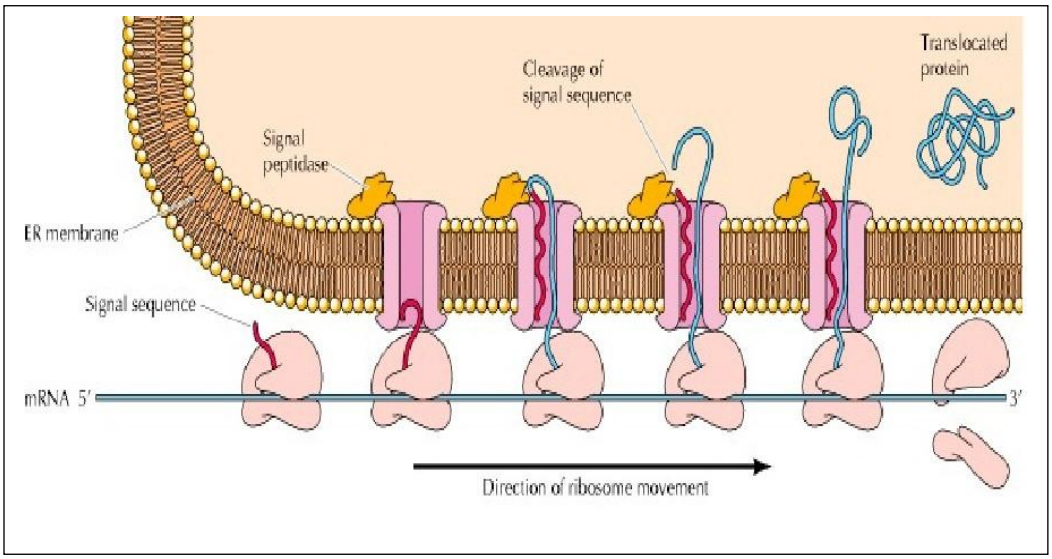


Figure 7. The role of signal sequences in membrane translocation. Signal sequences target the translocation of polypeptide chains across the plasma membrane of bacteria or into the endoplasmic reticulum of eukaryotic cells (shown here). The signal sequence, a stretch of hydrophobic amino acids at the amino terminus of the polypeptide chain, inserts into a membrane channel as it emerges from the ribosome. The rest of the polypeptide is then translocated through the channel and the signal sequence is cleaved by the action of signal peptidase, releasing the mature translocated protein.

In other important instances of proteolytic processing, active enzymes or hormones form via cleavage of larger precursors. Insulin, which is synthesized as a longer precursor polypeptide, is a good example. Insulin forms by two cleavages. The initial precursor (preproinsulin) contains an amino-terminal signal sequence that targets the polypeptide chain to the endoplasmic reticulum . Removal of the signal sequence during transfer to the endoplasmic reticulum yields a second precursor, called proinsulin. This precursor is then converted to insulin, which consists of two chains held together by disulfide bonds, by proteolytic removal of an internal peptide. Other proteins activated by similar cleavage processes include digestive enzymes and the proteins involved in blood clotting.

It is interesting to note that the proteins of many animal viruses are derived from the cleavage of larger precursors. One particularly important example of the role of proteolysis in virus replication is provided by HIV. In the replication of HIV, a virus-encoded protease cleaves precursor polypeptides to form the viral structural proteins. Because of its central role in virus replication, the HIV protease (in addition to reverse transcriptase) is an important target for the development of drugs used for treating AIDS. Indeed, such protease inhibitors are now among the most effective agents available for combating this disease

Glycosylation

Many proteins, particularly in eukaryotic cells, are modified by the addition of carbohydrates, a process called glycosylation. The proteins to which carbohydrate chains have been added (called glycoproteins) are usually secreted or localized to the cell surface, although some nuclear and cytosolic proteins are also glycosylated. The carbohydrate moieties of glycoproteins play important roles in protein folding in the endoplasmic reticulum, in the targeting of proteins for delivery to the appropriate intracellular compartments, and as recognition sites in cell-cell interactions.

Glycoproteins are classified as either N-linked or O-linked, depending on the site of attachment of the carbohydrate side chain. In N-linked glycoproteins, the carbohydrate is attached to the nitrogen atom in the side chain of asparagine. In O-linked glycoproteins, the oxygen atom in the side chain of serine or threonine is the site of carbohydrate attachment.

The sugars directly attached to these positions are usually either N-acetylglucosamine or N-acetylgalactosamine, respectively.

Most glycoproteins in eukaryotic cells are destined either for secretion or for incorporation into the plasma membrane. These proteins are usually transferred into the endoplasmic reticulum (with the cleavage of a signal sequence) while their translation is still in progress. Glycosylation is also initiated in the endoplasmic reticulum before translation is complete. The first step is the transfer of a common oligosaccharide consisting of 14 sugar residues (2 N-acetylglucosamine, 3 glucose, and 9 mannose) to an asparagine residue of the growing polypeptide chain (Figure 9). The oligosaccharide is assembled within the endoplasmic reticulum on a lipid carrier (dolichol phosphate). It is then transferred as an intact unit to an acceptor asparagine (Asn) residue within the sequence Asn-X-Ser or Asn-X-Thr (where X is any amino acid other than proline).

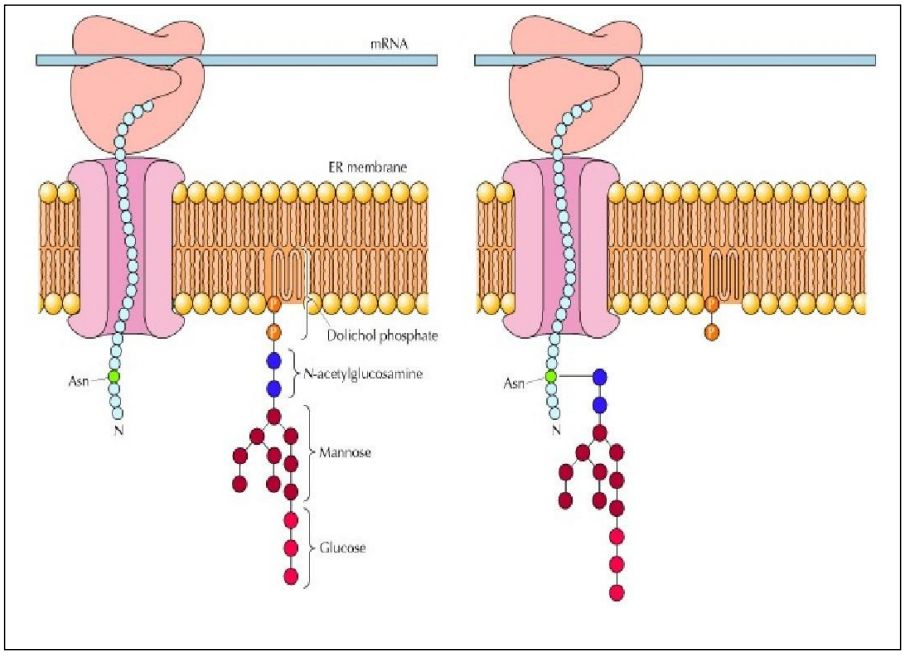


Figure 9.Synthesis of N-linked glycoproteins. The first step in glycosylation is the addition of an oligosaccharide consisting of 14 sugar residues to a growing polypeptide chain in the endoplasmic reticulum (ER). The oligosaccharide (which consists of two N-acetylglucosamine, nine mannose, and three glucose residues) is assembled on a lipid carrier (dolichol phosphate) in the ER membrane. It is then transferred as a unit to asparagine residue of the polypeptide.

In further processing, the common N-linked oligosaccharide is modified. Three glucose residues and one mannose are removed while the glycoprotein is in the endoplasmic reticulum. The oligosaccharide is then further modified in the Golgi apparatus, to which

glycoproteins are transferred from the endoplasmic reticulum. These modifications include both the removal and addition of carbohydrate residues as the glycoprotein is transported through the compartments of the Golgi. The N- linked oligosaccharides of different glycoproteins are processed to different extents, depending on both the enzymes present in different cells and on the accessibility of the oligosaccharide to the enzymes that catalyse its modification. Glycoproteins with inaccessible oligosaccharides do not have new sugars added to them in the Golgi. The relatively simple oligosaccharides of these glycoproteins are called high- mannose oligosaccharides because they contain a high proportion of mannose residues, similar to the common oligosaccharide originally added in the endoplasmic reticulum. In contrast, glycoproteins with accessible oligosaccharides are processed more extensively, resulting in the formation of a variety of complex oligosaccharides.

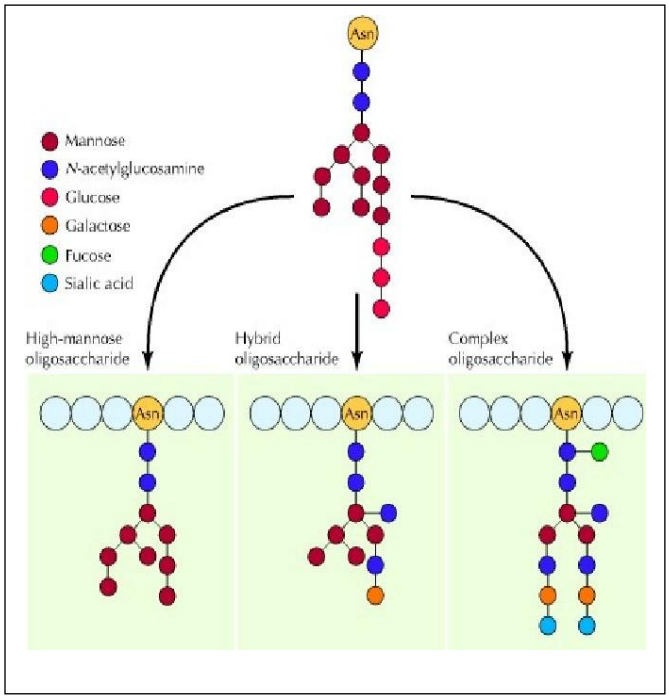


Figure 10.Examples of N-linked oligosaccharides. Various oligosaccharides form from further modifications of the common 14-sugar unit initially added in the endoplasmic reticulum (see Figure 7.26). In high-mannose oligosaccharides, the glucose residues and some mannose residues are removed, but no other sugars are added. In the synthesis of complex oligosaccharides, more mannose residues are removed and other sugars are added. Hybrid oligosaccharides are intermediate between high-mannose and complex oligosaccharides. The structures shown are representative examples.

O-linked oligosaccharides are also added within the Golgi apparatus. In contrast to the N-linked oligosaccharides, O-linked oligosaccharides are formed by the addition of one sugar at

a time and usually consist of only a few residues (Figure 11). Many cytoplasmic and nuclear proteins, including a variety of transcription factors, are also modified by the addition of single O-linked N-acetylglucosamine residues, catalyzed by a different enzyme system. However, the roles of carbohydrates in the function of these cytoplasmic and nuclear glycoproteins are not yet understood.

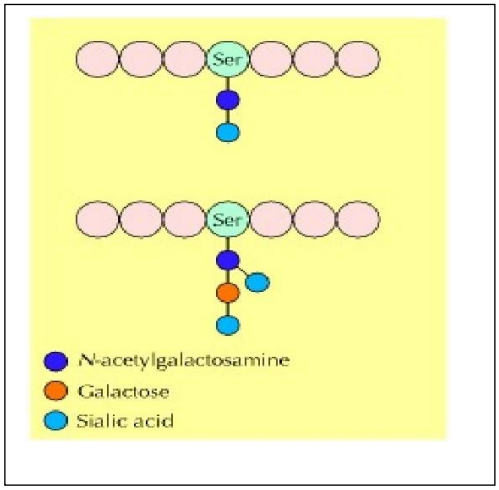


Figure 11.Examples of O-linked oligosaccharides

Attachment of Lipids

Some proteins in eukaryotic cells are modified by the attachment of lipids to the polypeptide chain. Such modifications frequently target and anchor these proteins to the plasma membrane, with which the hydrophobic lipid is able to interact. Three general types of lipid additions—N- myristoylation, prenylation, and palmitoylation—are common in eukaryotic proteins associated with the cytosolic face of the plasma membrane. A fourth type of modification, the addition of glycolipids, plays an important role in anchoring some cell surface proteins to the extracellular face of the plasma membrane.

In some proteins, a fatty acid is attached to the amino terminus of the growing polypeptide chain during translation. In this process, called N-myristoylation, myristic acid (a 14-carbon fatty acid) is attached to an N-terminal glycine residue (Figure 12). The glycine is usually the second amino acid incorporated into the polypeptide chain; the initiator methionine is removed by proteolysis before fatty acid addition. Many proteins that are modified by N-myristoylation are associated with the inner face of the plasma membrane, and the role of the fatty acid in this association has been clearly demonstrated by analysis of mutant proteins in which the N-terminal glycine is changed to an alanine. This substitution prevents myristoylation and blocks the function of the mutant proteins by inhibiting their membrane association.

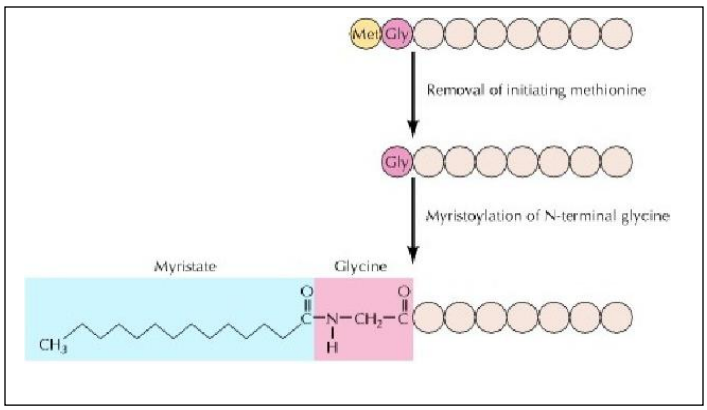


Figure 12.Addition of a fatty acid by N-myristoylation. The initiating methionine is removed, leaving glycine at the N terminus of the polypeptide chain. Myristic acid (a 14-carbon fatty acid) is then added.

Lipids can also be attached to the side chains of cysteine, serine, and threonine residues. One important example of this type of modification is prenylation, in which specific types of lipids (prenyl groups) are attached to the sulfur atoms in the side chains of cysteine residues located near the C terminus of the polypeptide chain (Figure 7.3130). Many plasma membrane– associated proteins involved in the control of cell growth and differentiation are modified in this way, including the Ras oncogene proteins, which are responsible for the uncontrolled growth of many human cancers. Prenylation of these proteins proceeds by three steps. First, the prenyl group is added to a cysteine located three amino acids from the carboxy terminus of the polypeptide chain. The prenyl groups added in this reaction are either farnesyl (15 carbons, as shown in Figure 7.30) or geranylgeranyl (20 carbons). The amino acids following the cysteine residue are then removed, leaving cysteine at the carboxy terminus. Finally, a methyl group is added to the carboxyl group of the C-terminal cysteine residue.

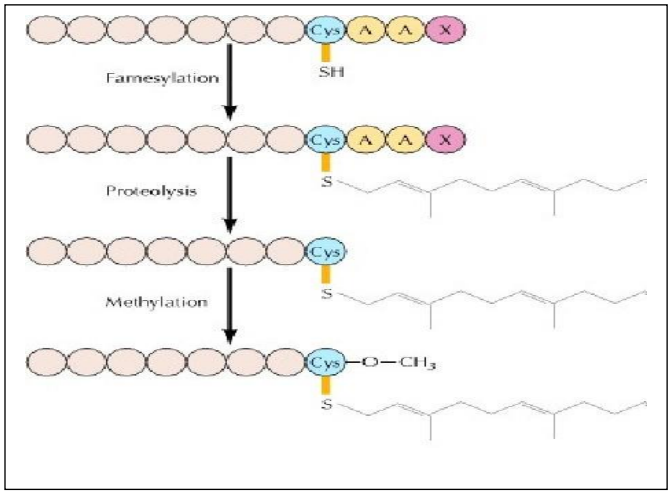


Figure 13.Prenylation of a C-terminal cysteine residue. The type of prenylation shown affects Ras proteins and proteins of the nuclear envelope (nuclear lamins). These proteins terminate with a cysteine residue (Cys) followed by two aliphatic amino acids (A) and any other amino acid (X) at the C terminus. The first step in their modification is addition of the 15-carbon farnesyl group to the side chain of cysteine (farnesylation). This step is followed by proteolytic removal of the three C-terminal amino acids and methylation of the cysteine, which is now at the C terminus

The biological significance of prenylation is indicated by the fact that mutations of the critical cysteine block the membrane association and function of Ras proteins. Because farnesylation is a relatively rare modification of cellular proteins, interest in this reaction has been stimulated by the possibility that inhibitors of the key enzyme (farnesyltransferase) might prove useful as drugs for the treatment of cancers that involve Ras proteins. Such inhibitors of farnesylation have been found to interfere with the growth of cancer cells in experimental models and are undergoing evaluation of their efficacy against human tumors in clinical trials.

In the third type of fatty acid modification, palmitoylation, palmitic acid (a 16-carbon fatty acid) is added to sulfur atoms of the side chains of internal cysteine residues. Like N-myristoylation and prenylation, palmitoylation plays an important role in the association of some proteins with the cytosolic face of the plasma membrane.

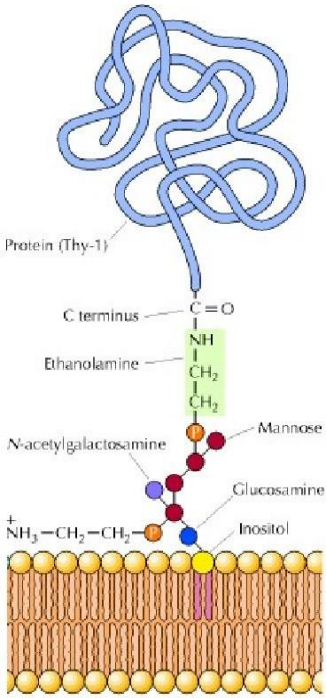


Figure 14.Structure of a GPI anchor.

Finally, lipids linked to oligosaccharides (glycolipids) are added to the C-terminal carboxyl groups of some proteins, where they serve as anchors that attach the proteins to the external face of the plasma membrane. Because the glycolipids attached to these proteins contain phosphatidylinositol, they are usually called glycosylphosphatidylinositol, or GPI, anchors (Figure 14). The oligosaccharide portions of GPI anchors are attached to the terminal carboxyl group of polypeptide chains. The inositol head group of phosphatidylinositol is in turn attached to the oligosaccharide, so the carbohydrate serves as a bridge between the protein and the fatty acid chains of the phospholipid. The GPI anchors are synthesized and added to proteins as a preassembled unit within the endoplasmic reticulum. Their addition is accompanied by cleavage of a peptide consisting of about 20 amino acids from the C terminus of the polypeptide chain. The modified protein is then transported to the cell surface, where the fatty acid chains of the GPI anchor mediate its attachment to the plasma membrane.

Probable questions:

1. How Chaperones helps in protein folding?
2. Describe the role of disulphideisomerase in protein folding.
3. What is protein cleavage how it occurs in the cell?
4. How proteins are glycosylated?
5. How lipids get attached to protein?

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

UNIT-VIII

Modes of cell communications; Signaling molecules and receptors; Signal transduction and amplification; Response to signals-Gene expression, Cellular growth and metabolism

Objective: In this unit we will discuss about cell signaling system. How cell respond to various stimulus and how cell signaling affects cellular growth and metabolism.

Introduction:

Eukaryotic cells and bacteria release a large number of signals and establish communication. The method of action is binding the signals with the protein receptors present on surface of large cell and triggering a series of intracellular reactions called intracellular signaling or signal transduction.

Besides, many signals (steroids and bacterial autoinducer) enter the cell, interact with signaling system and establish signal transduction. For establishing intracellular signaling, one must fully understand the operation of any cell from their origin to death.

The first signaling molecule (cyclic adenosine monophosphate or cAMP) was known during early 1960s. However, importance of intracellular signaling could be realized after the discovery of changes made by mutagenesis in signaling pathway which results in cellular transformation which is now called as cancer. One could understand their function by mutagenizing their cellularfunction.

It is now known that bacteria can change the eukaryotic cell signaling and invade the cells. The bacterial toxins can hijack the control of host cells. Similarly complexity of bacterial signaling is also known. Now an enormous amount of information is available on cell signaling and signal transduction pathway both in prokaryotes and eukaryotes.

Cells Communication:

All living cells communicate with each other through one or other form of signal. These signals may be environmental factors or may be produced by other cells.

Even in unicellular organisms like yeast chemical signals are released to stop proliferation and prepare for sexual reproduction. These are called peptide secreting factors (PSFs).

In multicellular organisms, cell-to-cell signalling may be divided into 4-categories:

1. Autocrine Signalling:

The signal molecules released by a cell act on the cell itself inducing it to respond.

2. Paracrine Signalling:

The signal molecule released by one cell acts locally on neighbouring cells.

3. Endocrine Signalling:

The signals are released by an endocrine gland in the form of hormones, which are carried to other cells by the blood and finally perceived by the target cells.

4. Synaptic Signalling:

In the organisms like mammals and other complex multicellular organisms short range signalling is not sufficient. For these organisms specialized cells have evolved, which are called neurons. These cells are involved in signalling between the two parts of bodies. Neurons receive, conduct and transmit signals at their junctions called synapse.

In all these instances, it is necessary for the target cell to have specific target molecules which recognise and find signal molecule. This binding initiates a series of events ultimately causing the cell to respond in the desired manner.

In a cell, the signal molecule may produce the following effects:

1. It may cause an immediate change in the metabolism of the cell.

2. It may bring about immediate change in the electrical potential across the cell membrane.

3. It may influence the process of gene expression.

Signal Transduction:
Transmission of impulse or transduction of signal to the target cell may be done directly or indirectly.

1. Direct Signal Transduction:

In this type of transduction, the signal molecules are directly transmitted to the nucleus, where they bind with specific segments of the genes and influence their transcription.

2. Indirect Signal Transduction:

In this process, the signal molecule (also called ligand in this case) does not enter the cell directly. Instead, it binds with a receptor molecule located at the cell-surface of the target cell. Cell-surface receptors are trans-membrane proteins that span across the plasma membrane.

Formation of ligand-receptor complex brings about a change in the conformation of receptor molecule. This change is transmitted in the cytoplasm where it generates a series of intracellular signals which are transmitted to genes in the nucleus. This pathway involving the operation of different proteins between cell membrane and nucleus is called signal transduction pathway.

The Signaling System:

Signaling system is very complex which may be compared to electronic circuits. You know that electronic system is such that can integrate, modulate and amplify inputs and generate output signals when switched on or switched off after getting suitable signals. The signaling systems include few basic type of modules. There are four main processes, but the signaling system uses the one or more processes.

The types of signaling modules used in intracellular signaling are:

(a) Receptor kinases (e.g. tyrosine kinase, serine kinase, histidinekinase),

(b) Receptor non-kinases (e.g. serpentine, cytokine, His-Aspphosphorelay),

(c) Protein kinase (intracellular enzymes e.g. cyclin families, Aspkinase),

(d) Lipid modifying intracellular enzymes (e.g. p13K, p15K,PLC),

(e) Cyclic nucleotides (e.g. cAMP, cGMP),and

(f) Metal ions (e.g.Ca^{++})

The four main processes are as follows:

(a) Protein phosphorylation bykinases,

(b) Small molecule and protein interaction, often involvingphosphates,

(c) Protein-protein interaction, often mediated by common motifs (specific protein sequences) which frequently results in membrane recruitment when one component is tethered to a membrane,and

(d) Protein and DNA interaction that promotes gene expression or geneinhibition.

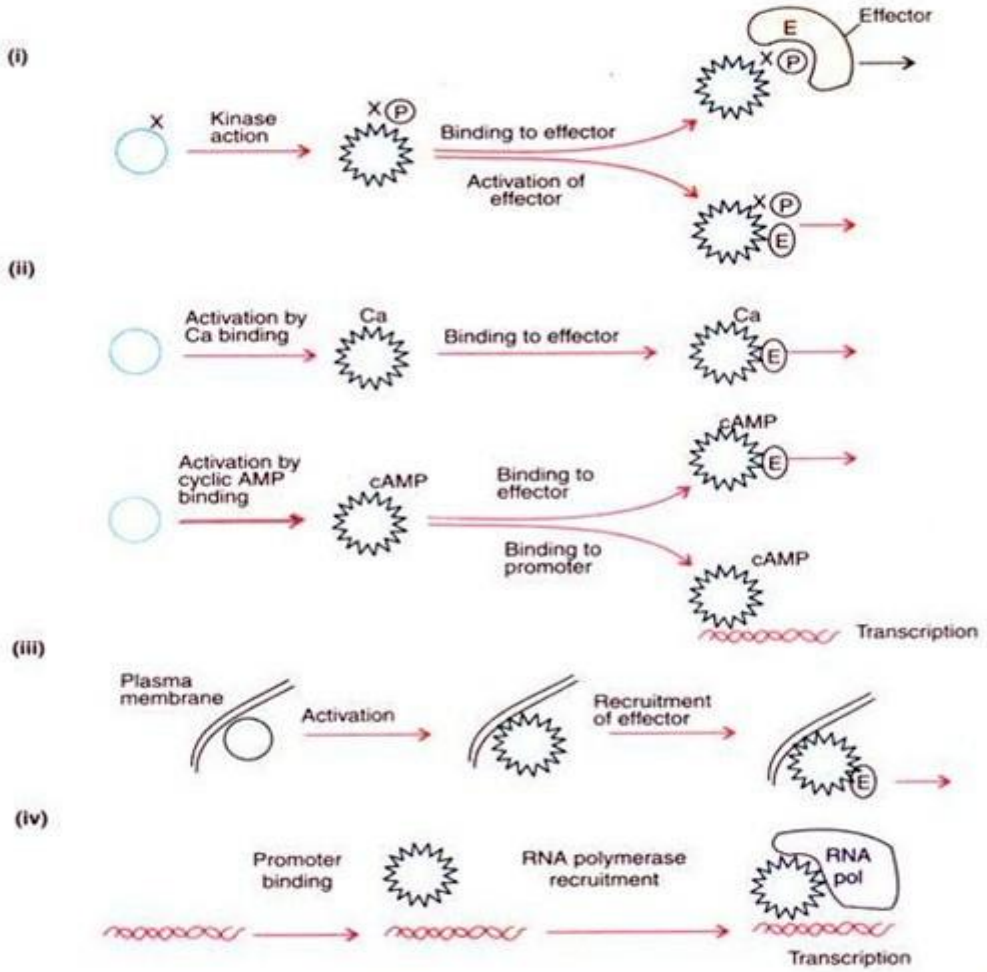


Fig. 27.8: Signalling molecules that use different types of interactions. *(i)* protein phosphorylation, where X=Tyr, Ser, Thr, His or Asp. *(ii)* interaction between small molecules and proteins, e.g. Ca^{++}, or cAMP. *(iii)* interaction between protein and protein, and *(iv)* interaction between protein and DNA which regulates transcription.

The Basic Building Blocks used in Signalling:

(a) Protein Phosphorylation: Protein phosphorylation is closely linked to cellular signaling. It exits in all signaling modules. The terminal y-phosphate is directly transferred from ARP (in some cases) to an acceptor protein by a protein kinase. The activity of the acceptor is modified example mitogen-activated protein (MAP) kinases in eukaryotes and histidyl-aspartyl phosphorelay in bacteria.

In some cases, indirect phosphorylation of protein also occurs (e.g. in G protein when binding of GTP activates their function, while GDP binding inactivates). There are secondary messengers which are used in intracellular signaling such as phosphorylated inositol's or cyclic nucleotides (cAMP, cGMP).

Kinases are regulated by any of a number of mechanisms: threonine and/or tyrosine phosphorylation, legend occupancy resulting in autophosphorylation or interaction with small molecules (e.g. cAMP or Ca^{++}).

i. HistidineKinases:

These are found in bacteria, lower eukaryotes and plants as trans membrane protein. They are stimulated to undergo self-phosphorylation by ligand occupancy.

ii. Protein Phosphatases:

Proteins which remove phosphate groups from proteins are called protein phosphatases. Protein kinases add phosphate group to proteins and play a key role in activation of signals.Specificphosphatases e.g.dephosphorylatephosphotyrosineand phosphoserine/phosphothreonme play a key role in control of proliferation, differentiation and cell cycle Phosphoproteins take part in signaling. They moderate the phosphorylation status by regulating the balance of phosphatases andkinases.

(b) Nucleotide-BindingProteins:

The three nucleotides (GTP, cGMP and cAMP) play a major role in the intracellular signaling.

iii. GTP-BindingProteins:

There is a set of eukaryotic proteins (G proteins) that show GTPase activity. They bind to GTP and remove the terminal phosphate of GTP and produce GDP bound to G protein. This cycle operates similar to ATP and ADP cycles.

When GDP dissociates from the G protein and GTP binds again, the cycle is repeated (Fig.

27. 9). G proteins are of two type: the heterotrimeric G proteins (the dominating proteins), and the small G proteins or membrane of Ras super family (the intermediate member of the signaling pathways).

The heterotrimeric G proteins consist of three different subunits- α, β and y subunits The a-subunit has GTP-binding domain; hence Gβ has a role in signal transduction.The βy subunits transmit signals by non-covalent interaction with effector molecules. Activation of G proteins and Association of α-subunit from βy subunits are given in Fig. 27.9. GTPase activity results in after binding the Gα subunit with GDP and subsequent association with Gβy and down regulation

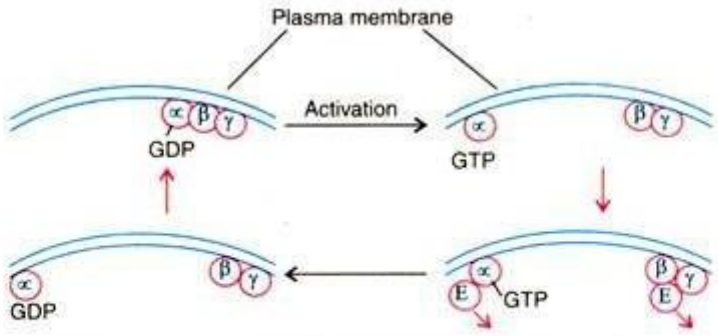


Fig. 27.9 : The function of membrane-bound heterotrimeric G proteins having α, β and γ subunits.

The small G proteins (Ras super family or p21 family) play a key role in many cellular functions such as proliferation and differentiation (Ras family), cytoskeletal organization (Rho) and nuclear membrane transport (Ran). The activity of small G proteins is modulated after interaction with several classes of proteins (Fig.27.10).

GDP-dissociation inhibitors (GDI) inhibit the loss of bound GDP and keep the G proteins in an inactive form to attenuate signaling from the activated G proteins. GTPase activity IS stimulated by GTPase-activating proteins (GAP). The removal of the bound GDP IS helped by guaninenucleotide exchange factors (GEF) which enable the GTP to bind and activate G proteins. Some of these factors have shown to be proto-oncogenes.

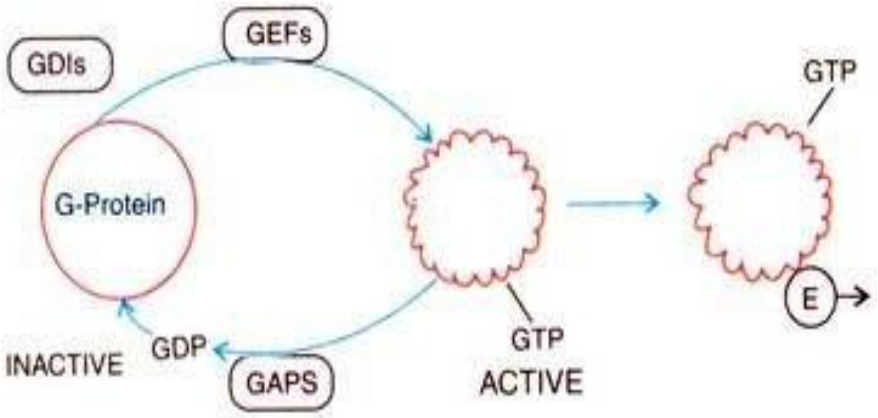


Fig. 27.10 : Functioning of small G proteins.

iv. Cyclin Nucleotide-BindingProteins:

In 1950s, cAMP was identified as the first intracellular signaling molecules. It mediates hormone action and acts as molecules transmitting the primary signal that has been received at the cell membrane). The cAMP mediates the response to chemo-attractants. The adenylate cyclase and guanylate cyclase regulate the concentration of cAMP and cGMP respectively.

The soluble bacterial adenylate cyclases produce cAMP which binds to c AMP receptor protein (CRP) and activate them. CRP is a transcription factor. The cAMP influences the expression of many of genes. Consequently bacteria become able to express metabolic enzymes which are required during growth. The cAMP also regulates the expression of the other genes which can cause pathogenesis.In eukaryotes heterotrimeric G proteins regulate the membrane-bound adenylate cyclases which produce cAMP. G proteins are coupled to transmembrane receptors. The cAMP-dependent protein kinases (protein kinase A) are the main effects of the cAMP signals.

While in the inactive form, protein kinase A consists of a dimer of regulatory (A) subunits and two catalytic (C) sub- units. The molecules of cAMP bind to reach R subunits and induceconformational changes. Consequently activated C subunits are released. This activated protein kinase A phosphorylates many substrates on serine or threonine (Fig. 27.11).

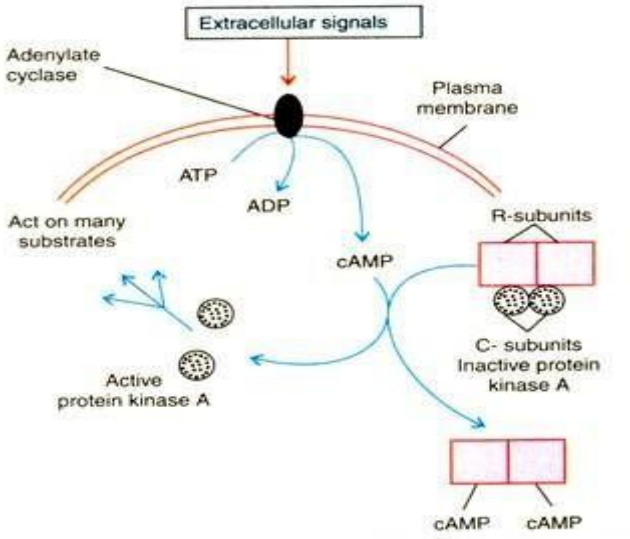


Fig. 27.11 : Production of cAMP and activation of protein kinase A.

Both the cycles work in eukaryotes by direct binding to proteins which form cation channels. Binding events result in opening of the channel.The G-protein-linked cell surface receptor generates small intracellular mediators thought cAMP pathways (Fig. 27.12).

(c) Role of Intracellular Concentration of Ca^{++} in CellSignaling:

Calcium is found in Cytoplasm and maintained in a very low concentration(10-100 nM). But its concentration varies with cell cycle, exogenous source or release from the stores. It gets complexes in membrane bound vesicles acting as stores. A highly specific protein calmodulin (CaM) binds to Ca^{++} and transmit the signal. Ca-binding to CaM brings about changes in conformation of CaM.

ConsequentlyCaM interacts with many effectors including CaM-modulated kinase. The most extensively studied CaM is the phosphatase calcineurin which is associated with several cellular activities such as NO synthesis, apoptosis, and induction of T lymphocytes. In eukaryotic cells Ca^{++} acts as a second messenger. Fig. 27.12 shows the two major pathways by which G-protein- linked cell surface receptors generate small intracellular mediators.

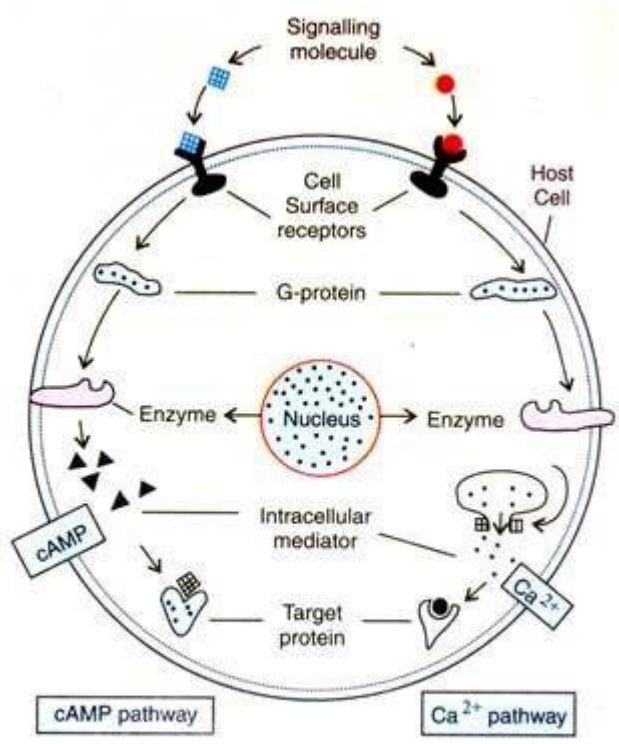


Fig. 27.12 : Generation of small intracellular mediators by G-protein-linked cell.

Role of Phosphorylated Lipids in Cell Signaling:

In eukaryotes lipids are involved in signaling process. Cellular phospholipases attack the lipid moieties of the membrane to produce different types of signaling molecules. For example, phosphatidylinositol lipids play a role in cellular stimulation. They have inositol as head, the six- membered carbon rings with a -OH group on each carbon.

On the basis of phosphorylation status of inositol head group, several phosphatidylinositols are found in the cells. The activity of three enzymes triggers their signaling role. These are:phosphoinositide 5′-kinase (P15p, phosphoinositide 3′-kinase (P13K), and phospholipase C (PLC). Extracellular signals regulate all these enzymes.

(d) Regulation ofTranscription:

Both types of cells are able to respond to any signal by changing their gene expression. In a signaling pathway the end point acts as signal. Regulators causing changes in expression of many genes in bacteria are called ‘global regulators’. In prokaryotes post-transcriptional

events regulate expression of many of the transcriptional factors for example cAMP-mediated CRP- DNA interactions.

In prokaryotes, phosphorylation or protein-protein interactions regulate the control of transcriptional factors and also select the other factors to the promotors. Besides, some other factors also get translocated from cytoplasm to the nucleus and regulate transcription.

(e) Role of Cell Membrane inSignaling:

Cell membrane acts as boundary of the cell through which extracellular signal has to enter. In bacteria histidine kinases acts as receptor and directs signals across the membrane. Besides, there are many signal molecules which are associated with cell membrane because the end effect is membrane-associated.

The components can be well organized in three-dimensional way in cell membrane. The signaling components recruit the other molecules to the membrane where they interact with other factors. For example, GTP-bound Ras activates Raf kinase to recruit Raf to the membrane where the membrane-bound kinase activates it throughphosphorylation.

3. Prokaryotic SignallingMechanisms:

Intracellular signaling is very complex like electronic circuit. Genome size of different bacteria varies and those organisms work according to genes present in them.

In bacteria the generic mechanism of regulation is called signaling systems which includes:

a. The histidyl-aspartae phosphorylation systems (the main module of bacteria used to receive and process incoming signals such as chemotaxis, response to osmolarity, oxygen and phosphate, and virulencesystem),

c. The cAMP and CRP (involved in regulation of hundreds of genes. The cAMP is controlled at transcriptional and post-transcriptional levels. Binding of CRP- cAMP complex induces gene expression).

4. Eukaryotic SignallingPathway:

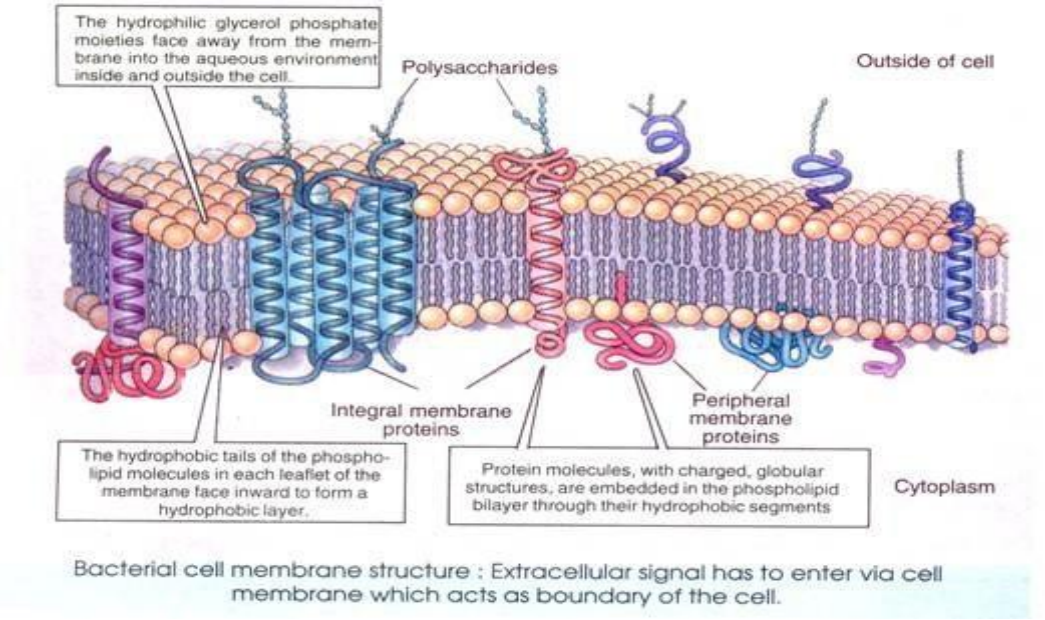
Earlier it was thought that signaling process in eukaryotes was very complex to understand in molecular terms. Fragmented understanding about individual components could be known.The knowledge of signaling expanded with the development of new techniques such

as genome sequencing, increasing number of reagents (isolated components, specific probes e.g. antibodies for individual components and selective inhibitors).

In spite of all these, no pathway has been fully elucidated. The best characterized pathway is the Ras activation and MAP kinases of which several details are unclarified. They are interconnected and cannot work without reference to others.

The Phospholipase C/Inositol Triphosphate Pathway:

The phospholipase C, beta or gamma is activated by membrane signaling events and cleaves PIP2 to produce diacylyglycerol (DAG) and inositol triphosphate (IP_3). These activates the release of Ca^{++} ions and results in activation of protein kinase C (PKC), which phosphorylates many additional protein substrates.



Bacterial cell membrane structure : Extracellular signal has to enter via cell membrane which acts as boundary of the cell.

The Adenylate Cyclase, cAMP and Protein Kinase A Pathway:

Adenylate cyclase is activated at the membrane by interaction with the activated heterotrimeric G protein G_5. The cAMP is generated and binds to and activates protein kinase A (PKA), which phosphorylates manysubstrates.

Integrin's, the Rho Family and Organization of Cytoskeletal:

The integrins are the signalling molecules that interact with the extracellular matrix on the outside of the cell and various proteins-linked to actin on the cells interior. The proteins involved include α-actin, lalin, tensin, vinculin andpavilin.

A local adhesion is formed upon activation that includes focal adhesion kinase (FAK). The Src kinase is recruited and several proteins in the complex are activated by phosphorylation by Src and FAC. These signals lead to the Ras/Raf, Rho signaling pathway and to cytoskeletal rearrangement.In eukaryotes, the central role of signaling pathway of a cell is to define its phenotype and function. The increasing novel knowledge about the components of signaling pathways and the types of genes which they interact are already being applied in new strategy to combat the cancer. For example, genetically engineered viruses are attempted to grow in such cells that lack functional p53 and kill thesecells.

There are about 2000-5000 signal transduction proteins in mammalian cells. Bacteria have capacity to utilize eukaryotic signaling pathway during the process of infection. These findings make a line between the signaling pathways involved in infection and the other responsible for the pathology in diseases such as cancer and inflammation.

Eicosanoids (Gr. ecosn = 20):

These are derived from arachidonic acid, a C-20 fatty acid with 4 double bonds, e.g., prostaglandins, thrombaxanes and leukotriene's. These are called local hormones because they are short lived and have autocrine and paracrine effect.

Cytokine receptors:

Cytokine receptors are receptors that bind cytokines. In recent years, the cytokine receptors have come to demand the attention of more investigators than cytokines themselves, partly because of their remarkable characteristics, and partly because a deficiency of cytokine receptors has now been directly linked to certain debilitating immunodeficiency states. In this regard, and also because the redundancy and pleiotropy of cytokines are a consequence of their homologous receptors, many authorities are now of the opinion that a classification of cytokine receptors would be more clinically and experimentally useful.

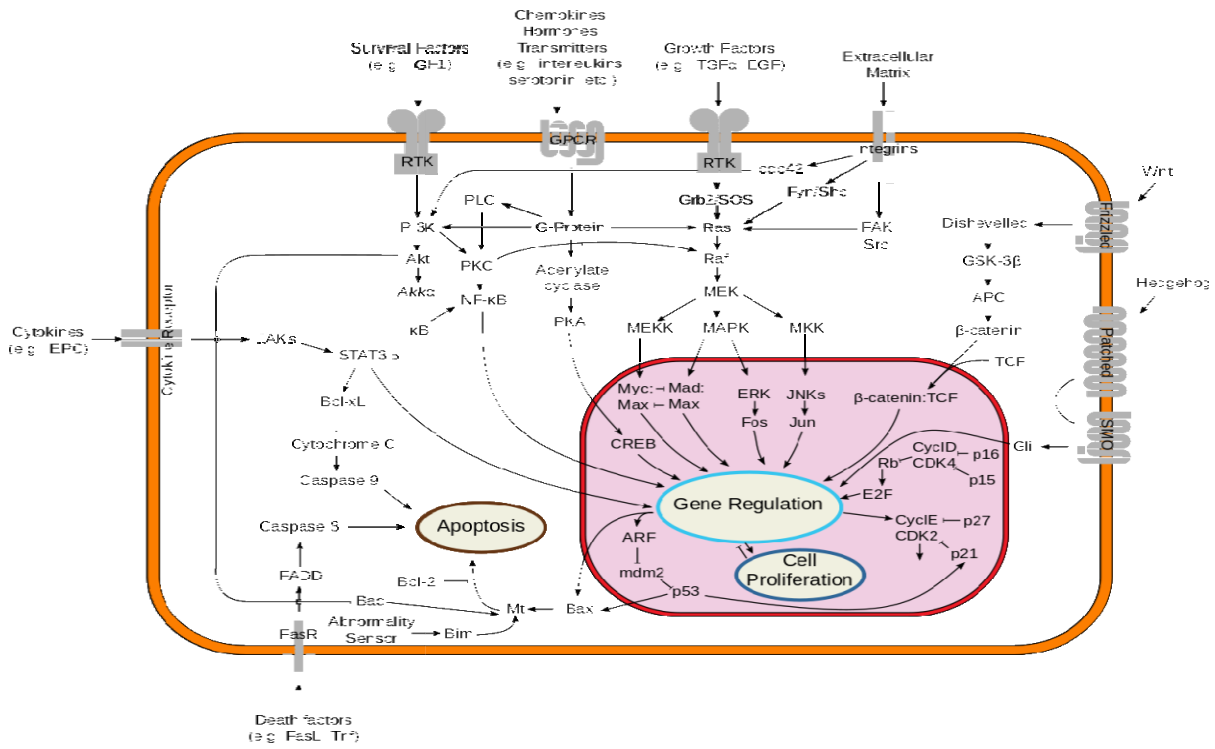


Fig : Signal transduction. (Cytokine receptor at center left.)

Classification of Cytokine Receptors

A classification of cytokine receptors based on their three- dimensional structure has been attempted. (Such a classification, though seemingly cumbersome, provides several unique perspectives for attractive pharmacotherapeutic targets.)

Type I cytokine receptorswhose members have certain conserved motifs in their extracellular amino-acid domain. The IL-2 receptor belongs to this chain, whose γ chain (commonto several other cytokines) deficiency is directly responsible for the X-linked form of Severe Combined Immunodeficiency (X-SCID).

Type II cytokine receptors, whose members are receptors mainly for interferons.

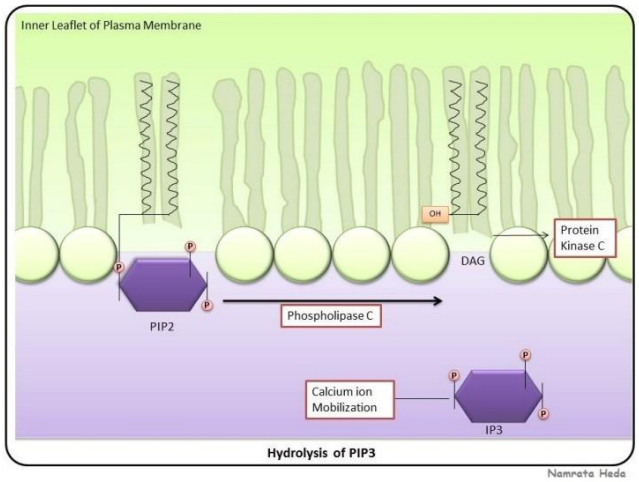
Immunoglobulin (Ig) superfamily, which are ubiquitously present throughout several cells and tissues of the vertebrate body

Tumor necrosis factor receptor family, whose members share a cysteine-rich common extracellular binding domain, and includes several other non-cytokine ligands like receptors, CD40, CD27 and CD30, besides the ligands on which the family is named (TNF).

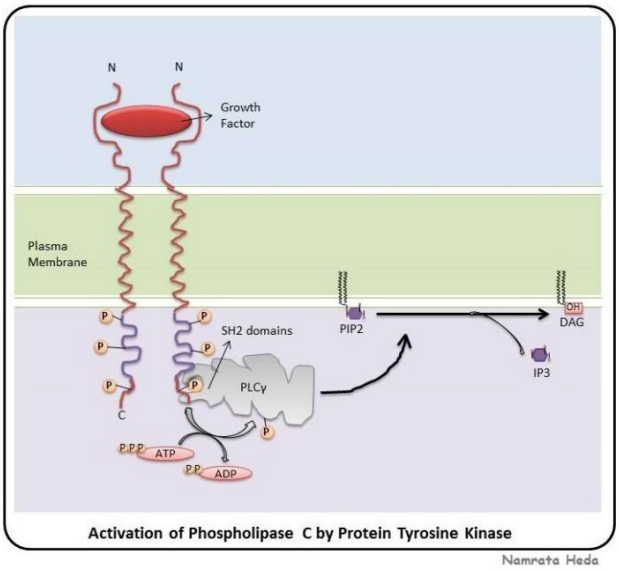
Chemokine receptors two of which acting as binding proteins for HIV (CXCR4 and CCR5). They are G protein coupled receptors.

Phospholipids and Ca ion mediated signaling :

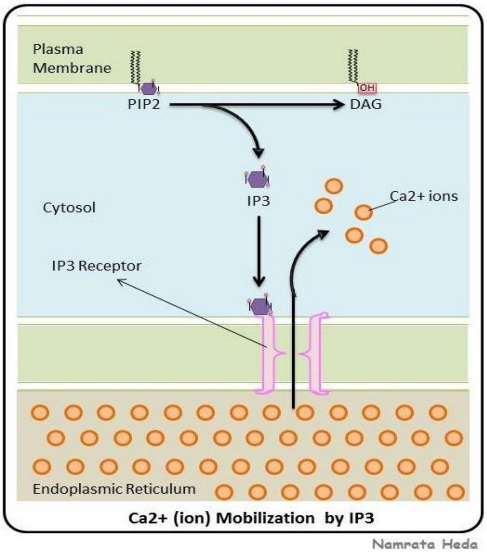
Phosphatidyl inositol 4,5-bisphosphate abbreviated as PIP2 is a phospholipid present in the inner leaflet of the bilayer of the plasma membrane. The second messengers are derived from this small component (phospholipid) and the pathway is based on these messengers.



The hydrolysis of PIP2 takes place by the enzyme phospholipase C as can be seen in the adjacent figure. It is interesting to note that the enzyme phospholipase C is ultimately activated by G- protein coupled receptors (GPCRs) or protein tyrosine kinases. This is so because one form of phospholipase C (PLC-β) is stimulated by G proteins while another form of phospholipase C (PLC-γ) contains SH2 domains (as can be seen in the figure shown below) and hence it associates with activated receptor protein tyrosine kinases. This interaction helps PLC-γtolocalize to plasma membrane and also leads to its phosphorylation. This tyrosine phosphorylation increases PLC-γ activity, which in turn stimulates hydrolysis ofPIP2.

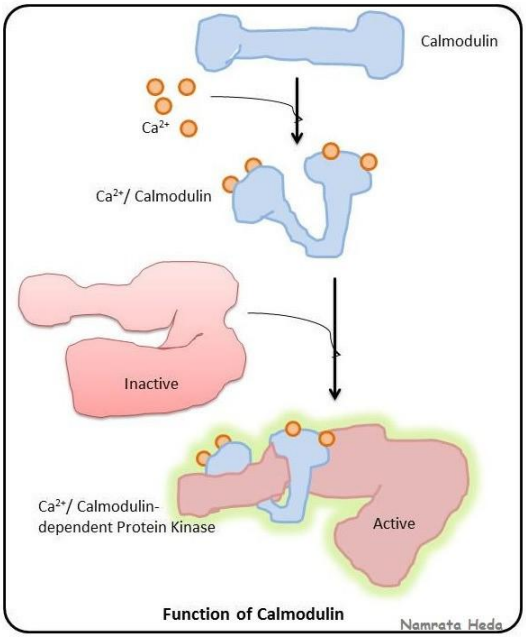


The hydrolysis of PIP2 produces two distinct second messengers as diacylglycerol and inositol 1,4,5-triphosphate which is abbreviated as IP3. Both these messengers stimulate different downstream signaling pathways thereby triggering two distinct cascades of intracellular signaling. Diacylglycerol stimulates protein kinase C mobilization while IP3 stimulates Ca2+(ions) mobilization. The diacylglycerol as second messenger activates serine/threonine kinases which belongs to the protein kinase C family which play an important role in cell growth anddifferentiation.IP3, another second messenger is released into the cytosol and it acts to release the Ca2+(ions) from intracellular stores. The level of the Ca2+(ions) inside the cell is very low and is maintained by pumping through Ca2+(ion) pumps across the plasma membrane.



