SEMESTER IV

17BTU403 GENOMICS AND PROTEOMICS 4H - 4C

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

Scope: Genomics and Proteomics deals with the genome and proteome structure and function in living system.
Objective: This paper will enable the students to learn the basics and lay strong foundation in understanding the genome and protein structure and functions.

UNIT-I

Introduction to Genomics, Gene and Pseudogenes, Gene structure, DNA sequencing methods – manual and automated: Maxam and Gilbert and Sangers method. Pyrosequencing, Genome Sequencing: Shotgun and Hierarchical (clone contig) methods, Computer tools for sequencing projects: Genome sequence assembly software.

UNIT-II

Managing and Distributing Genome Data: Web based servers and software for genome analysis: ENSEMBL, VISTA, UCSC Genome Browser, NCBI genome. Selected Model Organisms' Genomes and Databases.

UNIT-III

Genomic mapping: Genetic markers – VNTR, mini and micro satellites, STS, SNPs, ESTs. Types of genome maps, Mapping techniques – Physical and genetic mapping, Map resources, Practical uses genome maps.

UNIT-IV

Introduction to protein structure, Chemical properties of proteins. Physical interactions that determine the property of proteins. Short-range interactions, electrostatic forces, van der Waal interactions, hydrogen bonds, Hydrophobic interactions. Determination of sizes -Sedimentation analysis, gel filteration, Native PAGE, SDS-PAGE. Determination of covalent structures – Edman degradation.

UNIT-V

Introduction to Proteomics, Analysis of proteomes. 2D-PAGE. Sample preparation, solubilization, reduction, resolution. Reproducibility of 2D-PAGE. Mass spectrometry based methods for protein identification. De novo sequencing using mass spectrometricdata.

References

1. Benjamin Lewin, (2006).*Genes IX*. Johns and BartlettPublisher.
2. Primrose, S.B. (1987). *Modern Biotechnology* (2nd ed.). BlackwellPublishing.
3. Glick, B.R., Pasternak, J.J., & Patten, C.L.(2010). *Molecular Biotechnology: Principles and Applications of Recombinant DNA* (4th ed.).
4. Sambrook & Russell (3rd ed.). (1989). *Molecular Cloning: A Laboratory Manual* (Vols. 1to3). Cold Spring Harbor Laboratory Press
5. Primrose, S.B., Twyman, R.M. & Old, R.W. (2001). *Principles of Gene Manipulation* (6th ed.). Blackwell Science.
6. Snustad, D.P., &Simmons, M.J. (2009). *Principles of Genetics* (5th ed.). John Wiley and Sons Inc.
7. Klug, W.S., Cummings, M.R., & Spencer, C.A. (2009). *Concepts of Genetics* (9th ed.). Benjamin Cummings.
8. Russell, P. J. (2009). *iGenetics- A Molecular Approach* (3rd ed.). Benjamin Cummings.
9. Glick, B.R., & Pasternak, J.J. (2003). *Molecular Biotechnology- Principles and Applications of recombinant DNA*. Washington: ASMPress.

KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University Established Under Section 3 of UGC Act 1956)

Coimbatore – 641 021.

LECTURE PLAN
DEPARTMENT OF BIOTECHNOLOGY

STAFF NAME : T. SIVARAMAN
SUBJECT NAME : GENOMICS AND PROTEOMICS
SUBJECT CODE : 17BTU403
SEMESTER : IV
CLASS : II B.Sc., (BT)

S. No.	Lecture Duration Period	Topics to be Covered	Support Material/Page No.
		UNIT - I	
1	1	Introduction – Genes and pseudogenes, Gene Structure	T1: 74 - 91
2	1	Gene structures	T1: 74 - 91
3	1	Genome sequencing Maxam and Gilbert Method	T1: 91 - 98
4	1	Chain termination methods	R1: 1 - 36
5	1	Automated DNA Sequencing and Shotgun sequencing	R1: 1 - 36
6	1	Pyrosequencing	R1: 1 - 36
7	1	Hierarchical (clone contig) methods and computational tools	T1: 74 - 123
8	1	Genome sequence assembly software	T1: 74 - 123
	Total No. of Hours Planned For Unit I = 08		
		UNIT - II	
1	1	Web based servers and software for genome analysis	T3: 286 - 308
2	1	ENSEMBL	W1
3	1	VISTA	W2
4	1	UCSC Genome Browser	W3
5	1	NCBI genome	W4

6	1	Selected Model Organisms' Genomes and Databases	T2: 67 - 116
7	1	FlyBase/OMIM	T2: 67 - 116
	Total No. of Hours Planned For Unit II = 07		
		UNIT – III	
1	1	VNTR, mini and micro satellites	R1: 113 - 129 T3: 286 - 308
2	1	STS, SNP, ESTs.	R1: 113 - 129
3	1	Genome Mapping	R1: 98 - 99
4	1	Physical mapping	R1: 209 - 216
5	1	Genetic mapping	R1:209 - 216
6	1	Map resources	T3: 41 - 188
7	1	Practical uses - genome maps	R2: 143 - 148
8	1	Genome mapping and genetic markers overview	
	Total No. of Hours Planned For Unit III = 08		
		UNIT - IV	
1	1	Introduction to protein structure	T4: 219
2	1	Chemical properties of proteins	T4: 220 - 230
3	1	Physical interactions and properties	T2: 67 - 116
4	1	Van der Waal interactions, hydrogen bonds, Hydrophobic interactions, electrostatic forces	T4: 258 - 263
5	1	Edman degradation	T4: 264 - 271
6	1	Sedimentation analysis, Gel filtration	W5
7	1	Native PAGE, SDS-PAGE	W6
8	1	Determination of covalent structures	T4: 231 - 257
	Total No. of Hours Planned For Unit V = 08		
		UNIT - V	
1	1	Introduction to Proteomics	W7
2	1	2D-PAGE	T5: 1 - 34

3	1	Reproducibility of 2D-PAGE	T5: 1 - 34
4	1	EI-MS	T5: 35 - 74
5	1	ESI-MS	T5: 35 - 74
6	1	MALDI-MS	T5: 35 - 74
7	1	De novo sequencing using mass spectrometric data	T5: 35 - 74
8	1	Unit V revision	
9	1	ESE QP discussion	
	Total No. of Hours Planned For Unit V = 09		
	Total No. of Planned Hours = 40		

TEXT BOOKS

T1: Genomics and Cloning (East-West Press, 2004) – Kumar, HD.
T2: Introduction to Bioinformatics (Oxford University Press, Second edition, 2007) – Arthur M Lesk.
T3: Bioinformatics concepts, Skills and Applications (CBS Publishers & Distributors, Second edition, 2007) – Rastogi, SC., Namita, M and Parag, R.
T4: Biochemistry (John Wiley & Sons, 2011) – Donald Voet and Judith Voet.
T5: Proteomics (Kluwer Academic Publishers, 2002) – Timothy, P.

REFERENCE BOOKS

R1: Genomics (Bioscience Publishers, 2008) – Bhatt, S.
R2: Principles of Genome analysis (Blackwell publishing, 2003), Primrose, SB & Twyman, RM.

WEBSITES

W1: https://www.ensembl.org/index.html
W2: http://genome.lbl.gov/vista/index.html
W3: https://genome.ucsc.edu/
W4: https://www.ncbi.nlm.nih.gov/
W5: https://en.wikipedia.org/wiki/Size-exclusion_chromatography
W6: https://en.wikipedia.org/wiki/Polyacrylamide_gel_electrophoresis
W7: https://en.wikipedia.org/wiki/Proteomics

UNIT - I

SYLLABUS

Introduction to Genomics: Gene and Pseudogenes, Gene structure, DNA sequencing methods – manual and automated: Maxam and Gilbert and Sanger’s method. Pyrosequencing, Genome Sequencing: Shotgun and Hierarchical (clone contig) methods, Computer tools for sequencing projects: Genome sequence assembly software.

