Semester -III

17BTP304 GENOMICS AND PROTEOMICS 4H – 4C

Total hours/week: **L:4 T:0 P:0** **Marks:** Internal:**40** External:**60** Total: **100**

Scope: Gaining a thorough knowledge on genome and proteome would help the students to explore this technology in array designing and analysis.

Objective: To teach the students a through knowledge on genome and proteome identification, analysis and applications. To educate students on stand alone and online software for genetic studies.

UNIT - I

Genome Sequencing: Gene and pseudogenes, Gene structure, Genomes, Sequencing Genomes- methodology, chain termination method, chemical degradation method, automated DNA sequencing, shotgun sequencing and assembly of contiguous DNA sequence. cDNA and Genomic library construction.

UNIT - II

Genomic Mapping: Different types of genome maps and their practical uses, Genetic and Physical mapping techniques. Map resources. Practical uses of genome maps, Association mapping, Haplotypes. Genetic Markers - Mini and Micro satellite, STS and EST, SNPs,.

UNIT - III

Gene Expressions and Microarrays: Expression systems - Bacteria, Yeast and Viral. Concepts of microarrays, spotter analysis, Normalization –total intensity, using regression techniques, ratio statistics. Clustering Gene expression profiles-hierarchical, single-linkage, complete linkage, and average linkage. Tools for microarray analysis-MADAM, spot finder, SAGE Applications of Microarrays- Bioinformatics challenges in micro array design and analysis.

UNIT - IV

Experimental Proteomics: Proteome analysis- 2D gel electrophoresis: general strategy, immobilized pH gradients, sample preparation, isoelectric focussing, staining, transfer of proteins from gels, image acquisition and analysis of gels. 2DE databases.

UNIT - V

Analytical Proteomics: RP-HPLC, Mass Spectrometry – ESI MS and MALDI techniques and applications. Characterization of protein complexes – Protein - DNA, Protein-protein interactions, yeast two-hybrid system and protein micro-arrays – biomarkers.

References

Brown, T.A. (2006). *Genomes*. Singapore: John Wiley & Sons.

Cantor, C.R., & Smith, C. L. (1999). *Genomics: The Science and Technology behind the Human Genome Project*. Singapore: John Wiley and Sons.

Primrose, S.B., & Twyman, R.M. (2003). *Principles of Genome Analysis*. Oxford: Blackwell Publishing,.

Reiner, W., & Naven, T. (2002). *Proteomics in Practice*. Weinheim: Wiley – VCH.

Gibson, W., & Muse, V. (2009). *A Primer of Genome Science*. 3rd edition. New York: Sinauer Associates Inc. Publishers.

Stekal, D. (2003). *Microarray Bioinformatics*. Cambridge: Cambridge University Press.

Liebler, L.H. (2007). *Introduction to Proteomics, Tools for the New Biology*. 2nd edition. New Jersey: Humana Press.

Richard, P.S. (2004). *Proteins and Proteomics. A Laboratory Manual*. New York: Cold Spring Harbor Laboratory Press.

Pennington, S., & Dunn, M.J. (2001). *Proteomics: From Sequence to Function*. Oxford: Bios Scientific Pub.Ltd.

Bourne, P.E., & Weissig, H. (2003). *Structural Bioinformatics*. Singapore: John Wiley & Sons.

UNIT -I

Genome Sequencing: Gene and pseudogenes, Gene structure, Genomes, Sequencing Genomes- methodology, chain termination method, chemical degradation method, automated DNA sequencing, shotgun sequencing and assembly of contiguous DNA sequence. cDNA and Genomic library construction.

UNIT I

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Introduction – Genes and pseudogenes	2	11	R1: 12,13
Genome sequencing methods, Chain termination methods, Chemical degradation methods, Automated DNA Sequencing and Shotgun sequencing	2		T1: 20-36
Assembly of contiguous DNA sequence	2		T1: 20-36
Library construction	1		R1: 362-369
Genomics and cDNA library	2		R1: 362-369
Unit I Revision	1		
Unit I Class test	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.
R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.
W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs
W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/
W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html

UNIT -II

Genomic Mapping: Different types of genome maps and their practical uses, Genetic and Physical mapping techniques. Map resources. Practical uses of genome maps, Association mapping, Haplotypes. Genetic Markers - Mini and Micro satellite, STS and EST, SNPs,.

UNIT II

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Genome Mapping	2	11	T1: 209-216
Map resources and practical uses	2		T1: 209-216
Genetic markers	1		T1: 98-99
Mini and Macro satellites, VNTR, STS	2		T1: 113-129
EST & SNP	2		T1:113-129
Unit II Revision	1		
Unit I1 Class test	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.
R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.
W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs
W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/
W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html

UNIT -III

Gene Expressions and Microarrays: Expression systems - Bacteria, Yeast and Viral. Concepts of microarrays, spotter analysis, Normalization –total intensity, using regression techniques, ratio statistics. Clustering Gene expression profiles-hierarchical, single-linkage, complete linkage, and average linkage. Tools for microarray analysis-MADAM, spot finder, SAGE Applications of Microarrays- Bioinformatics challenges in micro array design and analysis.

UNIT III

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Gene expression	1	11	T2: 107-131
Microarray	1		T1: 22-23
Concept of gradient system, Sample preparation, processing and staining	2		T1: 614-615 W1
	1		W2
Tools for analysis – MADAM, Spot finder.	1		W3 & W4
SAGE, Applications	1		R2: 143-147
Bioinformatics challenges in microarray design and techniques	2		R2: 147-148
Unit III Revision	1		
Unit III Class test	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.
R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.
W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs
W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/
W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html

Unit IV

Experimental Proteomics: Proteome analysis- 2D gel electrophoresis: general strategy, immobilized pH gradients, sample preparation, isoelectric focussing, staining, transfer of proteins from gels, image acquisition and analysis of gels. 2DE databases.

Unit IV – Lecture Plan

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Experimental proteomics	2	**11**	T3: 1-34
2D Gel electrophoresis	3		R2: 449-452
Concept and Gradient system, Sample preparation, Processing and staining	1		R2: 449-452
Image acquisition analysis	1		R2: 449-452
Database for 2D Gel electrophoresis	2		W5
Unit IV Revision	1		
Unit IV Revision	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.

R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM. W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs

W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/ W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html

W5: www.world-2dpage.expasy.org/swiss-2dpage./does/ch2d-details.html
W6: www.ch.embert.org/coursEmnet/PropoA/slides/mass.spectrometry
W7: https://schuck/proteininteraction.chapter02.html

Unit V

Analytical Proteomics: RP-HPLC, Mass Spectrometry – ESI MS and MALDI techniques and applications. Characterization of protein complexes – Protein - DNA, Protein-protein interactions, yeast two-hybrid system and protein micro-arrays – biomarkers.

Unit V – Lecture Plan

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Analytical proteomics RP-HPLC	1		W6
ESI-MS	2		R2: 454 - 465
MALDI-MS	2		T3: 1-74
Characterization of protein complexes	1		W7
Protein – protein interaction	1		W7
Protein –DNA interaction	1	**15**	W7
Yeast 2 Hybrid system	1		R2 481-487
Protein microarrays, Biomarkers	2		R1 615-616
Unit V Class test	1		
ESE QP discussion	1		
ESE QP discussion	1		
ESE QP discussion	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.

R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM. W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs

W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/ W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html

W5: www.world-2dpage.expasy.org/swiss-2dpage./does/ch2d-details.html
W6: www.ch.embert.org/coursEmnet/PropoA/slides/mass.spectrometry
W7: https://schuck/proteininteraction.chapter02.html

UNIT -I

Genome Sequencing: Gene and pseudogenes, Gene structure, Genomes, Sequencing Genomes- methodology, chain termination method, chemical degradation method, automated DNA sequencing, shotgun sequencing and assembly of contiguous DNA sequence. cDNA and Genomic library construction.

UNIT I

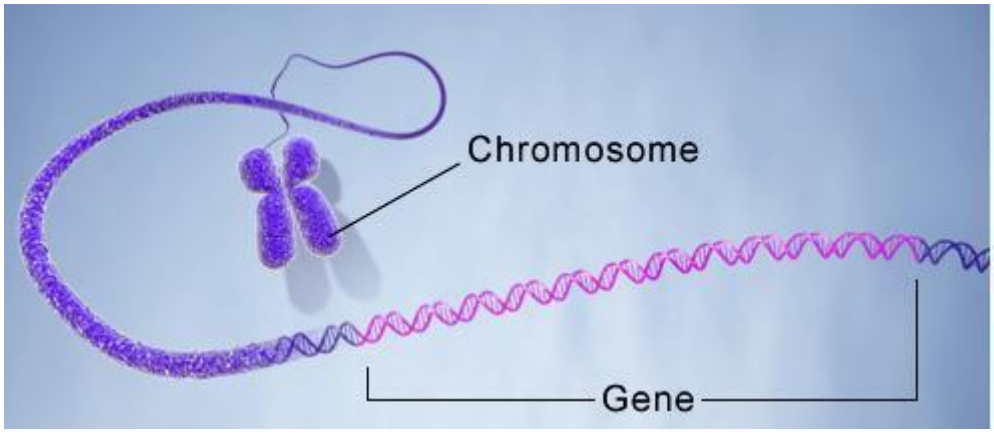
Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Introduction – Genes and pseudogenes	2	11	R1: 12,13
Genome sequencing methods, Chain termination methods, Chemical degradation methods, Automated DNA Sequencing and Shotgun sequencing	2		T1: 20-36
Assembly of contiguous DNA sequence	2		T1: 20-36
Library construction	1		R1: 362-369
Genomics and cDNA library	2		R1: 362-369
Unit I Revision	1		
Unit I Class test	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.
R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.
W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs
W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/
W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html

Genes

- A gene is the basic physical and functional unit of heredity.
- Genes, which are made up of DNA, act as instructions to make molecules called proteins.
- In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes.
- Every person has two copies of each gene, one inherited from each parent.
- Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people.
- Alleles are forms of the same gene with small differences in their sequence of DNA bases.
- These small differences contribute to each person‟s unique physical features.
- Genes are a section of DNA that are in charge of different functions like making proteins. Long strands of DNA with lots of genes make up chromosomes. DNA molecules are found in chromosomes. Chromosomes are located inside of the nucleus of cells.
- Each chromosome is one long single molecule of DNA. This DNA contains important genetic information.
- Chromosomes have a unique structure, which helps to keep the DNA tightly wrapped around the proteins called histones. If the DNA molecules were not bound by the histones, they would be too long to fit inside of the cell.
- Genes vary in complexity. In humans, they range in size from a few hundred DNA bases to more than 2 million bases.
- Different living things have different shapes and numbers of chromosomes. Humans have 23 pairs of chromosomes, or a total of 46. A donkey has 31 pairs of chromosomes, a hedgehog has 44, and a fruit fly has just 4.
- DNA contains the biological instructions that make each species unique.



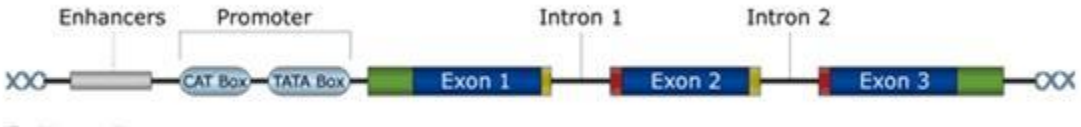
- DNA contains the biological instructions that make each species unique.
- DNA is passed from adult organisms to their offspring during reproduction. The building blocks of DNA are called nucleotides. Nucleotides have three parts: A phosphate group, a sugar group and one of four types of nitrogen bases.
- A gene consists of a long combination of four different nucleotide bases, or chemicals. There are many possible combinations.
- The four nucleotides are:
    1. A (adenine)
    2. C (cytosine)
    3. G (guanine)
    4. T (thymine)
- Different combinations of the letters ACGT give people different characteristics. For example, a person with the combination ATCGTT may have blue eyes, while somebody with the combination ATCGCT may have brown eyes.
- Genes decide almost everything about a living being. One or more genes can affect a specific trait. Genes may interact with an individual's environment too and change what the gene makes.
- Genes affect hundreds of internal and external factors, such as whether a person will get a particular color of eyes or what diseases they may develop.
- Some diseases, such as sickle-cell anemia and Huntington's disease, are inherited, and these are also affected by genes.

How we get our genes

- A gene is a basic unit of heredity in a living organism. Genes come from our parents. We may inherit our physical traits and the likelihood of getting certain diseases and conditions from a parent.
- Genes contain the data needed to build and maintain cells and pass genetic information to offspring.
- Each cell contains two sets of chromosomes: One set comes from the mother and the other comes from the father. The male sperm and the female egg carry a single set of 23 chromosomes each, including 22 autosomes plus an X or Y sex chromosome.
- A female inherits an X chromosome from each parent, but a male inherits an X chromosome from their mother and a Y chromosome from their father.

Genes consist of three types of nucleotide sequence:

- coding regions, called exons, which specify a sequence of amino acids
- non-coding regions, called introns, which do not specify amino acids
- regulatory sequences, which play a role in determining when and where the protein is made (and how much is made)



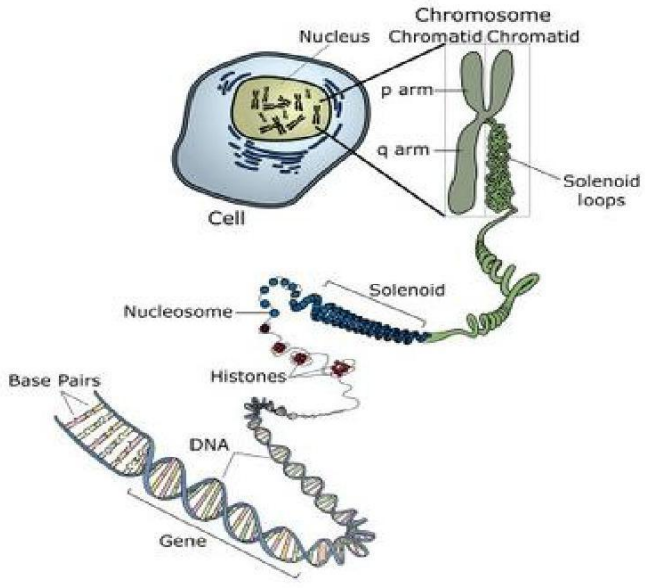
The structural components of a gene

A human being has 20,000 to 25,000 genes located on 46 chromosomes (23 pairs). These genes are known, collectively, as the human genome.

Chromosomes

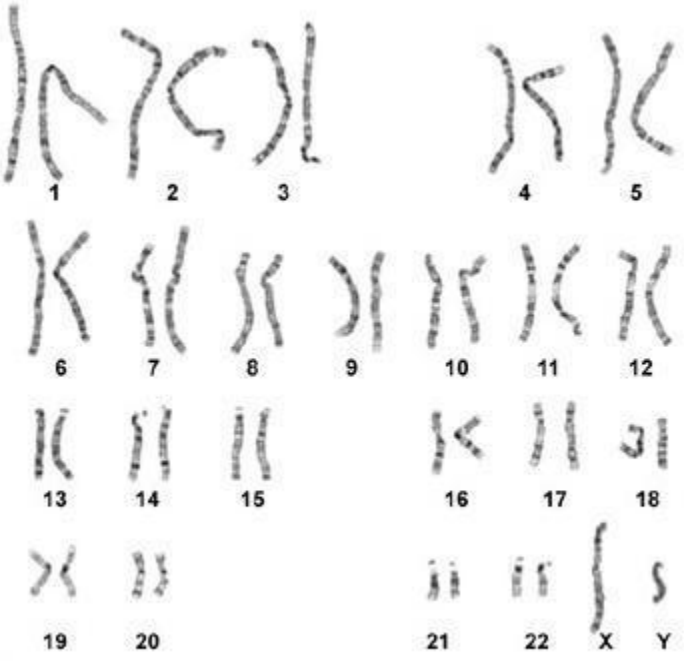
Eukaryotic chromosomes

- The label eukaryote is taken from the Greek for 'true nucleus', and eukaryotes (all organisms except viruses, Eubacteria and Archaea) are defined by the possession of a nucleus and other membrane-bound cell organelles.
- The nucleus of each cell in our bodies contains approximately 1.8 metres of DNA in total, although each strand is less than one millionth of a centimetre thick.
- This DNA is tightly packed into structures called chromosomes, which consist of long chains of DNA and associated proteins. In eukaryotes, DNA molecules are tightly wound around proteins - called histone proteins - which provide structural support and play a role in controlling the activities of the genes. A strand 150 to 200 nucleotides long is wrapped twice around a core of eight histone proteins to form a structure called a nucleosome.
- The histone octamer at the centre of the nucleosome is formed from two units each of histones H2A, H2B, H3, and H4. The chains of histones are coiled in turn to form a solenoid, which is stabilised by the histone H1. Further coiling of the solenoids forms the structure of the chromosome proper.
- Each chromosome has a p arm and a q arm. The p arm (from the French word 'petit', meaning small) is the short arm, and the q arm (the next letter in the alphabet) is the long arm. In their replicated form, each chromosome consists of two chromatids.



Chromosome unraveling to show the base pairings of the DNA

- The chromosomes - and the DNA they contain - are copied as part of the cell cycle, and passed to daughter cells through the processes of mitosis and meiosis.
- Human beings have 46 chromosomes, consisting of 22 pairs of autosomes and a pair of sex chromosomes: two X sex chromosomes for females (XX) and an X and Y sex chromosome for males (XY). One member of each pair of chromosomes comes from the mother (through the egg cell); one member of each pair comes from the father (through the sperm cell).
- A photograph of the chromosomes in a cell is known as a karyotype. The autosomes are numbered 1-22 in decreasing size order.



Karyotype of a human male

Prokaryotic chromosomes

- The prokaryotes (Greek for 'before nucleus' - including Eubacteria and Archaea) lack a discrete nucleus, and the chromosomes of prokaryotic cells are not enclosed by a separate membrane.
- Most bacteria contain a single, circular chromosome. (There are exceptions: some bacteria - for example, the genus Streptomyces - possess linear chromosomes, and Vibrio cholerae, the causative agent of cholera, has two circular chromosomes.)
- The chromosome - together with ribosomes and proteins associated with gene expression - is located in a region of the cell cytoplasm known as the nucleoid.
- The genomes of prokaryotes are compact compared with those of eukaryotes, as they lack introns, and the genes tend to be expressed in groups known as operons.
- The circular chromosome of the bacterium Escherichia coli consists of a DNA molecule approximately 4.6 million nucleotides long.
- In addition to the main chromosome, bacteria are also characterised by the presence of extra-chromosomal genetic elements called plasmids.
- These relatively small circular DNA molecules usually contain genes that are not essential to growth or reproduction.

The Human Genome Project (HGP)

- The Human Genome Project (HGP) is a major scientific research project. It is the largest single research activity ever carried out in modern science.
- It aims to determine the sequence of the chemical pairs that make up human DNA and to identify and map the 20,000 to 25,000 or so genes that make up the human genome.
- The HGP has opened the door to a wide range of genetic tests.
- The project was started in 1990 by a group of international researchers, the United States' National Institutes of Health (NIH) and the Department of Energy.
- The goal was to sequence 3 billion letters, or base pairs, in the human genome, that make up the complete set of DNA in the human body.

- By doing this, the scientists hoped to provide researchers with powerful tools, not only to understand the genetic factors in human disease, but also to open the door for new strategies for diagnosis, treatment, and prevention.
- The HGP was completed in 2003, and all the data generated is available for free access on the internet. Apart from humans, the HGP also looked at other organisms and animals, such as the fruit fly and E. coli.
- Over three billion nucleotide combinations, or combinations of ACGT, have been found in the human genome, or the collection of genetic features that can make up the human body.
- Mapping the human genome brings scientists closer to developing effective treatments for hundreds of diseases.
- The project has fueled the discovery of more than 1,800 disease genes. This has made it easier for researchers to find a gene that is suspected of causing an inherited disease in a matter of days. Before this research was carried out, it could have taken years to find the gene.

Eukaryotic gene structure

- The structure of eukaryotic genes includes features not found in prokaryotes.
- Most of these relate to post-transcriptional modification of pre-mRNAs to produce mature mRNA ready for translation into protein. Eukaryotic genes typically have more regulatory elements to control gene expression compared to prokaryotes.
- This is particularly true in multicellular eukaryotes, humans for example, where gene expression varies widely among different tissues.
- A key feature ofthe structure of eukaryotic genes is that their transcripts are typically subdivided into exon and intron regions. Exon regions are retained in the final mature mRNA molecule, while intron regions are spliced out (excised) during post-transcriptional processing.
- Indeed, the intron regions of a gene can be considerably longer than the exon regions. Once spliced together, the exons form a single continuous protein-coding regions, and the splice boundaries are not detectable.

- Eukaryotic post-transcriptional processing also adds a 5' cap to the start of the mRNA and a poly-adenosine tailto the end ofthe mRNA. These additions stabilise the mRNA and direct its transport from the nucleus to the cytoplasm, although neither of these features are directly encoded in the structure of a gene.

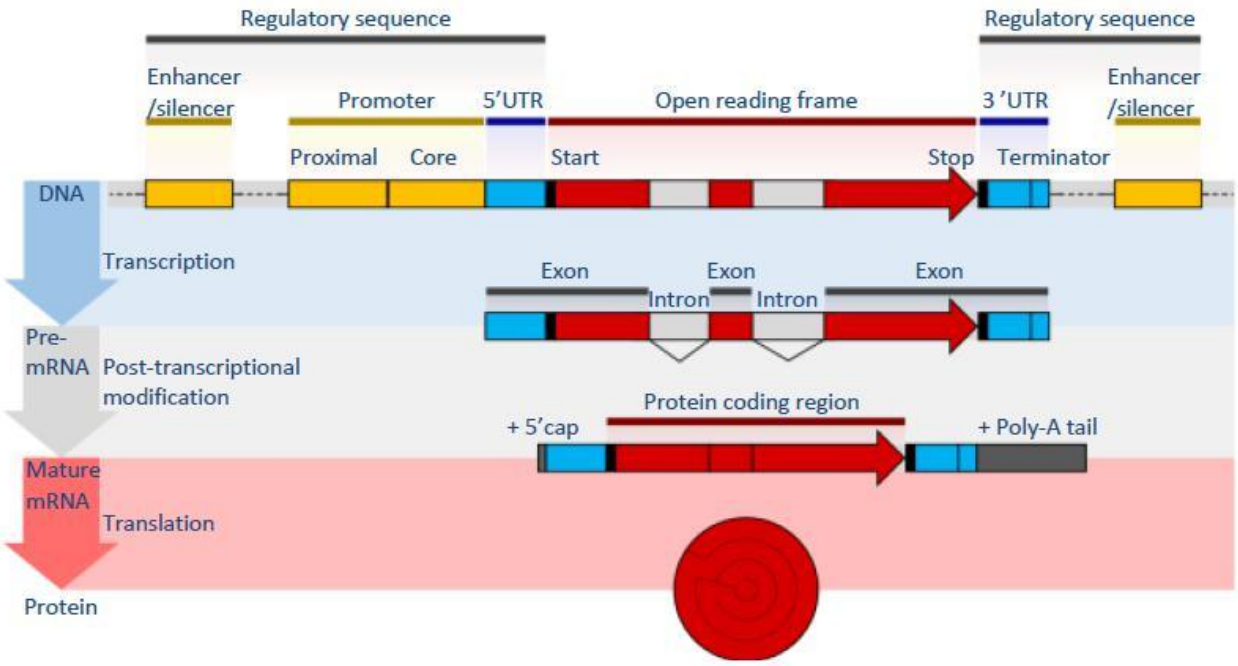


Figure The structure of a eukaryotic protein-coding gene. Regulatory sequence controls when and where expression occurs for the protein coding region (red). Promoter and enhancer regions (yellow) regulate the transcription of the gene into a pre-mRNA which is modified to remove introns (light grey) and add a 5' cap and poly-A tail (dark grey). The mRNA 5' and 3' untranslated regions (blue) regulate translation into the final protein product.

Prokaryotic gene structure

- The overall organisation of prokaryotic genes is markedly different from that of the eukaryotes.
- The most obvious difference is that prokaryotic ORFs are often grouped into a polycistronic operon under the control of a shared set of regulatory sequences.

- These ORFs are all transcribed onto the same mRNA and so are co-regulated and often serve related functions.
- Each ORF typically has its own ribosome binding site (RBS) so that ribosomes simultaneously
- translate ORFs on the same mRNA.
- Some operons also display translational coupling, where the translation rates of multiple ORFs within an operon are linked.
- This can occur when the ribosome remains attached at the end of an ORF and simply translocates along to the next without the need for a new RBS.
- Translational coupling is also observed when translation of an ORF affects the accessibility of the next RBS through changes in RNA secondary structure.
- Having multiple ORFs on a single mRNA is only possible in prokaryotes because their transcription and translation take place at the same time and in the same subcellular location.

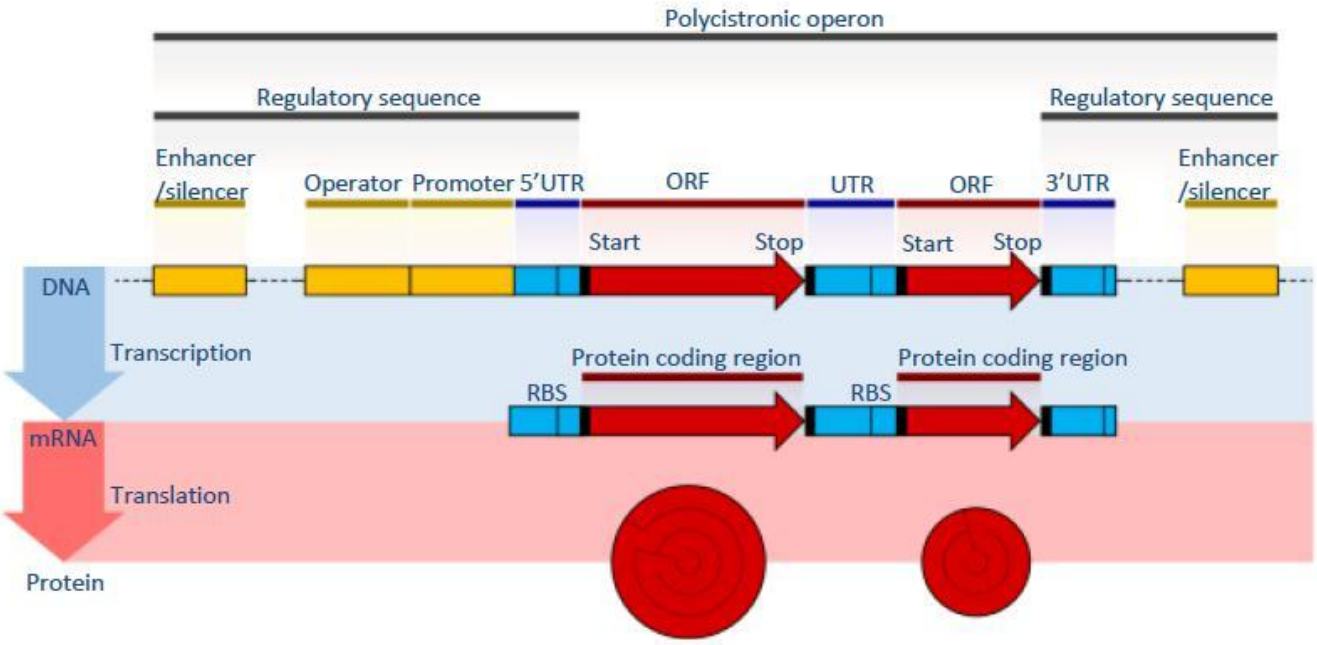


Figure The structure of a prokaryotic operon of protein-coding genes. Regulatory sequence controls when expression occurs for the multiple protein coding regions (red). Promoter, operator and enhancer regions (yellow) regulate the transcription of the gene into an mRNA. The mRNA untranslated regions (blue) regulate translation into the final protein products.

Genome

- A genome is an organism‟s complete set of DNA, including all of its genes. Each genome contains all of the information needed to build and maintain that organism.
- In humans, a copy of the entire genome more than 3 billion DNA base pairs is contained in all cells that have a nucleus.

Pseudogenes

- Pseudogenes are ubiquitous and abundant in genomes. Pseudogenes were once called “genomic fossils” and treated as “junk DNA” several years.
- Nevertheless, it has been recognized that some pseudogenes play essential roles in gene regulation of their parent genes, and many pseudogenes are transcribed into RNA.
- Pseudogene transcripts may also form small interfering RNA or decrease cellular miRNA concentration.
- Thus, pseudogenes regulate tumor suppressors and oncogenes. Their essential functions draw the attention of our research group in my current work on heat shock protein 90: a chaperone of oncogenes.
- The paper reviews our current knowledge on pseudogenes and evaluates preliminary results of the chaperone data.
- Current efforts to understand pseudogenes interactions help to understand the functions of a genome.

History of pseudogenes

- Sequencing human genome brought several debates about noncoding sequences. So what is the role of the noncoding parts since protein coding exons compromise only around two percent of the whole genome sequence?

- The noncoding regions are transposable elements, structural variants, segmental duplications, simple and tandem repeats, conserved noncoding elements, functional noncoding RNAs, regulatory elements, and pseudogenes.
- Annotation of these noncoding regions through functional genomics and sequence analysis helps our understanding of genomics.
- Noncoding regions of human genome in general were thought to be nonfunctional and "junk," or of no purpose DNA. Nowadays, scientists are conceding that junk DNA terminology is far from true since recent studies indicate that they have some regulatory roles.
- This work focuses on pseudogenes of junk DNA. Pseudogenes are gene copies that have coding-sequence deficiencies like frameshifts and premature stop codons but resemble functional genes.
- The first pseudogene was reported for 5S DNA of *Xenopus laevis*, coding for oocyte-type 5S RNA, in 1977, and several pseudogenes have been reported and described for a variety of species including plants, insects, and bacteria.
- Currently, approximately twenty thousand pseudogenes are estimated which is comparable to the number of protein-coding genes (around 27000) in human.
- Current knowledge of these genes remains poorly understood, and many sequences once believed defunct are in fact functional RNA genes and play roles in gene silencing either by forming siRNAs or by changing mRNA levels of functional protein-coding gene.
- Several studies focused on the pseudogene population and their regulatory roles as the function of more pseudogenes is being uncovered.
- It is interesting to compare and contrast genes from a variety of organisms to determine their adaptation for survival. Pseudogenes provide a record of all changes in the genome of a particular organism.

Types of pseudogenes

- Pseudogenes can be categorized in two forms: unprocessed and processed. Unprocessed pseudogenes can also be subcategorized as unitary and duplicated.

- Pseudogenes originate from decay of genes that originated from duplication through evolution. The decays include point mutations, insertions, deletions, misplaced stop codons, or frameshifts of a gene.
- The decay may occur during duplication, and these disablements may cause loss of a gene function. Loss of productivity, expression of RNA or protein coding ability, results in the production of unprocessed pseudogenes. A unique subfamily of unprocessed pseudogenes are described by Zhang et al. Formation of nonduplicated unprocessed pseudogenes is named "unitary" pseudogenes.
- In unitary type of pseudogenes, a single copy parent gene becomes nonfunctional. Unprocessed and duplicated pseudogenes keep their intron-exon structure. Processed pseudogenes are formed through retrotransposition. Retrotransposition occurs by reintegration of a cDNA, a reverse transcribed mRNA transcript, into the genome at a new location. The double-stranded sequences of processed pseudogenes are generated from single-strand RNA by RNA polymerase II rather than the RNA polymerase III. Therefore, processed pseudogenes lack introns, 5′ promoter sequence, and have flanking direct repeats and 3′ polyadenylation tag.
- The overall distribution of most pseudogenes is completely random, duplicated, and processed pseudogenes are found in the same or on different chromosome of their parent genes.
- Duplication of DNA segments explains the generation of gene families from a common ancestral gene. The dynamic nature of genome cause changes in its composition with time.
- Various pseudogenes have certain conserved mutations in different species. Conservation of pseudogenes was explored in human, chimpanzee, mouse, rat, dog, and cow.
- Pseudogenes from different species have point mutations and even specific types of mutations at certain gene locations. The shared mutations in different organisms are thought to depend on common descent or evolutionary ancestry.

- The locus of insertion of a pseudogene determines its evolution. Deleterious insertions will be selected, and the pseudogene will be lost; however, pseudogenes with other nondeleterious mutations persist and evolve over time.
- Processed pseudogenes evolve more rapidly than their functional paralogs and undergo genetic drift with random mutations, deletions, and insertions. Established pseudogenes can pass to next generation and may partially be duplicated to give a second pseudogene. Thus, pseudogenes provide a powerful tool for phylogenetic studies to investigate genome evolution.

DNA sequencing

Methods of sequencing:

Two different methods are now routinely used for determination of DNA sequences

1. Chemical degradation method
2. Chain termination method

Maxam and Gilbert's Chemical Degradation Method

- developed in the late 1970‟s, by Allan Maxam and Walter Gilbert
- first method to determine the sequence of a DNA molecule of up to 500 bp.
- based on chemical modification of DNA and subsequent cleavage at specific bases.

In this method, following steps are involved:

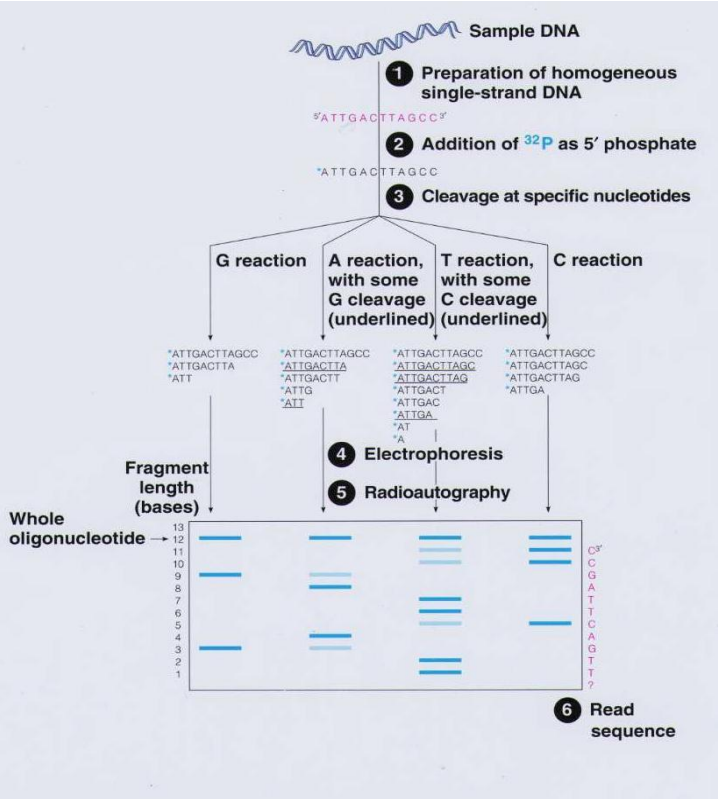
(i) Label the 3'ends of DNA with 32p.

(ii) Separate two strands, both labelled at 3'ends.

(iii) Divide the mixture in four samples, each treated with a different reagent having the property of destroying either only G, or only C, or 'A and G' or 'T and C. The concentration of reagent is so adjusted that 50% of target base is destroyed, so that fragments of different sizes having 32p are produced.

(iv) Electrophoreses each of the four samples in four different lanes of the gel.

(v) Autoradiograph the gel and determine the sequence from positions of bands in four lanes. vi) Read the sequence based on autoradiography.



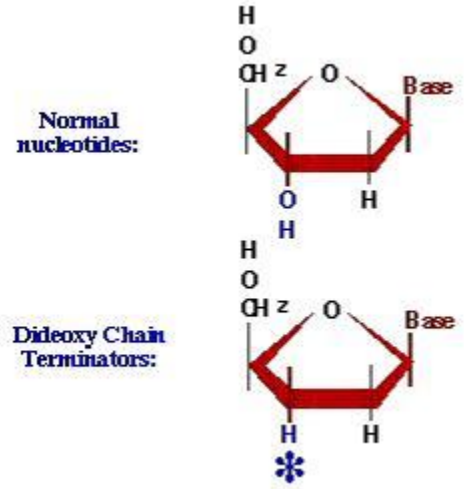
Disadvantages of this method:

This technology has the disadvantage of relying on toxic chemicals.

Chain termination method

- Developed by Fred Sanger
- Also called as Dideoxynucleotide Synthetic Method
- utilizing single stranded DNA as template for DNA synthesis, in which 2',3' dideoxynucleotides were incorporated leading to termination of DNA synthesis.
- These dideoxynucleotides are used as triphosphate and can be incorporated in a growing chain, but these dideoxynucleotides terminate synthesis, since they can not form a phosphodiester bond with next incoming deoxynucleotide triphosphate (dNTP).

Dideoxynucleotides:



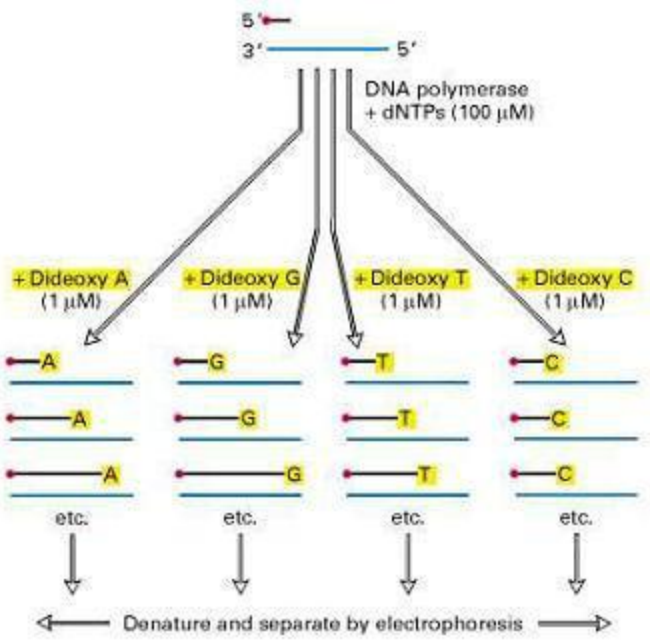
Following steps are involved in Sanger's dideoxy method for DNA sequencing:

(i) Four reaction tubes are set up, each containing single stranded DNA sample to be sequenced, all four dNTPs (radioactively labelled) and an enzyme for DNA synthesis (DNA polymerase I).

Each tube also contains a small amount (much smaller amount relative to four dNTPs) of one of the four ddNTP, so that four tubes have each a different ddNTP, bringing about termination at a specific base adenine (A), cytosine (C), guanine (G) and thymine (T).

ii) The fragments, generated by random incorporation of ddNTP leading to termination of reaction, are then separated by electrophoresis on a high resolution Polyacrylamide gel.