The Ca2+(ions) are pumped into the ER and hence ER is considered to be the store of intracellular Ca2+(ions). Here, IP3 binds to the receptors in the ER membrane as can be seen in the adjacent diagram. These receptors are ligand-gated ion channels and hence, there is efflux of Ca2+(ions) into thecytosol.This increase of Ca2+(ions) in the cytosol has an effect on variety of proteins like protein kinases. For example, there are some members of protein kinase C (PKC) family that requires Ca2+(ions) as well as diacylglycerol for their functioning. Hence, these PKC family members are regulated by both IP3 and diacylglycerol.

Calmodulin is another very important protein to mention while we are studying about Ca2+(ions). The word 'calmodulin' means - cal(cium) + modul(ate) + in(g). Thus, calmodulin is 'calcium modulating' protein that mediates most of the activities of Ca2+(ions). Calmodulin is dumbbell shaped protein which has four Ca2+(ions) binding sites (figure is shown below). When the Ca2+(ions) concentration in the cell increases, calmodulin is activated. This active Ca2+/calmodulin complex then binds to a variety of target proteins, like Ca2+ion/calmodulin - dependent protein kinases thereby rendering them active. The examples of Ca2+ion/calmodulin dependent-protein kinases are: myosin light-chain kinase and members.



When there is a change in plasma membrane's potential i.e.; when there is membrane depolarization, the voltage-gated Ca2+ ion channels are opened in the plasma membrane. Because of the opening, there is influx of Ca2+(ions) from the extracellular fluid into the cytosol of the cell. This increase in the levels of Ca2+(ions) further triggers the opening of the another receptor called the ryanodine receptor in the plasma membrane which further releases the Ca2+(ions) from the intracellular stores. This increase in the Ca2+(ions) results in triggering the release of neurotransmitter. Hence, we can say that Ca2+ ion plays an important role in converting electric signals to chemical signals. In muscle cells, the ryanodine receptors on the sarcoplasmic reticulum. These receptors maybe opened directly when there is membrane depolarization.

Probable Questions:

1. Describe the role of protein phosphorylation in signal transduction.
2. Describe role of Intracellular Concentration of Ca^{++} in CellSignaling.
3. Describe the role of Phosphorylated Lipids in Cell Signaling.
4. Classify cytokine receptors in details.
5. Write down the role of calmodulin in signal transduction.

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7^{th} Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

UNIT-IX

Cell death; DNA damage and repair signaling ; Extra cellular matrix and cell signaling; Signaling crosstalk.

Objective: In this unit we will discuss about DNA damage and repair signaling and also about extra cellular matrix and cell signaling. Signaling crosstalk will also be discussed in this unit.

Introduction

From bacteria to mammalian cells, damaged DNA is sensed and targeted by DNA repair pathways. In eukaryotes, kinases play a central role in coordinating the DNA damage response. DNA damage signaling kinases were identified over two decades ago and linked to the cell cycle checkpoint concept proposed by Weinert and Hartwell in 1988. Connections between the DNA damage signaling kinases and DNA repair were scant at first, and the initial perception was that the importance of these kinases for genome integrity was largely an indirect effect of their roles in checkpoints, DNA replication, and transcription. As more substrates of DNA damage signaling kinases were identified, it became clear that they directly regulate a wide range of DNA repair factors.

From pathway to network

In eukaryotes, kinases play a central role in the DNA damage response, from sensing DNA damage to regulating cellular processes. Work in yeast and mammalian systems in the 1990s identified an evolutionarily conserved set of DNA damage signaling kinases, including phosphatidylinositol 3′ kinase (PI3K)-related kinases (PIKKs) and PIKK-regulated downstream kinases. These kinases were found to be involved in cell cycle control (Allen *et al*, 1994; Carr, 1995; Morrow *et al*, 1995; Savitsky *et al*, 1995; Cimprich *et al*, 1996; Furnari *et al*, 1997; Peng *et al*, 1997; Sanchez *et al*, 1997) and were linked to the "cell cycle checkpoint" concept proposed by Weinert and Hartwell (Weinert & Hartwell, 1988; Hartwell & Weinert, 1989). Subsequent work in the late 1990s and early 2000s revealed how these kinases establish the checkpoint and control processes beyond the cell cycle, such as apoptosis, transcription, and DNA replication (Santocanale&Diffley, 1998; Sun *et al*, 1998; Zhou &Elledge, 2000). In 2007, the use of mass spectrometry (MS)-based proteomics allowed a more systematic analysis of the network of phosphorylation events triggered by DNA damage signaling kinases (Matsuoka *et al*, 2007; Smolka *et al*, 2007). As a result, the perception that DNA damage signaling kinases operate within a simple signaling pathway (Fig 1; the classical "linear" depiction of DNA damage signaling) evolved to a more comprehensive view in which

DNA damage signaling kinases function in an elaborate signaling network comprised of hundreds of substrates.

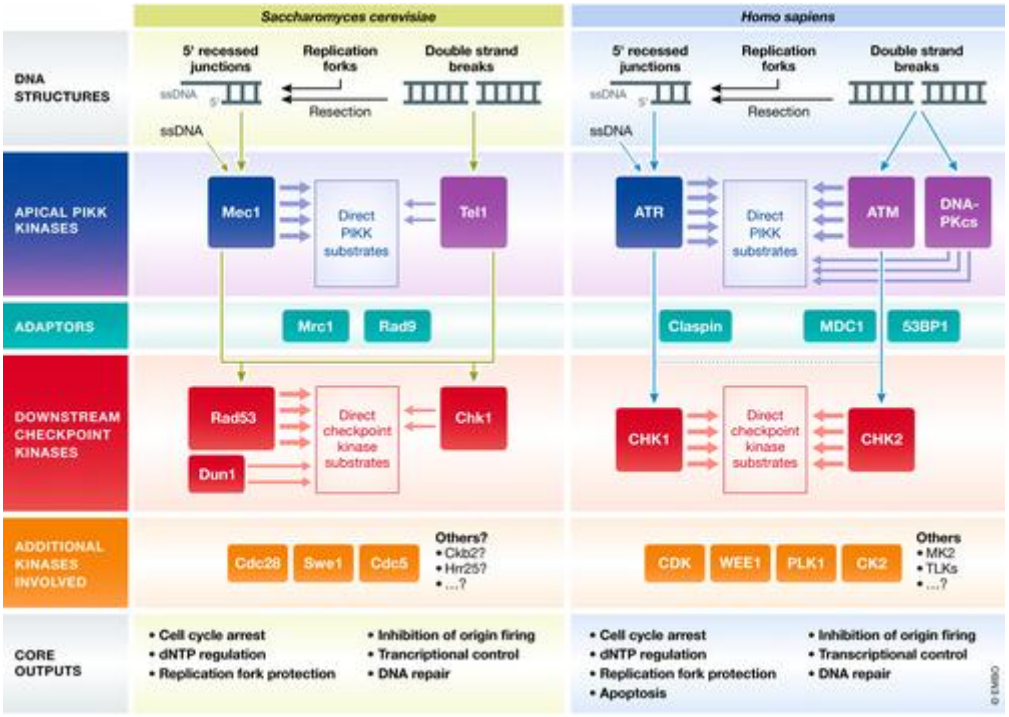


Figure 1. DNA damage signaling via PIKKs and checkpoint kinases in budding yeast and humans

DNA damage signaling is initiated at DNA structures that form during DNA damage or replication stress, including single-strand DNA (ssDNA) and broken DNA ends. The apical PIKKs are recruited to these structures and become activated to initiate downstream signaling. Mec1/ATR is recruited to RPA-coated ssDNA, while Tel1/ATM and DNA-PKcs initially associate with DNA ends formed by double-strand breaks. Adaptor proteins are often required to mediate the transfer of phosphorylation from apical to downstream checkpoint kinases. Apical and downstream checkpoint kinases function coordinately to mediate cellular responses to DNA damage, either directly or through the regulation of additional downstream kinases. PIKKs also target an extensive network of substrates independently of downstream checkpoint kinases.

From checkpoint to DNA repair

After the discovery of DNA damage signaling kinases, mechanistic links between these kinases and the DNA repair machinery were virtually non-existent. It was not immediately appreciated that these kinases directly target and regulate the DNA repair

machinery. Now that dozens of DNA repair proteins have been shown to be phosphorylated by DNA damage signaling kinases, there is little doubt that active and direct control of the DNA repair machinery is a core function of DNA damage signaling kinases. However, the precise mechanisms by which these kinases control the action of these substrates remain incompletely understood and represent a significant knowledge gap in the field. Here, we review our current understanding of the integrated action of DNA damage signaling kinases. We delineate the key substrates for checkpoint and DNA repair in budding yeast and highlight the potential parallels in humans. Based on the accumulated knowledge of the mechanisms of substrate regulation, we discuss how our understanding of the action of DNA damage signaling kinases in genome maintenance has evolved over the last 30 years.

DNA damage signaling kinases

DNA damage signaling kinases have been traditionally categorized as either apical or effector kinases. The apical PIKKs, or ATR, ATM, and DNA-PKcs in mammals and Mec1 and Tel1 in budding yeast (Fig 1; see Table 1 for gene name overview in model organisms), associate with DNA structures that form as byproducts of DNA damage or replication stress, including single-strand DNA (ssDNA) and broken DNA ends (Fig 1). In the canonical mode of action, Mec1/ATR is recruited to ssDNA, whereas Tel1/ATM and DNA-PKcs associate with the ends of double-strand DNA (dsDNA) breaks. The mechanism of activation for each of these kinases is different, but recruitment to damaged DNA is often a requirement (for details on the mechanism of how kinases associate with DNA structures, as well as co-factors involved, refer to the following reviews: Blackford& Jackson, 2017; Di Domenico *et al*, 2014; Saldivar *et al*, 2017; Zou, 2013). Activated apical kinases transfer stimulatory phosphorylation to the downstream checkpoint kinases (Rad53, Chk1, and Dun1 in yeast; CHK2 and CHK1 in mammals), which catalyze phosphorylation events that mediate cellular responses to DNA damage as part of the canonical DNA damage checkpoint (Fig 1). Whereas Rad53 mediates nearly all checkpoint-related functions in budding yeast, with Chk1 playing only a minor role (Sanchez *et al*, 1999), a more balanced division of labor exists for CHK1 and CHK2 in humans. Human CHK2 shares sequence and structural similarities with yeast Rad53, including an FHA domain (Matsuoka *et al*, 1998), but the functional similarities are limited to the DSB signaling response. Human CHK1 and yeast Rad53 play a crucial role in the replication stress response, but they share little or no sequence/structural similarity.

Adaptor proteins as substrates for downstream signaling activation

In budding yeast, the transfer of phosphorylation from apical to downstream checkpoint kinases requires the checkpoint adaptor proteins Rad9 and Mrc1, which recruit the downstream checkpoint kinases in proximity to the apical kinase. Mec1 and Tel1 promote the recruitment of the Rad9 adaptor by phosphorylating lesion-proximal

substrates, such as histone H2A (Downs *et al*, 2000) and the 9-1-1 complex (Paciotti *et al*, 1998), which are then recognized directly or indirectly by Rad9 (Fig 2; Toh *et al*, 2006; Hammet *et al*, 2007; Pfander&Diffley, 2011). Once recruited, Rad9 is phosphorylated by Mec1 or Tel1 (Emili, 1998; Vialard *et al*, 1998), which promotes its oligomerization (Soulier& Lowndes, 1999; Usui *et al*, 2009) and further stabilization on DNA (Naiki *et al*, 2004). Mec1- and Tel1-mediated phosphorylation of Rad9 creates docking sites for the recruitment of the downstream effector kinase Rad53 (Fig 2; Gilbert *et al*, 2001; Schwartz *et al*, 2002; Sun *et al*, 1998), which, upon recruitment to Rad9, is phosphorylated and activated by Mec1 or Tel1 (Sanchez *et al*, 1996; Sun *et al*, 1996). Chk1 also relies on Rad9 for its activation by Mec1 (Blankley & Lydall, 2004); however, unlike Rad53, the phosphorylation events that facilitate the recruitment of Chk1 to Rad9 are catalyzed by cyclin-dependent kinase (CDK; Fig 2; Abreu *et al*, 2013). Rad53 can also be activated via the Mrc1 adaptor (Alcasabas *et al*, 2001; Osborn &Elledge, 2003). Mrc1, being an intrinsic component of the replisome, is already "on-site" for mediating activation, obviating the need for a devoted recruitment mechanism, as is the case with Rad9. In fact, Mrc1 mediates a more rapid response to replication stress than Rad9-dependent DNA damage signaling (Pardo *et al*, 2017; Bacal *et al*, 2018). Similar to Rad9 and Mrc1, Sgs1 has been proposed to mediate Rad53 recruitment in a manner that depends upon its phosphorylation by Mec1 (Hegnauer *et al*, 2012). Mec1 and Tel1 also facilitate DNA damage signaling activation and propagation by recruiting chromatin modifiers and remodelers near sites of DNA damage (van Attikum *et al*, 2004; Downs *et al*, 2004; Morrison *et al*, 2004, 2007), which may help de-condense chromatin in a way that permits adaptor assembly.

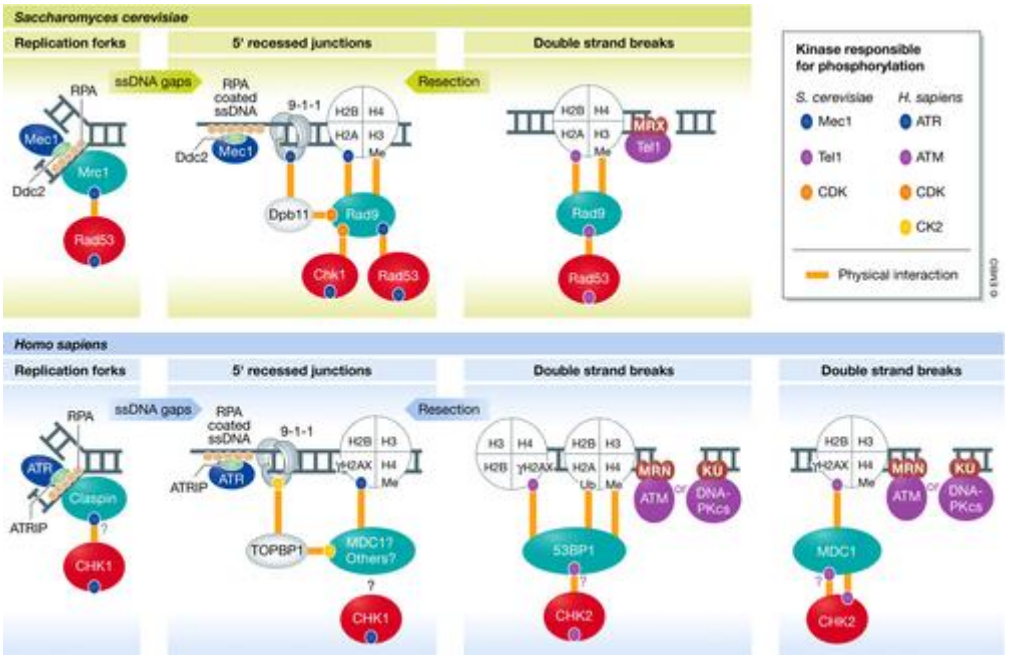


Figure 2. Recruitment of DNA damage signaling kinases and adaptor proteins to DNA lesions: conserved features between budding yeast and humans

Phosphorylation and adaptor proteins play a key role in the recruitment of downstream checkpoint kinases. The colored ovals indicate phosphorylation events mediated by DNA damage signaling kinases (see kinase key). The orange lines indicate protein–protein interactions promoted by the indicated phosphorylation events (also methylation (me) or ubiquitylation (Ub)). Activation of the downstream checkpoint kinases by the apical PIKK kinases requires adaptor proteins (outlined in green). In most cases, these adaptor proteins act as scaffolds to directly bind to and recruit the downstream checkpoint kinase. The model, mostly based on extensive work in yeast, posits that the recruitment of the downstream checkpoint kinase to the proximity of the apical PIKK kinase enables the phosphorylation and activation of the downstream checkpoint kinase. In addition to activating the downstream checkpoint kinase, phosphorylation events mediated by the apical PIKK kinases are critical for scaffold assembly, often promoting protein–protein interactions. Accordingly, a conserved feature of several adaptor proteins in budding yeast and humans is the presence of protein domains responsible for binding phosphorylated proteins (FHA and BRCT domains). Notably, other kinases such as CDK and CK2 also catalyze phosphorylation events involved in adaptor recruitment, although these events are often not induced by DNA damage. For DNA-PKcs, while this kinase has been implicated in the phosphorylation of H2AX and 53BP1, it does not seem to be involved in CHK2 phosphorylation.

Similar as in yeast, vertebrate ATM and ATR utilize checkpoint adaptor proteins to mediate their transfer of phosphorylation to the checkpoint effector kinases. ATR primarily relies on Claspin, the homolog of yeast Mrc1, to mediate the activation of CHK1 (Kumagai& Dunphy, 2000). Like Mrc1, Claspin associates with the replisome (Lee *et al*, 2003) and recruits CHK1 upon exposure to replication stress. Also similar to yeast, the Claspin–CHK1 interaction depends on ATR activity (Kumagai& Dunphy, 2003; Lindsey-Boltz *et al*, 2009). While it is currently unknown whether mammalian ATR directly phosphorylates Claspin, *Xenopus* ATR has been shown to directly phosphorylate Claspin at threonines 817 and 819, which are critical for CHK1 recruitment (Fig 2; the uncertainty of ATR's direct phosphorylation of Claspin in humans is denoted by "?"; Yoo *et al*, 2006; in-depth discussion of the similarities and differences between Claspin orthologs can be found here: Smits *et al*, 2019). Once bound to CHK1, Claspin is thought to both stabilize CHK1 and tether it in proximity to ATR, allowing for extensive phosphorylation and full activation of the effector kinase (Liu *et al*, 2006). While ATR utilizes Claspin to facilitate its phosphorylation of CHK1, the phosphorylation of many other ATR substrates does not require Claspin, suggesting that, like in yeast, the adaptor proteins are primarily responsible for facilitating Mec1/ATR's phosphorylation of the effector kinases and are not required for the phosphorylation of other substrates (like many of those depicted below in Fig 4).

Two mammalian adaptors have been linked to the ATM–CHK2 signaling axis: MDC1 and 53BP1. Despite extensive research on these proteins, it remains unclear precisely how they function in transducing ATM signaling toward CHK2 activation. While 53BP1 is the functional ortholog of yeast Rad9, MDC1 also shares functional similarities with Rad9 in the context of the DNA damage signaling response. Similar to Rad9, MDC1 possesses BRCT domains that directly bind to phosphorylated histone H2AX (Fig 2; γH2AX; analogous to histone H2A phosphorylated at the C-terminal SQ site in yeast; Stucki *et al*, 2005). Once associated with γH2AX at DNA breaks, MDC1 is phosphorylated by ATM and contributes to CHK2 activation (Goldberg *et al*, 2003; Peng & Chen, 2003; Stewart *et al*, 2003; Wu *et al*, 2008). MDC1 has also been shown to be important for CHK1 activation through its interaction with the TOPBP1 scaffold (Wang *et al*, 2011; Leung *et al*, 2013), which may functionally resemble the Rad9–Dpb11 interaction in budding yeast (Fig 2).

ATM promotes the recruitment of 53BP1 through two distinct mechanisms. Similar to yeast Rad9, 53BP1 recognizes γH2AX through its C-terminal pair of BRCT domains (Fig 2), an interaction which has been controversial, but has recently gained additional support (Baldock *et al*, 2015; Kleiner *et al*, 2015). Nonetheless, the prominent mechanism of 53BP1 recruitment to DNA breaks involves ATM- and γH2AX-mediated recruitment of MDC1, which becomes phosphorylated by ATM and recruits the E3 ubiquitin ligases RNF8 and RNF168, leading to ubiquitylation of H2A that is recognized directly by 53BP1 (Fig 2; for a detailed review, Hustedt& Durocher, 2016). Notably, this ubiquitylation-dependent recruiting mechanism is absent in budding yeast. Consistent

with an adaptor function for 53BP1, it interacts with CHK2, and the loss of 53BP1 results in reduced ATM-mediated phosphorylation of CHK2 in response to low doses of ionizing radiation (IR; Wang *et al*, 2002). However, 53BP1 appears to regulate activation of the checkpoint through a more complex mechanism, as the physical interaction between CHK2 and 53BP1 rapidly decreases upon IR radiation rather than becoming stabilized (Wang *et al*, 2002), in contrast to the Rad9–Rad53 interaction, which increases after DNA damage in budding yeast. Both 53BP1 and Rad9 play a key role in the control of DNA end resection, highlighting the connection and coordination between checkpoint signaling and the regulation of DNA repair (discussed in detail later in this review; Lazzaro *et al*, 2008; Bunting *et al*, 2010; Chapman *et al*, 2013; Zimmermann *et al*, 2013; Ferrari *et al*, 2015; Liu *et al*, 2017). In both cases, phosphorylation of Rad9/53BP1 by Mec1 or Tel1 in yeast or ATM in humans is essential for suppressing DNA end resection (Bothmer *et al*, 2011; Ferrari *et al*, 2015), and it is possible that 53BP1's function in preventing resection contributes to the stabilization of ATM at breaks, which may indirectly promote ATM–CHK2 signaling.

Post-recruitment events in downstream checkpoint kinase activation

Recruitment to sites of DNA lesions via checkpoint adaptors enables the downstream checkpoint kinases to be directly phosphorylated by the upstream PIKKs, triggering initial kinase activation and subsequent autophosphorylation for further kinase activation. In the case of Rad53, for example, initial phosphorylation by Mec1 or Tel1 promotes its kinase activity by interfering with a kinase auto-inhibitory domain (Fiorani *et al*, 2008), which then stimulates Rad53 to phosphorylate other Rad9-bound Rad53 molecules (Jia-Lin Ma & Stern, 2008). Such trans-autophosphorylation events contribute to dissociate Rad53 from Rad9 and prevent further Rad9 oligomerization (Usui *et al*, 2009). Notably, overexpression of Rad53 in bacteria cells, which lack PIKKs and checkpoint adaptors, results in hyper-activated Rad53 (Gilbert *et al*, 2001). This finding supports a model whereby increased local concentration of Rad53 is enough for activation, with adaptors building increased local concentration at the site of lesions and PIKKs facilitating the initial trigger by reducing the minimal concentration threshold required for activation. In mammals, the apical kinases ATR and ATM drive the key events leading to activation of CHK1 and CHK2, respectively. ATR- and ATM-mediated phosphorylation not only recruits CHK1 and CHK2 to sites of DNA lesions, but also directly phosphorylates these downstream kinases to promote their activation. Like in yeast, these priming phosphorylation events are mainly required to relieve inhibitory domains or to drive monomer-to-oligomer kinase transition (reviewed in Bartek & Lukas, 2003). ATM-mediated phosphorylation of CHK2 at threonine 68, an established marker of CHK2 activation, allows for the dimerization of two inactive CHK2 monomers and for their subsequent trans- and cis-phosphorylation (Ahn *et al*, 2000; Xu *et al*, 2002; Schwarz *et al*, 2003). CHK2 dimerization is a transient state, since the multiple trans- and cis-autophosphorylation events promote rapid dimer dissociation, leading to full active monomers (Ahn &Prives, 2002; Xu *et al*, 2002; Cai *et al*, 2009).

Similar to CHK2, CHK1 activation is a multistep process that requires ATR phosphorylation at serine 317 and serine 345; however, unlike Rad53 or CHK2, CHK1 activation does not appear to involve dimerization or oligomerization (Liu *et al*, 2000; Zhao &Piwnica-Worms, 2001).

Substrates mediating the core DNA damage signaling responses:

Once activated, DNA damage signaling kinases mediate hallmark responses that include the arrest of the cell cycle, inhibition of origin firing, protection and restart of stalled replication forks, induction of a transcriptional response, initiation of apoptosis, and control of dNTP levels. More recent work has demonstrated that the DNA damage signaling kinases also regulate a range of other processes, such as autophagy, gene gating, chromosome mobility, transcription–replication conflicts, and many more whose mechanistic connections to DNA damage signaling and degrees of conservation across eukaryotes remain less clear. Here, we focus on a select set of core conserved functions of the DNA damage signaling kinases with defined substrates, delineating the parallels between budding yeast and humans (Fig 3).

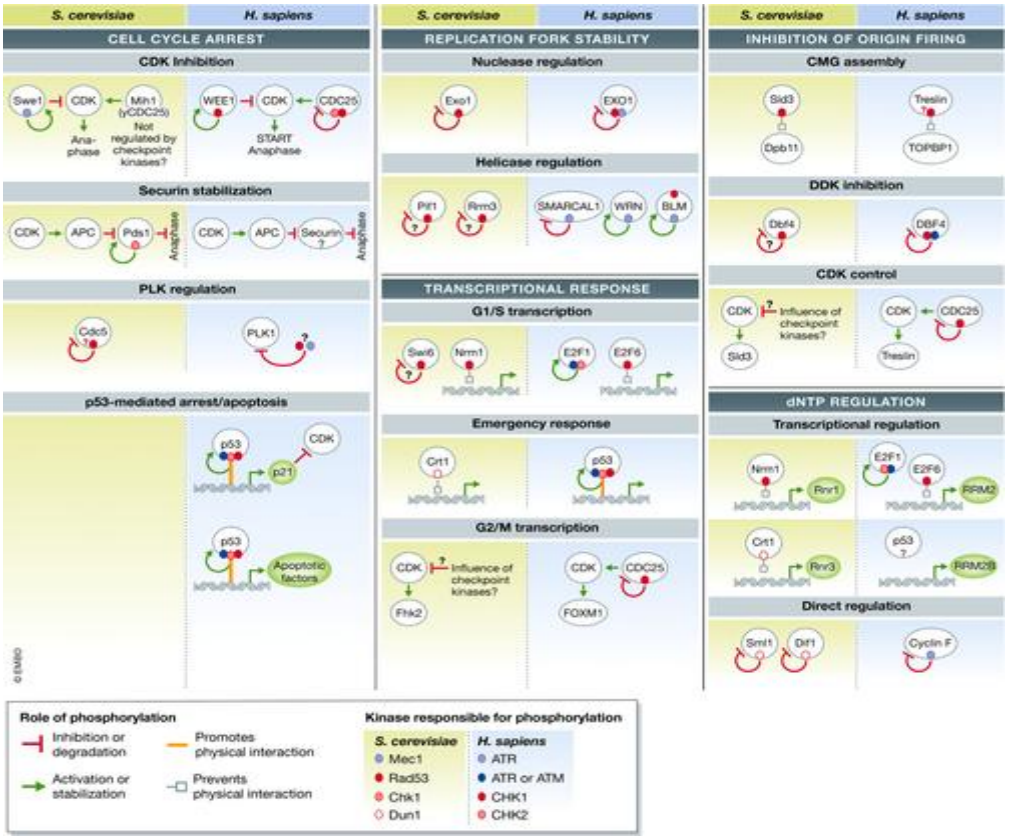


Figure 3. Yeast-to-human parallels in core checkpoint responses mediated by DNA damage signaling

Substrate map highlighting the phosphorylation events involved in core DNA damage signaling responses in yeast and humans (see text for detailed discussion of each substrate). Conserved or functionally analogous phosphorylation events are positioned parallel to each another. The colored ovals indicate phosphorylation events mediated by DNA damage signaling kinases (see "kinase-dependency" key). The arrows or lines that emanate from the colored ovals represent the role phosphorylation plays in regulating that protein (see "role of phosphorylation" key). Question marks indicate uncertainty, either in the functionality of the phosphorylation event or in the identity of the kinase or substrate. Arrows that impinge on CDK demonstrate how DNA damage signaling can indirectly inhibit CDK activity.

Cell cycle control

The most classical and widely known function of DNA damage signaling kinases is the imposition of a cell cycle arrest that prevents entry into mitosis. Exactly how this arrest is imposed in budding yeast remains elusive. The paradigm for how cell cycle arrest occurs comes from a series of works conducted in fission yeast and metazoans in 1997 (Furnari *et al*, 1997; Peng *et al*, 1997; Sanchez *et al*, 1997; Weinert, 1997), which revealed that DNA damage signaling inhibits CDC25, a phosphatase that removes inhibitory phosphorylation at a key tyrosine residue in mitotic-CDK (M-CDK; Fig 3). In addition, DNA damage signaling kinases stimulate WEE1 (O'Connell *et al*, 1997; Boddy *et al*, 1998; Lee *et al*, 2001), a kinase responsible for phosphorylating the same inhibitory tyrosine site (Fig 3; Mueller *et al*, 1995). While DNA damage signaling in budding yeast does not appear to impinge upon the Cdc25 phosphatase homolog Mih1, the budding yeast kinase Swe1 (the WEE1 homolog) is regulated similar to its counterpart in higher eukaryotes and fission yeast. Swe1 is likely phosphorylated and activated by DNA damage signaling, which results in inhibition of M-CDK (Fig 3; Edenberg *et al*, 2015; Palou *et al*, 2015). DNA damage signaling in budding yeast is also able to suppress M-CDK activity through additional redundant mechanisms that remain unclear (Palou *et al*, 2015). In addition to inhibiting M-CDK-dependent activation of the anaphase-promoting complex (APC/C), DNA damage signaling in budding yeast more directly inhibits the onset of anaphase through Chk1, which phosphorylates and stabilizes the anaphase-inhibiting Pds1/securin protein (Fig 3; Sanchez *et al*, 1999; Wang *et al*, 2001), preventing the separation of sister chromatids. Moreover, Rad53 can influence the mitotic spindle assembly checkpoint (SAC) through inhibition of the polo-like kinase Cdc5 (Fig 3; Sanchez *et al*, 1999; Valerio-Santiago *et al*, 2013; Zhang *et al*, 2009). Rad53-dependent inactivation of Cdc5, particularly in response to telomere damage, prevents Cdc5's phosphorylation of the spindle assembly checkpoint protein Bfa1 (Valerio-Santiago *et al*, 2013); however, it is unclear whether Rad53 phosphorylates Cdc5 directly (Fig 3). The inhibition of Cdc5 by DNA damage signaling also indirectly influences the resolution of joint molecules prior to M phase, as Cdc5 activity is important for promoting the activity of the Mus81 resolvase (Szakal&Branzei, 2013). Cdc5 has also been implicated in the down-regulation of the

DNA damage response (Donnianni *et al*, 2010; Vidanes *et al*, 2010). A more in-depth review exploring Cdc5's complex relationship with the DNA damage response in yeast can be found in Botchkarev and Haber (2018). Similar to its budding yeast counterpart, the mammalian polo-like kinase PLK1 is also inhibited by DNA damage signaling, albeit in an indirect manner (Fig 3; Bruinsma *et al*, 2017; Qin *et al*, 2013).

A critical mediator of cell cycle arrest in humans is the p53 transcription factor, whose classical function is to trigger the apoptotic program (reviewed in Chen, 2016). p53 is directly phosphorylated and stabilized by all the DNA damage signaling kinases, eliciting a p53-dependent transcriptional response that impacts the DNA damage response (reviewed in Kruse & Gu, 2009). p53-mediated expression of the CDK inhibitor protein p21 represents the primary mechanism by which p53 blocks progression through the cell cycle (Fig 3; Harper *et al*, 1993, 1995). Apical kinases in mammalian DNA damage signaling also phosphorylate the p53 inhibitor Mdm2, impairing its ability to promote p53 nuclear export and its subsequent degradation (Mayo *et al*, 1997; Maya *et al*, 2001; Shinozaki *et al*, 2003). Unlike most factors covered in Fig 3, *Saccharomyces cerevisiae* does not have clear p53 or Mdm2 homologs. That said, for reasons discussed in the section on dNTP regulation, the DNA damage signaling-mediated control of p53 might functionally resemble the control of the Crt1 transcription regulator by Rad53 and Dun1 (Huang *et al*, 1998).

Fork stability and protection

During DNA replication, stalled replication forks may be targeted and degraded by nucleases. Since fork degradation pathways impair replication fork restart after stalling, leading to persistent DNA lesions, they are considered a major driving force of genomic instability (recently reviewed in Pasero&Vindigni, 2017; Patel & Weiss, 2018). In *S. cerevisiae,* Rad53 is believed to play a major role in fork stability. It is worth mentioning that Rad53 is essential, and that the lethality of rad53 cells can be rescued by deletion of *SML1*, an inhibitor of the ribonucleotide reductase enzyme responsible for catalyzing the rate-limiting step in DNA precursor synthesis (Zhao *et al*, 1998). Nonetheless, even in the absence of *SML1*, Rad53 mutants are exquisitely sensitive to chemical agents that damage DNA or stall DNA replication forks (Allen *et al*, 1994; Sanchez *et al*, 1999; Gunjan &Verreault, 2003). This sensitivity to DNA damaging agents has been primarily attributed to Rad53's role in stabilizing and restarting stalled replication forks (reviewed in detail here: Segurado&Tercero, 2009). Rad53 protects stalled replication forks from nucleolytic processing by phosphorylating and inhibiting the Exo1 exonuclease (Fig 3; Cotta-Ramusino *et al*, 2005; Morin *et al*, 2008; Segurado&Diffley, 2008). In addition, Rad53 regulates the Pif1 and Rrm3 helicases, potentially limiting replication fork reversal and subsequent fork degradation (Fig 3; Rossi *et al*, 2015). However, precisely how DNA damage signaling maintains replication fork integrity is unknown and represents a fundamental knowledge gap in the field. Recent work has revealed that Rad53 phosphorylates the replicative helicase

component Cdc45, which in turn recruits and stabilizes Rad53 at replication complexes (Can *et al*, 2018). Identification of additional key substrates and recruiting mechanisms will be necessary to deconstruct Rad53's functions in maintaining fork stability, which may involve a range of redundant phosphorylation events in more than one essential replisome protein.

Mammalian DNA damage signaling kinases also play key roles in protecting stalled replication forks (Fig 3). Human EXO1 is phosphorylated by ATR, which promotes ubiquitylation-dependent degradation of EXO1 and prevents chromosome fragmentation due to unrestrained EXO1 processing activity (El-Shemerly *et al*, 2008; Tomimatsu *et al*, 2017). CHK1 also phosphorylates EXO1 directly on serine 746, creating a docking site for binding to 14-3-3 proteins, which prevent recruitment of EXO1 to chromatin and limit EXO1 action at stalled forks (Engels *et al*, 2011; Li *et al*, 2019). In addition, ATR governs the recruitment and/or stability of several helicases important for remodeling the stalled fork and promoting fork restart. For example, ATR phosphorylates the Werner syndrome helicase WRN at multiple S/T-Q sites, promoting WRN-RPA co-localization at replication stress sites (Ammazzalorso *et al*, 2010). Mutation of these ATR sites causes stalled fork breakage and severely impacts the resumption of DNA replication. Several studies have demonstrated that the Bloom syndrome helicase, BLM, is directly phosphorylated by ATR and that this phosphorylation is important for both promoting replication fork restart after HU-mediated arrest and suppressing dormant origin firing (Davies *et al*, 2004, 2007). CHK1 may also constitutively phosphorylate BLM to prevent proteasome-dependent BLM degradation, suppress chromatin bridge formation, and promote an interaction with 53BP1 (Sengupta *et al*, 2004; Tripathi *et al*, 2007; Kaur *et al*, 2010; Petsalaki *et al*, 2014). Interestingly, in yeast, Rad53 and Sgs1 (a BLM ortholog) physically interact (Hegnauer *et al*, 2012), and Sgs1 has been reported to display a strong genetic interaction with Rad9, 53BP1's yeast homolog (Nielsen *et al*, 2013). Finally, David Cortez's group described an additional mechanism of fork stability regulation through the ATR's phosphorylation of the SMARCAL1 helicase (Couch *et al*, 2013). ATR-dependent phosphorylation of SMARCAL1 inhibits fork remodeling activity, preventing subsequent pathological fork degradation by active nucleases (Couch *et al*, 2013; Kolinjivadi *et al*, 2017). Collectively, these examples highlight how yeast and human DNA damage signaling kinases protect replication fork integrity via the direct regulation of nucleases and helicases.

Inhibition of origin firing

One mechanism for preventing genome instability during replication stress is to inhibit the firing of late origins (reviewed in McIntosh & Blow, 2012; Yekezare *et al*, 2013). DNA damage signaling kinases control origin firing through two main pathways, which are conserved from yeast to higher eukaryotes (Fig 3). During genotoxic stress in S phase, Rad53 phosphorylates Dbf4, a subunit of the Dbf4-dependent kinase (DDK)

complex, and the Sld3 component of pre-RCs (pre-replication complexes) to inhibit the firing of late-replicating origins (Lopez-Mosqueda *et al*, 2010; Zegerman&Diffley, 2010), thereby slowing the progression of DNA synthesis and preventing the exhaustion of RPA (Toledo *et al*, 2013, 2017). While the mechanism by which Rad53 phosphorylation regulates DDK is unknown, the phosphorylation of Sld3 by Rad53 is thought to prevent the recruitment of Dpb11 to primed replication origins (Lopez-Mosqueda *et al*, 2010; Zegerman&Diffley, 2010). In vertebrates, the Dpb11 ortholog TOPBP1 docks to CDK2-phosphorylated Treslin (the Sld3 ortholog) to mediate origin firing (Kumagai *et al*, 2010; Boos *et al*, 2011). Conditions that perturb DNA replication and trigger CHK1 activation have been shown to disrupt the Treslin–TOPBP1 interaction (Boos *et al*, 2011) in a mechanism analogous to Rad53-dependent control of the Sld3–Dpb11 interaction in *S. cerevisiae*. CHK1 has also been shown to directly phosphorylate Treslin in *Xenopus* egg extracts, inhibiting DNA replication (Guo *et al*, 2015). Metazoan ATR/ATM and CHK1 phosphorylate the DBF4 subunit of DDK, inhibiting pre-RC assembly and origin firing (Costanzo *et al*, 2003; Heffernan *et al*, 2007; Lee *et al*, 2012). CHK1 is also assumed to inhibit origin firing indirectly by suppressing CDK activity via the inhibition of CDC25 (Shechter *et al*, 2004; Sorensen &Syljuasen, 2012). Consistent with this assumption, inhibition of CDK rescues replication stress sensitivity of cells lacking ATR or CHK1 signaling, likely by reducing origin firing and preventing RPA exhaustion (Toledo *et al*, 2013; Dungrawala *et al*, 2015). Paradoxically, Mec1 (and potentially ATR/ATM) also promotes DNA replication by phosphorylating core components of the replisome, such as the MCM helicase (Cortez *et al*, 2004; Yoo *et al*, 2004; Randell *et al*, 2010).

Transcriptional control

In yeast, Rad53 plays a central role in the transcriptional response to DNA damage and replication stress. In G1, Rad53 was proposed to influence the timing of START by phosphorylating Swi6, a component of the MBF/SBF transcription factor (analog of human E2F; Sidorova & Breeden, 1997, 2003). Rad53 can also regulate the transcription of MBF/SBF targets at the G1/S transition and during S phase through phosphorylation and inhibition of the Nrm1 transcriptional repressor protein, promoting transcription of the largest set of co-regulated genes in the DNA damage response (Fig 3; Bastos de Oliveira *et al*, 2012; de Bruin *et al*, 2008; Travesa *et al*, 2012). In addition, Dun1 up-regulates the transcription of a specific set of DNA damage-induced genes, including subunits of ribonucleotide reductase (RNR), by inducing the phosphorylation and inactivation of the transcriptional repressor protein, Crt1 (Huang *et al*, 1998; Fig 3). In-depth transcriptome analyses performed in budding yeast suggest that DNA damage signaling kinases may impinge upon other, as yet unknown, transcription factors (Jaehnig *et al*, 2013).

DNA damage signaling kinases profoundly impact multiple transcriptional programs in mammals (Fig 3). The E2F pathway (analogous to yeast SBF/MBF) represents a major

transcriptional system regulated by DNA damage signaling kinases. ATM-, ATR-, and CHK2-dependent phosphorylation activates E2F1 (Lin *et al*, 2001; Stevens *et al*, 2003), whereas CHK1 inhibits the transcriptional repressor E2F6 through direct phosphorylation (Bertoli *et al*, 2013a). Interestingly, CHK1's inhibition of E2F6 resembles the Rad53-dependent inhibition of the *S. cerevisiae* transcriptional repressor Nrm1, as discussed above (more details can be found here: Bertoli *et al*, 2013b). Similar to yeast, E2F-dependent transcription in humans includes a large set of genes that play a role in DNA replication, cell cycle, and DNA repair (reviewed here: Bracken *et al*, 2004; Poppy Roworth *et al*, 2015). Not surprisingly, inhibition of the ATR–CHK1 pathway causes major remodeling of the proteome via control of the E2F transcriptional circuit and strongly impacts the DNA repair machinery (Kim *et al*, 2018).

Finally, recent work in mammalian cells has revealed a role for ATR in enforcing a checkpoint between the S and G2 phases of the cell cycle. ATR suppresses CDK-dependent phosphorylation and activation of the FOXM1 transcription factor by limiting CDK activity during S phase (Fig 3), thereby preventing a premature expression of G2-specific genes (Saldivar *et al*, 2018). While the function of the FOXM1 transcription factor is conserved in yeast (Pic-Taylor *et al*, 2004; Murakami *et al*, 2010), it is unclear whether the budding yeast checkpoint kinases also restrict the expression of G2/M-specific transcripts through an analogous mechanism (Fig 3).

Regulation of dNTP production

In budding yeast, the regulation of dNTP production is considered an essential function of DNA damage signaling. This is supported by the fact that the lethality caused by deletion of *MEC1* or *RAD53* can be suppressed by deletion of *SML1,* an inhibitor of the ribonucleotide reductase enzyme complex, which catalyzes the limiting step in dNTP production (Zhao *et al*, 1998). The Rad53–Dun1 signaling axis induces the activity of the RNR complex by controlling several of its subunits via distinct mechanisms (Allen *et al*, 1994; Bashkirov *et al*, 2003; Fig 3). Once activated by Rad53, Dun1 directly phosphorylates Sml1, resulting in its degradation and the activation of the Rnr1 subunit (Zhao *et al*, 2001; Zhao & Rothstein, 2002; Lee *et al*, 2008). In addition, Dun1-mediated phosphorylation of Dif1, an *SML1* paralog, results in Dif1's degradation and the subsequent export of the Rnr2–Rnr4 subunits from the nucleus to the cytosol, which enables the formation of the active RNR complex (Lee *et al*, 2008). Dun1 also up-regulates the transcription of *RNR3* by inducing the phosphorylation and inactivation of the transcriptional repressor protein, Crt1 (Huang *et al*, 1998; Fig 3). Rad53 also plays a Dun1-independent role in RNR control. For example, Rad53's phosphorylation of the transcriptional repressor protein Nrm1, upon exposure to replication stress, results in the MBF-mediated induction of RNR1 transcription (Bastos de Oliveira *et al*, 2012; Huang *et al*, 1998; Travesa *et al*, 2012; Fig 3). Since cells lacking Dun1 are viable, in contrast to *rad53Δ* cells, it is possible that the essential function of Rad53 is due to both its Dun1-dependent and Dun1-independent roles in modulating RNR or, alternatively, a

combination of Rad53's function in dNTP regulation (via Dun1) and other roles in, for example, replication fork stability.

Despite the critical importance of dNTP regulation in yeast, no clear Sml1, Dif1, or Dun1 orthologs or functional analogs have been identified so far in mammalian cells. Nevertheless, it is intriguing that an extra genomic copy of the RRM2 gene (RNR regulatory subunit) can extend the lifespan of ATR-deficient mice (Lopez-Contreras *et al*, 2015). Whether ATR directly phosphorylates RRM2 is to date not clear, although ATR may promote RRM2 stabilization by regulating the degradation of the Cyclin F subunit of the SCF ubiquitin ligase complex (Fig 3), which prevents SCF-mediated degradation of RRM2 (D'Angiolella *et al*, 2012). RRM2 transcription is also induced by E2F and thus can be indirectly up-regulated by the ATR–CHK1 kinases (as discussed in the section "transcriptional control"; Ishida *et al*, 2001). Interestingly, p53 promotes transcription of the RRM2B subunit of human RNR (Tanaka *et al*, 2000), and it is tempting to speculate on a potential analogy to the role of yeast Crt1 in inducing RNR3 transcription (Huang *et al*, 1998). In response to hypoxia, this p53-dependent induction of RRM2B is important for maintaining the fidelity of DNA replication in low oxygen conditions (Foskolou *et al*, 2017). However, a direct link to the DNA damage signaling kinases has yet to be made.

Control of DNA repair by DNA damage signaling kinases

The control of DNA repair is a core function of DNA damage signaling kinases. Mounting genetic evidence from budding yeast illustrates the central role of DNA damage signaling to maintaining genome integrity. For example, using a genetic assay to assess chromosome instability in yeast, known as the gross chromosomal rearrangement (GCR) assay, the Kolodner Lab showed that cells lacking Mec1 and Tel1 exhibit extremely high rates of GCRs (Myung *et al*, 2001). In fact, *mec1Δ tel1Δ* cells are some of the most genetically unstable strains isolated to date. These cells undergo GCRs at a rate of five orders of magnitude higher than wild-type cells (Myung *et al*, 2001). As discussed below, current evidence supports a model whereby the ability of Mec1 and Tel1 to suppress GCRs and related forms of genetic instabilities is associated with the role these kinases play in directly targeting and regulating components of the DNA repair machinery. In mammals, ATM and ATR are also responsible for maintaining chromosome stability (Xu *et al*, 1996; Brown & Baltimore, 2000); as shown in Fig 4, ATM and ATR phosphorylate and regulate a large set of DNA repair factors. Importantly, the high levels of genetic instability in yeast or mammalian cells lacking Mec1/ATR and Tel1/ATM have not yet been recapitulated by any combination of mutations of phosphorylation sites, suggesting that the key substrates through which these kinases suppress genomic instabilities remain unclear.

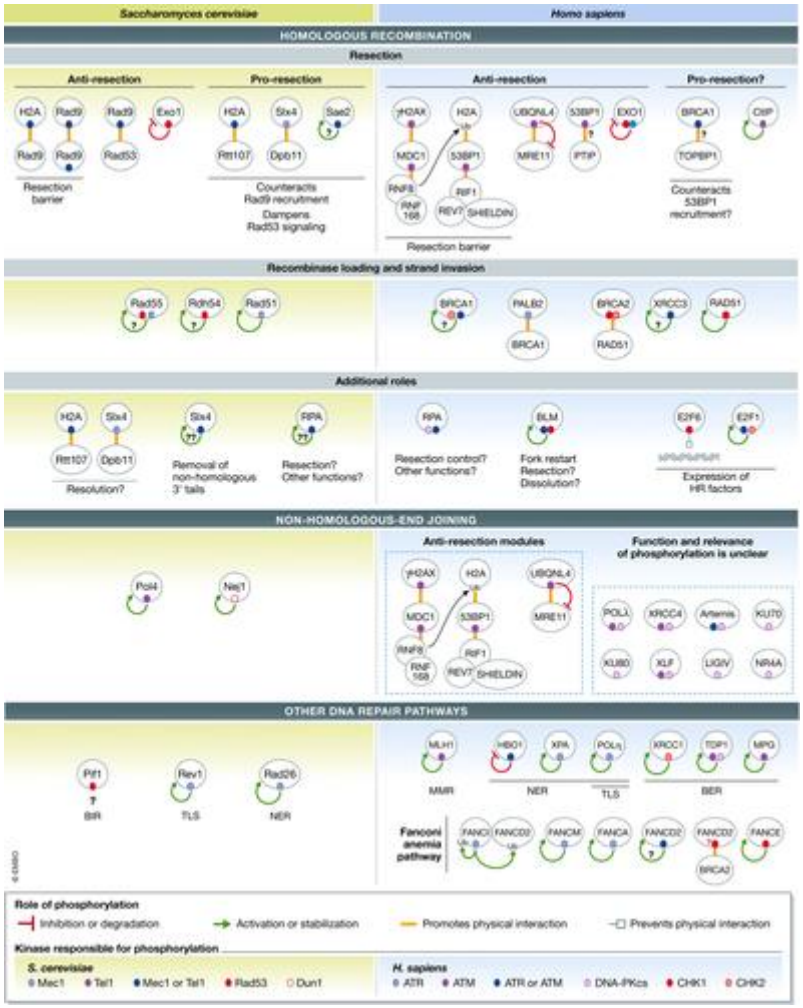


Figure 4. Yeast-to-human parallels in the regulation of DNA repair proteins by DNA damage signaling: substrates and mechanisms

Substrate map cataloging the DNA damage signaling events regulating DNA repair proteins (see text for detailed discussion of each substrate). Conserved or analogous substrates involved in related DNA repair pathways are positioned parallel to each other. The colored ovals indicate phosphorylation events mediated by DNA damage signaling kinases (see "kinase-dependency" key). The arrows or lines that emanate from the colored nodes represent the role phosphorylation plays in regulating that protein (see "role of phosphorylation" key). Question marks indicate uncertainty, either in the functionality of the phosphorylation event or in the identity of the kinase or substrate.

Phosphorylation control of the homologous recombination machinery: substrates and mechanisms

In both yeast and humans, DNA damage signaling kinases phosphorylate proteins in homologous recombination (HR)-directed repair and have been shown to play an important regulatory role. For example, Mec1 phosphorylates Rtt107, Slx4, and Sgs1, all of which are known to play multiple roles in HR-mediated repair, from the control of resection to the dissolution or resolution of joint chromosome structures (Sarbajna& West, 2014; Hang & Zhao, 2016; Cussiol *et al*, 2017; Guervilly& Gaillard, 2018). Recent work from our laboratory has revealed that phosphorylation of these proteins correlates with the ability of cells to suppress GCRs, suggesting that proper control of HR repair is essential for preventing chromosomal rearrangements (Lanz *et al*, 2018). Phosphorylation of Slx4 by Mec1 promotes DNA end resection, the first step in the HR process (Dibitetto *et al*, 2016; Liu *et al*, 2017). Mechanistically, Mec1 phosphorylation mediates the interaction between Slx4 and Dpb11, a multi-BRCT domain scaffold that bridges Slx4 to the 9-1-1 clamp loaded at 5′ recessed junctions (Ohouo *et al*, 2010; Cussiol *et al*, 2015; Fig 4). Such stabilization of Slx4 at DNA lesions, which is also dependent on Rtt107's interaction with phosphorylated H2A (Fig 4), is believed to counteract a resection block promoted by Rad9, thereby allowing long-range resection to occur (Dibitetto *et al*, 2016; Liu *et al*, 2017). Importantly, the role of Rad9 in counteracting resection relies on its oligomerization and Rad53 signaling (Clerici *et al*, 2014; Ferrari *et al*, 2015; Gobbini *et al*, 2015), which are both dependent on Mec1-mediated phosphorylation events. Therefore, Mec1 plays opposing roles in HR, inhibiting resection via the Rad9–Rad53 signaling axis or promoting resection by mediating the Slx4–Dpb11 interaction (Fig 4).

Mec1's phosphorylation of Slx4 has also been linked to a role in single-strand annealing (SSA), specifically in regulating cleavage of 3′ non-homologous DNA tails by Rad1–Rad10 (Toh *et al*, 2010). In another example, which supports a key role for Mec1 in the control of HR, phosphorylation of multiple S/T-Q residues in Sae2, a pro-resection factor homologous to human CtIP, contributes to Mec1-mediated GCR suppression (Liang *et al*, 2015). Furthermore, Mec1 phosphorylates the strand exchange factor Rad55, and Rad53 promotes the phosphorylation and DNA binding of the Tid1/Rdh54 translocase (Fig 4; Bashkirov *et al*, 2006; Ferrari *et al*, 2013; Herzberg *et al*, 2006). Mec1 also phosphorylates Rad51 at serine 192, likely supporting its ATPase activity and capacity for strand invasion (Fig 4; Flott *et al*, 2011).

The ssDNA binding protein RPA, which plays a critical role in coating resected DNA ends during the homologous recombination process, is one of the most established targets of the PIKKs. In both yeast and humans, RPA, through an interaction with Rad52, is thought to promote the loading of Rad51 on ssDNA (Park *et al*, 1996; Sugiyama *et al*, 1998; Sugiyama &Kowalczykowski, 2002). How PIKK phosphorylation regulates RPA function is not well understood, although studies point to a role in

homologous recombination. RPA phosphorylation was initially proposed to promote the interaction between RPA and Rad52 (Deng *et al*, 2009); however, recent work using single-molecule imaging reports that RPA phosphorylation inhibits DNA end resection via inhibition of the BLM helicase (Soniat *et al*, 2019). RPA phosphorylation in human cells has also been reported to impact DNA synthesis under stress (Vassin *et al*, 2009), the recruitment of factors to stalled replication forks (Murphy *et al*, 2014), and the imposition of the cell cycle checkpoint (Vassin *et al*, 2009). In budding yeast, mutation of the residues in RPA targeted by Mec1 and Tel1 has yet to yield an observable phenotype. Overall, the accumulated evidence supports a model whereby DNA damage signaling plays a key role in the control of the HR machinery by targeting multiple components that participate in discrete steps of the HR process (Fig 4).