Genome

Genomics is an interdisciplinary field of science within the field of molecular biology. A genome is a complete set of DNA within a single cell of an organism, and as such, focuses on the structure, function, evolution, and mapping of genomes. Genomics aims at the collective characterization and quantification of genes, which direct the production of proteins with the assistance of enzymes and messenger molecules. Genomics also involves the sequencing and analysis of genomes. Advances in genomics have triggered a revolution in discovery-based research to understand even the most currently complex biological systems such as the brain. In contrast to genetics, which refers to the study of individual genes and their roles in inheritance, genomics uses high throughput DNA sequencing and bioinformatics to assemble, and analyze the function and structure of entire genomes. The field also includes studies of intragenomic (within the genome) phenomena such as heterosis (hybrid vigour), epistasis (effect of one gene on another), pleiotropy (one gene affecting more than one trait) and other interactions between loci and alleles within the genome. Advances in genomics have triggered a revolution in systems biology which facilitates the understanding of complex biological systems such as the brain. From the Greek ΓΕΝ *gen*, "gene" (gamma, epsilon, nu, epsilon) meaning "become, create, creation, birth", and subsequent variants: genealogy, genesis, genetics, genic, genomere, genotype, genus etc. While the word *genome* (from the German *Genom*, attributed to Hans Winkler) was in use in English as early as 1926,]the term *genomics* was coined by Tom Roderick, a geneticist at the Jackson Laboratory (Bar Harbor, Maine), over beer at a meeting held in Maryland on the mapping of the human genome in 1986.

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 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.

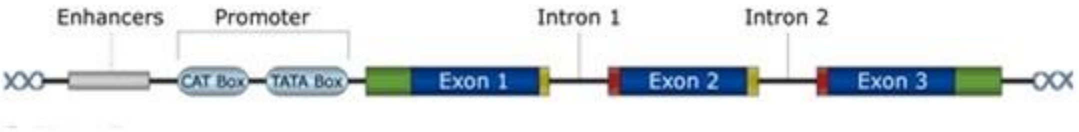
Genes consist of three types of nucleotide sequence:

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)

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.



The structural components of a gene

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.

Pseudogenes

Pseudogenes are segments of DNA that are related to real genes. Pseudogenes have lost at least some functionality, relative to the complete gene, in cellular gene expression or protein-coding ability. Pseudogenes often result from the accumulation of multiple mutations within a gene whose product is not required for the survival of the organism, but can also be caused by genomic copy number variation (CNV) where segments of 1+ kb are duplicated or deleted. Although not *fully* functional, pseudogenes may be functional, similar to other kinds of noncoding DNA, which can perform regulatory functions. The "pseudo" in "pseudogene" implies a variation in sequence relative to the parent coding gene, but does not necessarily indicate pseudo-function. Despite being non-coding, many pseudogenes have important roles in normal physiology and abnormal pathology. Although some pseudogenes do not have introns or promoters (such pseudogenes are copied from messenger RNA and incorporated into the chromosome, and are called "processed pseudogenes"), others have some gene-like features such as promoters, CpG islands, and splice sites. They are different from normal genes due to either a lack of protein-coding ability resulting from a variety of disabling mutations (e.g. premature stop codons or frameshifts), a lack of transcription, or their inability to encode RNA (such as with ribosomal RNA pseudogenes). The term "pseudogene" was coined in 1977 by Jacq et al.Because pseudogenes were initially thought of as the last stop for genomic material that could be removed from the genome, they were often labeled as junk DNA. Nonetheless, pseudogenes contain biological and evolutionary histories within their sequences. This is due to a pseudogene's shared ancestry with a functional gene: in the same way that Darwin thought of two species as possibly having a shared common ancestry followed by millions of years of evolutionary divergence, a pseudogene and its associated functional gene also share a common ancestor and have diverged as separate genetic entities over millions of years.

Pseudogenes are usually characterized by a combination of homology to a known gene and loss of some functionality. That is, although every pseudogene has a DNA sequence that is similar to some functional gene, they are usually unable to produce functional final protein products. Pseudogenes are sometimes difficult to identify and characterize in genomes, because the two requirements of homology and loss of functionality are usually implied through sequence alignments rather than biologically proven.

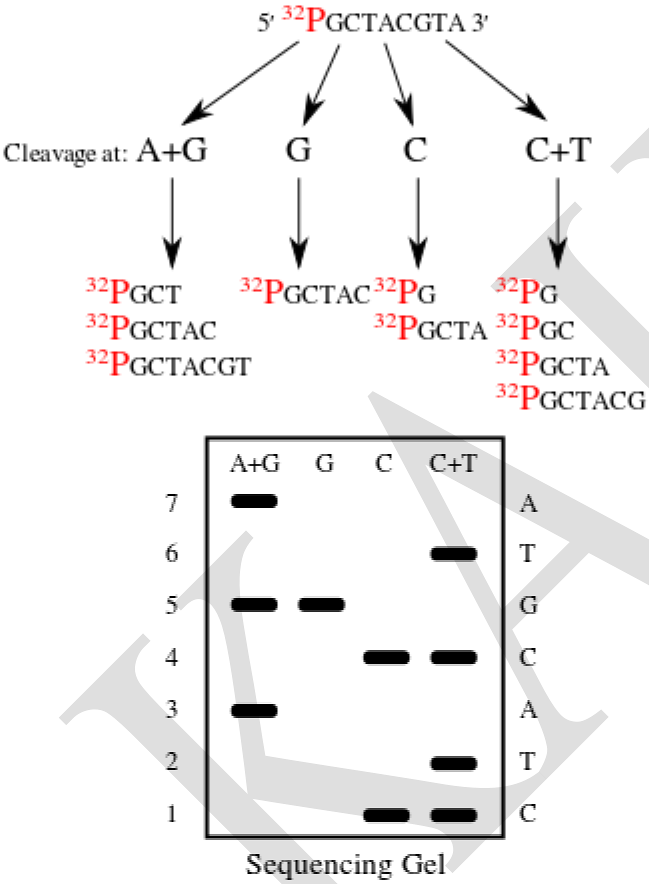
Homology is implied by sequence identity between the DNA sequences of the pseudogene and parent gene. After aligning the two sequences, the percentage of identical base pairs is computed. A high sequence identity means that it is highly likely that these two sequences diverged from a common ancestral sequence (are homologous), and highly unlikely that these two sequences have evolved independently (see Convergent evolution).

Non-functionality can manifest itself in many ways. Normally, a gene must go through several steps to a fully functional protein: Transcription, pre-mRNA processing, translation, and protein folding are all required parts of this process. If any of these steps fails, then the sequence may be considered nonfunctional. In high-throughput pseudogene identification, the most commonly identified disablements are premature stop codons and frameshifts, which almost universally prevent the translation of a functional protein product.

Pseudogenes for RNA genes are usually more difficult to discover as they do not need to be translated and thus do not have "reading frames". Pseudogenes can complicate molecular genetic studies. For example, amplification of a gene by PCR may simultaneously amplify a pseudogene that shares similar sequences. This is known as PCR bias or amplification bias. Similarly, pseudogenes are sometimes annotated as genes in genome sequences. Processed pseudogenes often pose a problem for gene prediction programs, often being misidentified as real genes or exons. It has been proposed that identification of processed pseudogenes can help improve the accuracy of gene prediction methods. Recently 140 human pseudogenes have been shown to be translated. However, the function, if any, of the protein products is unknown.