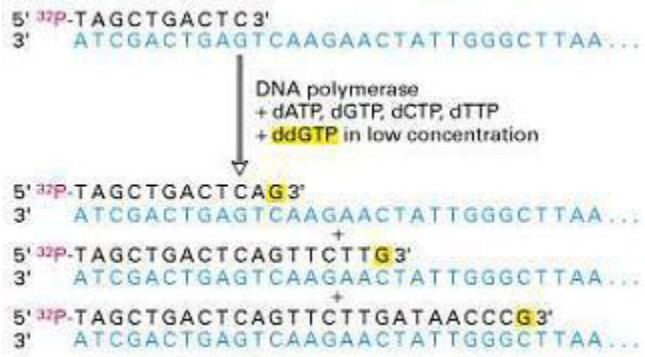
This is done for all the four reaction mixtures on adjoining lanes in the gel.



(iii) The gel is used for autoradiography so that the position of different bands in each lane can be visualized.

(iv) The bands on autoradiogram can be used for getting the DNA sequence.

Example



Now, variations of the above dideoxy method was developed as automatic sequencers.

In this automatic sequencer approach,

- a different fluorescent dye is tagged to the oligonucleotide primer in each of the four reaction tubes (blue for: A, red for C, etc.).
- The four reaction mixtures are pooled and electrophoreses together in a single Polyacrylamide of the tube.
- A high sensitivity fluorescence detector, placed near the bottom of the tube, measures the amount, of each fluorophore as a function of time.
- The sequence is determined from the temporal order of peaks corresponding to four different dyes.

The chain termination method of DNA sequencing can only be used for short strands (100 to 1000 base pairs), longer sequences must be subdivided into smaller fragments, and subsequently re-assembled to give the overall sequence.

Automated DNA sequencing

- Automated DNA sequencing utilizes a fluorescent dye to label the nucleotides instead of a radioactive isotope. The fluorescent dye is not an environmentally hazardous chemical and has no special handling or disposal requirements. Instead of using X-ray film to read the sequence, a laser is used to stimulate the fluorescent dye.
- The fluorescent emissions are collected on a charge coupled device that is able to determine he wavelength. The Perkin-Elmer Applied Biosystems (ABI) DNA sequencers are designed to discriminate all four fluorescent dye wavelengths simultaneously, which allows for complete DNA sequencing in one lane on the gel. Varying degrees of automation are also available. For full automation, all that is required is to load a sample tray with template DNA; the equipment performs the labeling and analysis. The other option is to perform the labeling reactions with fluorescent dyes, load the samples onto a gel, and place the gel into the DNA sequencer. The equipment performs the separation and analysis. The system automatically identifies the nucleotide sequence and saves the

information on the computer. Thus, only a review of the data is necessary to ensure no anomalies were misidentified by the computer.

- The greatest obstacle to researchers when converting from manual to automatic DNA sequencing is being required to learn the use of computer software necessary to interpret the results. Automated DNA sequencing equipment can eliminate the need for radioactive isotopes to label DNA, thereby reducing the volume of low-level radioactive waste generated on campus.
- As a general approximation, one template of manual DNA sequencing will produce 83 mL of liquid waste and 0.167 gallon of solid waste. As a result, every 45 templates processed by automated DNA sequencing reduces the amount of manual DNA sequencing.
- The time saved is due to not having to perform autoradiography or associated tasks required for working with radioactive materials such as radiation surveys, inventory/disposal documentation, etc. Finally, automated DNA sequencing provides more reliable research results than manual DNA sequencing, thus maintaining the integrity of the research.

Shortgun sequencing

- Shotgun sequencing involves randomly breaking up DNA sequences into lots of small pieces and then reassembling the sequence by looking for regions of overlap.
- Large, mammalian genomes are particularly difficult to clone, sequence and assemble because of their size and structural complexity. As a result clone-by-clone sequencing, although reliable and methodical, takes a very long time.
- With the emergence of cheaper sequencing and more sophisticated computer programs, researchers have therefore relied on whole genome shotgun sequencing to tackle larger, more complex genomes.
- Shotgun sequencing was originally used by Fred Sanger and his colleagues to sequence small genomes such as those of viruses and bacteria.

- Whole genome shotgun sequencing bypasses the time-consuming mapping and cloning steps that make clone-by-clone sequencing so slow.
- In whole genome shotgun sequencing the entire genome is broken up into small fragments of DNA for sequencing.
- These fragments are often of varying sizes, ranging from 2-20 kilobases (2,000-20,000 base pairs) to 200-300 kilobases (200,000-300,000 base pairs).
- These fragments are sequenced to determine the order of the DNA bases, A, C, G and T.
- The sequenced fragments are then assembled together by computer programs that find where fragments overlap.
- You can imagine shotgun sequencing as being a bit like shredding multiple copies of a book (which in this case is a genome), mixing up all the fragments and then reassembling the original text (genome) by finding fragments with text that overlap and piecing the book back together again.
- This method of genome sequencing was used by Craig Venter, founder of the private company Celera Genomics, to sequence the human genome.
- Venter wanted to sequence the human genome faster than the publicly funded effort and felt this was the best way. To assemble the sequence Venter used the clone-by-clone publically available data from the Human Genome Project.
- Now, as technologies are improving, whole genome shotgun sequencing is being used to improve the accuracy of existing genome sequences, such as the reference human genome.
- It is used to remove errors, fill in gaps or correct parts of the sequence that were originally assembled incorrectly when clone-by-clone sequencing was used.
- As a consequence the reference human genome is constantly being improved to ensure that the genome sequence is of the highest possible standard.

Advantages of shotgun sequencing

- By removing the mapping stages, whole genome shotgun sequencing is a much faster process than clone-by-clone sequencing.

- Whole genome shotgun sequencing uses a fraction of the DNA that clone-by-clone sequencing needs.
- Whole genome shotgun sequencing is particularly efficient if there is an existing reference sequence. It is much easier to assemble the genome sequence by aligning it to an existing reference genome.
- Shotgun sequencing is much faster and less expensive than methods requiring a genetic map.

Disadvantages of shotgun sequencing

- Vast amounts of computing power and sophisticated software are required to assemble shotgun sequences together. To sequence the genome from a mammal (billions of bases long), you need about 60 million individual DNA sequence reads.
- Errors in assembly are more likely to be made because a genetic map is not used. However these errors are generally easier to resolve than in other methods and minimised if a reference genome can be used.
- Whole genome shotgun sequencing can only really be carried out if a reference genome is already available, otherwise assembly is very difficult without an existing genome to match it to.
- Whole genome shotgun sequencing can also lead to errors which need to be resolved by other, more labour-intensive types of sequencing, such as clone-by-clone sequencing.
- Repetitive genomes and sequences can be more difficult to assemble.

Assembly of contiguous DNA sequence

- The next question to address is how the master sequence of a chromosome, possibly several tens of Mb in length, can be assembled from the multitude of short sequences generated by chain termination sequencing.
- The relatively short genomes of prokaryotes can be assembled by shotgun sequencing, but that this approach might lead to errors if applied to larger eukaryotic genomes.

- The whole-genome shotgun method, which uses a map to aid assembly of the master sequence, has been used with the fruit-fly and human genomes, but it is generally accepted that a greater degree of accuracy is achieved with the clone contig approach, in which the genome is broken down into segments, each with a known position on the genome map, before sequencing is carried out. We will start by examining how shotgun sequencing has been applied to prokaryotic genomes.
- Samples of each clone in row A of the first microtiter tray are mixed together and a single PCR carried out. This is repeated for every row of every tray – 80 PCRs in all.
- Samples of each clone in column 1 of the first microtiter tray are mixed together and a single PCR carried out. This is repeated for every column of every tray – 120 PCRs in all.
- Clones from well A1 of each of the ten microtiter trays are mixed together and a single PCR carried out. This is repeated for every well – 96 PCRs in all.
- These 296 PCRs provide enough information to identify which of the 960 clones give products and which do not. Ambiguities arise only if a substantial number of clones turn out to be positive.

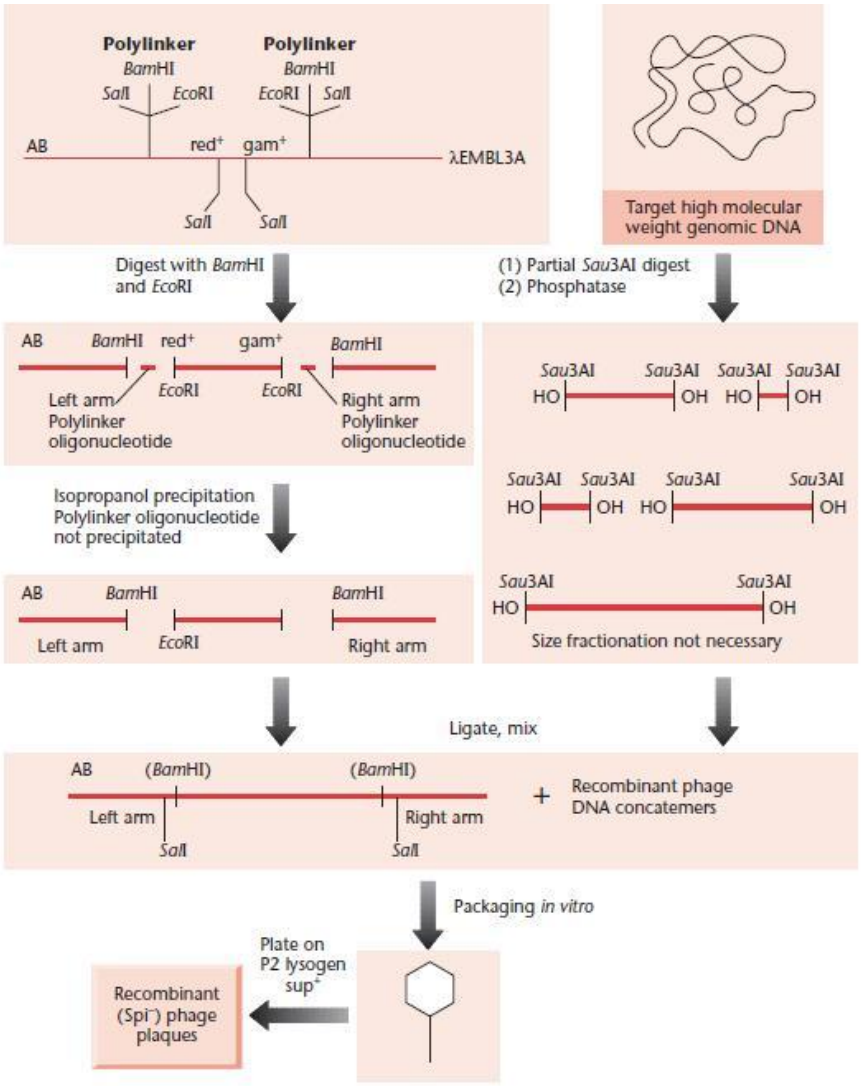
Genomic Library

- A genomic library is a collection of plasmid clones or phage lysates containing recombinant DNA molecules so that the sum total of DNA inserts in this collection, ideally, represents the entire genome of the concerned organism. However, inspite of all the care taken in the production of genomic libraries.

Certain DNA fragments should be expected to be under or over represented or even missing. There are several possible reasons for this, and at present they can not be taken care of.

3.2.Construction of A Genomic Library

- For preparation of a genomic library, the total genomic DNA of an organism is extracted.
- The DNA is broken into fragments of appropriate size either by mechanical shearing (this generates blunt ended fragments), sonication, or by using a suitable restriction endonuclease for partial digestion of the DNA.
- Complete digestion is avoided since it generates fragments that are too heterogeneous in size.



- For partial digestion, restriction enzymes having 4-base (tetrameric) recognition sequences are employed in preference to those having 6-base (hexameric) target sites. This is because a given 4-base recognition site is expected to' occur every 44 (= 256) base pairs in a DNA molecule, while a 6 base target site would occur only after every 46 (=

4096) base pairs. (It is assumed here that the arrangement of the 4 bases in DNA molecules is random).

- Therefore, the fragments produced in partial digests with enzymes having 4 base recognition sites are more likely to be of appropriate size for cloning than those generated by enzymes having 6 base recognition sites.
- Single or mixed digestions with the enzymes AluI, HaeIII or Sau3A have been used for constructing genomic libraries.
- The use of restriction enzymes has the advantage that the same set of fragments are obtained from a DNA each time a specific enzyme is used, and many of the enzymes; produce cohesive ends.
- The partial digests of genomic DNA are subjected to agarose gel electrophoresis or sucrose gradient centrifugation for separation from the mixture of fragment of appropriate size.
- These fragments are then inserted into a suitable vector for cloning.
- This constitutes the shotgun approach to gene cloning. In principle, any vector can be used, but A vectors and cosmids have been the most commonly used since DNA inserts of upto 23-25 kb (kilobase pairs) can be cloned in these vectors. The vectors containing the inserts are cloned in a suitable bacterial host.

Genomic DNA libraries *Producing representative genomiclibraries in* ë *cloning vectors*

let us suppose that to clone a single-copy gene from the human genome.

How might this be achieved?

- simply digest total human DNA with a restriction endonuclease, such as *Eco*RI, insert the fragments into a suitable phage-ë vector and then attempt to isolate the desired

clone. **How many recombinants have to screen in order to isolate the right one?**

- Assuming *Eco*RI gives, on average, fragments of about 4 kb,

- and given that the size of the human haploid genome is 2.8 × 106 kb, we can see that over 7×10^5 independent recombinants must be prepared and screened in order to have a reasonable chance of including the desired sequence.

There are two problems with the above approach.

- First, the gene may be cut internally one or more times by *Eco*RI so that it is not obtained as a single fragment.
- These problems can be overcome by cloning *random* DNA fragments of a large size (for ë replacement vectors, approximately 20 kb).
- Since the DNA is randomly fragmented, there will be no systematic exclusion of any sequence.
- Furthermore, clones will overlap one another, allowing the sequence of very large genes to be assembled and giving an opportunity to „walk" from one clone to an adjacent one.
- Because of the larger size of each cloned DNA fragment, fewer clones are required for a complete or nearly complete library.

How many clones are required?

Let *n* be the size of the genome relative to a single cloned fragment. Thus, for the human genome (2.8×10^6 kb) and an average cloned fragment size of 20 kb, $n = 1.4 \times 10^5$.

The number of independent recombinants required in the library must be greater than *n*, because sampling variation will lead to the inclusion of some sequences several times and the exclusion of other sequences in a library of just *n* recombinants.

Clarke and Carbon (1976) have derived a formula that relates the probability (P) of including any DNA sequence in a random library of N independent recombinants:

$$N = \frac{\ln(1-P)}{\ln\left(1-\frac{1}{n}\right)}$$

Therefore, to achieve a 95% probability ($P = 0.95$) of including any particular sequence in a random human genomic DNA library of 20 kb fragment size:

$$N = \frac{\ln(1-0.95)}{\ln\left(1-\frac{1}{1.4\times10^5}\right)} = 4.2\times10^5$$

Notice that a considerably higher number of recombinants is required to achieve a 99% probability, for here $N = 6.5 \times 10^5$.

How can appropriately sized random fragments be produced?

Various methods are available.

- Random breakage by mechanical shearing is appropriate because the average fragment size can be controlled, but insertion of the resulting fragments into vectors requires additional steps.
- The more commonly used procedure involves restriction endonucleases. In the strategy devised by Maniatis *et al*. (1978)
- the target DNA is digested with a mixture of *two* restriction enzymes.

- These enzymes have tetranucleotide recognition sites, which therefore occur frequently in the target DNA and in a complete double-digest would produce fragments averaging less than 1 kb.
- However, only a partial restriction digest is carried out, and therefore the majority of the fragments are large (in the range 10–30 kb).
- Given that the chances of cutting at each of the available restriction sites are more or less equivalent, such a reaction effectively produces a random set of overlapping fragments.
- These can be size-fractionated, e.g. by gel electrophoresis, so as to give a random population of fragments.

Cloning of cDNA

cDNA Library

- A cDNA library is a population of bacterial transformants or phage lysates in which each mRNA isolated from an organism or tissue is represented as its cDNA insertion in a plasmid or a phage vector.
- The frequency of a specific cDNA in such a library would ordinarily depend on the frequency of the concerned mRNA in the tissue/organism in question.

Preparation of cDNA

- cDNA is the copy or complementary DNA produced by using mRNA (usually) as a template. In fact, any RNA molecule can be used to produce cDNA.
- DNA copy of an RNA molecule is produced by the enzyme reverse transcriptase (RNA dependent DNA polymerase; discovered by **Temin and Baltimore** in 1970) generally obtained from avian mycloblastosis virus (AMV).
- This enzyme performs similar reactions as DNA polymerase, and has an absolute requirement for a primer with a free 3' -OH.
- When eukaryotic mRNA is used as a template, a poly T oligonucleotide (more specifically, oligodeoxynucleotide) is conveniently used as the primer since these mRNAs have a poly-A tail at their 3' ends.

- But special tricks are required to utilize primers for other RNAs, e.g., prokaryotic mRNA, rRNA, RNA virus genomes, etc.
- For example, a poly A tail may be added to 3' end of the RNA to make it analogous to eukaryotic mRNA (oligo-T is now used as primer); this reaction is catalyzed by the enzyme poly A polymerase.
- The appropriate oligonucleotide primer (oligo- T for eukaryotic mRNA) is annealed with the mRNA; this primer will base-pair to the 3'-end of mRNA.
- Reverse transcriptase extends the 3'-end of the primer using mRNA molecule as a template. This produces a RNA.DNA hybrid molecule, the DNA strand being the cDNA.
- The RNA strand is digested either by RNase H or alkaline hydrolysis; this frees the single-stranded cDNA.
- Curiously, the 3'-end of this cDNA serves as its own primer and provides the free 3'-OH required for the synthesis of its complementary strand; therefore, a primer is not required for this step.
- The complementary strand of cDNA single strand is synthesized by either the reverse transcriptase itself or by E. coli DNA polymerase; this generates a hairpin loop in the cDNA. The hairpin loop is cleaved by a single strand specific nuclease to yield a regular DNA duplex.

Problems in cDNA Preparation

- Usually the double strand cDNA preparations are always a mixture of different kinds of molecules due to problems in copying of the RNA and also because even highly purified mRNAs are never absolutely pure.
- Physical and chemical methods are incapable of resolving these mixtures.
- Therefore, the cDNA mixture itself is used for cloning and the desired cDNA is identified and isolated in pure form from the appropriate bacterial clone.

Isolation of mRNA

For isolation of mRNA, total RNA is first extracted from a suitable organism/tissue. The amount of desired mRNA in this sample is then increased by using one of the several procedures, some

of which are listed below. Chromatography on poly U sepharose or oligo T cellulose enriches the reparation with mRNAs of all kinds.

When the protein produced by a gene is known, it is purified and used to produce antibodies specific to it. These antibodies are used to precipitate the polysomes (mRNAs associated with ribosomes. and newly synthesized polypeptide chains), and the mRNA is isolated from them.

Some genes are expressed only in specific tissues, e.g., seed storage proteins in developing seeds, chicken ovalbumin gene in oviduct, etc. Therefore, mRNA preparations from such tissues are exceptionally rich in the concerned mRNA or may even contain only this mRNA. Use of cDNA is absolutely essential when the expression of an eukaryotic gene is required in a prokaryote, e.g., a bacterium. This is because eukaryotic genes have introns, which must be removed from their transcripts to yield mature mRNA, and bacteria do not possess the enzymes necessary for removal of introns.

Transformation of rDNAs to host cell

Transfer of rDNA into a suitable host cell is an important step in recombinant DNA technology looking for maintenance and expression of a foreign DNA in that cell.

The rDNA is safe in the cell and replicates independently of the chromosomal DNA of the cell. While doing so, the desired foreign gene present in the rDNA expresses its characters in the host cell. The host cell may be a bacterium or plant cell or animal cell. The cell containing an rDNA is known as transformed cell or transformant or recombinant.

Scientists have developed several methods to transfer genes into different types of host cells. The right method to be selected however depends on the type of gene cloning vector and nature of host cell, i.e. whether it is a bacterial cell or plant cell or animal cell.

The rDNAs can be introduced into host cells by *Transformation, Biolistics, Transduction, Electroporation, Transfection, Liposome fusion, Microinjection.*

Possible Questions

1. Define Gene and Genome. Explain pseudogenes on evolutionary basis.
2. Explain about cDNA library construction and screening.
3. Explain about the genomic library construction and screening.
4. Explain about chain termination method of sequencing.
5. Write about steps involved in Genomic DNA library construction and screening.
6. Explain in detail about DNA sequencing methods.
7. Give note on shot-gun sequencing and assembly of contiguous DNA sequences.
8. Diagrammatically explain prokaryotic genome.
9. Compare chain termination method with chemical degradation method of sequencing.

UNIT -II

Genomic Mapping: Different types of genome maps and their practical uses, Genetic and Physical mapping techniques. Map resources. Practical uses of genome maps, Association mapping, Haplotypes. Genetic Markers - Mini and Micro satellite, STS and EST, SNPs,.

UNIT II

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Genome Mapping	2	11	T1: 209-216
Map resources and practical uses	2		T1: 209-216
Genetic markers	1		T1: 98-99
Mini and Macro satellites, VNTR, STS	2		T1: 113-129
EST & SNP	2		T1:113-129
Unit II Revision	1		
Unit I1 Class test	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.
R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.
W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs
W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/
W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html

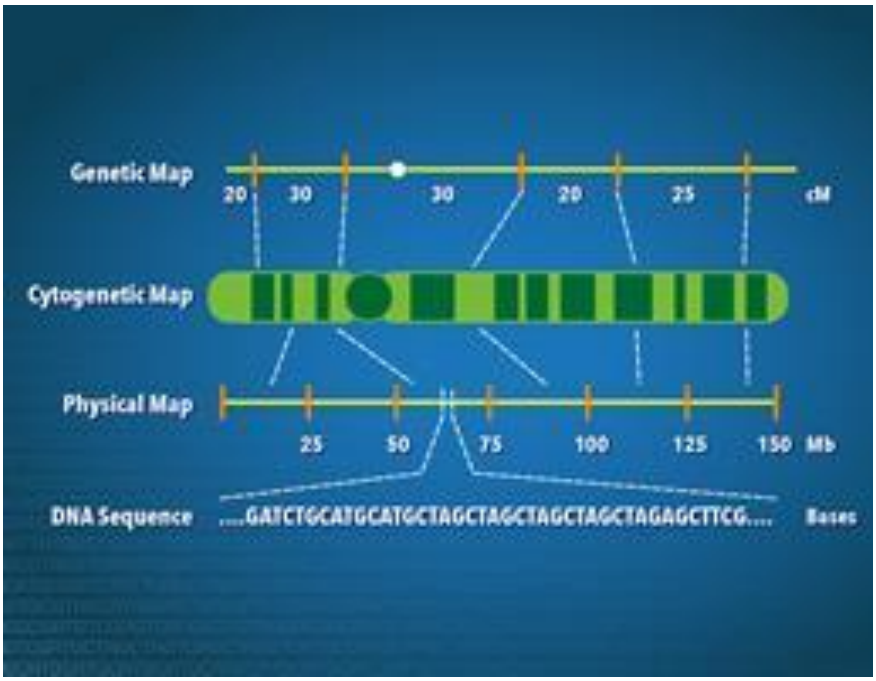
Genome Mapping

- Among the main goals of the Human Genome Project (HGP) was to develop new, better and cheaper tools to identify new genes and to understand their function.
- One of these tools is genetic mapping. Genetic mapping - also called linkage mapping - can offer firm evidence that a disease transmitted from parent to child is linked to one or more genes. Mapping also provides clues about which chromosome contains the gene and precisely where the gene lies on that chromosome.
- Genetic maps have been used successfully to find the gene responsible for relatively rare, single-gene inherited disorders such as cystic fibrosis and Duchenne muscular dystrophy.
- Genetic maps are also useful in guiding scientists to the many genes that are believed to play a role in the development of more common disorders such as asthma, heart disease, diabetes, cancer, and psychiatric conditions.
- In 1911, by Thomas Hunt Morgan, gene for eye-color was located on the X chromosome of fruit fly.
- Shortly after that, E.B. Wilson attributed the sex-linkedν genes responsible for color-blindness and hemophilia in human beings to be located on the X-chromosome, similar to the many X-linked factors being described by the Morgan group in flies.
- It wasn't until 1968 that an autosomal assignment ofν linkage was made by Donahue--- "Duffy" was assigned to Chromosome #1.
- "Gene mapping" refers to the mapping of genes to specific locations on chromosomes. It is a critical step in the understanding of genetic diseases.
- There are two types of gene mapping:
    - ✓ Genetic Mapping - using linkage analysis to determine the relative position between two genes on a chromosome.
    - ✓ Physical Mapping - using all available techniques or information to determine the absolute position of a gene on a chromosome.

Genetic mapping

- Uses genetic techniques to construct maps showing the positions of genes and other sequence features on a genome.
- Requires informative markers – polymorphic and a population with known relationships.
- Best if measured between "close" markers.ν
- Unit of distance in genetic maps = centiMorgans, cMν

- 1 cM = 1% chance of recombination between markersv
- Genetic techniques include crossbreeding experiments or, in the case of humans, the examination of family histories (pedigrees).



Markers for genetic mapping

- The first genetic maps, constructed in the organisms such as the fruit fly, used genes as markers.
- The only genes that could be studied were those specifying phenotypes that were distinguishable by visual examination.
- Eg. Eye color, height.
- Some organisms have very few visual characteristics so gene mapping with these organisms has to rely on biochemical phenotypes/

Biochemical markers in Human

- In human the biochemical phenotypes that can be scored by blood typing.
- These include the standard blood groups such as the ABO series and also 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 if all the family members have the same allele for the gene being studied then no useful information can be obtained.

Drawbacks of using gene as marker

- 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**.**

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).

Restriction Fragment Length Polymorphism (RFLP)

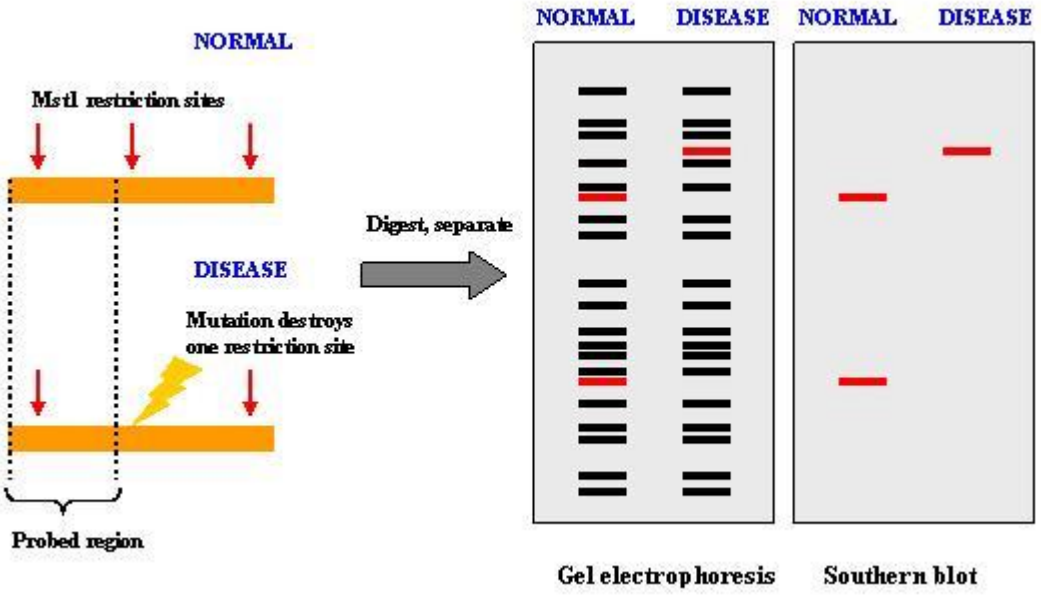
Introduction

- Restriction Fragment Length Polymorphism (RFLP) is a difference in homologous DNA sequences that can be detected by the presence of fragments of different lengths after digestion of the DNA samples in question with specific restriction endonucleases. RFLP, as a molecular marker, is specific to a single clone/restriction enzyme combination.
- Most RFLP markers are co-dominant (both alleles in heterozygous sample will be detected) and highly locus-specific.
- An RFLP probe is a labeled DNA sequence that hybridizes with one or more fragments of the digested DNA sample after they were separated by gel electrophoresis, thus revealing a unique blotting pattern characteristic to a specific genotype at a specific locus. Short, single- or low-copy genomic DNA or cDNA clones are typically used as RFLP probes.
- The RFLP probes are frequently used in genome mapping and in variation analysis (genotyping, forensics, paternity tests, hereditary disease diagnostics, etc.).

Developing RFLP probes

- Total DNA is digested with a methylation-sensitive enzyme (for example, *Pst*I), thereby enriching the library for single- or low-copy expressed sequences (*Pst*I clones are based on the suggestion that expressed genes are not methylated).

- The digested DNA is size-fractionated on a preparative agarose gel, and fragments ranging from 500 to 2000 bp are excised, eluted and cloned into a plasmid vector (for example, pUC18).
- Digests of the plasmids are screened to check for inserts.
- Southern blots of the inserts can be probed with total sheared DNA to select clones that hybridize to single- and low-copy sequences.
- The probes are screened for RFLPs using genomic DNA of different genotypes digested with restriction endonucleases. Typically, in species with moderate to high polymorphism rates, two to four restriction endonucleases are used such as *Eco*RI , *Eco*RV, and *Hin*dIII. In species with low polymorphism rates, additional restriction endonucleases can be tested to increase the chance of finding polymorphism.



Simple sequence length polymorphisms (SSLPs)

- SSLPs are arrays of repeat sequences that display length variations, different alleles containing different numbers of repeat units.
- Unlike RFLPs that can have only two alleles, SSLPs can be multi-allelic as each SSLP can have a number of different length variants.
- There are two types of SSLP, both of which were described in Minisatellites, also known as variable number of tandem repeats (VNTRs), in which the repeat unit is up to 25 bp in length.

- Microsatellites or simple tandem repeats (STRs), whose repeats are shorter, usually dinucleotide or tetranucleotide units.

Single nucleotide polymorphisms (SNPs)

- Single nucleotide polymorphisms, frequently called SNPs (pronounced "snips"), are the most common type of genetic variation among people.
- Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA.
- SNPs occur normally throughout a person‟s DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome.
- Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease.
- When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene‟s function.
- Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health.
- Researchers have found SNPs that may help predict an individual‟s response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases.
- SNPs can also be used to track the inheritance of disease genes within families.
- Future studies will work to identify SNPs associated with complex diseases such as heart disease, diabetes, and cancer.

Oligonucleotide hybridization

- Oligonucleotide hybridization can therefore discriminate between the two alleles of an SNP.
- Various screening strategies have been devised including DNA chip technology and solution hybridization techniques.

Linkage analysis is the basis of genetic mapping

- Chromosomes are inherited as intact units, so it was reasoned that the alleles of some pairs of genes will be inherited together because they are on the same chromosome.
- This is the principle of genetic linkage, Pairs of genes were either inherited independently, as expected for genes in different chromosomes, or, if they

showed linkage, then it was only partial linkage: sometimes they were inherited together and sometimes they were not.

- 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.

The LOD score

- The LOD score often used for linkage analysis in human populations, and also in animal and plant populations.
- Computerized LOD score analysis is a simple way to analyze complex family pedigrees in order to determine the linkage between Mendelian traits (or between a trait and a marker, or two markers).
- The method briefly, works as follows:
  - ✓ Establish a pedigree
  - ✓ Make a number of estimates of recombination frequency
  - ✓ Calculate a LOD score for each estimate
  - ✓ The estimate with the highest LOD score will be considered the best estimate
- The LOD score is calculated as follows:

$$LOD = Z = \text{Log10} \frac{\text{probability of birth sequence with a given linkage}}{\text{probability of birth sequence with no linkage}}$$

- By convention, a LOD score greater than 3.0 is considered evidence for linkage.
- On the other hand, a LOD score less than -2.0 is considered evidence to exclude linkage.

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.
  - ✓ Genetic maps have limited accuracy.
- Relies upon observable experimental outcomes
  - ✓ hybridization
  - ✓ amplification
- May or may not have a distance measure.

Physical mapping techniques

- Restriction mapping, which locates the relative positions on a DNA molecule of the recognition sequences for restriction endonucleases.
- Fluorescentin situhybridization (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.

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.
- Restriction mapping is a method used to map an unknown segment of DNA by breaking it into pieces and then identifying the locations of the breakpoints.
- This method relies upon the use of proteins called restriction enzymes, which can cut, or digest, DNA molecules at short, specific sequences called restriction sites.
- After a DNA segment has been digested using a restriction enzyme, the resulting fragments can be examined using a laboratory method called gel electrophoresis, which is used to separate pieces of DNA according to their size.
- One common method for constructing a restriction map involves digesting the unknown DNA sample in three ways.
- Here, two portions of the DNA sample are individually digested with different restriction enzymes, and a third portion of the DNA sample is double-digested with both restriction enzymes at the same time.
- Next, each digestion sample is separated using gel electrophoresis, and the sizes of the DNA fragments are recorded. The total length of the fragments in each digestion will be equal.
- However, because the length of each individual DNA fragment depends upon the positions of its restriction sites, each restriction site can be mapped according to the lengths of the fragments.
- The information from the double-digestion is particularly useful for correctly mapping the sites. The final drawing of the DNA segment that shows the positions of the restriction sites is called a restriction map.

Limitations of Restriction mapping

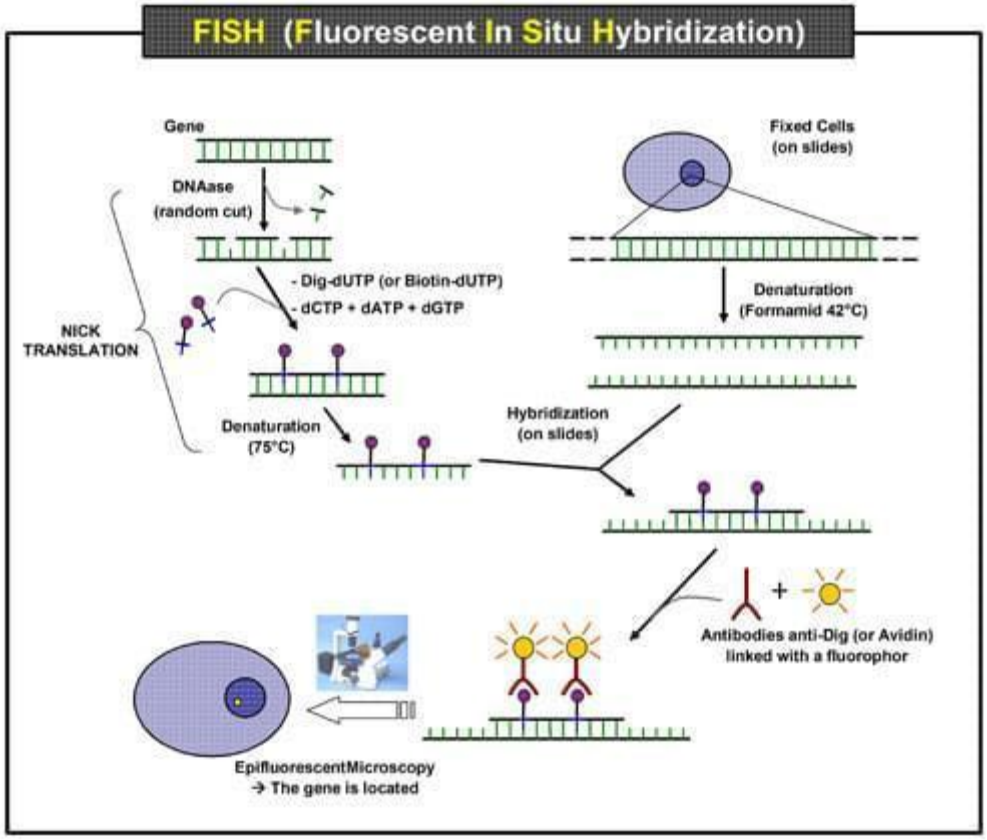
- Restriction mapping is more applicable to small rather than large molecules, with the upper limit for the technique depending on the frequency of the restriction sites in the molecule being mapped.
- In practice, if a DNA molecule is less than 50 kb in length it is usually possible to construct an unambiguous restriction map for a selection of enzymes with six-nucleotide recognition sequences.
- The limitations of restriction mapping can be eased slightly by choosing enzymes expected to have infrequent cut sites (rare cutter) in the target DNA molecule.

Rare cutters

- These „rare cutters' fall into two categories:
  - ✓ Enzymes with seven- or eight-nucleotide recognition sequences
  - ✓ Enzymes whose recognition sequences contain motifs that are rare in the target DNA

Fluorescence *in situ* hybridization (FISH)

- It is a kind of cytogenetic technique which uses fluorescent probes binding parts of the chromosome to show a high degree of sequence complementarity. Fluorescence microscopy can be used to find out where the fluorescent probe bound to the chromosome.
- This technique provides a novel way for researchers to visualize and map the genetic material in an individual cell, including specific genes or portions of genes.
- It is an important tool for understanding a variety of chromosomal abnormalities and other genetic mutations. Different from most other techniques used for chromosomes study, FISH has no need to be performed on cells that are actively dividing, which makes it a very versatile procedure.



Methodology

FISH is useful for example, to help a researcher identify where a particular gene falls within an individual's chromosomes. Here's how it works:

- Make a probe complementary to the known sequence. When making the probe, label it with a fluorescent marker, e.g. fluorescein, by incorporating nucleotides that have the marker attached to them.
- Put the chromosomes on a microscope slide and denature them.
- Denature the probe and add it to the microscope slide, allowing the probe hybridize to its complementary site.

- Wash off the excess probe and observe the chromosomes under a fluorescent microscope. The probe will show as one or more fluorescent signals in the microscope, depending on how many sites it can hybridize to.

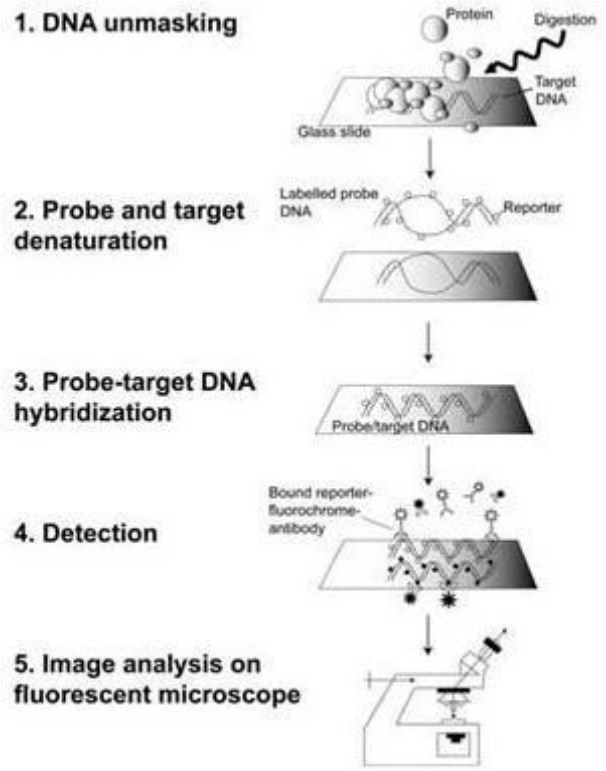


Fig. The five basic steps of FISH

Probes used in FISH

Generally, researchers use three different types of FISH probes, each of which has a different application:

1. Locus specific probes:

It bind to a particular region of a chromosome. This type of probe is useful when researchers have isolated a small portion of a gene and want to determine on which chromosome the gene is located.