ATR is a key regulator of homologous recombination-mediated repair

ATR's primary action occurs in S phase, a period during the cell cycle where a sister DNA template is available for homology-directed repair, and it functions mainly as a pro-HR kinase. Depletion or inhibition of ATR impairs the ability of cells to utilize HR and leads to synergistic cell death with replication stress-inducing drugs (Wang *et al*, 2004a; Vriend *et al*, 2016; Yazinski *et al*, 2017; Kim *et al*, 2018). One model proposes that during HR-mediated repair, an ATM-to-ATR transition occurs, where ATM initiates resection and triggers the ATR activation that governs the later steps of homologous recombination (Cuadrado *et al*, 2006; Shiotani& Zou, 2009). Our laboratory proposed that ATR drives HR by promoting the stabilization of the pro-resection factor BRCA1 at DNA lesions via interaction with TOPBP1 (human ortholog of Dpb11), which may counteract recruitment of the anti-resection factor 53BP1 to sites of DNA lesions (Liu *et al*, 2017). It is tempting to speculate that this mechanism could be analogous to the Mec1-mediated Dpb11–Slx4 interaction in yeast, which is also important for DSB resection (Dibitetto *et al*, 2016; Liu *et al*, 2017; Ohouo *et al*, 2010; Fig 4). A recent study from the Zou Lab provided additional insights into how ATR promotes HR-mediated repair. They showed that while initial DSB resection requires CDK activity, later steps in HR require a "CDK-to-ATR switch" to promote proper recruitment of the key HR factors PALB2 and BRCA2, which are required for strand invasion. Mechanistically, ATR activated at resected ends recruits PALB2–BRCA2 to DNA damage sites by phosphorylating PALB2 on serine 59 to promote its interaction with BRCA1 (Buisson *et al*, 2017). At the same time, ATR inhibits CDK to prevent a CDK-dependent phosphorylation at PALB2's serine 64 that inhibits the BRCA1–PALB2 interaction (Buisson *et al*, 2017). Finally, the control of E2F transcription via ATR–CHK1 signaling strongly impacts the ability of cells to utilize HR-mediated repair by ensuring proper expression of key components of the HR machinery (Kim *et al*, 2018). Similar to yeast, RAD51 deposition and RAD51-mediated strand invasion are regulated by DNA damage signaling kinases. CHK1 has been shown to drive the formation of RAD51 foci at stalled replication forks, likely through direct phosphorylation (Sorensen *et al*, 2005). CHK1 and CHK2 may also be important to regulate the association of RAD51 with BRCA2

(Bahassi *et al*, 2008). Moreover, RAD51C, a putative homolog of yeast Rad55, associates with XRCC3, whose phosphorylation by ATM and ATR in response to IR is important for HR regulation (Somyajit *et al*, 2013).

ATM and DNA-PKcs in resection control

ATM signaling blocks DNA end resection via recruitment of an anti-resection complex formed by 53BP1-RIF1-REV7-SHIELDIN (Setiaputra& Durocher, 2019). This process is achieved through a defined order of events: phosphorylation of H2AX and MDC1, followed by recruitment of the RNF8 and RNF168 ubiquitin ligases, followed by ubiquitylation of H2A and recruitment of 53BP1, whose phosphorylation by ATM serves as a docking platform for RIF1 and, indirectly, the REV7–SHIELDIN complex (Fig 4). In addition, ATM limits resection by phosphorylating the proteasomal shuttle factor UBQLN4, which leads to degradation of MRE11, a component of the MRN complex required for initial steps in resection (Jachimowicz *et al*, 2019). Paradoxically, ATM may also promote resection by phosphorylating CtIP, a key factor required for resection initiation (Shibata *et al*, 2011). DNA-PKcs also plays opposing roles in resection control. Whereas binding of DNA-PKcs to DNA ends has a major anti-resection function by preventing recruitment of the EXO1 nuclease, DNA-PKcs' autophosphorylation is required to promote dissociation from DNA ends, thereby allowing EXO1 binding and resection (Zhou & Paull, 2013). More work is needed to better understand and define the extent to which the actions of ATM and DNA-PKcs impact overall resection in cells.

ATM and DNA-PKcs promote NHEJ

ATM and DNA-PKcs are key players in the control of non-homologous end joining (NHEJ), a pathway for DSB repair in which DNA ends are ligated without the need for a homologous template. ATM and DNA-PKcs play partially redundant roles in promoting NHEJ, as impairment of both kinases results in a stronger NHEJ defect compared with impairment of one of the kinases (Gapud&Sleckman, 2011; Gapud *et al*, 2011; Zha *et al*, 2011). In addition, combined ATM and DNA-PKcs deficiency leads to embryonic lethality in mice and a more severe DSB repair defect during immunoglobulin class-switch recombination than loss of only one of the kinases (Callen *et al*, 2009). The ability of ATM and DNA-PKcs to limit DNA end resection is a key part of their pro-NHEJ function; however, these kinases likely regulate additional processes to promote NHEJ, since chemical inhibition of both kinases severely inhibits NHEJ without inducing resection due to persistent DNA-PKcs at break ends. ATM and DNA-PKcs redundantly phosphorylate several NHEJ repair components, such as XLF, Artemis, and POLλ (Zhang *et al*, 2004; Goodarzi *et al*, 2006; Yu *et al*, 2008; Sastre-Moreno *et al*, 2017) (Fig 4). In addition, DNA-PKcs solely phosphorylates the XRCC4, KU70, KU80, and LIG-IV core components (Chan *et al*, 1999; Lee *et al*, 2004; Wang *et al*, 2004b; Douglas *et al*, 2005; Amiri Moghani *et al*, 2018). However, mutation of these phosphorylation sites, which are dependent on ATM and/or DNA-PKcs, does

not impair NHEJ or phenocopy the loss or inhibition of ATM and DNA-PKcs function. Notably, phosphorylation of a non-conventional DNA-PKcs substrate, the transcription factor NR4A, was proposed to be important for NHEJ, as demonstrated by a phosphosite mutation, but the impact of the mutation in NHEJ remains unclear (Malewicz *et al*, 2011).

In *S. cerevisiae*, little is known about the crosstalk between DNA damage signaling kinases and NHEJ repair factors. Tel1 phosphorylates Pol4, stimulating its gap-filling activity and preventing the appearance of deleterious chromosome translocations (Ruiz *et al*, 2013). Another report showed that the Dun1 kinase is also important for NHEJ by phosphorylating Nej1, a key activator of the NHEJ ligase Dnl4 (Ahnesorg& Jackson, 2007). DNA damage signaling kinases in budding yeast do not appear to impinge upon the NHEJ pathway as extensively as they do in humans and may reflect budding yeast's strong preference for homologous recombination-mediated DNA repair (Jasin& Rothstein, 2013).

Other DNA repair pathways

While DNA damage signaling kinases appear to extensively impinge upon multiple factors and steps of recombinational DNA repair, they also target and regulate other repair mechanisms (Fig 4), suggesting that their action is context-dependent and, in some cases, important in repair pathway choice. In budding yeast, Mec1 phosphorylates the nucleotide excision repair (NER) protein Rad26 and regulates the transcription-coupled NER mode of repair (Taschner *et al*, 2010). In humans, ATR also controls NER by phosphorylating XPA (Wu *et al*, 2006) and, in conjunction with ATM, the histone acetyltransferase HBO1. HBO1 phosphorylation stabilizes its interaction with DDB2, promoting its proteasome-mediated degradation (Matsunuma *et al*, 2016). In addition, ATR phosphorylates POLη at serine 601, facilitating post-replication repair and promoting checkpoint activation at UV damage sites (Gohler *et al*, 2011). In yeast, Mec1 regulates the translesion synthesis polymerase Rev1 potentially through direct phosphorylation (Pages *et al*, 2009). This complements earlier studies, where it was shown that Mec1 is essential for Rev1 binding to chromosomes (Hirano & Sugimoto, 2006; Sabbioneda *et al*, 2007). DNA damage signaling kinases are also important for the execution of mismatch repair (MMR) and base excision repair (BER). ATM phosphorylates the MMR protein MLH1, contributing to its stabilization after DNA damage (Romeo *et al*, 2011). Also, ATM directly phosphorylates the BER proteins TDP1 (Das *et al*, 2009) and MPG (Agnihotri *et al*, 2014), enhancing their DNA repair activity and promoting cell survival after exposure to alkylating agents. DNA-PKcs is also able to phosphorylate TDP1 at serine 81 (Das *et al*, 2009), allowing interaction with the DNA repair protein XRCC1, which is itself a CHK2 target for BER regulation (Chou *et al*, 2008). Finally, Rad53 was shown to be important in break-induced replication (BIR) through the phosphorylation of the Pif1 helicase (Vasianovich *et al*, 2014).

Fanconi anemia pathway

The Fanconi anemia pathway is a genetic network involved in the repair of DNA interstrand crosslinks (ICLs) in the genome (Niraj *et al*, 2019). It combines the action of nucleotide excision repair and homologous recombination factors, several of which are regulated by ATR (Fig 4). ATR plays a crucial role at an early step in the Fanconi anemia pathway by phosphorylating the FANCI protein and inducing ubiquitylation of the FANCI–FANCD2 complex (Ishiai *et al*, 2008; Shigechi *et al*, 2012). The ubiquitylated FANCI–FANCD2 complex serves as a key scaffold for the recruitment of several DNA repair proteins, including structure-specific nucleases and TLS polymerases involved in ICL repair (Knipscheer *et al*, 2009). In addition to promoting repair and replication fork start, phosphorylation of FANCI by ATR modulates DNA replication by inhibiting dormant origin firing (Chen *et al*, 2015). FANCD2 is also phosphorylated by ATM and ATR (Ho *et al*, 2006), whereas CHK1 phosphorylation is important for the FANCD2–BRCA2 interaction (Zhi *et al*, 2009). CHK1 phosphorylates FANCE at threonine 346 and serine 374 to promote resistance to mitomycin C (Wang *et al*, 2007). Additionally, ICL resistance has been linked to the ATR-mediated phosphorylation of FANCA at serine 1449 and FANCM at serine 1045 (Collins *et al*, 2009; Singh *et al*, 2013; Fig 4).

Coordination of checkpoint and DNA repair: where are we 30 years later?

Following the checkpoint paper by Weinert and Hartwell in 1988 (Weinert & Hartwell, 1988), it was not at all clear how eukaryotic cells sense DNA damage and couple lesion detection to the control of the cell cycle. At that time, Weinert and Hartwell defined checkpoint as "control mechanisms enforcing dependency in the cell cycle", implying the existence of some undefined regulatory mechanism. While the 1988 paper marks the identification of the first checkpoint factor in budding yeast (Rad9), the checkpoint concept largely preceded the identification of most key components mediating lesion detection and signaling. It was not even clear whether kinases were involved; in fact, the word "kinase" is absent from both the checkpoint papers of 1988 and 1989 (Weinert & Hartwell, 1988; Hartwell & Weinert, 1989). Identification of the apical PIKK kinases and downstream checkpoint kinases, as well as their involvement in mediating DNA damage responses, would not become evident until the mid-1990s. Decades later, our understanding of both the mechanisms of lesion detection and the signaling circuitry connecting lesion detection to regulation of biological processes has radically changed. The field has defined many of the key proteins required for lesion detection and uncovered an extensive network of kinase substrates. The concept that checkpoint is a simple mechanism coupling lesion detection to cell cycle arrest also dramatically changed to incorporate both a range of biological outputs and a network of functional and physical connections. Nonetheless, at its essence, the central concept that DNA damage checkpoints coordinate cell cycle with DNA repair continues to hold true.

A unified model for the coordination of checkpoint and repair in the DSB response

When Weinert and Hartwell first proposed the checkpoint concept, it would be considered improbable that the circuitry responsible for promoting cell cycle arrest also directly and actively controls the DNA repair machinery. However, over the last 30 years, accumulated knowledge from work in yeast and mammals points to a key role for the apical PIKKs and downstream checkpoint kinases in the spatiotemporal coordination of checkpoint responses and DNA repair transactions. The response to DSBs epitomizes such intricate coordination. Mounting evidence supports a model whereby early response to DSBs involves rapid activation of apical PIKK kinases and the establishment of an emergency response that inhibits improper nuclease action (and resection) to prevent improper processing of DNA ends. These early response actions include cell cycle arrest, general inhibition of origin firing (if in S phase), and extensive transcriptional re-programming. Following establishment of a robust checkpoint response and early resection block, cells must decide when to initiate DNA repair, including which repair pathway to utilize and when to resume the cell cycle. This decision is simpler in G1 cells, as the engagement of NHEJ for the repair of DSBs is highly preferred and can occur as a natural consequence of the resection block. However, for many DNA lesions in S phase and G2, the use of HR-mediated repair is the pathway of choice for error-free repair. Cells must, therefore, oppose the initial anti-resection effects of DNA damage signaling to initiate resection, the first step in HR. Given the importance of the decision to initiate (or not to initiate) resection for repair pathway choice, this step is under intense regulation by phosphorylation. In yeast, Mec1 promotes resection by phosphorylating Slx4 and likely Sae2, opposing its own anti-resection function via Rad53 activation. Interestingly, such dual and opposite functions for Mec1 could be key to providing a highly controllable system in which resection is spatiotemporally fine-tuned. This is particularly important, as the extent of resection impacts which HR-mediated mechanisms will follow [e.g., canonical HR, single-strand annealing, synthesis-dependent strand annealing (SDSA), and break-induced replication (BIR)]. In humans, current evidence points to ATR, ATM, and DNA-PKcs playing an important role in the regulation of resection, although much less is understood compared with yeast. In both yeast and humans, phosphatases are expected to play a key role in counter-balancing the phosphorylation events that impose the early resection block and influence the DNA repair process. The balance between DNA damage signaling kinases and the phosphatases that oppose them is better appreciated for the more canonical aspects of the checkpoint, but more work is needed to understand how this balance affects choices made during the DNA repair process.

In sum, apical kinases perform highly elaborate actions in the spatiotemporal control of checkpoint responses and DNA repair. Since checkpoint functions and early resection blockades can counteract HR-mediated DNA repair mechanisms, the ability of cells to modulate DNA damage signaling is essential for repair pathway control, and, in

particular, proper execution of HR-mediated repair. More work is needed to better understand how DNA damage signaling kinases monitor and control subsequent steps in HR (including strand invasion, homology search, and resolution/dissolution), as well as how such regulation ensures the fidelity of HR and prevents the erroneous recombination that can give rise to genetic instability.

What's next? Toward a holistic view of the DNA damage signaling network

In budding yeast and humans, DNA damage signaling kinases target dozens, if not hundreds, of DNA repair proteins, thereby modulating repair pathways. As described above, our understanding of the mechanisms by which these kinases coordinate DNA repair machineries is incomplete. When considering a function associated with the action of a kinase, the gold standard is to demonstrate that the function is impaired by mutation of a discrete set of phosphorylatable residues within a substrate protein(s). In almost all the cases discussed above, the DNA repair-related phenotypes associated with the disruption of kinase function are far stronger than the phenotypes associated with the introduction of phosphosite mutations on the kinase substrates. Thus, either the proper combination of phosphosite mutations has not yet been uncovered, or there are still many more undiscovered phosphorylation events mediated by the DNA damage signaling kinases that are critical for the control of DNA repair. In some cases, fundamental phenotypes associated with the loss of DNA damage signaling still lack defined substrates and underlying molecular mechanisms. For example, a mechanistic understanding of the genetic instability (mostly monitored as accumulation of GCR) observed upon loss of Mec1 and Tel1 is unknown and represents a significant gap in our understanding of DNA damage signaling in budding yeast. In humans, the lethality and chromosomal fragmentation observed upon inhibition of ATR also lack a clear mechanistic underpinning. Furthermore, while it is established that inhibition of ATM and DNA-PKcs strongly impair NHEJ, it is unclear which phosphorylation events are disrupted upon inhibition and in what ways they are necessary for NHEJ.

Mass spectrometry as a tool to study DNA damage signaling kinases: strengths, weaknesses, and future directions

Mass spectrometry has been instrumental for discovering *in vivo* substrates of the DNA damage signaling kinases (Matsuoka *et al*, 2007; Smolka *et al*, 2007; Bastos de Oliveira *et al*, 2015; Wagner *et al*, 2016; Lanz *et al*, 2018). The ability to quantitatively assess the phosphoproteome of an organism opened the door for unbiased identification of kinase substrates. Phosphoproteomic screens, performed in budding yeast and humans, have identified hundreds of phosphorylation events catalyzed by the DNA damage signaling kinases. Reassuringly, many of these phosphorylation events map to previously known substrate proteins. Excitingly, most of these events occur in substrate proteins not yet studied, many of which are associated with a broad range of

nuclear processes, including DNA repair. However, the scope of the DNA damage signaling network revealed by phosphoproteomics raises the question of how many phosphorylation events have tangible biological significance. The ability to distinguish functional phosphorylation from kinase promiscuity represents the primary challenge of large-scale phosphoproteomic datasets. For such an inquiry, it is imperative to generate mutant strains that either lack or constitutively mimic the phosphorylated residues in the substrate protein, with the ultimate goal to phenocopy the effects of the kinase's action or inaction. However, this is equivalent to finding a needle in a haystack, and better strategies are needed to enhance our ability to efficiently and systematically predict functional sites and, when necessary, to dissect functional redundancy and combinatorial effects.

Identification of *in vivo* kinase substrates using mass spectrometry is often performed using quantitative MS-based approaches, primary SILAC, which provide a highly quantitative comparison between the phosphoproteomic profiles of two different cell populations (Bastos de Oliveira *et al*, 2015, 2018). Budding yeast represents an ideal system for mapping DNA damage signaling using mass spectrometry due to its relative simplicity compared with the mammalian system. A range of genetic tools and mutants enable powerful genetic–proteomic strategies to deconstruct the action of apical and downstream kinases. For example, loss of Mec1/ATR results in lethality in most eukaryotic systems, but suppressor mutations that rescue the lethality of *mec1Δ* cells (such as *SML1* deletion) are available in yeast and are vital for the identification of proteins containing Mec1/Tel1- and Rad53-dependent phosphorylation using phosphoproteomics (Smolka *et al*, 2007; Bastos de Oliveira *et al*, 2015; Lanz *et al*, 2018).

Phosphoproteomics has also been used to identify targets of the DNA damage signaling kinases in mammalian cell lines (Matsuoka *et al*, 2007; Stokes *et al*, 2007; Bennetzen *et al*, 2010; Pines *et al*, 2011; Beli *et al*, 2012; Kirkpatrick *et al*, 2013; Wagner *et al*, 2016). The sheer complexity of the human phosphoproteome makes in-depth phosphoproteomic analyses more challenging than in yeast. Nonetheless, phosphoproteomic studies, such as the 2007 landmark paper by the Elledge group (Matsuoka *et al*, 2007), have revealed hundreds of potential substrates of DNA damage signaling kinases. One limitation of this work is that it relied on the use of phospho-specific antibodies for the SQ/TQ phosphosite motif, so it did not identify substrates of the downstream DDC kinases CHK1 and CHK2, nor did it identify potentially direct ATR or ATM substrates phosphorylated at non-canonical motifs (non-SQ/TQ). Unlike in yeast, matching a substrate to a specific DNA damage signaling kinase is still a challenge, although the recent development of specific chemical inhibitors of ATR, ATM, and CHK1 has allowed more efficient assignment of kinase dependency. In fact, recent studies have made use of these chemical inhibitors to find ATR-, ATM-, and CHK1-dependent phosphorylation events *in vivo* (Blasius *et al*, 2011; Wagner *et al*, 2016). Of note, the substrates of DNA-PKcs, a third, human-specific PI3 kinase-like kinase, have yet to be extensively profiled by MS, although low-throughput studies have uncovered substrate

proteins associated with NHEJ (DNA-PKcs is reviewed in detail here: Blackford& Jackson, 2017).

An alternative approach to screening for substrates of a kinase-of-interest involves the mutation of gate-keeping residues in the kinase's ATP-binding pocket, which enables the accommodation and subsequent utilization of bulky ATP analogs (Hertz *et al*, 2010). The addition of an analog-compatible kinase to human cell extracts results in the *in vitro* modification of its substrate proteins, which are then isolated and identified based on the covalent attachment of analog phosphate groups. This chemical-genetics approach (commonly referred to as the "Shokat method") has been used to map the signaling network of several CDKs (Blethrow *et al*, 2008; Chi *et al*, 2008) and, more recently, to identify CHK1 substrate proteins in human cell extracts (Blasius *et al*, 2011). However, these chemical-genetic strategies require the generation of mutant kinase alleles that are able to utilize bulky ATP analogs, and the kinase labeling reactions normally take place in cell lysates rather than *in vivo*. Also, efforts to apply the Shokat method to the upstream PIKKs have thus far been unsuccessful, as ATM and ATR appear to be less amenable to mutations in their ATP-binding pockets.

Overall, phosphoproteomic studies have revealed that the human DNA damage signaling network is extensive, consisting of many nuclear proteins, but also non-nuclear proteins. In addition to the established DNA repair-related substrates previously described in low-throughput studies, phosphoproteomic analyses have uncovered a large proportion of substrates involved in transcription and RNA processing, although, like in yeast, the functional significance of these phosphorylation events remains unclear. While the work done thus far in mammalian systems has yielded many candidate substrates, the DNA damage signaling kinases may operate differently in different cell types. Thus, it will be important for future investigations to thoroughly assess how the DNA damage signaling network varies in different cell types and pathological statuses. Attaining full coverage of the phosphoproteome is also a concern, as available phosphoproteome datasets do not cover all phosphorylation events in a cell due to technical limitations with mass spectrometry (Engholm-Keller & Larsen, 2013). Currently, however, instruments have become more sensitive, and the ability to use mass spectrometry to cover the full phosphoproteome may become a reality in the near future. Looking ahead, it will be crucial to develop more efficient approaches for identifying functional phosphorylation events within the large phosphoproteomic datasets. One possibility is to use structural prediction analysis to identify phosphorylation events that are likely to impair protein function (such as disrupting protein–protein or protein–DNA interactions) (preprint: Lanz *et al*, 2019). In conclusion, a holistic understanding of the role and action of DNA damage signaling kinases will require powerful technologies capable of quantitatively monitoring the full extent of the phosphoproteome in combination with systematic approaches for functional and mechanistic analyses.

Probable questions:

1. Discuss role of Adaptor proteins as substrates for downstream signaling activation.
2. Discuss Post-recruitment events in downstream checkpoint kinase activation.
3. Discuss core DNA damage signaling responses.
4. How cell cycle is controlled?
5. How dNTP production is regulated?
6. Discuss Control of DNA repair by DNA damage signaling kinases.
7. How ATM and DNA-PKcs control the resection?
8. Discuss Fanconi anemia pathway.
9. Discuss unified model for the coordination of checkpoint and repair in the DSB response.
10. How Mass spectrometry can be used to study DNA damage signaling kinases?

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

UNIT-X

Proteomics – Proteomes, expression analysis, Post-translational modification, 2D Gel Electrophoresis

Objective: In this unit we will discuss Proteomics, post translational modifications of protein and also about two-dimensional electrophoresis.

Basic Concepts of Proteomics:

The gene transcripts that an individual can make in a lifetime—termed as transcriptome (by analogy with the term genome)—refers to the haploid set of chromosomes carrying all the functional genes.

Similarly, all the proteins made by an organism are now grouped under the shade of proteomics. Proteomics involves the systematic study of proteins in order to provide a comprehensive view of the structure, function and role in the regulation of a biological system.

These include protein-protein interaction, protein modification, protein function and its localization studies. The aim of proteomics is not only to identify all the proteins in a cell but also to create a complete three-dimensional map of the cell indicating where proteins are located. Coupled with advances in bioinformatics, this approach to comprehensively describing biological systems will undoubtedly have a major impact on our understanding of the phenotype of both normal and diseased cells.

The proteome (term coined by Mark Wilkins in 1995) of a given cell is the total number of proteins at any given instant and it is highly dynamic in response to internal and external cues. Proteins can be modified by post-translational modifications, undergo translocations within the cell or be synthesized or degraded.

Therefore, the examination of proteins of a cell at a particular time reflects the immediate protein environment in which it is studied. A cellular proteome is the collection of proteins found in a particular cell type under the influence of a particular set of environmental conditions like exposure to hormone stimulation. A complete set of proteins from all of the various cellular proteomes will form an organism's complete proteome.

An interesting finding of the Human Genome Project is that there are far more proteins in the human proteome (~ 400,000 proteins) than there are protein-coding genes in the human genome (~ 22,000 genes). The large increase in protein diversity is thought to

be due to alternative splicing and post-translational modification of proteins. This indicates that protein diversity cannot be fully characterized by gene expression analysis alone. Proteomics, thus is a useful tool for characterizing cells and tissues of interest.

The first protein studies that can be called proteomics began with the introduction of two dimensional gel electrophoresis of *E. coli* proteins (O'Ferrall, 1975) followed by mouse and guinea pig protein studies (Ksole, 1975). Although 2-dimensional electrophoresis (2-DE) was a major step forward and many proteins could be separated and visualized by this technique but it was not enough for the protein identification through any sensitive protein sequencing technology.

After certain efforts the first major technology for the identification of protein was protein sequencing by Edman degradation (Edman, 1949). This technology was used for the identification of proteins from 2-D gels to create first 2D database (Celis et al. 1987). Another most important development in protein identification was Mass Spectrometry (MS) technology (Andersen et al. 2000). Protein sequencing by MS technology has been increased due to its sensitivity of analysis, tolerate protein complexes and amenable to high throughput operations.

Although several advancements have been made in protein identification (by MS or Edman sequencing) without having the database of large scale DNA sequencing of expressed sequences and genomic DNA, proteins could not be characterized because different protein isoforms can be generated from a single gene through several modifications (Fig. 18.1). And the majority of DNA and protein sequences have been accumulated within a short period of time.

In 1995, the sequencing of the genome of an organism was done for the first time in Haemophilus influenzae (Fleischmann et al. 1995). Till date, sequencing of several other eukaryotic genomes have been completed viz. Arabidopsis thaliana (Tabata, 2000), Sachcharomyces cerevisiae (Goffeau, 1996), Caenorhabditis elegans (Abbott, 1998), Oryza (Matsumoto, 2001) and human (Venter, 2001).

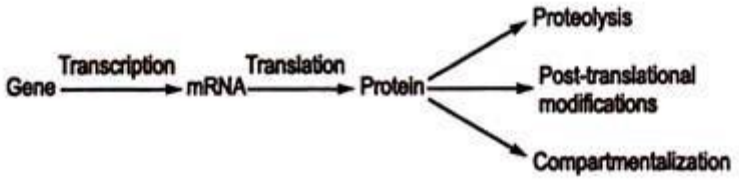


Fig. 18·1 : *Diagrammatic representation of a gene expression showing formation of many protein isoforms from a single gene. After transcription of the gene, mRNA is alternatively spliced or edited to form a mature mRNA that is translated to the protein. Proteins can be regulated by additional mechanism of proteolysis, compartmentalization and certain other modifications*

For protein expression profiling, a common procedure is the analysis of mRNA by different methods including serial analysis of gene expression (SAGE) (Velculescu et al. 1995) and DNA microarray technology (Shalon, 1996). However, the level of transcription of a gene gives only a rough idea of the real level of expression of that gene.

An mRNA may be produced in abundance, but at the same time degraded rapidly, or translated inefficiently keeping the amount of protein minimum. Proteins having been formed are subjected to post-translational modifications also. Different post-translational modifications or proteolysis and compartmentalization regulate the protein functions in the cell (Fig. 18.1).

The average number of proteins formed per gene was predicted to be one or two in bacterium, three in yeast and three or more in humans (Wilkins et al. 1996). In response to extra-cellular responses, a number of proteins undergo post-translational modifications. Protein phosphorylation is an important signaling mechanism and dis-regulation of protein kinase and phosphatase can result oncogenesis (Hunter, 1995).

Through proteome analysis, changes in the modifications of many proteins expressed by a cell can be analyzed after translation. Another important feature of a protein is its localization in the cell. The mis-localization of proteins is known to have an adverse effect on cellular function (cystic fibrosis) (Drumm and Collins, 1993). The cell growth, programmed cell death and the decision to proceed through the cell cycle are all regulated by signal transduction through protein complexes (Pippin et al. 1993). The protein interaction can be detected by using yeast two-hybrid system (Rain et al. 2001).

Types of Proteomics:

i. Structural Proteomics:

One of the main targets of proteomics investigation is to map the structure of protein complexes or the proteins present in a specific cellular organelle known as cell map or structural proteins. Structural proteomics attempt to identify all the proteins within a protein complex and characterization all protein-protein interactions. Isolation of specific protein complex by purification can simplify the proteomic analysis.

ii. Functional Proteomics:

It mainly includes isolation of protein complexes or the use of protein ligands to isolate specific types of proteins. It allows selected groups of proteins to be studied its chracteristics which can provide important information about protein signalling and disease mechanism etc.

To Understand a Proteome, Three Distinct Type of Analysis must be Carried Out:

(1) Protein-expression proteomics is the quantitative study of the protein expression of the entire proteome or sub-proteome of two samples that differ by some variable. Identification of novel proteins in signal transduction and disease specific proteins are major outcome of this approach.

(2) Structural proteomics attempts to identify all the proteins within a complex or organelle, determine their localization, and characterize all protein-protein interactions. The major goal of these studies is to map out the structure of protein complexes or cellular organelle proteins (Blackstock and Weir, 1999).

(3) Functional proteomics allows the study of a selected group of proteins responsible in signaling pathways, diseases and protein-protein interactions. This may be possible by isolating the specific sub-proteomes by affinity-chromatography for further analysis (Fig. 18.2):

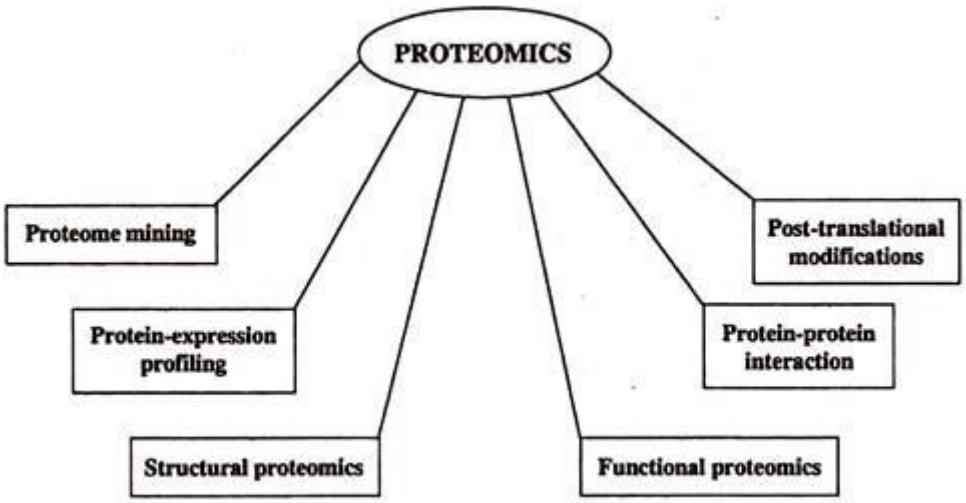


Fig. 18·2 : *Understanding the applications of proteomics (Graves and Haystead, 2002)*

Technology of Proteomics:

Measurement of the level of a gene transcript does not necessarily give clear picture of protein products formed. Therefore, for the measurement of real gene expression, the proteins should be analyzed. Before the identification and measurement of the activity, all the proteins in a proteome for any instant should be separated from each other.

A Typical Proteomics Experiment (e. g. Protein Expression Profiling) can be Divided into the following Categories:

(i) Separation and isolation of protein

(ii) The acquisition of protein structural information for protein identification and characterization

(iii) Database utilization.

(i) Protein Separation and Isolation:

An essential component of proteomics is the protein electrophoresis, the most effective way to resolve a complex mixture of proteins. Two types of electrophoresis are available as one and two-dimensional electrophoresis. In one dimensional gel electrophoresis (1-DE), proteins are resolved on the basis of their molecular masses. Proteins are stable enough during 1-DE due to their solubility in sodium dodycyl sulphate (SDS). Proteins with molecular mass of 10-300 kDa can be easily separated through 1-DE.

But with complex protein mixtures, results with 1-DE are limited, so for more complex protein mixture such as crude cell lysate, the best separation tool available is two dimensional gel electrophoresis (2-DE) (O'Ferrall, 1975). Here, proteins are separated according to their net charges in first dimension and according to their molecular masses in second dimension.

As a single 2-DE gel can resolve thousands of proteins, it remains a powerful tool for the cataloging of proteins. Two-dimensional electrophoresis has the ability to resolve proteins that have gone under some post-translational modifications as well as protein expression of any two samples can be compared quantitatively and qualitatively. Recently pH gradients have been introduced to 2-DE which greatly improved the reproducibility of this technique (Bjellqvist et al. 1993).

However, few problems with 2-DE still remain to be solved. Despite efforts to automate protein analysis by 2-DE, it is still a labour-intensive and time-consuming process. Another major limitation of 2-DE is the inability to detect low copy number proteins when a total cell lysate is analyzed (Link et al. 1997; Shevchenko et al. 1996) as well as inefficiency to speed up the in-gel digestion process also.

Therefore, alternatives have been searched to bypass protein gel electrophoresis. One approach is proteolytic digestion of protein mixture to convert them into peptides and then purify the peptides before subjecting them to analysis by mass spectrometry (MS). Peptide purification has been simplified through liquid chromatography (Link et al. 1999; McCormack et al. 1997), capillary electrophoresis (Figeys et al. 1999; Tong et al. 1999) and reverse phase chromatography (Opiteck et al. 1997).

Recently, Juan et al. (2005) have developed a new approach to speed up the protein identification process utilizing 'microwave' technology. Proteins excised from the gels are subjected to trypsin digestion by microwave irradiation, which rapidly produces peptides fragments. These fragments could be analyzed by MALDI (Matrix Assisted Laser Desorption/Ionization). Despite much downstream research on certain alternatives to 2-DE, this is the most widely utilized technique for proteome studies.

(ii) Acquisition of Protein Structures: Protein Identification:

Edman Sequencing (ES):
One of the earliest methods used for protein identification was micro sequencing by Edman chemistry to obtain N-terminal amino acid sequences. This technique was introduced by Edman in 1949. In Edman sequencing, N-terminal of a protein is sequenced to determine its true start site. Edman sequencing is more applicable sequencing method for the identification of proteins separated by SDS-Polyacrylamide gel electrophoresis.

This method has been used extensively in the starting years of proteomics but certain limitations have emerged in recent time. One of the major limitations is the N-terminal modification of proteins. If any protein is blocked on N-terminal before sequencing, then it is very difficult to identify the protein.

To overcome this problem a novel approach of mixed peptide sequencing (Damer et al. 1998) has been employed recently. In this approach, a protein is converted into peptides by cleavage with cyanogen bromide (CNBr) or skatole followed by the Edman sequencing of peptides.

Mass Spectrometry (MS):

The most significant breakthrough in proteomics has been the mass spectrometric identification of gel-separated proteins. Due to its high sensitivity levels, identification of proteins in protein complexes/mixtures and high throughput, this technique has been proved far better than ES.

In mass spectrometry, proteins are digested into peptides in the gel itself by suitable protease such as trypsin, because proteins, as such, are difficult to elute out from the gels. Moreover, molecular weight of proteins is not usually suitable for database

identification. In contrast, peptides can be eluted from the gels easily and matching of even a small set of peptides to the database is quite sufficient to identify a protein.

There are Two Main Approaches to Mass Spectrometric Protein Identification:

(i) "Electrospray ionization" (ESI) involves the fragmentation of individual peptides followed by direct ionization through electrospray in a tandem mass spectrometer. In ESI, a liquid sample flows from a microcapillary tube into the orifice of the mass spectrometer, where a potential difference between the capillary and the inlet to the mass spectrometer results in the generation of a fine mist of charged droplets (Fenn et al. 1989; Hunt et al. 1981).

It has the ability to resolve peptides in a mixture, isolate one species at a time and dissociate it into amino or carboxy-terminal containing fragments designated 'b' and 'y', respectively.

(ii) In "Peptide mass mapping" approach (Henzel et al. 1993) the mass spectrum of the eluted peptide mixture is acquired, which result in a peptide mass fingerprint of the protein being studied. The mass spectrum is obtained by a relatively simple 'mass spectrometric method-matrix assisted laser desorption/ ionization' (MALDI).

In this approach, tryptic peptide mixture is analyzed because trypsin cleaves proteins at the amino acid arginine and lysine. As the tryptic peptides can be predicted theoretically for any protein, the predicted peptide masses can be compared with those obtained experimentally by MALDI analysis. If the sufficient number of peptide matches with the existing protein sequence in database, the accuracy for protein identification is high.

After the protease cleavages of the proteins, they are analyzed by mass analysis also. Mass analysis follows the conversion of proteins or peptides into molecular ions. These ions got separated in a mass spectrometer based on their mass/charge (m/z) ratio. It is determined by the time it takes for the ions to reach the detector. Hence the instrument is called a time of flight (TOF) instrument.

The relationship that allows the m/z ratio to be determined is $E = 1/2\ (m/z)v^2$. In this equation. E is the energy imparted on the charged ions as a result of the voltage that is applied by the instrument and V is the velocity of the ions down the flight path. As peptide ions are introduced into the collision chamber, they interact with collision gas and undergo fragmentation along the peptide backbone (Fig. 18.4).
Because all the ions are exposed to the same electric field, all similarly charged ions will have similar energies. Therefore, based on the above equation, ions that have larger mass must have lower velocities and hence will require longer times to reach the detector. Different steps involved in mass spectrometry are described in a flow chart in Fig. 18.3.

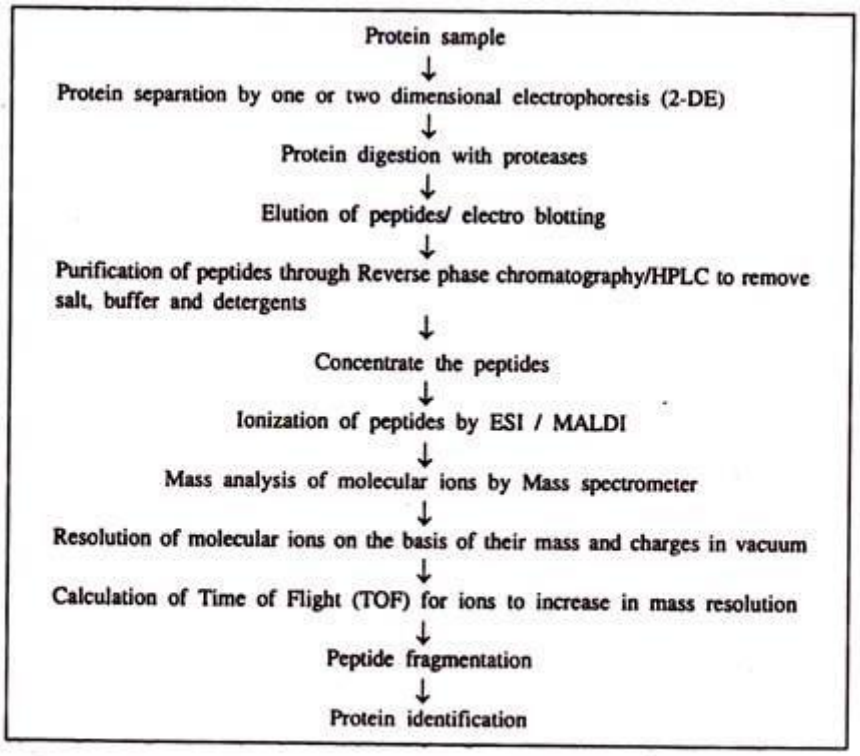


Fig. 18·3 : *A schematic representation of protein identification through Mass spectrometry. All the proteins present in the protein muixture of a cell lysate are identified with this method.*

(iii) Database Utilization:

Initially, sequencing of some proteins or peptides followed by the submission of sequences together created an assembly of proteins called protein database. Proteolytic digestion of many proteins are also predicted theoretically and deposited in database. Hence, at present, so much information has been accumulated that we can search for a homology between a new peptide sequence and the existing sequences in the database to identify the protein.

The major goal of database searching is to identify a large number of proteins—quickly and accurately. All the information accumulated through Edman sequencing or mass spectrometry are used to identify the proteins. In peptide mass fingerprinting database searching, the mass of a unknown peptide after proteolytic digestion is compared to the perdicted mass of peptide from theoretical digestion of proteins in database. In amino acid sequence database searching, the sequence of amino acids from a peptide is identified and can be used to search databases to find the protein from which it was derived.

Collection of protein sequence databases are thus designed to represent a partial list of an organism's genome, that is, the genes and all of the proteins they encode. The protein families are usually classified according to their evolutionary history inferred from sequence homology.

These databases are excellent tools for gene discovery, comparative genomics and molecular evolution. The purpose of database similarity searching is the sensitive detection of sequence homologues, regardless of the species relationship in order to infer similarity of function from similarity of sequence.

Recently, Chromatography-based proteomics is used to measure the concentration of low molecular weight peptides in complex mixtures such as plasma or sera. These technologies use time-of-flight (TOF) spectroscopy with matrix-assisted or surface-enhanced laser desorption/ionization to produce a spectrum of mass-to-charge (m/z) ratios that can be analysed in order to identify unique signatures from its chromatography pattern.

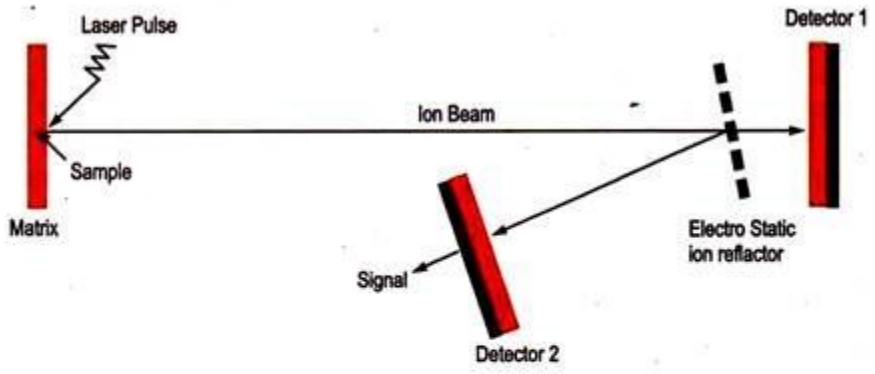


Fig. 18-4 : *Principle behind MALDI-TOF mass spectrometry. A sample is placed on the matrix and ionize by the laser beam. Due to the potential developed between the matrix and the sample, ions start moving towards the detector and get reflected by a reflector in the mid-way. Again after a flight in the tube the ions are detected by another detector. The time taken by these ions in the flight tubes depends on their masses. Therefore, we can calculate the ratio between the mass of an ion and the time of flight in the tube taken by that particular ion*

Applications of Proteomics:

1. Post-Translational Modifications:

Proteomics studies involve certain unique features as the ability to analyze post-translational modifications of proteins. These modifications can be phosphorylation, glycosylation and sulphation as well as some other modifications involved in the maintenance of the structure of a protein.

These modifications are very important for the activity, solubility and localization of proteins in the cell. Determination of protein modification is much more difficult rather than the identification of proteins. As for identification purpose, only few peptides are required for protease cleavages followed by database alignment of a known sequence of a peptide. But for determination of modification in a protein, much more material is

needed as all the peptides do not have the expected molecular mass need to be analyzed further.

For example, during protein phosphorylation events, phosphopeptides are 80 Da heavier than their unmodified counterparts. Therefore, it gives, rise to a specific fragment (PO^{3-} mass 79) bind to metal resins, get recognized by specific antibodies and later phosphate group can be removed by phosphatases (Clauser et al. 1999; Colledge and Scott, 1999). So protein of interest (post-translationally modified protein) can be detected by Western blotting with the help of antibodies or ^{32}P-labelling that recognize only the active state of molecules. Later, these spots can be identified by mass spectrometry.

2. Protein-Protein Interactions:

The major attribution of proteomics towards the development of protein interactions map of a cell is of immense value to understand the biology of a cell. The knowledge about the time of expression of a particular protein, its level of expression, and, finally, its interaction with another protein to form an intermediate for the performance of a specific biological function is currently available.

These intermediates can be exploited for therapeutic purposes also. An attractive way to study the protein-protein interactions is to purify the entire multi-protein complex by affinity based methods using GST-fusion proteins, antibodies, peptides etc.

The yeast two-hybrid system has emerged as a powerful tool to study protein-protein interactions (Haynes and Yates, 2000). According to Pandey and Mann (2000) it is a genetic method based on the modular structure of transcription factors in the close proximity of DNA binding domain to the activation domain induces increased transcription of a set of genes.

The yeast hybrid system uses ORFs fused to the DNA binding or activation domain of GAL4 such that increased transcription of a reporter gene results when the proteins encoded by two ORFs interact in the nucleus of the yeast cell. One of the main consequences of this is that once a positive interaction is detected, simply sequencing the relevant clones identifies the ORF. For this reason it is a generic method that is simple and amenable to high throughput screening of protein-protein interactions.

Phage display is a method where bacteriophage particles are made to express either a peptide or protein of interest fused to a capsid or coat protein. It can be used to screen for peptide epitopes, peptide ligands, enzyme substrate or single chain antibody fragments.

Another important method to detect protein-protein interactions involves the use of fluorescence resonance energy transfer (FRET) between fluorescent tags on interacting

proteins. FRET is a non-radioactive process whereby energy from an excited donor fluorophore is transferred to an acceptor fluorophore. After excitation of the first fluorophore, FRET is detected either by emission from the second fluorophore using appropriate filters or by alteration of the fluorescence lifetime of the donor.

A proteomics strategy of increasing importance involves the localization of proteins in cells as a necessary first step towards understanding protein function in complex cellular networks. The discovery of GFP (green fluorescent protein) and the development of its spectral variants has opened the door to analysis of proteins in living cells by use of the light microscope.

Large-scale approaches of localizing GFP-tagged proteins in cells have been performed in the genetically amenable yeast S. pombe (Ding et al. 2000) and in Drosophila (Morin et al. 2001). To localize proteins in mammalian cells, a strategy was developed that enables the systematic GFP tagging of ORFs from novel full-length cDNAs that are identified in genome projects.

3. Protein Expression Profiling:

The largest application of proteomics continues to be protein expression profiling. The expression levels of a protein sample could be measured by 2-DE or other novel technique such as isotope coded affinity tag (ICAT). Using these approaches the varying levels of expression of two different protein samples can also be analyzed.

This application of proteomics would be helpful in identifying the signaling mechanisms as well as disease specific proteins. With the help of 2-DE several proteins have been identified that are responsible for heart diseases and cancer (Celis et al. 1999). Proteomics helps in identifying the cancer cells from the non-cancerous cells due to the presence of differentially expressed proteins.

The technique of Isotope Coded Affinity Tag has developed new horizons in the field of proteomics. This involves the labeling of two different proteins from two different sources with two chemically identical reagents that differ in their masses due to isotope composition (Gygi et al. 1999). The biggest advantage of this technique is the elimination of protein quantitation by 2-DE. Therefore, high amount of protein sample can be used to enrich low abundance proteins.

Different methods have been used to probe genomic sets of proteins for biochemical activity. One method is called a biochemical genomics approach, which uses parallel biochemical analysis of a proteome comprised of pools of purified proteins in order to identify proteins and the corresponding ORFs responsible for a biochemical activity.

The second approach for analyzing genomic sets of proteins is the use of functional protein microarrays, in which individually purified proteins are separately spotted on a

surface such as a glass slide and then analyzed for activity. This approach has huge potential for rapid high-throughput analysis of proteomes and other large collections of proteins, and promises to transform the field of biochemical analysis.

4. Molecular Medicine:

With the help of the information available through clinical proteomics, several drugs have been designed. This aims to discover the proteins with medical relevance to identify a potential target for pharmaceutical development, a marker(s) for disease diagnosis or staging, and risk assessment—both for medical and environmental studies. Proteomic technologies will play an important role in drug discovery, diagnostics and molecular medicine because of the link between genes, proteins and disease.

As researchers study defective proteins that cause particular diseases, their findings will help develop new drugs that either alter the shape of a defective protein or mimic a missing one. Already, many of the best-selling drugs today either act by targeting proteins or are proteins themselves. Advances in proteomics may help scientists eventually create medications that are "personalized" for different individuals to be more effective and have fewer side effects. Current research is looking at protein families linked to disease including cancer, diabetes and heart disease.

Post Translational Modifications:

Protein post-translational modifications (PTMs) increase the functional diversity of the proteome by the covalent addition of functional groups or proteins, proteolytic cleavage of regulatory subunits, or degradation of entire proteins. These modifications include phosphorylation, glycosylation, ubiquitination, nitrosylation, methylation, acetylation, lipidation and proteolysis and influence almost all aspects of normal cell biology and pathogenesis. Therefore, identifying and understanding PTMs is critical in the study of cell biology and disease treatment and prevention.

Polypeptide chains like RNA transcripts are also modified after their synthesis. This additional processing is termed as post transcriptional modification.

These types of post translational modifications are important in achieving the functional status specific to any given protein. Because the final 3D structure of the molecule is closely elated to its specific function, the folding of protein is also important.

These complex biochemical processes are briefly described here for an understanding about overall process:

1. The N-terminus and C-terminus amino acids are usually removed or modified. The initial and terminal formylmethionine residue in bacterial polypeptide is usually removed enzymatically. In eukaryotes, initial methionine residue is removed and the amino group of the N-terminal residue is chemically modified.

2. Individual amino acid residues are sometimes modified, e.g., phosphate may be added to the hydroxyl groups of certain amino acids such as tyrosine. The process of phosphorylation is extremely important in regulating several cellular activities and is a result of the action of enzymes called kinases. In other proteins, methyl group may be added enzymatically.

3. In some proteins, carbohydrate side chains are sometimes added. Covalently added carbohydrates form a class of molecules called glycoproteins having antigenic properties.

4. Sometimes polypeptide chains are trimmed to make active protein molecules, e.g., insulin is produced as a large molecule and then trimmed to 51 amino acid molecules.

5. At the end terminal end of some proteins, a sequence of up to 30 amino acids is found that plays an important role in directing the protein to the location in the cell where it becomes functional. This is called a signal sequence and it determines the final destination of protein in the cell.

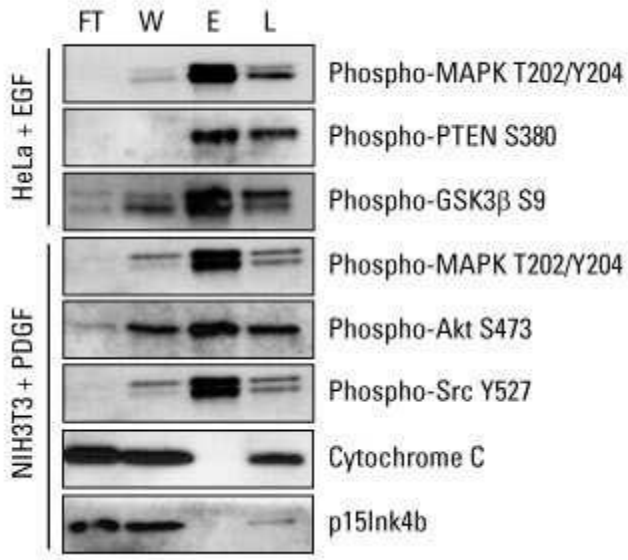
6. Polypeptide chains often complexes with metals. The quaternary and tertiary levels of protein structure often include and are dependent on metal atoms. The function of the protein is thus dependent on the molecular complex that includes both polypeptide chains and metal atoms, e.g., haemoglobin containing four iron atoms.

Post-translational modifications (PTMs)

As noted above, the large number of different PTMs precludes a thorough review of all possible protein modifications. Therefore, this overview only touches on a small number of the most common types of PTMs studied in protein research today. Furthermore, greater focus is placed on phosphorylation, glycosylation and ubiquitination, and therefore these PTMs are described in greater detail on pages dedicated to the respective PTM.

Phosphorylation

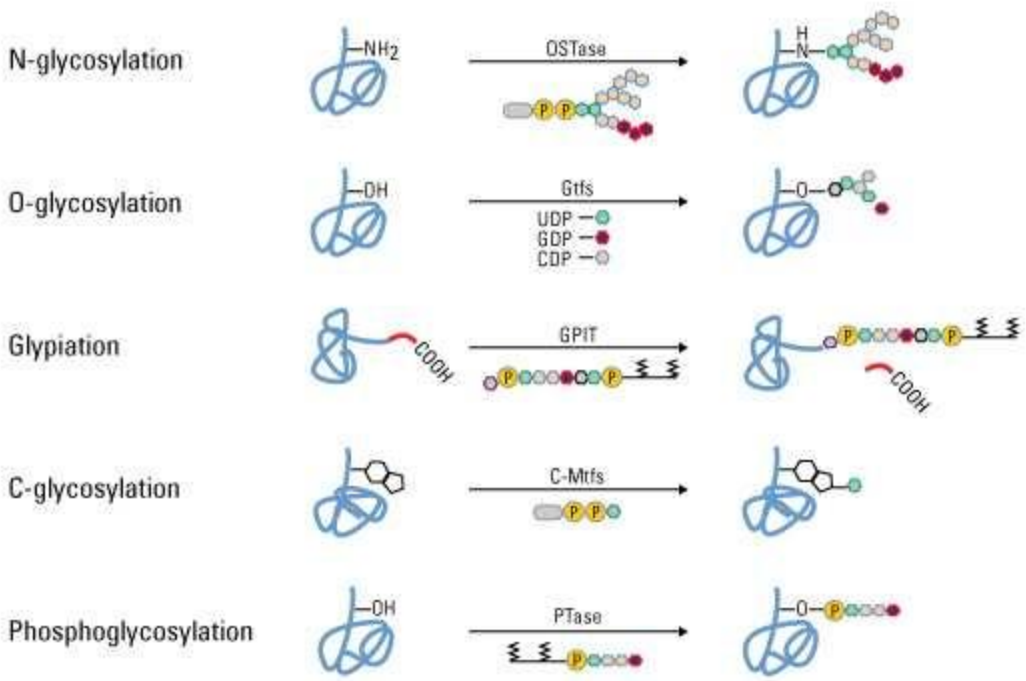
Reversible protein phosphorylation, principally on serine, threonine or tyrosine residues, is one of the most important and well-studied post-translational modifications. Phosphorylation plays critical roles in the regulation of many cellular processes, including cell cycle, growth, apoptosis and signal transduction pathways. In the following example, western blot analysis was used to evaluate phosphoprotein specificity in lysates obtained from serum-starved HeLa and NIH 3T3 cancer cell lines stimulated with epidermal growth factor (EGF) and platelet derived growth factor (PDGF), respectively.