Maxam–Gilbert sequencing

Maxam–Gilbert sequencing is a method of DNA sequencing developed by Allan Maxam and Walter Gilbert in 1976–1977. This method is based on nucleobase-specific partial chemical modification of DNA and subsequent cleavage of the DNA backbone at sites adjacent to the modified nucleotides.



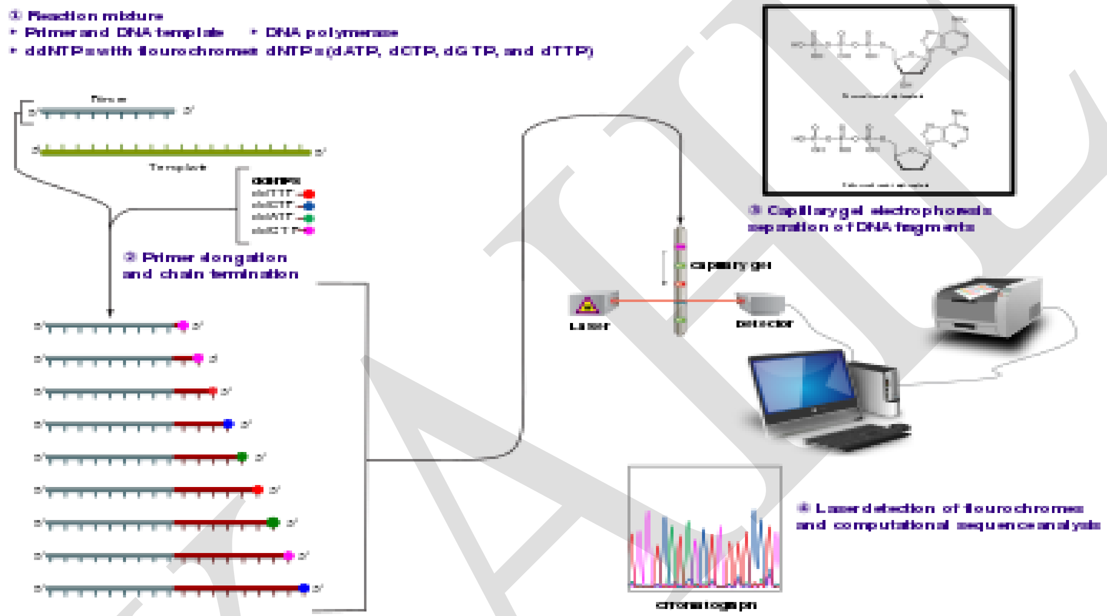
An example Maxam–Gilbert sequencing reaction. Cleaving the same tagged segment of DNA at different points yields tagged fragments of different sizes. The fragments may then be separated by gel electrophoresis. Maxam–Gilbert sequencing was the first widely adopted method for DNA sequencing, and, along with the Sanger dideoxy method, represents the first generation of DNA sequencing methods. Maxam–Gilbert sequencing is no longer in widespread use, having been supplanted by next-generation sequencing methods.

Although Maxam and Gilbert published their chemical sequencing method two years after Frederick Sanger and Alan Coulson published their work on plus-minus sequencing, Maxam–Gilbert sequencing rapidly became more popular, since purified DNA could be used directly, while the initial Sanger method required that each read start be cloned for production of single-stranded DNA. However, with the improvement of the chain-termination method (see below), Maxam–Gilbert sequencing has fallen out of favour due to its technical complexity prohibiting its use in standard molecular biology kits, extensive use of hazardous chemicals, and difficulties with scale-up. Maxam–Gilbert sequencing requires radioactive labeling at one 5′ end of the DNA fragment to be sequenced (typically by a kinase reaction using gamma-^{32}P ATP) and purification of the DNA. Chemical treatment generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). For example, the purines (A+G) are depurinated using formic acid, the guanines (and to some extent the adenines) are methylated by dimethyl sulfate, and the pyrimidines (C+T) are hydrolysed using hydrazine. The addition of salt (sodium chloride) to the hydrazine reaction inhibits the reaction of thymine for the C-only reaction. The modified DNAs may then be cleaved by hot piperidine; $(CH_2)_5NH$ at the position of the modified base. The concentration of the modifying chemicals is controlled to introduce on average one modification per DNA molecule. Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule. The fragments in the four reactions are electrophoresed side by side in denaturing acrylamide gels for size separation. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each showing the location of identical radiolabeled DNA molecules. From presence and absence of certain fragments the sequence may be inferred.

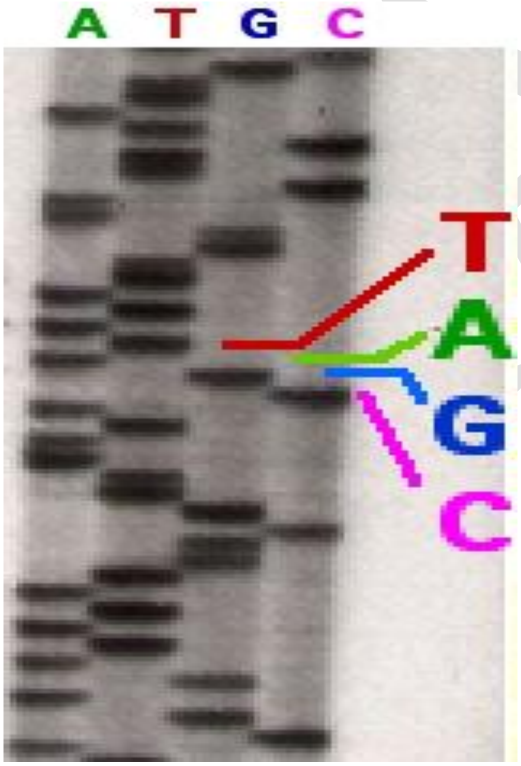
Sanger sequencing

Sanger sequencing is a method of DNA sequencing first commercialized by Applied Biosystems, based on the selective incorporation of chain-terminating dideoxynucleotides by

DNA polymerase during in vitro DNA replication. Developed by Frederick Sanger and colleagues in 1977, it was the most widely used sequencing method for approximately 40 years. More recently, higher volume Sanger sequencing has been supplanted by "Next-Gen" sequencing methods, especially for large-scale, automated genome analyses. However, the Sanger method remains in wide use, for smaller-scale projects, validation of Next-Gen results and for obtaining especially long contiguous DNA sequence reads (> 500 nucleotides).

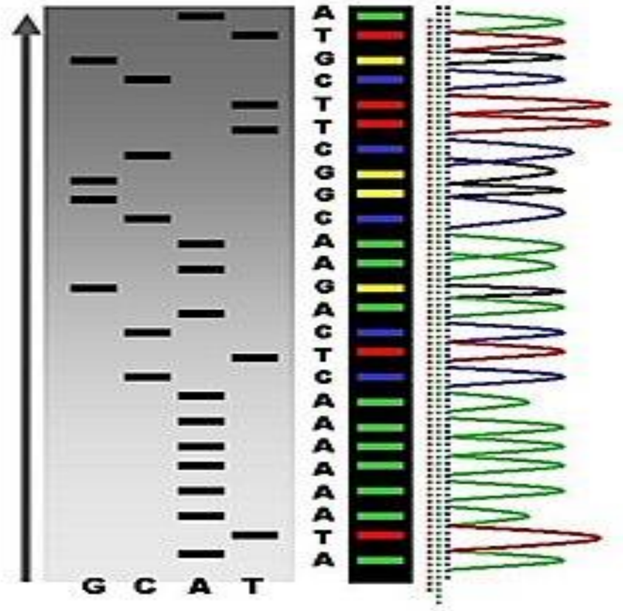

① Reaction mixture
• Primer and DNA template • DNA polymerase
• ddNTPs with flourochromes dNTPs (dATP, dCTP, dGTP, and dTTP)
Primer
Template
ddNTPs
ddTTP
ddCTP
ddATP
ddGTP
② Primer elongation and chain termination
③ Capillary gel electrophoresis separation of DNA fragments
capillary gel
Laser
Detector
④ Laser detection of flourochromes and computational sequence analysis
chromatograph