2. Alphoid or centromeric repeat probes:

They are generated from repetitive sequences found in the middle of each chromosome. Researchers use these probes to determine whether an individual has the correct number of chromosomes. These probes can also be used in combination with "locus specific probes" to determine whether an individual is missing genetic material from a particular chromosome.

3. Whole chromosome probes

They are actually collections of smaller probes, each of which binds to a different sequence along the length of a given chromosome. Using multiple probes labeled with a mixture of different fluorescent dyes, scientists are able to label each chromosome in its own unique color. The resulting full-color map of the chromosome is known as a spectral karyotype. Whole chromosome probes are particularly useful for examining chromosomal abnormalities, for example, when a piece of one chromosome is attached to the end of another chromosome.

Applications

- FISH is widely used for several diagnostic applications: identification of numerical and structural abnormalities,
- Characterization of marker chromosomes, monitoring the effects of therapy, detection of minimal residual disease
- Ttracking the origin of cells after bone marrow transplantation, identification of regions of deletion or amplification,
- Detection of chromosome abnormalities in non-dividing or terminally differentiated cells, determination of lineage involvement of clonal cells, etc.
- Moreover it has many applications in research: identification of non-random chromosome rearrangements, identification of translocation molecular breakpoint, identification of commonly deleted regions, gene mapping, characterization of somatic cells hybrids, identification of amplified genes, study the mechanism of rearrangements.
- FISH is also used to compare the genomes of two biological species to deduce evolutionary relationships.

Sequence Tagged Site (STS)

Sequence-tagged site (STS)

- It is a short region along the genome (200 to 300 bases long) whose exact sequence is found nowhere else in the genome.
- The uniqueness of the sequence is established by demonstrating that it can be uniquely amplified by the PCR.

- The DNA sequence of an STS may contain repetitive elements, sequences that appear elsewhere in the genome, but as long as the sequences at both ends of the site are unique, unique DNA primers complementary to those ends can be synthesized, the region amplified using PCR, and the specificity of the reaction demonstrated by gel electrophoresis of the amplified product.

Applications of STS

- STSs are very helpful for detecting microdeletions in some genes. For example, some STSs can be used in screening by PCR to detect microdeletions in Azoospermia (AZF) genes in infertile men.
- Identification of genes in elephants could provide additional information for evolutionary studies and for evaluating genetic diversity in existing elephant populations.
- Sequence tagged sites (STSs) were identified in the Asian and the African elephant for the following genes: melatonin receptor 1a (MTNR1A), retinoic acid receptor beta (RARB), and leptin receptor.

Map resources

Assembly

A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.

BioProject (formerly Genome Project)

A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.

CloneDB (formerly Clone Registry)

A database that integrates information about clones and libraries, including sequence data, map positions and distributor information.

Database of Genome Survey Sequences (dbGSS)

A division of GenBank that contains short single-pass reads of genomic DNA. dbGSS can be searched directly through the Nucleotide GSS Database.

Database of Genomic Structural Variation (dbVar)

The dbVar database has been developed to archive information associated with large scale genomic variation, including large insertions, deletions, translocations and inversions. In addition to archiving variation discovery, dbVar also stores associations of defined variants with phenotype information.

Genome

Contains sequence and map data from the whole genomes of over 1000 organisms. The genomes represent both completely sequenced organisms and those for which sequencing is in progress. All three main domains of life (bacteria, archaea, and eukaryota) are represented, as well as many viruses, phages, viroids, plasmids, and organelles.

Genome Reference Consortium (GRC)

The Genome Reference Consortium (GRC) maintains responsibility for the human and mouse reference genomes. Members consist of The Genome Center at Washington University, the Wellcome Trust Sanger Institute, the European Bioinformatics Institute (EBI) and the National Center for Biotechnology Information (NCBI). The GRC works to correct misrepresented loci and to close remaining assembly gaps. In addition, the GRC seeks to provide alternate assemblies for complex or structurally variant genomic loci. At the GRC website (http://www.genomereference.org), the public can view genomic regions currently under review, report genome-related problems and contact the GRC.

HIV-1, Human Protein Interaction Database

A database of known interactions of HIV-1 proteins with proteins from human hosts. It provides annotated bibliographies of published reports of protein interactions, with links to the corresponding PubMed records and sequence data.

Influenza Virus

A compilation of data from the NIAID Influenza Genome Sequencing Project and GenBank. It provides tools for flu sequence analysis, annotation and submission to GenBank. This resource also has links to other flu sequence resources, and publications and general information about flu viruses.

NCBI Pathogen Detection Project

A project involving the collection and analysis of bacterial pathogen genomic sequences originating from food, environmental and patient isolates. Currently, an automated pipeline clusters and identifies sequences supplied primarily by public health laboratories to assist in the investigation of foodborne disease outbreaks and discover potential sources of food contamination.

Nucleotide Database

A collection of nucleotide sequences from several sources, including GenBank, RefSeq, the Third Party Annotation (TPA) database, and PDB. Searching the Nucleotide Database will yield available results from each of its component databases.

PopSet

Database of related DNA sequences that originate from comparative studies: phylogenetic, population, environmental and, to a lesser degree, mutational. Each record in the database is a set of DNA sequences. For example, a population set provides information on genetic variation within an organism, while a phylogenetic set may contain sequences, and their alignment, of a single gene obtained from several related organisms.

Probe

A public registry of nucleic acid reagents designed for use in a wide variety of biomedical research applications, together with information on reagent distributors, probe effectiveness, and computed sequence similarities.

Retrovirus Resources

A collection of resources specifically designed to support the research of retroviruses, including a genotyping tool that uses the BLAST algorithm to identify the genotype of a query sequence; an alignment tool for global alignment of multiple sequences; an HIV-1 automatic sequence annotation tool; and annotated maps of numerous retroviruses viewable in GenBank, FASTA, and graphic formats, with links to associated sequence records.

SARS CoV

A summary of data for the SARS coronavirus (CoV), including links to the most recent sequence data and publications, links to other SARS related resources, and a pre-computed alignment of genome sequences from various isolates.

Sequence Read Archive (SRA)

The Sequence Read Archive (SRA) stores sequencing data from the next generation of sequencing platforms including Roche 454 GS System®, Illumina Genome Analyzer®, Life Technologies AB SOLiD System®, Helicos Biosciences Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Trace Archive

A repository of DNA sequence chromatograms (traces), base calls, and quality estimates for single-pass reads from various large-scale sequencing projects.

Viral Genomes

A wide range of resources, including a brief summary of the biology of viruses, links to viral genome sequences in Entrez Genome, and information about viral Reference Sequences, a collection of reference sequences for thousands of viral genomes.

Virus Variation

An extension of the Influenza Virus Resource to other organisms, providing an interface to download sequence sets of selected viruses, analysis tools, including virus-specific BLAST pages, and genome annotation pipelines.

FTP: Genome

This site contains genome sequence and mapping data for organisms in Entrez Genome. The data are organized in directories for single species or groups of species. Mapping data are collected in the directory MapView and are organized by species. See the README file in the root directory and the README files in the species subdirectories for detailed information.

FTP: Genome Mapping Data

Contains directories for each genome that include available mapping data for current and previous builds of that genome.

FTP: RefSeq

This site contains all nucleotide and protein sequence records in the Reference Sequence (RefSeq) collection. The ""release"" directory contains the most current release of the complete collection, while data for selected organisms (such as human, mouse and rat) are available in separate directories. Data are available in FASTA and flat file formats. See the README file for details.

FTP: SKY/M-Fish and CGH Data

This site contains SKY-CGH data in ASN.1, XML and EasySKYCGH formats. See the skycghreadme.txt file for more information.

FTP: Sequence Read Archive (SRA) Download Facility

This site contains next-generation sequencing data organized by the submitted sequencing project.

FTP: Trace Archive

This site contains the trace chromatogram data organized by species. Data include chromatogram, quality scores, FASTA sequences from automatic base calls, and other ancillary information in tab-delimited text as well as XML formats. See the README file for details.

FTP: Whole Genome Shotgun Sequences

This site contains whole genome shotgun sequence data organized by the 4-digit project code. Data include GenBank and GenPept flat files, quality scores and summary statistics. See the README.genbank.wgs file for more information.

A haplotype map of the human genome

- The planned Haplotype Map is the next logical step in mobilizing tools for gene discovery.
- The most common type of variation in the human genome is the single nucleotide polymorphism or SNP, a single-base difference at a genetic locus from person to person.
- Millions of SNPs have been found, making it imperative that we find efficient and cost-effective ways for using them.
- The Haplotype Map is based on the recognition that the development of genetic variation from ancestral chromosomes has not proceeded uniformly across the genome.
- Rather, there appear to be regions in which recombination is more likely to occur, thus shuffling the genetic deck at those points.
- There are other regions where is it less likely to occur, leaving relatively large blocks intact. These blocks or haplotypes can be identified by a small number of SNPs.
- Wise use of genetic markers will be enhanced by knowing the boundaries of these blocks. To be sure, a clear haplotype structure may not be apparent everywhere in the genome, but knowledge of the haplotype structure of the genome will speed the search for loci that confer disease risk.
- The Hap Map should help us use genetic markers wisely, to speed up (and to make affordable) association studies based on candidate genes and ultimately, whole-genome

association studies. Without the Hap Map, the choice of markers for association studies will remain more or less a matter of guesswork.

Association Mapping

- Association mapping (genetics), also known as "linkage disequilibrium mapping", is a method of mapping quantitative trait loci (QTLs) that takes advantage of historic linkage disequilibrium to link phenotypes (observable characteristics) to genotypes (the genetic constitution of organisms), uncovering genetic associations.
- Association mapping is based on the idea that traits that have entered a population only recently will still be linked to the surrounding genetic sequence of the original evolutionary ancestor, or in other words, will more often be found within a given haplotype, than outside of it.
- It is most often performed by scanning the entire genome for significant associations between a panel of SNPs (which, in many cases are spotted onto glass slides to create " SNP chips") and a particular phenotype.
- These associations must then be independently verified in order to show that they either (a) contribute to the trait of interest directly, or (b) are linked to/ in linkage disequilibrium with a quantitative trait locus (QTL) that contributes to the trait of interest.
- The advantage of association mapping is that it can map quantitative traits with high resolution in a way that is statistically very powerful.
- Association mapping, however, also requires extensive knowledge of SNPs within the genome of the organism of interest, and is therefore difficult to perform in species that have not been well studied or do not have well-annotated genomes.

Benefits of Genetic Mapping

The techniques developed for genetic mapping have had great impact on the life sciences, and particularly in medicine. But genetic mapping technologies also have useful applications in other fields. Commercialization of the fruits of genomics research promises immense opportunities for industry. A round-up of genetic mapping applications would include (but not be limited to) the areas below.

Medicine

Scientists have become more proficient in genetic sequencing - the detailed genetic maps that help locate the risk genes for a host of genetic diseases. The ability to investigate the root cause of diseases may one day allow medical researchers to develop strategies to avoid the environmental conditions that serve as triggers to disease, formulate customized drugs, and techniques for gene therapy.

Agricultural Applications

Knowledge of the genetic maps of plants and animals leads to the development of agricultural crops and animal breeds that are more nutritious, productive and can better resist diseases, insects and drought. Researchers can breed special plants that help clean up wastes that are difficult to break down.

Energy and the Environment

Genetic maps of microbes enable researchers to harness the power of bacteria for producing energy from bio-fuels, reducing toxic waste, and developing environment-friendly products and industrial processes.

Forensics

We are already familiar with the use of genetic mapping in crime investigations, paternity tests, and identification. The technique can also be used in organ transplants to achieve better matches between recipients and donors, thus minimizing the risks of complications and maximizing the use of donated healthy organs, a scarce resource. For more delectable applications, genetic mapping can authenticate the origins of consumer goods like caviar, fruits, and wine or the pedigree of livestock and animal breeds.

Genetic Markers

- Genetic markers are useful in identification of various genetic variations. The development of DNA-based genetic markers has had a revolutionary impact on genetic studies.
- With DNA markers, it is theoretically possible to observe and exploit genetic variation in the entire genome. These markers can be used to study the evolutionary relationships among individuals.
- Popular genetic markers include allozymes, mitochondrial DNA, RFLP, RAPD, AFLP, microsatellite, SNP, and EST markers.
- The application of DNA markers has allowed rapid progress in investigations of genetic variability and inbreeding, parentage assignments, species and strain identification, and the construction of high-resolution genetic linkage maps for aquaculture species.
- The advent of next-generation sequencing (NGS) has revolutionized genomic and transcriptomic approaches to biology.
- The new sequencing tools are also valuable for the discovery, validation and assessment of genetic markers in populations. This review focuses on importance and uses of genetic markers with advent of modern technologies.

Minisatellite

- Minisatellites have been found in association with important features of human genome biology such as gene regulation, chromosomal fragile sites, and imprinting. Our knowledge of minisatellite biology has greatly increased in the past 10 years owing to the identification and careful analysis of human hypermutable minisatellites, experimental models in yeast, and recent in vitro studies of minisatellite recombination properties.
- In parallel, minisatellites have been put forward as potential biomarkers for the monitoring of genotoxic agents such as ionizing radiation.
- We summarize and discuss recent observations on minisatellites. In addition we take advantage of recent whole chromosome sequence data releases to provide a unifying view which may facilitate the annotation of tandem repeat sequences.
- Minisatellites are usually defined as the repetition in tandem of a short (6- to 100-bp) motif spanning 0.5 kb to several kilobases.
- Although the first examples described 20 years ago were of human origin, (Wyman and White 1980), similar DNA structures have been found in many organisms including bacteria.
- Comparisons of the repeat units in classical minisatellites led early on to the notion of consensus or core sequences, which exhibit some similarities with the χ sequence of λ phage (GCTGTGG). In general, the majority of classical minisatellites are GC rich, with a strong strand asymmetry.

Microsatellite

- Microsatellites or Single Sequence Repeats (SSRs) are extensively employed in plant genetics studies, using both low and high throughput genotyping approaches.
- Motivated by the importance of these sequences over the last decades this review aims to address some theoretical aspects of SSRs, including definition, characterization and biological function.
- The methodologies for the development of SSR loci, genotyping and their applications as molecular markers are also reviewed.
- Finally, two data surveys are presented. The first was conducted using the main database of Web of Science, prospecting for articles published over the period from 2010 to 2015, resulting in approximately 930 records.

- The second survey was focused on papers that aimed at SSR marker development, published in the American Journal of Botany's Primer Notes and Protocols in Plant Sciences (over 2013 up to 2015), resulting in a total of 87 publications.
- This scenario confirms the current relevance of SSRs and indicates their continuous utilization in plant science.

Difference between Microsatellite and Minisatellite

Minisatellites	Microsatellites
a) Hypervriable family Repeat size:10-60bp Total Size:1000-20,000 bp b) Telomeric family: Repeat size: 6 bp Total size: 1000-20000 bp	Repeat size: 1-4 bp Total sites: Less than 1000 bp
Share a common core sequence (motif) GGGCAGGANG (where N is any base), dispersed, VNTRs usually TTAGGG and repeated about a thousand times protects chromosome ends.	Repeats A and CA are the most common Dispersed throughout genome.
Complexity of Array: Heterogeneous	**Complexity of Array:** Homogeneous

Sequence-Tagged Site (STS)

- It is a relatively short, easily PCR-amplified sequence (200 to 500 bp) which can be specifically amplified by PCR and detected in the presence of all other genomic sequences and whose location in the genome is mapped.
- The STS concept was introduced by Olson et al (1989). In assessing the likely impact of the Polymerase Chain Reaction (PCR) on human genome research, they recognized that single-copy DNA sequences of known map location could serve as markers for genetic and physical mapping of genes along the chromosome.
- The advantage of STSs over other mapping landmarks is that the means of testing for the presence of a particular STS can be completely described as information in a database: anyone who wishes to make copies of the marker would simply look up the STS in the database, synthesize the specified primers, and run the PCR under specified conditions to amplify the STS from genomic DNA.

- STS-based PCR produces a simple and reproducible pattern on agarose or polyacrylamide gel. In most cases STS markers are co-dominant, i.e., allow heterorozygotes to be distinguished from the two homozygotes.
- The DNA sequence of an STS may contain repetitive elements, sequences that appear elsewhere in the genome, but as long as the sequences at both ends of the site are unique and conserved, researches can uniquely identify this portion of genome using tools usually present in any laboratory.
- Thus, in broad sense, STS include such markers as microsatellites (SSRs, STMS or SSRPs), SCARs, CAPs, and ISSRs.

Expressed Sequence Tag (EST)

- It is a short stretch of DNA sequence that is used to identify an expressed gene. Although EST sequences are usually only 200 to 500 nucleotides in length, this is generally sufficient to identify the full-length complementary DNA (cDNA).
- ESTs are generated by sequencing a single segment of random clones from a cDNA library. A single sequencing reaction and automation of DNA isolation, sequencing, and analysis have allowed the rapid determination of many ESTs.
- Now, the majority of the sequences in sequence databases are ESTs. Although most ESTs have been isolated from humans, a large number of ESTs have been isolated from model organisms, such as Caenorhabditis elegans , Drosophila, rice, and Arabidopsis.
- ESTs are also being isolated from more exotic organisms, such as Entamoeba histolytica and Leishmania major promastigotes .
- ESTs have numerous uses, from genetic mapping to analyzing gene expression, and the number of ESTs isolated from different organisms will continue to rise rapidly.

Single nucleotide polymorphisms

- It is frequently called SNPs (pronounced “snips”), are the most common type of genetic variation among people.
- Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA.
- SNPs occur normally throughout a person‟s DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome.
- Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease.

- When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene‟s function.
- Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health.
- Researchers have found SNPs that may help predict an individual‟s response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases.
- SNPs can also be used to track the inheritance of disease genes within families. Future studies will work to identify SNPs associated with complex diseases such as heart disease, diabetes, and cancer.

Possible Questions

1. Explain: Genetic and physical mapping techniques.
2. Define SNPs and micro satellite genetic markers.
3. Give a detailed note on physical mapping techniques with its practical uses.
4. Write note on genetic mapping techniques and the practical uses of genetic mapping.
5. What are physical mapping techniques? Give its practical uses.
6. Briefly describe about the applications of genetic mapping.
7. Explain in detail about Yeast expression systems.
8. Explain in detail about viral expression systems.
9. Explain in detail about bacterial expression systems.
10. Define genetic markers. Explain about SNP in detail.
11. Explain Genetic mapping techniques and the practical uses of genetic mapping.
12. Explain physical mapping techniques and the practical uses of Map resources.
13. What are genetic markers? Give a detailed note on mini and micro satellite .

UNIT -III

Gene Expressions and Microarrays: Expression systems - Bacteria, Yeast and Viral. Concepts of microarrays, spotter analysis, Normalization –total intensity, using regression techniques, ratio statistics. Clustering Gene expression profiles-hierarchical, single-linkage, complete linkage, and average linkage. Tools for microarray analysis- MADAM, spot finder, SAGE Applications of Microarrays- Bioinformatics challenges in micro array design and analysis.

UNIT III

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Gene expression	1	11	T2: 107-131
Microarray	1		T1: 22-23
Concept of gradient system, Sample preparation, processing and staining	2		T1: 614-615 W1
	1		W2
Tools for analysis – MADAM, Spot finder.	1		W3 & W4
SAGE, Applications	1		R2: 143-147
Bioinformatics challenges in microarray design and techniques	2		R2: 147-148
Unit III Revision	1		
Unit III Class test	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.
R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.
W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs
W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/
W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html

Bacterial Expression System

The general procedure for expression of cloned genes in *E. coli* involves

- the insertion of the coding region of interest into a vector, usually a plasmid, so the region is efficiently transcribed and translated.
- Since eukaryotic genes do not contain the proper signals for transcription initiation, ribosome recognition, translation initiation, and translation termination, these signals need to be supplied by the vector.

Common promoters used in bacterial expression vectors are:

λPL, lac, tac and T7

Of these, the lac and tac promoters are the most widely used.

Control elements that regulate expression from these promoters are supplied by either the host or the vector.

The λPL promoter is controlled by a mutant cI repressor protein that is non-functional at 42°C.

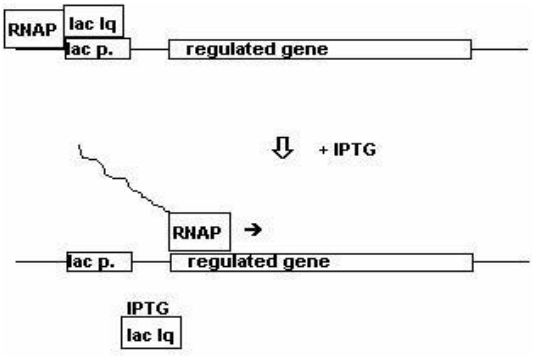
- At temperatures less that 42°C the repressor protein (cI) binds to the promoter and prevents expression.
- To induce expression the temperature is raised to that the repressor becomes nonfunctional and expression is now permitted.
- One problem with this system is that heat-shock proteins may also be induced.

The T7 promoter is from the T7 phage

- is only transcribed by the T7 RNA polymerase (T7 gene 1).
- The host cell must also contain the T7 gene 1 in order to express from this promoter.
- To regulate expression from this promoter it is necessary to be able to regulate the expression of the T7 gene 1.

The lac and tac promoters

- are controlled by the lac repressor.



- The lac repressor binds to the lac promoter and prevents RNA polymerase from transcribing the gene. Host strains with lacIq gene express the lac repressor at 10-fold higher concentrations than the normal
- lacI gene.
- IPTG (isopropyl-1-thio-β-D-galactopyanoside, an analog of lactose) binds to the repressor and prevents its interaction with the lac promoter and allowing RNA polymerase to transcribe the regulated gene.
- The tac promoter is a fusion of trp and lac promoters and is also regulated by IPTG.

The production of recombinant fusion proteins often stabilizes the expression of foreign proteins in *E. coli*. Several plasmids that express recombinant proteins as fusion proteins have been developed (Table).

Fusion Protein	Affinity Matrix
Glutathione-Stransferase	glutathione agarose

Thioredoxin	phenylarsine oxide-agarose
Maltose Binding Protein	amyloseagarose
Six Histidine Residues	Ni-agarose

In addition to stabilizing the recombinant protein, the fusion partner is often exploited for affinity purification or for the analysis of the recombinant protein.

- Recombinant proteins expressed at high levels will sometimes form insoluble aggregates known a inclusion bodies.
- In some applications it is possible to take advantage of this phenomenon.
- For example, the inclusion bodies can be isolated by differential centrifugation and solubilized under denaturing conditions (eg., urea). It is sometimes possible to renature the protein and regain activity. In addition, fusions with *E. coli* thioredoxin can circumvent inclusion body formation.

Bacterial expression vectors

pBR322 - An ideal plasmid vector must have the following functions:

(1) minimum amount of DNA,
(2) relaxed replication control,
(3) at least two selectable markers,
(4) only one (unique) recognition site for at least one restriction endonuclease, and
(5) for easy selection of the recombinant DNA, this unique restriction site must be located within one of the two selectable

markers. The name pBR denotes the following:

p signifies plasmid,

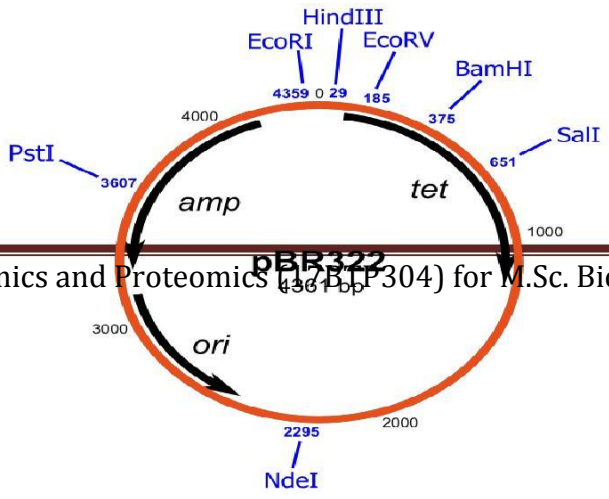
B is from Boliver, and

R is from Rodriguez, the two initials of the scientist who developed pBR322.

pBR322

- is the most popular and
- most widely used plasmid of 4363 bp;
- It has the replication module of *E. coli* plasmid Col El.
- This module has been incorporated in many other plasmid vectors since it permits plasmid replication even when chromosome replication and cell division are inhibited by amino acid starvation or chloramphemicol.
- Under such conditions, each cell accumulates several thousand copies of the plasmid so that one litre of bacterial culture easily yields a milligram of plasmid DNA.
- It has two selectable markers (tetracycline, tetr, and ampicillin, amp', resistance genes), and unique recognition sites for 12 different restriction enzymes (two unique sites, PstI and PvuI, are located within the amp' gene, and 4, e.g., BamHI, SalI, etc., are within tetr gene).

  The presence of restriction sites within the markers tetr and ampr permits an easy selection for cells transformed with the recombinant pBR322.
- Insertion of the DNA fragment into the plasmid using restriction enzyme PstI or PvuI places the DNA insert within the gene amp'; this makes amp' nonfunctional.
- Bacterial cells containing such a recombinant pBR322 will be unable to grow in the presence of ampicillin, but will grow on tetracycline.
- Similarly, when restriction enzyme BamHI or SalI is used, the DNA insert is placed within the gene tetr making it nonfunctional.



- Bacterial cells possessing such a recombinant pBR322 will, therefore, grow on ampicillin but not on tetracycline.
- This feature allows an easy selection of a single bacterial cell having recombinant pBR322 from among 108 other types of cells.
- Transformed E. coli cells are first plated on an agar medium containing the antibiotic within the resistance gene for which the DNA fragment is not inserted, i.e., for which the bacterial cells having the recombinant DNA are expected to be resistant.
- This eliminates nontransformed bacterial cells; the resulting bacterial colonies will posses either recombinant or unaltered pBR322.
- The colonies so obtained are then replicaplated on agar plates containing the other antibiotic (within the resistance gene for which the DNA insert is placed);
- all the colonies that develop on this plate will contain the unaltered pBR322.
- Therefore, the antibiotic sensitive colonies are identified and recovered from the master plate; these colonies will have the recombinant pBR322. This entire process may take up to 2 days.

pUC18-19 –

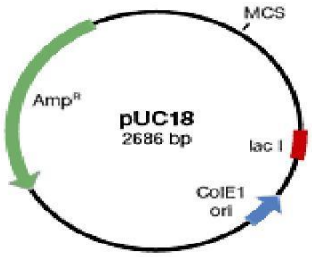
- pUC gets its name from University of California.
- It is a derivative of pBR322 and is much smaller (-2.7 kb);
- it has all the essential parts of pBR322, e.g.,

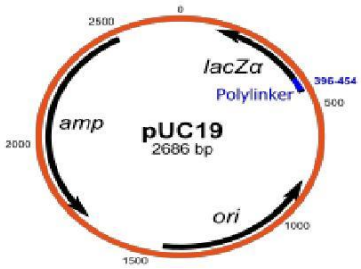
(1) ampicillin resistance gene and

(2) Col E1 origin.

(3) The second scrobale marker is due to *E. coli* gene lacZα encoding the a fragment of β-galactosidase, the enzyme that hydrolyses lactose.

(4) The *E.coli* strains, e.g., JMIO3, JM1O9, used as hosts for the pUC series vectors have the lacZα deleted from their lacZ genes.





- When pUC enters such an *E.coli* cell, the host genome and the plasmid encode for different parts of the β-galactosidase enzyme, which interact with each other to produce the active enzyme enabling these cells to hydrolyse lactose. β -galactosidase also hydrolyses X-gal (5-Bromo-4-chloro-3-indolyl-p-D-galactoside) to yield a blue dye.
- Therefore appropriate lacZ- *E. coli* cells transformed by the pUC vectors behave as lacZ+ and produce blue coloured colonies on a X-gal containing medium.
- A poly linker sequence located within the lacZα provides several (10 in case of pUC18/pUC19) unique restriction sites for DNA insertion.
- The polylinker sequence by itself does not interfere with lacZα ' expression.
- But when a DNA insert is placed within it, lacZα expression is prevented.
- Vectors pUC18 and pUC19 are identical, except for the orientation of the polylinker sequence, which is oriented in the opposite directions in the two vectors.
- The unique restriction sites used for integration of DNA inserts into pUC vectors interrupt the lacZα fragment so that appropriate
- E. coli cells possessing recombinant pUC DNA are β-galactosidase deficient and, as a result,

- produce white colonies on X-gal medium.
- Therefore, appropriate E.coli cells transformed with pUC recombinant DNA are grown on ampicillin, X-gal and IPG (isopropyl- β D-thiogalactoside; it serves as inducer of β -galactosidase, while X-gal itself can not) containing medium to eliminate non transformed cells. The white colonies are selected as they contain the recombinant DNA (in contrast, blue colonies will contain the unaltered vector). The other vectors in pUC series are pUC 8, pUC 9, pUC 12, pUC 13, etc.

The pUC series vectors offer the following adantages over pBR322:

(1) each E. coli cell produces up to 700 copies without any treatment,

(2) cells containing recombinant DNA are selected in a single step,

(3) the sites for DNA insert integration are confined to the poly linker, which permits the use of two restriction anzymes to open the vector, and

(4) they also allow sequencing of the DNA insert.

Viral Expression system

Viral vectors vary in their efficiency and toxicity and each are ideally suited to different applications but they share common traits that differentiate them from viruses not suited as vectors.

Lentiviral vectors

- Lentiviruses are used for either transient or stable gene expression, infecting both dividing and nondividing cells and integrating into the host cell‘s genome (however, lentiviruses deficient in integrase are available if needed). Insert capacity for lentiviruses (i.e., the size of the gene to be inserted) is up to about 5 kb, depending on the engineered constructs—overall, lower than other vectors (see below).

- Lentiviral vector tools are available for a variety of experimental conditions. For example, Dharmacon (part of GE Healthcare‘s Life Sciences business) provides a range of lentiviral products for gene knockdown or knockout. These include the Dharmacon™ Edit-R™ CRISPR-Cas9 platform (for gene editing) and the Precision

LentiORFs (human open-reading frames in lentiviral expression vectors), says Melissa Kelley, R&D manager for Dharmacon, GE Healthcare.

- Transduction with lentivirus tends to be very efficient (usually near 100%), so Kelley says sometimes you need to be more concerned with your multiplicity of infection (MOI) level, depending on your experiment.

- A lower MOI (less than 1) is important when you want cells to be infected by only one lentiviral particle (for example, in pooled screening applications or in the creation of stable cell lines). ―In this case, you don't necessarily want a high transduction efficiency—you want to have only single integrations into cells,‖ says Kelley.

  Subsequent selection of transduced cells by antibiotic resistance or fluorescent reporters can yield a population of 100% cells transduced with single integrations. Conversely, other situations may call for a high MOI, such as a short-term gene knockdown experiment.

- Recent studies using lentiviral vector-based hematopoietic stem cell gene therapy have shown progress in treating X-linked severe combined immunodeficiency (a genetic disorder popularly known as ―bubble boy disease‖), as well as Wiskott-Aldrich Syndrome (a genetic disease of the immune system). In addition, lentiviral gene therapy for the blood disease beta-thalassemia resulted in patients being able to stop regular blood transfusions.

Adenoviral vectors

- Adenoviral vectors infect both dividing and nondividing cells, and they don't integrate into host cells, so expression is transient. The packaging capacity of adenovirus vectors typically accommodates larger inserts (from 8 kb to 36 kb, when using helper-dependent vectors). Adenoviruses also tend to have high levels of protein expression, which can be valuable depending on the selected application.

- For viral vector production and purification, Sartorius Stedim Biotech offers a range of bioprocess tools, from low-volume production (with the ambr® 15 fermentation microscale bioreactor system) to large-scale production (with the BIOSTAT® STR bioreactor). The company's Vivapure® AdenoPACK™ and LentiSELECT virus purification kits are used in basic research, and many of the tools are used by ―companies

producing commercial products that are in clinical trials,‖ says Kim Bure, director of regenerative medicine at Sartorius Stedim Biotech.

- AMS Biotechnology also offers many viral vector tools, including lentivirus, adenovirus, and AAV — and a large collection of human open-reading frame cDNA clones in adenovirus and lentivirus vectors. Although many of their tools are used in basic research, ―we always work together with researchers to help them move the projects into applied studies, if this is their intention,‖ says technical sales representative Maja

  Petkovic, who notes that the persistence of basic research is helping to fuel the progress of therapeutic viral vectors. ―Continuing development in changing virus envelope and tissue targeting specificity can help further increase the efficacy of viral gene therapy and reduce the side effects,‖ says Petkovic.

- Adenoviral vectors have recently been used as a delivery vehicle to fight prostate cancer.

  Tumor cells are transduced with a gene that prompts the patient‘s immune system to attack the tumor cells. Because the tumor cells are essentially led to self-destruction, the technique is known as ―suicide gene therapy.‖

AAV vectors

- AAV vectors infect both dividing and nondividing cells for stable expression, are nonintegrating and nonpathogenic to humans (which is convenient for therapeutic applications). AAV vectors don‘t insert randomly but rather at a specific site in the human genome. A potential disadvantage, depending on your application, is that AAV vectors have a smaller insert capacity (less than 3 kb) compared with either lenti- or adenoviruses.

- Cell Biolabs offers AAV (and other viral) vector tools for research use, such as the ViraDuctin™ AAV Transduction Kit, which uses proprietary reagents that increase AAV transduction efficiency. As viral vectors emerge as a clinical tool, Ken Rosser, director of marketing and sales at Cell Biolabs, notes that it‘s important to consider the effects that different viral vectors can have on patients‘ immune systems.

- ―The lowest possible immune response is desirable, both to ensure proper delivery of the gene and to minimize complications,‖ says Rosser. ―This gives AAV vectors an advantage, since they have been demonstrated to elicit lower immune responses compared to both lentivirus and adenovirus.‖

- Being less immunogenic makes AAV vectors good candidates for therapeutic applications, such as the introduction of a normal gene into the eyes of patients with retinal pigmentosa, which gradually leads to blindness, if untreated. Recent clinical trials using AAV vectors have improved or restored patients‘ vision. At the University of

  Pittsburgh School of Medicine, a clinical trial is also underway using AAV to introduce the gene for a key enzyme into patients with Parkinson‘s Disease. The enzyme is responsible for converting the drug levodopa into the neurotransmitter dopamine, which prolongs the positive effects of the drug as neurons deteriorate.

Viral vectors are choosen because:

1. They‘re safe

Owing to the deletion of the component of the viral genome critical in viral replication, viruses will not reproduce uncontrollably. A helper virus can be used as a companion to facilitate replication if desired. There is some concern over insertional mutagenesis leading to cell malignant transformation of cells and tumor development however.

2. They‘re minimally cytotoxic

If a minimally toxic viral vector is chosen, it has very low effects on the physiology or the cell housing the product of the transfection. Some cell lines are more cytotoxic than others, so care needs to be used when selecting your viral vector.

3. They‘re genetically stable

This allows for confidence that the virus will not mutate and affect the reproducibility of an experiment but limits the selection of viruses that are appropriate for such experiments.

4. They can be broadly targeting or highly cell specific

You can target a specific cell type or a group or open the floor up to a range of cells. This gives you a lot more flexibility in your experimental design and increases the applications of this technique.

5. Selection markers can be used

You can easily find your cells that have been transfected when selection markers are used. These selection markers can include fluorescent proteins that light up the cells that have been successfully transfected or there are antibiotics resistant markers and the classic blue-white colony technique.

Choosing Your Viral Vector

Various properties such as virus size, insertion size, expression, and max titer should be considered when deciding which cell type to choose. This wonderful chart from Life Technologies shows how each of the more common viral vectors has its own advantages, best suited applications and problems.

Viral System	Size	DNA insert size	Max titer (particles/mL)	Infection	Expression	Drawbacks
Adenovirus	36 kb (dsDNA)	8 kb	1×10^{13}	Dividing and non-dividing cells	Transient	Elicits strong antiviral immune response
Retrovirus	7–11 kb (ssRNA)	8 kb	1×10^{9}	Dividing cells	Stable	Insertional mutagenesis potential
Lentivirus	8 kb (ssRNA)	9 kb	1×10^{9}	Dividing and non-dividing cells	Stable	Insertional mutagenesis potential
Adeno-associated virus	8.5 kb (ssDNA)	5 kb	1×10^{11}	Dividing and non-dividing cells	Stable; site-specific integration	Requires helper virus for replication; difficult to produce pure viral stocks
Baculovirus	80–180 kb (dsDNA)	no known upper limit	2×10^{8}	Dividing and non-dividing cells	Transient or stable	Limited mammalian host range
Vaccinia virus	190 kb (dsDNA)	25 kb	3×10^{9}	Dividing cells	Transient	Potential cytopathic effects
Herpex simplex virus	150 kb (dsDNA)	30–40 kb	1×10^{9}	Dividing and non-dividing cells	Transient	No gene expression during latent infection

Yeast expression system

- Expression of proteins in yeast is a common alternative to prokaryotic and higher eukaryotic expression. Yeast cells offer many of the advantages of producing proteins in microbes (growth speed, easy genetic manipulation, low cost media) while offering some of the attributes of higher eukaryotic systems (post translational modifications, secretory expression). Several yeast protein expression systems exist in organisms from the genera *Saccharomyces*, *Pichia*, *Kluyveromyces*, *Hansenula* and *Yarrowia*.

- At the heart of any yeast expression system is the expression vector. Vectors that integrate into the host chromosome are most widely used because of their mitotic stability in the absence of a selection. However, episomal expression vectors exist for some yeast systems. Expression vectors typically contain a strong yeast promoter/terminator and a yeast selectable marker cassette. Most yeast vectors can be propagated and amplified in *E. coli* to facilitate cloning and as such, also contain an *E. coli* replication origin and ampicillin selectable marker. Finally, many yeast expression vectors include the ability to optionally clone a gene downstream of an efficient secretion leader (usually that of mating factor) that efficiently directs a heterologous protein to become secreted from the cell.

- A growing number of engineered yeast strains are becoming available for protein expression. Strains have been described that increase yield of secreted proteins, improve the performance of certain affinity tags, reduce proteolysis, define the composition of N-glycans, and permit non-native amino acids (e.g. selenomethionine) into proteins have been described.