Highly pure phosphoprotein enrichment from complex biological samples. Western blot analysis was performed with the Thermo Scientific Pierce Phosphoprotein Enrichment Kit, and cell lysates were prepared according to the kit instructions to enrich for phosphoproteins. Protein detection was achieved using phospho-specific antibodies that recognize key regulatory proteins involved in growth factor signaling. Cytochrome C (pI 9.6) and p15Ink4b (pI 5.5) served as negative controls for nonspecific binding of non-phosphorylated proteins. FT = flow-through fraction, W = pooled wash fractions, E = pooled elution fractions and L = non-enriched total cell extract.

Glycosylation

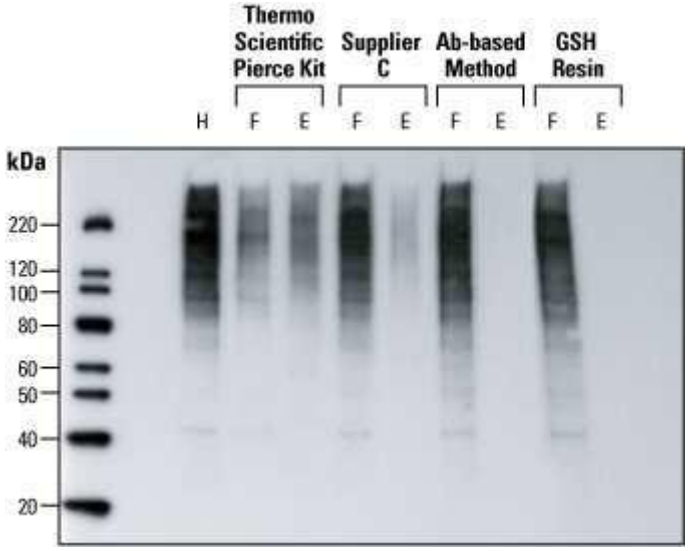
Protein glycosylation is acknowledged as one of the major post-translational modifications, with significant effects on protein folding, conformation, distribution, stability and activity. Glycosylation encompasses a diverse selection of sugar-moiety additions to proteins that ranges from simple monosaccharide modifications of nuclear transcription factors to highly complex branched polysaccharide changes of cell surface receptors. Carbohydrates in the form of asparagine-linked (N-linked) or serine/threonine-linked (O-linked) oligosaccharides are major structural components of many cell surface and secreted proteins.



Types of glycosylation. Glycopeptide bonds can be categorized into specific groups based on the nature of the sugar–peptide bond and the oligosaccharide attached, including N-, O- and C-linked glycosylation, glypiation and phosphoglycosylation.

Ubiquitination

Ubiquitin is an 8-kDa polypeptide consisting of 76 amino acids that is appended to the Îµ-NH2 of lysine in target proteins via the C-terminal glycine of ubiquitin. Following an initial monoubiquitination event, the formation of a ubiquitin polymer may occur, and polyubiquitinated proteins are then recognized by the 26S proteasome that catalyzes the degradation of the ubiquitinated protein and the recycling of ubiquitin. The following experiment provides an example of methods used to detect ubiquitinated proteins.



Detection of ubiquitin in HeLa cell lysates. Western blot analysis was performed to compare four methods for detecting ubiquitin protein in HeLa cell lysates. After epoxomicin-treatment, HeLa cells lysates (150 µg) were processed by four different methods. The resulting flow-through (F) and elution (E) fractions were volume-normalized to the original unprocessed lysate (H) and identical volumes electrophoresed for western blot detection. Compared to Supplier C's kit and an antibody-based method, the Thermo Scientific Pierce Ubiquitin Enrichment Kit yielded more ubiquitinated protein in the elution fraction (and less protein in the flow-through fraction), indicating significantly better enrichment of ubiquitinated proteins. GSH Resin is a negative control for comparison.

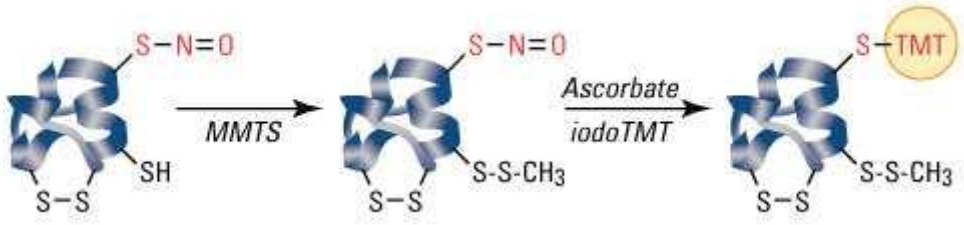
S-nitrosylation

Nitric oxide (NO) is produced by three isoforms of nitric oxide synthase (NOS), and it is a chemical messenger that reacts with free cysteine residues to form S-nitrothiols (SNOs). S-nitrosylation is a critical PTM used by cells to stabilize proteins, regulate gene expression and provide NO donors, and the generation, localization, activation and catabolism of SNOs are tightly regulated.

S-nitrosylation is a reversible reaction, and SNOs have a short half-life in the cytoplasm because of the host of reducing enzymes, including glutathione (GSH) and thioredoxin, that denitrosylate proteins. Therefore, SNOs are often stored in membranes, vesicles, the interstitial space and lipophilic protein folds to protect them from denitrosylation. For example, caspases, which mediate apoptosis, are stored in the mitochondrial intermembrane space as SNOs. In response to extra- or intracellular cues, the caspases are released into the cytoplasm, and the highly reducing environment rapidly denitrosylates the proteins, resulting in caspase activation and the induction of apoptosis.

S-nitrosylation is not a random event, and only specific cysteine residues are S-nitrosylated. Because proteins may contain multiple cysteines and due to the labile nature of SNOs, S-nitrosylated cysteines can be difficult to detect and distinguish from non-S-nitrosylated amino acids. The biotin switch assay, developed by Jaffrey et al., is a common method of detecting SNOs, and the steps of the assay are listed below:

- All free cysteines are blocked.
- All remaining cysteines (presumably only those that are denitrosylated) are denitrosylated.
- The now-free thiol groups are then biotinylated.
- Biotinylated proteins are detected by SDS-PAGE and western blot analysis or mass spectrometry.



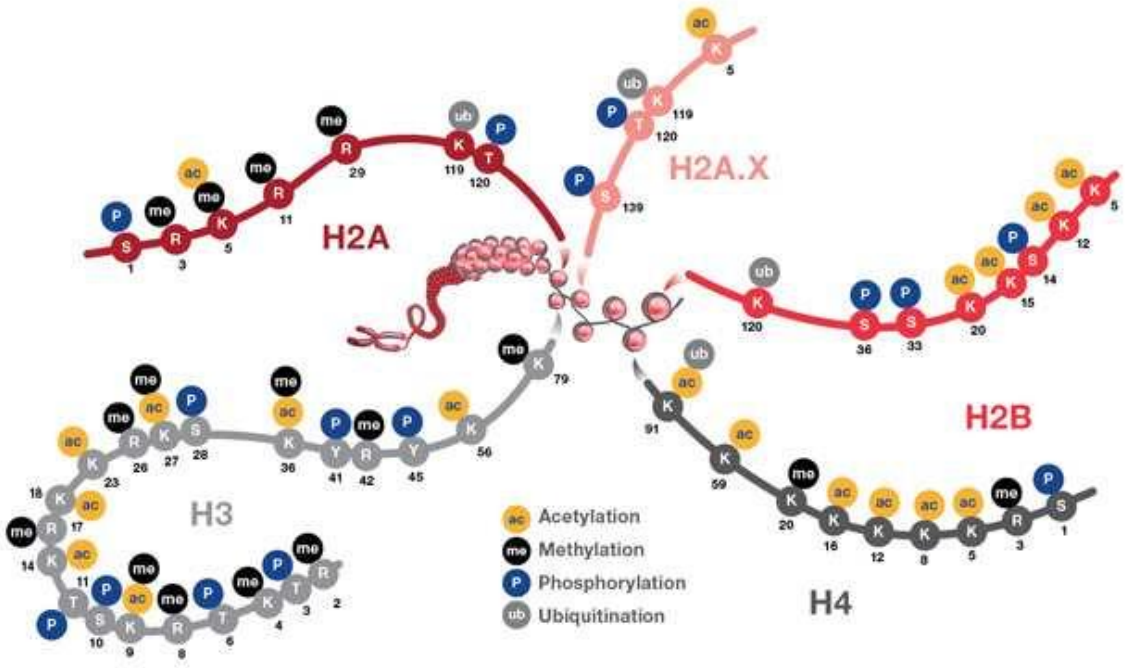
Reaction scheme for labeling and detection of S-nitrosylation with S-Nitrosylation Western Blot Kit. Samples are first reacted with MMTS to block free sulfhydryls in S-nitrosylated proteins. The S-nitrosocysteines are then selectively reduced with ascorbate for labeling with the Thermo Scientific iodoTMTzero Label Reagent. Subsequently, the supplied anti-TMT antibody is used to detect the TMT-labeled proteins in a western blot.

Methylation

The transfer of one-carbon methyl groups to nitrogen or oxygen (N- and O-methylation, respectively) to amino acid side chains increases the hydrophobicity of the protein and can neutralize a negative amino acid charge when bound to carboxylic acids. Methylation is mediated by methyltransferases, and S-adenosyl methionine (SAM) is the primary methyl group donor.

Methylation occurs so often that SAM has been suggested to be the most used substrate in enzymatic reactions after ATP. Additionally, while N-methylation is irreversible, O-methylation is potentially reversible. Methylation is a well-known mechanism of epigenetic regulation, as histone methylation and demethylation influences the availability of DNA for transcription. Amino acid residues can be conjugated to a single methyl group or multiple methyl groups to increase the effects of modification.

The figure below provides an illustration of PMTs associated with nucleosome core particles.



Representation showing post-translational modifications associated with histone particles. Nucleosomes are represented by red spheres wrapped by DNA (shown in gray). Also depicted are the positions of PTMs located on the histone proteins H2A (and H2A.X), H2B, H3, and H4. These PTMs impact gene expression by altering chromatin structure and recruiting histone modifiers. PTM events mediate diverse biological functions such as transcriptional activation and inactivation, chromosome packaging, and DNA damage and repair processes.

N-acetylation

N-acetylation, or the transfer of an acetyl group to nitrogen, occurs in almost all eukaryotic proteins through both irreversible and reversible mechanisms. N-terminal acetylation requires the cleavage of the N-terminal methionine by methionine aminopeptidase (MAP) before replacing the amino acid with an acetyl group from acetyl-CoA by N-acetyltransferase (NAT) enzymes. This type of acetylation is co-translational, in that N-terminus is acetylated on growing polypeptide chains that are still attached to the ribosome. While 80 to 90% of eukaryotic proteins are acetylated in this manner, the exact biological significance is still unclear.

Acetylation at the ε-NH2 of lysine (termed lysine acetylation) on histone N-termini is a common method of regulating gene transcription. Histone acetylation is a reversible event that reduces chromosomal condensation to promote transcription, and the

acetylation of these lysine residues is regulated by transcription factors that contain histone acetyltransferase (HAT) activity. While transcription factors with HAT activity act as transcription co-activators, histone deacetylase (HDAC) enzymes are co-repressors that reverse the effects of acetylation by reducing the level of lysine acetylation and increasing chromosomal condensation.

Sirtuins (silent information regulator) are a group of NAD-dependent deacetylases that target histones. As their name implies, they maintain gene silencing by hypoacetylating histones and have been reported to aid in maintaining genomic stability.

While acetylation was first detected in histones, cytoplasmic proteins have been reported to also be acetylated, and therefore acetylation seems to play a greater role in cell biology than simply transcriptional regulation. Furthermore, crosstalk between acetylation and other post-translational modifications, including phosphorylation, ubiquitination and methylation, can modify the biological function of the acetylated protein.

Protein acetylation can be detected by chromatin immunoprecipitation (ChIP) using acetyllysine-specific antibodies or by mass spectrometry, where an increase in histone by 42 mass units represents a single acetylation.

Lipidation

Lipidation is a method to target proteins to membranes in organelles (endoplasmic reticulum [ER], Golgi apparatus, mitochondria), vesicles (endosomes, lysosomes) and the plasma membrane. The four types of lipidation are:

- C-terminal glycosyl phosphatidylinositol (GPI) anchor
- N-terminal myristoylation
- S-myristoylation
- S-prenylation

Each type of modification gives proteins distinct membrane affinities, although all types of lipidation increase the hydrophobicity of a protein and thus its affinity for membranes. The different types of lipidation are also not mutually exclusive, in that two or more lipids can be attached to a given protein.

GPI anchors tether cell surface proteins to the plasma membrane. These hydrophobic moieties are prepared in the ER, where they are then added to the nascent protein en bloc. GPI-anchored proteins are often localized to cholesterol- and sphingolipid-rich lipid rafts, which act as signaling platforms on the plasma membrane. This type of modification is reversible, as the GPI anchor can be released from the protein by phosphoinositol-specific phospholipase C. Indeed, this lipase is used in the detection of

GPI-anchored proteins to release GPI-anchored proteins from membranes for gel separation and analysis by mass spectrometry.

N-myristoylation is a method to give proteins a hydrophobic handle for membrane localization. The myristoyl group is a 14-carbon saturated fatty acid (C14), which gives the protein sufficient hydrophobicity and affinity for membranes, but not enough to permanently anchor the protein in the membrane. N-myristoylation can therefore act as a conformational localization switch in which protein conformational changes influence the availability of the handle for membrane attachment. Because of this conditional localization, signal proteins that selectively localize to membrane, such as Src-family kinases, are N-myristoylated.

N-myristoylation is facilitated specifically by N-myristoyltransferase (NMT) and uses myristoyl-CoA as the substrate to attach the myristoyl group to the N-terminal glycine. Because methionine is the N-terminal amino acid of all eukaryotic proteins, this PTM requires methionine cleavage by the above-mentioned MAP prior to addition of the myristoyl group; this represents one example of multiple PTMs on a single protein.

S-palmitoylation adds a C16 palmitoyl group from palmitoyl-CoA to the thiolate side chain of cysteine residues via palmitoyl acyltransferases (PATs). Because of the longer hydrophobic group, this anchor can permanently anchor the protein to the membrane. This localization can be reversed, though, by thioesterases that break the link between the protein and the anchor; thus, S-palmitoylation is used as an on/off switch to regulate membrane localization. S-palmitoylation is often used to strengthen other types of lipidation, such as myristoylation or farnesylation (see below). S-palmitoylated proteins also selectively concentrate at lipid rafts.

S-prenylation covalently adds a farnesyl (C15) or geranylgeranyl (C20) group to specific cysteine residues within five amino acids from the C-terminus via farnesyl transferase (FT) or geranylgeranyl transferases (GGT I and II). Unlike S-palmitoylation, S-prenylation is hydrolytically stable. Approximately 2% of all proteins are prenylated, including all members of the Ras superfamily. This group of molecular switches is farnesylated, geranylgeranylated or a combination of both. Additionally, these proteins have specific 4-amino acid motifs at the C-terminus that determine the type of prenylation at single or dual cysteines. Prenylation occurs in the ER and is often part of a stepwise process of PTMs that is followed by proteolytic cleavage by Rce1 and methylation by isoprenyl cysteine methyltransferase (ICMT).

Proteolysis

Peptide bonds are indefinitely stable under physiological conditions, and therefore cells require some mechanism to break these bonds. Proteases comprise a family of enzymes

that cleave the peptide bonds of proteins and are critical in antigen processing, apoptosis, surface protein shedding and cell signaling.

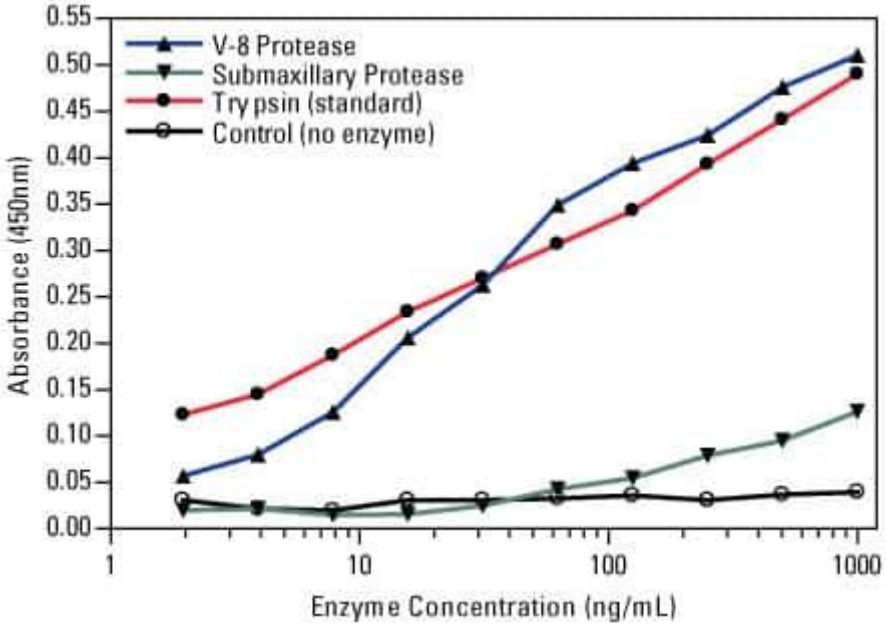
The family of over 11,000 proteases varies in substrate specificity, mechanism of peptide cleavage, location in the cell and the length of activity. While this variation suggests a wide array of functionalities, proteases can generally be separated into groups based on the type of proteolysis. Degradative proteolysis is critical to remove unassembled protein subunits and misfolded proteins and to maintain protein concentrations at homeostatic concentrations by reducing a given protein to the level of small peptides and single amino acids. Proteases also play a biosynthetic role in cell biology that includes cleaving signal peptides from nascent proteins and activating zymogens, which are inactive enzyme precursors that require cleavage at specific sites for enzyme function. In this respect, proteases act as molecular switches to regulate enzyme activity.

Proteolysis is a thermodynamically favorable and irreversible reaction. Therefore, protease activity is tightly regulated to avoid uncontrolled proteolysis through temporal and/or spatial control mechanisms including regulation by cleavage in cis or trans and compartmentalization (e.g., proteasomes, lysosomes).

The diverse family of proteases can be classified by the site of action, such as aminopeptidases and carboxypeptidase, which cleave at the amino or carboxy terminus of a protein, respectively. Another type of classification is based on the active site groups of a given protease that are involved in proteolysis. Based on this classification strategy, greater than 90% of known proteases fall into one of four categories as follows:

- Serine proteases
- Cysteine proteases
- Aspartic acid proteases
- Zinc metalloproteases

The following representative example demonstrates the performance of a commercially available protease assay.



Colorimetric protease assay response curves. The Thermo Scientific Pierce Colorimetric Protease Assay Kit was used to measure the activity of V-8 protease and submaxillary protease for digestion of casein substrate by comparison to the supplied trypsin standard.

Two Dimensional Gel Electrophoresis:

Two-dimensional gel electrophoresis (2-DE) is considered a powerful tool for proteomics work. It is used for separation and fractionation of complex protein mixtures from biological samples. 2-DE separates proteins depending on two different steps: the first one is called isoelectric focusing (IEF) which separates proteins according to isoelectric points (pI); the second step is SDS-polyacrylamide gel electrophoresis (SDS-PAGE) which separates proteins based on the molecular weights(relative molecular weight, Mr). Thus, thousands of proteins can be separated, and the information about IEF and molecular weights can be obtained. 2-DE is a widely used method for protein analysis with the ability to separate thousands of proteins at one time. It can provide direct visual information of changes in protein/post-translational modifications (PTMs) abundance. It can also be used to other analysis, such as whole proteome analysis, detection of biomarkers, drug discovery, and so on. There are several steps for successful 2-DE analysis.

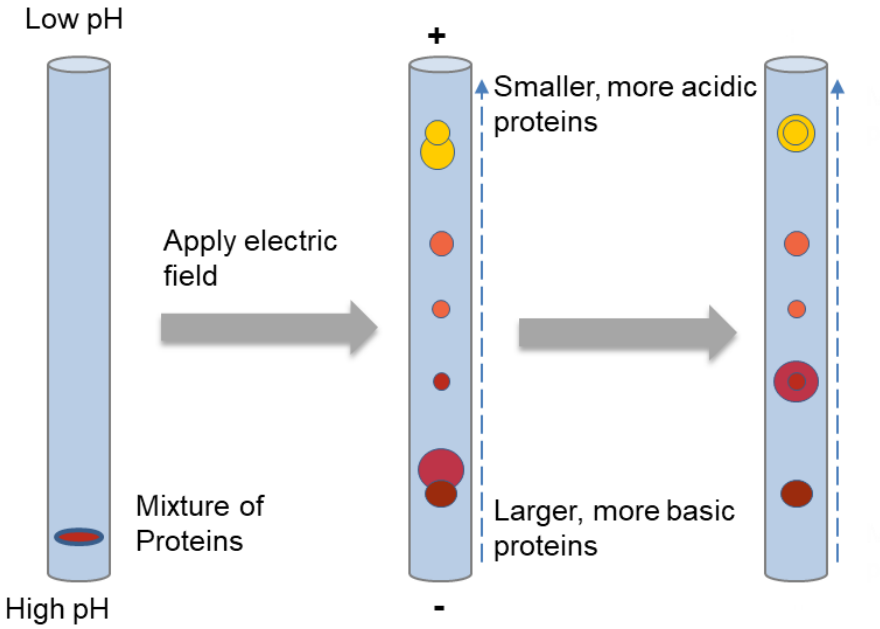
Sample preparation

Speaking of sample preparation, the native samples need to be converted to a physicochemical state suitable for the first dimension IEF and keep the native charge and Mr of the constituent proteins. In order to get good results, appropriate sample

preparation is essential. The sample preparation is various because of the difference of the types and origins of proteins. Ideally, the process will result in the complete solubilization, disaggregation, denaturation, and reduction of the proteins in the sample.

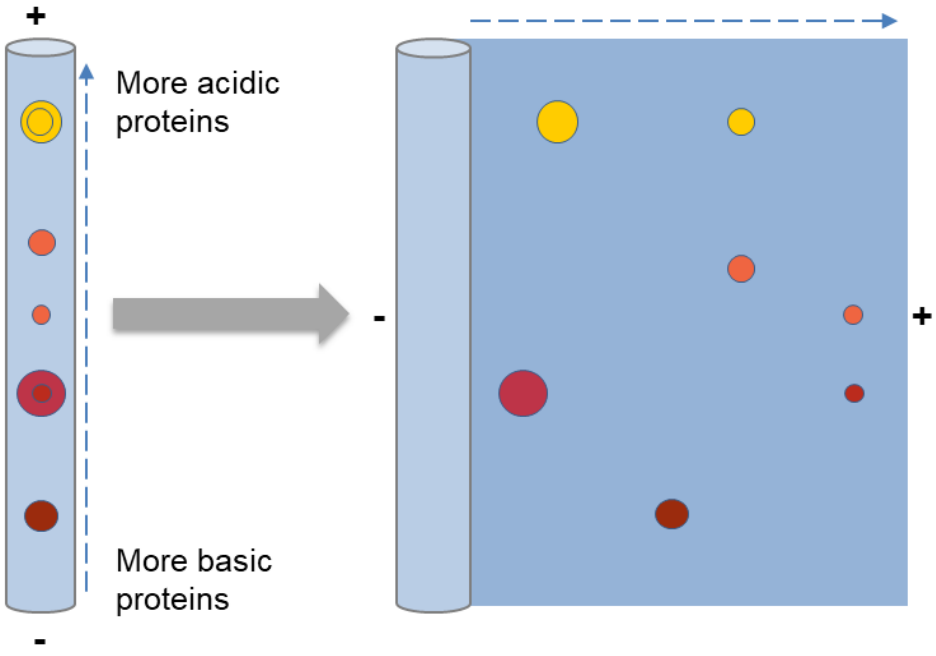
First dimension: Isoelectric Focusing (IEF)

As we mentioned above, the first dimension separate proteins depending on pI of proteins. Proteins are amphoteric molecules and the positive, negative, or zero net charge they carry depending on the pH of the surroundings. The isoelectric point (pI) is defined as the pH of a solution at which the net charge of the protein becomes zero. A protein with a positive net charge will migrate toward the cathode, becoming less positively charged until reaching its pI. While a protein with a negative net charge will migrate toward the anode, becoming less negatively charged until it also reaches its pI.



To be specific, a protein mixture is loaded at the basic end of the pH gradient gel. After applying an electric field, the proteins are separated depending on charges, focusing at positions where the pl value is equivalent to the surrounding pH. Larger proteins will move more slowly through the gel, but with sufficient time will catch up with small proteins of equal charge.

Second dimension: SDS-PAGE



After the first dimension, the second dimension separation can be performed on flatbed or vertical systems on a slab gel. The second dimension is often performed by SDS-PAGE (SDS-polyacrylamide gel electrophoresis), which is an electrophoretic method for separating polypeptides according to their molecular weights (Mr). This method often contains four steps, including preparation of the gel, the equilibrium of the immobilized pH gradient (IPG) strips in SDS buffer, placing the equilibrated IPG strip on the SDS gel, and finally handling the electrophoresis.

SDS can make proteins denaturing and bind to the backbone at a constant molar ratio. When applying SDS and a reducing agent, like a DTT which can cleave disulfide bonds, proteins unfold into linear chains with negative charge proportional to the polypeptide chain length. Polyacrylamide forms a mesh-like matrix which is appropriate for separating proteins. When proteins are separated by SDS-PAGE, smaller proteins migrate faster since the less resistance.

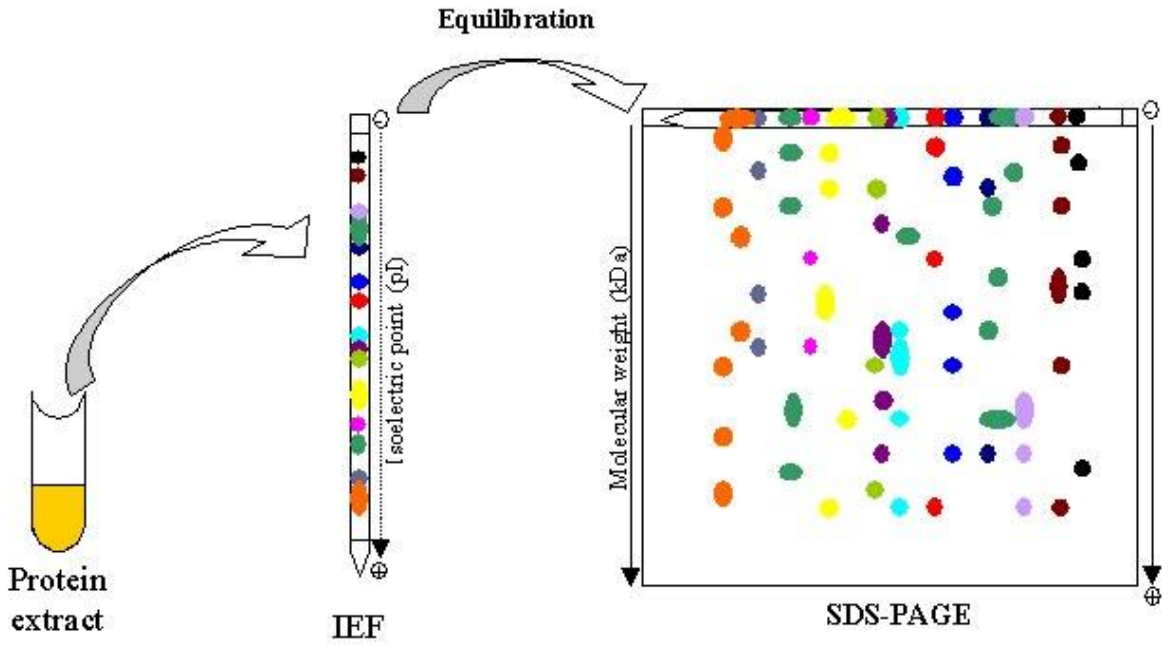
Visualization of results: Staining

There are various methods for visualization of proteins, but the most commonly used are Coomassie Blue staining and silver staining. Silver staining is a sensitive and non-radioactive method. The principle of silver staining is quite simple. The amino acid side chains can bind to silver ions, primary the sulfhydryl and carboxyl groups of proteins, followed by reduction to free metallic silver. As a result, the protein bands are visualized as spots where the reduction occurs. Silver staining is suitable for low protein levels because of its sensitivity (in the very low ng range).

Coomassie Blue staining is a relatively simple method and more quantitative than silver staining. It is suitable to detect protein bands containing about 0.2 μg or more proteins. The Coomassie dye binds to proteins to form a protein-dye complex through Van der Waals attractions. There are two kinds of Coomassie dyes, R250 and G-250.

Further analysis of protein.

There are almost no possibilities to detect the appearance of a few new spots or the disappearance of single spots in large studies with several thousand spots. In addition, evaluation of two gels by manual comparison is also impossible. Therefore, it is necessary to detect differences and obtain information from gels by image collection hardware and image evaluation software. There is some 2D gel analysis software, such as Melanie, PDQuest, Progenesis, REDFIN etc. And the gels can be used for the identification and other applications by mass spectrometry.



Two-dimensional electrophoresis (2-D electrophoresis) is a powerful and widely used method for the analysis of complex protein mixtures extracted from cells, tissues, or other biological samples. This technique sorts proteins according to two independent properties in two discrete steps: the first-dimension step, isoelectric focusing (IEF), separates proteins according to their isoelectric points (pi); the second-dimension step, SDS-polyacrylamide gel electrophoresis (SDS-PAGE), separates proteins according to their molecular weights (M_r, relative molecular weight).

Each spot on the resulting two-dimensional array corresponds to a single protein species in the sample. Thousands of different proteins can thus be separated, and information such as the protein pi, the apparent molecular weight, and the amount of each protein are obtained.

Two-dimensional electrophoresis was first introduced by P. H. O'Farrell and J. Klose in 1975. In the original technique, the first-dimension separation was performed in carrier ampholyte-containing polyacrylamide gels cast in narrow tubes. A. Gorg and colleagues developed the currently employed 2-D technique, where carrier ampholyte-generated pH gradients have been replaced with immobilized pH gradients and tube gels replaced with gels supported by a plastic backing.

A large and growing application of 2-D electrophoresis is "proteome analysis." The analysis involves the systematic separation, identification, and quantification of many proteins simultaneously from a single sample. Two-dimensional electrophoresis is used in this technique due to its unparalleled ability to separate thousands of proteins simultaneously.

Two-dimensional electrophoresis is also unique in its ability to detect post- and co-translational modifications, which cannot be predicted from the genome sequence. Applications of 2-D electrophoresis include proteome analysis, cell differentiation, and detection of disease markers, monitoring therapies, drug discovery, cancer research, purity checks, and micro-scale protein purification.

The 2-D process begins with sample preparation. Proper sample preparation is absolutely essential for a good 2-D result. The next step in the 2-D process is IPG (Isoelectric pH gradient) strip rehydration. IPG strips are provided dry and must be rehydrated with the appropriate additives prior to IEF (Immunoelectrophoresis).

First-dimension IEF is performed on a flatbed system at very high voltages with active temperature control. Next, strip equilibration in SDS-containing buffer prepares the sample for the second-dimension separation. Following equilibration, the strip is placed on the second-dimension gel for SDS-PAGE. The final steps are visualization and analysis of the resultant two-dimensional array of spots.

In summary, the experimental sequence for 2-D electrophoresis is:

1. Sample preparation

2. IPG strip rehydration

3. IEF

4. IPG strip equilibration

5. SDS-PAGE

6. Visualization

7. Analysis.

Sample Preparation:

Due to the great diversity of protein sample types and origins, only general guidelines for sample preparation are provided in this section. The optimal procedure must be determined empirically for each sample type. Ideally, the process will result in the complete solubilization, disaggregation, denaturation, and reduction of the proteins in the sample.

When developing a sample preparation strategy, it is important to have a clear idea of what is desired in the final 2-D result. Is the goal to view as many proteins as possible, or is only a subset of the proteins in the sample of potential interest? Which is more important – complete sample representation, or a clear, reproducible pattern?

Additional sample preparation steps can improve the quality of the final result, but each additional step can result in the selective loss of protein species. The trade-off between improved sample quality and complete protein representation must, therefore, be carefully considered.

In order to characterize specific proteins in a complex protein mixture, the proteins of interest must be completely soluble under electrophoresis conditions. Different treatments and conditions are required to solubilize different types of protein samples; some proteins are naturally found in complexes with membranes, nucleic acids, or other proteins, some proteins form various nonspecific aggregates, and some proteins precipitate when removed from their normal environment.

The effectiveness of solubilization depends on the choice of cell disruption method, protein concentration and dissolution method, choice of detergents, and composition of the sample solution. If any of these steps are not optimized for a particular sample, separations may be incomplete or distorted and information may be lost.

To fully analyze all intracellular proteins, the cells must be effectively disrupted. Choice of disruption method depends on whether the sample is from cells, solid tissue, or other biological material and whether the analysis is targeting all proteins or just a particular subcellular fraction. Both gentle and vigorous lysis methods are discussed below.

Proteases may be liberated upon cell disruption. Proteolysis greatly complicates analysis of the 2-D result, thus the protein sample should be protected from proteolysis during cell disruption and subsequent preparation. Proteases are less active at lower temperatures, so sample preparation at as low a temperature as possible is recommended.

In addition, proteolysis can often be inhibited by preparing the sample in the presence of tris base, sodium carbonate, or basic carrier ampholyte mixtures. These approaches alone are often sufficient protection against proteolysis. However, some proteases may

retain some activity even under these conditions. In these cases, protease inhibitors may be used.

Individual protease inhibitors are only active against specific classes of proteases, so it is usually advisable to use a combination of protease inhibitors. Broad range protease inhibitor "cocktails" are available from a number of commercial sources.

If only a subset of the proteins in a tissue or cell type is of interest, pre-fractionation can be employed during sample preparation. If proteins from one particular subcellular compartment (e.g., nuclei, mitochondria, plasma membrane) are desired, the organelle of interest can be purified by differential centrifugation or other means prior to solubilization of proteins for 2-D electrophoresis.

The sample can also be pre-fractionated by solubility under different extraction conditions prior to 2-D electrophoresis. Precipitation of the proteins in the sample and removal of interfering substances are optional steps. The decision to employ these steps depends on the nature of the sample and the experimental goal. Precipitation procedures, which are used both to concentrate the sample and to separate the proteins from potentially interfering substances, are described below.

No precipitation technique is completely efficient and some proteins may not readily re-suspend following precipitation. Thus, employing a precipitation step during sample preparation can alter the protein profile of a sample. Precipitation and re-suspension should be avoided if the aim of a 2- D experiment is complete and accurate representation of all the proteins in a sample.

If sample preparation requires precipitation, typically only one of the following precipitation techniques is employed:

1. Ammonium sulphate precipitation:

("Salting out") In the presence of high salt concentrations, proteins tend to aggregate and precipitate out of solution. Many potential contaminants (e.g., nucleic acids) will remain in solution.

2. TCA precipitation:

TCA (trichloroacetic acid) is a very effective protein precipitant.

3. Acetone precipitation:

This organic solvent is commonly used to precipitate proteins. Many organic-soluble contaminants (e.g., detergents, lipids) will remain in solution.

4. Precipitation with TCA in acetone:

The combination of TCA and acetone is commonly used to precipitate proteins during sample preparation for 2-D electrophoresis and is more effective than either TCA or acetone alone.

5. Precipitation with ammonium acetate in methanol following phenol extraction:
This technique has proven useful with plant samples containing high levels of interfering substances. Non-protein impurities in the sample can interfere with separation and subsequent visualization of the 2-D result, so sample preparation can include steps to rid the sample of these substances.

In general, it is advisable to keep sample preparation as simple as possible. A sample with low protein concentrations and a high salt concentration, for example, could be diluted normally and analyzed, or desalted, then concentrated by lyophilization, or precipitated with TCA and ice-cold acetone and re-solubilized with rehydration solution.

The first option of simply diluting the sample with rehydration solution may be sufficient. If problems with protein concentration or interfering substances are otherwise insurmountable, then precipitation or removal steps may be necessary. The composition of the sample solution is particularly critical for 2-D because solubilisation treatments for the first-dimension separation must not affect the protein pi, nor leave the sample in a highly conductive solution. In general, concentrated urea's as well as one or more detergents are used. Sample solution composition is discussed herewith.

These always include urea and one or more detergents. Complete denaturation ensures that each protein is present in only one configuration and that aggregation and intermolecular interaction are avoided. The lysis solution, which contains urea and the zwitterionic detergent CHAPS, has been found to be effective for solubilizing a wide range of samples.

Reductant and IPG buffer are also frequently added to the sample solution to enhance sample solubility. IEF performed under denaturing conditions gives the highest resolution and the cleanest results. Urea, a neutral chaotrope, is used as the denaturant in the first- dimension of 2-D electrophoresis. Urea solubilizes and unfolds most proteins to their fully random conformation, with all ionizable groups exposed to solution. Recently, the use of thiourea in addition to urea has been found to further improve solubilization, particularly of membrane proteins. A non-ionic or zwitterionic detergent is always included in the sample solution to ensure complete sample solubilization and to prevent aggregation through hydrophobic interactions. Originally, either of two similar non-ionic detergents, NP-40 or Triton X-100, was used.

When difficulties in achieving full sample solubilization are encountered, the anionic detergent SDS can be used as a solubilizing agent. SDS is a very effective protein solubilizer, but as because it is charged and forms complexes with proteins, it cannot be used as the sole detergent for solubilizing samples for 2-D electrophoresis.

A widely used method for negating the interfering effect of SDS is dilution of the sample into a solution containing an excess of CHAPS, Triton X-100, or NP-40. The final concentration of SDS should be 0.25% or lower and the ratio of the excess detergent to SDS should be at least 8:1.Reducing agents are frequently included in the sample solution to break any disulphide bonds present and to maintain all proteins in their fully reduced state.

The most commonly used reductant is dithiothreitol (DTT) at concentrations ranging from 20 to 100 mM.Dithioerythreitol (DTE) is similar to DTT and can also be used as a reducing agent. Originally, 2-mercaptoethanol was used as a reductant, but higher concentrations of the reductant are required and inherent impurities may result in artifacts. More recently, the non-thiol reductant tributyl phosphine (TBP), at a concentration of 2 mM, has been used as a reductant for 2-D samples.

However, due to the limited solubility and instability of TBP in solution, a thiol reductant such as DTT should be used to maintain proteins in their reduced state through rehydration and first-dimension IEF, if TBP is employed as a reductant during sample preparation. Carrier ampholytes or IPG buffer (up to 2% (v/v)) can be included in the sample solution.

They enhance protein solubility by minimizing protein aggregation due to charge-charge interactions. In some cases, buffers or bases (e.g., 40 mM Tris base) are added to the sample solution. This is done when basic conditions are required for full solubilization or to minimize proteolysis. However, introduction of such ionic compounds can result in first-dimension disturbances.

Bases or buffers should be diluted to 5 mM or lower prior to loading the sample onto first-dimension IEF. A sample should remain in sample solution at room temperature for at least 30 min for full denaturation and solubilization prior to centrifugation and subsequent sample application. Heating of the sample in the presence of detergent can aid in solubilization, but should only be done prior to the addition of urea. Sonication helps speed up solubilization, particularly from material that is otherwise difficult to re-suspend.

First-dimension Isoelectric Focusing (IEF):

A useful first-dimension separation requires selecting a first-dimension pH range appropriate for the sample, as well as a suitable sample application method. IEF is an electrophoretic method that separates proteins according to their isoelectric points (pi). Proteins are amphoteric molecules; they carry either positive, negative, or zero net charge, depending on the pH of their surroundings (Fig. 8.25).

The net charge of a protein is the sum of all the negative and positive charges of its amino acid side- chains and amino- and carboxyl-ter- mini. The isoelectric point (pi) is

the specific pH at which the net charge of the protein is zero. Proteins are positively charged at pH values below their pi and negatively charged at pH values above their pi. If the net charge of a protein is plotted versus the pH of its environment, the resulting curve intersects the x-axis at the isoelectric point (Fig. 8.25).

The presence of a pH gradient is critical to the IEF technique. In a pH gradient, under the influence of an electric field, a protein will move to the position in the gradient where its net charge is zero. A protein with a positive net charge will migrate toward the cathode, becoming progressively less positively charged as it moves through the pH gradient until it reaches its pi.

A protein with a negative net charge will migrate toward the anode, becoming less negatively charged until it also reaches zero net charge. If a protein should diffuse away from its pi, it immediately gains charge and migrates back. This is the focusing effect of IEF, which concentrates proteins at their pls and allows proteins to be separated on the basis of very small charge differences.

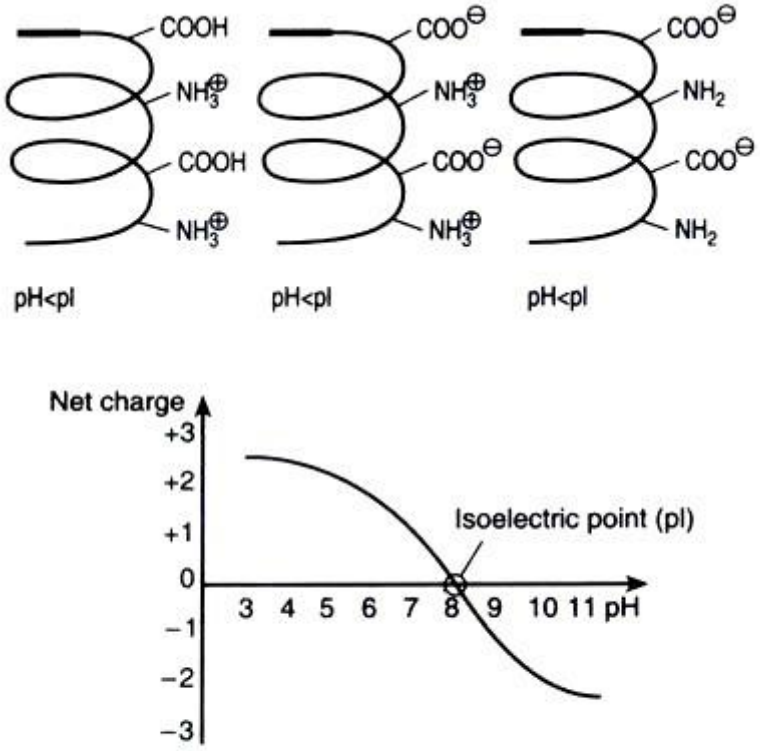


Fig. 8.25: Plot of the net charge of a protein versus the pH of its environment. The point of intersection of the curve at the x-axis represents the isoelectric point of the protein.

The resolution is determined by the slope of the pH gradient and the electric field strength. IEF is, therefore, performed at high voltages (typically in excess of 1 000 V). When the proteins have reached their final positions in the pH gradient, there is very

little ionic movement in the system, resulting in a very low final current (typically below 1 mA). IEF of a given sample in a given electrophoresis system is generally performed for a constant number of volt-hours.

The original method for first-dimension IEF depended on carrier ampholyte-generated pH gradients in polyacrylamide gel rods in tubes. Carrier ampholytes are small, soluble, amphoteric molecules with a high buffering capacity near their pi. Commercial carrier ampholyte mixtures are comprised of hundreds of individual polymeric species with pis spanning a specific pH range.

When a voltage is applied across a carrier ampholyte mixture, the carrier ampholytes with the highest pi (and the most negative charge) move toward the anode and the carrier ampholytes with the lowest pi (and the most positive charge) move toward the cathode. The other carrier ampholytes align themselves between the extremes, according to their pis, and buffer their environment to the corresponding pHs. The result is a continuous pH gradient.

Although this basic method has been used in hundreds of 2-D electrophoresis studies, it has several limitations that have prevented its more widespread application:

i. Carrier ampholytes are mixed polymers that are not well characterized and suffer from batch-to-batch manufacturing variations. These variations reduce the reproducibility of the first-dimension separation.

ii. Carrier ampholyte pH gradients are unstable and have a tendency to drift, usually toward the cathode, over time. Gradient drift adversely affects reproducibility by introducing a time variable. Gradient drift also causes a flattening of the pH gradient at each end, particularly above pH 9, rendering the 2-D technique less useful at pH extremes.

iii. The soft polyacrylamide tube gels have low mechanical stability. The gel rods may stretch or break, affecting reproducibility. Results are often dependent on the skill of the operator.

As a result of the limitations and problems with carrier ampholyte pH gradients, immobilized pH gradients were developed .An immobilized pH gradient (IPG) is created by covalently incorporating a gradient of acidic and basic buffering groups into a polyacrylamide gel at the time it is cast. The buffers are a set of well-characterized molecules, each with a single acidic or basic buffering group linked to an acrylamide monomer.

The general structure of immobilized reagents is:
R = weakly acidic or basic buffering group. Immobilized pH gradients are formed using two solutions, one containing a relatively acidic mixture of acrylamide buffers and the other containing a relatively basic mixture. The concentrations of the various buffers in the two solutions define the range and shape of the pH gradient produced.

Both solutions contain acrylamide monomers and catalysts. During polymerization, the acrylamide portion of the buffers copolymerizes with the acrylamide and bisacrylamide monomers to form a polyacrylamide gel. Fig. 8.26 is a graphic representation of the polyacrylamide matrix with attached buffering groups.

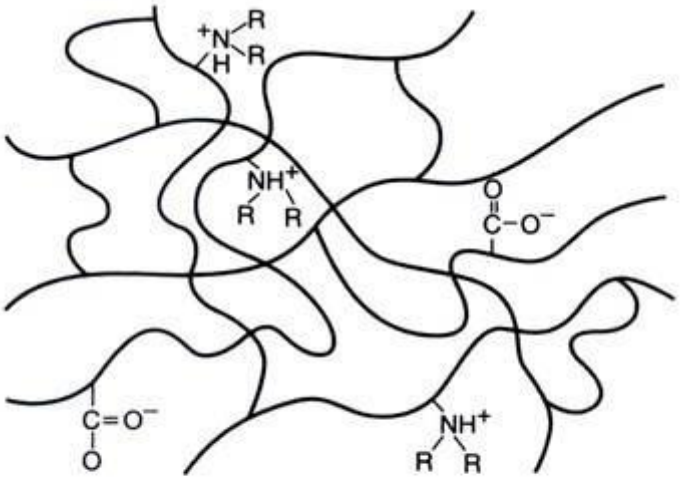


Fig. 8.26: Immobilized pH gradient polyacrylamide gel matrix showing attached buffering groups.

IPG strips must be rehydrated prior to IEF. The IPG strips are rehydrated in a solution containing the necessary additives and, optionally, the sample proteins. IEF is performed by gradually increasing the voltage across the IPG strips to at least 3500 V and maintaining this voltage for at least several thousand volt-hours.

After IEF, the IPG strips are equilibrated in equilibration solution and applied onto flatbed or vertical SDS-polyacrylamide gels. When IPG strips are used for the first-dimension separation, the resultant 2-D maps are superior in terms of resolution and reproducibility. IPG strips are a marked improvement over tube gels with carrier ampholyte-generated pH gradients.

Ready-made IPG strips are commercially available. It is always advisable to choose shorter strips for fast screening or when the most abundant proteins are of interest and to use longer strips for maximal resolution and loading capacity.

Sample can be applied either by including it in the rehydration solution (rehydration loading) or by applying it directly to the rehydrated IPG strip via sample cups, sample wells, or paper bridge. Usually rehydration loading is preferable.

Advantages to this mode of application include the following:

i. Rehydration loading allows larger quantities of protein to be loaded and separated.

ii. Rehydration loading allows more dilute samples to be loaded.

iii. Because there is no discrete application point, this method eliminates the formation of precipitates at the application point that often occurs when loading with sample cups.

iv. The rehydration loading method is technically simpler, avoiding problems of leakage that can occur when using sample cups.

v. There are, however, cases when one might prefer to load the sample following rehydration, immediately prior to IEF, e.g., if proteolysis or other protein modifications are a concern, overnight rehydration with sample may not be desired.

The IPG strips are rehydrated in the especially designed rehydration tray or cassette. Rehydration solution, which may or may not include the sample, is applied to the reservoir slots of the rehydration tray and then the IPG strips are soaked individually.

The choice of the most appropriate rehydration solution for the sample will depend on its specific protein solubility requirements, but a typical solution generally contains urea, non-ionic or zwitterionic detergent, dithiothreitol (DTT), Pharmalytes. The sample may also be included. The role of each component is described below, as well as the recommended concentration range.

Urea solubilizes and denatures proteins, unfolding them to expose internal ionizable amino acids. Commonly 8 M urea is used, but the concentration can be increased to 9 M or 9.8 M, if necessary for complete sample solubilization. Thiourea, in addition to urea, can be used to further improve protein solubilization.

Detergent solubilizes hydrophobic proteins and minimizes protein aggregation. The detergent must have zero net charge—use only non-ionic and zwitterionic detergents. CHAPS, Triton X-100, or NP-40 in the range of 0.5 to 4% are most commonly used.

Reductant cleaves disulphide bonds to allow proteins to unfold completely. DTT or DTE, (20 to 100 mM) are commonly used. 2-Mercaptoethanol is not recommended, because higher concentrations are required, and impurities may result in artifacts. Tributyl phosphine (TBP) is not recommended as reductant for IEF due to its low solubility and poor stability in rehydration solution. Reductants should be added directly before use.

Pharmalyte (carrier ampholyte mixtures) improve separations, particularly with high sample loads. Carrier ampholyte mixtures enhance protein solubility and produce more uniform conductivity across the pH gradient without disturbing IEF or affecting the shape of the gradient.

Tracking dye (bromophenol blue) allows IEF progress to be monitored at the beginning of the protocol. If the tracking dye does not migrate toward the anode, no current is flowing. As isoelectric focusing proceeds, the bromophenol blue tracking dye migrates toward the anode. Note that the dye front leaves the IPG strip well before focusing is complete, so clearing of the dye is no indication that the sample is focused. If the dye does not migrate, no current is flowing.

Sample Application by Cup Loading:

If the sample was not applied by means of the rehydration solution, it can be applied using the sample cups, immediately prior to isoelectric focusing. When sample cups are used, the sample load limits are lower and more specific. Higher sample volumes and protein amounts can be applied with paper bridges, which are placed between the anodic or cathodic end of the IPG strip and the electrode strip. A large sample volume requires a large paper pad applied at the other side to absorb excess water.

In a typical IEF protocol, voltage is gradually increased to the final desired focusing voltage, which is held for up to several hours. With cup loading, a low initial voltage minimizes sample aggregation and a low initial voltage generally allows the parallel separation of samples with differing salt concentrations.

The main factors determining the required voltage-hours (Vh) are the length of the IPG strips and the pH gradient used. Sample composition, rehydration solution composition, and sample application mode influence the required voltage-hours.

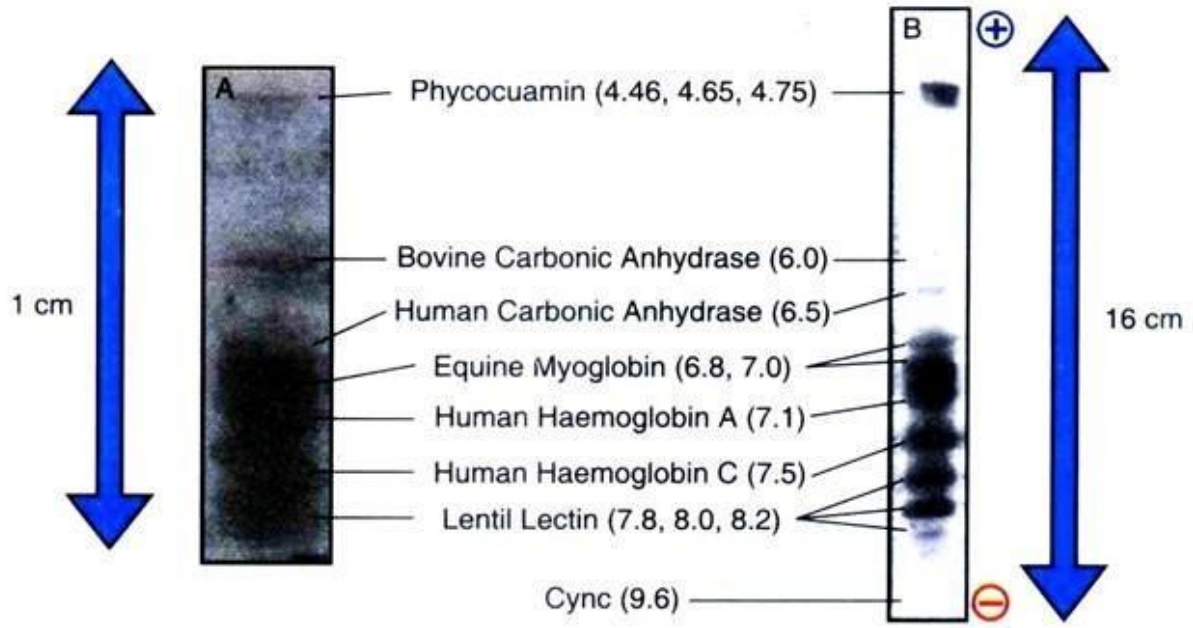


Fig. 8.27: Scanned pictures of IEF electrophoresis on (A) 1 cm^2 and (B) 16 cm^2 gels. Only one channel of the whole gel is shown.