The classical chain-termination method requires a single-stranded DNA template, a DNA primer, a DNA polymerase, normal deoxynucleosidetriphosphates (dNTPs), and modified di-deoxynucleotidetriphosphates (ddNTPs), the latter of which terminate DNA strand elongation. These chain-terminating nucleotides lack a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, causing DNA polymerase to cease extension of DNA when a modified ddNTP is incorporated. The ddNTPs may be radioactively or fluorescently labeled for detection in automated sequencing machines. The DNA sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only one of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP), while the other added nucleotides are ordinary ones. The dideoxynucleotide is added to be approximately 100-fold lower in concentration than the corresponding deoxynucleotide (e.g. 0.005mM ddATP : 0.5mM dATP) allowing for enough fragments to be produced while still transcribing the complete sequence. Putting it in a more sensible order, four separate reactions are needed in this process to test all four ddNTPs. Following rounds of template DNA extension from the bound primer, the resulting DNA fragments are heat denatured and separated by size using gel electrophoresis. In the original publication of 1977,[2] the formation of base-paired loops of ssDNA was a cause of serious difficulty in resolving bands at some locations. This is frequently performed using a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C). The DNA bands may then be visualized by autoradiography or UV light and the DNA sequence can be directly read off the X-ray film or gel image.



In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes, from bottom to top, are then used to read the DNA sequence. DNA fragments are labelled with a radioactive or fluorescent tag on the primer (1), in the new DNA strand with a labeled dNTP, or with a labeled ddNTP.

Technical variations of chain-termination sequencing include tagging with nucleotides containing radioactive phosphorus for radiolabelling, or using a primer labeled at the 5' end with a fluorescent dye. Dye-primer sequencing facilitates reading in an optical system for faster and more economical analysis and automation. The later development by Leroy Hood and coworkers of fluorescently labeled ddNTPs and primers set the stage for automated, high-throughput DNA sequencing.

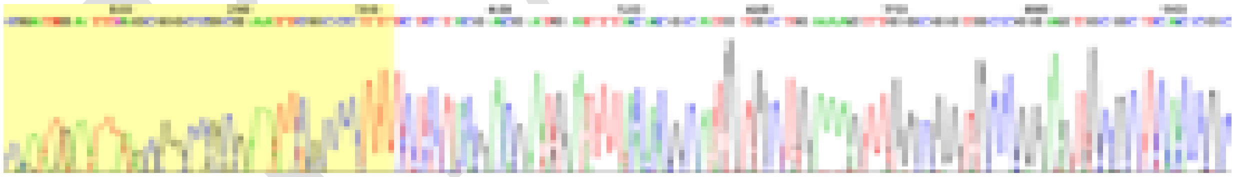


Sequence ladder by radioactive sequencing compared to fluorescent peaks

Chain-termination methods have greatly simplified DNA sequencing. For example, chain-termination-based kits are commercially available that contain the reagents needed for sequencing, pre-aliquoted and ready to use. Limitations include non-specific binding of the primer to the DNA, affecting accurate read-out of the DNA sequence, and DNA secondary structures affecting the fidelity of the sequence.

Dye-terminator sequencing utilizes labeling of the chain terminator ddNTPs, which permits sequencing in a single reaction, rather than four reactions as in the labelled-primer method. In dye-terminator sequencing, each of the four dideoxynucleotide chain terminators is labelled with fluorescent dyes, each of which emits light at different wavelengths.

Owing to its greater expediency and speed, dye-terminator sequencing is now the mainstay in automated sequencing. Its limitations include dye effects due to differences in the incorporation of the dye-labeled chain terminators into the DNA fragment, resulting in unequal peak heights and shapes in the electronic DNA sequence trace chromatogram after capillary electrophoresis (see figure to the left). This problem has been addressed with the use of modified DNA polymerase enzyme systems and dyes that minimize incorporation variability, as well as methods for eliminating "dye blobs". The dye-terminator sequencing method, along with automated high-throughput DNA sequence analyzers, is now being used for the vast majority of sequencing projects.



Automated DNA-sequencing instruments (DNA sequencers) can sequence up to 384 DNA samples in a single batch. Batch runs may occur up to 24 times a day. DNA sequencers separate strands by size (or length) using capillary electrophoresis, they detect and record dye fluorescence, and output data as fluorescent peak trace chromatograms. Sequencing reactions (thermocycling and labelling), cleanup and re-suspension of samples in a buffer solution are performed separately, before loading samples onto the sequencer. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (which are generally located at the ends of the sequence). The accuracy of such algorithms is inferior to visual examination by a human operator, but is adequate for automated processing of large sequence data sets. Common challenges of DNA sequencing with the Sanger method include poor quality in the first 15-40 bases of the sequence due to primer binding and deteriorating quality of sequencing traces after 700-900 bases. Base calling software such as Phred typically provides an estimate of quality to aid in trimming of low-quality regions of sequences.

In cases where DNA fragments are cloned before sequencing, the resulting sequence may contain parts of the cloning vector. In contrast, PCR-based cloning and next-generation sequencing technologies based on pyrosequencing often avoid using cloning vectors. Recently, one-step Sanger sequencing (combined amplification and sequencing) methods such as Ampliseq and SeqSharp have been developed that allow rapid sequencing of target genes without cloning or prior amplification. Current methods can directly sequence only relatively short (300-1000 nucleotides long) DNA fragments in a single reaction. The main obstacle to sequencing DNA fragments above this size limit is insufficient power of separation for resolving large DNA fragments that differ in length by only one nucleotide.

Shotgun 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.

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.

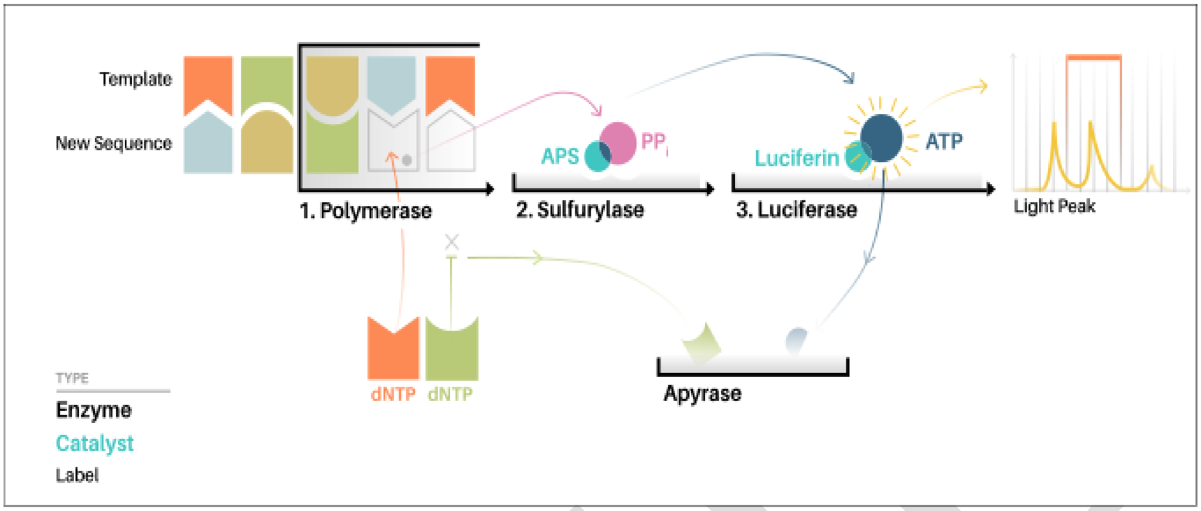
Pyrosequencing

Pyrosequencing is a method of DNA sequencing (determining the order of nucleotides in DNA) based on the "sequencing by synthesis" principle, in which the sequencing is performed by detecting the nucleotide incorporated by a DNA polymerase. Pyrosequencing relies on light detection based on a chain reaction when pyrophosphate is released.