- One yeast system that is commonly used for protein expression is *Kluyveromyces lactis.*

  K. lactis was one of the first yeasts to be used as an expression host in the early 1980‘s. It was developed for use in the food industry and was harnessed for industrial production of bovine prochymosin, an important cheese manufacturing enzyme. In addition, *K. lactis* and *S. cerevisiae* are the only two yeasts classified by the NIH as Host-Vector I systems, an important biosafety designation, making it an attractive host for use in regulated industries. The system was adapted for life sciences research by NEB and a variety of vectors and host strains in a frozen competent format suitable are offered.

Microarray technique

- Microarray analysis has emerged in the last few years as a flexible method for analyzing large numbers of nucleic acid fragments in parallel.

- Its origins can be traced to several different disciplines and techniques. Microarrays can be seen as a continued development of molecular biology hybridization methods, as an extension of the use of fluorescence microscopy in cell biology, as well as a diagnostic assay using capture to solid surface as a way to reduce the amount of analytes needed.

- The convergence of ideas and principles utilized in these fields, together with technological advancements in preparing miniaturized collections of nucleic acids on solid supports, have all contributed to the emergence of microarray and microchip technologies. In molecular biology, analysis of nucleic acids by hybridization is a universally adopted key method for analysis.

- Filter-based dot blot analysis has been used for a long time as a convenient method for analyzing multiple samples by hybridization. Classical gene expression analysis methods such as Northern blotting, reverse transcriptase polymerase chain reaction (RT-PCR) and nuclease protection assays, are best suited for analyzing a limited number of genes and samples at a time.

- By reversing the Northern blotting principle so that the labelled moiety is derived from the mRNA sample and the immobilized fractions are the known sequences traditionally used as probes, filter-based gene expression analysis has enabled simultaneous determination of expression levels of thousands of genes in one experiment.

- Because of the ease of use of these filter-based methods and their compatibility with general lab equipment, these macroarrays have been widely adopted for gene expression studies.

- One disadvantage to using this method has been the relatively large size and the autofluorescence of the membrane, which prevents efficient use of multiplexed fluorescent probes and subsequently limits the number of samples that can be analyzed in each experiment

- A microarray is a pattern of ssDNA probes which are immobilized on a surface (called a chip or a slide). The probe sequences are designed and placed on an array in a regular pattern of spots.

- The chip or slide is usually made of glass or nylon and is manufactured using technologies developed for silicon computer chips. Each microarray chip is arranged as a checkerboard of 105 or 106 spots or features, each spot containing millions of copies of a unique DNA probe (often 25 nt long).

- Like Southern & northern blots, microarrays use hybridization to detect a specific DNA or RNA in a sample. But whereas a Southern blot uses a single probe to search a complex DNA mixture, a DNA microarray uses a million different probes, fixed on a solid surface, to probe such a mixture.

- The exact sequence of the probes at each feature/location on the chip is known. Wherever some of the sample DNA hybridizes to the probe in a particular spot, the hybridization can be detected because the target DNA is labeled (and unbound target is washed away).

- Therefore one can determine which of the million different probe sequences are present in the target. {NOTE: In a Southern, the target DNA is immobilized on a membrane; in a microarray, the probes are fixed to the slide or chip. In a Southern, the probe is labeled; in a microarray, the DNA being studied is labeled.} Additionally, the amount of signal directly depends on the quantity of labeled target DNA.

- Thus microarrays can give a quantitative description of how much of a particular sequence is present in the target DNA. This is particularly useful for studying gene expression, one common application of microarray technology. Obviously, microarrays must be read mechanically, using a laser and detector.

- Good software for interpreting the raw data is crucial (as one can imagine a long list of sources of error in reading the individual spots, including nonspecific hybridization and background fluorescence).

- To study gene expression, mRNA is isolated from the cells of interest and converted into labeled cDNA. This cDNA is then washed over a microarray carrying features representing all the genes that could possibly be expressed in those cells.

- If hybridization occurs to a certain feature, it means the gene is expressed. Signal intensity at that feature/spot indicates how strongly the gene is expressed (as it is a sign of how much mRNA was present in the original sample).

- One can therefore study gene expression in an entire cell (not just for one or two genes) under various conditions, over time, or in normal vs. diseased cells. Microarrays are sensitive enough to detect single base differences, mutations, or SNPs (single nucleotide polymorphisms).

- This makes them useful for a wide range of applications, for example: identifying strains of viruses; identifying contamination of food products with cells from other plants or

animals; detecting a panel of mutations in a patient‘s cancer cells that may influence the disease‘s response to treatment. Protein microarrays are also being developed to allow massive screening for interactions between proteins on the microarray, and other proteins, substrates, or ligands.

Principles of microarray analysis

Despite the variety of technical solutions that have been developed for performing microarray analysis, all are miniaturized hybridization assays for studying thousands of nucleic acid fragments simultaneously. All microarray systems (Fig 2) share the following key components:

- the array, which contains immobilized nucleic acid sequences, or _targets‘
- one or more labelled samples or _probes‘, that are hybridized with the microarray
- a detection system that quantitates the hybridization signal

Probe labelling

- The microarray sample that is being analyzed, whether it is mRNA for a gene expression study or DNA derived from genomic analysis, is converted to a labelled population of nucleic acids, the probe.
- These probes frequently consist of several thousands of different labelled nucleic acid fragments. The complexity of microarray hybridization over 10 000 different labelled fragments interrogating up to 100 000 different immobilized sequences is greater than that encountered in other routine molecular biology experiments.
- Therefore, this hybridization should be carried out under conditions that do not promote annealing of non-complementary fragments. Fluorescent dyes, and especially the cyanine dyes Cy3 and Cy5, have been adopted as the predominant label in microarray analysis.
- Fluorescence has the advantage of permitting the detection of two or more different signals in one experiment. This has allowed investigators to perform comparative analysis of two or more samples on one microarray.
- It has also increased the accuracy and throughput of microarray analysis over filter-based macroarrays, in which only one radioactively labelled sample can be conveniently analyzed at a time.

Microarray hybridization

- In microarray hybridization, the labelled fragments in the probe are expected to form duplexes with their immobilized complementary targets.

- This requires that the nucleic acids are single-stranded and accessible to each other. The number of duplexes formed reflects the relative number of each specific fragment in the probe, as long as the amount of immobilized target nucleic acid is in excess and not limiting the kinetics of hybridization.

- Two or more samples labelled with different fluorescent dyes can be hybridized simultaneously, resulting in simultaneous hybridization taking place at each target spot. By measuring the different fluorescent signals associated with each spot, the relative abundance of specific sequences in each of the samples can be determined.

Scanning and data analysis

- Microarray scanners typically contain two different lasers that emit light at wavelengths that are suitable for exciting the fluorescent dyes used as labels.

- A confocal microscope attached to a detector system records the emitted light from each of the microarray spots, allowing high-resolution detection of the hybridization signals.

- Despite their small size, microarrays generate large quantities of data even from a single experiment. As a typical experiment will involve the use of several analyzed samples on replicate arrays, the use of computerized data processing is necessary in order to handle the amount of data generated and to gain maximum information from the experiment.

- This can be achieved by specialized software that extracts primary data from scanned microarray slide images, normalizes this data to remove the influence of experimental variation, and finally manipulates the data so that biologically meaningful conclusions can be made.

Applications of microarray analysis

- The versatility of microarray analysis is confirmed by its rapid emergence as a general molecular biology analytical technique.

- Increasing numbers of researchers within academic institutions and industrial laboratories are now exploiting this technology in diverse biomedical disciplines.

- Microarrays have not become a replacement to established techniques, but more a novel, high-power approach to perform analyses that were previously time consuming.

- By using information derived from the several complete or near complete genome sequences, including the human genome, it is now possible to perform genome-wide experiments using microarray technology. This has already been demonstrated for S. cerevisiae where all the expressed genes are known.

- As microarrays can contain thousands of targets, both characterized and uncharacterized, experiments can be conducted without prior hypotheses.

- This combined with the millions of data points that are possible to analyze in one experiment, microarray analysis has enabled global analysis of biological processes.

- Gene expression analysis, genome analysis, and drug discovery have been three of the main areas in which microarray analysis has been applied so far.

Gene expression analysis

- Gene expression analysis examines the composition of cellular messenger RNA populations. The identity of transcripts that make up these populations and their expression levels are informative of cell state and activity of genes and, as the precursors of translated proteins, changes in mRNA levels are related to changes in the proteome.

Gene expression analysis with microarrays

- A typical microarray gene expression analysis experiment compares the relative expression levels of specific transcripts in two samples. One of these samples is a control and the other is derived from cells whose response or status is being investigated.

- Each of these samples is labelled with a different fluorescent dye, and equal amounts of the labelled samples are combined and hybridized with the microarray.

- The fluorescent signals corresponding to the two dyes are measured independently from each spot after hybridization. After normalization, the intensity of the two hybridization signals can be compared. Equal signal from both samples suggests equal expression in both samples.

- Microarray analysis does not give information about absolute gene expression levels in the samples. This is because the intensity of the fluorescent signals is not only proportional to the number of hybridized fragments but also to the length of these fragments and the number of fluorescent labels each fragment carries, i.e. labelling density.

- As these are determined by the unique nucleotide sequence of each gene and transcript, they will vary from gene to gene. If two samples have been labelled under similar conditions, the length and labelling density of specific transcripts will be similar in the two samples, making it possible to compare the relative abundance of the transcripts in the two samples.

- A strong hybridization signal from microarray analysis does not necessarily correspond to a highly expressed gene; it could be derived, for example, from a gene that is expressed at a relatively low level but yields long, highly-labelled probe fragments. Gene expression analysis with microarrays has been applied to numerous mammalian tissues, plants, yeast, and bacteria alike (1, 5, 6, 7, 8).

- These studies have examined the effects of treating cells with chemicals, the consequences of over-expression of regulatory factors in transfected cells, and compared mutant strains with parental strains to delineate functional pathways. In cancer research microarrays have been used to find gene expression changes in transformed cells and metastases, to identify diagnostic markers, and to classify tumors based on their gene expression profiles (9, 10, 11).

Clustering of gene expression profiles

- It is a widely used approach for finding macroscopic data structure. A complication in such analyses is that not all genes are informative for forming clusters and different clusters might have different transcription regulation.

- Driven by these considerations, we present a novel two-stage clustering approach. The first stage identifies informative genes by adaptive variable selection using pseudo-samples modeled by a high dimensional multigroup ANOVA model.

- Variables are selected using a rescaled spike and slab Bayesian hierarchical model having a special selective shrinkage property. The second stage uses output from the first stage for clustering. We demonstrate why selective shrinkage occurs, and by extension, why it is useful for the clustering paradigm.

- We analyze a human gene atlas expression dataset where the question of interest is to look for tissue-specific transcription regulation and investigate whether tissues can be grouped together due to similar genomic control.

Hierarchical Cluster Analysis

- Hierarchical Cluster Analysis Due to the scarcity of psychological research employing the general technique of cluster analysis, researchers may not fully understand the utility of cluster analysis and the application of the clustering technique to their data.

- There are two main methods: hierarchical and non-hierarchical cluster analysis. Hierarchical clustering combines cases into homogeneous clusters by merging them together one at a time in a series of sequential steps (Blei & Lafferty, 2009).

- Nonhierarchical techniques (e.g., k-means clustering) first establish an initial set of cluster means and then assign each case to the closest cluster mean (Morissette & Chartier, 2013).

- The present paper focuses on hierarchical clustering, though both clustering methods have the same goal of increasing within-group homogeneity and between-groups heterogeneity.

- At each step in the hierarchical procedure, either a new cluster is formed or one case joins a previously grouped cluster. Each step is irreversible meaning that cases cannot be subsequently reassigned to a different cluster.

- This makes the initial clustering steps highly influential because the first clusters generated will be compared to all of the remaining cases.

- The alternate method of non-hierarchical clustering requires the researcher to establish a priori the number of clusters in the final solution. If there is uncertainty about the total number of clusters in the dataset, the analysis must be re-run for each possible solution.

- In this situation, hierarchical clustering is preferred as it inherently allows one to compare the clustering result with an increasing number of clusters; no decision about the final number of clusters needs to be made a priori. Hierarchical cluster analysis can be conceptualized as being agglomerative or divisive.

- Agglomerative hierarchical clustering separates each case into its own individual cluster in the first step so that the initial number of clusters equals the total number of cases

(Norusis, 2010). At successive steps, similar cases–or clusters–are merged together (as described above) until every case is grouped into one single cluster.

- Divisive hierarchical clustering works in the reverse manner with every case starting in one large cluster and gradually being separated into groups of clusters until each case is in an individual cluster. This latter technique, divisive clustering, is rarely utilized because of its heavy computational load (for a discussion on divisive methods, see Wilmink & Uytterschaut, 1984).

- The focus of the present paper is on the method of hierarchical agglomerative cluster analysis and this method is defined by two choices: the measurement of distance between cases and the type of linkage between clusters.

Single-Link, Complete-Link & Average-Link Clustering

- Hierarchical clustering treats each data point as a singleton cluster, and then successively merges clusters until all points have been merged into a single remaining cluster. A hierarchical clustering is often represented as a dendrogram (from Manning et al. 1999).

- In complete-link (or complete linkage) hierarchical clustering, we merge in each step the two clusters whose merger has the smallest diameter (or: the two clusters with the smallest maximum pairwise distance).

- In single-link (or single linkage) hierarchical clustering, we merge in each step the two clusters whose two closest members have the smallest distance (or: the two clusters with the smallest minimumpairwise distance).

- Complete-link clustering can also be described using the concept of clique. Let dn be the diameter of the cluster created in step n of complete-link clustering. Define graph G(n) as the graph that links all data points with a distance of at most dn. Then the clusters after step n are the cliques of G(n). This motivates the term complete-link clustering.

- Single-link clustering can also be described in graph theoretical terms. If dn is the distance of the two clusters merged in step n, and G(n) is the graph that links all data points with a distance of at most dn, then the clusters after step n are the connected components of G(n). A single-link clustering also closely corresponds to a weighted graph's minimum spanning tree.

- Average-link (or group average) clustering (defined below) is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.

Complete-link clustering

- The worst case time complexity of complete-link clustering is at most O(n^2 log n).
- One O(n^2 log n) algorithm is to compute the n^2 distance metric and then sort the distances for each data point (overall time: O(n^2 log n)).
- After each merge iteration, the distance metric can be updated in O(n). We pick the next pair to merge by finding the smallest distance that is still eligible for merging.
- If we do this by traversing the n sorted lists of distances, then, by the end of clustering, we will have done n^2 traversal steps. Adding all this up gives you O(n^2 log n).

Single-link clustering

- The time complexity of single-link clustering is O(n^2). We first compute all distances in O(n^2). While doing this we also find the smallest distance for each data point and keep them in a next-best-merge array.
- In each of the n-1 merging steps we then find the smallest distance in the next-best-merge array. We merge the two identified clusters, and update the distance matrix in O(n). Finally, we update the next-best-merge array in O(n) in each step.
- We can do the latter in O(n) because if the best merge partner for k before merging i and j was either i or j, then after merging i and j the best merge partner for k is the merger of i and j.
- Complete-link clustering is harder than single-link clustering because the last sentence does not hold for complete-link clustering: in complete-link clustering, if the best merge partner for k before merging i and j was either i or j, then after merging i and j the best merge partner for k can be a cluster different from the merger of i and j.
- The reason for this difference between single-link and complete-link is that distance defined as the distance of the two closest members is a local property that is not affected by merging; distance defined as the diameter of a cluster is a non-local property that can change during merging.

Average-link clustering

- Average-link clustering merges in each iteration the pair of clusters with the highest cohesion. If our data points are represented as normalized vectors in a Euclidean space, we can define the cohesion G of a cluster C as the average dot product:

- G(C) = 1/[n(n-1)] (gamma(C)-n)

1. where
   n = !C!,

2. gamma(C) = sum (v in C) sum (w in C) <v,w>

3. <v,w> is the dot product of v and w.

Tools for microarray data analysis

- A typical microarray experiment results in series of images, depending on the experimental design and number of samples. Software analyses the images to obtain the intensity at each spot and quantify the expression for each transcript.
- This is followed by normalization, and then various data analysis techniques are applied on the data. The whole analysis pipeline requires a large number of software to accurately handle the massive amount of data.
- Fortunately, there are large number of freely available and commercial software to churn the massive amount of data to manageable sets of differentially expressed genes, functions, and pathways. This chapter describes the software and tools which can be used to analyze the gene expression data right from the image analysis to gene list, ontology, and pathways.

Madam (Microarray Data Manager)

- It is a suite of tools used to upload,download, and display a plethora of microarray data to and from a database management system (MySql).
-
- Working as an interface for the MySql, Madam allows scientists and researchers to manage their microarray data efficiently to meet the requirement of experiment annotation and data mining.
- Madam implemented in Java, facilitates the entry of data into a relational database. MADAM guides users through the microarray process from RNA procurement to data analysis, offering intelligent forms to simplify the tracking of experimental parameters and results that are essential for the interpretation of expression results in downstream analyses.

- Canned reports provide information on RNA samples, studies, slide maps and other pertinent data and a general SQL query window allows freeform access to the underlying database.

- MADAM also serves as a platform for launching other data entry and management tools. Through the use of these integrated modules, users can view and score PCR plates, design experiments and studies, and track laboratory materials. Although not yet supported, MADAM is being adapted to read and write MAGE-ML, the XML data exchange format being developed by an international consortium of leading public databases and microarray research centers.

- A MAGE-ML version of MADAM should be available by the end of this year and will facilitate submission of microarray data to public repositories such as Array Express and GEO.

Serial Analysis of Gene Expression (SAGE)

- Serial analysis of gene expression (SAGE), a functional genomics technique, can be used for global profiling of gene transcripts.

- It relies on the preparation and sequencing of cDNA concatemers, but it does not require prior knowledge of the genes to be assayed (as with microarrays)

- . Once analyzed, SAGE data provide both a qualitative and quantitative assessment of potentially every transcript present in a particular cell or tissue type.

- In this chapter, we describe the fundamental principles of SAGE, describe a complete protocol for the generation of SAGE libraries, and show how it has been employed to generate the first SAGE reference data set of the mouse myocardium.

- Following the protocols described here, investigators should be able to generate unique mouse heart SAGE libraries, which can be directly compared with our reference library. This permits the identification of transcripts that are differentially expressed as a function of time, age, genetic background or transgenic state, among other factors.

- SAGE is thus a powerful technique that permits a comprehensive analysis of changes in mRNA abundance. The results provide a snapshot of altered patterns of gene expression in response to any genetic or environmental stimulus that can be used to generate new biological hypotheses or test existing paradigms.

Spot finder

- It is a multi-channel image analysis tool.
- This tool was designed for the rapid, reproducible, user-aided analysis of microarray images and the quantification of gene expression. Spotfinder reads paired 16-bit or 8-bit TIFF image files generated by most microarray scanners.
- Semi-automatic grid construction defines the areas of the slide where spots are expected. Automatic and manual grid adjustments help to ensure that each rectangular grid cell is centered on a spot.
- A number of segmentation methods available define the boundaries between each spot and the surrounding local background. Local background correction is available as default choice.
- Reads paired 16-bit or 8-bit TIFF image files generated by most microarray scanners. TIGR Spotfinder is an image-processing software created for analysis of the image files generated in microarray expression studies.
- It uses a fast and reproducible algorithm to identify the spots in the array and provide quantification of expression levels.
- Spotfinder was designed for the rapid, reproducible and computer-aided analysis of microarray images and the quantification of gene expression. Spotfinder is part of TM4 Microarray software.

Applications of microarray

Gene expression analysis

- The predominate application of DNA microarrays has been to measure gene expression levels. In this application, RNA is extracted from the cells of interest and either, labeled directly, converted to a labeled cDNA or converted to a T7 RNA promoter tailed cDNA which is further converted to cRNA through the Eberwine amplification process.
- A wide variety of methods have been developed for labeling of the cDNA or cRNA including: incorporation of fluorescently labeled nucleotides during the synthesis, incorporation of biotin labeled nucleotide which is subsequently stained fluorescently labeled streptavidin, incorporation of a modified reactive nucleotide to which a

fluorescent tag is added later, and a variety of signal amplification methods (an early review of different labeling methods is provided in.

- The two most frequently used methods are the incorporation of fluorescently labeled nucleotides in the cRNA or cDNA synthesis step or the incorporation of a biotin labeled nucleotide in the cRNA synthesis step (as is done by Affymetrix).

Transcription factor binding analysis

- Microarrays have also been used in combination with chromatin immunoprecipitation to determine the binding sites of transcription factors.
- In brief, transcription factors (TFs) are cross linked to DNA with formaldehyde and the DNA is fragmented.
- The TF(s) of interest (with the DNA to which they were boud still attached) are affinity purified using either an antibody to the TF or by tagging the transcription factor with peptide that‘s amenable to affinity chromatography (for example a FLAG-, HIS-, myc or HA-tag). After purification, the DNA is released from the TF, amplified, labeled and hybridized to the array. This technique is commonly referred to as ―ChIP-chip‖ for Chromatin Immuno-Precipitation on a ―chip‖ or microarray.
- As TF‘s often bind quite a distance away from the genes that they regulate, the design of the array and size distribution of the fragment length are interrelated. E.g. the array must contain probes that will interrogate the region of DNA bound to the transcription factor.
- For bacteria or yeast, the intergenic regions are fairly small and the same arrays used for gene expression work can be applied to ChIP-chip. For mammalian genomes, the intergenic regions are large and the TF often bind many kbp away from the gene of interest. Hence, for mammalian genomes, oligo arrays with oligo‘s spaced evenly across the entire genome are typically used for ChIP-chip experiments.

Genotyping

Microarrays have been widely used as single-nucleotide-polymorphism (SNP) genotyping platforms. Several alternative approaches have been used to detect SNP‘s but the most commonly used are allele discrimination by hybridization as used by Affymetri, allele specific extension and ligation to a ―bar-code‖ oligo which is hybridized to a universal array (the Illumina ―Golden Gate Assay‖ or approaches in which the arrayed DNA is extended across the SNP in a single nucleotide extension reaction (the Arrayed Primer Extension assay of Kurg et.al. (Kurg et al., 2000) or the Infinium Assay of Illumina. In order to reduce this background, Affymetrix developed a PCR based approach to reduce genomic complexity.

Medical diagnosis and treatment

The use of microarray in the field of oncology is well known. Apart from this, it is used to study various cardiovascular, inflammatory, and infectious diseases, as well as psychiatric disorders.

The application of microarray in the medical field can be categorized into four types:

a. **The discovery of target:** in this application, the microarray is used to compare diseased tissues/cells with healthy tissues/cells to find the characteristics of a particular disease. This helps in finding the genes responsible for that disease.

b. **The discovery of drugs and leads:** after the target has been discovered, microarrays can be used to screen potential compounds and identify the toxicity of the lead compound that will help in deciding proper medication for the patient. The study of antibodies, as well as microorganisms like bacteria and virus, also helps in the discovery of more effective antibiotics and vaccines.

c. **Diagnostics and prognostics:** the microarray is widely used to know the state of disease, type of tumor and other factors important for the patient. As mentioned above, it is used to diagnose a number of diseases and infections, most notably cancer.

d. **Pharmacogenomics and theranostics:** the microarray technique can be used to decide a patient‘s treatment and therapy on the basis of his/her genetic makeup. Thus, it helps in carrying out personalized treatments than using generalized ones. It can also help in controlling side effects of medications.

Biotechnological and other researches

- Microarray may aid in the identification of new genes, as well as in studying their function and expressions in different conditions.

- It includes the determination of the genetic sequence of all types of organisms such as human beings, mice, as well as microbes. It also helps in conducting studies in the field of agriculture; for example, it can be used for studies related to pest control.

- Furthermore, oligonucleotide microarray technology is used to analyze exposure of toxic occupational biomarkers. This helps in classifying toxins into different categories based on the responses of the biomarkers and study the risk factors in the environment.

- It can be used in understanding the catabolism of xenobiotics, which can enhance bioremediation processes with a direct impact on pollution control and environmental organization.

- With the use of species-specific arrays, scientists may also conduct studies on the evolution of different species. This helps in getting more information about genetic mutation and the idea of common ancestry.

Crime and security

- Microarray has a prominent role in detecting ―Biological Warfare Agents‖ (BWAs):

  BWAs are microorganisms or toxins produced by them that are intentionally dispersed by terrorists to spread diseases in man and other organisms. The microarray provides a platform that provides fast, sensitive, and simultaneous identification of these agents. This is very much useful for national security and protection of life.

- Also, it is important in forensic analysis. SNP microarrays (a type of DNA microarray) are used in forensic analysis to get the details of DNA that will be helpful for investigation purposes. The recent discovery of the abundance of SNPs along with the ease of automation and miniaturization of the detection techniques paved the way for the implementation of microarray in forensic science.

.

Possible Questions

1. Explain the concept of microarrays and give its applications.
2. Explain: SAGE technique and its applications.
3. Explain bacterial and yeast expression systems in detail.
4. Give a detailed note on Genetic markers.
5. What is microarray? Explain about DNA microarray technique with its applications.
6. Discuss about microarray analysis tools- MADAM and Spot finder.
7. Give a detailed note on microarray technique.

8. Explain about the concept of microarrays, normalization and clustering analysis of microarrays.
9. List out the applications of microarray technique.

Genomics and Proteomics (17BTP304)

Unit IV

Experimental Proteomics: Proteome analysis- 2D gel electrophoresis: general strategy, immobilized pH gradients, sample preparation, isoelectric focussing, staining, transfer of proteins from gels, image acquisition and analysis of gels. 2DE databases.

Unit IV – Lecture Plan

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Experimental proteomics	2	**11**	T3: 1-34
2D Gel electrophoresis	3		R2: 449-452
Concept and Gradient system, Sample preparation, Processing and staining	1		R2: 449-452
Image acquisition analysis	1		R2: 449-452
Database for 2D Gel electrophoresis	2		W5
Unit IV Revision	1		
Unit IV Revision	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.
R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.
W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs
W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/ W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html
W5: www.world-2dpage.expasy.org/swiss-2dpage./does/ch2d-details.html
W6: www.ch.embert.org/coursEmnet/PropoA/slides/mass.spectrometry W7: https://schuck/proteininteraction.chapter02.html

Two-dimensional gel electrophoresis

Two-dimensional gel electrophoresis, abbreviated as 2-DE or 2-D electrophoresis, is a form of gel electrophoresis commonly used to analyze proteins. Mixtures of proteins are separated by two properties in two dimensions on 2D gels. 2-DE was first independently introduced by O'Farrell and Klose in 1975.

Basis for separation

2-D electrophoresis begins with electrophoresis in the first dimension and then separates the molecules perpendicularly from the first to create an electropherogram in the second dimension. In electrophoresis in the first dimension, molecules are separated linearly according to their isoelectric point. In the second dimension, the molecules are then separated at 90 degrees from the first electropherogram according to molecular mass. Since it is unlikely that two molecules will be similar in two distinct properties, molecules are more effectively separated in 2-D electrophoresis than in 1-D electrophoresis.

The two dimensions that proteins are separated into using this technique can be isoelectric point, protein complex mass in the native state, and protein mass. Separation of the proteins by isoelectric point is called isoelectric focusing (IEF). Thereby, a gradient of pH is applied to a gel and an electric potential is applied across the gel, making one end more positive than the other. At all pH values other than their isoelectric point, proteins will be charged. If they are positively charged, they will be pulled towards the more negative end of the gel and if they are negatively charged they will be pulled to the more positive end of the gel. The proteins applied in the first dimension will move along the gel and will accumulate at their isoelectric point; that is, the point at which the overall charge on the protein is 0 (a neutral charge).

For the analysis of the functioning of proteins in a cell, the knowledge of their cooperation is essential. Most often proteins act together in complexes to be fully functional. The analysis of this sub organelle organisation of the cell requires techniques conserving the native state of the protein complexes. In native polyacrylamide gel electrophoresis (native PAGE), proteins remain in their native state and are separated in the electric field following their mass and the mass of their complexes respectively. To obtain a separation by size and not by net

charge, as in IEF, an additional charge is transferred to the proteins by the use of Coomassie Brilliant Blue or lithium dodecyl sulfate. After completion of the first dimension the complexes are destroyed by applying the denaturing SDS-PAGE in the second dimension, where the proteins of which the complexes are composed of are separated by their mass.

Before separating the proteins by mass, they are treated with sodium dodecyl sulfate (SDS) along with other reagents (SDS-PAGE in 1-D). This denatures the proteins (that is, it unfolds them into long, straight molecules) and binds a number of SDS molecules roughly proportional to the protein's length. Because a protein's length (when unfolded) is roughly proportional to its mass, this is equivalent to saying that it attaches a number of SDS molecules roughly proportional to the protein's mass. Since the SDS molecules are negatively charged, the result of this is that all of the proteins will have approximately the same mass-to-charge ratio as each other. In addition, proteins will not migrate when they have no charge (a result of the isoelectric focusing step) therefore the coating of the protein in SDS (negatively charged) allows migration of the proteins in the second dimension (SDS-PAGE, it is not compatible for use in the first dimension as it is charged and a nonionic or zwitterionic detergent needs to be used). In the second dimension, an electric potential is again applied, but at a 90 degree angle from the first field. The proteins will be attracted to the more positive side of the gel (because SDS is negatively charged) proportionally to their mass-to-charge ratio. As previously explained, this ratio will be nearly the same for all proteins. The proteins' progress will be slowed by frictional forces. The gel therefore acts like a molecular sieve when the current is applied, separating the proteins on the basis of their molecular weight with larger proteins being retained higher in the gel and smaller proteins being able to pass through the sieve and reach lower regions of the gel.

Detecting proteins

The result of this is a gel with proteins spread out on its surface. These proteins can then be detected by a variety of means, but the most commonly used stains are silver and Coomassie Brilliant Blue staining. In the former case, a silver colloid is applied to the gel. The silver binds to cysteine groups within the protein. The silver is darkened by exposure to ultra-violet light. The

amount of silver can be related to the darkness, and therefore the amount of protein at a given location on the gel. This measurement can only give approximate amounts, but is adequate for most purposes. Silver staining is 100x more sensitive than Coomassie Brilliant Blue with a 40-fold range of linearity.

Molecules other than proteins can be separated by 2D electrophoresis. In supercoiling assays, coiled DNA is separated in the first dimension and denatured by a DNA intercalator (such as ethidium bromide or the less carcinogenic chloroquine) in the second. This is comparable to the combination of native PAGE /SDS-PAGE in protein separation.

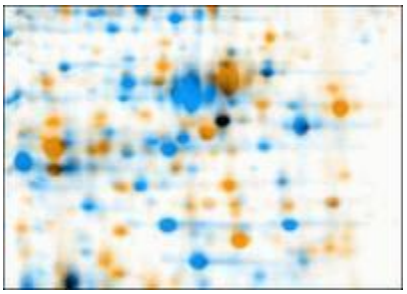
Common techniques

IPG-DALT

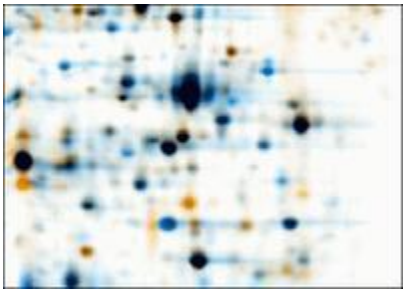
A common technique is to use an Immobilized pH gradient (IPG) in the first dimension. This technique is referred to as **IPG-DALT**. The sample is first separated onto IPG gel (which is commercially available) then the gel is cut into slices for each sample which is then equilibrated in SDS-mercaptoethanol and applied to an SDS-PAGE gel for resolution in the second dimension. Typically IPG-DALT is not used for quantification of proteins due to the loss of low molecular weight components during the transfer to the SDS-PAGE gel.

IEF SDS-PAGE

2D gel analysis software



Warping: Images of two 2D electrophoresis gels, overlaid with Delta2D. First image is colored in orange, second one colored in blue. Due to running differences, corresponding spots do not overlap.



Warping: Images of two 2D electrophoresis gels after warping. First image is colored in orange, second one colored in blue. Corresponding spots overlap after warping. Common spots are colored black, orange spots are only present (or much stronger) on the first image, blue spots are only present (or much stronger) on the second image.

In quantitative proteomics, these tools primarily analyze bio-markers by quantifying individual proteins, and showing the separation between one or more protein "spots" on a scanned image of a 2-DE gel. Additionally, these tools match spots between gels of similar samples to show, for example, proteomic differences between early and advanced stages of an illness. Software packages include Delta2D, ImageMaster, Melanie, PDQuest, Progenesis and REDFIN – among others. [*citation needed*] While this technology is widely utilized, the intelligence has not been perfected. For example, while PDQuest and Progenesis tend to agree on the quantification and analysis of well-defined well-separated protein spots, they deliver different results and analysis tendencies with less-defined less-separated spots. Challenges for automatic software-based analysis include incompletely separated (overlapping) spots (less-defined and/or separated), weak spots / noise (e.g., "ghost spots"), running differences between gels (e.g., protein migrates to different positions on different gels), unmatched/undetected spots, leading to missing values, mismatched spots , errors in quantification (several distinct spots may be erroneously detected as a single spot by the software and/or parts of a spot may be excluded from quantification), and differences in software algorithms and therefore analysis tendencies

Protein Gel Staining Methods

Once protein bands have been separated by electrophoresis, they can be visualized using different methods of in-gel detection, each with particular advantages and disadvantages. Over the past several decades, demand for improved sensitivity for small sample sizes and compatibility with downstream applications and detection instrumentation have driven the development of several basic staining methods. Here we discuss the general principles of protein gel staining and describe several staining methods.

General principles of gel staining

The first step after performing denaturing polyacrylamide gel electrophoresis (SDS-PAGE) is to disassemble the gel cassette and place the thin polyacrylamide gel in a tray filled with water or buffer. The electrophoresed proteins exist as concentrated "bands" embedded within each lane of the porous polyacrylamide gel matrix. Typically, the proteins are still bound to the anionic detergent (SDS), and the entire gel matrix is saturated in running buffer.

To make the proteins visible, a protein-specific, dye-binding or color-producing chemical reaction must be performed on the proteins within the gel. Depending on the particular chemistry of the stain, various steps are necessary to retain, or fix, the proteins in the gel matrix and to facilitate the necessary chemical reaction. All steps are done in solution, i.e., with the gel suspended in a tray filled with one liquid reagent or another.

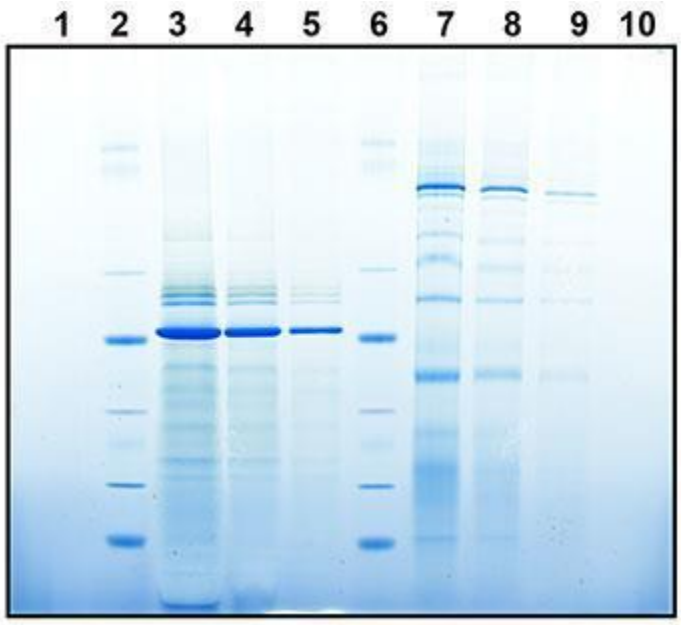
Given the common constraints of this format, most staining methods involve some version of the same general incubation steps:

- A water wash to remove electrophoresis buffers from the gel matrix
- An acid or alcohol wash to condition or fix the gel to limit diffusion of protein bands from the matrix
- Treatment with the staining reagent to allow the dye or chemical to diffuse into the gel and bind to (or react with) the proteins
- Destaining to remove excess dye from the gel matrix background

Depending on the particular staining method, two or more of these functions can be accomplished with one step. For example, a dye reagent that is formulated in an acidic buffer can effectively fix and stain in one step. Conversely, certain functions require several steps. For example, silver staining requires both a staining reagent step and a developer step to produce the colored reaction product.

Coomassie dye stains

The most common method of in-gel protein detection is staining with Coomassie dye. Several recipes for Coomassie staining reagents exist in the literature and use either the G-250 ("colloidal") or R-250 form of the dye. Colloidal Coomassie can be formulated to effectively stain proteins within 1 hour and requires only water (no methanol or acetic acid) for destaining. In acidic conditions, Coomassie dye binds to basic and hydrophobic residues of proteins, changing in color from a dull reddish-brown to intense blue (see previous images on this page). As with all staining methods, Coomassie staining detects some proteins better than others, based on the chemistry of action and differences in protein composition. Thus, Coomassie staining can detect as little as 8–10 ng per band for some proteins and 25 ng per band for most proteins.



Gel staining with Coomassie dye

Two-fold dilutions of protein extracts were run on an Invitrogen™ NativePAGE™ 3–12% Bis-Tris Protein Gel using a Mini Gel Tank. Following electrophoresis, the gel was stained with Coomassie dye and imaged using a flatbed scanner. Lanes 1 and 10: blank; lanes 2 and 6: 5 µL Invitrogen™ NativeMark™ Unstained Protein Standard; lanes 3, 4 and 5: 10, 5, and 2.5 µg spinach chloroplast extract; lanes 7, 8, and 9: 10, 5, and 2.5 µg bovine mitochondrial extract.

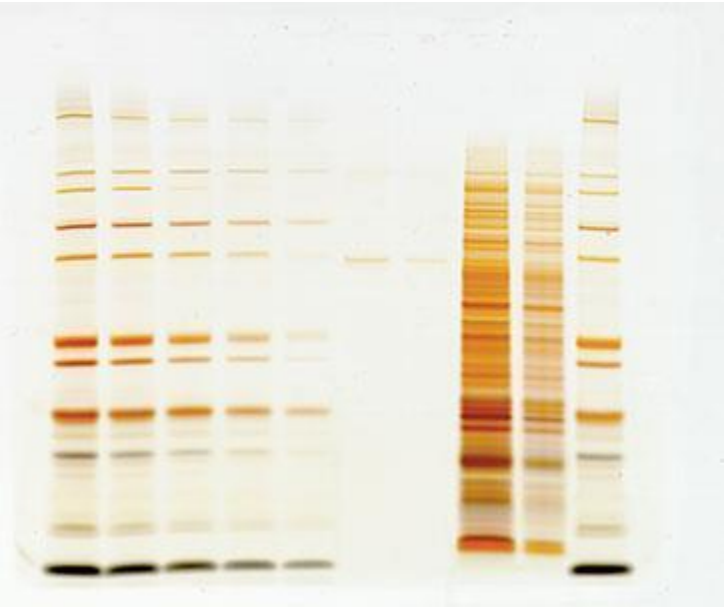
Coomassie dye staining is especially convenient because it involves a single ready-to-use reagent and does not permanently chemically modify the target proteins. An initial water wash step is necessary to remove residual SDS, which interferes with dye binding. Then the staining reagent is added, usually for about 1 hour; finally, a water or simple methanol:acetic acid destaining step is used to wash away excess unbound dye from the gel matrix. Because no chemical modification occurs, excised protein bands can be completely destained and the proteins recovered for analysis by mass spectrometry or sequencing. Coomassie staining and other traditional staining methods require several long incubation and wash steps. To expedite the staining process, more rapid staining protocols have been developed using powered (electrophoretic) devices such as the Thermo Scientific™ Pierce™ Power Stainer.