Focusing for substantially longer than recommended will cause horizontal streaking and loss of proteins. This phenomenon is called "over-focusing". Therefore, focusing time should be reduced to the minimum necessary.

Second-Dimension SDS-PAGE:

After IEF, the second-dimension separation can be performed on various flatbed or vertical systems.

SDS-PAGE consists of four steps:

(1) Preparing the second-dimension gel,

(2) Equilibrating the IPG strip(s) in SDS buffer,

(3) Placing the equilibrated IPG strip on the SDS gel, and

(4) Electrophoresis.

IPG Strip Equilibration:

The equilibration step saturates the IPG strip with the SDS buffer system required for the second- dimension separation. The equilibration solution contains buffer, urea, glycerol, reductant, SDS, and dye. An additional equilibration step replaces the reductant with iodoacetamide. Equilibration is always performed immediately prior to the second-dimension run, never prior to storage of the IPG strips at -40 °C or lower.

Once equilibrated, place the IPG strips using forceps, gel-side down horizontally on vertical SDS-PAGE gel for electrophoresis. Electrophoresis is performed at constant current in two steps. During the initial migration and stacking period, the current is approximately half of the value required for the separation. Stop electrophoresis when the dye front is approximately 1 mm from the bottom of the gel. After electrophoresis, remove gels from their gel cassettes in preparation for staining or blotting.

Visualization of Results:

Most detection methods used for SDS gels can be applied to second-dimension gels.

However, the following features are desired:

i. High sensitive

ii. Wide linear range for quantification

iii. Compatibility with mass spectrometry

iv. Low toxicity and environmentally safe

v. Environmentally friendly.

Because none of the existing techniques can meet all these requirements, a 2-D electrophoresis laboratory may need to have more than one of the following methods in its repertoire.

Autoradiography and fluorography are the most sensitive detection methods (down to 200 fg protein). To employ these techniques, the sample must consist of protein radiolabelled in vivo using either ^{35}S, ^{14}C, ^{3}H or, in the case of phosphoproteins, ^{32}P. For auto-radiographic detection, the gel is simply dried and exposed to X-ray film or—for quicker results and superior dynamic range of quantification—to a storage phosphor screen. Fluorography is a technique that provides extra sensitivity by impregnating the gel in a scintillant such as PPO (2, 4-diphenyloxazole) prior to drying.

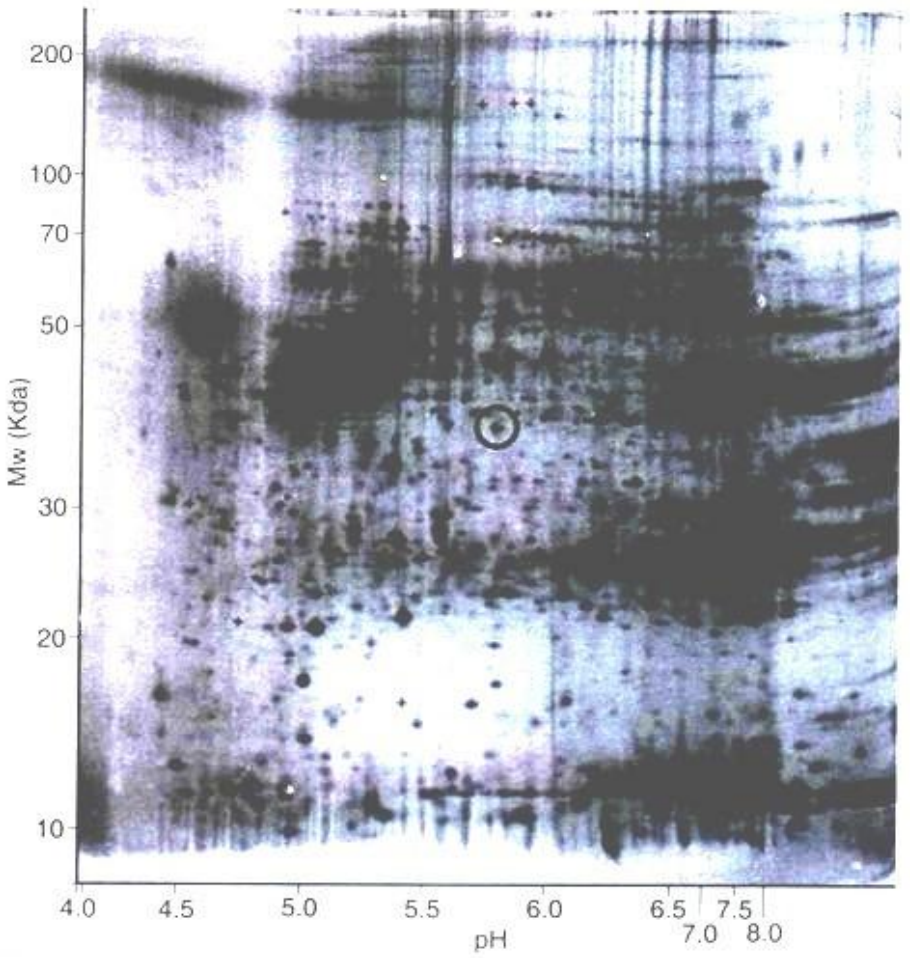


Fig. 8.28: A 2D gel (*Courtsey*: Dr D Dash, Deptt. of Biochemistry, IMS, BHU)

Silver staining is the most sensitive non-radioactive method (below 1 ng). Silver staining is a complex, multi-step process utilizing numerous reagents for which quality is critical. It is, therefore, often advantageous to purchase these reagents in the form of a dedicated kit, in which the reagents are quality assured specifically for the silver-staining application. By omitting glutardialdehyde from the sensitizer and formaldehyde from the silver nitrate solution the method becomes compatible with mass spectrometry analysis, however, at the expense of sensitivity.

Coomassie staining, although 50- to 100-fold less sensitive than silver staining, is a relatively simple method and more quantitative than silver staining. Coomassie blue is preferable when relative amounts of protein are to be determined by densitometry. Colloidal staining methods are recommended, because they show the highest sensitivity, down to 100 ng/protein spot.

Negative Zinc—Imidazole staining has a detection limit of approx. 15 ng protein/spot and is well compatible with mass spectrometry, but it is a poor quantification technique. Fluorescent labelling and fluorescent staining with dyes have a sensitivity in-between colloidal Coomassie and Silver Staining.

These techniques require fluorescence scanners, but they are compatible with mass spectrometry and show a wide dynamic range for quantification. Apart from staining, Second-dimension gels can be blotted onto a nitrocellulose or PVDF membrane for immunochemical detection of specific proteins or chemical micro sequencing.

Preserving the Gels:

The gels are optimally stored in sheet protectors after soaking them in 10% v/v glycerol for 30 min. Un-backed gels are shrunk back to their original sizes by soaking them in 30% (v/v) methanol or ethanol/4% glycerol until they match their original sizes. For autoradiography the gels are dried onto strong filter paper with a vacuum drier or in-between two sheets of wet cellophane sealed at ends.

Further Analysis of Protein Spots:

a. Picking the spots:

Robotic systems are available that automatically picks selected protein spots from stained or de-stained gels using a pick list from the image analysis, and transfers them into micro-plates for further analysis.

b. Digestion of the proteins:

The gel plugs are automatically digested in the computer controlled Digester; the supernatant peptides are mixed with MALDI matrix material and spotted onto MALDI slides using robotic spotter.

c. MALDI-ToF mass spectrometry:

In the MALDI-ToF mass spectrometer, a laser beam is fired into the dried peptide-matrix spots for ionization of the peptides. After accurate determination of the peptide masses, databases are searched for identification of the original proteins.

Probable questions:

1. Differentiate structural and functional proteomnics.
2. Discuss Edmann Degradation.
3. How mass spectrophotometry can be used in protein identification?
4. Discuss the applications of proteomics.
5. How protein protein interactions can be studied?
6. Briefly discuss post translational modifications of proteins.
7. How phosphorylation affects protein structure?
8. How glycosylation affects protein structure?
9. How ubiquitanation affects protein structure?
10. How n-nitrosylation affects protein structure?
11. How methylation affects protein structure?
12. How lipidation affects protein structure?
13. How proteolysis can be assessed?
14. Describe iso-electric focussing.

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

UNIT-XI

Transcriptome, Transcriptome analysis, DNA microarray expression profiling, Data processing and presentation, RNASequencing.

Objective: In this unit we will discuss about Transcriptome analysis, DNA microarray expression profiling, Data processing and presentation, RNA Sequencing.

Introduction:

After the genome sequences are being completed, the new questions arise about the functional roles of different genes; the cellular processes in which they participate; mechanism by which the genes regulate the interaction of genes and gene products; changes in level of gene expression in different cell types and states.

To answer all these questions, the new area of science has got emerged which is transcriptomics. The transcription of genes to produce RNA is the first stage of gene expression. Although mRNA is not the ultimate product of a gene, but it is the first step of gene regulation and information about the transcript levels which is needed for understanding gene regulatory networks.

The transcriptome is the complete set of mRNA transcripts produced by the genome at any one time. Unlike the genome, the transcriptome is extremely dynamic, all the cells of an organism contain same genome but the transcriptome varies considerably in different cells at different circumstances due to different patterns of gene expression.

Techniques for Transcriptome Analysis:

High throughput techniques based on DNA chip/microarray technology (i.e., cDNA microarrays, oligo microarrays), cDNA- AFLP (cDNA-amplified fragment length polymorphism) analysis, SAGE (serial analysis of gene expression) and a new technique MPSS (massively parallel signature sequencing) are used for transcriptome analysis.

The cDNA microarray technique is based on the ability of the mRNA molecule to bind specifically to or hybridize to, its original DNA coding sequence in the form of a cDNA template spotted on an array.DNA chip is prepared on a silicon or glass based surface with regions of known sequence of chosen target DNA, which can hybridize with an unknown labelled DNA sample.

Besides using cDNA clones as probes on an array, oligonucleotides of around 20 nucleotides can also be used as probe. Microarray experiments allow for comparison of gene expression profiles between two mRNA samples (e.g., treatment vs. control, or treatment 1 vs. treatment 2).

The most important advantage of microarray-based technology is that large data sets from different experiments can be combined together in a single database, which allows gene expression profiles from either different samples or samples from different treatments to be compared with each other and analysed together.

Significance of Transcriptomics:

As the transcriptome includes all mRNA transcripts in the cell, it reflects the genes that are being actively expressed at any given time, with the exception of mRNA degradation phenomenon such as transcriptional attenuation. The study of transcriptomics examines the expression level of mRNA in a given cell population.

Many DNA sequences that have been isolated shown to have no known functions. However, if they show similar expression patterns to a characterized gene, it is likely that their functions are similar.

It is sometimes possible to identify conserved regulatory sequences of such genes. Ultimately, these studies promise to expand the size of existing gene families, reveal new patterns of coordinated gene expression across gene families and uncover entirely new categories of genes.

Furthermore, the product of any one gene usually interacts with those of many others, therefore, transcriptomics will provide precise knowledge on coordination among genes and their inter-relationships.

It will also help to understand the integration of gene expression and function at the cellular level, revealing how multiple gene products work together to produce physical and chemical responses to both static and changing cellular needs.

Microarray Analysis:

A DNA microarray (also commonly known as **gene** or **genome** chip, DNA chip, or gene array) is a collection of microscopic DNA spots, commonly representing single genes, arrayed on a solid surface by covalent attachment to chemically suitable matrices.
DNA arrays are different from other types of microarray, only in that they either measure DNA or use DNA as part of its detection system.

Qualitative or quantitative measurements with DNA microarrays utilize the selective nature of DNA-DNA or DNA-RNA hybridization under high-stringency conditions and fluorophore-based detection. DNA arrays are commonly used for expression profiling, i.e., monitoring expression levels of thousands of genes simultaneously, or for comparative genomic hybridization.

Arrays of DNA can either be spatially arranged, as in commonly known gene or genome chip, DNA chip, or gene array, or can be specific DNA sequences tagged or labelled such that they can be independently identified in solution. The traditional solid-phase array is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip.

The affixed DNA segments are known as probes (although some sources will use different nomenclature such as reporters), thousands of which can be placed in known locations on a single DNA microarray. Microarray technology evolved from Southern blotting, whereby fragmented DNA is attached to a substrate and then probed with a known gene or fragment.

Principle of DNA microarray:

- DNA microarray technology was originated from Southern blotting, in which fragmented DNA is attached to a substrate and then probed with a known DNA sequence.
- DNA microarray is based on principle of hybridization between the nucleic acid strands.
- Complementary nucleic acid sequences have the characteristic to specifically pair to each other by the formation of hydrogen bonds between complementary nucleotide base pairs.
- Unknown sample of DNA sequence is termed as sample or target and the known sequence of DNA molecule is called as probe.
- Fluorescent dyes are used for labelling the samples and at least 2 samples are hybridized to the chip.

- A large number of complementary base pairs in nucleotide sequence is suggestive of tighter non-covalent bonding between the two strands.
- Following the washing off of non-specific bonding sequences, only strongly paired strands will stay hybridized.
- Thus, the fluorescent labeled target sequences that pairs to the probe releases a signal that relies on the strength of the hybridization detected by the number of paired bases, hybridization conditions, and washing after hybridization.
- DNA microarrays employs relative quantization in which the comparison of same character is done under two different conditions and the identification of that character is known by its position.
- After completion of the hybridization, the surface of chip can be examined both qualitatively and quantitatively by use of autoradiography, laser scanning, fluorescence detection device, enzyme detection system.
- The presence of one genomic or cDNA sequence in 1,00,000 or more can be screened in a single hybridization by using DNA microarray.
-

Types of DNA microarray:

1. cDNA based microarrays
2. Oligonucleotide based microarrays

1. cDNA based microarrays:

- cDNA is used for the preparation of chips.
- cDNAs are amplified by PCR.
- It is a high throughput technique.
- It is highly parallel RNA expression assay technique that allows quantitative analysis of RNAs transcribed from both known and unknown genes.

2. Oligonucleotide based microarrays:

- In this type, the spotted probes contains of short, chemically synthesized sequences, 20-25 mers/gene.
- Shorter probe lengths allows less errors during probe synthesis and enables the interrogation of small genomic regions, plus polymorphisms
- Despite being easier to produce than dsDNA probes, oligonucleotide probes need to be carefully designed so that all probes acquire similar melting temperatures (within 5^0 c) and eliminate palindromic sequences.
- The probe's attachment to the glass slides takes place by the covalent linkage as electrostatic immobilization and cross-linking can result in significant loss of probes during wash steps due to their small size.

- The coupling of probes to the microarray surface takes place via modified 5′ to 3′ ends on coated slides that provide functional groups (epoxy or aldehyde)

Applications of Microarray:

1. mRNA or gene expression profiling:
Monitoring expression levels for thousands of genes simultaneously is relevant to many areas of biology and medicine, such as studying treatments, disease, and developmental stages. For example, microarrays can be used to identify disease genes by comparing gene expression in diseased and normal cells.

2. Comparative genomic hybridization (Array CGH):
Assessing large genomic rearrangements.

3. SNP detection arrays:
Looking for single nucleotide polymorphism in the genome of populations.

4. Chromatin immunoprecipitation (ChIP) studies:
Determining protein binding site occupancy throughout the genome, employing ChlP-on-chip technology.

Fabrication:

Microarrays can be fabricated using a variety of technologies, including printing with fine-pointed pins onto glass slides, photolithography using pre-made masks, photolithography using dynamic micro-mirror devices, ink-jet printing, or electrochemistry on microelectrode arrays.

DNA microarrays can be used to detect RNAs that may or may not be translated into active proteins. Scientists refer to this kind of analysis as “expression analysis” or expression profiling. Since there can be tens of thousands of distinct probes on an array, each microarray experiment can accomplish the equivalent number of genetic tests in parallel.Arrays have, therefore, dramatically accelerated many types of investigations. The use of microarrays for gene expression profiling was first published in 1995 (Science) and the first complete eukaryotic genome {Saccharomyces cerevisiae) on a microarray was published in 1997 (Science).

1. Spotted Microarrays:

In spotted microarrays (or two-channel or two-colour microarrays), the probes are oligonucleotides, cDNA or small fragments of PCR products that correspond to mRNAs and are spotted onto the microarray surface. This type of array is typically hybridized with cDNA from two samples to be compared (e.g., diseased tissue versus healthy tissue) that are labelled with two different fluorophores (e.g., Rhodamine (Cyanine 5, red) and Fluorescein (Cyanine 3, green)).

The two samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores. Relative intensities of each fluorophore are then used to identify up-regulated and down-regulated genes in ratio-based analysis. Absolute levels of gene expression cannot be determined in the two-colour array, but relative differences in expression among different spots (= genes) can be estimated with some oligonucleotide arrays.

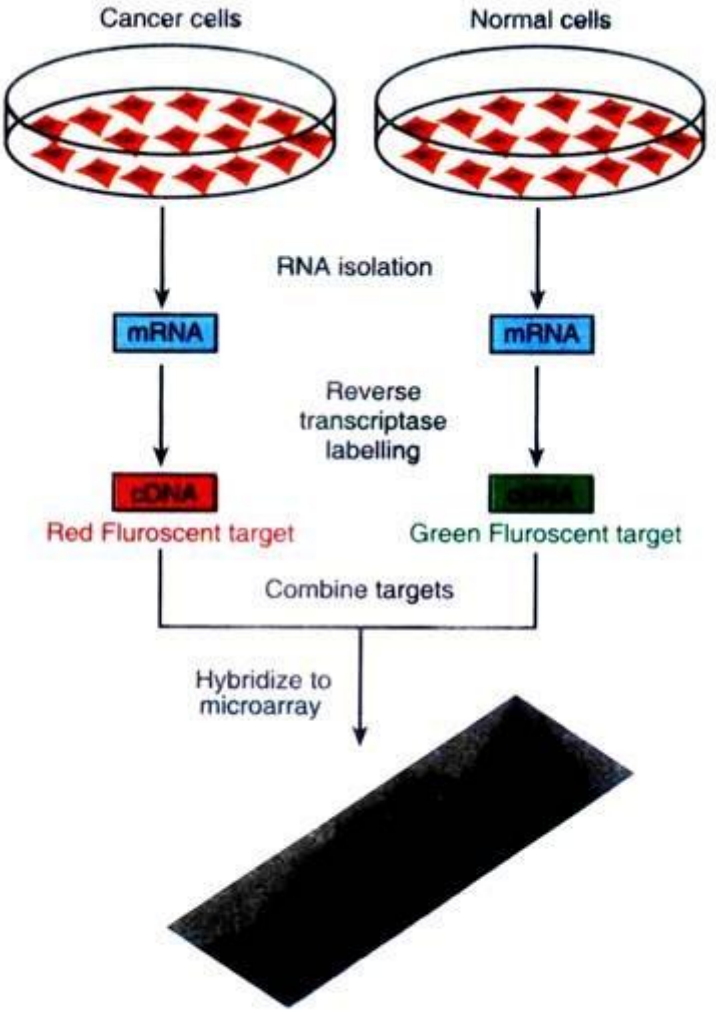


Fig. 16.2: Diagram of typical dual-colour microarray experiment

2. Oligonucleotide Microarrays:

In oligonucleotide microarrays (or single-channel microarrays), the probes are designed to match parts of the sequence of known or predicted mRNAs. There are commercially available designs that cover complete genomes from companies such as GE Healthcare, Affymetrix, Ocimum Bio-solutions, or Agilent. These microarrays give estimations of the absolute value of gene expression and, therefore, the comparison of two conditions requires the use of two separate micro- arrays.

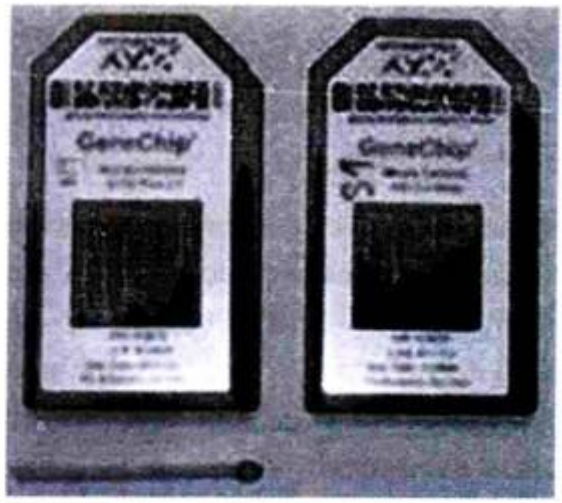


Fig. 16.3: Two Affymetrix chips

Oligonucleotide Arrays can be either produced by piezoelectric deposition with full length oligonucleotides or in situ synthesis. Long Oligonucleotide Arrays are composed of 60-mers, or 50-mers and are produced by ink-jet printing on a silica substrate. Short Oligonucleotide Arrays are composed of 25-mer or 30-mer and are produced by photolithographic synthesis (Affymetrix) on a silica substrate or piezoelectric deposition (GE Healthcare) on an acrylamide matrix.

More recently, Maskless Array Synthesis from NimbleGen Systems has combined flexibility with large numbers of probes. Arrays can contain up to 390,000 spots, from a custom array design. New array formats are being developed to study specific pathways or disease states for a systems biology approach.

Oligonucleotide microarrays often contain control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes.

Genotyping Microarrays:

DNA microarrays can also be used to read the sequence of a genome in particular positions. SNP microarrays are a particular type of DNA microarrays that are used to identify genetic variation in individuals and across populations.

Short oligonucleotide arrays can be used to identify the single nucleotide polymorphisms (SNPs) that are thought to be responsible for genetic variation and the source of susceptibility to genetically caused diseases. Generally termed genotyping applications, DNA microarrays may be used in this fashion for forensic applications, rapidly discovering or measuring genetic predisposition to disease, or identifying DNA-based drug candidates.

These SNP microarrays are also being used to profile somatic mutations in cancer, specifically loss of heterozygosity events and amplifications and deletions of regions of DNA. Amplifications and deletions can also be detected using comparative genomic hybridization, or aCGH, in conjunction with microarrays, but may be limited in detecting novel Copy Number Polymorphisms, or CNPs, by probe coverage.

Re-sequencing arrays have also been developed to sequence portions of the genome in individuals. These arrays may be used to evaluate germ line mutations in individuals, or somatic mutations in cancers. Genome tiling arrays include overlapping oligonucleotides designed to blanket an entire genomic region of interest. Many companies have successfully designed tiling arrays that cover whole human chromosomes.

Microarrays and Bioinformatics:

1. Experimental Design:

Due to the biological complexity of gene expression, the considerations of experimental design that are discussed in the expression profiling article are of critical importance if statistically and biologically valid conclusions are to be drawn from the data.

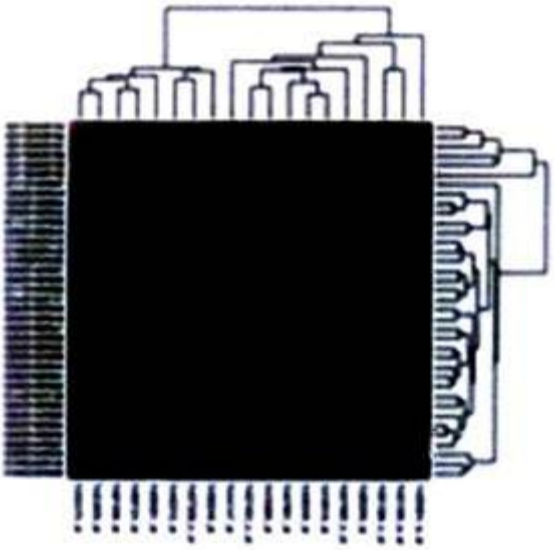


Fig. 16.4: Gene expression values from microarray experiments can be represented as heat maps to visualize the result of data analysis

There are three main elements to consider when designing a microarray experiment.First, replication of the biological samples is essential for drawing conclusions from the experiment.

Second, technical replicates (two RNA samples obtained from each experimental unit) help to ensure precision and allow for testing differences within treatment groups. The technical replicates may be two independent RNA extractions or two aliquots of the same extraction.

Third, spots of each cDNA clone or oligonucleotide are present at least as duplicates on the microarray slide, to provide a measure of technical precision in each hybridization. It is critical that information about the sample preparation and handling is discussed in order to help identify the independent units in the experiment as well as to avoid inflated estimates of significance.

2. Standardization:

The lack of standardization in arrays presents an interoperability problem in bioinformatics, which hinders the exchange of array data. Various grass-roots open-source projects are attempting to facilitate the exchange and analysis of data produced with non-proprietary chips.

1. The “Minimum Information about a Microarray Experiment” (MIAME) checklist helps define the level of detail that should exist and is being adopted by many journals as a requirement for the submission of papers incorporating microarray results. MIAME describes the minimum required information for complying experiments, but not its format. Thus, as of 2007, whilst many formats can support the MIAME requirements there is no format which permits verification of complete semantic compliance.

2. The “MicroArray Quality Control (MAQC) Project” is being conducted by the FDA to develop standards and quality control metrics which will eventually allow the use of MicroArray data in drug discovery, clinical practice and regulatory decision-making.

3. The MicroArray and Gene Expression (MAGE) group is working on the standardization of the representation of gene expression data and relevant annotations.

3. Statistical Analysis:

The analysis of DNA microarrays poses a large number of statistical problems, including the normalization of the data. There are dozens of proposed normalization methods in the published literature; as in many other cases where authorities disagree, a sound conservative approach is to try a number of popular normalization methods and compare the conclusions reached; how sensitive are the main conclusions to the method chosen?

From a hypothesis-testing perspective, the large number of genes present on a single array means that the experimenter must take into account a multiple testing problem; even if the statistical P-value assigned to a given gene indicates that it is extremely unlikely that differential expression of this gene was due to random rather than treatment effects, the very high number of genes on an array makes it likely that differential expression of some genes represents false positives or false negatives.Statistical methods tailored to microarray analyses have recently become available that assess statistical power based on the variation present in the data and the number of experimental replicates, and can help minimize type I and type II errors in the analyses.

A basic difference between microarray data analysis and much traditional biomedical research is the dimensionality of the data. A large clinical study might collect 100 data items per patient for thousands of patients. A medium-size microarray study will obtain many thousands of numbers per sample for perhaps a hundred samples. Many analysis techniques treat each sample as a single point in a space with thousands of dimensions, then attempt by various techniques to reduce the dimensionality of the data to something humans can visualize.

4. Relation between Probe and Gene:

The relation between a probe and the mRNA that it is expected to detect is problematic. On the one hand, some mRNAs may cross-hybridize probes in the array that are supposed to detect another mRNA. On the other hand, probes that are designed to detect the mRNA of a particular gene may be relying on genomic EST information that is incorrectly associated with that gene.

Online Microarray Data Analysis Programs and Tools:

Several Open Directory Project categories list online microarray data analysis programs and tools:

i. Bioinformatics: Online Services:

Gene Expression and Regulation at the Open Directory Project

ii. Gene Expression:

Databases at the Open Directory Project

iii. Gene Expression:

Software at the Open Directory Project

iv. Data Mining:

Tool Vendors at the Open Directory Project

v. Bio-conductor:

Open source and open development software project for the analysis and comprehension of genomic data

vi. Genevestigator:

Web-based database and analysis tool to study gene expression across large sets of tissues, developmental stages, drugs, stimuli, and genetic modifications.

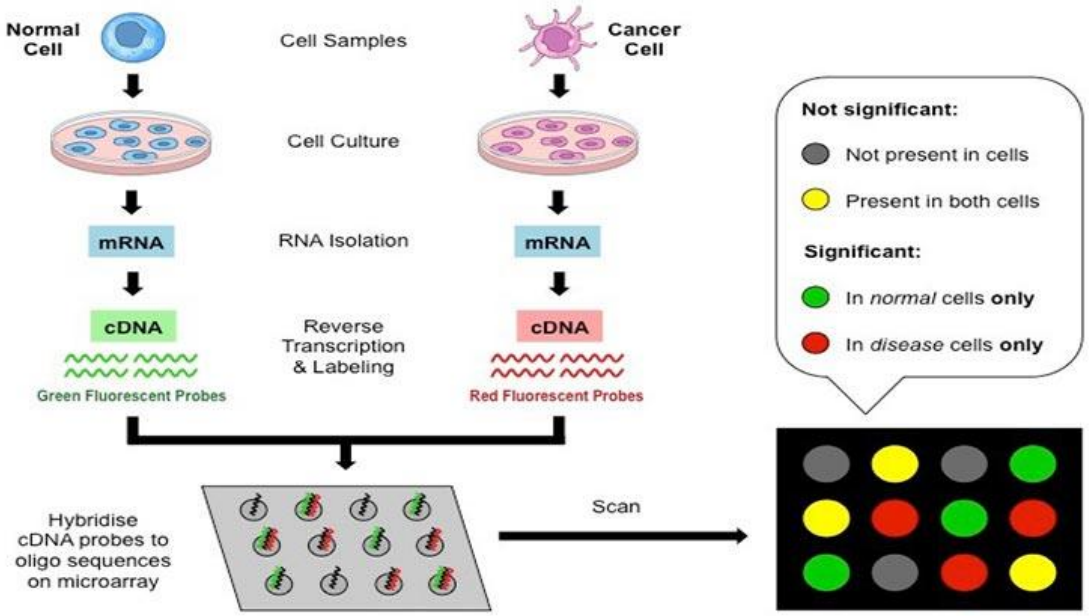
Requirements of DNA microarray:

- DNA chip
- Fluorescent dyes
- Fluorescent labelled target/sample
- Probes
- Scanner

Steps involved in cDNA based microarray:

1. Sample collection
2. Isolation of mRNA

3. Creation of labeled cDNA
4. Hybridization
5. Collection and analysis



Sample collection:

- A sample can be any cell/tissue that we desire to conduct our study on.
- Generally, 2 types of samples are collected, i.e. healthy and infected cells, for comparing and obtaining the results.

Isolation of mRNA:

- The extraction of RNA from a sample is performed by using a column or solvent like phenol-chloroform.
- mRNA is isolated from the extracted RNA leaving behind rRNA and tRNA.
- As mRNA has a poly-A tail, column beads with poly-T tails are employed to bind mRNA.
- Following the extraction, buffer is used to rinse the column inorder to isolate mRNA from the beads.

Creation of labeled cDNA:

- Reverse transcription of mRNA yields cDNA.
- Both the samples are then integrated with different fluorescent dyes for the production of fluorescent cDNA strands which allows to differentiate the sample category of the cDNAs.

Hybridization:

- The labeled cDNAs from both the samples are placed on the DNA microarray which permits the hybridization of each cDNA to its complementary strand.
- Then they are thoroughly washed to remove unpaired sequences.

Collection and analysis:

- Microarray scanner is used to collect the data.
- The scanner contains a laser, a computer and a camera. The laser is responsible for exciting the fluorescence of the cDNA, generating signals.
- The camera records the images produced at the time laser scans the array.
- Then computer stores the data and yields results instantly. The data are now analyzed.
- The distinct intensity of the colors for each spot determines the character of the gene in that particular spot.
-

Applications of DNA microarray technique:

- Drug discovery
- Study of functional genomics
- DNA sequencing
- Gene expression profiling
- Study of proteomics
- Diagnostics and genetic engineering
- Toxicological researches
- Pharmacogenomics and theranostics

RNA sequencing:

RNA sequencing (RNA-Seq) uses the capabilities of high-throughput sequencing methods to provide insight into the transcriptome of a cell. Compared to previous Sanger sequencing- and microarray-based methods, RNA-Seq provides far higher coverage and greater resolution of the dynamic nature of the transcriptome. Beyond quantifying gene expression, the data generated by RNA-Seq facilitate the discovery of novel transcripts, identification of alternatively spliced genes, and detection of allele-specific expression. Recent advances in the RNA-Seq workflow, from sample preparation to library construction to data analysis, have enabled researchers to further elucidate the functional complexity of the transcription. In addition to polyadenylated messenger RNA (mRNA) transcripts, RNA-Seq can be applied to investigate different populations of RNA, including total RNA, pre-mRNA, and noncoding RNA, such as

microRNA and long ncRNA. This article provides an introduction to RNA-Seq methods, including applications, experimental design, and technical challenges.

The central dogma of molecular biology outlines the flow of information that is stored in genes as DNA, transcribed into RNA, and finally translated into proteins (Crick 1958; Crick 1970). The ultimate expression of this genetic information modified by environmental factors characterizes the phenotype of an organism. The transcription of a subset of genes into complementary RNA molecules specifies a cell's identity and regulates the biological activities within the cell. Collectively defined as the transcriptome, these RNA molecules are essential for interpreting the functional elements of the genome and understanding development and disease.

The transcriptome has a high degree of complexity and encompasses multiple types of coding and noncoding RNA species. Historically, RNA molecules were relegated as a simple intermediate between genes and proteins, as encapsulated in the central dogma of molecular biology. Therefore, messenger RNA (mRNA) molecules were the most frequently studied RNA species because they encoded proteins via the genetic code. In addition to protein-coding mRNA, there is a diverse group of noncoding RNA (ncRNA) molecules that are functional. Previously, most known ncRNAs fulfilled basic cellular functions, such as ribosomal RNAs and transfer RNAs involved in mRNA translation, small nuclear RNA (snRNAs) involved in splicing, and small nucleolar RNAs (snoRNAs) involved in the modification of rRNAs (Mattick and Makunin 2006). More recently, novel classes of RNA have been discovered, enhancing the repertoire of ncRNAs. For instance, one such class of ncRNAs is small noncoding RNAs, which include microRNA (miRNA) and piwi-interacting RNA (piRNA), both of which regulate gene expression at the posttranscriptional level (Stefani and Slack 2008). Another noteworthy class of ncRNAs is long noncoding RNAs (lncRNAs). As a functional class, lncRNAs were first described in mice during the large-scale sequencing of cDNA libraries (Okazaki et al. 2002). A myriad of molecular functions have been discovered for lncRNAs, including chromatin remodeling, transcriptional control, and posttranscriptional processing, although the vast majority are not fully characterized (Guttman et al. 2009; Mercer et al. 2009; Wilusz et al. 2009).

Initial gene expression studies relied on low-throughput methods, such as northern blots and quantitative polymerase chain reaction (qPCR), that are limited to measuring single transcripts. Over the last two decades, methods have evolved to enable genome-wide quantification of gene expression, or better known as transcriptomics. The first transcriptomics studies were performed using hybridization-based microarray technologies, which provide a high-throughput option at relatively low cost (Schena et al. 1995). However, these methods have several limitations: the requirement for a priori knowledge of the sequences being interrogated; problematic cross-hybridization artifacts in the analysis of highly similar sequences; and limited ability to accurately

quantify lowly expressed and very highly expressed genes (Casneuf et al. 2007; Shendure 2008). In contrast to hybridization-based methods, sequence-based approaches have been developed to elucidate the transcriptome by directly determining the transcript sequence. Initially, the generation of expressed sequence tag (EST) libraries by Sanger sequencing of complementary DNA (cDNA) was used in gene expression studies, but this approach is relatively low-throughput and not ideal for quantifying transcripts (Adams et al. 1991, 1995; Itoh et al. 1994). To overcome these technical constraints, tag-based methods such as serial analysis of gene expression (SAGE) and cap analysis gene expression (CAGE) were developed to enable higher throughput and more precise quantification of expression levels. By quantifying the number of tagged sequences, which directly corresponded to the number of mRNA transcripts, these tag-based methods provide a distinct advantage over measuring analog-style intensities as in array-based methods (Velculescu et al. 1995; Shiraki et al. 2003). However, these assays are insensitive to measuring expression levels of splice isoforms and cannot be used for novel gene discovery. In addition, the laborious cloning of sequence tags, the high cost of automated Sanger sequencing, and the requirement for large amounts of input RNA have greatly limited its use.

The development of high-throughput next-generation sequencing (NGS) has revolutionized transcriptomics by enabling RNA analysis through the sequencing of complementary DNA (cDNA) (Wang et al. 2009). This method, termed RNA sequencing (RNA-Seq), has distinct advantages over previous approaches and has revolutionized our understanding of the complex different physiological and pathological conditions. In this article we will provide an introduction to RNA sequencing and analysis using next-generation sequencing methods and discusses how to apply these advances for more comprehensive and detailed transcriptome analyses.

Transcriptome Sequencing

The introduction of high-throughput next-generation sequencing (NGS) technologies revolutionized transcriptomics. This technological development eliminated many challenges posed by hybridization-based microarrays and Sanger sequencing-based approaches that were previously used for measuring gene expression. A typical RNA-Seq experiment consists of isolating RNA, converting it to complementary DNA (cDNA), preparing the sequencing library, and sequencing it on an NGS platform (Fig. 1). However, many experimental details, dependent on a researcher's objectives, should be considered before performing RNA-Seq. These include the use of biological and technical replicates, depth of sequencing, and desired coverage across the transcriptome. In some cases, these experimental options will have minimal impact on the quality of the data. However, in many cases the researcher must carefully design the experiment, placing a priority on the balance between high-quality results and the time and monetary investment.

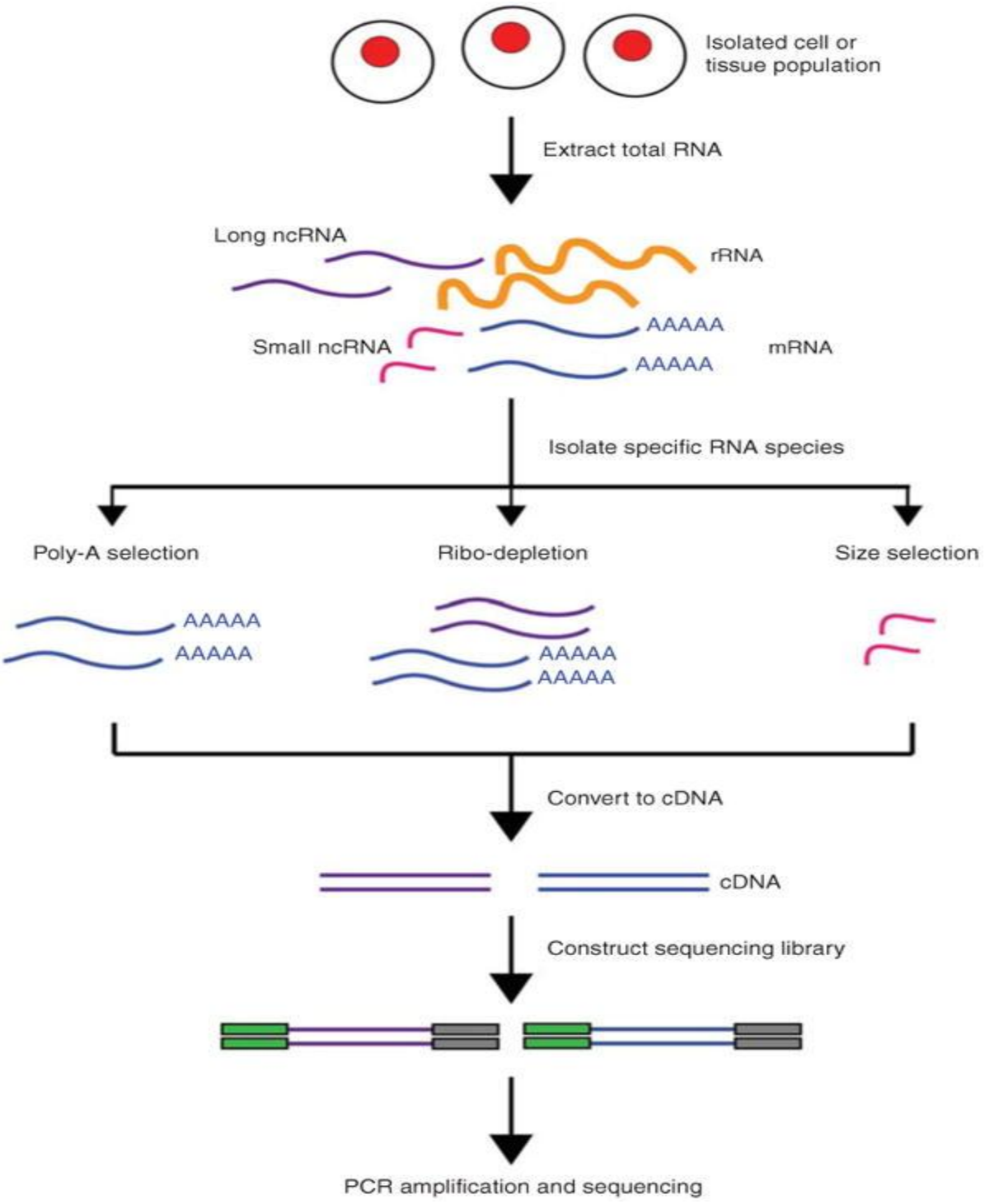


Figure: Overview of RNA-Seq. First, RNA is extracted from the biological material of choice (e.g., cells, tissues). Second, subsets of RNA molecules are isolated using a specific protocol, such as the poly-A selection protocol to enrich for polyadenylated transcripts or a ribo-depletion protocol to remove ribosomal RNAs. Next, the RNA is converted to complementary DNA (cDNA) by reverse transcription and sequencing adaptors are ligated to the ends of the cDNA fragments. Following amplification by PCR, the RNA-Seq library is ready for sequencing.

Isolation of RNA

The first step in transcriptome sequencing is the isolation of RNA from a biological sample. To ensure a successful RNA-Seq experiment, the RNA should be of sufficient quality to produce a library for sequencing. The quality of RNA is typically measured using an Agilent Bioanalyzer, which produces an RNA Integrity Number (RIN) between 1 and 10 with 10 being the highest quality samples showing the least degradation. The RIN estimates sample integrity using gel electrophoresis and analysis of the ratios of 28S to 18S ribosomal bands. Note that the RIN measures are based on mammalian organisms and certain species with abnormal ribosomal ratios (i.e., insects) may erroneously generate poor RIN numbers. Low-quality RNA (RIN < 6) can substantially affect the sequencing results (e.g., uneven gene coverage, 3′–5′ transcript bias, etc.) and lead to erroneous biological conclusions. Therefore, high-quality RNA is essential for successful RNA-Seq experiments. Unfortunately, high-quality RNA samples may not be available in some cases, such as human autopsy samples or paraffin embedded tissues, and the effect of degraded RNA on the sequencing results should be carefully considered (Tomita et al. 2004; Thompson et al. 2007; Rudloff et al. 2010).

Library Preparation Methods

Following RNA isolation, the next step in transcriptome sequencing is the creation of an RNA-Seq library, which can vary by the selection of RNA species and between NGS platforms. The construction of sequencing libraries principally involves isolating the desired RNA molecules, reverse-transcribing the RNA to cDNA, fragmenting or amplifying randomly primed cDNA molecules, and ligating sequencing adaptors. Within these basic steps, there are several choices in library construction and experimental design that must be carefully made depending on the specific needs of the researcher. Additionally, the accuracy of detection for specific types of RNAs is largely dependent on the nature of the library construction. Although there are a few basic steps for preparing RNA-Seq libraries, each stage can be manipulated to enhance the detection of certain transcripts while limiting the ability to detect other transcripts.

Selection of RNA Species

Before constructing RNA-Seq libraries, one must choose an appropriate library preparation protocol that will enrich or deplete a “total” RNA sample for particular RNA species. The total RNA pool includes ribosomal RNA (rRNA), precursor messenger RNA (pre-mRNA), mRNA, and various classes of noncoding RNA (ncRNA). In most cell types, the majority of RNA molecules are rRNA, typically accounting for over 95% of the total cellular RNA. If the rRNA transcripts are not removed before library construction, they will consume the bulk of the sequencing reads, reducing the overall depth of sequence

coverage and thus limiting the detection of other less-abundant RNAs. Because the efficient removal of rRNA is critical for successful transcriptome profiling, many protocols focus on enriching for mRNA molecules before library construction by selecting for polyadenylated (poly-A) RNAs. In this approach, the 3′ poly-A tail of mRNA molecules is targeted using poly-T oligos that are covalently attached to a given substrate (e.g., magnetic beads). Alternatively, researchers can selectively deplete rRNA using commercially available kits, such as RiboMinus (Life Technologies) or RiboZero (Epicentre). This latter method facilitates the accurate quantification of noncoding RNA species, which may be polyadenylated and thus excluded from poly-A libraries. Lastly, highly abundant RNA can be removed by denaturing and re-annealing double-stranded cDNA in the presence of duplex-specific nucleases that preferentially digest the most abundant species, which re-anneal as double-stranded molecules more rapidly than less-abundant molecules (Christodoulou et al. 2011). This method can also be used to remove other highly abundant mRNA transcripts in samples, such as hemoglobin in whole blood, immunoglobulins in mature B cells, and insulin in pancreatic beta cells.

A comprehensive understanding of the technical biases and limitations surrounding each methodological approach is essential for selecting the best method for library preparation. For example, poly-A libraries are the superior choice if one is solely interested in coding RNA molecules. Conversely, ribo-depletion libraries are a more appropriate choice for accurately quantifying noncoding RNA as well as pre-mRNA that has not been posttranscriptionally modified. Furthermore, moderate differences exist between ribo-depletion protocols, such as the efficiency of rRNA removal and differential coverage of small genes, which should be investigated before selecting a method (Huang et al. 2011).

In addition to the selective depletion of specific RNA species, new approaches have been developed to selectively enrich for regions of interest. These approaches include methods employing PCR-based approaches, hybrid capture, in-solution capture, and molecular inversion probes (Querfurth et al. 2012). The hybridization-based in solution capture involves a set of biotinylated RNA baits transcribed from DNA template oligo libraries that contain sequences corresponding to particular genes of interest. The RNA baits are combined with the RNA-Seq library where they hybridize to RNA sequences that are complementary to the baits, and the bounded complexes are recovered using streptavidin-coated beads. The resulting RNA-Seq library is now enriched for sequences corresponding to the baits and yet retains its gene expression information despite the removal of other RNA species (Levin et al. 2009). The approach enables researchers to reduce sequencing costs by sequencing selected regions in a greater number of samples.

Selection of Small RNA Species

Complementing the library preparation protocols discussed above, more specific protocols have been developed to selectively target small RNA species, which are key regulators of gene expression. Small RNA species include microRNA (miRNA), small interfering RNA (siRNA), and piwi-interacting RNA (piRNA). Because small RNAs are lowly abundant, short in length (15–30 nt), and lack polyadenylation, a separate strategy is often preferred to profile these RNA species (Morin et al. 2010). Similar to total RNA isolation, commercially available extraction kits have been developed to isolate small RNA species. Most kits involve isolation of small RNAs by size fractionation using gel electrophoresis. Size fractionation of small RNAs requires involves running the total RNA on a gel, cutting a gel slice in the 14–30 nucleotide region, and purifying the gel slice. For higher concentrations of small RNAs, the excised gel slice can be concentrated by ethanol precipitation. An alternative to gel electrophoresis is the use of silica spin columns, which bind and elute small RNAs from a silica column. After isolation of small RNAs species from total RNA, the RNA is ready for cDNA synthesis and primer ligation.

cDNA Synthesis

Universal to all RNA-Seq preparation methods is the conversion of RNA into cDNA because most sequencing technologies require DNA libraries. Most protocols for cDNA synthesis create libraries that were uniformly derived from each cDNA strand, thus representing the parent mRNA strand and its complement. In this conventional approach, the strand orientation of the original RNA is lost as the sequencing reads derived from each cDNA strand are indistinguishable in an effort to maximize efficiency of reverse transcription. However, strand information can be particularly valuable for distinguishing overlapping transcripts on opposite strands, which is critical for de novo transcript discovery (Parkhomchuk et al. 2009; Vivancos et al. 2010; Mills et al. 2013). Therefore, alternative library preparation protocols have since been developed that yield strand-specific reads. One strategy to preserve strand information is to ligate adapters in predetermined directions to single-stranded RNA or the first-strand of cDNA (Lister et al. 2008). Unfortunately, this approach is laborious and results in coverage bias at both the 5′ and 3′ ends of cDNA molecules. The preferred strategy to preserve strandedness is to incorporate a chemical label such as deoxy-UTP (dUTP) during synthesis of the second-strand cDNA that can be specifically removed by enzymatic digestion (Parkhomchuk et al. 2009). During library construction, this facilitates distinguishing the second-strand cDNA from the first strand. Although this approach is favored, the validity of antisense transcripts near highly expressed genes should be measured with caution because a small amount of reads (~1%) have been observed from the opposite strand (Zeng and Mortazavi 2012).

Multiplexing

Another consideration for constructing cost-effective RNA-Seq libraries is assaying multiple indexed samples in a single sequencing lane. The large number of reads that can be generated per sequencing run (e.g., a single lane of an Illumina HiSeq 2500 generates up to 750 million paired-end reads) permits the analysis of increasingly complex samples. However, increasingly high sequencing depths provide diminishing returns for lower complexity samples, resulting in oversampling with minimal improvement in data quality (Smith et al. 2010). Therefore, an affordable and efficient solution is to introduce unique 6-bp indices, also known as "barcodes," to each RNA-Seq library. This enables the pooling and sequencing of multiple samples in the same sequencing reaction because the barcodes identify which sample the read originated from. Depending on the application, adequate transcriptome coverage can be attained for 2–20 samples (Birney et al. 2007; Blencowe et al. 2009). To detect transcripts of moderate to high abundance, ~30–40 million reads are required to accurately quantify gene expression. To obtain coverage over the full-sequence diversity of complex transcript libraries, including rare and lowly-expressed transcripts, up to 500 million reads is required (Fu et al. 2014). As such, for any given study it is important to consider the level of sequencing depth required to answer experimental questions with confidence while efficiently using NGS resources.

Quantitative Standards

Although RNA-Seq is a widely used technique for transcriptome profiling, the rapid development of sequencing technologies and methods raises questions about the performance of different platforms and protocols. Variation in RNA-Seq data can be attributed to an assortment of factors, ranging from the NGS platform used to the quality of input RNA to the individual performing the experiment. To control for these sources of technical variability, many laboratories use positive controls or "spike-ins" for sequencing libraries. The External RNA Controls Consortium (ERCC) developed a set of universal RNA synthetic spike-in standards for microarray and RNA-Seq experiments (Jiang et al. 2011; Zook et al. 2012). The spike-ins consist of a set of 96 DNA plasmids with 273–2022 bp standard sequences inserted into a vector of ~2800 bp. The spike-in standard sequences are added to sequencing libraries at different concentrations to assess coverage, quantification, and sensitivity. These RNA standards serve as an effective quality control tool for separating technical variability from biological variability detected in differential transcriptome profiling studies.

Selection of Tissue or Cell Populations

When beginning an RNA-Seq experiment, one of the initial considerations is the choice of biological material to be used for library construction and sequencing. This choice is not trivial considering there are hundreds of cell types in over 200 different tissues that make up greater than 50 unique organs in humans alone. In addition to spatial (e.g., cell- and tissue-type) specificity, gene expression shows temporal specificity, such that different developmental stages will show unique expression signatures. Ultimately, the biological material chosen will be dependent on both the experimental goals and feasibility. For example, the tissue of choice for an investigation of unique gene expression signatures in colon cancer, the tissue choice is clear. However, for research studies investigating variation in gene expression across individuals in a population, the choice of biological material is less apparent and will likely depend on the feasibility of obtaining the biological samples (e.g., blood draws are less invasive and easier to perform than tissue biopsies).

Handling Tissue Heterogeneity

Another consideration when selecting the biological source of RNA is the heterogeneity of tissues. The accuracy of gene expression quantification is dependent on the purity of samples. In fact, the heterogeneity can substantially impact estimations of transcript abundances in samples composed of multiple cell types. Most tissue samples isolated from the human body are heterogeneous by nature. Furthermore, pathological tissue samples are often composed of disease-state cells surrounded by normal cells. To isolate distinct cell types, experimental methods have been developed, including laser-capture microdissection and cell purification. Laser-capture microdissection enables the isolation of cell types that are morphologically distinguishable under direct microscopic visualization (Emmert-Buck et al. 1996). Although this technique yields high-quality RNA, the total yield is low and requires PCR amplification, thereby introducing amplification biases and creating less distinguishable expression profiles across different cell types (Kube et al. 2007). Cell purification and enrichment protocols are also available, such as differential centrifugation and fluorescence-activated cell sorting (Cantor et al. 1975). In conjunction with RNA-Seq, these experimental methods have overcome previous technical limitations and enable researchers to uncover unique expression signatures across specific cell-types and developmental stages (Moran et al. 2012; Nica et al. 2013). In addition to these experimental methods, in silico probabilistic models can be applied in downstream analysis to differentiate the transcript abundances of distinct cells from RNA-Seq data of heterogeneous tissue samples (Erkkila et al. 2010; Li and Xie 2013). Interestingly, in some cases, the sample heterogeneity can have advantages in transcriptome profiling by identifying novel pathways, implicating cellular origins of disease, or identifying previously unknown pathological sites (Alizadeh et al. 2000; Khan et al. 2001; Sorlie et al. 2001).

Single-Cell Transcriptomics

Beyond tissue heterogeneity, considerable evidence indicates that cell-to-cell variability in gene expression is ubiquitous, even within phenotypically homogeneous cell populations (Huang 2009). Unfortunately, conventional RNA-Seq studies do not capture the transcriptomic composition of individual cells. The transcriptome of a single cell is highly dynamic, reflecting its functionality and responses to ever-changing stimuli. In addition to cellular heterogeneity resulting from regulation, individual cells show transcriptional "noise" that arises from the kinetics of mRNA synthesis and decay (Yang et al. 2003; Sun et al. 2012). Furthermore, genes that show mutually exclusive expression in individual cells may be observed as genes showing co-expression in expression analyses of bulk cell populations.