The principle of Pyrosequencing was first described in 1993 by Bertil Pettersson, Mathias Uhlen and Pål Nyren by combining the solid phase sequencing method using streptavidin coated magnetic beads with recombinant DNA polymerase lacking 3´to 5´exonuclease activity (proof-reading) and luminescence detection using the firefly luciferase enzyme. A mixture of three enzymes (DNA polymerase, ATP sulfurylase and firefly luciferase) and a nucleotide (dNTP) are added to single stranded DNA to be sequenced and the incorporation of nucleotide is followed by measuring the light emitted. The intensity of the light determines if 0, 1 or more nucleotides have been incorporated, thus showing how many complementary nucleotides are present on the template strand. The nucleotide mixture is removed before the next nucleotide mixture is added. This process is repeated with each of the four nucleotides until the DNA sequence of the single stranded template is determined.

A second solution-based method for Pyrosequencing was described in 1998 by Mostafa Ronaghi, Mathias Uhlen and Pål Nyren. In this alternative method, an additional enzyme apyrase is introduced to remove nucleotides that are not incorporated by the DNA polymerase. This enabled the enzyme mixture including the DNA polymerase, the luciferase and the apyrase to be added at the start and kept throughout the procedure, thus providing a simple set-up suitable for automation. An automated instrument based on this principle was introduced to the market the following year by the company Pyrosequencing.

A third microfluidic variant of the Pyrosequencing method was described in 2005 by Jonathan Rothberg and co-workers at the company 454 Life Sciences. This alternative approach for Pyrosequencing was based on the original principal of attaching the DNA to be sequenced to a solid support and they showed that sequencing could be performed in a highly parallel manner using a microfabricated microarray. This allowed for high-throughput DNA sequencing and an automated instrument was introduced to the market. This became the first next generation sequencing instrument starting a new era in genomics research, with rapidly falling prices for DNA sequencing allowing whole genome sequencing at affordable prices.



"Sequencing by synthesis" involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. The pyrosequencing method is based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step. The template DNA is immobile, and solutions of A, C, G, and T nucleotides are sequentially added and removed from the reaction. Light is produced only when the nucleotide solution complements the first unpaired base of the template. The sequence of solutions which produce chemiluminescent signals allows the determination of the sequence of the template.

For the solution-based version of Pyrosequencing, the single-strand DNA (ssDNA) template is hybridized to a sequencing primer and incubated with the enzymes DNA polymerase, ATP sulfurylase, luciferase and apyrase, and with the substrates adenosine 5´ phosphosulfate (APS) and luciferin.

1. The addition of one of the four deoxynucleotide triphosphates (dNTPs) (dATPαS, which is not a substrate for a luciferase, is added instead of dATP to avoid noise) initiates the second step. DNA polymerase incorporates the correct, complementary dNTPs onto the template. This incorporation releases pyrophosphate (PPi).

2. ATP sulfurylase converts PPi to ATP in the presence of adenosine 5´ phosphosulfate. This ATP acts as a substrate for the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP. The light produced in the luciferase-catalyzed reaction is detected by a camera and analyzed in a pyrogram.
3. Unincorporated nucleotides and ATP are degraded by the apyrase, and the reaction can restart with another nucleotide.

Currently, a limitation of the method is that the lengths of individual reads of DNA sequence are in the neighborhood of 300-500 nucleotides, shorter than the 800-1000 obtainable with chain termination methods (e.g. Sanger sequencing). This can make the process of genome assembly more difficult, particularly for sequences containing a large amount of repetitive DNA.

The company Pyrosequencing AB in Uppsala, Sweden was founded with venture capital provided by HealthCap in order to commercialize machinery and reagents for sequencing short stretches of DNA using the pyrosequencing technique. Pyrosequencing AB was listed on the Stockholm Stock Exchange in 1999. It was renamed to Biotage in 2003. The pyrosequencing business line was acquired by Qiagen in 2008. Pyrosequencing technology was further licensed to 454 Life Sciences. 454 developed an array-based pyrosequencing technology which has emerged as a platform for large-scale DNA sequencing. Most notable are the applications for genome sequencing and metagenomics. Roche announced the discontinuation of the 454 sequencing platform in 2013.

Possible questions

1 Mark questions

1. The term 'genomics' was coined by
a) Tom Roderick b) Anselme Payen c) Wilhelm Johannsen d) Paulien Hogeweg

2. How many different types of chemcial treatments are required in Maxam-Gilbert method?
a) 1 b) 2 c) 3 d) 4

3. The HGP was completed in
a) 2002 b) 2003 c) 2004 d) 2005

4. In Maxam-gilbert method, chemical used for cytosine alteration is
a) Formic acid b) hydrazine c) Dimethyl sulphate d) piperdine

5. What do you mean by 'epistasis'?
a) structural genes b) functional genes c) effect of one gene on another
d) all the above

6. What is pleiotropy?
a) one gene affecting only one trait b) gene affecting no traits
c) one gene affecting more than one trait d) two genes affecting one trait

7. The principle of sanger's method relies on
a) use of chemicals for base specific cleavage b) use of dNTPs for chain termination
c) use of ddNTPs for chain termination d) use of For P32 Chain termination

8. An open reading frame (ORF) is
a) the sequence of a complete genome b) a plasmid vector used in genomic sequencing
c) a possible gene predicted by DNA sequencing d) a fragment of a genome

9. Proteomics is
a) a branch of quantum physics b) the study of algal genomes
c) the study of the entire collection of proteins expressed by an organism
d) study of entire set of genomes

10. Genomic libraries are useful for obtaining what product?
a) periodicals on genomics research b) collections of isolated genes

c) instructional information on how to locate the exact site of the gene of interest
d) information relating to primers and PCR

11. Protein coding genes can be identified by
a) transposon tagging b) ORF scanning c) Zoo-blotting d) Nuclease S1 Mapping

2 Marks questions

1. What are genes?

2. What are pseudogenes?

3. Write a short note on ‘Shot-gun’ method.

4. What do you mean by ‘epistasis’?

5. What do you mean by ‘pleiotropy’?

6. Write any two unique applications of HGP?

6/8 Marks questions

1. How will you determine a short DNA sequence by ‘Maxam – Gilbert’ method?

2. Describe the ‘Chain termination method’ in a systematic manner with an example.

3. Enumerate various steps involved in a ‘Genomic library construction’.

4. Discuss the merits and limitation of ‘Maxam - Gilbert Method’.

5. What method will you employ to sequencing single-strand DNA fragment composed of 5 bases (5´-AGCTT-3´)? Justify your answer.

6. Explain the role of bioinformatics tools on sequencing polynucleotide chains.

7. Analyze strategies, merits and imitations of ‘Pyrosequencing Method’ for analyzing DNA sequencing.

8. Describe a next-generation DNA sequencing methods in detail manner.

UNIT - II

SYLLABUS

Managing and Distributing Genome Data: Web based servers and software for genome analysis: ENSEMBL, VISTA, UCSC Genome Browser, NCBI genome. Selected Model Organisms' Genomes and Databases.

Ensembl genome database project

Ensembl genome database project is a joint scientific project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, which was launched in 1999 in response to the imminent completion of the Human Genome Project. After 10 years in existence, Ensembl's aim remains to provide a centralized resource for geneticists, molecular biologists and other researchers studying the genomes of our own species and other vertebrates and model organisms. Ensembl is one of several well-known genome browsers for the retrieval of genomic information. Similar databases and browsers are found at NCBI and the University of California, Santa Cruz (UCSC).