The Thermo Scientific™ Pierce™ Power Stainer.

This powered device enables rapid (6–11 min) Coomassie dye staining of proteins in polyacrylamide gels, including the removal of unbound stain, in a single step. The small, easy-to-use device consists of the Pierce™ Power Station and Pierce™ Power Stain Cassette, which accommodates up to two mini gels or one midi gel at a time. The staining procedure is designed exclusively for use with Pierce™ Power Staining Kits.

Silver stains

Silver staining is the most sensitive colorimetric method for detecting total protein. The technique involves the deposition of metallic silver onto the surface of a gel at the locations of protein bands. Silver ions (from silver nitrate in the staining reagent) interact and bind with certain protein functional groups. The strongest interactions occur with carboxylic acid groups (Asp and Glu), imidazole (His), sulfhydryls (Cys), and amines (Lys). Various sensitizer and enhancer reagents are essential for controlling the specificity and efficiency of silver ion binding to proteins and effective conversion (development) of the bound silver to metallic silver. The development process is essentially the same as for photographic film: silver ions are reduced to metallic silver, resulting in a brown-black color.

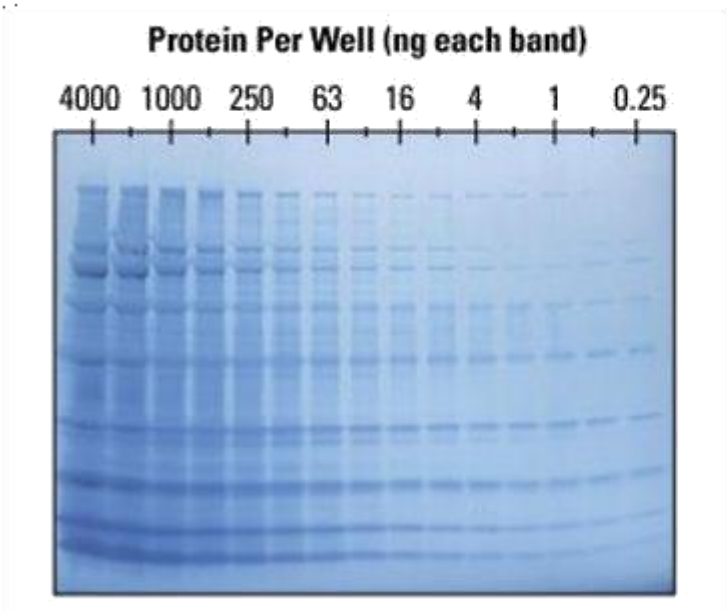


Gel staining with silver stain

Samples were separated on an Invitrogen™ NuPAGE™ 4–12% Bis-Tris Protein Gel and stained with the Invitrogen™ SilverXpress™ Kit. Lanes 1–5: Invitrogen™ Mark12™ Unstained Standard (blend of 12 purified proteins), serial 2-fold dilutions ranging from 1:4 to 1:64; lane 6: 1.6 ng BSA; lane 7: 0.8 ng BSA; lane 8: *E. coli* lysate diluted 1:20; lane 9: *E. coli* lysate diluted 1:80; lane 10: replicate of lane 1. Silver staining protocols require several steps, which are affected by reagent quality as well as incubation times and thickness of the gel. An advantage of commercially available silver staining kits is that the formulations and protocols are optimized and consistently manufactured, helping to maximize consistency of results from experiment to experiment. Kits with optimized protocols are robust and easy to use, detecting less than 0.5 ng of protein in typical gels. Silver stains use either glutaraldehyde or formaldehyde as the enhancer. These reagents can cause chemical crosslinking of the proteins in the gel matrix, limiting compatibility with destaining and elution methods for analysis by mass spectrometry (MS). Therefore, optimization of sensitivity vs. protein recoverability is critical when employing silver staining as part of an MS workflow. Silver stain formulations can be made such that protein bands stain black, blue-brown, red, or yellow, depending on their charge and other characteristics. This is particularly useful for differentiating overlapping spots on 2D gels.

Zinc stains

Zinc staining is unlike all other staining methods. Instead of staining the proteins, this procedure stains all areas of the polyacrylamide gel in which there are no proteins. Zinc ions complex with imidazole, which precipitates in the gel matrix except where SDS-saturated proteins are located. The milky-white precipitate renders the background opaque while the protein bands remain clear. The process is short (about 15 minutes), and the gel can be photographed by viewing it over a dark background. Zinc staining is as sensitive as typical silver staining (detects less than 1 ng of protein), and no fixation steps are required. Furthermore, the stain is easily removed, making this method compatible with MS or western blotting.



Gel staining with zinc stain

A 2-fold dilution series of a protein mixture was separated by protein gel electrophoresis using a 15-well mini gel. Subsequently the gel was stained using the Thermo Scientific™ Pierce™ Zinc Reversible Stain Kit, and then photographed with the gel placed over a dark blue background. The sensitivity on this gel is 0.25 ng, as indicated by the bands that are visible in the last lane.

Fluorescent dye stains

Recent improvements in fluorescence imaging instruments and fluorescent applications have resulted in greater demand for fluorescent stains. A number of fluorescent stains for total protein have been introduced in recent years. Newer fluorescent total-protein stains provide exceptional fluorescent staining performance with fast and easy procedures. The most useful are those whose excitation and emission maxima correspond to common filter sets and laser settings of popular fluorescence imaging instruments.

Protein gel stained with fluorescent dyes

2D gel stained with Invitrogen™ SYPRO™ Ruby protein gel stain and Invitrogen™ Pro-Q™ Emerald 300 reagent. Cohn fractions II and III from cow plasma were combined and resolved on a 2D gel. The gel was first stained with Pro-Q Emerald 300 reagent (left), followed

by staining with the SYPRO Ruby stain (right). Most fluorescent stains involve simple dye-binding mechanisms rather than chemical reactions that alter protein functional groups. Therefore, most are compatible with destaining and protein recovery methods for downstream analysis by MS. Accordingly, these stains are frequently used in both 1D and 2D applications.

Functional group–specific stains

Sometimes it is desirable to detect a subset of proteins rather than all of the proteins in a sample. Glycoproteins and phosphoproteins are classified as such on the basis of a particular chemical moiety (i.e., polysaccharides and phosphate groups, respectively). When a dye-binding or color-producing chemistry can be designed to detect one of these functional groups, it can be used as the basis for a specific gel stain.

Phosphoprotein and total protein visualization in a 2D gel

Protein lysates obtained from a Jurkat T-cell lymphoma line were separated by 2D gel electrophoresis and subsequently stained with Invitrogen™ Pro-Q™ Diamond phosphoprotein gel stain (blue) followed by SYPRO Ruby protein gel stain (red). The gel was dried and imaged on an FLA-3000 scanner (Fuji). Shown is a digitally pseudocolored composite overlaid image.

Proteins that have been post-translationally modified by glycosylation can be detected by a procedure that involves chemical activation of the carbohydrate into a reactive group. The method works by fixing the proteins in the gel and then oxidizing the sugar residues with sodium meta-periodate. The resulting aldehyde groups can then be reacted with an amine-containing dye. In older literature, this method is known as the periodate acid–Schiff (PAS) technique. A subsequent reduction step stabilizes the dye–protein bond. Both colorimetric and fluorescent dyes have been used for this technique, and glycoprotein stain kits are available commercially.

Various protein gel staining methods, both colorimetric and fluorescent, have also been developed to detect phosphorylated proteins and His-tagged fusion proteins. For instance, certain gel stains selectively stain phosphoproteins and His-tags in acrylamide gels, without the need for blotting or phosphoprotein-specific or His-tag–specific antibodies and western blot analysis.

Isoelectric focusing

Isoelectric focusing (IEF), also known as electrofocusing, is a technique for separating different molecules by differences in their isoelectric point (pI). It is a type of zone electrophoresis, usually performed on proteins in a gel, that takes advantage of the fact that overall charge on the molecule of interest is a function of the pH of its surroundings.

IEF involves adding an ampholyte solution into immobilized pH gradient (IPG) gels. IPGs are the acrylamide gel matrix co-polymerized with the pH gradient, which result in completely stable gradients except the most alkaline (>12) pH values. The immobilized pH gradient is obtained by the continuous change in the ratio of *Immobilines*. An Immobiline is a weak acid or base defined by its pK value.

A protein that is in a pH region below its isoelectric point (pI) will be positively charged and so will migrate towards the cathode (negatively charged electrode). As it migrates through a gradient of increasing pH, however, the protein's overall charge will decrease until the protein reaches the pH region that corresponds to its pI. At this point it has no net charge and so migration ceases (as there is no electrical attraction towards either electrode). As a result, the proteins become focused into sharp stationary bands with each protein positioned at a point in the pH gradient corresponding to its pI. The technique is capable of extremely high resolution with proteins differing by a single charge being fractionated into separate bands.

Molecules to be focused are distributed over a medium that has a pH gradient (usually created by aliphatic ampholytes). An electric current is passed through the medium, creating a "positive" anode and "negative" cathode end. Negatively charged molecules migrate through the pH gradient in the medium toward the "positive" end while positively charged molecules move toward the "negative" end. As a particle moves towards the pole opposite of its charge it moves through the changing pH gradient until it reaches a point in which the pH of that molecules isoelectric point is reached. At this point the molecule no longer has a net electric charge (due to the protonation or deprotonation of the associated functional groups) and as such will not proceed any further within the gel. The gradient is established before adding the particles of

interest by first subjecting a solution of small molecules such as polyampholytes with varying pI values to electrophoresis.

The method is applied particularly often in the study of proteins, which separate based on their relative content of acidic and basic residues, whose value is represented by the pI. Proteins are introduced into an Immobilized pH gradient gel composed of polyacrylamide, starch, or agarose where a pH gradient has been established. Gels with large pores are usually used in this process to eliminate any "sieving" effects, or artifacts in the pI caused by differing migration rates for proteins of differing sizes. Isoelectric focusing can resolve proteins that differ in pI value by as little as 0.01. Isoelectric focusing is the first step in two-dimensional gel electrophoresis, in which proteins are first separated by their pI and then further separated by molecular weight through SDS-PAGE.

Living cells

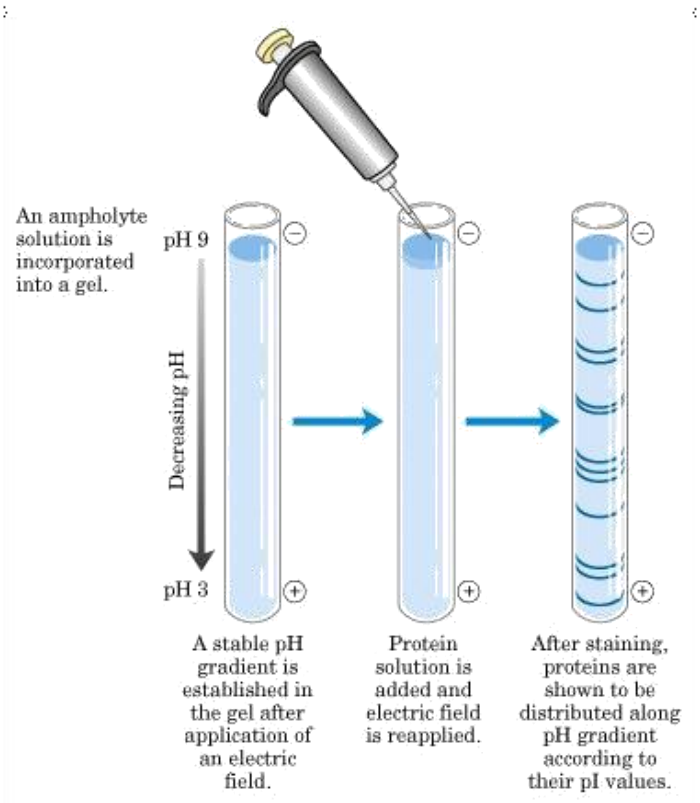
According to some opinions, living eukaryotic cells perform isoelectric focusing of proteins in their interior to overcome a limitation of the rate of metabolic reaction by diffusion of enzymes and their reactants, and to regulate the rate of particular biochemical processes. By concentrating the enzymes of particular metabolic pathways into distinct and small regions of its interior, the cell can increase the rate of particular biochemical pathways by several orders of magnitude. By modification of the isoelectric point (pI) of molecules of an enzyme by, e.g., phosphorylation or dephosphorylation, the cell can transfer molecules of the enzyme between different parts of its interior, to switch on or switch off particular biochemical processes.

Microfluidic chip based

Microchip based electrophoresis is a promising alternative to capillary electrophoresis since it has the potential to provide rapid protein analysis, straightforward integration with other microfluidic unit operations, whole channel detection, nitrocellulose films, smaller sample sizes and lower fabrication costs.

Multi-junction

The increased demand for faster and easy-to-use protein separation tools has accelerate the evolution of IEF towards in-solution separations. In this context, a multi-junction IEF system was developed to perform fast and gel-free IEF separations. The multi-junction IEF system utilizes a series of vessels with a capillary passing through each vessel. Part of the capillary in each vessel is replaced by a semipermeable membrane. The vessels contain buffer solutions with different pH values, so that a pH gradient is effectively established inside the capillary. The buffer solution in each vessel has an electrical contact with a voltage divider connected to a high-voltage power supply, which established electrical field along the capillary. When a sample (a mixture of peptides or proteins) is injected in the capillary, the presence of the electrical field and the pH gradient separates these molecules according to their isoelectric points. The multi-junction IEF system has been used to separate tryptic peptide mixtures for two-dimensional proteomics and blood plasma proteins from Alzheimer‟s disease patients for biomarker discovery.


An ampholyte solution is incorporated into a gel.
pH 9
Decreasing pH
pH 3
A stable pH gradient is established in the gel after application of an electric field.
Protein solution is added and electric field is reapplied.
After staining, proteins are shown to be distributed along pH gradient according to their pI values.

SWISS-2DPAGE contains data on proteins identified on various 2-D PAGE and SDS-PAGE reference maps. You can locate these proteins on the 2-D PAGE maps or display the region of a 2-D PAGE map where one might expect to find a protein from UniProtKB/Swiss-Prot [More details / References / Linking to SWISS-2DPAGE / Commercial users.

Access to SWISS-2DPAGE

- + **[How to use this interface]**
http://world-2dpage.expasy.org/swiss-2dpage?de • by description (any word in the ID, DE,

GN and KW lines)
http://world-2dpage.expasy.org/swiss-2dpage?ac • by accession number (AC lines)
http://world-2dpage.expasy.org/swiss-2dpage/viewer • by clicking on a spot: select one of our

2-D PAGE or SDS-PAGE reference maps, click on a spot and then get the corresponding information from the SWISS-2DPAGE database.

- by author (RA lines)
http://world-2dpage.expasy.org/swiss-2dpage?spot • by spot serial number (2D and 1D lines)
- by experimental pI/Mw range
http://world-2dpage.expasy.org/swiss-2dpage?ident • by experimental identification methods
- by full text search
- retrieve all the protein entries identified on a given reference map
- user defined / complex queries (SRS like)

SWISS-2DPAGE documents

- + **[Facts and statistics]**

- User manual
http://world-2dpage.expasy.org/swiss-2dpage/docs/relnotes.html • Release notes (September 26, 2006)

- FAQ (Frequently Asked Questions about SWISS-2DPAGE)
- **Protocols:**
  - Technical information about 2-D PAGE (IPG's, silver staining, protocols, etc)
- **Figure captions of SWISS-2DPAGE maps available from publications:**
  - Human CSF, ELC, HEPG2, HEPG2SP, LIVER, LYMPHOMA, PLASMA, PLATELET, RBC, U937, CEC, KIDNEY.
  - *Dictyostelium discoideum*, *Escherichia coli*, *Saccharomyces cerevisiae*.

Services

- ftp://ftp.expasy.org/databases/swiss-2dpage • Downloading SWISS-2DPAGE by FTP
- Proteomics Core Facility at University

Software

- ImageMaster / Melanie - Software package for 2-D PAGE analysis

of Geneva - Get your 2-D Gels performed according to Swiss standards

- Make2D-DB package ver. 3.10 new - A package preparing the data and the programs necessary to build a federated 2-DE database on one's own web site.

Gateways to other 2-D PAGE related servers and services

- The World-2DPAGE Constellation - Promote and publish gel-based proteomics data through the ExPASy server
  - WORLD-2DPAGE List – Index to other Federated 2-D PAGE databases
  - World-2DPAGE Portal - A dynamic portal to query simultaneously world-wide gel-based proteomics databases
  - World-2DPAGE Repository - A public standards-compliant repository for gel-based proteomics data published in the literature

The **World-2DPAGE Constellation** is composed of many 2D-PAGE resources and tools:

- **SWISS-2DPAGE** -- The world famous reference database providing extensive annotated data for over 4000 identified spots, 1265 proteins, covering 36 reference maps of 7 different species. (*Last update: November 2011*).
- **World-2DPAGE Repository** -- A public standards-compliant repository for gel-based Proteomics data linked to protein identification published in the literature. The repository is continuously extended, puting together a large collection of multi-species reference maps, with thousands of identified spots.
- **World-2DPAGE Portal** -- A dynamic portal to query simultaneously world-wide gel-based Proteomics databases.

  This portal can be seen as a virtual unique database accessing a very large collection of Proteomics data provided by many renowned institutions (*Last update: December 2011*).
- **World-2DPAGE List** -- A list of database and services grouped by species and classified in categories according to their implementation of the rules defining a federated

2-DE database. It currently lists up to 60 databases, totalizing nearly 400 gel images (*Last update: September 2011*).

- **Make2D-DB package ver. 3.10.2** -- The open source tools package to build, acces and integrate federated 2D-PAGE databases, portals or data repositories on one's own web site.
- **Melanie Viewer ver. 7.0** -- The Melanie Viewer is made available free of charge to help researchers visualize gels and related data obtained through the use of the full version of Melanie.
- **MIAPEGelDB** -- Produce MIAPE-compliant documents describing your gel experiments.

- WORLD-2DPAGE List
Index to 2-D PAGE databases and services

- This page contains references to known 2-D PAGE database servers, as well as to 2-D PAGE related servers and services. You may also find related information in the Proteomics Links page of the Swiss Proteomics Society.
- For a reference to federated 2-D PAGE database, see Appel et Al., Electrophoresis 17, 1996, 540-546, 1996.
    - [2-DE databases][Services & Software][Journals & Societies][Meetings]

2-DE databases

Guidelines for building a federated 2-DE database
Add or remove a link from this list

Database servers: federated partially federated other

How to link to specific entry of these databases

Multi species	
World-2DPAGE Portal	A dynamic portal to query simultaneously World-Wide gel-based proteomics databases

World-2DPAGE Repository			A public standards-compliant repository for gel-based proteomics data published in the literature
SWISS-2DPAGE	SIB Swiss Institute of Bioinformatics	Human	Liver, Plasma, HepG2, Red blood cells, Lymphoma, HepG2 Secreted Proteins, Cerebrospinal Fluid, Macrophage Like Cell Line (U937), Erythroleukemia Cell, Platelet, Kidney, Promyelocytic Leukemia Cells (HL60), Colorectal epithelia cells, Colorectal adenocarcinoma cell line (DL-1), Soluble nuclear proteins and matrix from liver tissue, Nucleoli from Hela cells (1D and 2D), Lymphocytes
		Mus musculus	Liver, Gastrocnemius muscle, Pancreatic islet cells, Brown adipose tissue, White adipose tissue, Soluble nuclear proteins and matrix from liver tissue
SIENA-2DPAGE	Department of Molecular Biology, University of Siena, Italy	Human	Breast ductal carcinoma and histologically normal tissue, Amniotic fluid

UCD 2D-PAGE	The UCD Conway Institute Proteome Research Centre (CIPRC), University College Dublin, Ireland	Human	Brain cortex, Heart ventricle, Endothelial cell, Neutrophil,
		Dog	Heart ventricle
2-DE database	Max Planck Institute for Infection Biology, Berlin, Germany	Human	Jurkat T-cells, synovial fluid exosomes
		Mouse	Mammary gland, urea soluble lens proteins
		Rat	liver and muscle 20S proteasome subtype II & IV
			Bacillus amyloliquefaciens, Bacillus anthracis, Bartonella henselae, Borrelia garinii, Chlamydia pneumoniae, Chlamydophila pneumoniae, Francisella tularensis, Helicobacter pylori, Leishmania mexicana, Mycobacterium bovis, Mycobacterium tuberculosis, Mycoplasma pneumoniae, Paracoccus denitrificans, Vibrio cholerae
Mammalia			
DOSAC-COBS 2D-PAGE	University of Palermo, Italy	Human	various breast cancer samples (cells, serum, etc.), dermal fibroblasts, mammary cells, lymphocytes, U937 and HL60 cell lines

RAT HEART-2DPAGE	German Heart Institute, Berlin, Germany	Rat	Heart
HEART-2DPAGE	German Heart Institute, Berlin, Germany	Human	Heart (ventricle, atrium)
HSC-2DPAGE	Heart Science Centre, Harefield Hospital, UK	Human Rat, Dog	Heart (ventricle), endothelial cells Heart (ventricle)
HP-2DPAGE - Heart 2-DE Database	MDC & MPIIB, Berlin, Germany	Human	Heart (ventricule)
USC-OGP 2-DE database	University of Santiago de Compostela (USC), ES	Human	Platelet (various pI ranges), Uveal melanoma
REPRODUCTION-	Nanjing Medical	Human	Ovary, Testis

2DPAGE	University, P. R. China	Mouse	Testis
Washington University Inner Ear Protein Database	Washington University, USA	Human	Perilymph
		Guinea pig	Endolymph, organ of Corti, otoconia, tectorial membrane
Toothprint - 2-DE gels of dental tissues	University of Otago, Dunedin, New Zealand	Rat	Enamel
JPLS Proteomics database	Ludwig Institute for Cancer Research, Melbourne, Australia	Human	Breast carcinoma cell line, Placenta, Colorectal cell line
		Mouse	colonic crypts
2-DE Map of Cerebrospinal Fluid Proteome	Lee Research Group at Cornell University, USA	Human	Cerebrospinal Fluid
2-DE maps	Proteomics and	Cow	Sera (during dioestrus), positive acute-

	Protein Structure Study Group, at Milano University, Italy		phase reactants detected in serum from a cow with udder edema (inset C)
		Rat	Serum (control Sprague-Dawley (CD) female and male rats, turpentine-treated rats, diabetic rats, stroke-prone rats), urine (control female and male rats, stroke-prone rats), cerebrospinal fluid (control and stroke-prone rats)
		Mouse	Serum
HUVEC (Proteomics of Human Umbilical Vein Endothelial Cell)	Hopital Saint-Antoine, Paris, France	Human	Umbilical vein endothelial cell
Tooth Pulp 2-D	Proteomics Core Facility of the Biocenter Oulu, University of Oulu, Finland	Human	Tooth pulp tissue (healthy and carious)
Vascular Cell database	Department of Cardiac & Vascular Sciences, University of London, UK	Mouse	aortic smooth muscle cells, Sca-1+ progenitor cells, Embryonic Stem Cells

LECB 2D PAGE Gel Image Data Sets	CCR Nanobiology Program, USA	Human	leukemias (AML, ALL, CLL, HCL and other), HL-60 cell line, Molt-4 cell line, serum (Fetal Alchohol Syndrome, Rett Syndrome), Urine (cadmium-exposure, mercury-exposure) - 500 gel images

Plant

- Proteome database - Tabacco BY-2 cell suspension culture, Musa meristem, Arabidopsis root at University of Antwerp, Belgium
- Rice Proteome Database at National Institute of Agrobiological Sciences, Japan
- Arabidopsis 2D gel at Institut des Sciences du Végétal, CNRS - Gif-sur-Yvette, France
- Arabidopsis Seed Proteome at INRA, Lyon, France
- GABI primary database - 2-DE database for 8 different *Arabidospis thaliana* tissues (inflorescence, stem, primary leaf, leaf, seed, seedling, silique, root) of the Genomanalyse im biologischen System Pflanze, Germany
- NASC Proteomics database for Arabidopsis data at the Nottingham Arabidopsis Stock Centre, University of Nottingham, UK
- Plant Proteomics database (PROTICdb) at UMR de Genetique Vegetale du Moulon, Gif-sur-Yvette, France
- 2-D gels for maritime pine at INRA, Bordeaux, France
- 2-D PAGE of Medicago truncatula and other plants at The Samuel Roberts Noble Foundation, Ardmore, USA
- The GelMap Arabidopsis mitochondrial proteome project at Institute for Plant Genetics, Leibniz Universität Hannover, Germany
- Mt Proteomics - the symbiotic proteomics of Medicago truncatula roots at INRA, Epoisses, France

Yeast

- COMPLUYEAST-2DPAGE - 2-DE database (*Candida albicans*, *Saccharomyces cerevisiae*) at Universidad complutense Madrid, Spain
- YMP - Yeast Mitochondrial Proteome at Proteomics Core Facility of the Biocenter Oulu, University of Oulu, Finland
- YPM - Yeast Protein Map at Institut de Biochimie et Génétique Cellulaires, Bordeaux, France

Possible questions

Objective questions

1. Electrophoretic mobility is directly proportional to
(a) Field strength
(b) Molecular weight
(c) Molecular structures
(d) Solvent viscosity

2. Electrophoretic mobility is inversely proportional to
(a) Electrostatic potential
(b) Overall charge of proteins
(c) Molecular structures
(d) Field strength

3. The pH at which a protein assume zero net charge is its
(a) pK_1
(b) pK_2
(c) Pk_R
(d) pI

4. Separation of protein molecules in SDS-PAGE is on the basis of
(a) Shapes.
(b) Size
(c) Charge
(d) Shape, size and Charge.

5. Separation of protein molecules in IEF is on the basis of
(a) Shapes.
(b) Size
(c) Charge
(d) Shape, size and Charge.

6. How does acrylamide affect pore sizes of SDS-PAGE?
(a) When acrylamide is less, pore sizes are bigger.
(b) When acrylamide is less, pore sizes are smaller.
(c) When acrylamide is more, pore sizes are bigger.
(d) When acrylamide is more, pore sizes are smaller.

7. What are ampholytes?
(a) Poly amino and poly carboxylic compounds.
(b) Poly amino compounds.
(c) Poly carboxylic compounds.
(d) None of the above.

8. The SDS-PAGE is very useful to determine
(a) Approximate molecular weight of a protein.
(b) Structural stability of a protein.
(c) Overall shape of a protein.
(d) Overall net charge of a protein

2 Marks Questions

1. Write any two unique features of ampholytes.
2. What are the components in the SDS-PAGE „loading buffer".
3. What is isotachophoresis?
4. How will you optimize the pore size in PAG?
5. Why is BSA used in „Western blotting" after baking process?

6/8/10 Marks questions

1. Explain an electrophoresis method to identify proteins from a mixture containing five molecules which are similar in their size and shape but differing from each other in their pI values (3.5, 5.5, 7.0, 9.4 and 11.2).
2. Write on essay on principles, experimental set-up and applications of Isoelectric focussing electrophoresis technique.

3. Write on essay on principles, experimental set-up and applications of SDS-PAGE

Genomics and Proteomics (17BTP304)

Unit V

Analytical Proteomics: RP-HPLC, Mass Spectrometry – ESI MS and MALDI techniques and applications. Characterization of protein complexes – Protein - DNA, Protein-protein interactions, yeast two-hybrid system and protein micro-arrays – biomarkers.

Unit V – Lecture Plan

Topic to be covered	Duration (Hours)	Total hours	Books: Pages
Analytical proteomics RP-HPLC	1		W6
ESI-MS	2		R2: 454 - 465
MALDI-MS	2		T3: 1-74
Characterization of protein complexes	1		W7
Protein – protein interaction	1		W7
Protein –DNA interaction	1	**15**	W7
Yeast 2 Hybrid system	1		R2 481-487
Protein microarrays, Biomarkers	2		R1 615-616
Unit V Class test	1		
ESE QP discussion	1		
ESE QP discussion	1		
ESE QP discussion	1		

References

T1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
T2: Genomes (John Wiley & Sons, 2002) – Brown, T.A.
T3: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.
R1: Molecular Biology (W.H.Freeman & Company, New York, 5th Edition, 2000) – Berk, LHA
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.
W1: www.cs.columbia.edu/4761/notes07/chapter5.2/microarray/pdfs
W2: www.chagall.med.cornell.edu/12MT/MA-tools.pdf/ W3: www.tms.org/madam.html
W4: www.tms.org/spotfinder.html
W5: www.world-2dpage.expasy.org/swiss-2dpage./does/ch2d-details.html
W6: www.ch.embert.org/coursEmnet/PropoA/slides/mass.spectrometry W7: https://schuck/proteininteraction.chapter02.html

RP-HPLC

High-performance liquid chromatography (HPLC; formerly referred to as high-pressure liquid chromatography), is a technique in analytical chemistry used to separate, identify, and quantify each component in a mixture. It relies on pumps to pass a pressurized liquid solvent containing the sample mixture through a column filled with a solid adsorbent material. Each component in the sample interacts slightly differently with the adsorbent material, causing different flow rates for the different components and leading to the separation of the components as they flow out the column.

HPLC has been used for manufacturing (*e.g.* during the production process of pharmaceutical and biological products), legal (*e.g.* detecting performance enhancement drugs in urine), research (*e.g.* separating the components of a complex biological sample, or of similar synthetic chemicals from each other), and medical (*e.g.* detecting vitamin D levels in blood serum) purposes.

Chromatography can be described as a mass transfer process involving adsorption. HPLC relies on pumps to pass a pressurized liquid and a sample mixture through a column filled with adsorbent, leading to the separation of the sample components. The active component of the column, the adsorbent, is typically a granular material made of solid particles (*e.g.* silica, polymers, etc.), 2–50 μm in size. The components of the sample mixture are separated from each other due to their different degrees of interaction with the adsorbent particles. The pressurized liquid is typically a mixture of solvents (*e.g.* water, acetonitrile and/or methanol) and is referred to as a "mobile phase". Its composition and temperature play a major role in the separation process by influencing the interactions taking place between sample components and adsorbent. These interactions are physical in nature, such as hydrophobic (dispersive), dipole–dipole and ionic, most often a combination.

HPLC is distinguished from traditional ("low pressure") liquid chromatography because operational pressures are significantly higher (50–350 bar), while ordinary liquid chromatography typically relies on the force of gravity to pass the mobile phase through the column. Due to the small sample amount separated in analytical HPLC, typical column

dimensions are 2.1–4.6 mm diameter, and 30–250 mm length. Also HPLC columns are made with smaller sorbent particles (2–50 μm in average particle size). This gives HPLC superior resolving power (the ability to distinguish between compounds) when separating mixtures, which makes it a popular chromatographic technique.

The schematic of an HPLC instrument typically includes a degasser, sampler, pumps, and a detector. The sampler brings the sample mixture into the mobile phase stream which carries it into the column. The pumps deliver the desired flow and composition of the mobile phase through the column. The detector generates a signal proportional to the amount of sample component emerging from the column, hence allowing for quantitative analysis of the sample components. A digital microprocessor and user software control the HPLC instrument and provide data analysis. Some models of mechanical pumps in a HPLC instrument can mix multiple solvents together in ratios changing in time, generating a composition gradient in the mobile phase. Various detectors are in common use, such as UV/Vis, photodiode array (PDA) or based on mass spectrometry. Most HPLC instruments also have a column oven that allows for adjusting the temperature at which the separation is performed.

The sample mixture to be separated and analyzed is introduced, in a discrete small volume (typically microliters), into the stream of mobile phase percolating through the column. The components of the sample move through the column at different velocities, which are a function of specific physical interactions with the adsorbent (also called stationary phase). The velocity of each component depends on its chemical nature, on the nature of the stationary phase (column) and on the composition of the mobile phase. The time at which a specific analyte elutes (emerges from the column) is called its retention time. The retention time measured under particular conditions is an identifying characteristic of a given analyte.

Many different types of columns are available, filled with adsorbents varying in particle size, and in the nature of their surface ("surface chemistry"). The use of smaller particle size packing materials requires the use of higher operational pressure ("backpressure") and typically improves chromatographic resolution (*i.e.* the degree of separation between consecutive analytes emerging from the column). Sorbent particles may be hydrophobic or polar in nature.

Common mobile phases used include any miscible combination of water with various organic solvents (the most common are acetonitrile and methanol). Some HPLC techniques use water-free mobile phases (see Normal-phase chromatography below). The aqueous component of the mobile phase may contain acids (such as formic, phosphoric or trifluoroacetic acid) or salts to assist in the separation of the sample components. The composition of the mobile phase may be kept constant ("isocratic elution mode") or varied ("gradient elution mode") during the chromatographic analysis. Isocratic elution is typically effective in the separation of sample components that are very different in their affinity for the stationary phase. In gradient elution the composition of the mobile phase is varied typically from low to high eluting strength. The eluting strength of the mobile phase is reflected by analyte retention times with high eluting strength producing fast elution (=short retention times). A typical gradient profile in reversed phase chromatography might start at 5% acetonitrile (in water or aqueous buffer) and progress linearly to 95% acetonitrile over 5–25 minutes. Periods of constant mobile phase composition may be part of any gradient profile. For example, the mobile phase composition may be kept constant at 5% acetonitrile for 1–3 min, followed by a linear change up to 95% acetonitrile.

A rotary fraction collector collecting HPLC output. The system is being used to isolate a fraction containing Complex I from *E. coli* plasma membranes. About 50 litres of bacteria were needed to isolate this amount. The chosen composition of the mobile phase (also called eluent) depends on the intensity of interactions between various sample components ("analytes") and stationary phase (*e.g.* hydrophobic interactions in reversed-phase HPLC). Depending on their affinity for the stationary and mobile phases analytes partition between the two during the separation process taking place in the column. This partitioning process is similar to that which occurs during a liquid–liquid extraction but is continuous, not step-wise. In this example, using a water/acetonitrile gradient, more hydrophobic components will elute (come off the column) late, once the mobile phase gets more concentrated in acetonitrile (*i.e.* in a mobile phase of higher eluting strength). The choice of mobile phase components, additives (such as salts or acids) and gradient conditions depends on the nature of the column and sample components. Often a series of trial runs is performed with the sample in order to find the HPLC method which gives adequate separation.

Reversed phase HPLC (RP-HPLC) has a non-polar stationary phase and an aqueous, moderately polar mobile phase. One common stationary phase is a silica which has been surface-modified with RMe_2SiCl, where R is a straight chain alkyl group such as $C_{18}H_{37}$ or C_8H_{17}. With such stationary phases, retention time is longer for molecules which are less polar, while polar molecules elute more readily (early in the analysis). An investigator can increase retention times by adding more water to the mobile phase; thereby making the affinity of the hydrophobic analyte for the hydrophobic stationary phase stronger relative to the now more hydrophilic mobile phase. Similarly, an investigator can decrease retention time by adding more organic solvent to the eluent. RP-HPLC is so commonly used that it is often incorrectly referred to as "HPLC" without further specification. The pharmaceutical industry regularly employs RP-HPLC to qualify drugs before their release.

RP-HPLC operates on the principle of hydrophobic interactions, which originates from the high symmetry in the dipolar water structure and plays the most important role in all processes in life science. RP-HPLC allows the measurement of these interactive forces. The binding of the analyte to the stationary phase is proportional to the contact surface area around the non-polar segment of the analyte molecule upon association with the ligand on the stationary phase. This solvophobic effect is dominated by the force of water for "cavity-reduction" around the analyte and the C_{18}-chain versus the complex of both. The energy released in this process is proportional to the surface tension of the eluent (water: 7.3×10^{-6} J/cm², methanol: 2.2×10^{-6} J/cm²) and to the hydrophobic surface of the analyte and the ligand respectively. The retention can be decreased by adding a less polar solvent (methanol, acetonitrile) into the mobile phase to reduce the surface tension of water. Gradient elution uses this effect by automatically reducing the polarity and the surface tension of the aqueous mobile phase during the course of the analysis.

Structural properties of the analyte molecule play an important role in its retention characteristics. In general, an analyte with a larger hydrophobic surface area (C–H, C–C, and generally non-polar atomic bonds, such as S-S and others) is retained longer because it is non-interacting with the water structure. On the other hand, analytes with higher polar surface area (conferred by the presence of polar groups, such as -OH, $-NH_2$, COO^- or $-NH_3^+$ in their structure) are less retained as they are better integrated into water. Such interactions are subject to steric

effects in that very large molecules may have only restricted access to the pores of the stationary phase, where the interactions with surface ligands (alkyl chains) take place. Such surface hindrance typically results in less retention. Retention time increases with hydrophobic (non-polar) surface area. Branched chain compounds elute more rapidly than their corresponding linear isomers because the overall surface area is decreased. Similarly organic compounds with single C–C bonds elute later than those with a C=C or C–C triple bond, as the double or triple bond is shorter than a single C–C bond. Aside from mobile phase surface tension (organizational strength in eluent structure), other mobile phase modifiers can affect analyte retention. For example, the addition of inorganic salts causes a moderate linear increase in the surface tension of aqueous solutions (ca. 1.5×10^{-7} J/cm² per Mol for NaCl, 2.5×10^{-7} J/cm² per Mol for $(NH_4)_2SO_4$), and because the entropy of the analyte-solvent interface is controlled by surface tension, the addition of salts tend to increase the retention time. This technique is used for mild separation and recovery of proteins and protection of their biological activity in protein analysis (hydrophobic interaction chromatography, HIC).

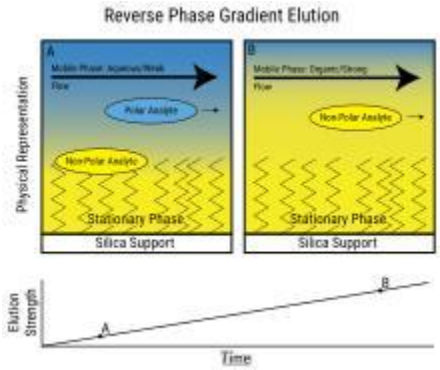
Another important factor is the mobile phase pH since it can change the hydrophobic character of the analyte. For this reason most methods use a buffering agent, such as sodium phosphate, to control the pH. Buffers serve multiple purposes: control of pH, neutralize the charge on the silica surface of the stationary phase and act as ion pairing agents to neutralize analyte charge. Ammonium formate is commonly added in mass spectrometry to improve detection of certain analytes by the formation of analyte-ammonium adducts. A volatile organic acid such as acetic acid, or most commonly formic acid, is often added to the mobile phase if mass spectrometry is used to analyze the column effluent. Trifluoroacetic acid is used infrequently in mass spectrometry applications due to its persistence in the detector and solvent delivery system, but can be effective in improving retention of analytes such as carboxylic acids in applications utilizing other detectors, as it is a fairly strong organic acid. The effects of acids and buffers vary by application but generally improve chromatographic resolution.