To uncover cell-to-cell variation within populations, significant efforts have been invested in developing single-cell RNA-Seq methods. The biggest challenge has been extending the limits of library preparation to accommodate extremely low input RNA. A human cell contains <1 pg of mRNA (Kawasaki 2004), whereas most sequencing protocols such as Illumina's TruSeq RNA-Seq kit recommends 400 ng to 1 μg of input RNA material. Various single-cell RNA amplification methods have been developed to accommodate less input RNA (Tang et al. 2009, 2010; Hashimshony et al. 2012; Islam et al. 2012; Picelli et al. 2013; Sasagawa et al. 2013; Shalek et al. 2013). The key limiting factors in the detection of transcripts in single cells are cDNA synthesis and PCR amplification. The efficiency of RNA-to-cDNA conversion is imperfect, estimated to be as low as 5%–25% of all transcripts (Islam et al. 2012). In addition, PCR amplification methods do not linearly amplify transcript and are prone to introduce biases based on the nucleic acid composition of different transcripts, ultimately altering the relative abundance of these transcripts in the sequencing library. Methods that avoid PCR amplification steps, such as CEL-Seq, through linear in vitro amplification of the transcriptome can avoid these biases (Hashimshony et al. 2012). In addition, the use of nanoliter-scale reaction volumes with microfluidic devices as opposed to microliter-scale reactions can reduce biases that arise during sample preparation (Wu et al. 2014). Although single-cell methods are still under active development, quantitative assessments of these techniques indicate that obtaining accurate transcriptome measurements by single-cell RNA-Seq is possible after accounting for technical noise (Brennecke et al. 2013; Wu et al. 2014). These methods will undoubtedly be important for uncovering oscillatory and heterogeneous gene expression within single-cell types, as well as identifying cell-specific biomarkers that further our understanding of biology across many physiological and pathological conditions.

Sequencing Platforms for Transcriptomics

When designing an RNA-Seq experiment, the selection of a sequencing platform is important and dependent on the experimental goals. Currently, several NGS platforms are commercially available and other platforms are under active technological development (Metzker 2010). The majority of high-throughput sequencing platforms use a sequencing-by-synthesis method to sequence tens of millions of sequence clusters in parallel. The NGS platforms can often be categorized as either ensemble-based (i.e. sequencing many identical copies of a DNA molecule) or single-molecule-based (i.e. sequencing a single DNA molecule). The differences between these sequencing techniques and platforms can affect downstream analysis and interpretation of the sequencing data.

In recent years, the sequencing industry has been dominated by Illumina, which applies an ensemble-based sequencing-by-synthesis approach (Bentley et al. 2008). Using fluorescently labeled reversible-terminator nucleotides, DNA molecules are clonally amplified while immobilized on the surface of a glass flowcell. Because molecules are clonally amplified, this approach provides the relative RNA expression levels of genes. To remove potential PCR-amplification biases, PCR controls and specific steps in the downstream computational analysis are required. One major benefit of ensemble-based platforms is low sequencing error rates (<1%) dominated by single mismatches. Low error rates are particularly important for sequencing miRNAs, whose relatively small sizes result in misalignment or loss of reads if error rates are too high. Currently, the Illumina HiSeq platform is the most commonly applied next-generation sequencing technology for RNA-Seq and has set the standard for NGS sequencing. The platform has two flow cells, each providing eight separate lanes for sequencing reactions to occur. The sequencing reactions can take between 1.5 and 12 d to complete, depending on the total read length of the library. Even more recently, Illumina released the MiSeq, a desktop sequencer with lower throughput but faster turnaround (generates ~30 million paired-end reads in 24 h). The simplified workflow of the MiSeq instrument offers rapid turnaround time for transcriptome sequencing on a smaller scale.

Single-molecule-based platforms such as PacBio enable single-molecule real-time (SMRT) sequencing (Eid et al. 2009). This approach uses DNA polymerase to perform uninterrupted template-directed synthesis using fluorescently labeled nucleosides. As each base is enzymatically incorporated into a growing DNA strand, a distinctive pulse of fluorescence is detected in real-time by zero-mode waveguide nanostructure arrays. An advantage of SMRT is that it does not include a PCR amplification step, thereby avoiding amplification bias and improving uniform coverage across the transcriptome. Another advantage of this sequencing approach is the ability to produce extraordinarily long reads with average lengths of 4200 to 8500 bp, which greatly improves the detection of novel transcript structures (Au et al. 2013; Sharon et al. 2013). A critical

disadvantage of SMRT is a high rate of errors (~5%) that are predominately characterized by insertions and deletions (Carneiro et al. 2012); the high error rate results in misalignment and loss of sequencing reads due to the difficulty of matching erroneous reads to the reference genome.

Another important consideration for choosing a sequencing platform is transcriptome assembly. Transcriptome assembly, which is discussed in greater detail later, is necessary to transform a collection of short sequencing reads into a set of full-length transcripts. In general, longer sequencing reads make it simpler to accurately and unambiguously assemble transcripts, as well as identify splicing isoforms. The extremely long reads generated by the PacBio platform are ideal for de novo transcriptome assembly in which the reads are not aligned to a reference transcriptome. The longer reads will facilitate an accurate detection of alternative splice isoforms, which may not be discovered with shorter reads. Moleculo, a company acquired by Illumina, has developed long-read sequencing technology capable of producing 8500 bp reads. Although it has yet to be widely adopted for transcriptome sequencing, the long reads aid transcriptome assembly. Lastly, Illumina has developed protocols for its desktops MiSeq to sequence slightly longer reads (up to 350 bp). Although much shorter than PacBio and Moleculo reads, the longer MiSeq reads can also be used to improve both de novo and reference transcriptome assembly.

Transcriptome Analysis

Gene expression profiling by RNA-Seq provides an unprecedented high-resolution view of the global transcriptional landscape. As the sequencing technologies and protocol methodologies continually evolve, new informatics challenges and applications develop. Beyond surveying gene expression levels, RNA-Seq can also be applied to discover novel gene structures, alternatively spliced isoforms, and allele-specific expression (ASE). In addition, genetic studies of gene expression using RNA-Seq have observed genetically correlated variability in expression, splicing, and ASE (Montgomery et al. 2010; Pickrell et al. 2010; Battle et al. 2013; Lappalainen et al. 2013). This section will introduce how expression data are analyzed to provide greater insight into the extensive complexity of transcriptomes.

RNA-Sequencing Data Analysis Workflow

The conventional pipeline for RNA-Seq data includes generating FASTQ-format files contains reads sequenced from an NGS platform, aligning these reads to an annotated reference genome, and quantifying expression of genes (Fig. 2). Although basic sequencing analysis tools are more accessible than ever, RNA-Seq analysis presents

unique computational challenges not encountered in other sequencing-based analyses and requires specific consideration to the biases inherent in expression data.

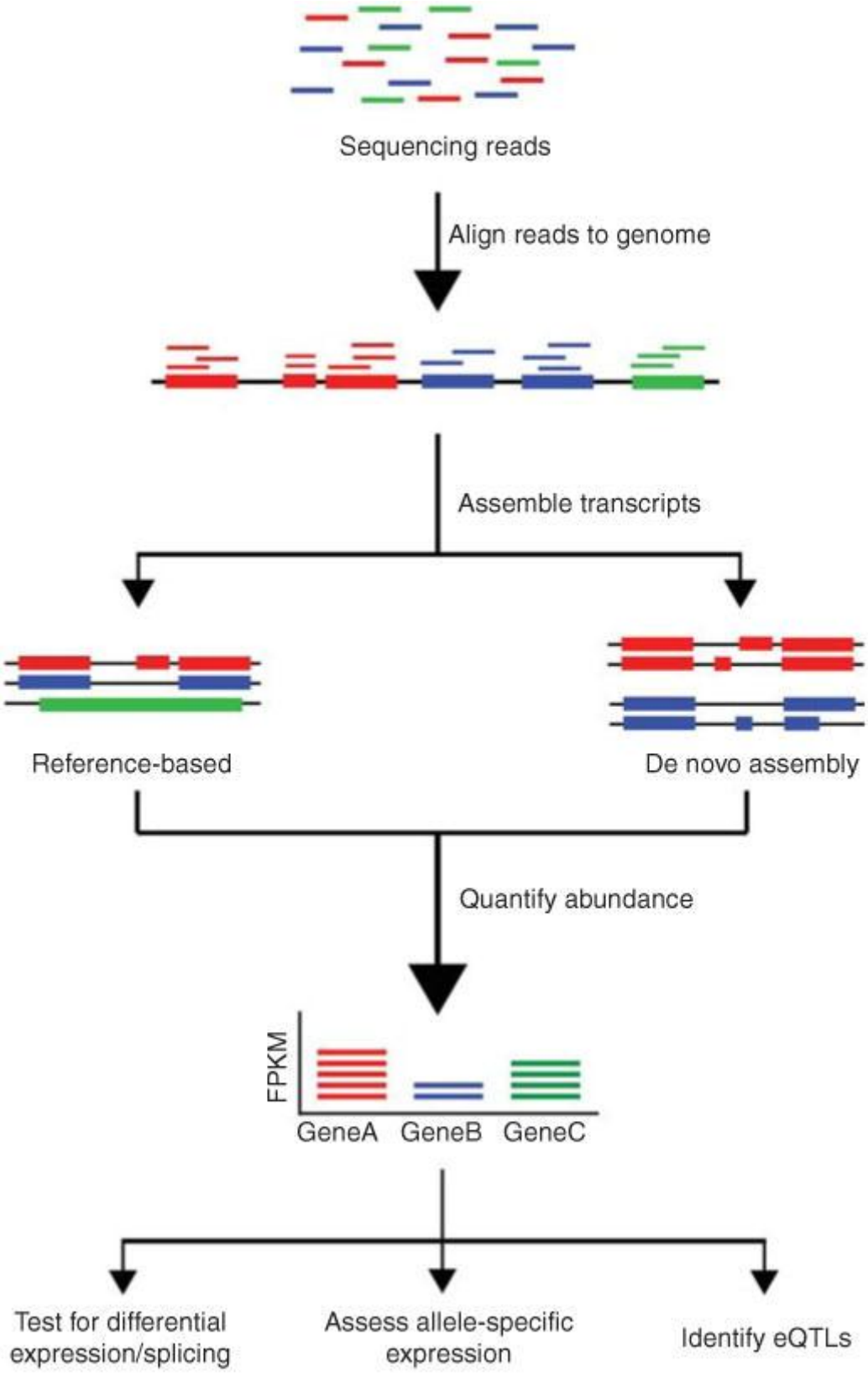


Figure: Overview of RNA-Seq data analysis. Following typical RNA-Seq experiments, reads are first aligned to a reference genome. Second, the reads may be assembled into transcripts using reference transcript annotations or de novo assembly approaches. Next, the expression level of each gene is estimated by counting the number of reads that align to each exon or full-length transcript. Downstream analyses with RNA-Seq data include testing for differential expression between samples, detecting allele-specific expression, and identifying expression quantitative trait loci (eQTLs).

Read Alignment

Mapping RNA-Seq reads to the genome is considerably more challenging than mapping DNA sequencing reads because many reads map across splice junctions. In fact, conventional read mapping algorithms, such as Bowtie (Langmead et al. 2009) and BWA (Li and Durbin 2009), are not recommended for mapping RNA-Seq reads to the reference genome because of their inability to handle spliced transcripts. One approach to resolving this problem is to supplement the reference genome with sequences derived from exon–exon splice junctions acquired from known gene annotations (Mortazavi et al. 2008). A preferred strategy is to map reads with a "splicing-aware" aligner that can recognize the difference between a read aligning across an exon–intron boundary and a read with a short insertion. As RNA-Seq data have become more widely used, a number of splicing-aware mapping tools have been developed specifically for mapping transcriptome data. The more commonly used RNA-Seq alignment tools include GSNAP (Wu and Nacu 2010), MapSplice (Wang et al. 2010a), RUM (Grant et al. 2011), STAR (Dobin et al. 2013), and TopHat (Trapnell et al. 2009) (Table 2). Each aligner has different advantages in terms of performance, speed, and memory utilization. Selecting the best aligner to use depends on these metrics and the overall objectives of the RNA-Seq study. Efforts to systematically evaluate the performance of RNA-Seq aligners have been initiated by GENCODE's RNA-Seq Genome Annotation Assessment Project3 (RGASP3), which has found major performance difference between alignments tools on numerous benchmarks, including alignment yield, basewise accuracy, mismatch and gap placement, and exon junction discovery (Engstrom et al. 2013).

Transcript Assembly and Quantification

After RNA-Seq reads are aligned, the mapped reads can be assembled into transcripts. The majority of computational programs infer transcript models from the accumulation of read alignments to the reference genome (Trapnell et al. 2010; Li et al. 2011; Roberts et al. 2011a; Mezlini et al. 2013) (Table 2). An alternative approach for transcript assembly is de novo reconstruction, in which contiguous transcript sequences are assembled with the use of a reference genome or annotations (Robertson et al.

2010; Grabherr et al. 2011; Schulz et al. 2012). The reconstruction of transcripts from short-read data is a major challenge and a gold standard method for transcript assembly does not exist. The nature of the transcriptome (e.g., gene complexity, degree of polymorphisms, alternative splicing, dynamic range of expression), common technological challenges (e.g., sequencing errors), and features of the bioinformatics workflow (e.g., gene annotation, inference of isoforms) can substantially affect transcriptome assembly quality. RGASP3 has initiated efforts to evaluate computational methods for transcriptome reconstruction and has found that most algorithms can identify discrete transcript components, but the assembly of complete transcript structures remains a major challenge (Steijger et al. 2013).

A common downstream feature of transcript reconstruction software is the estimation of gene expression levels. Computational tools such as Cufflinks (Trapnell et al. 2010), FluxCapacitor (Montgomery et al. 2010; Griebel et al. 2012), and MISO (Katz et al. 2010), quantify expression by counting the number of reads that map to full-length transcripts (Table 2). Alternative approaches, such as HTSeq, can quantify expression without assembling transcripts by counting the number of reads that map to an exon (Anders et al. 2013). To accurately estimate gene expression, read counts must be normalized to correct for systematic variability, such as library fragment size, sequence composition bias, and read depth (Oshlack and Wakefield 2009; Roberts et al. 2011b). To account for these sources of variability, the reads per kilobase of transcripts per million mapped reads (RPKM) metric normalizes a transcript's read count by both the gene length and the total number of mapped reads in the sample. For paired end-reads, a metric that normalizes for sources of variances in transcript quantification is the paired fragments per kilobase of transcript per million mapped reads (FPKM) metric, which accounts for the dependency between paired-end reads in the RPKM estimate (Trapnell et al. 2010). Another technical challenge for transcript quantification is the mapping of reads to multiple transcripts that are a result of genes with multiple isoforms or close paralogs. One solution to correct for this “read assignment uncertainty” is to exclude all reads that do not map uniquely, as in Alexa-Seq (Griffith et al. 2010). However, this strategy is far from ideal for genes lacking unique exons. An alternative strategy used by Cufflinks (Trapnell et al. 2012), and MISO (Katz et al. 2010) is to construct a likelihood function that models the sequencing experiment and estimates the maximum likelihood that a read maps to a particular isoform.

Considerations for miRNA Sequencing Analysis

The general approach for analysis of miRNA sequencing data is similar to approaches discussed for mRNA. To identify known miRNAs, the sequencing reads can be mapped to a specific database, such as miRBase, a repository containing over 24,500 miRNA loci from 206 species in its latest release (v21) in June 2014 (Kozomara and Griffiths-Jones 2014). In addition, several tools have been developed to facilitate analysis of miRNAs

including the commonly used tools miRanalyzer (Hackenberg et al. 2011) and miRDeep (An et al. 2013). MiRanalyzer can detect known miRNAs annotated on miRBase as well as predict novel miRNAs using a machine-learning approach based on the random forest method with a broad range of features. Similarly, miRDeep is able to identify known miRNAs and predict novel miRNAs using properties of miRNA biogenesis to score the compatibility of the position and frequency of sequenced RNA from the secondary structure of precursor miRNAs. Although miRDeep and miRanalyzer contain modules for target prediction, expression quantification, and differential expression, the methods developed for mRNA quantification and differential expression can also be applied to miRNA data (Eminaga et al. 2013).

Quality Assessment and Technical Considerations

At each stage in the RNA-Seq analysis pipeline, careful consideration should be applied to identifying and correcting for various sources of bias. Bias can arise throughout the RNA-Seq experimental pipeline, including during RNA extraction, sample preparation, library construction, sequencing, and read mapping (Kleinman and Majewski 2012; Lin et al. 2012; Pickrell et al. 2012; 't Hoen et al. 2013). First, the quality of the raw sequence data in FASTQ-format files should be evaluated to ensure high-quality reads. User-friendly software tools designed to generate quality overviews include the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit), the FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), and the RobiNA package (Lohse et al. 2012). Several important parameters that should be evaluated include the sequence diversity of reads, adaptor contamination, base qualities, nucleotide composition, and percentage of called bases. These technical artifacts can arise at the sequencing stage or during the construction of the RNA-Seq. For example, the 5′ read end, derived from either end of a double-stranded cDNA fragment, shows higher error rate due to mispriming events introduced by the random oligos during the RNA-Seq library construction protocol (Lin et al. 2012). If possible, actions to correct for these biases should be performed, such as trimming the ends of reads, to expedite the speed and improve the quality of the read alignments.

After aligning the reads, additional parameters should be assessed to account for biases that arise at the read mapping stage. These parameters include the percentage of reads mapped to the transcriptome, the percentage of reads with a mapped mate pair, the coverage bias at the 5′- and 3′-ends, and the chromosomal distribution of reads. One of the most common sources of mapping errors for RNA-Seq data occurs when a read spans the splicing junction of an alternatively spliced gene. A misalignment can be easily introduced due to ambiguous mapping of the read end to one of the two (or more) possible exons and is especially common when reads are mapped to a reference transcriptome that contains an incomplete annotation of isoforms (Kleinman and Majewski 2012; Pickrell et al. 2012). If genotype information is available, the integrity of

the samples should also be evaluated by investigating the correlation of single-nucleotide variants (SNVs) between the DNA and RNA reads ('t Hoen et al. 2013). The concordance between the DNA and RNA sequencing data may provide insight into sample swaps or sample mixtures caused accidentally as a result of personnel or equipment error. In the case of a swapped sample, more discordant variants would be observed between the DNA and RNA sequencing data. In the case of a mixture of samples, more significant patterns of allele-specific expression would be observed than expected for a single individual as a result of more combinations of heterozygous and homozygous sites that would skew the alleles beyond the expected 1:1 allelic ratio.

Differential Gene Expression

A primary objective of many gene expression experiments is to detect transcripts showing differential expression across various conditions. Extensive statistical approaches have been developed to test for differential expression with microarray data, where the continuous probe intensities across replicates can be approximated by a normal distribution (Cui and Churchill 2003; Smyth 2004; Grant et al. 2005). Although in principle these approaches are also applicable to RNA-Seq data, different statistical models must be considered for discrete read counts that do not fit a normal distribution. Early RNA-Seq studies suggested that the distribution of read counts across replicates fit a Poisson distribution, which formed the basis for modeling RNA-Seq count data (Marioni et al. 2008). However, further studies indicated that biological variability is not captured by the Poisson assumption, resulting in high false-positive rates due to underestimation of sampling error (Anders and Huber 2010; Langmead et al. 2010; Robinson and Oshlack 2010). Hence, negative binomial distribution models that take into account overdispersion or extra-Poisson variation have been shown to best fit the distribution of read counts across biological replicates.

To model the count-based nature of RNA-Seq data, complex statistical models have been developed to handle sources of variability that model overdispersion across technical and biological replicates. One source of variability is differences in sequencing read depth, which can artificially create differences between samples. For instance, differences in read depth will result in the samples appearing more divergent if raw read counts between genes are compared. To correct for this, it is advantageous to transform raw read count data to FPKM or RPKM values in differential expression analyses. Although this correction metric is commonly used in place of read counts, the presence of several highly expressed genes in a particular sample can significantly alter the RPKM and FPKM values. For example, a highly expressed gene can “absorb” many reads, consequently repressing the read counts for other genes and artificially inflating gene expression variation. To account for this bias, several statistical models have been proposed that use the highly expressed genes as model covariates (Robinson and Oshlack 2010). Another source of variability that has been observed is that the

distribution of sequencing reads is unequal across genes. Therefore, a two-parameter generalized Poisson model that simultaneously considers read depth and sequencing bias as independent parameters was developed and shown to improve RNA-Seq analysis (Srivastava and Chen 2010). More complex normalization methods have also been developed to account for hidden covariates without removing significant biological variability. For example, the probabilistic estimation of expression residuals (PEER) framework (Stegle et al. 2012) and the hidden covariates with prior (HCP) framework (Mostafavi et al. 2013) are methods that use a Bayesian approach to infer hidden covariates and remove their effects from expression data.

To detect differential expression, a variety of statistical methods have been designed specifically for RNA-Seq data. A popular tool to detect differential expression is Cuffdiff, which is part of the Tuxedo suite of tools (Bowtie, Tophat, and Cufflinks) developed to analyze RNA-Seq data (Trapnell et al. 2013). In addition to Cuffdiff, several other packages support testing differential expression, including baySeq (Hardcastle and Kelly 2010), DESeq (Anders and Huber 2010), DEGseq (Wang et al. 2010b), and edgeR (Robinson et al. 2010) (Table 2). Although these packages can assign significance to differentially expressed transcripts, the biological observations should be carefully interpreted. Each model makes specific assumptions that may be violated in the context of the observed data; therefore, an understanding of the model parameters and their constraints is critical for drawing meaningful and accurate biological conclusions (Bullard et al. 2010). Furthermore, replicates in RNA-Seq experiments are crucial for measuring variability and improving estimations for the model parameters (Tarazona et al. 2011; Glaus et al. 2012). Biological replicates (e.g., cells grown on two different plates under the same conditions) are preferred to technical replicates (e.g., one RNA-Seq library sequenced on two different lanes), which show little variation. Although the number of replicates required per condition is an open research question, a minimum of three replicates per sample has been suggested (Auer and Doerge 2010). In many cases, multiplexed RNA-Seq libraries can be used to add biological replicates without increasing sequencing costs (if sequenced at a lower depth) and will greatly improve the robustness of the experimental design (Liu et al. 2014). Additionally, the accuracy of measurements of differential gene expression can be further improved by using ERCC spike-in controls to distinguish technical variation from biological variation.

Allele-Specific Expression

A major advantage of RNA-Seq is the ability to profile transcriptome dynamics at a single-nucleotide resolution. Therefore, the sequenced transcript reads can provide coverage across heterozygous sites, representing transcription from both the maternal and paternal alleles. If a sufficient number of reads cover a heterozygous site within a gene, the null hypothesis is that the ratio of maternal to paternal alleles is balanced. Significant deviation from this expectation suggests allele-specific expression (ASE).

Potential mechanisms for ASE include genetic variation (e.g., single-nucleotide polymorphism in a *cis*-regulatory region upstream of a gene) and epigenetic effects (e.g., genomic imprinting, methylation, histone modifications, etc.). Early studies showed that allele-specific differences can affect up to 30% of loci within an individual (Ge et al. 2009) and are caused by both common and rare genetic variants (Pastinen 2010). Studies have also applied ASE to identify expression modifiers of protein-coding variation (Lappalainen et al. 2011; Montgomery et al. 2011), effects of loss-of-function variation (MacArthur et al. 2012), and differences between pathogenic and healthy tissues (Tuch et al. 2010). Furthermore, ASE studies using single-cell transcriptomics have uncovered a stochastic pattern of allelic expression that may contribute to variable expressivity, a novel perspective which may have fundamental implications for variable disease penetrance and severity (Deng et al. 2014).

Conventional workflows to detect ASE involve counting reads containing each allele at heterozygous sites and applying a statistical test, such as the binomial test or the Fisher's exact test (Degner et al. 2009; Rozowsky et al. 2011; Wei and Wang 2013). However, more rigorous statistical approaches are necessary to overcome technical challenges involved in ASE detection. These challenges include read-mapping bias, sampling variance, overdispersion at extreme read depths, alternatively spliced alleles, insertions and deletions (indels), and genotyping errors. To account for overdispersion, one approach is to model allelic read counts using a beta-binomial distribution at individual loci (Sun 2012); however, accurate estimation of the overdispersion parameter requires replicates and, in our experience, major source of bias come from site-specific mapping differences. Another strategy is to use a hierarchical Bayesian model that combines information across loci, as well as across replicates and technologies, to make global and site-specific inferences for ASE (Skelly et al. 2011). To assess reference-allele mapping bias, the number of mismatches in reads containing the nonreference allele should be assessed as increased bias is observed with greater sequence divergence between alleles (Stevenson et al. 2013). To correct for read-mapping bias, an enhanced reference genome can be constructed that masks all SNP positions or includes the alternative alleles at polymorphic loci (Degner et al. 2009; Satya et al. 2012). Statistical methods to better address these technical biases are under active development and are expected to foster further improvements in ASE detection.

Expression Quantitative Trait Loci

Another prominent direction of RNA-Seq studies has been the integration of expression data with other types of biological information, such as genotyping data. The combination of RNA-Seq with genetic variation data has enabled the identification of genetic loci correlated with gene expression variation, also known as expression quantitative trait loci (eQTLs). This expression variation caused by common and rare

variants is postulated to contribute to phenotypic variation and susceptibility to complex disease across individuals (Majewski and Pastinen 2011). The goal of eQTL analysis is to identify associations that will uncover underlying biological processes, discover genetic variants causing disease, and determine causal pathways. Initial eQTL studies using RNA-Seq data identified a greater number of statistically significant eQTLs than had been identified by microarray studies (Montgomery et al. 2010; Pickrell et al. 2010). Most of the eQTLs identified directly influenced gene expression in an allele-specific manner and were located near transcriptional start sites, indicating that eQTLs could modulate expression directly, or in cis. Later studies identified *trans*-eQTLs, which are variants that affect the expression of a distant gene (>1 Mb) by modifying the activity or expression of upstream factors that regulate the gene (Fehrmann et al. 2011; Battle et al. 2013; Westra et al. 2013). Although *trans*-eQTLs show weaker effects and present validation difficulties, they can potentially reveal previously unknown pathways in gene regulation networks.

RNA-Seq has revolutionized QTL analyses because it enables association analyses of more than just gene expression levels alone. For example, RNA-Seq provides unprecedented opportunity to investigate variations in splicing by profiling alternately spliced isoforms of a gene. This has enabled the identification of variants influencing the quantitative expression of alternatively spliced isoforms commonly referred to as splicing-QTLs (sQTLs) (Lalonde et al. 2011). In addition, specific RNA-Seq library constructions (e.g., ribo-depleted) have enabled the detection of eQTLs affecting other RNA species; recent studies have identified variants affecting the expression of various ncRNAs, including long intergenic noncoding RNAs (Montgomery et al. 2010; Gamazon et al. 2012; Kumar et al. 2013; Popadin et al. 2013). The expanding potential of RNA-Seq to associate phenotypic variations with genetic variation offers an enhanced understanding of gene regulation.

Traditional eQTL mapping methods that were developed for microarray data use linear models such as linear regression and ANOVA to associate genetic variants with gene expression (Kendziorski and Wang 2006). These methods have been directly applied to RNA-Seq data following appropriate normalization of total read counts. Most eQTL studies perform separate testing for each transcript-SNP pair using linear regression and ANOVA models to detect significant association. Nonlinear approaches have also been developed to test associations, such as generalized linear and mixed models, Bayesian regression (Servin and Stephens 2007). Alternative models, such as Merlin, have also been developed to detect eQTLs from expression data that include related individuals using pedigree data (Abecasis et al. 2002). In addition, several methods have been developed to simultaneously test the effect of multiple SNPs on the expression of a single gene using Bayesian methods (Lee et al. 2008). To further improve on the detection of causal regulatory variants, several studies have integrated ASE information with eQTL analysis. These studies showed that genetic variants showing allele-specific

effects and identified as eQTLs show higher enrichment in functional annotations and provide stronger evidence of *cis*-regulatory impact (Battle et al. 2013; Lappalainen et al. 2013; Sun and Hu 2013). Because high-throughput sequencing has created genotype data sets featuring millions of SNPs and expression data sets featuring tens of thousands of transcripts, the task of testing billions of transcript-SNP pairs in eQTL analysis can be computationally intensive. To mitigate this computational burden, software has been developed such as Matrix eQTL to efficiently test the associations by modeling the effect of genotype as either additive linear (least squares model) or categorical (ANOVA model) (Shabalin 2012). Because of the large number of tests performed, it is important to correct for multiple-testing by calculating the false discovery rate (Benjamini and Hochberg 1995; Yekutieli and Benjamini 1999) or resampling using bootstrap or permutation procedures (Karlsson 2006; Zhang et al. 2012).

However, the design and interpretation of eQTL studies is not straightforward. Many complications result from the complexity of gene regulation, which shows both spatial (cell and tissue location) specificity as well as temporal (developmental stage) specificity. For instance, several studies have performed eQTL analysis across multiple tissues, indicating that genetic regulatory elements can have tissue-specific effects (Petretto et al. 2006; Schadt et al. 2008; Dimas et al. 2009; Kwan et al. 2009; Grundberg et al. 2012; Flutre et al. 2013). Therefore, future eQTL analyses should test for SNP-transcript associations in well-defined cell types that are relevant to the trait of interest (Lonsdale et al. 2013). For example, a study detecting eQTLs in cardiovascular disease should use heart tissue while a study interested in autoimmune disease should use whole blood. Another major consideration for eQTL studies is accounting for population structure and elucidating the causal variants (Stranger et al. 2012). The structure of genomic variation can vary significantly between populations and will influence the resolution of any genetic association study (Frazer et al. 2007; Altshuler et al. 2010). Furthermore, if substantial linkage disequilibrium (LD) exists within the genome, the associated genetic variant is often "tagging" the causal variant rather than acting as the causal regulatory variant itself. As eQTL studies integrate data across different populations and use population-scale genome sequencing, the ability to elucidate causal variants will greatly improve (Montgomery et al. 2010; Lappalainen et al. 2013).

Probable Questions:

1. Define Transcriptome. How transcriptome is analyzed?
2. What are the significance of transcriptome?
3. Describe principles of Microarray.
4. Differentiate cDNA based microarray and Oligonucleotide based microarrays
5. What are the applications of Microarray?
6. Describe spotted microarray with suitable diagram.
7. Name some online tools used in microarray data analysis.
8. Describe the steps of DNA microarray with suitable diagram.
9. What are the applications of DNA Microarray?

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

UNIT-XII

Protein sequencing: Protein structure analysis,protein-protein interaction, Protein-DNA interaction

Objective: In this unit we will discuss how protein sequencing is performed. We will also discuss various analytical tools by which protein-protein and protein-DNA interaction are studied.

Protein sequencing is the practical process of determining the amino acid sequence of all or part of a protein or peptide. This may serve to identify the protein or characterize its post-translational modifications. Typically, partial sequencing of a protein provides sufficient information (one or more sequence tags) to identify it with reference to databases of protein sequences derived from the conceptual translation of genes.

The two major direct methods of protein sequencing are mass spectrometry and Edman degradation using a protein sequenator (sequencer). Mass spectrometry methods are now the most widely used for protein sequencing and identification but Edman degradation remains a valuable tool for characterizing a protein's *N*-terminus.

Determining amino acid composition

It is often desirable to know the unordered amino acid composition of a protein prior to attempting to find the ordered sequence, as this knowledge can be used to facilitate the discovery of errors in the sequencing process or to distinguish between ambiguous results. Knowledge of the frequency of certain amino acids may also be used to choose which protease to use for digestion of the protein. The misincorporation of low levels of non-standard amino acids (e.g.norleucine) into proteins may also be determined. A generalized method often referred to as *amino acid analysis*[2] for determining amino acid frequency is as follows:

1. Hydrolyse a known quantity of protein into its constituent amino acids.
2. Separate and quantify the amino acids in some way.

Hydrolysis

Hydrolysis is done by heating a sample of the protein in 6 M hydrochloric acid to 100–110 °C for 24 hours or longer. Proteins with many bulky hydrophobic groups may require longer heating periods. However, these conditions are so vigorous that some amino acids (serine, threonine, tyrosine, tryptophan, glutamine, and cysteine) are degraded. To circumvent this problem, Biochemistry Online suggests heating separate samples for different times, analysing each resulting solution, and extrapolating back to zero hydrolysis time. Rastall suggests a variety of reagents to prevent or reduce degradation, such as thiol reagents or phenol to protect tryptophan and tyrosine from attack by chlorine, and pre-oxidising cysteine. He also suggests measuring the quantity of ammonia evolved to determine the extent of amide hydrolysis.

Separation and quantitation

The amino acids can be separated by ion-exchange chromatography then derivatized to facilitate their detection. More commonly, the amino acids are derivatized then resolved by reversed phase HPLC.

An example of the ion-exchange chromatography is given by the NTRC using sulfonated polystyrene as a matrix, adding the amino acids in acid solution and passing a buffer of steadily increasing pH through the column. Amino acids are eluted when the pH reaches their respective isoelectric points. Once the amino acids have been separated, their respective quantities are determined by adding a reagent that will form a coloured derivative. If the amounts of amino acids are in excess of 10 nmol, ninhydrin can be used for this; it gives a yellow colour when reacted with proline, and a vivid purple with other amino acids. The concentration of amino acid is proportional to the absorbance of the resulting solution. With very small quantities, down to 10 pmol, fluorescent derivatives can be formed using reagents such as ortho-phthaldehyde (OPA) or fluorescamine.

Pre-column derivatization may use the Edman reagent to produce a derivative that is detected by UV light. Greater sensitivity is achieved using a reagent that generates a fluorescent derivative. The derivatized amino acids are subjected to reversed phase chromatography, typically using a C8 or C18 silica column and an optimised elution gradient. The eluting amino acids are detected using a UV or fluorescence detector and the peak areas compared with those for derivatised standards in order to quantify each amino acid in the sample.

N-terminal amino acid analysis

Determining which amino acid forms the *N*-terminus of a peptide chain is useful for two reasons: to aid the ordering of individual peptide fragments' sequences into a whole chain, and because the first round of Edman degradation is often contaminated by

impurities and therefore does not give an accurate determination of the *N*-terminal amino acid. A generalised method for *N*-terminal amino acid analysis follows:

1. React the peptide with a reagent that will selectively label the terminal amino acid.
2. Hydrolyse the protein.
3. Determine the amino acid by chromatography and comparison with standards.

There are many different reagents which can be used to label terminal amino acids. They all react with amine groups and will therefore also bind to amine groups in the side chains of amino acids such as lysine - for this reason it is necessary to be careful in interpreting chromatograms to ensure that the right spot is chosen. Two of the more common reagents are **Sanger's reagent** (1-fluoro-2,4-dinitrobenzene) and dansyl derivatives such as dansyl chloride. Phenylisothiocyanate, the reagent for the Edman degradation, can also be used. The same questions apply here as in the determination of amino acid composition, with the exception that no stain is needed, as the reagents produce coloured derivatives and only qualitative analysis is required. So the amino acid does not have to be eluted from the chromatography column, just compared with a standard. Another consideration to take into account is that, since any amine groups will have reacted with the labelling reagent, ion exchange chromatography cannot be used, and thin-layer chromatography or high-pressure liquid chromatography should be used instead.

C-terminal amino acid analysis

The number of methods available for C-terminal amino acid analysis is much smaller than the number of available methods of N-terminal analysis. The most common method is to add carboxypeptidases to a solution of the protein, take samples at regular intervals, and determine the terminal amino acid by analysing a plot of amino acid concentrations against time. This method will be very useful in the case of polypeptides and protein-blocked N termini. C-terminal sequencing would greatly help in verifying the primary structures of proteins predicted from DNA sequences and to detect any posttranslational processing of gene products from known codon sequences.

Edman degradation

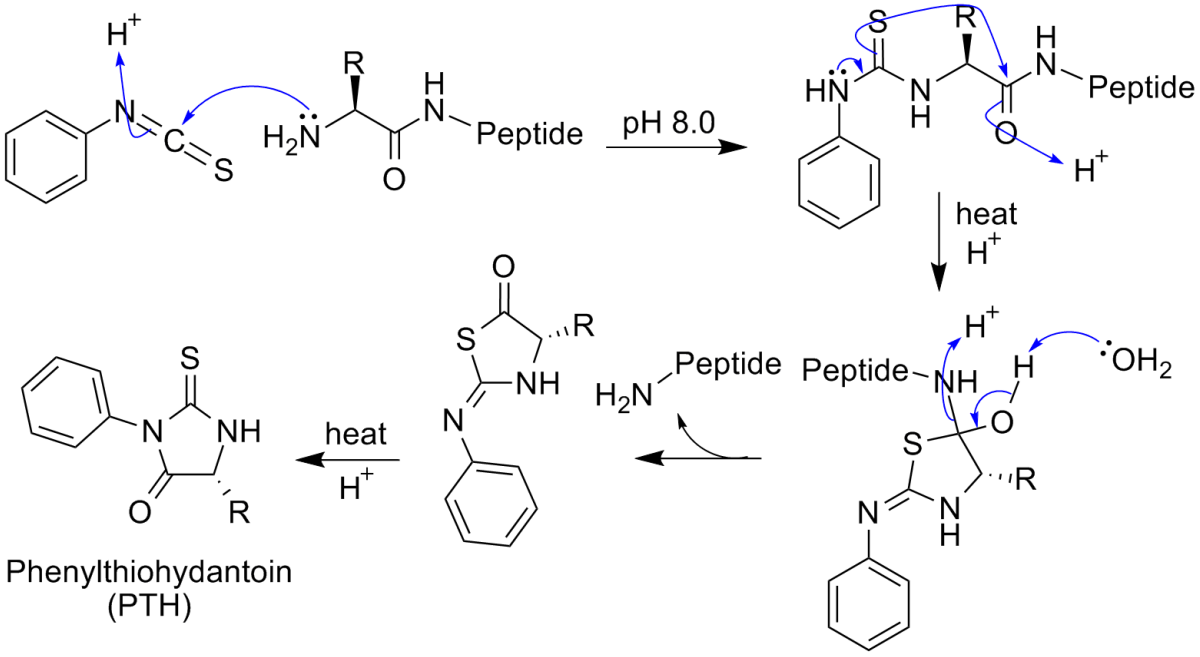
Edman degradation, developed by Pehr Edman, is a method of sequencing amino acids in a peptide.[1] In this method, the amino-terminal residue is labelled and cleaved from the peptide without disrupting the peptide bonds between other amino acid residues. The Edman degradation is a very important reaction for protein sequencing, because it allows the ordered amino acid composition of a protein to be discovered. Automated Edman sequencers are now in widespread use, and are able to sequence peptides up to approximately 50 amino acids long. A reaction scheme for sequencing a

protein by the Edman degradation follows; some of the steps are elaborated on subsequently.

1. Break any disulphide bridges in the protein with a reducing agent like 2-mercaptoethanol. A protecting group such as iodoacetic acid may be necessary to prevent the bonds from re-forming.
2. Separate and purify the individual chains of the protein complex, if there are more than one.
3. Determine the amino acid composition of each chain.
4. Determine the terminal amino acids of each chain.
5. Break each chain into fragments under 50 amino acids long.
6. Separate and purify the fragments.
7. Determine the sequence of each fragment.
8. Repeat with a different pattern of cleavage.
9. Construct the sequence of the overall protein.

Mechanism

Phenyl isothiocyanate is reacted with an uncharged N-terminal amino group, under mildly alkaline conditions, to form a cyclical *phenylthiocarbamoyl* derivative. Then, under acidic conditions, this derivative of the terminal amino acid is cleaved as a thiazolinone derivative. The thiazolinone amino acid is then selectively extracted into an organic solvent and treated with acid to form the more stable phenylthiohydantoin (PTH)- amino acid derivative that can be identified by using chromatography or electrophoresis. This procedure can then be repeated again to identify the next amino acid. A major drawback to this technique is that the peptides being sequenced in this manner cannot have more than 50 to 60 residues (and in practice, under 30). The peptide length is limited due to the cyclical derivatization not always going to completion. The derivatization problem can be resolved by cleaving large peptides into smaller peptides before proceeding with the reaction. It is able to accurately sequence up to 30 amino acids with modern machines capable of over 99% efficiency per amino acid. An advantage of the Edman degradation is that it only uses 10 - 100 pico-moles of peptide for the sequencing process. The Edman degradation reaction was automated in 1967 by Edman and Beggs to speed up the process and 100 automated devices were in use worldwide by 1973.



Limitations

Because the Edman degradation proceeds from the N-terminus of the protein, it will not work if the N-terminus has been chemically modified (e.g. by acetylation or formation of pyroglutamic acid). Sequencing will stop if a non-α-amino acid is encountered (e.g. isoaspartic acid), since the favored five-membered ring intermediate is unable to be formed. Edman degradation is generally not useful to determine the positions of disulfide bridges. It also requires peptide amounts of 1 picomole or above for discernible results.

Digestion into peptide fragments

Peptides longer than about 50–70 amino acid long cannot be sequenced reliably by the Edman degradation. Because of this, long protein chains need to be broken up into small fragments that can then be sequenced individually. Digestion is done either by endopeptidases such as trypsin or pepsin or by chemical reagents such as cyanogen bromide. Different enzymes give different cleavage patterns, and the overlap between fragments can be used to construct an overall sequence.

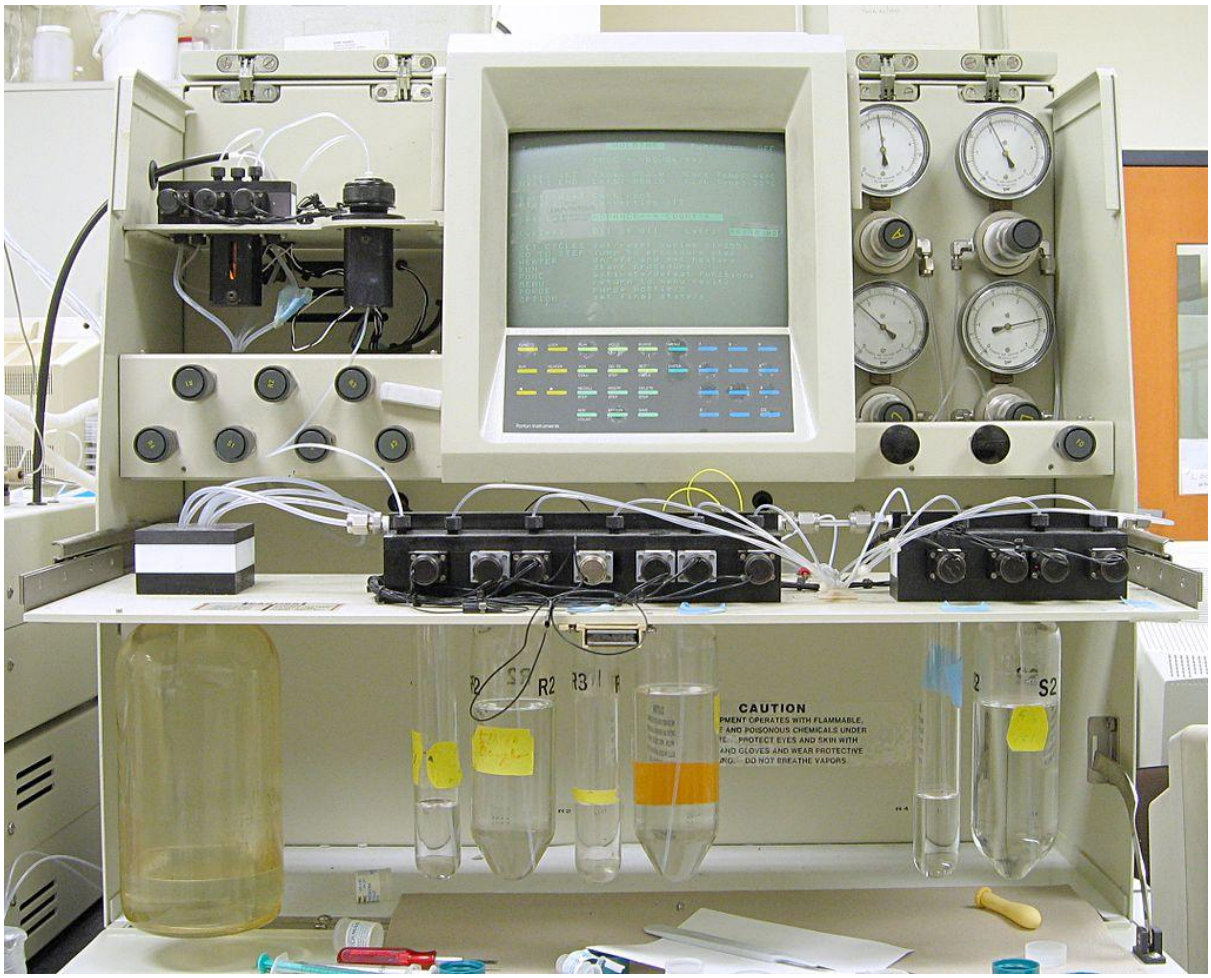
Reaction:

The peptide to be sequenced is adsorbed onto a solid surface. One common substrate is glass fibre coated with polybrene, a cationic polymer. The Edman reagent, phenylisothiocyanate (PITC), is added to the adsorbed peptide, together with a mildly basic buffer solution of 12% trimethylamine. This reacts with the amine group of the N-terminal amino acid.

The terminal amino acid can then be selectively detached by the addition of anhydrous acid. The derivative then isomerises to give a substituted phenylthiohydantoin, which can be washed off and identified by chromatography, and the cycle can be repeated. The efficiency of each step is about 98%, which allows about 50 amino acids to be reliably determined.

Protein sequencer:

A **protein sequenator** is a machine that performs Edman degradation in an automated manner. A sample of the protein or peptide is immobilized in the reaction vessel of the protein sequenator and the Edman degradation is performed. Each cycle releases and derivatises one amino acid from the protein or peptide's *N*-terminus and the released amino-acid derivative is then identified by HPLC. The sequencing process is done repetitively for the whole polypeptide until the entire measurable sequence is established or for a pre-determined number of cycles.



Identification by mass spectrometry

Protein identification is the process of assigning a name to a protein of interest (POI), based on its amino-acid sequence. Typically, only part of the protein's sequence needs to be determined experimentally in order to identify the protein with reference to databases of protein sequences deduced from the DNA sequences of their genes. Further protein characterization may include confirmation of the actual N- and C-termini of the POI, determination of sequence variants and identification of any post-translational modifications present.

Proteolytic digests:

A general scheme for protein identification is described.

1. The POI is isolated, typically by SDS-PAGE or chromatography.
2. The isolated POI may be chemically modified to stabilise Cysteine residues (e.g. S-amidomethylation or S-carboxymethylation).
3. The POI is digested with a specific protease to generate peptides. Trypsin, which cleaves selectively on the C-terminal side of Lysine or Arginine residues, is the most commonly used protease. Its advantages include i) the frequency of Lys and Arg residues in proteins, ii) the high specificity of the enzyme, iii) the stability of the enzyme and iv) the suitability of tryptic peptides for mass spectrometry.
4. The peptides may be desalted to remove ionizable contaminants and subjected to MALDI-TOF mass spectrometry. Direct measurement of the masses of the peptides may provide sufficient information to identify the protein (see Peptide mass fingerprinting) but further fragmentation of the peptides inside the mass spectrometer is often used to gain information about the peptides' sequences. Alternatively, peptides may be desalted and separated by reversed phase HPLC and introduced into a mass spectrometer via an ESI source. LC-ESI-MS may provide more information than MALDI-MS for protein identification but uses more instrument time.
5. Depending on the type of mass spectrometer, fragmentation of peptide ions may occur via a variety of mechanisms such as collision-induced dissociation (CID) or post-source decay (PSD). In each case, the pattern of fragment ions of a peptide provides information about its sequence.
6. Information including the measured mass of the putative peptide ions and those of their fragment ions is then matched against calculated mass values from the conceptual (in-silico) proteolysis and fragmentation of databases of protein sequences. A successful match will be found if its score exceeds a threshold based on the analysis parameters. Even if the actual protein is not represented in the database, error-tolerant matching

allows for the putative identification of a protein based on similarity to homologous proteins. A variety of software packages are available to perform this analysis.

7. Software packages usually generate a report showing the identity (accession code) of each identified protein, its matching score, and provide a measure of the relative strength of the matching where multiple proteins are identified.
8. A diagram of the matched peptides on the sequence of the identified protein is often used to show the sequence coverage (% of the protein detected as peptides). Where the POI is thought to be significantly smaller than the matched protein, the diagram may suggest whether the POI is an N- or C-terminal fragment of the identified protein.

De novo sequencing:

The pattern of fragmentation of a peptide allows for direct determination of its sequence by *de novo* sequencing. This sequence may be used to match databases of protein sequences or to investigate post-translational or chemical modifications. It may provide additional evidence for protein identifications performed as above.

N- and C-termini:

The peptides matched during protein identification do not necessarily include the N- or C-termini predicted for the matched protein. This may result from the N- or C-terminal peptides being difficult to identify by MS (e.g. being either too short or too long), being post-translationally modified (e.g. N-terminal acetylation) or genuinely differing from the prediction. Post-translational modifications or truncated termini may be identified by closer examination of the data (i.e. *de novo* sequencing). A repeat digest using a protease of different specificity may also be useful.

Post-translational modifications:

Whilst detailed comparison of the MS data with predictions based on the known protein sequence may be used to define post-translational modifications, targeted approaches to data acquisition may also be used. For instance, specific enrichment of phosphopeptides may assist in identifying phosphorylation sites in a protein. Alternative methods of peptide fragmentation in the mass spectrometer, such as ETD or ECD, may give complementary sequence information.

Whole-mass determination:

The protein's whole mass is the sum of the masses of its amino-acid residues plus the mass of a water molecule and adjusted for any post-translational modifications.

Although proteins ionize less well than the peptides derived from them, a protein in solution may be able to be subjected to ESI-MS and its mass measured to an accuracy of 1 part in 20,000 or better. This is often sufficient to confirm the termini (thus that the protein's measured mass matches that predicted from its sequence) and infer the presence or absence of many post-translational modifications.

Limitations:

Proteolysis does not always yield a set of readily analyzable peptides covering the entire sequence of POI. The fragmentation of peptides in the mass spectrometer often does not yield ions corresponding to cleavage at each peptide bond. Thus, the deduced sequence for each peptide is not necessarily complete. The standard methods of fragmentation do not distinguish between leucine and isoleucine residues since they are isomeric.

Because the Edman degradation proceeds from the N-terminus of the protein, it will not work if the N-terminus has been chemically modified (e.g. by acetylation or formation of Pyroglutamic acid). Edman degradation is generally not useful to determine the positions of disulphide bridges. It also requires peptide amounts of 1 picomole or above for discernible results, making it less sensitive than mass spectrometry.

Predicting from DNA/RNA sequences:

In biology, proteins are produced by translation of messenger RNA (mRNA) with the protein sequence deriving from the sequence of codons in the mRNA. The mRNA is itself formed by the transcription of genes and may be further modified. These processes are sufficiently understood to use computer algorithms to automate predictions of protein sequences from DNA sequences, such as from whole-genome DNA-sequencing projects, and have led to the generation of large databases of protein sequences such as UniProt. Predicted protein sequences are an important resource for protein identification by mass spectrometry.

Historically, short protein sequences (10 to 15 residues) determined by Edman degradation were back-translated into DNA sequences that could be used as probes or primers to isolate molecular clones of the corresponding gene or complementary DNA. The sequence of the cloned DNA was then determined and used to deduce the full amino-acid sequence of the protein.

Bioinformatics tools:

Bioinformatics tools exist to assist with interpretation of mass spectra (see de novo peptide sequencing), to compare or analyse protein sequences (see sequence analysis), or search databases using peptide or protein sequences (see BLAST).

Applications to cryptography:

The difficulty of protein sequencing was recently proposed as a basis for creating k-time programs, programs that run exactly k times before self-destructing. Such a thing is impossible to build purely in software because all software is inherently clonable an unlimited number of times.

Protein Protein Interaction:

Protein-protein interaction plays key role in predicting the protein function of target protein and drug ability of molecules. The majority of genes and proteins realize resulting phenotype functions as a set of interactions. The *in vitro* and *in vivo* methods like affinity purification, Y2H (yeast 2 hybrid), TAP (tandem affinity purification), and so forth have their own limitations like cost, time, and so forth, and the resultant data sets are noisy and have more false positives to annotate the function of drug molecules. Thus, *in silico* methods which include sequence-based approaches, structure-based approaches, chromosome proximity, gene fusion, *in silico* 2 hybrid, phylogenetic tree, phylogenetic profile, and gene expression-based approaches were developed. Elucidation of protein interaction networks also contributes greatly to the analysis of signal transduction pathways. Recent developments have also led to the construction of networks having all the protein-protein interactions using computational methods for signaling pathways and protein complex identification in specific diseases.

Protein-protein interactions (PPIs) handle a wide range of biological processes, including cell-to-cell interactions and metabolic and developmental control. Protein-protein interaction is becoming one of the major objectives of system biology. Noncovalent contacts between the residue side chains are the basis for protein folding, protein assembly, and PPI. These contacts induce a variety of interactions and associations among the proteins. Based on their contrasting structural and functional characteristics, PPIs can be classified in several ways. On the basis of their interaction surface, they may be homo- or heterooligomeric; as judged by their stability, they may be obligate or nonobligate; as measured by their persistence, they may be transient or permanent. A given PPI may be a combination of these three specific pairs. The transient interactions would form signalling pathways while permanent interactions will form a stable protein complex.