The human genome consists of three billion base pairs, which code for approximately 20,000–25,000 genes. However the genome alone is of little use, unless the locations and relationships of individual genes can be identified. One option is manual annotation, whereby a team of scientists tries to locate genes using experimental data from scientific journals and public databases. However this is a slow, painstaking task. The alternative, known as automated annotation, is to use the power of computers to do the complex pattern-matching of protein to DNA.

In the Ensembl project, sequence data are fed into the gene annotation system (a collection of software "pipelines" written in Perl) which creates a set of predicted gene locations and saves them in a MySQL database for subsequent analysis and display. Ensembl makes these data freely accessible to the world research community. All the data and code produced by the Ensembl project is available to download and there is also a publicly accessible database server allowing remote access. In addition, the Ensembl website provides computer-generated visual displays of much of the data.

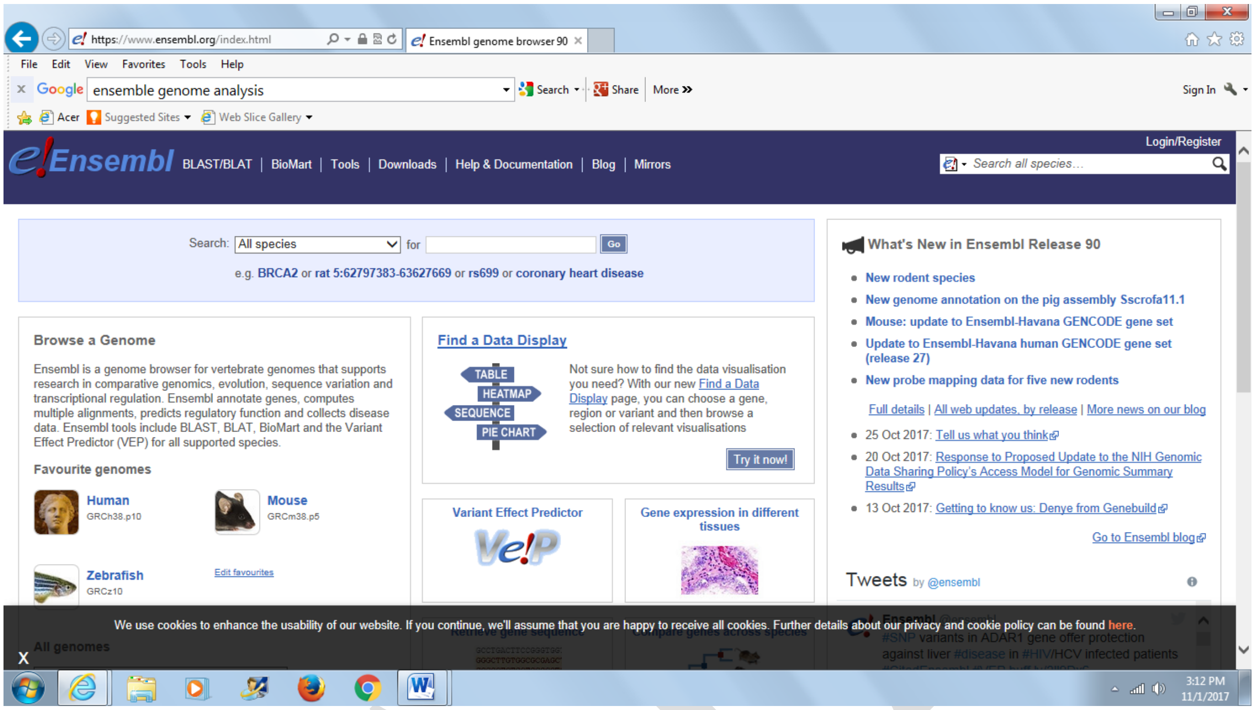
Over time the project has expanded to include additional species (including key model organisms such as mouse, fruitfly and zebrafish) as well as a wider range of genomic data, including genetic variations and regulatory features. Since April 2009, a sister project, Ensembl Genomes, has extended the scope of Ensembl into invertebrate metazoa, plants, fungi, bacteria, and protists, whilst the original project continues to focus on vertebrates.

Central to the Ensembl concept is the ability to automatically generate graphical views of the alignment of genes and other genomic data against a reference genome. These are shown as data tracks, and individual tracks can be turned on and off, allowing the user to customize the display to suit their research interests. The interface also enables the user to zoom in to a region or move along the genome in either direction.

Other displays show data at varying levels of resolution, from whole karyotypes down to text-based representations of DNA and amino acid sequences, or present other types of display such as trees of similar genes (homologues) across a range of species. The graphics are complemented by tabular displays, and in many cases data can be exported directly from the page in a variety of standard file formats such as FASTA.

Externally produced data can also be added to the display, either via a DAS (Distributed Annotation System) server on the internet, or by uploading a suitable file in one of the supported formats, such as BAM, BED, or PSL.

Graphics are generated using a suite of custom Perl modules based on GD, the standard Perl graphics display library.



Search: All species for

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Favourite genomes

Human GRCh38.p10

Mouse GRCm38.p5

Zebrafish GRCz10

Edit favourites

All genomes

Find a Data Display

Not sure how to find the data visualisation you need? With our new Find a Data Display page, you can choose a gene, region or variant and then browse a selection of relevant visualisations

Variant Effect Predictor

Gene expression in different tissues

What's New in Ensembl Release 90

- New rodent species
- New genome annotation on the pig assembly Sscrofa11.1
- Mouse: update to Ensembl-Havana GENCODE gene set
- Update to Ensembl-Havana human GENCODE gene set (release 27)
- New probe mapping data for five new rodents

Full details | All web updates, by release | More news on our blog

- 25 Oct 2017: Tell us what you think
- 20 Oct 2017: Response to Proposed Update to the NIH Genomic Data Sharing Policy's Access Model for Genomic Summary Results
- 13 Oct 2017: Getting to know us: Denye from Genebuild

Go to Ensembl blog

Tweets by @ensembl

We use cookies to enhance the usability of our website. If you continue, we'll assume that you are happy to receive all cookies. Further details about our privacy and cookie policy can be found here.

Alternative access method

In addition to its website, Ensembl provides a Perl API (Application Programming Interface) that models biological objects such as genes and proteins, allowing simple scripts to be written to retrieve data of interest. The same API is used internally by the web interface to display the data. It is divided in sections like the core API, the compara API (for comparative genomics data), the variation API (for accessing SNPs, SNVs, CNVs), and the functional genomics API (to access regulatory data). The Ensembl website provides extensive information on how to install and use the API.

This software can be used to access the public MySQL database, avoiding the need to download enormous datasets. The users could even choose to retrieve data from the MySQL with direct SQL queries, but this requires an extensive knowledge of the current database schema. Large datasets can be retrieved using the BioMart data-mining tool. It provides a web interface for downloading datasets using complex queries. Last, there is an FTP server which can be used to download entire MySQL databases as well some selected data sets in other formats.

VISTA (comparative genomics)

VISTA is a collection of databases, tools, and servers that permit extensive comparative genomics analyses.

The VISTA family of tools is developed and hosted at Genomics Division of Lawrence Berkeley National Laboratory. This collaborative effort is supported by the Programs for Genomic Applications grant from the NHLBI/NIH and the Office of Biological and Environmental Research, Office of Science, US Department of Energy.

It was developed from modules supplied by developers at UC Berkeley, Stanford, and UC Davis, and based partly on the AVID Global Alignment program.

There are multiple VISTA servers, each allowing different types of searches.