Reversed phase columns are quite difficult to damage compared with normal silica columns; however, many reversed phase columns consist of alkyl derivatized silica particles and should never be used with aqueous bases as these will destroy the underlying silica particle. They

can be used with aqueous acid, but the column should not be exposed to the acid for too long, as it can corrode the metal parts of the HPLC equipment. RP-HPLC columns should be flushed with clean solvent after use to remove residual acids or buffers, and stored in an appropriate composition of solvent. The metal content of HPLC columns must be kept low if the best possible ability to separate substances is to be retained. A good test for the metal content of a column is to inject a sample which is a mixture of 2,2'- and 4,4'- bipyridine. Because the 2,2'-bipy can chelate the metal, the shape of the peak for the 2,2'-bipy will be distorted (tailed) when metal ions are present on the surface of the silica.

A separation in which the mobile phase composition remains constant throughout the procedure is termed *isocratic* (meaning *constant composition*). The word was coined by Csaba Horvath who was one of the pioneers of HPLC. The mobile phase composition does not have to remain constant. A separation in which the mobile phase composition is changed during the separation process is described as a *gradient elution* .[12] One example is a gradient starting at 10% methanol and ending at 90% methanol after 20 minutes. The two components of the mobile phase are typically termed "A" and "B"; *A* is the "weak" solvent which allows the solute to elute only slowly, while *B* is the "strong" solvent which rapidly elutes the solutes from the column. In reversed-phase chromatography, solvent *A* is often water or an aqueous buffer, while *B* is an organic solvent miscible with water, such as acetonitrile, methanol, THF, or isopropanol.

In isocratic elution, peak width increases with retention time linearly according to the equation for N, the number of theoretical plates. This leads to the disadvantage that late-eluting peaks get very flat and broad. Their shape and width may keep them from being recognized as peaks.



A schematic of gradient elution. Increasing mobile phase strength sequentially elutes analytes having varying interaction strength with the stationary phase. Gradient elution decreases the retention of the later-eluting components so that they elute faster, giving narrower (and taller) peaks for most components. This also improves the peak shape for tailed peaks, as the increasing concentration of the organic eluent pushes the tailing part of a peak forward. This also increases the peak height (the peak looks "sharper"), which is important in trace analysis. The gradient program may include sudden "step" increases in the percentage of the organic component, or different slopes at different times – all according to the desire for optimum separation in minimum time. In isocratic elution, the selectivity does not change if the column dimensions (length and inner diameter) change – that is, the peaks elute in the same order. In gradient elution, the elution order may change as the dimensions or flow rate change. The driving force in reversed phase chromatography originates in the high order of the water structure. The role of the *organic component of the mobile phase* is to reduce this high order and thus *reduce the retarding strength of the aqueous component.*

Applications

As briefly mentioned, HPLC has many applications in both laboratory and clinical science. It is a common technique used in pharmaceutical development, as it is a dependable way to obtain and ensure product purity. While HPLC can produce extremely high quality (pure) products, it is not always the primary method used in the production of bulk drug materials. According to the European pharmacopoeia, HPLC is used in only 15.5% of syntheses. However, it plays a role in 44% of syntheses in the United States pharmacopoeia. This could possibly be due to differences in monetary and time constraints, as HPLC on a large scale can be an expensive technique. An increase in specificity, precision, and accuracy that occurs with HPLC unfortunately corresponds to an increase in cost. This technique is also used for detection of illicit drugs in urine. The most common method of drug detection is an immunoassay. This method is much more convenient. However, convenience comes at the cost of specificity and coverage of a wide range of drugs. As HPLC is a method of determining (and possibly increasing) purity, using HPLC alone in evaluating concentrations of drugs is somewhat insufficient. With this, HPLC in this context is often performed in conjunction with mass

spectrometry. Using liquid chromatography instead of gas chromatography in conjunction with MS circumvents the necessity for derivitizing with acetylating or alkylation agents, which can be a burdensome extra step. This technique has been used to detect a variety of agents like doping agents, drug metabolites, glucuronide conjugates, amphetamines, opioids, cocaine, BZDs, ketamine, LSD, cannabis, and pesticides Performing HPLC in conjunction with Mass spectrometry reduces the absolute need for standardizing HPLC experimental runs.

Research

Similar assays can be performed for research purposes, detecting concentrations of potential clinical candidates like anti-fungal and asthma drugs. This technique is obviously useful in observing multiple species in collected samples, as well, but requires the use of standard solutions when information about species identity is sought out. It is used as a method to confirm results of synthesis reactions, as purity is essential in this type of research. However, mass spectrometry is still the more reliable way to identify species.

Medical

Medical use of HPLC can include drug analysis, but falls more closely under the category of nutrient analysis. While urine is the most common medium for analyzing drug concentrations, blood serum is the sample collected for most medical analyses with HPLC. Other methods of detection of molecules that are useful for clinical studies have been tested against HPLC, namely immunoassays. In one example of this, competitive protein binding assays (CPBA) and HPLC were compared for sensitivity in detection of vitamin D. Useful for diagnosing vitamin D deficiencies in children, it was found that sensitivity and specificity of this CPBA reached only 40% and 60%, respectively, of the capacity of HPLC. While an expensive tool, the accuracy of HPLC is nearly unparalleled.

Mass Spectrometry

Mass spectrometry (MS) is an analytical technique that measures the mass-to-charge ratio of charged particles. It is used for determining masses of particles, for determining the elemental composition of a sample or molecule. The MS principle consists of ionizing chemical compounds to generate charged molecules or molecule fragments and measurement of their mass-to-charge ratios by using the one of a variety of techniques (e.g EI/CI/ESI/APCI/MALDI).

Electron Ionization (EI) technique

EI is the most appropriate technique for relatively small (m.w.<600) neutral organic molecules which can easily be promoted to the gas phase without decomposition, i.e. volatile. Since EI samples are thermally desorbed to the gas phase and subjected to the high energy of EI, analytes must be thermally stable. The gas phase molecules enter into the ion source where they are bombarded with free electrons emitted from a filament (Figure 1). The electrons bombard the molecules causing a hard ionization that fragments the molecule, and turn into positively charged particles called ions. This is important because the particles must be charged to pass through the analyser. As the ions continue from the source, they travel through an analyser (electromagnetic/quadrupole/the ion trap) that filters the ions based on mass to charge ratio. The filter continuously scans through the range of masses as the stream of ions come from the ion source. A detector counts the number of ions with a specific mass. This information is sent to a computer and a mass spectrum is created. The mass spectrum is a graph of the number of ions with different masses that traveled through the analyser.

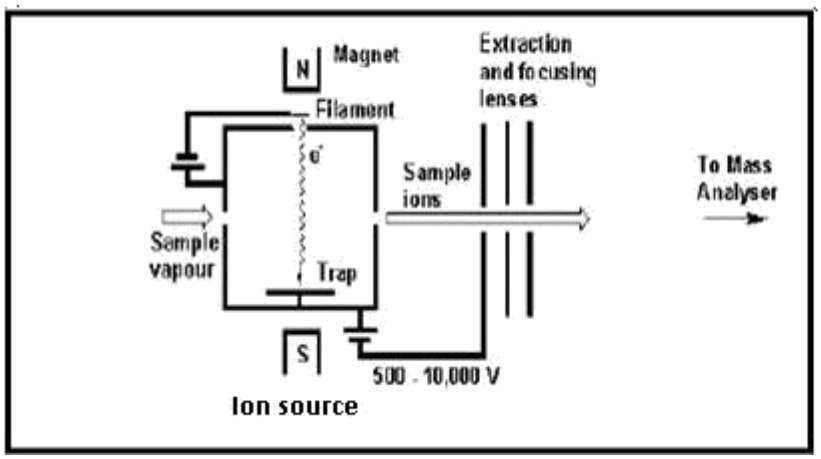


Figure 1. Schematic diagram of EIMS

Chemical Ionization (CI) technique

CI technique is especially useful when no molecular ion is observed in EI mass spectrum of a compound, and also in the case of confirming the molecular weight of a compound. CI technique uses nearly the same ion source device as in EI, except, CI uses tight ion source, and a reagent gas (Figure 2).

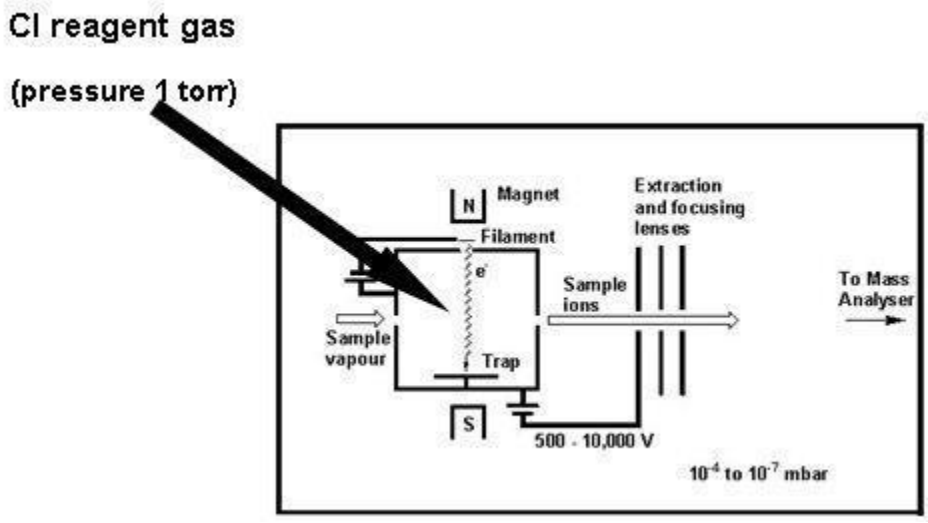


Figure 2. Schematic diagram of CI interface

Reagent gas (e.g. methane, iso-butane and ammonia) is first subjected to electron impact to yield reagent gas ions. These initial reagent gas ions further undergo ion-molecule reactions with neutral reagent molecules (G) to yield reagent selective ions (reagent plasma, e.g., GH+). When sample is introduced, the sample molecules (M) undergo ion-molecule reactions with reagent plasma to produce sample ions. In general, reagent gas molecules are present in the ratio of about 100:1 with respect to sample molecules. Pseudo-molecular ions, [M+H]+ (positive ion mode) or [M-H]- (negative ion mode) are often observed. Unlike in EI method, the CI process is soft ionization and yields abundant quasi-molecular ions, with less fragment ions.

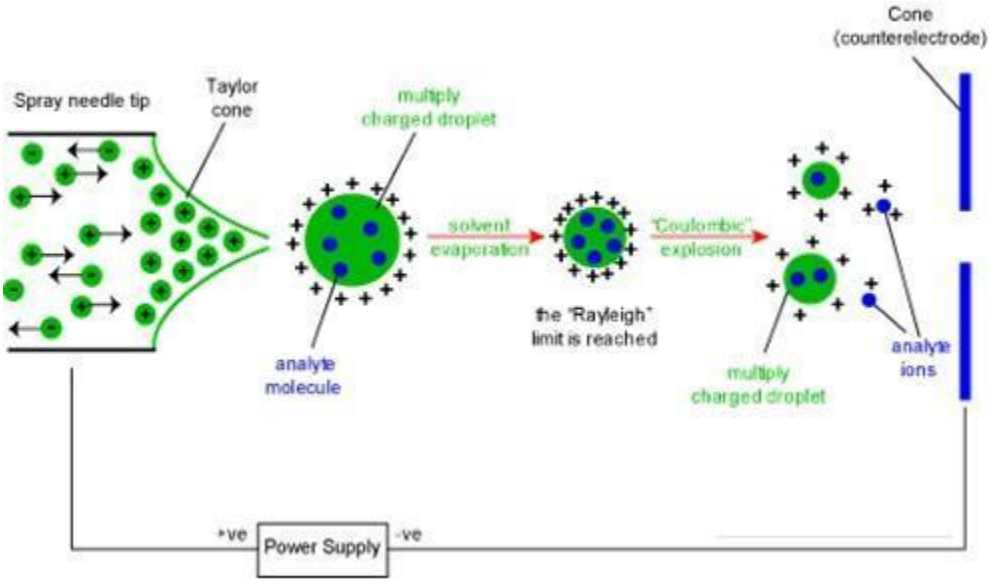
Positive ion mode: GH+ + M ------> MH+ + G

Negative ion mode: [G-H]- + M ------> [M-H]- + G

The fragmentation pattern of protonated molecules obtained under CI conditions may be different from that of the molecular ions observed under EI conditions. In CI mass spectrometry the molecules of a vaporized sample are ionized by a set of reagent ions (reagent plasma) in a series of ion-molecule reactions. The energy transferred by these reactions is lower than the energy imparted by electrons in EI source, and therefore fragmentation of the sample molecules is greatly decreased. For this reason CI mass spectrometry has been finding increasing use as a tool for the molecular weight confirmation and for elucidation of structure of variety of organic compounds including differentiation of isomeric compounds. Generally hydrogen (H2), methane (CH4), isobutane (iso-C4H10) and ammonia (NH3) are used as reagent gases in CI mass spectrometry; with all these CI gases the compounds form protonated molecule ion in their CI spectra.

Electrospray Ionization (ESI)

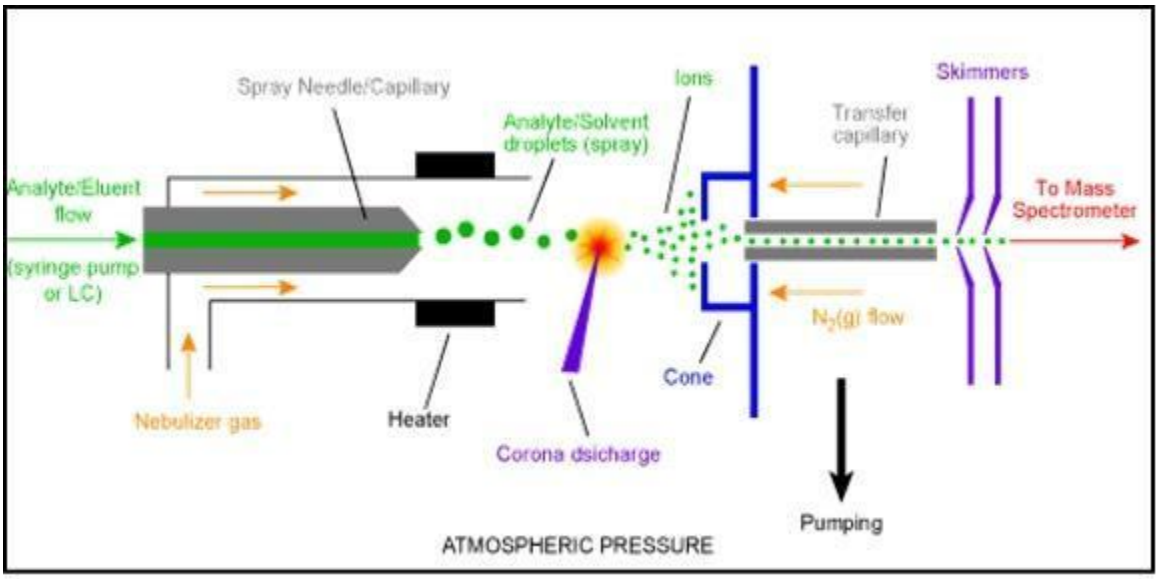
ESI technique involves spraying of a solution of the sample through a highly charged needle so-called capillary which is at atmospheric pressure (Figure 3). The spraying process can be streamlined by using a nebulizing gas. The charged droplets are produced in which the positive or negative ions are solvated with solvent molecules. Heat gas or a dry gas, usually called as desolvation gas is applied to the charged droplets to cause solvent evaporation. The desolvation process decreases the droplet size, leads to the columbic repulsion between the charges present in the droplet and further the droplet fission leads to the formation of individual gas phase analyte ions (that critical point known as the Rayleigh limit), that are guided through ion optics into the mass analyzer. ESI can produce singly or multiply charged ions. The number of charges retained by a particular analyte depends on several factors such as the size, chemical composition, and higher order structure of the analyte molecule, the solvent composition, the presence of co-solutes, and the instrument parameters. For small molecules (< 2000 Da) ESI typically generates singly, doubly, or triply charged ions, while for large molecules (> 2000 Da) ESI can produce a series of multiply charged ions.



ESI is very suitable for a wide range of biochemical compounds including peptides and proteins, lipids, oligosaccharides, oligonucleotides, bio-inorganic compounds, synthetic polymers, and intact non-covalent complexes.

Atmospheric pressure chemical ionization (APCI) technique

APCI has also become an important ionization source because it generates ions directly from solution and it is capable of analyzing relatively non-polar compounds. Similar to electrospray, the liquid effluent of APCI (Figure 4) is introduced directly into the ionization source. The droplets are not charged and the APCI source contains a heated vaporizer, which facilitates rapid desolvation/vaporization of the droplets. Vaporized sample molecules are carried through an ion-molecule reaction region at atmospheric pressure. APCI ionization originates from the solvent being excited or ionized from the corona discharge. Because the solvent ions are present at atmospheric pressure conditions, chemical ionization of analyte molecules is very efficient; at atmospheric pressure analyte molecules collide with the reagent ions frequently.



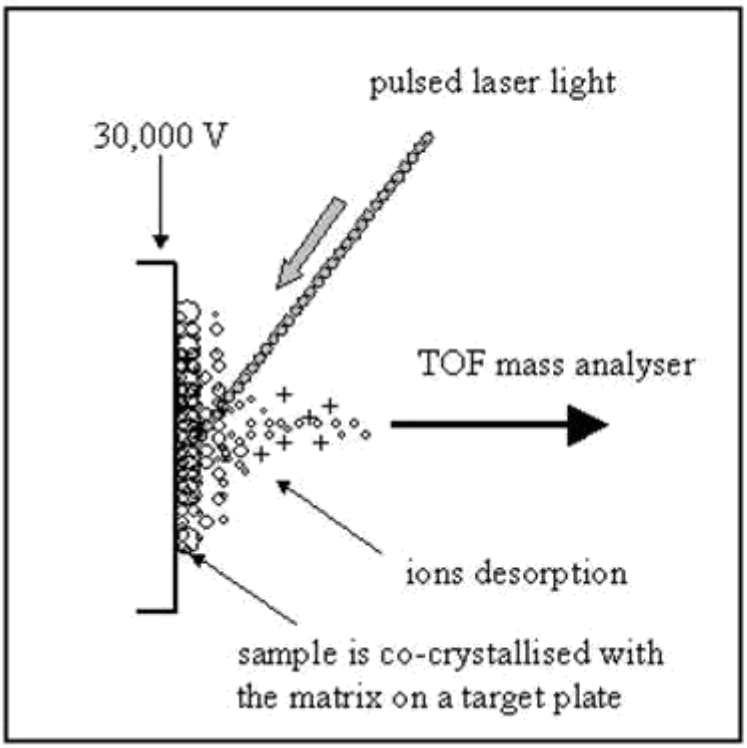
In general, proton transfer occurs in the positive mode to yield [M+H]+ ions. In the negative ion mode, either electron transfer or proton loss occurs to yield M- . or [M-H]- ions, respectively. The moderating influence of the solvent clusters on the reagent ions, and of the

high gas pressure, reduces fragmentation during ionization and results in primarily intact quasi-

molecular ions. Multiple charging is typically not observed presumably because the ionization process is more energetic than ESI.

MALDI technique

Matrix-assisted laser desorption/ionization (MALDI) is a technique to allows the high molecular weight compounds such as organic macro molecules and labile bimolecular into the gas phase as intact ions. MALDI is one of the recent developments of soft ionization techniques in the field of mass spectrometry. It can desorb intact analyte molecular ions with relative masses up to 300KDa. In MALDI-MS analysis, the analyte is first co-crystallized with a larger excess of a matrix compound (CHCA, DBA, Sinapic acid etc), after which, on laser radiation of this matrix-analyte preparation results in desorption of the matrix as a plume, which carries the analyte along with it into gas phase (Figure 5). Thus the matrix plays a key role by strongly absorbing the laser light energy and causing, indirectly, the analyte to vaporize. The matrix also serves as a proton donor and acceptor, acting to ionize analyte in both positive and negative ionization modes, respectively. The TOF analyzers are typically used with the MALDI ionization source.

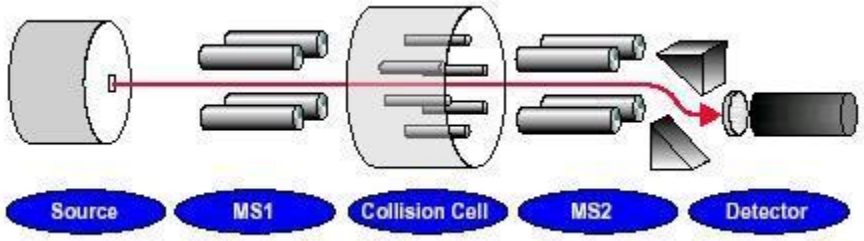


Schematic diagram of MALDI source

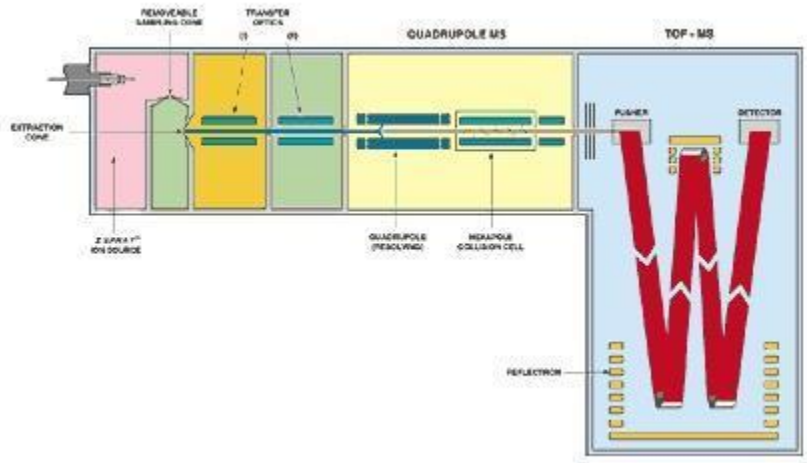
Tandem Mass spectrometry/Collision induced dissociation (CID)

The general approach of CID mass spectrometry (MS/MS), in modern terms known as product ion scanning using a triple quadrupole mass spectrometer is shown in Figure 6, and it is routinely used for primary structure determination. The first mass analyzer (MS1) is used to select ion of interest specifically from those transmitted by the ionization source. The precursor ion is passed into the collision cell where it undergoes low-energy collisions with an inert gas (argon or nitrogen) to induce fragmentation. The second mass analyser (MS2) acquires the m/z ratio for all the ions exiting the collision cell. For an MS/MS experiment performed using a triple-quadrupole equipped with an ESI source, a quadrupole/hexapole mass analyzer acts as the collision cell. Using the second quadrupole as a collision cell enables the re-focusing of scattered ions following fragmentation. Similarly, for MS/MS experiments using a quadrupole-time of flight (TOF) mass spectrometer, the precursor is selected using a quadrupole analyzer and the product ions that are formed in the collision cell are analyzed by TOF (Figure 7).

Precursor ion scanning is used to confirm the identity of compounds from a mixture that result in a common daughter ion. For precursor ion scanning, the second mass analyzer (MS2) is fixed to only monitor and transmit product ions of a specific m/z ratio. The first mass analyzer (MS1) is set to scan the whole mass range that includes all the precursor ions whose fragmentation would result in the selected daughter ion.

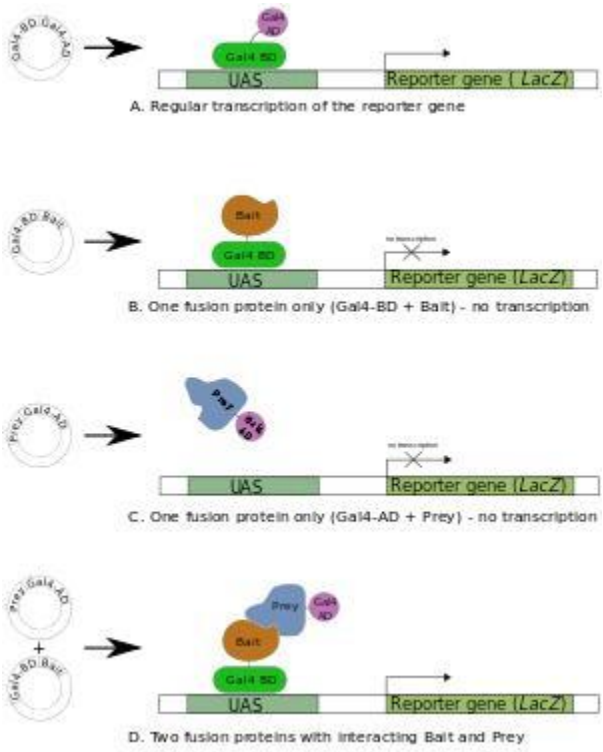


Schematic of a triple quadrupole mass spectrometer



Schematic of a quadrupole-TOF mass spectrometer

Two-hybrid screening



Overview of two-hybrid assay, checking for interactions between two proteins, called here *Bait* and *Prey*.

A. The *Gal4* transcription factor gene produces a two-domain protein (*BD* and *AD*) essential for transcription of the reporter gene (*LacZ*).

B,**C**. Two fusion proteins are prepared: *Gal4BD+Bait* and *Gal4AD+Prey*. Neither of them are usually sufficient to initiate transcription (of the reporter gene) alone.

D. When both fusion proteins are produced **and** the Bait part of the first fusion protein interacts with the Prey part of the second, transcription of the reporter gene occurs.

Two-hybrid screening (originally known as **yeast two-hybrid system** or **Y2H**) is a molecular biology technique used to discover protein–protein interactions (PPIs) and protein–DNA interactions [2][3] by testing for physical interactions (such as binding) between two proteins or a single protein and a DNA molecule, respectively.

The premise behind the test is the activation of downstream reporter gene(s) by the binding of a transcription factor onto an upstream activating sequence (UAS). For two-hybrid

screening, the transcription factor is split into two separate fragments, called the DNA-binding domain (DBD or often also abbreviated as BD) and activating domain (AD). The BD is the domain responsible for binding to the UAS and the AD is the domain responsible for the activation of transcription. [1][2] The Y2H is thus a protein-fragment complementation assay.

Pioneered by Stanley Fields and Ok-Kyu Song in 1989, the technique was originally designed to detect protein–protein interactions using the Gal4 transcriptional activator of the yeast *Saccharomyces cerevisiae*. The Gal4 protein activated transcription of a protein involved in galactose utilization, which formed the basis of selection. Since then, the same principle has been adapted to describe many alternative methods, including some that detect protein–DNA interactions or DNA-DNA interactions, as well as methods that use different host organisms such as *Escherichia coli* or mammalian cells instead of yeast.

The key to the two-hybrid screen is that in most eukaryotic transcription factors, the activating and binding domains are modular and can function in proximity to each other without direct binding. This means that even though the transcription factor is split into two fragments, it can still activate transcription when the two fragments are indirectly connected. The most common screening approach is the yeast two-hybrid assay. In this approach the researcher knows where each prey is located on the used medium (agar plates). Millions of potential interactions in several organisms have been screened in the latest decade using high-throughput screening systems (often using robots) and over thousands of interactions have been detected and categorized in databases as BioGRID. This system often utilizes a genetically engineered strain of yeast in which the biosynthesis of certain nutrients (usually amino acids or nucleic acids) is lacking. When grown on media that lacks these nutrients, the yeast fail to survive. This mutant yeast strain can be made to incorporate foreign DNA in the form of plasmids. In yeast two-hybrid screening, separate bait and prey plasmids are simultaneously introduced into the mutant yeast strain or a mating strategy is used to get both plasmids in one host cell.

The second high-throughput approach is the library screening approach. In this set up the bait and prey harboring cells are mated in a random order. After mating and selecting surviving cells on selective medium the scientist will sequence the isolated plasmids to see which prey

(DNA sequence) is interacting with the used bait. This approach has a lower rate of reproducibility and tends to yield higher amounts of false positives compared to the matrix approach. Plasmids are engineered to produce a protein product in which the DNA-binding domain (BD) fragment is fused onto a protein while another plasmid is engineered to produce a protein product in which the activation domain (AD) fragment is fused onto another protein. The protein fused to the BD may be referred to as the bait protein, and is typically a known protein the investigator is using to identify new binding partners. The protein fused to the AD may be referred to as the prey protein and can be either a single known protein or a library of known or unknown proteins. In this context, a library may consist of a collection of protein-encoding sequences that represent all the proteins expressed in a particular organism or tissue, or may be generated by synthesising random DNA sequences. Regardless of the source, they are subsequently incorporated into the protein-encoding sequence of a plasmid, which is then transfected into the cells chosen for the screening method. This technique, when using a library, assumes that each cell is transfected with no more than a single plasmid and that, therefore, each cell ultimately expresses no more than a single member from the protein library.

If the bait and prey proteins interact (i.e., bind), then the AD and BD of the transcription factor are indirectly connected, bringing the AD in proximity to the transcription start site and transcription of reporter gene(s) can occur. If the two proteins do not interact, there is no transcription of the reporter gene. In this way, a successful interaction between the fused protein is linked to a change in the cell phenotype. The challenge of separating cells that express proteins that happen to interact with their counterpart fusion proteins from those that do not, is addressed in the following section.

In any study, some of the protein domains, those under investigation, will be varied according to the goals of the study whereas other domains, those that are not themselves being investigated, will be kept constant. For example, in a two-hybrid study to select DNA-binding domains, the DNA-binding domain, BD, will be varied while the two interacting proteins, the bait and prey, must be kept constant to maintain a strong binding between the BD and AD. There are a number of domains from which to choose the BD, bait and prey and AD, if these are to remain constant. In protein–protein interaction investigations, the BD may be chosen from any of

many strong DNA-binding domains such as Zif268. A frequent choice of bait and prey domains are residues 263–352 of yeast Gal11P with a N342V mutation and residues 58–97 of yeast Gal4, respectively. These domains can be used in both yeast- and bacterial-based selection techniques and are known to bind together strongly.

The AD chosen must be able to activate transcription of the reporter gene, using the cell's own transcription machinery. Thus, the variety of ADs available for use in yeast-based techniques may not be suited to use in their bacterial-based analogues. The herpes simplex virus-derived AD, VP16 and yeast Gal4 AD have been used with success in yeast whilst a portion of the α-subunit of *E. coli* RNA polymerase has been utilised in *E. coli*-based methods. Whilst powerfully activating domains may allow greater sensitivity towards weaker interactions, conversely, a weaker AD may provide greater stringency.

Construction of expression plasmids

A number of engineered genetic sequences must be incorporated into the host cell to perform two-hybrid analysis or one of its derivative techniques. The considerations and methods used in the construction and delivery of these sequences differ according to the needs of the assay and the organism chosen as the experimental background. There are two broad categories of hybrid library: random libraries and cDNA-based libraries. A cDNA library is constituted by the cDNA produced through reverse transcription of mRNA collected from specific cells of types of cell. This library can be ligated into a construct so that it is attached to the BD or AD being used in the assay. A random library uses lengths of DNA of random sequence in place of these cDNA sections. A number of methods exist for the production of these random sequences, including cassette mutagenesis. Regardless of the source of the DNA library, it is ligated into the appropriate place in the relevant plasmid/phagemid using the appropriate restriction endonucleases.

***E. coli*-specific considerations**

By placing the hybrid proteins under the control of IPTG-inducible *lac* promoters, they are expressed only on media supplemented with IPTG. Further, by including different antibiotic resistance genes in each genetic construct, the growth of non-transformed cells is easily prevented through culture on media containing the corresponding antibiotics. This is particularly important for counter selection methods in which a *lack* of interaction is needed for cell survival. The reporter gene may be inserted into the *E. coli* genome by first inserting it into an episome, a type of plasmid with the ability to incorporate itself into the bacterial cell genome with a copy number of approximately one per cell. The hybrid expression phagemids can be electroporated into *E. coli* XL-1 Blue cells which after amplification and infection with VCS-M13 helper phage, will yield a stock of library phage. These phage will each contain one single-stranded member of the phagemid library. Once the selection has been performed, the primary structure of the proteins which display the appropriate characteristics must be determined. This is achieved by retrieval of the protein-encoding sequences (as originally inserted) from the cells showing the appropriate phenotype.

E. coli

The phagemid used to transform *E. coli* cells may be "rescued" from the selected cells by infecting them with VCS-M13 helper phage. The resulting phage particles that are produced contain the single-stranded phagemids and are used to infect XL-1 Blue cells. [2] The double-stranded phagemids are subsequently collected from these XL-1 Blue cells, essentially reversing the process used to produce the original library phage. Finally, the DNA sequences are determined through dideoxy sequencing.

Controlling sensitivity

The *Escherichia coli*-derived Tet-R repressor can be used in line with a conventional reporter gene and can be controlled by tetracycline or doxicycline (Tet-R inhibitors). Thus the expression of Tet-R is controlled by the standard two-hybrid system but the Tet-R in turn controls (represses) the expression of a previously mentioned reporter such as *HIS3*, through its Tet-R promoter. Tetracycline or its derivatives can then be used to regulate the sensitivity of a system utilising Tet-R. Sensitivity may also be controlled by varying the dependency of the cells

on their reporter genes. For example, this may be affected by altering the concentration of histidine in the growth medium for *his3*-dependent cells and altering the concentration of streptomycin for *aadA* dependent cells. Selection-gene-dependency may also be controlled by applying an inhibitor of the selection gene at a suitable concentration. 3-Amino-1,2,4-triazole (3-AT) for example, is a competitive inhibitor of the *HIS3*-gene product and may be used to titrate the minimum level of *HIS3* expression required for growth on histidine-deficient media. Sensitivity may also be modulated by varying the number of operator sequences in the reporter DNA.

Non-fusion proteins

A third, non-fusion protein may be co-expressed with two fusion proteins. Depending on the investigation, the third protein may modify one of the fusion proteins or mediate or interfere with their interaction. Co-expression of the third protein may be necessary for modification or activation of one or both of the fusion proteins. For example, *S. cerevisiae* possesses no endogenous tyrosine kinase. If an investigation involves a protein that requires tyrosine phosphorylation, the kinase must be supplied in the form of a tyrosine kinase gene. The non-fusion protein may mediate the interaction by binding both fusion proteins simultaneously, as in the case of ligand-dependent receptor dimerization. For a protein with an interacting partner, its functional homology to other proteins may be assessed by supplying the third protein in non-fusion form, which then may or may not compete with the fusion-protein for its binding partner. Binding between the third protein and the other fusion protein will interrupt the formation of the reporter expression activation complex and thus reduce reporter expression, leading to the distinguishing change in phenotype.

Split-ubiquitin yeast two-hybrid

One limitation of classic yeast two-hybrid screens is that they are limited to soluble proteins. It is therefore impossible to use them to study the protein–protein interactions between insoluble integral membrane proteins. The split-ubiquitin system provides a method for overcoming this limitation. In the split-ubiquitin system, two integral membrane proteins to be studied are fused to two different ubiquitin moieties: a C-terminal ubiquitin moiety ("Cub",

residues 35–76) and an N-terminal ubiquitin moiety ("Nub", residues 1–34). These fused proteins are called the bait and prey, respectively. In addition to being fused to an integral membrane protein, the Cub moiety is also fused to a transcription factor (TF) that can be cleaved off by ubiquitin specific proteases. Upon bait–prey interaction, Nub and Cub-moieties assemble, reconstituting the split-ubiquitin. The reconstituted split-ubiquitin molecule is recognized by ubiquitin specific proteases, which cleave off the transcription factor, allowing it to induce the transcription of reporter genes.

Fluorescent two-hybrid assay

Zolghadr and co-workers presented a fluorescent two-hybrid system that uses two hybrid proteins that are fused to different fluorescent proteins as well as LacI, the lac repressor. The structure of the fusion proteins looks like this: FP2-LacI-bait and FP1-prey where the bait and prey proteins interact and bring the fluorescent proteins (FP1 = GFP, FP2=mCherry) in close proximity at the binding site of the LacI protein in the host cell genome. [13] The system can also be used to screen for inhibitors of protein–protein interactions.

Enzymatic two-hybrid systems: KISS

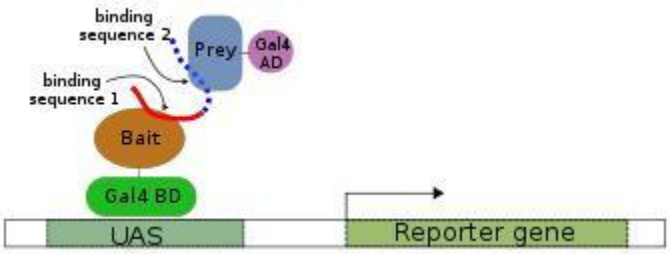
While the original Y2H system used a reconstituted transcription factor, other systems create enzymatic activities to detect PPIs. For instance, the KInase Substrate Sensor ("KISS"), is a mammalian two-hybrid approach has been designed to map intracellular PPIs. Here, a bait protein is fused to a kinase-containing portion of TYK2 and a prey is coupled to a gp130 cytokine receptor fragment. When bait and prey interact, TYK2 phosphorylates STAT3 docking sites on the prey chimera, which ultimately leads to activation of a reporter gene. [15]

One-hybrid

The one-hybrid variation of this technique is designed to investigate protein–DNA interactions and uses a single fusion protein in which the AD is linked directly to the binding domain. The binding domain in this case however is not necessarily of fixed sequence as in two-hybrid protein–protein analysis but may be constituted by a library. This library can be selected against the desired target sequence, which is inserted in the promoter region of the reporter gene construct. In a positive-selection system, a binding domain that successfully binds the UAS and allows transcription is thus selected.

Note that selection of DNA-binding domains is not necessarily performed using a one-hybrid system, but may also be performed using a two-hybrid system in which the binding domain is varied and the bait and prey proteins are kept constant.

Three-hybrid



Overview of three-hybrid assay.

RNA-protein interactions have been investigated through a three-hybrid variation of the two-hybrid technique. In this case, a hybrid RNA molecule serves to adjoin together the two protein fusion domains—which are not intended to interact with each other but rather the intermediary RNA molecule (through their RNA-binding domains). Techniques involving non-fusion proteins that perform a similar function, as described in the 'non-fusion proteins' section above, may also be referred to as three-hybrid methods.