Typically proteins hardly act as isolated species while performing their functions *in vivo*. It has been revealed that over 80% of proteins do not operate alone but in complexes. The substantial analysis of authenticated proteins reveals that the proteins involved in the same cellular processes are repeatedly found to be interacting with each other. The study of PPIs is also important to infer the protein function within the cell. The functionality of unidentified proteins can be predicted on the evidence of their

interaction with a protein, whose function is already revealed. The detailed study of PPIs has expedited the modelling of functional pathways to exemplify the molecular mechanisms of cellular processes. Characterizing the interactions of proteins in a given proteome will be phenomenal to figure out the biochemistry of the cell. The result of two or more proteins interacting with a definite functional objective can be established in several ways. The significant properties of PPIs have been marked by Phizicky and Fields.PPIs can(i)modify the kinetic properties of enzymes;(ii)act as a general mechanism to allow for substrate channelling;(iii)construct a new binding site for small effector molecules;(iv)inactivate or suppress a protein;(v)change the specificity of a protein for its substrate through interaction with different binding partners;(vi)serve a regulatory role in either upstream or downstream level.

Uncovering protein-protein interaction information helps in the identification of drug targets. Studies have shown that proteins with larger number of interactions (hubs) can include families of enzymes, transcription factors, and intrinsically disordered proteins, among others. However, PPIs involve more heterogeneous processes and the scope of their regulation is large. For more accurate understanding of their importance in the cell, one has to identify various interactions and determine the aftermath of the interactions.

In recent years, PPI data have been enhanced by guaranteed high-throughput experimental methods, such as two-hybrid systems, mass spectrometry, phage display, and protein chip technology. Comprehensive PPI networks have been built from these experimental resources. However, the voluminous nature of PPI data is imposing a challenge to laboratory validation. Computational analysis of PPI networks is increasingly becoming a mandatory tool to understand the functions of unexplored proteins. At present, protein-protein interaction (PPI) is one of the key topics for the development and progress of modern system's biology.

Classification of PPI Detection Methods

Protein-protein interaction detection methods are categorically classified into three types, namely, *in vitro*, *in vivo*, and *in silico* methods. In *in vitro* techniques, a given procedure is performed in a controlled environment outside a living organism. The *in vitro* methods in PPI detection are tandem affinity purification, affinity chromatography, coimmunoprecipitation, protein arrays, protein fragment complementation, phage display, X-ray crystallography, and NMR spectroscopy. In *in vivo* techniques, a given procedure is performed on the whole living organism itself. The *in vivo* methods in PPI detection are yeast two-hybrid (Y2H, Y3H) and synthetic lethality. *In silico* techniques are performed on a computer (or) via computer simulation. The *in silico* methods in PPI detection are sequence-based approaches, structure-based approaches, chromosome proximity, gene fusion, *in silico* 2 hybrid, mirror tree, phylogenetic tree, and gene expression-based approaches.

In Vitro Techniques to Predict Protein-Protein Interactions

TAP tagging was developed to study PPIs under the intrinsic conditions of the cell. Gavin et al. first attempted the TAP-tagging method in a high-throughput manner to analyse the yeast interactome. This method is based on the double tagging of the protein of interest on its chromosomal locus, followed by a two-step purification process. Proteins that remain associated with the target protein can then be examined and identified through SDS-PAGE followed by mass spectrometry analysis, thereby identifying the PPI collaborator of the original protein of interest. An important dominance of TAP-tagging is its ability to identify a wide variety of protein complexes and to test the activeness of monomeric or multimeric protein complexes that exist *in vivo* . The TAP when used with mass spectroscopy (MS) will identify protein interactions and protein complexes.

The advantage of the affinity chromatography is that it is highly responsive, can even detect weakest interactions in proteins, and also tests all the sample proteins equally for interaction with the coupled protein in the column. However, false positive results also arise in the column due to high specificity among proteins, even though they do not get involved in the cellular system. Thus protein interaction studies cannot fully rely on affinity chromatography and hence require other methods in order to crosscheck and verify results obtained. The affinity chromatography can also be associated with SDS-PAGE technique and mass spectroscopy in order to generate a high-throughput data.

Coimmunoprecipitation confirms interactions using a whole cell extract where proteins are present in their native form in a complex mixture of cellular components that may be required for successful interactions. In addition, use of eukaryotic cells enables posttranslational modification which may be essential for interaction and which would not occur in prokaryotic expression systems.

Protein microarrays are rapidly becoming established as a powerful means to detect proteins, monitor their expression levels, and probe protein interactions and functions. A protein microarray is a piece of glass on which various molecules of protein have been affixed at separate locations in an ordered manner. Protein microarrays have seen tremendous progress and interest at the moment and have become one of the active areas emerging in biotechnology. The objective behind protein microarray development is to achieve efficient and sensitive high-throughput protein analysis, carrying out large numbers of determinations in parallel by automated process.

Protein-fragment complementation assay is another method of proteomics for the identification of protein-protein interactions in biological systems. Protein-fragment complementation assays (PCAs) are a family of assays for detecting protein-protein interactions (PPIs) that have been introduced to provide simple and direct ways to study PPIs in any living cell, multicellular organism, or *in vitro*. PCAs can be used to detect PPI between proteins of any molecular weight and expressed at their

endogenous levels. The two choices for protein identification using a mass spectroscopy are peptide fingerprinting and shotgun proteomics. For peptide fingerprinting, the eluted complex is separated using SDS-PAGE. The gel is either Coomassie-stained or silver-stained and bands unique to the test sample and hopefully containing a single protein are excised, enzymatically digested, and analysed by mass spectrometry. The mass of these peptides is determined and matched to a peptide database to determine the source protein. The gel also provides a rough estimate of the molecular weight of the protein. Since only unique bands are cut out, background bands are not identified. Abundant background proteins may obscure target proteins while less abundant proteins may fall below the limits of detection by staining. This method works well with purified samples containing only a handful of proteins. Alternatively, for shotgun proteomics, the entire eluate, containing many proteins, is digested. Shotgun proteomics is currently the most powerful strategy for analysing such complicated mixtures.

There are different implementations of the phage display methodology as well as different applications. One of the most common approaches used is the M13 filamentous phage. The DNA encoding the protein of interest is ligated into the gene encoding one of the coat proteins of the virion. Normally, the process is followed by computational identification of potential interacting partners and a yeast two-hybrid validation step, but the method is a new born one.

X-ray crystallography is essentially a form of very high resolution microscopy, which enables visualization of protein structures at the atomic level and enhances the understanding of protein function. Specifically it shows how proteins interact with other molecules and the conformational changes in case of enzymes. Armed with this information, we can also design novel drugs that target a particular target protein.

In the recent past, the researchers have shown interest in the analysis of protein-protein interaction by nuclear magnetic resonance (NMR) spectroscopy. The location of binding interface is a crucial aspect in the protein interaction analysis. The basis for the NMR spectroscopy is that magnetically active nuclei oriented by a strong magnetic field absorb electromagnetic radiation at characteristic frequencies governed by their chemical environment.

In Vivo Techniques to Predict Protein-Protein Interactions

Y2H method is an *in vivo* method applied to the detection of PPIs . Two protein domains are required in the Y2H assay which will have two specific functions: (i) a DNA binding domain (DBD) that helps binding to DNA, and (ii) an activation domain (AD) responsible for activating transcription of DNA. Both domains are required for the transcription of a reporter gene. Y2H analysis allows the direct recognition of PPI between protein pairs. However, the method may incur a large number of false positive interactions. On the other hand, many true interactions may not be traced using Y2H

assay, leading to false negative results. In an Y2H assay, the interacting proteins must be localized to the nucleus, since proteins, which are less likely to be present in the nucleus are excluded because of their inability to activate reporter genes. Proteins, which need posttranslational modifications to carry out their functions, are unlikely to behave or interact normally in an Y2H experiment. Furthermore, if the proteins are not in their natural physiological environment, they may not fold properly to interact. During the last decade, Y2H has been enriched by designing new yeast strains containing multiple reporter genes and new expression vectors to facilitate the transformation of yeast cells with hybrid proteins. Other widely used techniques, such as bioluminescence resonance energy transfer (BRET), fluorescence resonance energy transfers (FRET), and bimolecular fluorescence complementation (BiFC), require extensive instrumentation. FRET uses time-correlated single-photon counting to predict protein interactions.

Synthetic lethality is an important type of *in vivo* genetic screening which tries to understand the mechanisms that allow phenotypic stability despite genetic variation, environmental changes, and random events such as mutations. This methodology produces mutations or deletions in two or more genes which are viable alone but cause lethality when combined together under certain conditions. Compared with the results obtained in the aforesaid methods, the relationships detected by synthetic lethality do not require necessity of physical interaction between the proteins. Therefore, we refer to this type of relationships as functional interactions.

In Silico Methods for the Prediction of Protein-Protein Interactions

The yeast two-hybrid (Y2H) system and other *in vitro* and *in vivo* approaches resulted in large-scale development of useful tools for the detection of protein-protein interactions (PPIs) between specified proteins that may occur in different combinations. However, the data generated through these approaches may not be reliable because of nonavailability of possible PPIs. In order to understand the total context of potential interactions, it is better to develop approaches that predict the full range of possible interactions between proteins.

A variety of *in silico* methods have been developed to support the interactions that have been detected by experimental approach. The computational methods for *in silico* prediction include sequence-based approaches, structure-based approaches, chromosome proximity, gene fusion, *in silico* 2 hybrid, mirror tree, phylogenetic tree, gene ontology, and gene expression-based approaches.

Structure-Based Prediction Approaches

The idea behind the structure-based method is to predict protein-protein interaction if two proteins have a similar structure. Therefore, if two proteins A and B can interact with each other, then there may be two other proteins and whose structures are similar

to those of proteins A and B; then it is implied that proteins and can also interact with each other. But most proteins may not be having known structures; the first step for this method is to guess the structure of the protein based on its sequence. This can be done in different ways. The PDB database offers useful tools and information resources for researchers to build the structure for a query protein. Using the multimeric threading approach, Lu et al have made 2,865 protein-protein interactions in yeast and 1,138 interactions have been confirmed in the DIP.

Recently, Hosur et al. developed a new algorithm to infer protein-protein interactions using structure-based approach. The Coev2Net algorithm, which is a three-step process, involves prediction of the binding interface, evaluation of the compatibility of the interface with an interface coevolution-based model, and evaluation of the confidence score for the interaction. The algorithm when applied to binary protein interactions has boosted the performance of the algorithm over existing methods. However, Zhang et al. have used three-dimensional structural information to predict PPIs with an accuracy and coverage that are superior to predictions based on nonstructural evidence.

Sequence-Based Prediction Approaches

Predictions of PPIs have been carried out by integrating evidence of known interactions with information regarding sequential homology. This approach is based on the concept that an interaction found in one species can be used to infer the interaction in other species. However, recently, Hosur et al. developed a new algorithm to predict protein-protein interactions using threading-based approach which takes sequences as input. The algorithm, iWARP (Interface Weighted RAPtor), which predicts whether two proteins interact by combining a novel linear programming approach for interface alignment with a boosting classifier for interaction prediction. Guilherme Valente et al. introduced a new method called Universal *In Silico* Predictor of Protein-Protein Interactions (UNISPPI), based on primary sequence information for classifying protein pairs as interacting or noninteracting proteins. Kernel methods are hybrid methods which use a combination of properties like protein sequences, gene ontologies, and so forth. However, there are two different methods under sequence-based criterion.

(1) Ortholog-Based Approach. The approach for sequence-based prediction is to transfer annotation from a functionally defined protein sequence to the target sequence based on the similarity. Annotation by similarity is based on the homologous nature of the query protein in the annotated protein databases using pairwise local sequence algorithm. Several proteins from an organism under study may share significant similarities with proteins involved in complex formation in other organisms.

The prediction process starts with the comparison of a probe gene or protein with those annotated proteins in other species. If the probe gene or protein has high similarity to the sequence of a gene or protein with known function in another species, it is assumed that the probe gene or protein has either the same function or similar properties. Most

subunits of protein complexes were annotated in that way. When the function is transferred from a characterized protein to an uncharacterized protein, ortholog and paralog concepts should be applied. Orthologs are the genes in different species that have evolved from a common ancestral gene by speciation. In contrast, paralogs usually refer to the genes related by duplication within a genome. In broad sense, orthologs will retain the functionality during the course of evolution, whereas paralogs may acquire new functions. Therefore, if two proteins—A and B—interact with each other, then the orthologs of A and B in a new species are also likely to interact with each other.

(2) Domain-Pairs-Based Approach. A domain is a distinct, compact, and stable protein structural unit that folds independently of other such units. But most of times, domains are defined as distinct regions of protein sequence that are highly conserved in the process of evolution. As individual structural and functional units, protein domains play an important role in the development of protein structural class prediction, protein subcellular location prediction, membrane protein type prediction, and enzyme class and subclass prediction.

Conventionally, protein domains are used for basic research and also for structure-based drug designing. In addition, domains are directly involved in the intermolecular interaction and hence must be fundamental to protein-protein interaction. Multiple studies have shown that domain-domain interactions (DDIs) from different experiments are more consistent than their corresponding PPIs. So, it is quite reliable to use the domains and their interactions for prediction of the protein-protein interactions and vice versa.

Chromosome Proximity/Gene Neighbourhood

With the ever increasing number of the completely sequenced genomes, the global context of genes and proteins in the completed genomes has provided the researchers with the enriched information needed for the protein-protein interaction detection. It is well known that the functionally related proteins tend to be organized very closely into regions on the genomes in prokaryotes, such as operons, the clusters of functionally related genes transcribed as a single mRNA. If the neighbourhood relationship is conserved across multiple genomes, then it will be more relevant for implying the potential possibility of the functional linkage among the proteins encoded by the related genes. And this evidence was applied to study the functional association of the corresponding proteins. This relationship was confirmed by the experimental results and shown to be more independent of relative gene orientation. Recently, it has been found that there is functional link among the adjacent bidirectional genes along the chromosome. Interestingly, in most cases, the relationship among adjacent bidirectionally transcribed genes with conserved gene orientation is that one gene encodes a transcriptional regulator and the other belongs to nonregulatory protein. It has been found that most of the regulators control the transcription of the diver gently

transcribed target gene/operon and automatically regulate their own biosynthesis as well. This relationship provides another way to predict the target processes and regulatory features for transcriptional regulators. One of the pitfalls of this method is that it is directly suitable for bacterial genome since gene neighbouring is conserved in the bacteria.

Gene Fusion:

Gene fusion, which is often called as Rosetta stone method, is based on the concept that some of the single-domain containing proteins in one organism can fuse to form a multidomain protein in other organisms. This domain fusion phenomenon indicates the functional association for those separate proteins, which are likely to form a protein complex. It has been shown that fusion events are particularly common in those proteins participating in the metabolic pathway. This method can be used to predict protein-protein interaction by using information of domain arrangements in different genomes. However, it can be applied only to those proteins in which the domain arrangement exists.

In Silico Two-Hybrid (I2h)

The method is based on the assumption that interacting proteins should undergo coevolution in order to keep the protein function reliable. In other words, if some of the key amino acids in one protein changed, the related amino acids in the other protein which interacts with the mutated counter partner should also make the compulsory mutations as well. During the analysis phase, the common genomes containing those two proteins will be identified through multiple sequence alignments and a correlation coefficient will be calculated for every pair of residues in the same protein and between the proteins. Accordingly, there are three different sets for the pairs: two from the intraprotein pairs and one from the interprotein pairs. The protein-protein interaction is inferred based on the difference from the distribution of correlation between the interacting partners and the individual proteins. Since I2h analysis is based on the prediction of physical closeness between residue pairs of the two individual proteins, the result from this method automatically indicates the possible physical interaction between the proteins.

Phylogenetic Tree:

Another important method for detection of interaction between the proteins is phylogenetic tree. The phylogenetic tree gives the evolution history of the protein. The mirror tree method predicts protein-protein interactions under the belief that the interacting proteins show similarity in molecular phylogenetic tree because of the

coevolution through the interaction. The underlying principle behind the method is that the coevolution between the interacting proteins can be reflected from the degree of similarity from the distance matrices of corresponding phylogenetic trees of the interacting proteins. The set of organisms common to the two proteins are selected from the multiple sequence alignments (MSA) and the results are used to construct the corresponding distance matrix for each protein. The BLAST scores could also be used to fill the matrices. Then the linear correlation is calculated among these distance matrices. High correlation scores indicate the similarity between the phylogenetic trees and therefore the proteins are considered to have the interaction relationship. The MirrorTree method is used to detect the coevolution relationship between proteins and the results are used to infer the possibility of their physical interaction.

Phylogenetic Profile:

The notion for this method is that the functionally linked proteins tend to coexist during the evolution of an organism. In other words, if two proteins have a functional linkage in a genome, there will be a strong pressure on them to be inherited together during evolution process. Thus, their corresponding orthologs in other genome will be preserved or dropped. Therefore, we can detect the presence or absence (cooccurrence) of proteins in the phylogenetic profile. A phylogenetic profile describes an occurrence of a certain protein in a set of genomes: if two proteins share the same phylogenetic profiling, this indicates that the two proteins have the functional linkage. In order to construct the phylogenetic profile, a predetermined threshold of BLASTP -value is used to detect the presence or absence of the homological proteins on the target genome with the source genomes. This method gives promising results in the detection of the functional linkage among the proteins and, at the same time, assigns the functions to query proteins. Even though the phylogenetic profile has shown great potential for building the functional linkage network on the full genome level, the following two pitfalls should be mentioned: one is that this method is based on full genome sequences and the other is that the functional linkage between proteins is detected by their phylogenetic profiling, so it is difficult to use the method for those essential proteins in the cell where no difference can be detected from the phylogenetic profile. Moreover, even though the increasing number in the source genome set can improve the prediction accuracy, there may be an upper limit for this method.

Many genomic events contribute to the noise during the coevolution, such as gene duplication or the possible loss of gene functions in the course of evolution, which could corrupt the phylogenetic profile of single genes. Phylogenetic-profile-based methods conceded satisfactory performance only on prokaryotes but not on eukaryotes.

Gene Expression:

The method takes the advantage of high-throughput detection of the whole gene transcription level in an organism. Gene expression means the quantification of the level at which a particular gene is expressed within a cell, tissue or organism under different experimental conditions and time intervals. By applying the clustering algorithms, different expression genes can be grouped together according to their expression levels, and the resultant gene expression under different experimental conditions can help to enunciate the functional relationships of the various genes. A lot of research has also been carried out to investigate the relationship between gene co-expression and protein interaction. Based on the yeast expression data and proteome data, proteins from the genes belonging to the common expression-profiling clusters are more likely to interact with each other than proteins from the genes belonging to different clusters. In other studies, it has been confirmed that adjacent genes tend to be expressed both in the eukaryotes and prokaryotes. The gene co-expression concept is an indirect way to infer the protein interaction, suggesting that it may not be appropriate for accurate detection of protein interactions. However, as a complementary approach, gene co-expression can be used to validate interactions generated from other experimental methods.

Comparison of Protein-Protein Interaction Methods

Each of the above methods has been applied to detect the protein-protein interaction in both the prokaryotes and eukaryotes. The results show that most of them fit better for the prokaryotes than eukaryotes. The significant increase for the coverage among those studies during the past several years could be mainly because of the increase in the number of genomes being decoded. This is because the more the number of genomes used in the study, the higher the coverage that the methods can reach. With the accumulation of fully sequenced genomes, the information content in the reference genome set is expected to increase. Accordingly, the prediction accuracy would increase with more genomes incorporated in the study. It can be anticipated that, with more and more genomes available in the future, the prediction potential will be improved and the corresponding combined methods will get higher coverage and accuracy. One thing that should be mentioned is that the selection of the standard used for the evaluation of the methods has a great impact on the coverage and accuracy. Besides the Operon and Swiss-Prot key word recovery used in the above studies, the KEGG has been used as the standard in Search Tool for the Retrieval of Interacting Genes (STRING) database [58]. It can be expected that the prediction coverage and accuracy will be different for each method under different standards. Obviously, the achieved highest coverage for the gene order method based on the operon standard indicates that the method is strongly related to operon.

Recent technological advances have allowed high-throughput measurements of protein-protein interactions in the cell, producing protein interaction networks for different species at a rapid pace. However, high-throughput methods like yeast two-hybrid, MS, and phage display have experienced high rates of noise and false positives. There are some verification methods to know the reliability of these high throughput interactions. They are Expression Profile Reliability (EPR index), Paralogous Verification Method (PVM) Protein Localization Method (PLM), and Interaction Generalities Measures IG1 and IG2. EPR method compares protein interaction with RNA expression profiles whereas PVM analyzes paralogs of interactors for comparison. The IG1 measure is based on the idea that interacting proteins that have no further interactions beyond level-1 are likely to be false positives. The IG2 measure uses the topology of interactions. Bayesian approaches have also been used for calculation of reliability. The PLM gives the true positives (TP) as interacting proteins, which need to be localized in the same cellular compartment or annotated to have a common cellular role. So, in order to counter these errors, many methods have been developed which provides confidence scores with each interaction. Also, the methods that assign scores to individual interactions generally perform better than those with the set of interactions obtained from an experiment or a database.

Computational Analysis of PPI Networks

A PPI network can be described as a heterogeneous network of proteins joined by interactions as edges. The computational analysis of PPI networks begins with the illustration of the PPI network arrangement. The simplest sketch takes the form of a mathematical graph consisting of nodes and edges. Protein is represented as a node in such a graph and the proteins that interact with it physically are represented as adjacent nodes connected by an edge. An examination of the network can yield a variety of results. For example, neighbouring proteins in the graph probably may share more the same functionality. In addition to the functionality, densely connected subgraphs in the network are likely to form protein complexes as a unit in certain biological processes. Thus, the functionality of a protein can be inferred by spotting at the proteins with which it interacts and the protein complexes to which it resides. The topological prediction of new interactions is a novel and useful option based exclusively on the structural information provided by the PPI network (PPIN) topology. Some algorithms like random layout algorithm, circular layout algorithm, hierarchical layout algorithm, and so forth are used to visualize the network for further analysis. Precisely, the computational analysis of PPI networks is challenging, with these major barriers being commonly confronted [4]:(1)the protein interactions are not stable;(2)one protein may have different roles to perform;(3)two proteins with distinct functions periodically interact with each other.

Role of PPI Networks in Proteomics

Predicting the protein functionality is one of the main objectives of the PPI network. Despite the recent comprehensive studies on yeast, there are still a number of functionally unclassified proteins in the yeast database which reflects the impending need to classify the proteins. The functional annotation of human proteins can provide a strong foundation for the complete understanding of cell mechanisms, information that is valuable for drug discovery and development. The increased availability of PPI networks has developed various computational methods to predict protein functions. The availability of reliable information on protein interactions and their presence in physiological and pathophysiological processes are critical for the development of protein-protein-interaction-based therapeutics. The compendium of all known protein-protein interactions (PPIs) for a given cell or organism is called the interactome.

Protein functions may be predicted on the basis of modularization algorithms. However, predictions found in this way may not be accurate because the accuracy of the modularization process itself is typically low. There are other methods which include the neighbour counting, Chi-square, Markov random field, Prodistin, and weighted-interactions-based method for the prediction of protein function. For greater accuracy, protein functions should be predicted directly from the topology or connectivity of PPI networks. Several topology-based approaches that predict protein function on the basis of PPI networks have been introduced. At the simplest level, the "neighbour counting method" predicts the function of an unexplored protein by the frequency of known functions of the immediate neighbour proteins. The majority of functions of the immediate neighbours can be statistically assessed. Recently, the number of common neighbours of the known protein and the unknown protein has been taken as the basis for the inference of function. The weighted-graph-mining-based protein function prediction is a novel approach in the area.

Protein-protein complex identification is the crucial step in finding the signal transduction pathways. Protein-protein complexes mostly consist of antibody-antigen and protease-inhibitor complexes. Crystallography is the major tool for determining protein complexes at atomic resolution.

The complete analysis of PPIs can enable better understanding of cellular organization, processes, and functions. The other applications of PPI Network include biological indispensability analysis, assessing the drug ability of molecular targets from network topology, estimation of interactions reliability, identification of domain-domain interactions, prediction of protein interactions, detection of proteins involved in disease pathways, delineation of frequent interaction network motifs, comparison between model organisms and humans, and protein complex identification.

Protein Interaction Databases

The massive quantity of experimental PPI data being generated on steady basis has led to the construction of computer-readable biological databases in order to organize and to process this data. For example, the biomolecular interaction network database (BIND) is created on an extensible specification system that permits an elaborate description of the manner in which the PPI data was derived experimentally, often including links directly to the concluding evidence from the literature. The database of interacting proteins (DIP) is another database of experimentally determined protein-protein binary interactions. The biological general repository for interaction datasets (BioGRID) is a database that contains protein and genetic interactions among thirteen different species. Interactions are regularly added through exhaustive curation of the primary literature to the databases. Interaction data is extracted from other databases including BIND and MIPS (Munich Information Center for protein se quences), as well as directly from large-scale experiments. HitPredict is a resource of high confidence protein-protein interactions from which we can get the total number of interactions in a species for a protein and can view all the interactions with confidence scores.

The Molecular Interaction (MINT) database is another database of experimentally derived PPI data extracted from the literature, with the added element of providing the weight of evidence for each interaction. The Human Protein Interaction Database (HPID) was designed to provide human protein interaction information precomputed from existing structural and experimental data. The information Hyperlinked over Proteins (iHOP) database can be searched to identify previously reported interactions in PubMed for a protein of interest. IntAct provides an open source database and toolkit for the storage, presentation, and analysis of protein interactions. The web interface provides both textual and graphical representations of protein interactions and allows exploring interaction networks in the context of the GO annotations of the interacting proteins. However, we have observed that the intersection and overlap between these source PPI databases is relatively small. Recently, the integration has been done and can be explored in the web server called APID (Agile Protein Interaction Data Analyzer) which is an interactive bioinformatics' web tool developed to allow exploration and analysis of currently known information about protein-protein interactions integrated and unified in a common and comparative platform. The Protein Interaction Network Analysis (PINA2.0) platform is a comprehensive web resource, which includes a database of unified protein-protein interaction data integrated from six manually curated public databases and a set of built-in tools for network construction, filtering, analysis, and visualization.

Methods to Study Protein- Protein Interactions:

The operation of the genome can be evaluated by the study of proteome. Thus, by studying the functions of proteins, it is possible to understand how the genome operates and how a dysfunctional genome activity can result in disease states such as cancer.

Proteomics broadly involves the methodology for characterizing the protein content of the cell. This can be done by protein electrophoresis, mass spectrometry etc.

Identification of protein-protein interaction is a recent approach to study proteome. The protein interaction maps can be constructed to understand the relation between the proteome and cellular biochemistry. Phage display and yeast two-hybrid system are commonly used to study protein- protein interactions.

Phage Display:

Phage display is a novel technique to evaluate genome activity with particular reference to identify proteins that interact with one another. It basically involves insertion of a foreign DNA into phage genome, and its expression as fusion product with a phage coat protein (Fig. 5.14A). This is followed by screening of test protein by phage display library. The technique is briefly described below.

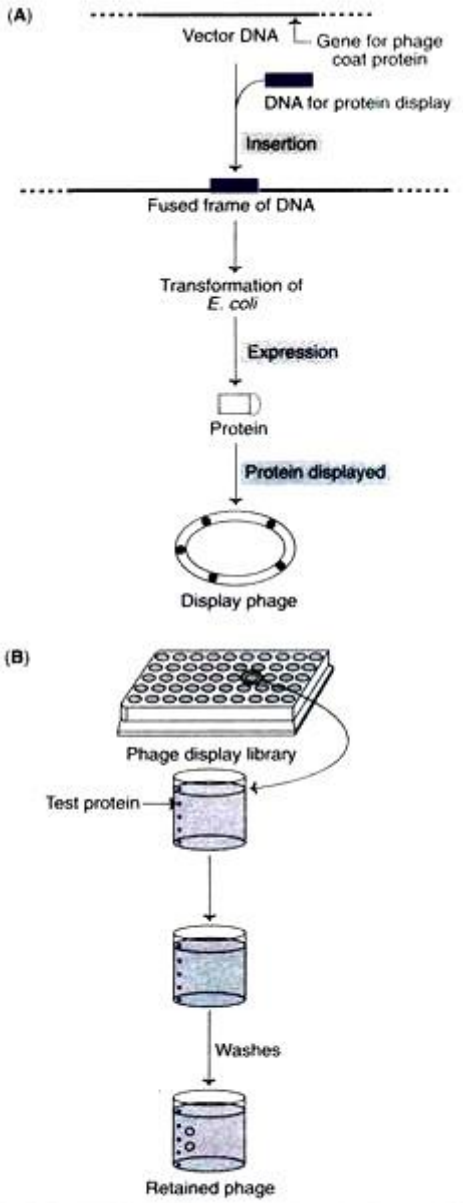


Fig. 5.14 : Elucidation of protein–protein interaction by phage display (A) Production of fusion protein displayed on phage (B) Screening of test protein by phage display library.

A special type of cloning vector such as a bacteriophage or filamentous bacteriophage (e.g. M13) are used for phage display. A fragment of DNA coding for the test protein is inserted into the vector DNA (adjacent to phage coat protein gene). After transformation of E. coli, this recombinant gene (fused frame of DNA) results in the synthesis of hybrid protein. The new protein is made up of the test protein fused with the phage coat protein. The phage particles produced in the transformed E. coli display the test protein in their coats.

The test protein interaction can be identified by using a phage display library. For this purpose, the test protein is immobilized within a well of a micro-titer tray, and the

phage display library added. After several washes, the phages that are retained in the well are those displaying a protein that interacts with the test protein.

Phage-displaying peptides can be isolated, based on their antibody-binding properties, by employing affinity chromatography. Several rounds of affinity chromatography and phage propagation can be used to enrich phages with desired proteins.

Phagemid display:

Phagemid in place of plasmid can also be used for the display of proteins. In fact, special types of phagemid display vectors have been developed for this purpose. Phage and phagemid display can be successfully used for selecting and engineering polypeptides with novel functions.

Yeast Two-Hybrid System:

When two proteins interact with each other, their corresponding genes are known as interacting genes. The yeast two-hybrid system uses a reporter gene to detect the physical interaction of a pair of proteins inside a yeast nucleus.

The two-hybrid method is based on the observation that most of the transcriptional proteins (i.e. the proteins involved in promoting transcription of a gene) contain two distinct domains—DNA binding domain and transcriptional activation domain. When these two domains are physically separated, the protein loses its activity. However, the same protein can be reactivated when the domains are brought together. These proteins can bind to DNA and activate transcription.

The target protein is fused to a DNA-binding domain to form a bait. When this target protein binds to another specifically designed protein namely the prey in the nucleus, they interact, which in turn switches on the expression of the reporter gene (Fig. 5.15). The reporter genes can be detected by growing the yeast on a selective medium.

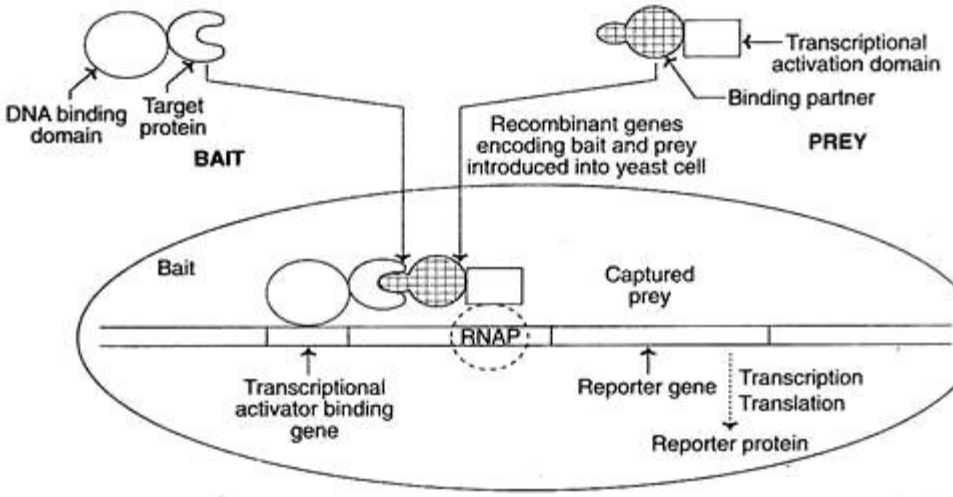


Fig. 5.15 : Elucidation of protein-protein interaction by yeast two-hybrid system (RNAP-RNA polymerase)

It is possible to generate the bait and prey fusion proteins by standard recombinant DNA techniques. A single baid protein is frequently used to fish out interacting partners among the collection of prey proteins. A large number of prey proteins can be produced by ligating DNA encoding the activation domain of a transcriptional activator to a misture of DNA-fragments from a cDNA library.

Yeast Three-Hybrid System:
The interactions between protein and RNA molecules can be investigated by using a technique known as yeast three-hybrid system.

DNA-Protein Interaction:

DNA footprinting is a method of investigating the sequence specificity of DNA-binding proteins in vitro. This technique can be used to study protein-DNA interactions both outside and within cells.

The regulation of transcription has been studied extensively, and yet there is still much that is unknown. Transcription factors and associated proteins that bind promoters, enhancers, or silencers to drive or repress transcription are fundamental to understanding the unique regulation of individual genes within the genome. Techniques like DNA footprinting help elucidate which proteins bind to these associated regions of DNA and unravel the complexities of transcriptional control.

History:

In 1978, David Galas and Albert Schmitz developed the DNA footprinting technique to study the binding specificity of the lac repressor protein. It was originally a modification of the Maxam-Gilbert chemical sequencing technique.

Methods:

The simplest application of this technique is to assess whether a given protein binds to a region of interest within a DNA molecule. Polymerase chain reaction (PCR) amplify and label region of interest that contains a potential protein-binding site, ideally amplicon is between 50 and 200 base pairs in length. Add protein of interest to a portion of the labeled template DNA; a portion should remain separate without protein, for later comparison. Add a cleavage agent to both portions of DNA template. The cleavage agent is a chemical or enzyme that will cut at random locations in a sequence independent manner. The reaction should occur just long enough to cut each DNA molecule in only one location. A protein that specifically binds a region within the DNA template will protect the DNA it is bound to from the cleavage agent. Run both samples side by side on a polyacrylamide gel electrophoresis. The portion of DNA template without protein will be cut at random locations, and thus when it is run on a gel, will produce a ladder-like distribution. The DNA template with the protein will result in ladder distribution with a break in it, the "footprint", where the DNA has been protected from the cleavage agent. Note: Maxam-Gilbert chemical DNA sequencing can be run alongside the samples on the polyacrylamide gel to allow the prediction of the exact location of ligand binding site.

Labelling:

The DNA template labelled at the 3' or 5' end, depending on the location of the binding site(s). Labels that can be used are: radioactivity and fluorescence. Radioactivity has been traditionally used to label DNA fragments for footprinting analysis, as the method was originally developed from the Maxam-Gilbert chemical sequencing technique. Radioactive labelling is very sensitive and is optimal for visualizing small amounts of DNA. Fluorescence is a desirable advancement due to the hazards of using radio-chemicals. However, it has been more difficult to optimize because it is not always sensitive enough to detect the low concentrations of the target DNA strands used in DNA footprinting experiments. Electrophoretic sequencing gels or capillary electrophoresis have been successful in analyzingfootprinting of fluorescent tagged fragments.

Cleavage agent:

A variety of cleavage agents can be chosen. a desirable agent is one that is sequence neutral, easy to use, and is easy to control. Unfortunately no available agents meet all of these standards, so an appropriate agent can be chosen, depending on your DNA sequence and ligand of interest. The following cleavage agents are described in detail: DNase I is a large protein that functions as a double-strand endonuclease. It binds the minor groove of DNA and cleaves the phosphodiester backbone. It is a good cleavage agent for footprinting because its size makes it easily physically hindered. Thus is more likely to have its action blocked by a bound protein on a DNA sequence. In addition, the DNase I enzyme is easily controlled by adding EDTA to stop the reaction. There are however some limitations in using DNase I. The enzyme does not cut DNA randomly; its activity is affected by local DNA structure and sequence and therefore results in an uneven ladder. This can limit the precision of predicting a protein's binding site on the DNA molecule. Hydroxyl radicals are created from the Fenton reaction, which involves reducing Fe^{2+} with H_2O_2 to form free hydroxyl molecules. These hydroxyl molecules react with the DNA backbone, resulting in a break. Due to their small size, the resulting DNA footprint has high resolution. Unlike DNase I they have no sequence dependence and result in a much more evenly distributed ladder. The negative aspect of using hydroxyl radicals is that they are more time consuming to use, due to a slower reaction and digestion time.[4] Ultraviolet irradiation can be used to excite nucleic acids and create photoreactions, which results in damaged bases in the DNA strand.[5] Photoreactions can include: single strand breaks, interactions between or within DNA strands, reactions with solvents, or crosslinks with proteins. The workflow for this method has an additional step, once both your protected and unprotected DNA have been treated, there is subsequent primer extension of the cleaved products.[6][7] The extension will terminate upon reaching a damaged base, and thus when the PCR products are run side-by-side on a gel; the protected sample will show an additional band where the DNA was crosslinked with a bound protein. Advantages of using UV are that it reacts very quickly and can therefore capture interactions that are only momentary. Additionally it can be applied to *in vivo* experiments, because UV can penetrate cell membranes. A disadvantage is that the gel can be difficult to interpret, as the bound protein does not protect the DNA, it merely alters the photoreactions in the vicinity.

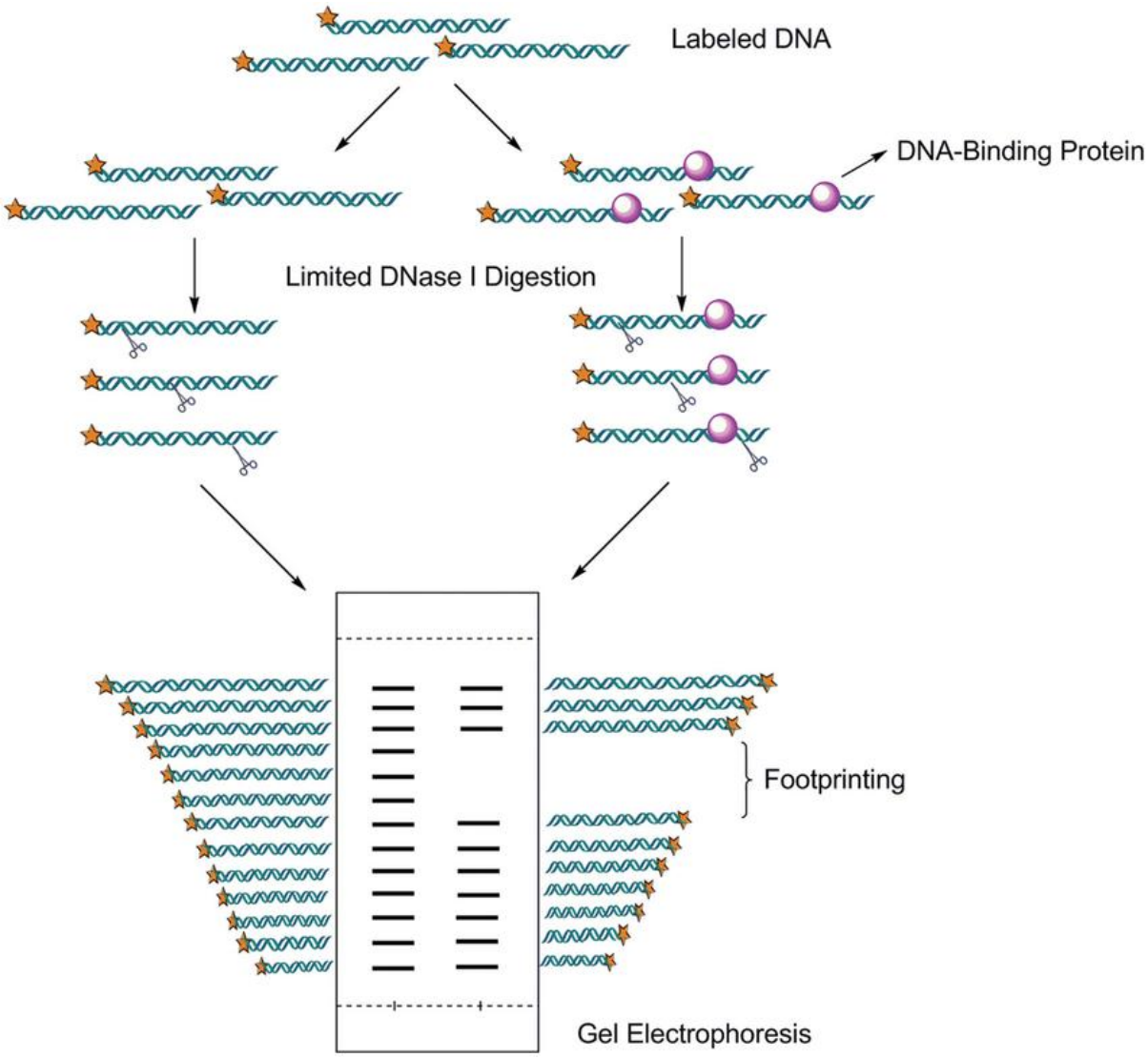


Figure: Dnase I footprinting. This Analysis involves endonuclease treatment of an end labeled DNA fragment bound to a protein. This technique relies on the fact that fragments of DNA that have DNA-binding proteins bound will move more slowly through an acrlyamide gel. The enzyme DNaseI will only cut exposed DNA. Limited digestion yields fragments terminating everywhere except in the footprint region, which is protected from digestion.

Advanced applications:

In vivo footprinting is a technique used to analyze the protein-DNA interactions that are occurring in a cell at a given time point. DNase I can be used as a cleavage agent if the cellular membrane has been permeabilized. However the most common cleavage agent used is UV irradiation because it penetrates the cell membrane without disrupting cell state and can thus capture interactions that are sensitive to cellular changes. Once the DNA has been cleaved or damaged by UV, the cells can be lysed and DNA purified for

analysis of a region of interest. Ligation-mediated PCR is an alternative method to footprint *in vivo*. Once a cleavage agent has been used on the genomic DNA, resulting in single strand breaks, and the DNA is isolated, a linker is added onto the break points. A region of interest is amplified between the linker and a gene-specific primer, and when run on a polyacrylamide gel, will have a footprint where a protein was bound.[11] *In vivo* footprinting combined with immunoprecipitation can be used to assess protein specificity at many locations throughout the genome. The DNA bound to a protein of interest can be immunoprecipitated with an antibody to that protein, and then specific region binding can be assessed using the DNA footprinting technique.

Quantitative footprinting

The DNA footprinting technique can be modified to assess the binding strength of a protein to a region of DNA. Using varying concentrations of the protein for the footprinting experiment, the appearance of the footprint can be observed as the concentrations increase and the proteins binding affinity can then be estimated.

Detection by capillary electrophoresis

To adapt the footprinting technique to updated detection methods, the labelled DNA fragments are detected by a capillary electrophoresis device instead of being run on a polyacrylamide gel. If the DNA fragment to be analyzed is produced by polymerase chain reaction (PCR), it is straightforward to couple a fluorescent molecule such as carboxyfluorescein (FAM) to the primers. This way, the fragments produced by DNaseI digestion will contain FAM, and will be detectable by the capillary electrophoresis machine. Typically, carboxytetramethyl-rhodamine (ROX)-labelled size standards are also added to the mixture of fragments to be analyzed. Binding sites of transcription factors have been successfully identified this way.

Genome-wide assays:

Next-generation sequencing has enabled a genome-wide approach to identify DNA footprints. Open chromatin assays such as DNase-Seq and FAIRE-Seq have proven to provide a robust regulatory landscape for many cell types. However, these assays require some downstream bioinformatics analyses in order to provide genome-wide DNA footprints. The computational tools proposed can be categorized in two classes: segmentation-based and site-centric approaches.

Segmentation-based methods are based on the application of Hidden Markov models or sliding window methods to segment the genome into open/closed chromatin region. Examples of such methods are: HINT, Boyle method[18] and Neph method. Site-centric methods, on the other hand, find footprints given the open chromatin profile around motif-predicted binding sites, i.e., regulatory regions predicted using DNA-protein sequence information (encoded in structures such as position weight matrix). Examples of these methods are CENTIPEDE and Cuellar-Partida method.

A **DNase footprinting assay** is a DNA footprinting technique from molecular biology/biochemistry that detects DNA-protein interaction using the fact that a protein bound to DNA will often protect that DNA from enzymatic cleavage. This makes it possible to locate a protein binding site on a particular DNA molecule. The method uses an enzyme, deoxyribonuclease (DNase, for short), to cut the radioactively end-labeled DNA, followed by gel electrophoresis to detect the resulting cleavage pattern.

For example, the DNA fragment of interest may be PCR amplified using a ^{32}P 5' labeled primer, with the result being many DNA molecules with a radioactive label on one end of one strand of each double stranded molecule. Cleavage by DNase will produce fragments. The fragments which are smaller with respect to the ^{32}P-labelled end will appear further on the gel than the longer fragments. The gel is then used to expose a special photographic film.

The cleavage pattern of the DNA in the absence of a DNA binding protein, typically referred to as free DNA, is compared to the cleavage pattern of DNA in the presence of a DNA binding protein. If the protein binds DNA, the binding site is protected from enzymatic cleavage. This protection will result in a clear area on the gel which is referred to as the "footprint".By varying the concentration of the DNA-binding protein, the binding affinity of the protein can be estimated according to the minimum concentration of protein at which a footprint is observed.This technique was developed by David Galas and Albert Schmitz at Geneva in 1977[

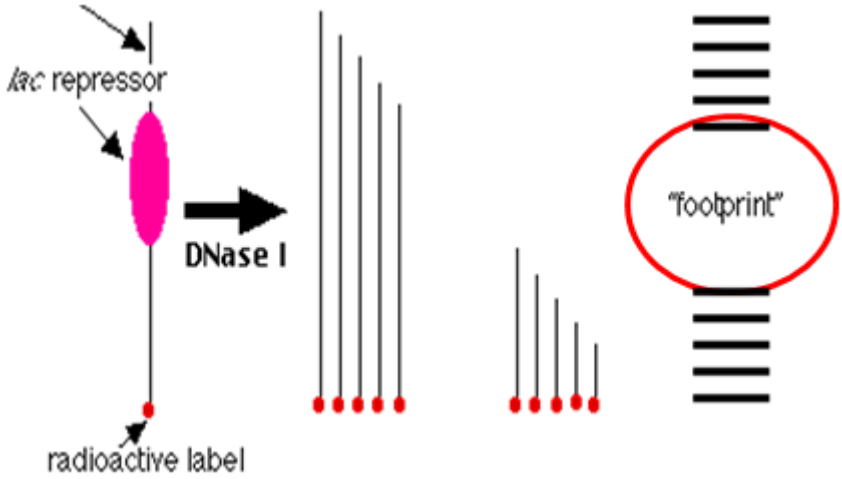


Figure: DNaseI Foot-printing assay

Probable Questions:

1. Describe how Edman degradation can be used for Protein sequence determination?
2. Describe different steps of protein sequencing by protein sequencer.
3. How protein structure can be speculated from DNA/ RNA sequence analysis?
4. How protein protein interaction can be detected?
5. Describe In Vitro Techniques to Predict Protein-Protein Interactions.
6. Describe In Vivo Techniques to Predict Protein-Protein Interactions
7. Describe In Silico Methods for the Prediction of Protein-Protein Interactions
8. Discuss about protein interaction database.
9. Define phage display and phagemid display.
10. How yeast two hybrid system can be used to determine protein protein interaction?
11. Describe different steps of DNA footprinting.

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

UNIT-XIII

Signaling defects, disease and therapeutic drugs: Signaling defects in human disease-Alzheimer's disease, Diabetes mellitus and Cysticfibrosis

Objective: In this unit we will learn about defects in signalling pathways related to particular diseases such as Alzheimer's disease, Diabetes mellitus and Cystic fibrosis.

Alzheimer's disease: The cognitive deficits characteristic of Alzheimer disease (AD) are attributed to the loss of synapses, and ultimately neurons (Stephan *et al.* 2001). Toxic proteins that accrue in the AD brain are believed to precipitate this demise through diagnostic lesions such as senile plaques and neurofibrillary tangles (Hyman and Tanzi 1992). The main proteinaceous constituent of plaques is β-amyloid (Aβ), a peptide derived by cleavage of the amyloid precursor protein (APP) (Glabe 2001). The amyloid hypothesis maintains that extracellular Aβ accumulation instigates a series of events that culminates in a collection of impairments associated with AD (Lee *et al.* 2004). For example, APP transgenic mice with elevated brain levels of Aβ exhibit synapse loss, behavioral abnormalities, and synaptic transmission deficits well before plaque formation (Hsia *et al.* 1999; Mucke *et al.* 2000).

Despite the wealth of evidence indicating that AD-related defects are consequences of Aβ, little is known about the mechanistic pathways leading downstream of Aβ to damage. Several groups have previously demonstrated in cell culture the significance of the C-terminus of APP, including evidence that certain cleavage products from that region influence gene expression and cell survival (Cao and Sudhof 2001; Galvan *et al.* 2002; Kim *et al.* 2003). In particular, APP contains an intracytoplasmic caspase cleavage site - VEVD at codons 661-664 - and cleavage at Asp664 liberates a cytotoxic C-terminal peptide, C31 (Lu *et al.* 2000, 2003). Several findings suggest the functionality of this cleavage site in the genesis of AD pathology. For example, Aβ treatment enhances cleavage at Asp664 and an Asp664 mutation attenuates Aβ-induced cell death (Lu *et al.* 2000, 2003). Furthermore, cleavage of APP at Asp664 is amplified in brains of AD subjects (Zhao *et al.* 2003) and APP transgenic mice (Galvan *et al.* 2006). In support of these findings, our group recently showed that introducing an Asp664 mutation [Asp → Ala; aspartate to alanine mutation at position 664 of APP (D664A)] prevents the neuropathological and behavioral deficits in an Alzheimer transgenic model, platelet-derived growth factor B-chain promoter-driven APP transgenic mice (PDAPP) mice (Galvan *et al.* 2006).

However, the mechanism(s) by which this neuroprotective effect is achieved, and in particular the resulting effect(s) on APP-mediated signal transduction, have not been explored. Based on data regarding (i) the pathophysiological actions of APP and those of its proteolytic products (Chang and Suh 2005; Reddy 2006), (ii) the knowledge that APP is a transmembrane protein similar to a signaling receptor (Okamoto *et al.* 1995), and (iii) identification of binding proteins at the C-terminus of APP (Chang *et al.* 2003; Russo *et al.* 2005), Aβ production and subsequent intracytoplasmic cleavage of APP may initiate a signaling cascade that is dependent on the Asp664 cleavage site and underlies the Aβ-directed neurodegeneration.

Although relatively little is known about signaling pathways activated by the APP cytoplasmic domain, the p21-activated kinase (PAK) family (subdivided into group I that includes isoforms PAK-1, PAK-2, and PAK-3 and group II that includes PAK-4, PAK-5, and PAK-6) has recently been identified in the pathogenesis of AD. In neuron cultures, PAK-3 interacts within the C100 region of APP and mediates C31-induced apoptosis (McPhie *et al.* 2003). In addition, PAK-1 and PAK-3 loss has been suggested to contribute to the dendritic spine defects and cognitive deficits seen in brains of AD subjects and APP Swedish mice (Zhao *et al.* 2006). Furthermore, PAK-5 binds Par-1 and suppresses τ hyperphosphorylation in early Alzheimer neurodegeneration (Matenia *et al.* 2005).The availability of Alzheimer model transgenic mice that are matched for their APP expression, Aβ production, and plaque number, but completely discordant with respect to AD-related sequelae such as synapse loss, electrophysiological abnormalities, and behavioral deficits (Galvan *et al.* 2006; Saganich *et al.* 2006), allows for the identification of underlying candidate signaling mechanisms.

II. Diabetes mellitus: Insulin resistance is present in one-quarter of the general population, predisposing these people to a wide range of diseases. Our aim was to identify cell-intrinsic determinants of insulin resistance in this population using induced pluripotent stem cell–derived (iPSC–derived) myoblasts (iMyos). We found that these cells exhibited a large network of altered protein phosphorylation in vitro. Integrating these data with data from type 2 diabetic revealed critical sites of conserved altered phosphorylation in IRS-1, AKT, mTOR, and TBC1D1 in addition to changes in protein phosphorylation involved in Rho/Rac signaling, chromatin organization, and RNA processing. There were also striking differences in the phosphoproteome in cells from men versus women. These sex-specific and insulin-resistance defects were linked to functional differences in downstream actions. Thus, there are cell-autonomous signaling alterations associated with insulin resistance within the general population and important differences between men and women, many of which also occur in diabetes, that contribute to differences in physiology and disease.

Insulin resistance is a major risk factor in the development of metabolic syndrome, type 2 diabetes (T2D), and cardiovascular disease. Indeed, the cardiometabolic syndrome

currently affects 20% to 30% of Westernized populations, and its prevalence continues to increase worldwide, with differing presentations in an age- and sex-specific manner. Although the impact of insulin resistance on glucose homeostasis and metabolic syndrome is well studied, 20% to 30% of nondiabetic people within the general population also have a substantial level of insulin resistance, and the molecular determinants underlying the insulin resistance in this population remain elusive. In individuals with a family history of T2D, insulin resistance precedes and predicts a high risk of developing the disease, whereas in individuals without a family history of diabetes, insulin resistance appears to be linked to increased risk for hyperlipidemia and accelerated atherosclerosis but not necessarily diabetes.