- mVISTA can be used to align and compare your sequences to those of multiple other species
- rVISTA (regulatory VISTA) combines transcription factor binding sites database search with a comparative sequence analysis,the discovery of possible regulatory transcription factor binding sites in regions of their genes of interest. It can be used directly or through mVISTA, Genome VISTA, or VISTA Browser. A database of tissue-specific human enhancers is available through VISTA Enhancer Browser.
- GenomeVISTA allows the comparison of sequences with whole genome assemblies. It will automatically find the ortholog, obtain the alignment and VISTA plot. It allows the viewing of an alignment together with pre-computed alignments of other species in the same interval.
- Phylo-VISTA allows the analysis of multiple DNA sequence alignments of sequences from different species while considering their phylogenic relationships.
- wgVISTA allows the alignment of sequences up to 10Mb long (finished or draft) including microbial whole-genome assemblies.

Researchers can use the VISTA Browser:

- to examine pre-computed alignments among a variety of species
- to submit sequences of their own (not limited by the species collection already in the database)

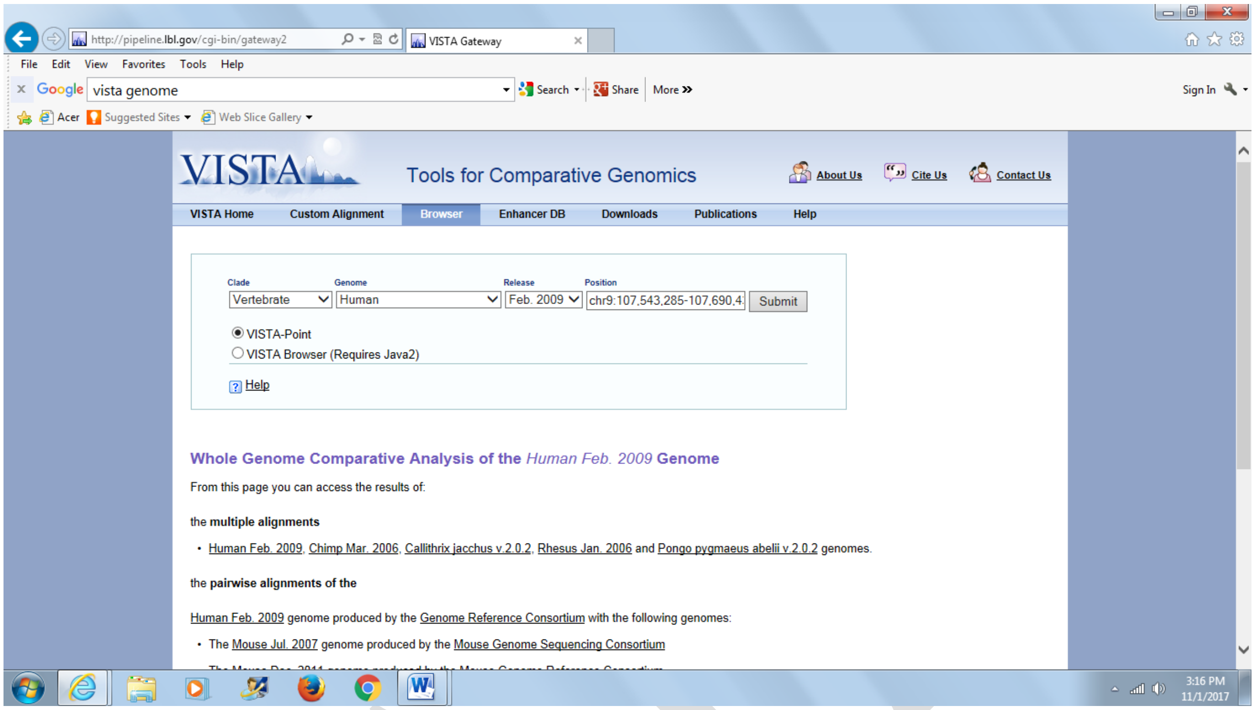
Genomes

There are more than 28 searchable genomes, including vertebrate, non-vertebrate, plants, fungi, algae, bacteria, and others. More are continually being added. These include:

- Human—Orangutan—Rhesus—Marmoset—Horse—Dog—Mouse—Rat—Chicken
- Drosophila spp.
- Arabidopsis—Rice—Sorghum
- E. coli—mycoplasma—nitrosomonas

Collaboration with other projects

Pre-computed full scaffold alignments for microbial genomes are available as the VISTA component of IMG (Integrated Microbial Genomes System) developed in the DOE (Department of Energy's) Joint Genome Institute.



Tools for Comparative Genomics

Clade: Vertebrate | Genome: Human | Release: Feb. 2009 | Position: chr9:107,543,285-107,690,4 | Submit

● VISTA-Point
○ VISTA Browser (Requires Java2)

Help

Whole Genome Comparative Analysis of the *Human Feb. 2009* Genome

From this page you can access the results of:

the **multiple alignments**

- Human Feb. 2009, Chimp Mar. 2006, Callithrix jacchus v.2.0.2, Rhesus Jan. 2006 and Pongo pygmaeus abelii v.2.0.2 genomes.

the **pairwise alignments of the**

Human Feb. 2009 genome produced by the Genome Reference Consortium with the following genomes:

- The Mouse Jul. 2007 genome produced by the Mouse Genome Sequencing Consortium

VISTA is a comprehensive suite of programs and databases for comparative analysis of genomic sequences. There are two ways of using VISTA - you can submit your own sequences and alignments for analysis (VISTA servers) or examine pre-computed whole-genome alignments of different species.

mVISTA

- mVISTA

  Align and compare your sequences from multiple species

- gVISTA

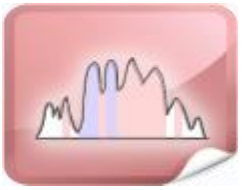
  Compare your sequences against whole-genome assemblies.

- wgVISTA

Align pair of sequences up to 10Mb long (finished or draft) including microbial whole-genome assemblies.

Precomputed Alignments

VISTA Browser



- VISTA-Point

  Access complete data and visual presentation of pairwise and multiple alignments of whole genome assemblies.
- VISTA Browser

  Examine pre-computed pairwise and multiple alignments of whole genome assemblies.
- Microbial Genomes

  Access pre-computed full scaffold alignments for microbial genomes through the VISTA component of IMG.

UCSC Genome Browser

The UCSC Genome Browser is an on-line genome browser hosted by the University of California, Santa Cruz (UCSC). It is an interactive website offering access to genome sequence data from a variety of vertebrate and invertebrate species and major model organisms, integrated with a large collection of aligned annotations. The Browser is a graphical viewer optimized to support fast interactive performance and is an open-source, web-based tool suite built on top of a MySQL database for rapid visualization, examination, and querying of the data at many levels. The Genome Browser Database, browsing tools, downloadable data files, and documentation can all be found on the UCSC Genome Bioinformatics website.

Initially built and still managed by Jim Kent, then a graduate student, and David Haussler, professor of Computer Science (now Biomolecular Engineering) at the University of

California, Santa Cruz in 2000, the UCSC Genome Browser began as a resource for the distribution of the initial fruits of the Human Genome Project. Funded by the Howard Hughes Medical Institute and the National Human Genome Research Institute, NHGRI (one of the US National Institutes of Health), the browser offered a graphical display of the first full-chromosome draft assembly of human genome sequence. Today the browser is used by geneticists, molecular biologists and physicians as well as students and teachers of evolution for access to genomic information.