One-two-hybrid

Simultaneous use of the one- and two-hybrid methods (that is, simultaneous protein–protein and protein–DNA interaction) is known as a one-two-hybrid approach and expected to increase the stringency of the screen. Although theoretically, any living cell might be used as the background to a two-hybrid analysis, there are practical considerations that dictate which is chosen. The chosen cell line should be relatively cheap and easy to culture and sufficiently robust to withstand application of the investigative methods and reagents. The latter is especially important for doing high-throughput studies. Therefore the yeast *S. cerevisiae* has been the main host organism for two-hybrid studies. However it is not always the ideal system to study interacting proteins from other organisms. Yeast cells often do not have the same post translational modifications, have a different codon use or lack certain proteins that are important for the correct expression of the proteins. To cope with these problems several novel two-hybrid systems have been developed. Depending on the system used agar plates or specific growth medium is used to grow the cells and allow selection for interaction. The most common used method is the agar plating one where cells are plated on selective medium to see of interaction takes place. Cells that have no interaction proteins should not survive on this selective medium.

The yeast *S. cerevisiae* was the model organism used during the two-hybrid technique's inception. It is commonly known as the Y2H system. It has several characteristics that make it a robust organism to host the interaction, including the ability to form tertiary protein structures, neutral internal pH, enhanced ability to form disulfide bonds and reduced-state glutathione among other cytosolic buffer factors, to maintain a hospitable internal environment. The yeast model can be manipulated through non-molecular techniques and its complete genome sequence is known. Yeast systems are tolerant of diverse culture conditions and harsh chemicals that could not be applied to mammalian tissue cultures. A number of yeast strains have been created specifically for Y2H screens, e.g. Y187 and AH109,both produced by Clontech. Yeast strains R2HMet and BK100 have also been used.

Candida albicans

C. albicans is a yeast with a particular feature: it translates the CUG codon into serine rather than leucine. Due to this different codon usage it is difficult to use the model system *S. cerevisiae* as a Y2H to check for protein-protein interactions using *C. albicans* genes. To provide a more native environment a *C. albicans* two-hybrid (C2H) system was developed. With this system protein-protein interactions can be studied in *C. albicans* itself. Bacterial *E. coli*-based two hybrid methods (abbreviated as B2H) have several characteristics that may make them preferable to yeast-based homologues. The higher transformation efficiency and faster rate of growth lends *E. coli* to the use of larger libraries (in excess of 10^8). A low false positive rate of approximately $3x10^{-8}$, the absence of requirement for a nuclear localisation signal to be included in the protein sequence and the ability to study proteins that would be toxic to yeast may also be major factors to consider when choosing an experimental background organism. It may be of note that the methylation activity of certain *E. coli* DNA methyltransferase proteins may interfere with some DNA-binding protein selections. If this is anticipated, the use of an *E. coli* strain that is defective for a particular methyltransferase may be an obvious solution. Important to mention is that bacteria are prokaryotic organisms and when studying eukaryotic protein-protein interactions (e.g. human proteins) the results need to be carefully approached.

Mammalian cells

In recent years a mammalian two hybrid (M2H) system has been designed to study mammalian protein-protein interactions in a cellular environment that closely mimics the native protein environment Transiently transfected mammalian cells are used in this system to find protein-protein interactions. Using a mammalian cell line to study mammalian protein-protein interactions gives the advantage of working in a more native context. [5] The post-translational modifications, phosphorylation, acylation and glycosylation are similar. The intracellular localization of the proteins is also more correct compared to using a yeast two hybrid system It is also possible with the mammalian two-hybrid system to study signal inputs. Another big advantage is that results can be obtained wishing 48 hours after transfection.

Arabidopsis thaliana

In 2005 a two hybrid system in plants was developed. Using protoplasts of *A. thaliana* protein-protein interactions can be studied in plants. This way the interactions can be studied in their native context. In this system the GAL4 AD and BD are under the control of the strong 35S promoter. Interaction is measured using a GUS reporter. In order to enable a high-throughput screening the vectors were made gateway compatible. The system is known as the protoplast two hybrid (P2H) system. The sea hare *A californica* is a model organism in neurobiology to study among others the molecular mechanisms of long-term memory. To study interactions, important in neurology, in a more native environment a two-hybrid system has been developed in *A californica* neurons. A GAL4 AD and BD are used in this system. An insect two-hybrid (I2H) system was developed in a silkworm cell line from the larva or caterpillar of the domesticated silk moth, *Bombyx mori* (BmN4 cells). This system uses the GAL4 BD and the activation domain of mouse NF-κB P65. Both are under the control of the OpIE2 promoter.

Applications

Determination of sequences crucial for interaction

By changing specific amino acids by mutating the corresponding DNA base-pairs in the plasmids used, the importance of those amino acid residues in maintaining the interaction can be determined.

After using bacterial cell-based method to select DNA-binding proteins, it is necessary to check the specificity of these domains as there is a limit to the extent to which the bacterial cell genome can act as a sink for domains with an affinity for other sequences (or indeed, a general affinity for DNA).

Drug and poison discovery

Protein–protein signalling interactions pose suitable therapeutic targets due to their specificity and pervasiveness. The random drug discovery approach uses compound banks that comprise random chemical structures, and requires a high-throughput method to test these structures in their intended target.

The cell chosen for the investigation can be specifically engineered to mirror the molecular aspect that the investigator intends to study and then used to identify new human or animal therapeutics or anti-pest agents.

Determination of protein function

By determination of the interaction partners of unknown proteins, the possible functions of these new proteins may be inferred. This can be done using a single known protein against a library of unknown proteins or conversely, by selecting from a library of known proteins using a single protein of unknown function.

Zinc finger protein selection

To select zinc finger proteins (ZFPs) for protein engineering, methods adapted from the two-hybrid screening technique have been used with success. A ZFP is itself a DNA-binding protein used in the construction of custom DNA-binding domains that bind to a desired DNA sequence.

By using a selection gene with the desired target sequence included in the UAS, and randomising the relevant amino acid sequences to produce a ZFP library, cells that host a DNA-ZFP interaction with the required characteristics can be selected. Each ZFP typically recognises only 3–4 base pairs, so to prevent recognition of sites outside the UAS, the randomised ZFP is engineered into a 'scaffold' consisting of another two ZFPs of constant sequence. The UAS is thus designed to include the target sequence of the constant scaffold in addition to the sequence for which a ZFP is selected.

Strengths

- Two-hybrid screens are low-tech; they can be carried out in any lab without sophisticated equipment.
- Two-hybrid screens can provide an important first hint for the identification of interaction partners.
- The assay is scalable, which makes it possible to screen for interactions among many proteins. Furthermore, it can be automated, and by using robots many proteins can be screened against thousands of potentially interacting proteins in a relatively short time. Two types of large screens are used: the library approach and the matrix approach.
- Yeast two-hybrid data can be of similar quality to data generated by the alternative approach of coaffinity purification followed by mass spectrometry (AP/MS).

Weaknesses

- The main criticism applied to the yeast two-hybrid screen of protein–protein interactions are the possibility of a high number of false positive (and false negative) identifications. The exact rate of false positive results is not known, but earlier estimates were as high as 70%. This also, partly, explains the often found very small overlap in results when using a (high throughput) two-hybrid screening, especially when using different experimental systems.

The reason for this high error rate lies in the characteristics of the screen:

- Certain assay variants overexpress the fusion proteins which may cause unnatural protein concentrations that lead to unspecific (false) positives.
- The hybrid proteins are fusion proteins; that is, the fused parts may inhibit certain interactions, especially if an interaction takes place at the N-terminus of a test protein (where the DNA-binding or activation domain is typically attached).
- An interaction may not happen in yeast, the typical host organism for Y2H. For instance, if a bacterial protein is tested in yeast, it may lack a chaperone for proper folding that is only present in its bacterial host. Moreover, a mammalian protein is sometimes not correctly modified in yeast (e.g., missing phosphorylation), which can also lead to false results.

- The Y2H takes place in the nucleus. If test proteins are not localized to the nucleus (because they have other localization signals) two interacting proteins may be found to be non-interacting.
- Some proteins might specifically interact when they are co-expressed in the yeast, although in reality they are never present in the same cell at the same time. However, in most cases it cannot be ruled out that such proteins are indeed expressed in certain cells or under certain circumstances.

Each of these points alone can give rise to false results. Due to the combined effects of all error sources yeast two-hybrid have to be interpreted with caution. The probability of generating false positives means that all interactions should be confirmed by a high confidence assay, for example co-immunoprecipitation of the endogenous proteins, which is difficult for large scale protein–protein interaction data. Alternatively, Y2H data can be verified using multiple Y2H variants or bioinformatics techniques. The latter test whether interacting proteins are expressed at the same time, share some common features (such as gene ontology annotations or certain network topologies), have homologous interactions in other species.

Protein microarray

A protein microarray (or protein chip) is a high-throughput method used to track the interactions and activities of proteins, and to determine their function, and determining function on a large scale. Its main advantage lies in the fact that large numbers of proteins can be tracked in parallel. The chip consists of a support surface such as a glass slide, nitrocellulose membrane, bead, or microtitre plate, to which an array of capture proteins is bound. Probe molecules, typically labeled with a fluorescent dye, are added to the array. Any reaction between the probe and the immobilised protein emits a fluorescent signal that is read by a laser scanner Protein microarrays are rapid, automated, economical, and highly sensitive, consuming small quantities of samples and reagents. The concept and methodology of protein microarrays was first introduced and illustrated in antibody microarrays (also referred to as antibody matrix) in 1983 in a scientific publication and a series of patents. [6] The high-throughput technology behind the protein microarray was relatively easy to develop since it is based on the technology developed for DNA microarrays, which have become the most widely used microarrays.

Protein microarrays were developed due to the limitations of using DNA microarrays for determining gene expression levels in proteomics. The quantity of mRNA in the cell often doesn't reflect the expression levels of the proteins they correspond to. Since it is usually the protein, rather than the mRNA, that has the functional role in cell response, a novel approach was needed. Additionally post-translational modifications, which are often critical for determining protein function, are not visible on DNA microarrays. [8] Protein microarrays replace traditional proteomics techniques such as 2D gel electrophoresis or chromatography, which were time consuming, labor-intensive and ill-suited for the analysis of low abundant proteins.

The proteins are arrayed onto a solid surface such as microscope slides, membranes, beads or microtitre plates. The function of this surface is to provide a support onto which proteins can be immobilized. It should demonstrate maximal binding properties, whilst maintaining the protein in its native conformation so that its binding ability is retained. Microscope slides made of glass or silicon are a popular choice since they are compatible with the easily obtained robotic arrays and laser scanners that have been developed for DNA

microarray technology. Nitrocellulose film slides are broadly accepted as the highest protein binding substrate for protein microarray applications.

The chosen solid surface is then covered with a coating that must serve the simultaneous functions of immobilising the protein, preventing its denaturation, orienting it in the appropriate direction so that its binding sites are accessible, and providing a hydrophilic environment in which the binding reaction can occur. In addition, it also needs to display minimal non-specific binding in order to minimize background noise in the detection systems. Furthermore, it needs to be compatible with different detection systems. Immobilising agents include layers of aluminium or gold, hydrophilic polymers, and polyacrylamide gels, or treatment with amines, aldehyde or epoxy. Thin-film technologies like physical vapour deposition (PVD) and chemical vapour deposition (CVD) are employed to apply the coating to the support surface.

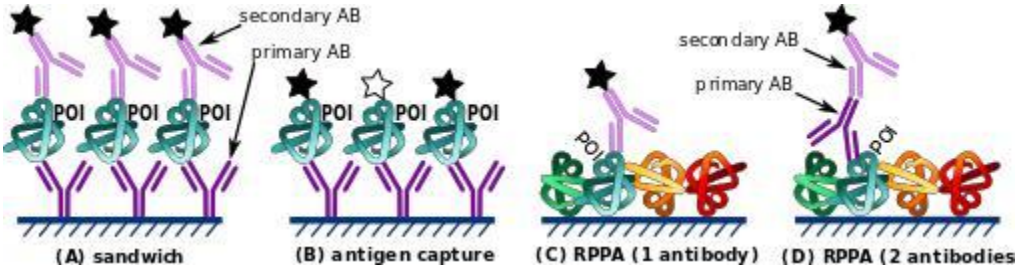
An aqueous environment is essential at all stages of array manufacture and operation to prevent protein denaturation. Therefore, sample buffers contain a high percent of glycerol(to lower the freezing point), and the humidity of the manufacturing environment is carefully regulated. Microwells have the dual advantage of providing an aqueous environment while preventing cross-contamination between samples.

In the most common type of protein array, robots place large numbers of proteins or their ligands onto a coated solid support in a pre-defined pattern. This is known as robotic contact printing or robotic spotting. Another fabrication method is ink-jetting, a drop-on-demand, non-contact method of dispersing the protein polymers onto the solid surface in the desired pattern. Piezoelectric spotting is a similar method to ink-jet printing. The printhead moves across the array, and at each spot uses electric stimulation to deliver the protein molecules onto the surface via tiny jets. This is also a non-contact process. [10] Photolithography is a fourth method of arraying the proteins onto the surface. Light is used in association with photomasks, opaque plates with holes or transparencies that allow light to shine through in a defined pattern. A series of chemical treatments then enables deposition of the protein in the desired pattern upon the material underneath the photomask.

The capture molecules arrayed on the solid surface may be antibodies, antigens, aptamers (nucleic acid-based ligands), affibodies (small molecules engineered to mimic monoclonal antibodies), or full length proteins. Sources of such proteins include cell-based expression systems for recombinant proteins, purification from natural sources, production in vitro by cell- free translation systems, and synthetic methods for peptides. Many of these methods can be automated for high throughput production but care must be taken to avoid conditions of synthesis or extraction that result in a denatured protein which, since it no longer recognizes its binding partner, renders the array useless.

Proteins are highly sensitive to changes in their microenvironment. This presents a challenge in maintaining protein arrays in a stable condition over extended periods of time. In situ methods involve on-chip synthesis of proteins as and when required, directly from the DNA using cell-free protein expression systems. Since DNA is a highly stable molecule it does not deteriorate over time and is therefore suited to long-term storage. This approach is also advantageous in that it circumvents the laborious and often costly processes of separate protein purification and DNA cloning, since proteins are made and immobilised simultaneously in a single step on the chip surface. Examples of in situ techniques are PISA (protein in situ array), NAPPA (nucleic acid programmable protein array) and DAPA (DNA array to protein array).

Types of arrays



Analytical microarrays are also known as capture arrays. In this technique, a library of antibodies, aptamers or affibodies is arrayed on the support surface. These are used as capture molecules since each binds specifically to a particular protein. The array is probed with a complex protein solution such as a cell lysate. Analysis of the resulting binding reactions using various detection systems can provide information about expression levels of particular proteins in the sample as well as measurements of binding affinities and specificities. This type of microarray is especially useful in comparing protein expression in different solutions. For instance the response of the cells to a particular factor can be identified by comparing the lysates of cells treated with specific substances or grown under certain conditions with the lysates of control cells. Another application is in the identification and profiling of diseased tissues.

Functional protein microarrays (also known as target protein arrays) are constructed by immobilising large numbers of purified proteins and are used to identify protein–protein, protein–DNA, protein– RNA, protein– phospholipid, and protein–small-molecule interactions, to assay enzymatic activity and to detect antibodies and demonstrate their specificity. They differ from analytical arrays in that functional protein arrays are composed of arrays containing full-length functional proteins or protein domains. These protein chips are used to study the biochemical activities of the entire proteome in a single experiment.

Reverse phase protein microarray (RPPA) involve complex samples, such as tissue lysates. Cells are isolated from various tissues of interest and are lysed. The lysate is arrayed onto the microarray and probed with antibodies against the target protein of interest. These antibodies are typically detected with chemiluminescent, fluorescent or colorimetric assays. Reference peptides are printed on the slides to allow for protein quantification of the sample lysates. RPAs allow for the determination of the presence of altered proteins or other agents that may be the result of disease. Specifically, post-translational modifications, which are typically altered as a result of disease can be detected using RPAs.

Detection

Protein array detection methods must give a high signal and a low background. The most common and widely used method for detection is fluorescence labeling which is highly sensitive,

safe and compatible with readily available microarray laser scanners. Other labels can be used, such as affinity, photochemical or radioisotope tags. These labels are attached to the probe itself and can interfere with the probe-target protein reaction. Therefore, a number of label free detection methods are available, such as surface plasmon resonance (SPR), carbon nanotubes, carbon nanowire sensors (where detection occurs via changes in conductance) and microelectromechanical system (MEMS) cantilevers. All these label free detection methods are relatively new and are not yet suitable for high-throughput protein interaction detection; however, they do offer much promise for the future.

Protein quantitation on nitrocellulose coated glass slides can use near-IR fluorescent detection. This limits interferences due to auto-fluorescence of the nitrocellulose at the UV wavelengths used for standard fluorescent detection probes.

Applications

There are five major areas where protein arrays are being applied: diagnostics, proteomics, protein functional analysis, antibody characterization, and treatment development Diagnostics involves the detection of antigens and antibodies in blood samples; the profiling of sera to discover new disease biomarkers; the monitoring of disease states and responses to therapy in personalized medicine; the monitoring of environment and food. Digital bioassay is a en example of using protein microarray for diagnostic purposes. In this technology, an array of microwells on a glass/polymer chip are seeded with magnetic beads (coated with fluorescent tagged antibodies), subjected to targeted antigens and then characterised by a microscope through counting fluorescing wells. A cost-effective fabrication platform (using OSTE polymers) for such microwell arrays has been recently demonstrated and the bio-assay model system has been successfully characterised.

Proteomics pertains to protein expression profiling i.e. which proteins are expressed in the lysate of a particular cell. Protein functional analysis is the identification of protein–protein interactions (e.g. identification of members of a protein complex), protein–phospholipid interactions, small molecule targets, enzymatic substrates (particularly the substrates of kinases) and receptor ligands. Antibody characterization is characterizing cross-reactivity, specificity and mapping epitopes. Treatment development involves the development of antigen-specific therapies for autoimmunity, cancer and allergies; the identification of small molecule targets that could potentially be used as new drugs.

Challenges

Despite the considerable investments made by several companies, proteins chips have yet to flood the market. Manufacturers have found that proteins are actually quite difficult to handle. A protein chip requires a lot more steps in its creation than does a DNA chip. Challenges include: 1) finding a surface and a method of attachment that allows the proteins to maintain their secondary or tertiary structure and thus their biological activity and their interactions with other molecules, 2) producing an array with a long shelf life so that the proteins on the chip do not denature over a short time, 3) identifying and isolating antibodies or other capture molecules against every protein in the human genome, 4) quantifying the levels of bound protein while assuring sensitivity and avoiding background noise, 5) extracting the detected protein from the chip in order to further analyze it, 6) reducing non-specific binding by the capture agents, 7) the capacity of the chip must be sufficient to allow as complete a representation of the proteome to be visualized as possible; abundant proteins overwhelm the detection of less abundant proteins such as signaling molecules and receptors, which are generally of more therapeutic interest.

BIOMARKERS

A biomarker, or biological marker, generally refers to a measurable indicator of some biological state or condition. The term is also occasionally used to refer to a substance whose detection indicates the presence of a living organism. Biomarkers are often measured and evaluated to examine normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. Biomarkers are used in many scientific fields. The widespread use of the term "biomarker" dates back to as early as 1980. The term "biological marker" was introduced in 1950s. In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention."

After a heart attack a number of different cardiac biomarkers can be measured to determine exactly when an attack occurred and how severe it was. In medicine, a biomarker can be a traceable substance that is introduced into an organism as a means to examine organ function or other aspects of health. For example, rubidium chloride is used as a radioactive isotope to evaluate perfusion of heart muscle.

It can also be a substance whose detection indicates a particular disease state, for example, the presence of an antibody may indicate an infection. More specifically, a biomarker indicates a change in expression or state of a protein that correlates with the risk or progression of a disease, or with the susceptibility of the disease to a given treatment.

Other biomarkers can be based on measures of the electrical activity of the brain (using Electroencephalography (so-called Quantitative electroencephalography (qEEG)) or Magnetoencephalography), or volumetric measures of certain brain regions (using Magnetic resonance imaging) or saliva testing of natural metabolites, such as saliva nitrite, a surrogate marker for nitric oxide. One example of a commonly used biomarker in medicine is prostate-specific antigen (PSA). This marker can be measured as a proxy of prostate size with rapid changes potentially indicating cancer. The most extreme case would be to detect mutant proteins as cancer specific biomarkers through Selected Reaction Monitoring (SRM), since mutant

proteins can only come from an existing tumor, thus providing ultimately the best specificity for medical purposes.

Biomarkers used for personalized medicine are typically categorized as either prognostic or predictive. An example is KRAS, an oncogene that encodes a GTPase involved in several signal transduction pathways. Prognostic biomarkers indicate the likelihood of patient outcome regardless of a specific treatment. Predictive biomarkers are used to help optimize ideal treatments, and indicates the likelihood of benefiting from a specific therapy. Biomarkers for precision oncology are typically utilized in the molecular diagnostics of chronic myeloid leukemia, colon, breast, and lung cancer, and in melanoma. HER2 is the most popular validated biomarker. The measurement of HER2 status is used to predict breast cancer treatment responses

This step allows the development of the most adapted protocol for routine use of the biomarker. Simultaneously, it is possible to confirm the relevance of the protocol with various methods (histology, PCR, ELISA, ...) and to define strata based on the results.

In cell biology, a biomarker is a molecule that allows the detection and isolation of a particular cell type (for example, the protein Oct-4 is used as a biomarker to identify embryonic stem cells) .[9]

In genetics, a biomarker (identified as genetic marker) is a DNA sequence that causes disease or is associated with susceptibility to disease. They can be used to create genetic maps of whatever organism is being studied.

A biomarker can be any kind of molecule indicating the existence, past or present, of living organisms. In the fields of geology and astrobiology, biomarkers, versus geomarkers, are also known as biosignatures.

Biomarkers are used to indicate an exposure to or the effect of xenobiotics which are present in the environment and in organisms. The biomarker may be an external substance itself (e.g. asbestos particles or NNK from tobacco), or a variant of the external substance processed by the body (a metabolite) that usually can be quantified.

In medicine, a biomarker is a measurable indicator of the severity or presence of some disease state. More generally a biomarker is anything that can be used as an indicator of a particular disease state or some other physiological state of an organism. A biomarker can be a substance that is introduced into an organism as a means to examine organ function or other aspects of health. For example, rubidium chloride is used in isotopic labeling to evaluate perfusion of heart muscle. It can also be a substance whose detection indicates a particular disease state, for example, the presence of an antibody may indicate an infection. More specifically, a biomarker indicates a change in expression or state of a protein that correlates with the risk or progression of a disease, or with the susceptibility of the disease to a given treatment. Biomarkers can be characteristic biological properties or molecules that can be detected and measured in parts of the body like the blood or tissue. They may indicate either normal or diseased processes in the body. Biomarkers can be specific cells, molecules, or genes, gene products, enzymes, or hormones. Complex organ functions or general characteristic changes in biological structures can also serve as biomarkers. Although the term biomarker is relatively new, biomarkers have been used in pre-clinical research and clinical diagnosis for a considerable time. For example, body temperature is a well-known biomarker for fever. Blood pressure is used to determine the risk of stroke. It is also widely known that cholesterol values are a biomarker and risk indicator for coronary and vascular disease, and that C-reactive protein (CRP) is a marker for inflammation. Biomarkers are useful in a number of ways, including measuring the progress of disease, evaluating the most effective therapeutic regimes for a particular cancer type, and establishing long-term susceptibility to cancer or its recurrence. The parameter can be chemical, physical or biological. In molecular terms biomarker is "the subset of markers that might be discovered using genomics, proteomics technologies or imaging technologies. Biomarkers play major roles in medicinal biology. Biomarkers help in early diagnosis, disease prevention, drug target identification, drug response etc. Several biomarkers have been identified for many diseases such as serum LDL for cholesterol, blood pressure, and P53 gene and MMPs as tumor markers for cancer.

Possible questions

Objective questions

1. What are ampholytes?
   (A) Poly amino and poly carboxylic compounds (B) Poly carboxylic compounds.
   (C) Poly amino compounds. (D) Poly hydroxyl compounds

2. In ESI-MS, the protein molecules would be
   (A) Fragmented (B) Condensed (C) Multiple charged (D) Uncharged

3. The mobile phase of RP-HPLC is
   (A) Non-polar (B) Polar (C) Neutral (D) All of the above

4. HPLC is generally operated at high pressure due to
   (A) Low theoretical plates (B) High theoretical plates (C) non-polar stationary phase (D) polar stationary phase

5. In MALDI-TOF, molecules are separated by subjecting to
   (A) Electric field only (B) Magnetic field only (C) Both electric and magnetic field
   (D) pH gradient

6. Isocratic elution in which the concentration of mobile phase is
   (A) Constant (B) Variable (C) Neither constant nor variable
   (D) Either constant or variable

7. The essential feature for a compound to be used as a matrix in MALDI is
   (A) Highly volatile (B) Highly stable (C) Highly soluble (D) Highly reactive

8. Biomolecules that are useful to detect certain diseases can be called as
   (A) Biomacromolecules (B) Biomarkers (C) Biosensors (D) Biocatalysts

2 Marks Questions

1. Write the unique feature of MALDI-TOF.
2. What is 'protein microarray'?
3. How are proteins immobilized in 'Protein microarray'?
4. What are affibodies?

6/8/10 Marks questions

1. Describe the principles of MALDI-MS/MALDI-MS-TOF and its applications on analyzing protein-protein interactions in a detailed manner.
2. Describe the principles of RP-HPLC and its applications on analyzing protein-protein interactions in a detailed manner.
3. Systematically enumerate the applications of Y2H methods on analyzing protein-protein interactions in a detailed manner.

Question	Option 1	Option 2	Option 3	Option 4	Answer
Unit 1					
DNA sequencing refers to	technique used to determine the sugar sequence in a DNA molecule	technique used to determine the phosphate sequence in a DNA molecule	technique used to determine the baase sequence in a DNA molecule	technique used to determine the amino acid sequence in a protein	technique used to determine the phosphate sequence in a DNA molecule
Which of the following is a chemical nucelotide sequencing method?	Sanger method	Maxam-Gilbert Method	Edmans method	Automated sequencing method	Maxam- Gilbert method
The enzyme used in Maxam-Gilbert method for P32 labelling of the DNA at 3' end is	polynucelotide kinase	alkaline phosphatase	exonuclease	terminal nucelotidyl transferase	polynucelotide kinase
How many different types of chemcial treatments are required in Maxam-Gilbert method?	1	2	3	4	4
Guanine specific cleavage in Maxam-Gilbert method	formic acid	hydrazine	Dimethyl sulphate	piperdine	Dimethyl sulphate
The principle of sanger's method relies on	use of chemicals for base specific cleavage	Use of dNTPs for chain termination	Use of ddNTPs for chain termination	use of For P32 Chain termination	Use of ddNTPs for chain termination
The sample in sanger's method after reaction are separated using	AGE	PAGE	PFGE	2-D gel electrophoresis	PAGE
Which of the following sequencing methods uses PCR for generating sequence templates?	Sanger's method	Sanger's method	LMPCR	LMPCR and automated DNA sequencing	LMPCR
Automated DNA sequencing is an improvement of Sanger's method where	ddNTPs are used for chain termination	PCR is used for making sequencing templates	Fluorescently labelled ddNTPs are used for chain termination	Fluorescently labelled ddNTPs are used for chain termination	Fluorescently labelled ddNTPs are used for chain termination
Which of the following is not a DNA sequencing method?	LMPCRR	Edmans method	Sanger's method	Maxam-Gilbert method	Edman's method
GFP protein was originally isolated from	*A.thaliana*	*A.victoria*	*C.elegans*	*D.melanogaster*	*A.victoria*
The first crop plant genome sequenced	Maize	*A.thaliana*	Rice	Barley	*A.thaliana*
The year of launching of Human Genome project and completion of rough draft of the sequence was in	1992-2000	1990-2000	1990-2001	1991-2001	1992-2000
The name Kary Mullis is associated with	RFLP	PCR	Chain termination method	RAPD	PCR
In Maxam-gilbert method, chemical used for cytosine alteration is	Formic acid	hydrazine	Dimethyl sulphate	piperdine	hydrazine
The chemical used for strand cleavage in Maxam-Gilbert method is	Piperdine	hydrazine	formic acid	DMSO	Piperdine
ddNTP is different from dNTP in having	H in place of OH in 3 position of dNTP	OH in place of H in 3 position of dNTP	OH in place of H 2 position of dNTP	CH3 in place of OH in 3 position of dNTP	H in place of OH in 3 position of dNTP
Sanger sequenced the first complete genome of	T4 phage	C. elegans	TMV	øX 174	øX 174
Pseudogenes are important to evolution because	they are silent genes that have been activated by mutations	they did not exist until recently	they occur entirely within multigene families	they tend to be transcribed more frequently	they are silent genes that have been activated by mutations
Tandem clusters occur most commonly in	satellite DNA	centromere	the nuclear organizer regions	plasmids	the nuclear organizer regions
A recombinant vector in a yeast genomic library may contain more than one gene. This is not the case wit the recombinant vectors in a yeast cDNA library because	every cDNA fragment cloned into each vector carries only one gene	different restriction enzymes are used	the host cells cannot be transformed with a vector carrying more than one gene	the vectors used in making a cDNA library are different	every cDNA fragment cloned into each vector carries only one gene
unlike a genomic DNA library particular care must be taken to ensure that the cDNA library represents every gene in the genome. Why is this the case?	some genes are transcribed only in certain cell types or at low levels	Host cells are more likely to reject vectors with cDNAs as compared to genomic DNA	It is harder to clone cDNA fragments as compared to genomic fragments in to the vectors	a cDNA library lacks introns	some genes are transcribed only in certain cell types or at low levels
Many cDNA libraries are used as expression libraries. The vector chosen for use in an expression library must have additional DNA sequence that is not required in the vector chosen for use in a genomic DNA library. What is the sequence?	an origin of replication	a second selectable marker	additional restriction enzyme sites	a promoter sepecific for the host organism	a promoter sepecific for the host organism
Whole-genome shotgun sequencing requires	Construction of a genomic library of plasmid vectors	Random sequencing of DNA fragments	Fragment alignment and gap closure	cloning in to vectors and sequencing using shot-gun	Fragment alignment and gap closure
what information is absent from the sequence of a cDNA clone?	intron sequence	promoter sequence	amino acid sequence of the encoded polypeptides	intron and promoter sequences	intron and promoter sequences
What is the function of the ampr gene in a vector?	To allow resistant transformants to grow in selective medium	To distinguish introns from exons	To allow viral replication	To screen for vectors with inserts	To allow resistant transformants to grow in selective medium
The lacz gene is sometimes included in a cloning vector. What is its purpose?	allow resistant transformants to grow in selective medium	distinguish introns from exons	allow viral replication	screen for vectors with inserts	screen for vectors with inserts
Which vector is appropriate for subcloning an insert of 3 Kb?	ampr plasmid	YAC	cDNA	cosmid	ampr plasmid
The goal of-----is to determine the location of specific genes with in the genome	cloning	annotaion	proteomics	genomics	annotaion
An open reading frame (ORF) is	the sequence of a compelte genome	a plasmid vector used in genomic sequencing	a possible gene predicted by DNA sequencing	a fragment of a genome	a possible gene predicted by DNA sequencing
The whole genome shot gun sequencing approach depends priamrily on	rapidly sequencing thousands of small randomly cloned fragments	methodical sequencing a few large cloned fragments of DNA	sequencing the bacterial chromosome while it is still intact	fragmenting the genome and cloning	rapidly sequencing thousands of small randomly cloned fragments
How are the four different bases distinguished in automated sequencing systems?	each base has a different radioactive tag	each base has a distinctive fluorescent tag	each base has a unique antibody bound to it	each base tagged with antigen to detect by electrophoresis	each base has a distinctive fluorescent tag
proteomics is	a branch of quantum physics	the study of algal genomes	the study of the entire collection of proteins expressed by an organism	study of entire set of genomes	the study of the entire collection of proteins expressed by an organism
The enzyme reverse transcriptase enables scientists to produce what product?	Restriction endonucleases	cDNA molecules	Restriction fragment length polymorphisms	mRNA transcripts	cDNA molecules
Oligo nucelotide gene rpobes are defined as what?	Enzymes that recognize and subsequently degrade foreign DNA	The pieces of DNA produced by restriction endonucleases	A short stretch of DNA of a known sequence that will bnase-pair with a complementary sequence	A piece of DNA to which new nucelotides are added during DNA sequencing	A short stretch of DNA of a known sequence that will base-pair with a complementary sequence
The techniques that utilizes probes to detect specific DNA sequences is known as what?	Southern blot	Northern blot	Northwestern blot	eastern blot	Southern blot
Which of the following statements regarding the polymerase chain reaction is intrue?	It can increase the amount of DNA in a sample	It has the potential of diagnosing an infection from a single copy of a gene	It utilizes DNA polymerases from psychrophilic organisms	It can amplify DNA of only a few base pairs up to a whole genome	It utilizes DNA polymerases from psychrophilic organisms
When a gene has to be studied in detail, selected gene is removed from an animal, plant and microorganisms, and is inserted into what?	A primer	an oligonucleotide	a palindrome	a vector	Vector
Genomic libraries are useful for obtaining what product?	Periodicals on genomics research	collections of isolated genes	Instructional information on how to locate the exact site of the gene of interest	Information relating to primers and PCR	collections of isolated genes
Good cloning vectors must possessall but which of the following qualitites?	They should possess their won origin of replication	They should be readily accepted by the cloning host	They should be easily manipulated	They should be resistant to restriction endonucleases	They should be resistant to restriction endonucleases
The insertion of a cloning vector into a cloning host typcially involves what process?	Transduction	polymerase chain reaction	Tranformation	Conjugation	Tranformation
Which of the following is NOT a step in whole genome shotgun sequencing?	library construction	sequencing of randomly produced fragments	fragment alignment and closure	library construction, sequencing of randomly produced fragments alignment and editing	library construction, sequencing of randomly produced fragments alignment and editing
DNA was first isolated in the	17th century	18th century	19th century	20th century	19th century
Sizes of genomes of free-living organisms have been found to range from approximately	2-200 Mbp	0.5-200,000 Mbp	100-200000 Mbp	0.5-1000 Mbp	0.5-200,000 Mbp
The size of the human genome is about	0.5 Mbp	3,000 Mbp	10 Mbp	500,000 Mbp	3000 Mbp
Most sequences in the human genome belong to	Genes	Pseudogenes	Interspersed repeats	Tandem repeats	Interspersed repeats
Which genome has the fewest introns?	Drosophila genome	human genome	Maize genome	E.coli genome	E.coli genome