At the cellular level, insulin signaling is initiated by ligand binding leading to conformational change and transautophosphorylation of the insulin receptor (IR), which leads to activation of the receptor and phosphorylation of IR substrates, such as the IRS proteins and Shc. As a result, 2 major downstream signaling cascades are initiated: the Ras/MAP kinase pathway and the PI3K/Akt pathway. Insulin signaling also activates serine/threonine protein kinase mTOR C1 (mTORC1) to regulate protein translation and cell growth, stimulates glucose transport, inactivates glycogen synthase kinase 3 (GSK3), regulating glycogen synthesis, activates atypical PKCs mediating lipid metabolism, and leads to phosphorylation of FoxO1 and FoxK1/FoxK2, which serve as transcriptional regulators of insulin action. The insulin-signaling events play an important role in the regulation of cellular metabolism and growth in the classical insulin-responsive tissues, such as the muscle, liver, and adipose tissue. Given that skeletal muscle is the largest organ in the body and the primary site of glucose disposal, skeletal muscle insulin resistance largely contributes to dysregulation of whole-body glucose homeostasis. The goal of the current study was to investigate the cell-autonomous determinants of insulin resistance and phosphorylation-mediated signaling within the nondiabetic population using myoblasts derived from induced pluripotent stem cells (iPSCs) in vitro.

III. Cystic Fibrosis: Defects in processing and trafficking of cystic fibrosis transmembrane conductance regulator. Cystic fibrosis (CF) is caused by inherited mutations in the gene encoding the cystic fibrosis transmembrane conductance regulator (CFTR), a cAMP-regulated chloride channel expressed in epithelial tissues. Most mutations in CF patients result in rapid intracellular degradation of the CFTR protein. While this defect is thought to result from abnormal protein folding, it is unclear how mutant and wild-type (WT) proteins differ in structure, how the cell is able to distinguish these differences, and how the fate of the mutant protein is determined. By examining the initial steps of CFTR assembly into the endoplasmic reticulum (ER) membrane, it has recently been shown that CFTR utilizes two redundant translocation pathways to direct N-terminus folding events. Mutations that block one pathway therefore do not alter transmembrane topology, but rather appear to

disrupt intracellular trafficking through perturbations in higher order tertiary structure. These studies suggest that cellular quality control machinery acts at least in part, by monitoring proper interactions between CFTR subdomains. The end result of this process is the conversion of misfolded CFTR into a membrane-bound, polyubiquitinated complex. This complex recruits cytosolic degradation machinery to the endoplasmic reticulum membrane where CFTR is degraded as it is extracted from the lipid bilayer. Understanding how cellular machinery mediates this process will be an important step in designing strategies to modify protein folding and degradation in CF and related ion channelopathies.

The cystic fibrosis transmembrane conductance regulator (CFTR) is a complex, polytopic membrane protein expressed in the apical membrane of selected epithelial cells. CFTR functions directly as a cAMP regulated chloride channel[1] and also regulates the activity of other membrane proteins including the epithelial sodium channel (ENaC) and the outwardly rectifying chloride channel (ORCC). It thus plays a key role in the movement of ions and water across epithelial tissues. Not surprisingly, CFTR disruption results in a pleiotropic phenotype. The most profound effects are insufficiency of the exocrine pancreas, increase in sweat chloride concentration, male infertility and recurrent pulmonary infections. In most cystic fibrosis (CF) patients, chronic airway inflammation results in progressive pulmonary scarring, reduced lung function and ultimately, death. While CFTR is also expressed in the kidney (proximal and distal tubules, cortical collecting duct and inner medullary collecting duct), only mild renal abnormalities are observed in CF patients. These include decreased ability to excrete a salt load, mild urinary concentrating defects, increased proximal sodium reabsorption and altered drug excretion. CFTR is also expressed in the apical membrane of renal cysts in patients with polycystic kidney disease, and may play a role in chloride and fluid secretion into the cyst lumen.

More than 800 mutations in the *CFTR* gene have been identified in CF patients. These are broadly grouped into four classes: (I) defective protein synthesis; (II) defective protein processing; (III) defective ion conduction; and (IV) defective regulation of channel gating. Defective processing is by far the most common mechanism of protein disruption, accounting for more than 2/3 of clinical CF cases. While these latter patients synthesize adequate amounts of functional CFTR protein, the mutant protein is rapidly degraded prior to reaching the plasma membrane. To understand the molecular basis of CF it will therefore be necessary to define the molecular mechanism(s) by which CFTR is folded, assembled and packaged into cellular membranes and trafficked through cells.

CFTR is a member of the ATP binding cassette (ABC) transporter superfamily. It contains two hydrophobic domains (each with six predicted transmembrane (TM) segments), two cytosolic nucleotide binding domains (NBDs) and a cytosolic regulatory (R) domain. Like most eukaryotic membrane proteins, CFTR is synthesized and assembled in the endoplasmic reticulum (ER). During the earliest steps in this process, nascent chain-ribosome complexes are targeted to the ER membrane, and TM segments

are precisely oriented and integrated into the lipid bilayer. Additional biogenesis events involve the packing of transmembrane helices, folding of cytosolic domains, and finally, assembly of these domains into a mature tertiary structure. This process is mediated by specialized cellular machinery that includes the Sec61 translocation complex and cytosolic (hsp70, hsp40) as well as ER (calnexin) chaperones that assist folding and prevent aggregation of folding intermediates. CFTR maturation is thus a stepwise and compartmentalized process that coordinates folding of different protein domains in the lipid environment of the ER membrane, the oxidizing environment of the ER lumen and the reducing environment of the cytosol Figure 1. At the center of this process is a stringent quality control mechanism capable of discriminating normally folded from abnormally folded proteins. Quality control machinery thus prevents misfolded CFTR from exiting the ER compartment and is responsible for its degradation via the cytosolic ubiquitin/proteasome pathway.

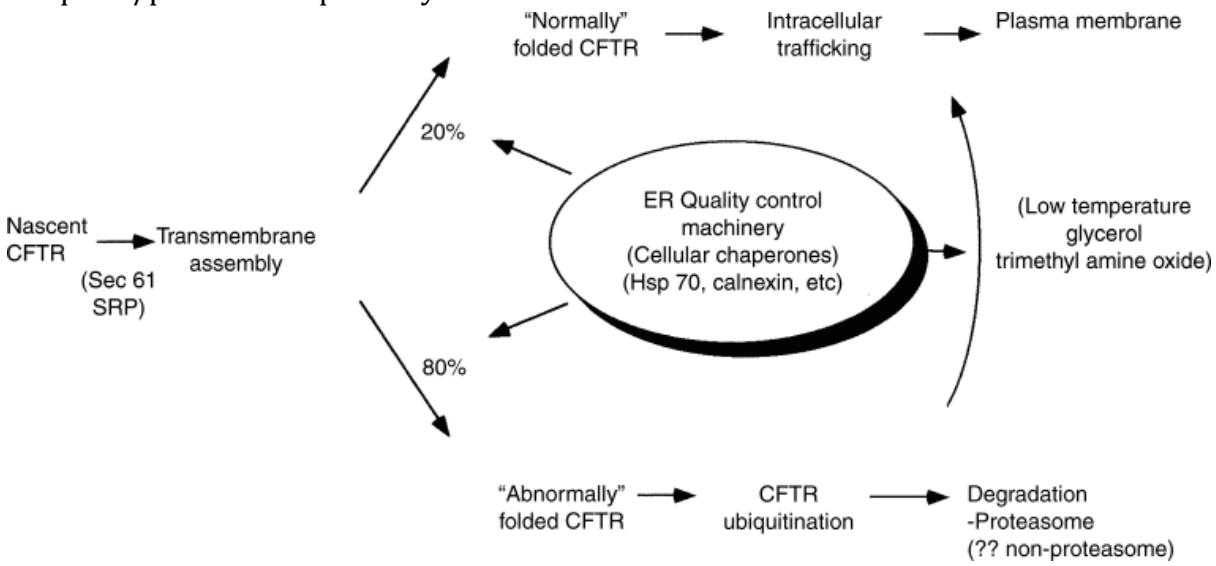


Figure: Cystic fibrosis transmembrane conductance regulator (CFTR) biogenesis and degradation. Synthesis of nascent CFTR begins on cytosolic ribosomes that are targeted via SRP to the Sec61 complex in the endoplasmic reticulum (ER) membrane. As translation continues, transmembrane topology is established in the ER membrane and folding and assembly of membranous, lumenal and cytosolic domains is facilitated by cellular chaperones. For unclear reasons, only 20% of WT CFTR is normally trafficked out of the ER compartment where it then undergoes gradual turnover. Most CFTR protein (~80%) fails to exit the ER, undergoes polyubiquitination and is rapidly degraded by the 26S proteasome. For many CFTR mutants (such as αδF508), ER degradation accounts for nearly 100% of newly synthesized protein despite the fact that the mutant protein forms functional chloride channels in the ER membrane. A major challenge in CF is therefore to devise methods to rescue CFTR into productive folding pathways. This strategy has been proved in principle using reduced temperature, and molecular chaperones such as glycerol, and trimethylamine oxide

MECHANISM OF CFTR ASSEMBLY INTO THE ER MEMBRANE:

Conventional models predict that the transmembrane topology of polytopic proteins is established through the action of sequential signal and stop transfer sequences as the nascent chain emerges from the ribosome. Signal sequences function to target nascent chains to the ER, facilitate ribosome binding to the membrane, and open a large aqueous translocation channel through which the elongating nascent chain moves to reach the ER lumen. Stop transfer sequences terminate ongoing translocation, disrupt the ribosome-membrane junction, close the translocon, and direct the hydrophobic TM helix laterally out of the translocon and into the lipid bilayer. A large number of molecular events must therefore be precisely coordinated as polytopic proteins such as CFTR are "stitched" into the ER membrane.

If CFTR followed a conventional biogenesis model, then the first TM segment (TM1) should encode signal sequence activity capable of orienting the N-terminus in the cytosol and the first extracellular loop (ECL1) in the ER lumen. Surprisingly, when TM1 signal sequence activity was tested in a defined heterologous cassette, TM1 was unable to efficiently initiate translocation or span the membrane. This suggested either that CFTR assembly into the ER membrane was inefficient, or that topogenic information in addition to TM1 was required for N-terminus transmembrane assembly. Subsequent analysis confirmed the latter prediction by demonstrating that TM2 also functioned as a signal sequence with translocation specificity complimentary to that of TM1. Moreover, by simultaneously disrupting signal sequence activities of TM1 and TM2, it was shown that TM2 was able to independently orient TM1 and ECL1 in the ER membrane after this region had been synthesized in the cytosol. This post-translational translocation activity of TM2 was ribosome dependent, indicating that TM2, like TM1, utilized established translocation machinery (for example, signal recognition particle) for ER targeting.

Mutagenesis studies of TM1 and TM2 thus define two alternate translocation pathways by which CFTR acquires its proper N-terminus transmembrane topology (diagrammed in Figure 2). For a minority of nascent chains, TM1 functions as a signal sequence to initiate translocation into the ER lumen. In these chains, TM2 stop transfer activity terminates ongoing translocation and establishes the membrane boundaries of TM1, TM2 and ECL1. In chains where TM1 fails to start translocation, however, TM1 and TM2 emerge from the ribosome into the cytosol where TM2 initiates translocation of its N-terminus flanking residues. Here, TM1 functions in the capacity of a stop transfer sequence and is positioned post-translationally into its proper orientation. In this manner, TM2 provides a backup mechanism for ensuring proper topology in chains where TM1-mediated translocation has failed.

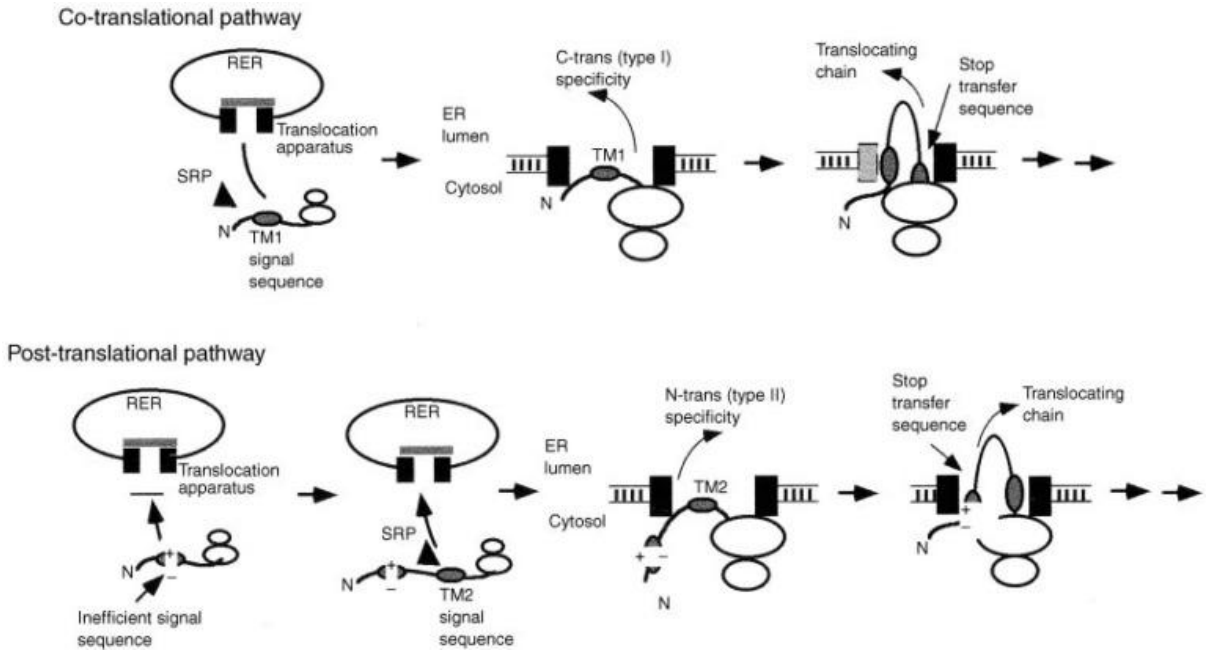


Figure **Alternate pathways for CFTR N-terminus transmembrane assembly.** In the conventional or cotranslational pathway, CFTR topology is established through sequential action of TM1 signal sequence activity and TM2 stop transfer activity. During this process the TM1 gates the translocon open and the nascent chain translocates into the ER lumen in an N→C terminus direction as it emerges from the ribosome. The post-translational pathway is utilized by most (>60%) of WT chains and essentially all G85E and G91R mutant chains. Here, TM2 acts as the initial signal sequence to start translocation which proceeds in a C→N terminus direction (designated by arrow). In both pathways, the final topology of the nascent chain is equivalent; TM1 and TM2 each span the membrane and the intervening peptide loop, ECL1, resides in the ER lumen. Ribosomes (open circles), TM segments (shaded ovals), translocon channel (black rectangles), and lipid bilayer are indicated.

Further analysis indicated that two charged residues located within the hydrophobic membrane spanning core of TM1 (E92 and K95) were responsible for the weak TM1 signal sequence activity. Mutating these residues to alanine markedly improved the ability of TM1 to direct translocation[13] but completely disrupted CFTR chloride channel activity in *Xenopus* oocytes (unpublished observations). Thus, for CFTR, structural features required for protein function (such as residues E92 and K95) directly conflict with structural features necessary to direct CFTR topology via the cotranslational pathway Figure 2. The presence of TM2 signal sequence activity, by providing an alternate mechanism to ensure CFTR topology, therefore enables TM1 to contain the necessary charged residues. This increased sequence diversity within TM1 would not have been possible if CFTR biogenesis were restricted solely to the conventional mode of biogenesis. While CFTR provides the first example of this type of redundancy in topogenic pathways, it seems likely that other polytopic proteins, particularly those with specialized structural requirements, will exhibit additional variations in transmembrane assembly.

EFFECTS OF INHERITED MUTATIONS ON CFTR TRANSMEMBRANE ASSEMBLY:

Two CF mutations, G85E and G91R, each introduce an additional charged residue within the hydrophobic core of TM1. These mutations also disrupted CFTR chloride efflux in microinjected *Xenopus* oocytes by preventing newly synthesized protein from exiting the ER[33]. This suggested that G85E and G91R CFTR mutants failed to fold properly and were recognized by ER quality control machinery similar to the common αδF508 mutant. To understand how charged residues within TM1 influenced CFTR folding, we compared N-terminus transmembrane assembly and topology in WT and mutant chains. Topologic analysis revealed that each mutation completely eliminated TM1 signal sequence activity but had no effect on CFTR topology. Thus, in these mutant chains, TM2 was entirely responsible for directing translocation of ECL1. More importantly, because mutant TM1 and TM2 each spanned the membrane in their native orientations, ER quality control machinery must have been able to detect the presence of the aberrant charged residues localized within the plane of the lipid bilayer. To determine how this might occur, WT and mutant CFTR constructs were truncated after the second transmembrane segment, the first transmembrane domain (TM6), NBD1 or the R domain. Expression of these constructs expressed in *Xenopus* oocytes demonstrated that cellular quality control machinery was effectively able to distinguish WT from mutant chains only after synthesis of the R domain had been completed.

These studies demonstrated that in order for CFTR to acquire a stable structure in the ER membrane, multiple protein domains must be synthesized and properly assembled. In addition, they indicated that G85E and G91R mutations likely interfered with late, rather than early, assembly events required for CFTR tertiary structure. A subtle alteration in the first transmembrane domain such as the insertion of a charged residue may thus indirectly influence folding interactions at distant sites in the molecule. This provides an intriguing model as to how mutant proteins might be recognized by ER quality control machinery, and how mutations in diverse regions of CFTR could give rise to similar trafficking phenotypes. If ER quality control machinery recognized structural interfaces between CFTR subdomains, then subtle structural changes that influence the strength and/or kinetics of these interactions could be recognized by quality control machinery in much the same manner as unassembled oligomeric subunits. ER quality control machinery would therefore not be required to recognize each local structural perturbation, but rather it might serve to monitor more global aspects of protein compaction.

Finally, it should be noted that “abnormal” protein folding in terms of ER quality control is operational and entirely based on a cellular response, namely ER associated degradation. While it is often tempting to view the acquisition of protein function as a criteria for “normal” folding, in the case of CFTR this is not necessarily correct. CFTR protein containing the αδF508 mutation is clearly capable of forming cAMP gated chloride channels with nearly normal conduction properties[35]. Yet essentially 100% of αδF508 CFTR is degraded in the ER. Conversely, WT CFTR protein truncated after the R

domain at residue #836 is nearly as stable as full length protein in *Xenopus* oocytes, yet its chloride channel activity is <5% of wild-type (unpublished observations). Thus, protein maturation in a functional sense may be distinct from structural maturation as determined by ER quality control machinery. This process is further complicated by observations that the efficiency of intracellular trafficking differs markedly between cell systems. In mammalian cells, 80% of WT and ~99% of αδ F508 CFTR is degraded in the ER, while in *Xenopus* oocytes, <10% of WT and ~80% of αδF508 is degraded in the ER. It is unknown whether these differences in intracellular trafficking reflect different folding efficiencies or alternatively, different stringencies in quality control systems. In either case, understanding the relationship between CFTR quality control and CFTR functional maturation will likely require detailed structural studies and the identification of cellular components responsible for discriminating subtle structural differences.

CFTR DEGRADATION BY THE UBIQUITIN/PROTEASOME PATHWAY:

The hallmark of abnormal CFTR processing and trafficking is rapid degradation of CFTR protein in a pre-Golgi and lysosome-independent compartment. Surprisingly, several studies have now demonstrated that the cytosolic ubiquitin-proteasome pathway plays a key role in ER associated degradation not only of CFTR, but also of a wide variety of misfolded secretory, bitopic and polytopic protein substrates[17,36]. In the ubiquitin proteasome pathway, substrates are first modified by covalent addition of multiple ubiquitin moieties through the action of cytosolic (and/or membrane bound) ubiquitin activating (E1), conjugating (E2), and ligating (E3) enzymes [reviewed in[37]. Polyubiquitinated proteins are then recognized by the cytosolic 26S proteasome complex; ubiquitin chains are removed; and the substrate is digested into small peptide fragments. These observations require that cytosolic degradation machinery gains access to proteins in the lumen (or membrane) of the ER, and suggest that translocation across the ER membrane is a bidirectional process that is regulated in part by the folded state of a given protein.

The degradation of polytopic proteins by cytosolic proteases poses an additional topologic challenge in that multiple transmembrane helices must be removed from the bilayer. Then, at what stage of biogenesis is CFTR recognized for degradation, and once recognized, how is CFTR delivered to the cytosolic proteolytic complex? Recently, Sato, Ward and Kopito used an in vitro expression system to demonstrate that CFTR ubiquitination might actually begin prior to the completion of protein synthesis[39]. Using a similar rabbit reticulocyte lysate-based expression system, we showed that full length and membrane integrated CFTR is also a substrate for polyubiquitination Figure 3. CFTR ubiquitination required cytosolic components as well as ATP. By allowing ubiquitination to occur in the presence of the proteasome inhibitor hemin, we demonstrated that ubiquitinated CFTR remained tightly associated with the ER membrane until it was degraded into trichloroacetic acid-soluble fragments.

Furthermore, pre-ubiquitinated, membrane-bound CFTR could be degraded only in the presence of additional cytosol.

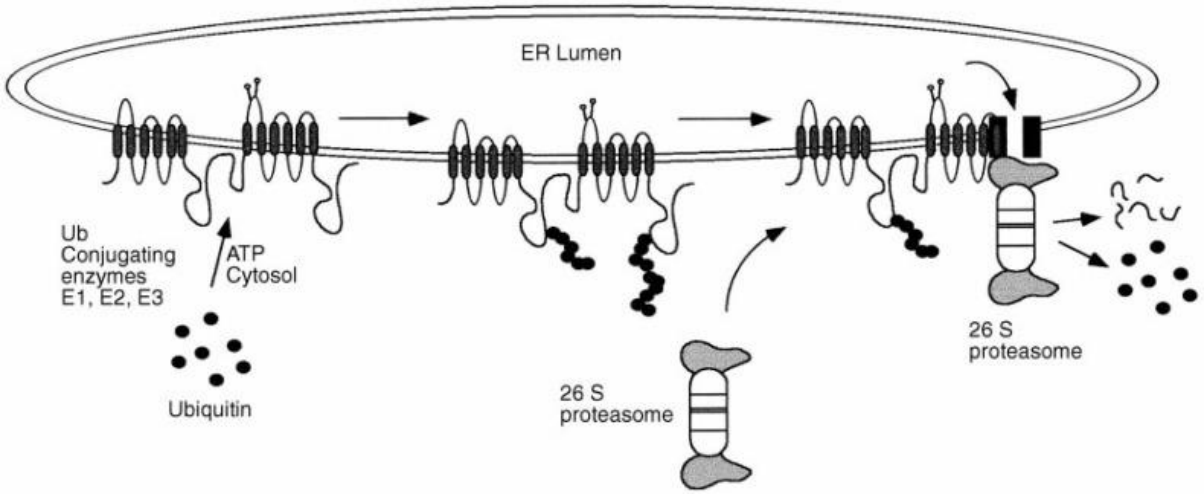


Figure . **Proteasome-mediated CFTR degradation**. Full length, membrane-integrated and glycosylated CFTR undergoes polyubiquitination at the ER membrane in an ATP- and cytosol-dependent manner. This complex remains tightly bound to the ER membrane and recruits cytosolic proteolytic machinery that includes the 26S proteasome. During degradation, ubiquitin moieties are removed and CFTR is cleaved into small (TCA soluble) peptide fragments as the protein is extracted from the ER membrane. One possibility is that ATPase activity within the 19S proteasome regulatory subunit facilitates CFTR unfolding and/or membrane extraction.

These results suggest that polyubiquitinated CFTR is involved in recruiting cytosolic degradation machinery directly to the ER membrane, consistent with studies by Rivett, Palmer and Knecht, which demonstrate that proteasomes are bound to the ER in living cells. Because hemin inhibits the proteasome by blocking ATPase activities within the 19S subunit (PA700), it is also possible that the unfolding activity of the proteasome itself might be involved in extracting CFTR from the membrane. This would explain our observation that CFTR degradation was tightly coupled to extraction of TM helices from the lipid bilayer. It would also explain why proteasome inhibitors such as peptide aldehydes and/or lactacystin that directly inactivate the catalytic active site in the proteasome, but do not effect ATPase activity, might give rise to cytosolic intermediates of ER degradation substrates.

CFTR PROCESSING AND TRAFFICKING AS A PARADIGM FOR ION CHANNELOPATHIES:

In the past decade medical research has uncovered the genetic basis for an expanding group of ion channelopathies. A significant challenge in the next decade will be to decipher the underlying cellular and metabolic pathways that are influenced by these

mutations. This will involve: (*1*) identifying the molecular components and steps that regulate normal biosynthetic processes; (*2*) defining how specific mutations influence these pathways; and (*3*) devising strategies for controlled manipulation of these pathways in human disease. In this regard, CF has served as a key example for understanding fundamental mechanisms of protein biogenesis, folding and degradation. It seems highly likely that these studies will have far reaching implications, and that futureefforts to understand and correct the CF defect at the cellular level will impact an ever growing variety of ion channelopathies.

Cystic fibrosis results from mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) chloride channel, leading to defective apical chloride transport. Patients also experience overactivation of inflammatory processes, including increased calcium signaling. Many investigations have described indirect effects of calcium signaling on CFTR or other calcium-activated chloride channels; here, we investigate the direct response of CFTR to calmodulin-mediated calcium signaling. We characterize an interaction between the regulatory region of CFTR and calmodulin, the major calcium signaling molecule, and report protein kinase A (PKA)-independent CFTR activation by calmodulin. We describe the competition between calmodulin binding and PKA phosphorylation and the differential effects of this competition for wild-type CFTR and the major F508del mutant, hinting at potential therapeutic strategies. Evidence of CFTR binding to isolated calmodulin domains/lobes suggests a mechanism for the role of CFTR as a molecular hub. Together, these data provide insights into how loss of active CFTR at the membrane can have additional consequences besides impaired chloride transport. In airway epithelia, transport of chloride into the lumen of the airways is followed by water movement and allows normal clearance of the respiratory system to prevent lung infections. This chloride flux is maintained by cAMP-activated cystic fibrosis transmembrane conductance regulator (CFTR) and calcium ion-activated chloride channels (CaCCs); the separation between activation pathways of these channels was recently questioned, however. In cystic fibrosis (CF), a disease caused by mutations in the CFTR gene, lack of normal fluid transport results in a dehydrated airway surface liquid (ASL) layer that leads to mucus accumulation and frequent lung infections. Calcium signaling is the major regulatory pathway of airway physiology because increased intracellular calcium is the primary signal for fluid secretion. Despite early investigations finding no evidence for calcium signaling directly activating CFTR, recent studies have focused on the interplay between the cAMP and calcium pathways in CFTR activation and function, including the role of protein kinase C (PKC) phosphorylation on CFTR activation and PKA activation by calcium-activated adenylate cyclases.

CFTR activation by the cAMP pathway is well established in the literature. Together with ATP binding by the nucleotide-binding domains (NBDs), phosphorylation of NBD1 and the 200-residue regulatory (R) region facilitates CFTR trafficking and channel

opening. R region is an intrinsically disordered segment of the CFTR that samples multiple conformations under physiological conditions. It is responsible for most of CFTR's regulatory intramolecular and intermolecular protein–protein interactions. Diverse binding elements of R region for various partners are conserved and controlled by phosphorylation. Nonphosphorylated R region interacts with NBD1 via helical elements, whereas PKA-phosphorylated R region loses helical propensity and binds to 14-3-3 via shorter extended segments. Because these binding segments are accessible and largely independent from each other, R region can interact with more than one partner at a time. Consequently, it can integrate different regulatory inputs via transient, dynamic interactions.

CFTR is located on the apical surface of epithelial cells and is a key component of a macromolecular signaling complex that involves sodium and potassium channels, anion exchangers, transporters, and other regulator proteins and molecules. Store operated calcium channel Orai1 was also found in the same microdomain; because mutations in CFTR affect Orai1 channel function, Orai1 is likely to be part of the same complex. Interaction between endoplasmic reticulum (ER)-resident stromal interaction molecule 1 (STIM1) and Orai1 form a membrane contact site, a critical junction between the ER and plasma membrane. This STIM1:Orai1 interaction activates the store operated calcium entry and localizes ER and plasma membrane calcium channels in close proximity to CFTR, resulting in local elevation of calcium levels. Calcium-regulated interaction between CFTR and other membrane proteins is thus very likely. One recent example demonstrates that the potassium channel KCa3.1 interacts with CFTR in a calcium-dependent manner.

Calcium signaling is tightly coupled to calmodulin, a multifunctional intermediate messenger that translates calcium signals to regulatory calcium-dependent protein–protein interactions with various targets (18, 19). It consists of two lobes, each with two calcium-binding EF-hand motifs, and a connecting linker region. Calmodulin recognizes very diverse substrate sequences categorized (e.g., IQ, 1–10, 1–14, and 1–16 motif classes) based on key residues of the interaction. Moreover, it can interact with its partners in two different binding conformations. In the canonical binding mode, calmodulin wraps around the substrate with high (nanomolar) affinity, whereas in the alternative binding conformation the two lobes can bind to two separate sequences independently and bridge over larger distances, usually with weaker (micromolar) affinity. Most calmodulin substrates use the canonical interaction mode, but, interestingly, membrane channels such as the calcium channel Orai1, the sodium channel NaV1.5, and the potassium channel Kv7.1 use the alternative calmodulin-binding mode.

Probable Questions:

1. Discuss signaling defects found in Alzehimer's disease.
2. Discuss signaling defects found in Cystic Fibrosis.
3. Discuss signaling defects found in Diabetes mellitus.

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

UNIT XIV

Human disease and therapeutic drugs targeting GPCR, JAK STAT and TLR signaling pathways

Objective: In this unit we will learn about therapeutic drugs which target various signaling pathways such as GPCR, JAK STAT and TLR signaling pathways.

G-Protein Coupled Receptor Pathways:

Estimates vary regarding the number of G protein-coupled receptors (GPCRs), the largest family of membrane receptors that are targeted by approved drugs, and the number of such drugs that target GPCRs. We review current knowledge regarding GPCRs as drug targets by integrating data from public databases (ChEMBL, Guide to PHARMACOLOGY, and DrugBank) and from the Broad Institute Drug Repurposing Hub. To account for discrepancies among these sources, we curated a list of GPCRs currently targeted by approved drugs. As of November 2017, 134 GPCRs are targets for drugs approved in the United States or European Union; 128 GPCRs are targets for drugs listed in the Food and Drug Administration Orange Book. We estimate that ~700 approved drugs target GPCRs, implying that approximately 35% of approved drugs target GPCRs. GPCRs and GPCR-related proteins, i.e., those upstream of or downstream from GPCRs, represent ~17% of all protein targets for approved drugs, with GPCRs themselves accounting for ~12%. As such, GPCRs constitute the largest family of proteins targeted by approved drugs. Drugs that currently target GPCRs and GPCR-related proteins are primarily small molecules and peptides. Since ~100 of the ~360 human endo-GPCRs (other than olfactory, taste, and visual GPCRs) are orphan receptors (lacking known physiologic agonists), the number of GPCR targets, the number of GPCR-targeted drugs, and perhaps the types of drugs will likely increase, thus further expanding this GPCR repertoire and the many roles of GPCR drugs in therapeutics. Based on their canonical structure, G protein-coupled receptors (GPCRs), which are sometimes termed heptahelical or 7-transmembrane receptors, are the largest family of membrane receptors in humans and numerous other species. In addition, GPCRs are considered the largest family of targets for approved drugs (Allen and Roth, 2011; Rask-Andersen et al., 2014; Santos et al., 2017). Numerous factors contribute to the wide utility of GPCR-targeted drugs, including their druggability, interaction with numerous types of chemical entities, and expression in the plasma membrane, which facilitates molecular interactions in the extracellular milieu. Scientific articles, grant applications, and lectures that describe findings related to GPCRs often note the therapeutic utility of GPCRs but differ widely in estimates of their contribution as targets for approved drugs, which generally range from ~20% to >50%.

JAK-STAT Pathways:

JAK/STAT signaling pathway is one of the important regulatory signaling cascades for the myriad of cellular processes initiated by various types of ligands such as growth factors, hormones, and cytokines. The physiological processes regulated by JAK/STAT signaling are immune regulation, cell proliferation, cell survival, apoptosis and hematopoiesis of myeloid and non-myeloid cells. Dysregulation of JAK/STAT signaling is reported in various immunological disorders, hematological and other solid malignancies through various oncogenic activation mutations in receptors, downstream mediators, and associated transcriptional factors such as STATs. STATs typically have a dual role when explored in the context of cancer. While several members of the STAT family are involved in malignancies, however, a few members which include STAT3 and STAT5 are linked to tumor initiation and progression. Other STAT members such as STAT1 and STAT2 are pivotal for antitumor defense and maintenance of an effective and long-term immune response through evolutionarily conserved programs. The effects of JAK/STAT signaling and the persistent activation of STATs in tumor cell survival; proliferation and invasion have made the JAK/STAT pathway an ideal target for drug development and cancer therapy. Therefore, understanding the intricate JAK/STAT signaling in the pathogenesis of solid malignancies needs extensive research. A better understanding of the functionally redundant roles of JAKs and STATs may provide a rationale for improving existing cancer therapies which have deleterious effects on normal cells and to identifying novel targets for therapeutic intervention in solid malignancies.

An intracellular signaling pathway is critical in regulating the cellular fate and modulates phenotypic modifications. JAK/STAT signaling is one such pathway that regulates embryonic development, stem cell maintenance, hematopoiesis and, inflammatory response (Bowman et al., 2018). JAK/STAT is a signal transduction pathway, which transmits the extracellular information or signals through a transmembrane protein, called Janus kinase (JAK). The JAK further directs the signal to an intracellular environment by phosphorylating the transcription factor known as STATs, which translocates into the nucleus to target the promoter region of a gene to regulate the mechanism of transcription (Fiebelkow et al., 2021). The evolutionarily conserved pathway in all eukaryotes, JAK/STAT is a principal signal transduction pathway in mammals for cytokines and growth factors (O'Shea et al., 2015). Structurally there are four members of the JAK family i.e., JAK1, JAK2, JAK3, Tyk2, and seven members of STAT, i.e., STAT1, STAT2, STAT3, STAT4, STAT5A, STAT5B, and STAT6 in mammals (Abroun et al., 2015; Alunno et al., 2019) (Figure 1). As the ligands bind to the cognate receptor, the two JAK's come closer and allow *trans*-phosphorylation of both the receptors as well as STATs at their conserved tyrosine residue present near to the C-terminal region (Morris et al., 2018). The phosphorylation of tyrosine at the C-terminal region is responsible for the dimerization of STATs which further enhances the interaction of a conserved domain called as SH2 domain (Morris et al., 2018). STAT is a transcription factor that resides in the cytoplasm. The phosphorylated STATs then enter the nucleus from the cytoplasm through a mediator called importin α-5 and Ran nuclear import (Seif et al., 2017). These dimerized STATs then bind to the particular regulatory sequence for the activation or repression of the targeted genes (Seif et al., 2017) (Figure

2). Thus, JAK/STAT is a signal transduction pathway that converts the extracellular signals into the transcriptional message thereby regulating physiological processes. Aberrant activation of intracellular signaling pathways confers malignant phenotype to genetically and metabolically altered cells (Flavahan et al., 2017). Many of these alterations occur in signaling pathways that control cell growth and division, cell death, cell fate, cell motility, tumor microenvironment, angiogenesis, and inflammation (Jing et al., 2019). Besides regulating physiological processes, JAK/STAT signaling cascade is implicated in various pathophysiological disorders including malignancies by altering the JAK/STAT signaling. In solid malignancies, these alterations characteristically support progression from a relatively benign group of proliferating cells (hyperplasia) to a mass of cells with abnormal morphology (dysplasia), cytological appearance, cellular organization, genomic integrity, and metabolic state (Gale et al., 2016). As a result of the structural and functional complexity of solid tumors, cancer therapies exhibit variable responses in distinct patients and cancer types (Cabrera et al., 2015). Although, the number of signaling pathways are deregulated in malignancies, however; besides being invariably activated in hematological malignancies, JAK/STAT signaling is, altered in many solid tumors and showed deregulated activation (Joshi et al., 2015). STATs are the effective downstream mediators of the JAK/STAT signaling cascade. STATs regulate the expression of a wide variety of genes both positively and negatively, identification of many of them is still imprecisely determined and warrants further investigation (Stark et al., 2018). The combination of STATs and the tissue in which they function at a given time often determines the subset of genes they control, which further contributes to the challenge of target gene identification. Constitutive phosphorylation of STAT1, STAT3, and STAT5 has been detected in many tumor cell models (Bellucci et al., 2015). The microarray-based expression analysis is used to comprehensively identify STAT target genes. Although STATs contributes to a malignant phenotype by regulating genes involved in cellular proliferation, survival, differentiation, angiogenesis, and invasion (Stark et al., 2018). However, how STATs execute these biological functions is critical for understanding the pathophysiology of solid tumors, signifying the need to understand the more precise role of JAK/STAT pathway in the pathobiology of solid malignancies. Despite this, some differences in STAT targets among cell types may be due to epigenetic variation among cells, which includes altered histone modifications or DNA methylation (Biswas and Rao, 2017; Garg et al., 2018). Thus, there is a need to identify the new target genes of STATs to fully understand how JAK/STATs can be used for the therapeutic intervention of solid tumors. Indeed, solid tumors contain aberrantly activated transcription factors—either through mutation of the transcription factor itself or through mutation of upstream signaling cascades leading to its activation (Qureshy et al., 2020; Gilmore, 2021). Therefore, understanding the genes regulated by STATs may provide insights into the pathophysiology of solid tumors in which they are inappropriately activated and may unravel novel targets for therapeutic intervention. Targeting intracellular signaling pathways has been a productive strategy for drug development, with several drugs acting on JAK-STAT signaling already in use and many more are being developed. In this review, we provide a comprehensive review of the role of the JAK/STAT pathway in solid tumors, clinical evidence of targeted agents, and also discuss the therapeutic intervention of JAK/STAT in solid tumors.

JAK/STAT is one of the versatile signaling pathways which has been extensively studied for its crucial role in tumor progression. The potential crosstalk of JAK/STAT with multiple alternative pathways has made it a promising target for the development of new lead molecules. Aberrant activation of JAK/STAT signaling has been frequently observed in a wide range of malignant neoplasms. A number of studies have provided compelling evidence that inhibition of the JAK/STAT pathway provides significant therapeutic benefits. Several JAK/STAT inhibitors (natural as well as synthetic) have been found to modulate the expression of molecules involved in the JAK/STAT signaling network. Preclinical studies in cancer models have shown the effect of various inhibitors of JAK/STAT signaling which resulted in inhibition of cellular proliferation and tumor progression.

Thus, JAK/STAT signaling appears to be an important target with the potential for a high therapeutic index. Although it takes more than two decades to get approval from FDA to target JAK/STAT signaling. In the coming future, we expect more specific compounds with the least deleterious effects on normal cells which target particularly the kinase activity of JAK/STAT mediators to attenuate its amplification in malignancies. However, the big question is whether next-generation compounds/inhibitors will be able to attenuate the functional part of mutated JAKs and leave wild-type JAKs unaffected. This approach of inhibition could decrease the deleterious effects of compounds/inhibitors on normal cells and will improve the drug efficacy. Besides various preclinical and experimental settings, effective STAT inhibitors in clinical settings are still a dream to come true. Although targeting transcriptional factors is not common, however, this approach will lead to more specific inhibition and in the coming future could be a promising therapeutic approach against JAK/STAT signaling in solid malignancies. However, it is important to validate the mechanism of action of small molecule inhibitors (SMIs) of the JAK-STAT pathway before conclusions can be drawn about their clinical use. The hope is to administer SMIs of JAK-STAT pathway with minimum risk to human subjects and that these inhibitors will produce a significant therapeutic effect against the solid malignancies.

TLR Signalling Pathways:

Toll-like receptors (TLRs) are sentinel receptors of the host innate immune system that recognize conserved ‘pathogen-associated molecular patterns’ of invading microbes, including viruses. The activation of TLRs establishes antiviral innate immune responses and coordinates the development of long-lasting adaptive immunity in order to control viral pathogenesis. However, microbe-induced damage to host tissues may release ‘danger-associated molecular patterns’ that also activate TLRs, leading to an overexuberant inflammatory response and, ultimately, to tissue damage. Thus, TLRs have proven to be promising targets as therapeutics for the treatment of viral infections that result in inflammatory damage or as adjuvants in order to enhance the efficacy of

vaccines. Here, we explore recent advances in TLR biology with a focus on novel drugs that target TLRs (agonists and antagonists) for antiviral therapy.

A significant advance in the field of immunology accompanied the identification of the two arms of immune responses as 'innate' and 'adaptive'. Initially, innate immunity was considered to be a relatively nonspecific and simple part of the overall immune response, while adaptive immunity was believed to provide antigen-specific protection from microbial and viral infection. However, accumulating evidence has clearly established that innate immune responses are the first line of defense against invading pathogens and also coordinate the development of a pathogen-specific adaptive immune response. Despite striking differences in terms of response timing, effector cells and recognition receptors on the cells of both of these systems, there are many cellular and molecular components in common that orchestrates highly specific, integrated responses against invasive pathogens and establishes long-term immune memory.

The primary receptors of innate immunity are a diverse set of germ line-encoded 'pattern-recognition receptors' (PRRs) that identify a broad spectrum of 'pathogen-associated molecular patterns' (PAMPs), which are diverse microbial structures of invading microorganisms, or 'danger-associated molecular patterns' (DAMPs), which are host-derived molecules released by stressed or injured cells. PAMPs include diverse microbial molecules, such as lipopolysaccharides (LPSs), lipopeptides, peptidoglycans, mannans, flagellin, bacterial and viral nucleic acids and viral envelope proteins, whereas examples of DAMPs include endogenous (host) components, such as histones, nucleic acids, uric acid crystals, cytochrome C, ATP, oxidized 1-palmitoyl-2-arachidonoyl-phosphaticylcholine and HMGB1, among others. In 1989, Janeway first proposed the concept of PRRs that recognize the molecular structures of microorganisms and link innate and adaptive immune responses. Two important discoveries strengthened Janeway's concept of PRRs: the proof of the importance of Toll-mediated signaling in the induction of antifungal peptides by *Drosophila* in response to infection; and the positional cloning of the *Lps* gene (now known to be *Tlr4*). Both C3H/HeJ and C57BL/10ScCr mice were shown to express mutations in this gene that led to LPS hyporesponsiveness in these two strains. The importance of this research resulted in the sharing of the 2011 Nobel Prize in Physiology or Medicine.

Among the various families of PRRs (e.g., Toll-like receptors [TLRs], Nod-like receptors, RIG-I-like receptors [RLRs], c-type lectin receptors and cytosolic DNA receptors), TLRs are one of the largest and best-studied families of PRRs. The study of TLR biology has provided molecular insights into how PRRs recognize PAMPs and DAMPs, and how this leads to the activation of signaling cascades that converge in order to induce the expression of proinflammatory cytokines, chemokines and antiviral interferons (IFNs). Moreover, the discovery of TLRs has also enabled the identification of other families of

innate immune receptors. Together, the knowledge generated in the field of TLRs over the past decade has resulted in a significant paradigm shift in our understanding of innate immunity and its role in the development of a long-lasting, pathogen-specific adaptive immune responses.

The evidence thus far points to a role for TLRs in immune and inflammatory diseases, including allergies, autoimmune disorders and cancer. To date, many viruses have been shown to activate the innate immune system through TLRs, assigning to TLRs an important role in controlling viral infections. The following sections provide examples of TLR–virus interactions and their outcomes, as well as recent advances in our understanding of the role of TLRs in antiviral innate immunity, with a focus on the studies designed for developing novel TLR-targeting drugs that exert antiviral activity or serve as adjuvants for vaccines. TLRs are evolutionarily conserved across a wide range of species; 10 human and 12 mouse TLRs have been identified. TLRs are type I transmembrane proteins composed of an N-terminal leucine-rich repeat (LRR) domain that enables the recognition of a wide variety of ligands, a single transmembrane-spanning domain and a conserved cytoplasmic Toll–IL-1 receptor resistance (TIR) domain for downstream signal transduction. Resolution of the crystal structure of TLR3 revealed that a horseshoe-shaped antigen binding core was formed by the LRR-containing domain and sequence homology analyses indicated that activation of all TLRs requires a common tertiary structure. Ten functional human TLRs (TLR1–10) and 12 functional mouse TLRs (TLR1–9 and TLR11–13) can be categorized by their subcellular localization. TLR1, 2, 4–6 and 10 localize at the cell surface, whereas TLR3, 7–9 and 11–13 reside in endosomes and/or the endoplasmic reticulum. Most TLRs form homodimers after ligand binding through their LRRs; however, TLR2 typically forms heterodimers with either TLR1, TLR6 or possibly with TLR10 and detects components of microbial cell walls and membranes, such as lipopeptides, peptidoglycan, porins and mannan, while TLR11 heterodimerizes with TLR12 in order to bind to the profilin protein of the parasite *Toxoplasma gondii* . TLR4 recognizes LPS from Gram-negative bacteria, the fusion (F) protein of respiratory syncytial virus (RSV), the mouse mammary tumor virus and Ebola virus glycoprotein. In addition, TLR4 also senses DAMPs, including oxidized 1-palmitoyl-2-arachidonoyl-phosphaticylcholine, which is a host oxidized phospholipid that is produced due to oxidative stress in response to acute lung injury by acid aspiration, infection by respiratory viruses or bacteria or exposure to microbial products, and HMGB1, which is a chromatin binding protein that is released upon pyroptosis. TLR5 detects flagellin, the major protein of bacterial flagella, whereas the ligand for TLR10 has not yet been identified. Homodimers of mouse TLR11 recognize components of uropathogenic *Escherichia coli*. TLR3, 7, 8 and 9 sense microbial nucleic acids: dsRNA is sensed by TLR3 and ssRNA by TLR7 and 8, while unmethylated CpG DNA is sensed by TLR9. Moreover, mouse TLR13 recognizes bacterial 23S ribosomal RNA.

TLR4 and, to some extent, TLR2 require coreceptor molecules in order to recognize microbial ligands. A noncovalently associated protein, MD-2, confers LPS responsiveness to TLR4. MD-2 binds the lipid A region of LPS in a deep hydrophobic pocket and interacts with the TLR4 ectodomain, which suggests that the MD-2–LPS complex is the essential ligand for TLR4. A second coreceptor for TLR4, CD14, transfers LPS monomers to MD-2 and increases the responsiveness of cells to LPS at low concentrations. The F protein of RSV also requires MD-2 for signaling through TLR4, an event that involves direct protein–protein interaction between MD-2 and the domain of the F protein that encompasses its hydrophobic fusion peptide. In addition, it has been shown that CD14 also acts as a coreceptor in order to activate TLR2 by mycobacterial lipoarabinomannan.

TLR signaling & downstream gene expression:

Pathogen-encoded ligand binding to TLR causes conformational changes and TLR dimerization that lead to the recruitment of cytosolic TIR domain-containing adapter proteins to the intracellular TIR domain of the TLR. The main adapter proteins include MyD88, TIRAP (also known as MAL), TRIF (also known as TICAM1) and TRAM (also known as TICAM2). The MyD88-dependent pathway is activated by all TLRs except TLR3, which only engages TRIF. TLR4 is the only TLR that activates both MyD88- and TRIF-dependent signaling pathways. CD14-dependent TLR4 internalization into endosomes from the plasma membrane facilitates induction of the TRIF signaling pathway. TIRAP was originally thought to act as a bridge to recruit MyD88 to TLR2 and TLR4, while TRAM recruits TRIF to TLR4. However, recent work by Kagan and colleagues suggest that TIRAP is more promiscuous. A fifth member of the TIR adapter group, SARM, interacts with TRIF and negatively regulates TLR3 and TLR4 signaling. A proposed sixth adapter is BCAP, which has a TIR-like domain and modulates B-cell activation by TLRs .

Engagement of TLRs by ligands causes a conformational change and the recruitment of adapters through TIR–TIR interactions, leading to the activation of a cascade of signal transduction molecules, including IRAKs, TRAF6 and TAK1, among others, leading to phosphorylation of the inhibitor of NF-κB kinase and the release of NF-κB transcription factors into the nucleus, which induces the expression of proinflammatory genes, such as *TNFA* and *IL6*. The MyD88-dependent pathway also results in the activation of MAPKs. By contrast, the TRIF-mediated signaling pathway involves the delayed activation of NF-κB and robust activation of IRF3, which is an important transcription factor for the induction of type I IFNs (primarily IFN-β in macrophages) and IFN-inducible genes. Endosomal TLRs, such as TLR7–9, engage the MyD88-dependent pathway and activate NF-κB and IRF7, which leads to the production of high levels of type I I. Taken together, activation of MAPKs and NF-κB is triggered by all TLRs from the plasma membrane and endosomes, whereas TLR-induced IRF3 (TLR3 and TLR4) and IRF7 (TLR7–9 and TLR13) activation is initiated only from the endosome. Activation of TLR signaling culminates in the expression of many secreted cytokines, such as IFNs,

TNF-α, IL-1, IL-6, IL-10, IL-12 and chemokines, as well as causing cell differentiation, proliferation or apoptosis. TLR–ligand interactions are complex and their outcomes depend on many factors, such as the differential expression of TLRs among different cell types, cell type-specific signaling pathways and the usage of varied adapters by different TLRs.

Since the discovery of TLR3 as the first receptor to recognize dsRNA, significant progress has been made in understanding TLR-mediated immune responses following different viral infections. The knowledge of virus-induced TLR signaling pathways has led to the development of novel therapeutics targeting TLRs as antiviral and anti-inflammatory therapies. This is a very dynamic field and has been growing rapidly in recent years. Approved use of the TLR7 agonist imiquimod for therapy against HPV-induced genital warts and the TLR4 agonist MPL as an adjuvant for vaccines against HPV and HBV are some of the successful examples of translational efforts. In addition, TLR3 and TLR7–9 agonists are showing very promising results for the treatment of viral infections. However, many challenges exist for the development of new TLR-based antiviral targets. Importantly, the interactions between virus and TLR signaling components are complex and the outcomes depend on the TLR, virus or the host species. In the case of vaccinia virus infection, TLR3-dependent responses were harmful, while TLR4-mediated immune responses proved to be protective in mice. Moreover, a single ligand can be sensed by different TLRs depending on the localization of the antigen, and this may generate overlapping, redundant responses. In the case of MCMV infection, TLR7 and TLR9 impart redundant functions for IFN, IL-12 p40 and TNF-α production by pDCs *in vivo* . In this case, redundancy of ligand sensing by TLRs should be taken into account in order to develop a single TLR-targeted treatment strategy. Some of the infections are resolved by the cooperation of multiple TLRs with each other or TLRs cooperating with other classes of PRRs. For example, TLR2 and TLR9 are both required for immunity against HSV-2, and TLR7 and TLR9 overlap to generate responses against MCMV. Both $TLR4^{-/-}$ mice and $PAR2^{-/-}$ mice are highly refractory to influenza infection. Furthermore, the coordinated recognition of rhinovirus, initially via TLR3 and later by RIG-I and MDA5, is required in order to induce antiviral responses within the bronchial epithelium. HSV infection is sensed by both TLR9 and RLRs, which synergize to induce type I IFN production. These observations indicate that complex cross-talk between different TLRs and between TLRs and other families of PRRs exist in order to resolve or mediate certain viral infections. It is of utmost importance for us to understand, characterize and take into account all of the possible interactions between TLRs or other families of PRRs in order to design and apply novel targets against TLR for efficacious treatment application.

Animal models provide primary valuable information about the safety, efficacy and molecular mechanisms of the selected drug targets. However, some of the TLR targets show important species-specific variations in the effects of certain drugs. RSV-infected BALB/c mice treated with poly-ICLC showed significantly reduced inflammation and

clinical scores for the disease. However, in contrast to the murine model, poly-ICLC treatment resulted in increased pathology during RSV infection in cotton rats [90]. In addition, due to important differences in the TLR signaling pathways in animal models and humans, many drugs tested with encouraging results in animal models failed to show much effect in human studies. In the natural condition, TLR-based immune responses to pathogen exposure are diversified on the basis of differences in cellular distributions at various anatomical sites and differential patterns of TLR expression among subsets of DCs and other antigen-presenting cells. Thus, TLR agonists can produce different responses on basis of the route of administration (e.g., contrary to intravenous administration), as subcutaneous CPG 7909 induced a Th1-like innate immune response. The differences in the response to TLR ligands administered by different routes pose challenges and opportunities for the development of TLR-based drugs and vaccines.

TLRs are fundamental sensors of the innate immune system. Thus, the activation or inhibition of TLR pathways by therapeutic TLR agonists or antagonists may cause potent harmful immune activation or unwanted immunosuppression. Moreover, it is difficult to predict efficacy and off-target effects in a large human population. To this end, Phase I safety trials of therapeutic TLR targets must be assessed for both their short- and long-term effects. Alternative dosing regimens or differential routes of administration may alter both the efficacy and safety of the drug in question. In addition, targeting therapeutic drugs to the relevant tissues or organ may be beneficial for the limitation of off-target effects. Furthermore, recent mouse and cotton rat data with eritoran, a TLR4 antagonist developed for the management of sepsis, showed promising results with regards to intervening in the inflammation associated with influenza infection, which opens the door for the possible treatment of other infectious agents where TLR4 senses DAMPs and initiates the cytokine storm. In summary, the development of TLR-targeted therapies in the form of agonists or antagonists offers exciting and promising new possibilities for the prevention of virus-induced infectious diseases or management of virus-induced harmful inflammatory responses.

Probable Question:

1. How GPCR are considered as target for therapeutic drugs.
2. Describe structure and function of JAK-STAT signaling system.
3. How JAK-STAT pathways are targeted for therapeutic purposes.
4. Discuss structure of TLR .
5. How TLR are used in viral infection.

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics .Verma and Agarwal.

Disclaimer

The study materials of this book have been collected from books, various e-books, journals and other e-sources.