Gene density can be high	In telomeres	anywhere on the chromosomes	in centromeres	in anaphase chromosomes	anywhere on the chromosomes
This vector can be used to construct genomic libraries	phasmid	phagmid	cos site	BAC	BAC
mRNA can be readily isolated from lysed eukaryotic cells by adding magnetic beads which have ----covalently attached	Oligo (dT)	Oligo (dG)	Oligo(dC)	Oligo (dA)	Oligo (dT)
This enzyme is used to dephsophorylate the vector	Terminal transferase	Alkaline phosphatase	Klenow enzyme	DNA ligase	Alkaline phosphatase
usually nucleic acid sequences are added to create sticky ends for clonig called as	linkers	sequencers	promoters	enhancers	linkers
Minisatellites are	10-40 bp sized short sequences with in genes	Short coding repetitive regions on the eukaryotic genome	short non-coding repetitive sequences present through out the chromosome	Are regions of chromosomes after secondary constriction	short non-coding repetitive sequences present through out the chromosome
Each individual has a unique DNA fingerprinting as individuals differ in	number of minisatellites or chromosomes	location, size and number of the minisatellites on chromsomes	number of fragmented DNA	fragemented number of clones	location, size and number of the minisatellites on chromsomes
Protein coding genes can be identified by	transposon tagging	ORF scanning	Zoo-blotting	Nuclease S1 Mapping	ORF scanning
The functions of the genes can be determined by	Gene inactivation	Exon trapping	Zoo-blotting	Northern analysis	Gene inactivation
Reporter genes	Indicate the presence of stress conditions	can often be detected by histochemical assays	are all of bacterial origin	are used to characterize proteomes	can often be detected by histochemical assays
Chromosoem walking	Is used in genetic mapping	Occurs in mitosis	requires a genomic library and done by PCR	can be sued to close physical sequence maps	requires a genomic library and done by PCR
Fluorescent in situ hybridization	requires deoxynucleotides	requires a labelled probe and used in physical mapping of genomes	requires a DNA polymerase	used in genetic mapping of genomes	requires a labelled probe and used in physical mapping of genomes
genes can be altered or replaced by	transposon tagging	RNA interference	Homologous recombination	gene rearrangement	Homologous recombination
RFLP involves	used to identify a specific pro	used to identify a specific DNA	used to a identify a specific RN	used to identify a both DN	used to identify a specific RNA
Location of quantitative genes on chromosomes are called	quantitative trait loci	qualitative trait loci	maps	construct or position of ger	quantitative trait loci
RAPD is a	DNA sequencing based meth	restrictioon digestion based method	PCR based method	Enzyme based method	PCR based method
Set of DNAs generated by using random primers in a PCR reaction is called	RAPD	RFLP	AFLP	insitu hybridization	RAPD
Unit II					
What type of genome map is the most ideal for understanding the nature of genes, what they code and their functions?	Linkage maps	Physical maps	Sequence Maps	fingerprint maps	Sequence maps
Which of the following statements regarding the findings of the human genome project is incorrect?	Humans share approximately 80 % of their gene sequence with mice	Genetic screening of families for inheritable diseases may become possible	Human chromosome are capable of walking	A large amount of human genome contains DNA sequences that do not code for cell protein	Human chromosome are capable of walking
Microsatellites are	frequently found in bacterial	always smaller than 10 bp	used as DNA markers	movable DNA elements	used as DNA markers
Specific biomolecules which show easily detectable difference among different strains of a species of among different species is termed as	DNA fingerprinting	molecular markers	moelcular scissors	RFLP	molecular markers
Molecular markers include	RFLP	PCR	alkaline phosphatase	phosphatase	RFLP
Molecular markers are used to construct	Chromosome maps	cytogenetic maps	physical maps	geographic maps	physical maps
The variation in the restriction DNA fragment lengths between individuals of a species is called	Retriction fragment length polymorphism	RAPD	AFLP	simple sequence repeats	Retriction fragment length polymorphism
RFLP involves	used to identify a specific protein	used to identify a sepcific DNA	used to identify a sepcific RNA	used to identify both DNA and RNA	used to identify a sepcific RNA
Locations of quantitative genes on chromosomes are called	Qualitative trait loci	Quantitative trait loci	maps	construct or position of gene	Quantitative trait loci
RAPD is a	DNA sequencing based method	Restriction digestion based method	PCR based method	enzyme based method	PCR based method
The set of DNAs generated by using random primers in a PCR reaction is called	RAPD	RFLP	AFLP	Insitu hybridization	RAPD
All statements are true regarding RFLP and RAPD except	RAPD is a quick method compared to RFLP	RFLP is more relible than RAPD	Species specific primers are required for RAPD	Radioactive probes are not required in RAPD	Species specific primers are required for RAPD
DNA of eukaryotic organisms has several repeating units of short sequences called	random repeats	tandem repeats	mini satellites	microsatellites	Tandem repeats
The variation in number of tandem repeats between two or more individuals is called	VNTRs	RFLP	SSRs	AFLP	VNTRs
Simple sequence repeats SSRs are	1-6 bp long ssequences distributed along the chromosomes	also called as satellite repeats	not specific in number and position	not used as markers	1-6 bp long ssequences distributed along the chromosomes
AFLP is a	not a PCR based methods	it is not associated with polymorphism of DNA	method that detects the presence or absence of a fragment	not specific in position and number	method that detects the presence or absence of a fragment
The variant fragment that distinguish one individual from another one is called	variant fragment	marking fragment	differing fragment	variable repeats	Marking fragment
Which of these is an example of cytopgenetic mapping?	hybridization studies in Drosophila determined the location of the yellow gene, which influences body color, to be near the tip of the X chromosome	Karyotyping analysis in human chromosome	Analysis of the chromosome of a transgenic strain using transgenic mice	banding pattern techniques	hybridization studies in Drosophila determined the location of the yellow gene, which influences body color, to be near the tip of the X chromosome
Cytologistics can use which of the following to describe locations of a gene at a specific palce on the chromosome?	DNA sequence of a chromosomal region	protein expression from a chromosomal region	Banding pattern of stained chromosomes	compariosn to markers located with in a few thoudsand bp	Banding pattern of stained chromosomes
Cytogenetic mapping:	Requires that gene have been cloned if in situ hybridization is to be used	Is a high resolution method of gene localization	Can be used to determine relative order of genes located very close to each other	banding pattern techniques	Requires that gene have been cloned if in situ hybridization is to be used
Which of the following questions could be easily answered by FISH?	Recombination occurs between species	Is the mutations due to a deletion of the entire gene?	How large DNA would be need to clone	position of genes	Is the mutations due to a deletion of the entire gene?
Which of these is a key characterisitic of a molecular marker?	It is a known gene	It is located at a known site on the chromosome	It is only sueful for linkage and physical mapping studies	positional analysis	It is located at a known site on the chromosome
A moelcular marker which is amplified by PCR and is polymorphic by length is	RFLP	VNTRs	AFLP	SNP	AFLP
A polymorphism is	Any change in the DNA sequence	The most common variation of a gene or marker sequence	The least common vaariation of a gene or marker sequence	Variation of gene or marker sequence present in > 1% of the population	Variation of gene or marker sequence present in > 1% of the population
Which of the these statements regarding RFLP analysis is correct?	RFLP analysis requires southern blotting for detection of fragments	RFLPs can identify single base pair changes at any site in the chromosomes	An RFLP typically rpoduces several different alleles	positional analysis of the gene	RFLP analysis requires southern blotting for detection of fragments
A monomorphic DNA segment is	A segment of DNA that exists in many forms in the population	A segment of DNA that controls a single gene function	A segment of DNA inherited in a dominant fashion	A segment of DNA shared by over 99 % of the population	A segment of DNA shared by over 99 % of the population
Linkage mapping can determine the distance between which of the following pairs of DNA sequences?	AFLPs and RFLPs	Two AFLPs	Two known genes	A known gene and any type of molecualr marker, AFLP, RFLPs	A known gene and any type of molecualr marker, AFLP, RFLPs
Which of the following would be a reasonable use of an RFLP map?	Identification of the exact location of an unknown gene along the chromosome	Description of the size of the DNA fragment required to clone the region containing an unknown gene	Identification of the region in which an unknown gene is located	positional analysis of genes	Identification of the region in which an unknown gene is located
Which of these genetic markers is most likely to be highly polymorphic	An RFLP	A Microsatellite	An SNP	All other forms of markers	A Microsatellite
Why might use of microsatellite in genetic mapping studies be an advantage over RFLPs?	A microsatellite are easier to detect	Microsatellites are more abundant than RFLPs	Microsatellites with out potential alleles	Easy to detect, more abundant and have more potential alleles	Easy to detect, more abundant and have more potential alleles

how are individual chromosomes identified in chromosome sorting techniques?	Relative level of fluorescence when stained with a dye mixture	Level of charge relativ eto size	Size of the molecule	shape and orientation	Relative level of fluorescence when stained with a dye mixture
Which of these describes a contig?	A complete genomci library including overlapping clones	A complete mRNA library	A chromosome specific library of overlapping clones	An ordered genomci library	A chromosome specific library of overlapping clones
What vector would be best suited for creating a contig of bovine chromosome 10?	λ phage	A plasmid	YAC	Cos site	YAC
which of the following would not be a critical characteristic of a YAC vector?	Telomeric sequences	A gene encoding a required structural proteins	AN origin of replication	A centromere	A gene encoding a required structural proteins
In cytogenetic and linkage mapping ,After identifed small area of chromosomes which contains the gene of interset, by contig maps the positional cloning of the gene is identified by	The clone will also contain the nearest marker mapped by linkage analysis	The DNA sequence of the lone will contain an ORF	Cytological hybridization of the clone produces a different pattern	Fragmental analysis	The DNA sequence of the lone will contain an ORF
Which of these is a reasonable use for the contigs and clones generated in the process of obtaining a physical map of the human genome?	Correlation of physicla and linkage map distances to determine recombination frequency	Identification of all the known alleles of a single disease-containing gene	Insertion in to cells affected with genetic disease in the process of gene therapy	Insertional cloning	Correlation of physical and linkage map distances to determine recombination frequency
Sequencing of genomes other than humans is potentially valuable because	Disorders of cellualr function can be studies in relatively simple model systems	Comparison of gene sequences between different species can allow prediction of desase causing mutations	genes may have a similar function in other species, giving us a palce to start with fucntional analysis of similar genes	positional cloning of evolutionary genes	genes may have a similar function in other species, giving us a palce to start with fucntional analysis of similar genes
A unit of measurement on physical maps is	kilobases	centomorgans	cytological bands	centimeters	Kilobases
A haplotype is	the set of polymorphic nucelotides dound together on a single chromosome	a genotype that is unique to non-african populations	a genotype that is only found in a single individual in a population	a set of diploid genotypes at two or more loci in an individual	the set of polymorphic nucelotides dound together on a single chromosome
Alternative splicing refers to	a differnec in the number of exons in two or more species	the production of two or more mRNAs from a single gene	regulation of two different genes by a single regulatory element	post-translational modification of the cleavage site of receptor proteins	the production of two or more mRNAs from a single gene
Tagging SNP are designed	to capture the majority of genetic variation in haplotype	reducing the number of sites that must be tested in a genomic scan and the genetic variation in haplotype	positional cloning	functional cloning	reducing the number of sites that must be tested in a genomic scan and the genetic variation in haplotype
If the map distane between genes A and B is 10 map units and the map distance between genes B and C is 25 map units, what is the map distance between genes A and C	15 map units	35 map units	Either 15 map units or 35 map units depending on the order of the genes.	The map distance between A and C cannot be predicted from the data	Either 15 map units or 35 map units depending on the order of the genes.
The pairwise map distances for four linked genes are as follwos: A-B=22 m.u., B-C= 7 m.u., C-D= 9 m.u., A-D=20 m.u., A-C= 29 m.u. What is the order of these four genes?	ABCD	ADBC	ABDC	BADC	ADBC
The measured distance between genes D and E in a two point test cross is 50 map units. What does this mean in physical terms?	D and E are on different pairs of chromosomes	D and E are linked and exactly 50 map units apart	D and E are linked and at least 50 map units aprat	Either a or c.	Either a or c.
plasmid vectors for cloning	can generally accommodate larger inserts than pahge vectors can	grow with in bacteria, and are present in bacterial colonies on an agar plate	can accommodate inserts of over 100 kilobases	burst bacteria and form plaques on a "lawn" bacteria	grow with in bacteria, and are present in bacterial colonies on an agar plate
Simple tandem repeat polymorphisms in humans are most useful for	solving criminal and paternity cases	reconstructing the relationships of humans and chimps	estimating matches for blood transfusions	estimating relationships of humans and neanderthals	solving criminal and paternity cases
Mapping of human chromosomes	has been restricted to the sex chromosomes because of small family sizes	proceeded much more successfully as large numbers of DNA markers became available	has shown that there are more genes on the Y than on the X chromosomes	has demonstrated that almost all of the DNA is invloved in coding of genes	proceeded much more successfully as large numbers of DNA markers became available
RFLP analysis is a technique that	uses hybridization to detect specific DNA restriction fragments in genomic DNA	is used to determine whether a gene is transcribed in specific cells	measures the transfer frequency of genes during conjugation	is used to amplify genes for producing useful products	uses hybridization to detect specific DNA restriction fragments in genomic DNA
The "sticky ends" generated by restriction enzymes allow:	Selection for plasmids lacking antibiotic resistance	easy identification of plasmids which carry an insert	replciation of transfer RNA with in bacterial cells	Pieces of DNA from different sources to hybridize to each other and to be joined together	Pieces of DNA from different sources to hybridize to each other and to be joined together
Simple sequence length polymorphisms	can be satellite DNA	are typed by oligonucleotide hybridization	are not mini satellite	are used in genetic mapping and a type of mini and micro satellite	are used in genetic mapping and a type of mini and micro satellite
The shotgun method	is used in analyzing transcriptomes	requires computers	is normally used with large genomes	is more accurate than clone contig method	requires computers
Microsatellites	are tandemly repeated sequences	are not very common in the human genome	are usually longer than 200 bp	are normally found at the end of chromosomes	are tandemly repeated sequences
Transcriptomes	Consist of RNA	consist of DNA	do not change	consist of proteins	consist of RNA
RFLPs	are used to determine the position of restriction sites in a genome	are used in physical mapping	are used in genetic mapping	usually occurs as multiple alleles in a genome	are used in genetic mapping
		web lab			
The term used to refer something performed on computer or computer simualtions	dry lab		invitro	insilico	insilico
what are the applicatiosn of genomics	predicting disease risk at	Genome wide dissociation analysis annotations	comparative analysis at the	predicting disease risk at	predicting disease risk at
which of the following is used for determining the location of secific genes with in a genome	Genomcis	genetic linkage and chi square	cloning	proteomics	Annotations
The tendency of genes to be transmitted togethr exemplies ------and a unit of gentic map distance is given by------	chromosomal interference and chi square		recombination and centimorgan	genetic linakge and centimorgan	genetic linakge and centimorgan
A visible site of crossing over is known as --------and recombination between genes can be modulated by------	centromere, chromosomal interference	centromere and map units	chaisma and chromosomal interference	chiasma and chi square	chaisma and chromosomal interference
An underlying assumption about linkage can be given by------and a statistical test of linkage is called-------	chromosomal interference and chi square	null hypothesis and chi square	null hypothesis and chromosomal interference	chi square and null hypothesis	null hypothesis and chi square
Unit III					
Any DNA molecule that has the ability to replicate in an appropriate host cell, to which the desire gene are integrated for cloning is called	Plasmid	linker	Vector	adapter	Vector
Which of the following statement is true	A vector should not have an origin of replication	a vector doesnot have selectable markers	a vector should have unique restriction sites	vector should not replicate	a vector should have unique restr
Expression vectors differ from a cloning vector in having	an origin of replication	suitable marker gene	unique restriction sites	control elements	control elements
Extra chromosomal, double stranded, circular DNA molecule present in bacteria which is widely used as vector is called	phagemid	cosmid	plasmid	bacterial vectors	plasmid

In pBR 322, pBR stands for	plasmid bacterial recombination	plasmid bacterial replication	plasmid boliver and Rodriguez	plasmid baltimore and Rodriguez	plasmid boliver and Rodriguez
Phage M13 vectors are widely used for	obtaining single stranded copies of cloned DNA suitable for DNA sequencing	obtaining double sranded copies of cloned DNA suitable for DNA sequencing	obtaining fragments of cloned DNA suitable for DNA sequencing	obtaining double stranded copies of cloned DNA suitable electrophoresis	obtaining single stranded copies of cloned DNA suitable for DNA sequencing
Cosmid is a plasmid with	a minimum of 250 bp of lambda DNA that includes Cos site	a minimum of 250 bp of M13 DNA that includes Cos site	a minimum of 100 bp of lambda DNA that includes Cos site	a minimum of 250 bp of T4 phage DNA that includes cos site	a minimum of 250 bp of lambda DNA that includes Cos site
Vectors designed to replicate in cells of two different species are called	phasmids	transfer vectors	shuttle vectors	phagmids	shuttle vectors
Autonomously replicating sequences (ARS) is a characteristic feature of	plasmid vectors	phage vectors	E.coli vectors	yeast vectors	yeast vectors
Which of the following are vectors for animals	SV 40 vectors and bovine papillomavirus vectors	CMV vectors and Gemini vectors	lambda phage and M13 phage	pUCvector	SV 40 vectors and bovine papillomavirus vectors
The vectors cpmmonly used for sequencing human genome	Yeast artificial chromosome (YA	plasmid	CMV vectors	M13 Vectors	Yeast artificial chromosome (YA
The most common vectors for plants are	SV 40 vectors and bovine papillomavirus vectors	Cauliflower mosaic virus (CMV vec	lambda phage and M13 phage	T4 phage vectors	Cauliflower mosaic virus (CMV vectors) and Gemini Vectors
Plasmids are transferred between bacteria by way of a	conjugator	pilus	transposer	snorkel	conjugator
Small solid supports onto which are spotted hundreds of thousands of tiny drops of DNA that can be used to screen gene expression	DNA microarrays	cloning library	southern blot	western blot	DNA microarrays
The first dimension of separation for two-dimensioanl electrophoresis is based on	molecular mass	isoelectric point	folding	binding	isoelectric point
The second dimension of sepration for two dimensional electrophoresis is based on	molecualr mass	isoelectric point	folding	binding	molecualr mass
Microarray analysis has allowed scientists to view what phenomenon?	the genome sequence in a cell	The cDNA of a cell	The RFLPs of a cell	The expression of specific genes in a cell	The expression of specific genes in a cell
A virus is made up of	protein coat and mitochondria	protein coat and nucleic acid	nucleic acid and cell membrane	protein coat and mitochondria	protein coat and nucleic acid
The protein coat of viruses that encloses the genetic material is called	virion	peplomers	capsomeres	capsid	capsid
An icosahedral capsid consists of	pentagonal capsomeres	hexagonal capsomeres	triangular pentagonal and nexagonal capsomeres	pentagonal capsomeres and hexagonal capsomeres	pentagonal capsomeres and hexagonal capsomeres
viral envelope is made up of	lipids	proteins	lipids and proteins	lipids and glycoproteins	lipids and glycoproteins
The DNA microarray technology that indicates which genes are transcribed is called	DNA variation screening	gene expression profiling	microarray comparative hybridization	antisense	gene expression profiling
The DNA microarryas technology that tracks deletions and amplifications of sepcific DNA sequence is called-----	DNA variation screening	gene expression profiling	microarray comparative genomic hybridization	antisense	microarray comparative genomic hybridization
Genome wise gene expression analysis is performed using	DNA microarrays	Northern analysis	Real time PCR	RT-PCR	DNA microarrays
In standard microarrays, the probes are attaached via surface engineering to a solid surface by a ------to a chemical matrix	metallic bond	covalent bond	ionic nond	aromaticity	covalent bond
The two Cy-labelled cDNAsamples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophors after excitation with a -----beam of a defined wavelength	Helim	laser	X-rays	UV-rays	laser
Microarray data processing uses-----	Generative topographic map	artificial intelligence	artificial neural network	self organizing map	self organizing map
DNA microarrays can be used to detect DNA and cDNA after treatment with------ enzyme	thymidine kinase	reverse transcriptase	integrase	glycolase	reverse transcriptase
which technique can demosntrate mRNA expression levels?	Southern blotting	EST sequencing	RT-PCR	Serial analysis of gene expression	Serial analysis of gene expression
Which of the followings are true for micorsatellite markers?	Have been found in the mitochodnrial genome never in the host genome	potential mRNA markers for cancer screening	can be used as markers in the studies of genetic linkage	they are simple repeats	can be used as markers in the studies of genetic linkage
RNA microarrays	Make use of SNPs	utilize microsatellites	monitor 1000s of geen simultaneously	monitor 100s of genes simultaneoulsy	monitor 1000s of geen simultaneously
A map unit refers to	The relative distance between genes on a chromosomes	the chromosomes that exchange part	the percentage of recombination	relative distance between genes and percentage of recombination	relative distance between genes and percentage of recombination
If computers were not able to access the entire genomic sequence of an organism, which of the following technique might allow determination of the gene sequence that encodes a particular protein?	PCR amplification of related gene sequences	hybridisation of a genomic library with a degenerate probe	production of synthetic peptides	checking in PAGE	hybridisation of a genomic library with a degenerate probe
which of these concerns would apply to functional protein microarrays but not antibody movement?	Antibodies recognize relatively short peptide sequences from a protein, and the presence of the same peptide sequence in several proteins is possible	Three dimensional structure of a protein is essential for correct function, and the process of microarray preparation may interfere with this structure	certain functions such as structural support of a cell, may not be amenable to analysis on a microarray	some concerns are applied to preotin microarray but not antibody microarray	certain functions such as structural support of a cell, may not be amenable to analysis on a microarray
How many potential ORFs are present in a DNA sequence?	One	Three dimensional structure of a pro	six	more than six	six
Linkage analysis	is used in physical mapping	is primarily used with prokaryotes	requires linked DNA markers	is based on recombination frequencies	is used in physical mapping
Radiation hybrids	are human cell lines	are produced by irradiation of UV light	can hold large piece of chromosomal DNA	have been used in mapping the yeast genome	are produced by irradiation of UV light
Physical mapping	requires large number of organisms	utilizes genomic libraries	detects polymorphic DNA sequences	uses RFLPs	uses RFLPs
A genome marker	must occur in multiple alleles	must be at least 1000 bp long	must be at least 50 bp long	msut be a unique DNA sequences	must be a unique DNA sequences
microarray analysis can be performed in order to	determine th activity of genes in a treatment sampl compared to an untreated control	to determine the postion of the gene	to determine order of the gene	to examine the genomic library	determine th activity of genes in a treatment sampl compared to an untreated control
Normalization of data is used to----	standardize data to a comparison value valid for all measured data points	To anable the analysis of sequencing data	to position the gene in the chromosome	to order the sequence of genes in a chromosome	standardize data to a comparison value valid for all measured data points
Normalization of microarray is carried out using different techniques one among them is	total linkage	average linkage	hierarchicial	total intensity	total intensity
Clustering of microarray data are carried out using	total intensity	regression techniques	ratio statistics	single-linkage	single-linkage
A p-value of < 0.05 means that.....	there is no statistically significant between two repeatedly measured samples using 5 % cut off	there is a statistically significant between two repeatedly measured samples using 5 % cut off	Null hypothesis is accepted	Devoid of any hypothesis	there is a statistically significant between two repeatedly measured samples using 5 % cut off
with microarray, it is possible	to look the expression of only asingle gene at a time	to look globally at the expression of all genes	to pinpoint all of the cis-regulatory sequences for all genes in one	distinguish between transcription rate and RNA stability	to look globally at the expression of all genes
Key findings from microarray experiments	Should be validated by other means	can always be accepted as being accurate	are impossible to verify using RT-PCR	never need to be accompanied by statistical analysis	Should be validated by other means

		k-means of clustering			
The follwoing is NOT a data mining method usually used for microarray analysis	Hierarchical clustering		principle component analysis	self organizing maps	principle component analysis
		there are no incorrect ways to	clustering of data		the only acceptable way of
	there is only one	cluster data	should be avoided	the only acceptable way	analyzing microarray data is by
which of the follwoing statements is true?	appropriate way to cluster microarray data		whenever possible	of analyzing microarray data is by cluster analysis	cluster analysis
	decrease	increase	not decrease	not stay the same	not decrease
If you change the correlation threshold for " cutting the tree" in hierarchical clustering from 0.8 to 0.5, you can certain that the number of genes per cluster will					
if you change the correlation threshold for " cutting the tree" in hierarchical clustering		increase			
from 0.95 to 0.2 , the number of clusters is likely to	decrease		stay the same	not stay the same	decrease
Adeno virus genome is-------in type	dsDNA	ssDNA	circular	circular RNA	dsDNA
adenoassociated virus genome is------in type	dsDNA	ssDNA	circular	circular RNA	ssDNA
Alpha virus genome type is-------	dsDNA	ssRNA	circular	circular RNA	ssRNA
Herpes virus genome type is----	dsDNA	ssDNA	circular	circular RNA	dsDNA
Genome size of Adeno virus is	39-40 kb	10 kb	5 kb	100 kb	39-40 kb
Genome integration nature of adeno virus	non-integration	integrating	transient	non-transient	non-integration
Expression pattern of adenovirus is----	transient	long lasting	potential long lasting	virulent	transient
Capsid of the herpes virus is	hexagonal	icosahedral	hexagonal	polygonal	icosahedral
Genome size of vaccinia virus is----	280 kb	300 kb	150 kb	120 kb	280 kb
Genome size of alpha virus is	60- 70 nm	12 nm	10 nm	40 nm	60- 70 nm
virion polymerase of Herpes virus is	Negative	Positive	Neutral	bivalent	Negative
Unit IV					
	Different forms of the same	up to about a hundred different	proteins with similar	resolving power of crude	resolving power of crude
In two dimensional gel electrophoresis	protein will tend to migrate at the same position	proteins can be distinguished from each other	functions are locate near to each other	separation of proteins is very high	separation of proteins is very high
	Proteins separated by isoelectric focusing cannot	Proteins separated by isoelectric focusing can be tested for biological		The separation of proteins by isoelectric focusing is only based on size	Proteins separated by isoelectric focusing can be
Which of the follwoing statements about isoelectric focusing is correct?	be tested for biological activity	activity	The separation of proteins by isoelectric focussing is only based on charge		tested for biological activity
		The individual bands become		the individual bands	
The first step in two dimensional gel electrophoresis generates a seris of protein bands	Proteins with similar	stained so that isoelectric focus		become visualized by	Proteins with similar isoelectric
by isoelectric focussing. In the second step, a strip of this gel is turned 90 degrees,	isoelectric points become	pattern can be visualized	the individual bands undergo	interacting with protein-	points become further separated
placed on the another gel containing SDS and electric current is again applied. in this	further separated according		a second, more intense	specific antibodis in the	according to their molecular
second step	to their molecular weight		isoelectric focussing	second gel.	weight
Name the surfactant used in 2D gel electrophoresis process	thiourea	7M urea	100 mM DTT	sulfobetaine	sulfobetaine
	comassie brilliant blue	congo red			comassie brilliant blue R250
Staining used in 2D gel electrophoresis	R250 staining		methylene blue	crystal violet	staining
The first dimension separation of proteins is based on	isoelectric focussing	gradient pH	amphipathic	hydrophobic	isoelectric focussing
		gradient pH			
The second dimension of sepration for two dimensional electrophoresis is based on	molecular mass		zwitter ion	amphipathic form	molecular mass
		Limited by pH range	Analysis and quantification	not for hydrophobic	
Advantages of two Dimensional gel electrophoresis	Good resolution of proteins		are difficult	proteins	Good resolution of proteins
		detection of post translational	Analysis and Quantification		Analysis and Quantification are
Disadvantages of 2D gel electrophoresis	Good resolution of proteins	modifications	are difficult	Not limited by pH range	difficult
	Proteins are denatured by				Proteins are denatured by the
In an SDS-PAGE	the SDS	not for hydrophobic proteins	smaller proteins migrate more	limited pH range	SDS
		using electron microscope only	measuring their molecular		
Proteins can be visualized directly in gels by	Staining them with the dye		weight	limited pH range	Staining them with the dye
	treated with a reducing	fractioned by electrophoresisthen	treated with an oxidizing	separated by their	
	agent and then anionic	treated with an oxidising agent	agent and then with anionic	difference in pH	treated with a reducing agent
	detergent followed by	follwoed by anionic detergent	detergent followed by		and then anionic detergent
In SDS-PAGE, the protein sample is first	fractionation by electrophoresis		fractionation by electrophoresis		follwoed by fractionation by eelctrophoresis
		relative content of negatively			
	Relative content of	charged residue		relative content of positively and negatively	relative content of positively
In isoelectric focussing, proteins are separated on the basis of their	positively charged residue		Size of the molecule	charged residue	and negatively charged residue
		Gel filtration chromatography	combining information from		combining information from
The subunit molecular weight as well as the number of subunits in the quaternary			SDS PAGE and gel filtration		SDS PAGE and gel filtration
structure can be determined	SDS-PAGE electrophoresis		chromatography	Isoelectric focussing	chromatography
	positively charged side	molecular weight	negatively charged side	different isoelectric	molecular weight
Proteins are separated in an SDS-PAGE experiment on the basis of their	chains		chains	points	
		A greater number of theoretical			
	More versatile in adjusting	plates	Applicable to wider range of	both liquid and gases	both liquid and gases samples
which of the following is not an advantage of HPLC over GC?	separation parameters		analytes	samples can be analysed	can be analysed
Analyte that are extremely hydrophobic (or) incompatible with water, a mobile phase	Reversed phase	Normal Phase chromatography	Non-aquoeus reversed phase	Aqueous Normal phase	Reversed phase
that contains no water is used. The Separation is termed as	chromatography		chromatography	chromatography	chromatography
		HPLC			HPLC
The elution is depends on "Boiling point" of analytes (very low B.P) observed in	PTGC		OTGC	CTGC	
Prevention of back flow during pump reciprocation done by	Adapter	Back diffusion	Check Value	Degassing	Adapter
Sieving is the mechanism of separation in	CEC	SFC	SEC	TIC	SFC
If the two analytes present in the sample are of similar chemical type; choose the correct		Adsorption chromatography			
chromatographic technique for	Partition chromatography		Chiral Chromatography	Ion pair chromatography	Partition chromatography
		It requires the crystallization of	Mass spectrometry involves		
	Mass spectrometry can be	proteins	ionized molecules in the		Mass spectrometry can be used
which of the follwoing statements about the use of mass spectrometry in protein	used to investigate post		liquid phase	It separates the proteins	to investigate post translational
investigation is correct?	translational modifications			on the basis of the size	modifications
The separation technique of charged molecules under the influence of electric current is					
called	Colony hybridization	electrophoresis	dot blot technique	western blotting	electrophoresis
For protein detection, most commonly used probe is	antibody	lectin	antigens	interferons	antibody
In isoelectric focussing proteins are separated on the basis of their	relative content of	relative content of negatively	size	relative content of	
	positively charged residue	charged residue only		positively and negatively	
	only			charged residues	relative content of positively
					and negatively charged residues
Isoelectric focusing is a common method for separation of proteins. In isoelectric					
focussing proteins are separated in the following gradient	A pH gradient	A salt gradient	A temperature gradient	A density gradient	A pH gradient
				A combination of SDS-	
To determine the molecular mass of the subunits and the number of each subunit present,			Gel filtration	PAGE and isoelectric	A combination of SDS-PAGE
the follwoing can be utilized.	SDS-PAGE	Isoelectri focussing	Chromatography	focussing	and isoelectric focussing
Which of the following protein features is utilized in the first dimension of 2D gel				The number of sulphur	
electrophoresis	Molecular weight	Isoelectric point	Protein conformation	bridges	Isoelectric point
Proteins are chains of amino acids held together by:				Planar dihedral bonds	
	Planar dihedral bonds	Peptide bonds that are formed	Peptide bond that are formed	formed between	Peptide bonds that are formed
	formed between C alpha	between the carboxyl group of one	between side chains and	subsequent amino acids in	between the carboxyl group of
	atoms in a reaction the	amino acid and tthe amino group of	central C alpha atom of each	a reaction that consume	one amino acid and tthe amino
	result in the release of water	the next amino acid	amino acid	water	group of the next amino acid
Molecules with higher solubility will migrate to	greater than Rf	lesser than Rf	equal to RF	none of the above	greater than Rf
Usually low moleculalr weight compounds are separated using …….. chromatography	adsoprtion	partition	column	thin layer	partition
Impurities present in paper are removed by washin with	1 N HCl	0.1 N HCl	0.01 N HCl	0.001N HCl	0.001N HCl
In paper chromatography, amino acids are viewed in purple or blue by spraying	bromine water	ninhydrin	methanol	ethanol	ninhydrin
The stationary phase used in column chromatography is known as	filter paper	gradient	gel matrix	none of the above	gel matrix
Gel matrix cellulose has ……. units.	β-1,4 linked galactose	β-1,4 linked fucose	β-1,4 linked arabinose	β-1,4 linked glucose	β-1,4 linked glucose
Gel matrix dextran has ……. units.	α-1,6 linked gluc ose	α-1,6 linked galact ose	α-1,6 linked fuc ose	α-1,6 linked arabinose	α-1,6 linked gluc ose
Gel matrix agarose has ……. units.	D-glucose	D- galact ose	L-fucose	L-arabinose	D- galact ose
Unit V					
Which of the following statements are not true	Transposons can be directed	Proteomes consist of proteins	Proteomes consist of proteins	Proteins can be identified b	Transposons can be directed to d

Protein-coding genes can be identified by	Transposon tagging	ORF scanning	Zoo-blotting	Nuclease S1 mapping	ORF scanning
The function of genes can be determined by	Gene inactivation	Homology search	Exon trapping	Zoo-blotting	Homology search
Expression of genes can be analyzed by	Microarrays	Northern analysis	Southern analysis	Protein profiling	Northern analysis
Polypeptides	Can fold into a double helix	Can have a tertiary structure	Can contain phosphate	Are synthesized in the nucl	Can have a tertiary structure
Reporter genes	Indicate the presence of stres	Are used to characterize proteomes	Are all of bacterial origin	Are used to delineate regul	Are used to delineate regulatory s
Microarrays	Are used for analysis of trans	Are made of glass	Contain RNA sequences	Are smaller than DNA chi	Are used for analysis of transcrip
ORF scanning	Is used to find exons	Is used to find intergenic sequences	Is used to find gene homologie	Is used to find protein-codi	Is used to find protein-coding ge
Phage display	Is used to analyze transcripto	Can identify protein-protein interacti	Requires a clone library	Requires a cDNA library	Can identify protein-protein inter
A codon bias	Is used in genome mapping	Is found in functional RNAs	Is found in intergenic regions	Is used to identify genes	Is used to identify genes
The human genome	Is the largest genome known	Was the first completely sequenced g	Contains more than 40,000 ge	Is divided into 24 chromos	Is the largest genome known to d
Mass spectrometry is used in	Transcriptome analysis	Proteome analysis	Protein seperation	Protein identification	Proteome analysis
Protein-protein interactions can be identified by	Phage display	Microarrays	Hierarchical clustering	Isoelectric focusing	Phage display
Genes can be altered or replaced by	Transposon tagging	RNA interference	DNA interference	Homologous recombination	
Centromeres	Contain satellite DNA	Contain many tightly packed genes	Contain no histones	Function in DNA replicatio	Contain satellite DNA
Most sequences in the human genome belong to	Genes	Pseudogenes	Gene fragments	Interspersed repeats	Interspersed repeats
Pseudogenes	Are nonfunctional genes	Are expressed genes	Are incomplete genes	Are complete genes	Are nonfunctional genes
Telomers are	minisatellites	found in bacteria	usually less than 1000 bp in si	found in middle of chromo	minisatellites
Histones	Are only found in eukaryotes	Are not part of chromatin	Are not found in telomeres	Are not found in centerome	Are only found in eukaryotes
Introns	Do not contain any functiona	Are also found in bacterial genomes	Are only found in coding regio	Occur in almost every gene	Are also found in bacterial genor
In the yeast two-hybrid system, which of the following statements is accurate: A reporter g	is fused to the activation don	is fused to the DNA binding domain	Requires the presence of Histi	Is expressed only if the test	Is expressed only if the tested pro
In phage display, what are the two main reasons for testing fusion of the protein library to	is fused to the activation don	The library protein may be secreted	The different coat proteins are	The degree of difficulty in	is fused to the activation domain
Which of the following describes an advantage of the yeast two-hybrid method for analysi	is fused to the activation don	The assay can screen for interaction	The assay only detects direct a	The assay secretes proteins	The assay can screen for interacti
A yeast two-hybrid assay is being used to monitor interaction between two proteins, X an	The DBD-X and AD-Y fusic	The DBD-X fusion protein is able to	X and Y are proteins that can	The yeast cells will be abl	The DBD-X and AD-Y fusion pr
Which of the following statements about amino acids is correct?	Amino acids are classified ac	Amino acids are uncharged at neutra	Amino acids in proteins are ma	Twenty four amino acids a	Amino acids are classified accor
Which type of bonding is responsible for the secondary structure of proteins?	Disulphide bridges between	Hydrogen bonding between the C=O	Peptide bonds between amino	Salt bridges between charg	Hydrogen bonding between the C
Which term below best defines the 'quaternary structure' of a protein?	The arrangement of two or m	The folding of the polypeptide backb	The interaction of amino acid	The sequence of amino aci	The arrangement of two or more
Which of the following statements about collagen is correct?	Collagen contains a high pro	Collagen is a globular, intracellular	Post-translational modificatior	The structure of collagen c	Collagen contains a high proport
Which of the following statements about haemoglobin is correct?	2,3-Bisphosphoglycerate (BI	Deoxygenated haemoglobin has a hi	Haemoglobin has a higher affi	One molecule of haemoglo	Deoxygenated haemoglobin has
Which of the following amino acids is mostly likely to disrupt an alpha helix?	Proline	Leucine	Glycine	Valine	Proline
Which of the following most accurately describes how secondary structures in proteins are	Through ionic bonds operati	Through covalent bonds joining diff	Through hydrogen bonds betw	Through hydrogen bonds j	Through hydrogen bonds joining
Which amino acid can form disulphide bonds?	Glycine	Proline	Glutamate	Cysteine	Cysteine
Which of the following pairs of amino acids might contribute to protein conformation by	Glycine and leucine.	Glutamate and lysine.	Phenylalanine and tyrosine.	Lysine and arginine.	Glutamate and lysine.
Which of the following best describes a protein domain?	The α-helical portion of a pr	A discrete region of polypeptide chai	The β-pleated sheet portion of	A feature that rarely occurs	A discrete region of polypeptide
As haemoglobin binds oxygen molecules, its affinity for oxygen increases, driving the bin	Catalysis	Saturation	Allostery	Isomerism	Allostery
Which of the following statements correctly describes the behaviour of the haemoglobin p	The haemoglobin protein dis	The haemoglobin protein lacks a hae	Neighbouring haemoglobin pr	The haemoglobin protein p	Neighbouring haemoglobin prote
Which of the following statements is wrong?	UV absorption is attributable	UV spectra provide information abo	IR absorption is attributable to	NMR spectrometers use ra	IR absorption is attributable to tr
Which is the correct order of increasing wave number of the stretching vibrations of (1) C	(4) < (3) < (2) < (1)	(3) < (4) < (2) < (1)	(3) < (4) < (1) < (2)	(4) < (3) < (1) < (2)	(3) < (4) < (1) < (2)
How many signals does the aldehyde $(CH_3)_3CCH_2CHO$ have in ^1H NMR and ^{13}C NMR s	five ^1H signals and six ^{13}C s	three ^1H signals and four ^{13}C signals	five ^1H signals and four ^{13}C si	three 1H signals and six 13	three ^1H signals and four ^{13}C sig
Which of hydrogens a-d in the following molecule gives a triplet signal in a normal ^1H NI	hydrogen a	hydrogen b	hydrogen c	hydrogen d	hydrogen c
Which hydrogen of 1-chloropent-2-ene shows the largest chemical (downfield) shift in its	the H on C1	the H on either C2 or C3	the H on C4	the H on C5	the H on either C2 or C3
Which carbon of (a)-(d) of hex-3-en-2-one shows the largest (most downfield) chemical sl	c1	**c2**	c4	c6	**c2**
Which of the following statements regarding IR spectroscopy is wrong?	Infrared radiation is higher i	Infrared spectra record the transmiss	Molecular vibrations are due t	Infrared spectra give infor	Infrared radiation is higher in en
Which of the following statements regarding NMR spectroscopy is wrong?	NMR signals towards the le	Chemical shifts are larger when the f	Chemical shifts are larger wh	A hydrogen signal splits i	Chemical shifts are larger when
Which of the following statements regarding mass spectrometry is wrong?	In a normal mass spectromet	Only cations can be detected by a no	A compound whose molecules	Molecular ion peaks alway	Molecular ion peaks always have
Which of the following statements is wrong?	A conventional mass spectro	A conventional mass spectrometer d	Conventional mass spectromet	A mass spectrum does not	A conventional mass spectromet
Absorption of radiation in the UV range attributable to n→π* electronic transitions is cha	Aromatic hydrocarbons.	Unsaturated carbonyl compounds.	Non-conjugated polyenes.	Conjugated polyenes.	Unsaturated carbonyl compounds
Which of the following statements is wrong.	The wavenumber of a band	Water is a good solvent for recording	Water is a good solvent for rec	Hydrogen bonding in hydr	Water is a good solvent for recor
Which is the correct order of increasing wave number of the stretching vibrations of (1) C	(1) < (2) ≈ (3) < (4)	(4) < (3) ≈ (2) < (1)	(3) < (4) ≈ (2) < (1)	(1) < (4) ≈ (2) < (3)	(1) < (4) ≈ (2) < (3)
How many signals does the unsaturated ketone $(CH_3)_2CHCH_2C(O)CH=CH_2$ have in ^1H N	five ^1H signals and six ^{13}C s	six ^1H signals and six ^{13}C signals	six ^1H signals and seven ^{13}C si	five ^1H signals and seven	six ^1H signals and six ^{13}C signal
Which of the following statements in the context of ^1H NMR spectroscopy is true?	Arene C-H chemical shift (δ	Arene C-H chemical shift (δ) values	Arene C-H signals are always	Arene C-H signals are alw	Arene C-H chemical shift (δ) va
Which carbon of (a)-(d) of hex-3-en-2-one has the smallest (most upfield) chemical shift i	**c1**	**c2**	c4	**c6**	**c6**
Which of (a)-(d) indicates the multiplicities for hydrogens on C1, C3, and C4 of butanone	Hs on C1, singlet; Hs on C3	Hs on C1, singlet; Hs on C3, triplet	Hs on C1, singlet; Hs on C3, c	Hs on C1, triplet; Hs on C	Hs on C1, singlet; Hs on C3, qua
Which of (a)-(d) indicates the correct order of carbon chemical shifts of the four carbons o	C_{Me} < C2 < C3 < C1	C_{Me} < C3 < C2 < C1	C_{Me} < C2 < C1 < C3	C_{Me} < C1 < C2 < C3	C_{Me} < C2 < C3 < C1
Which of the following statements regarding electron-impact mass spectrometry is true?	Samples need isotopic label	The base peak is formed by loss of c	Compounds must have a funct	A meaningful mass spectr	A meaningful mass spectrum ca
Which of the following statements regarding mass spectrometry is false?	The base peak of a simple ke	The molecular ion of carbonyl comp	The molecular ion peak of som	Structurally isomeric alkar	Structurally isomeric alkanes ca
Northern blots probes are	DNA and RNA	DNA	RNA and protein	RNA	RNA