In the years since its inception, the UCSC Browser has expanded to accommodate genome sequences of all vertebrate species and selected invertebrates for which high-coverage genomic sequences is available, now including 46 species. High coverage is necessary to allow overlap to guide the construction of larger contiguous regions. Genomic sequences with less coverage are included in multiple-alignment tracks on some browsers, but the fragmented nature of these assemblies does not make them suitable for building full featured browsers. (more below on multiple-alignment tracks). The species hosted with full-featured genome browsers are shown in the table.

great apes
human, baboon, bonobo, chimp, gibbon, gorilla, orangutan

non-ape primates
bushbaby, marmoset, mouse lemur, rhesus macaque, squirrel monkey, tarsier, tree shrew

non-primate mammals
mouse, alpaca, armadillo, cat, Chinese hamster, cow, dog, dolphin, elephant, ferret, guinea pig, hedgehog, horse, kangaroo rat, manatee, Minke whale, naked mole-rat, opossum, panda, pig, pika, platypus, rabbit, rat, rock hyrax, sheep, shrew, sloth, squirrel, Tasmanian devil, tenrec, wallaby, white rhinoceros

non-mammal chordates

American alligator, Atlantic cod, budgerigar, chicken, coelocanth, elephant shark, Fugu, lamprey, lizard, medaka, medium ground finch, Nile tilapia, painted turtle, stickleback, Tetraodon, turkey, Xenopus tropicalis, zebra finch, zebrafish
invertebrates
Caenorhabditis spp (5), Drosophila spp. (11), Ebola virus, honey bee, lancelet, mosquito, P. Pacificus, sea hare, sea squirt, sea urchin, yeast

The large amount of data about biological systems that is accumulating in the literature makes it necessary to collect and digest information using the tools of bioinformatics. The UCSC Genome Browser presents a diverse collection of annotation datasets (known as "tracks" and presented graphically), including mRNA alignments, mappings of DNA repeat elements, gene predictions, gene-expression data, disease-association data (representing the relationships of genes to diseases), and mappings of commercially available gene chips (e.g., Illumina and Agilent). The basic paradigm of display is to show the genome sequence in the horizontal dimension, and show graphical representations of the locations of the mRNAs, gene predictions, etc. Blocks of color along the coordinate axis show the locations of the alignments of the various data types. The ability to show this large variety of data types on a single coordinate axis makes the browser a handy tool for the vertical integration of the data.

To find a specific gene or genomic region, the user may type in the gene name, (e.g., BRCA1) an accession number for an RNA, the name of a genomic cytological band (e.g., 20p13 for band 13 on the short arm of chr20) or a chromosomal position (chr17:38,450,000-38,531,000 for the region around the gene BRCA1).

Presenting the data in the graphical format allows the browser to present link access to detailed information about any of the annotations. The gene details page of the UCSC Genes track provides a large number of links to more specific information about the gene at many other data resources, such as Online Mendelian Inheritance in Man (OMIM) and SwissProt.

Designed for the presentation of complex and voluminous data, the UCSC Browser is optimized for speed. By pre-aligning the 55 million RNAs of GenBank to each of the 81 genome assemblies (many of the 46 species have more than one assembly), the browser allows instant access to the alignments of any RNA to any of the hosted species.

Multiple gene products of FOXP2 gene (top) and evolutionary conservation shown in multiple alignment (bottom). The juxtaposition of the many types of data allow researchers to display exactly the combination of data that will answer specific questions. A pdf/postscript output functionality allows export of a camera-ready image for publication in academic journals.

One unique and useful feature that distinguishes the UCSC Browser from other genome browsers is the continuously variable nature of the display. Sequence of any size can be displayed, from a single DNA base up to the entire chromosome (human chr1 = 245 million bases, Mb) with full annotation tracks. Researchers can display a single gene, a single exon, or an entire chromosome band, showing dozens or hundreds of genes and any combination of the many annotations. A convenient drag-and-zoom feature allows the user to choose any region in the genome image and expand it to occupy the full screen.

Researchers may also use the browser to display their own data via the Custom Tracks tool. This feature allows users to upload a file of their own data and view the data in the context of the reference genome assembly. Users may also use the data hosted by UCSC, creating subsets of the data of their choosing with the Table Browser tool (such as only the SNPs that change the amino acid sequence of a protein) and display this specific subset of the data in the browser as a Custom Track.

Any browser view created by a user, including those containing Custom Tracks, may be shared with other users via the Saved Sessions tool.

Variation data

Many types of variation data are also displayed. For example, the entire contents of each release of the dbSNP database from NCBI are mapped to human, mouse and other genomes. This includes the fruits of the 1000 Genomes Project, as soon as they are released in dbSNP.

Other types of variation data include copy-number variation data (CNV) and human population allele frequencies from the HapMap project.

The Genome Browser offers a unique set of comparative-genomic data for most of the species hosted on the site. The comparative alignments give a graphical view of the evolutionary relationships among species. This makes it a useful tool both for the researcher, who can visualize regions of conservation among a group of species and make predictions about functional elements in unknown DNA regions, and in the classroom as a tool to illustrate one of the most compelling arguments for the evolution of species. The 44-way comparative track on the human assembly clearly shows that the farther one goes back in evolutionary time, the less sequence homology remains, but functionally important regions of the genome (e.g., exons and control elements, but not introns typically) are conserved much farther back in evolutionary time.

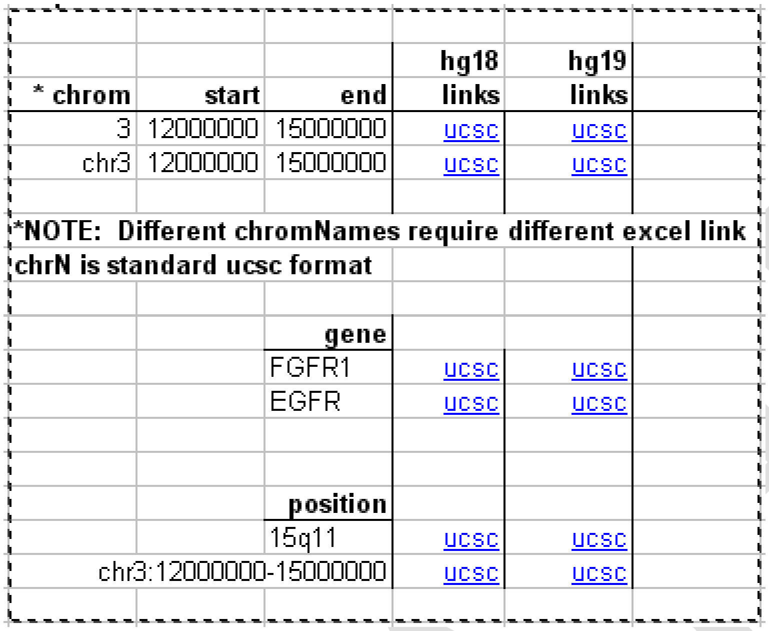
Analysis tools

More than simply a genome browser, the UCSC site hosts a set of genome analysis tools, including a full-featured GUI interface for mining the information in the browser database (the Table Browser), a fast sequence alignment tool (BLAT) that is also useful for simply finding sequences in the massive sequence (human genome = 2.8 billion bases, Gb) of any of the featured genomes.

A liftOver tool uses whole-genome alignments to allow conversion of sequences from one assembly to another or between species. The Genome Graphs tool allows users to view all chromosomes at once and display the results of genome-wide association studies (GWAS). The Gene Sorter displays genes grouped by parameters not linked to genome location, such as expression pattern in tissues.

Creating spreadsheet links to UCSC Genome Browser views

Many users of the Genome Browser gather data of their own in Excel spreadsheets and would like to create links to the Browser using data in the spreadsheet. For example, a clinical geneticist may have lists of regions for a patient that are duplicated or deleted, as determined by comparative genomic hybridization (CGH). These regions can be the source information for a browser view allowing access to each region with a single click.

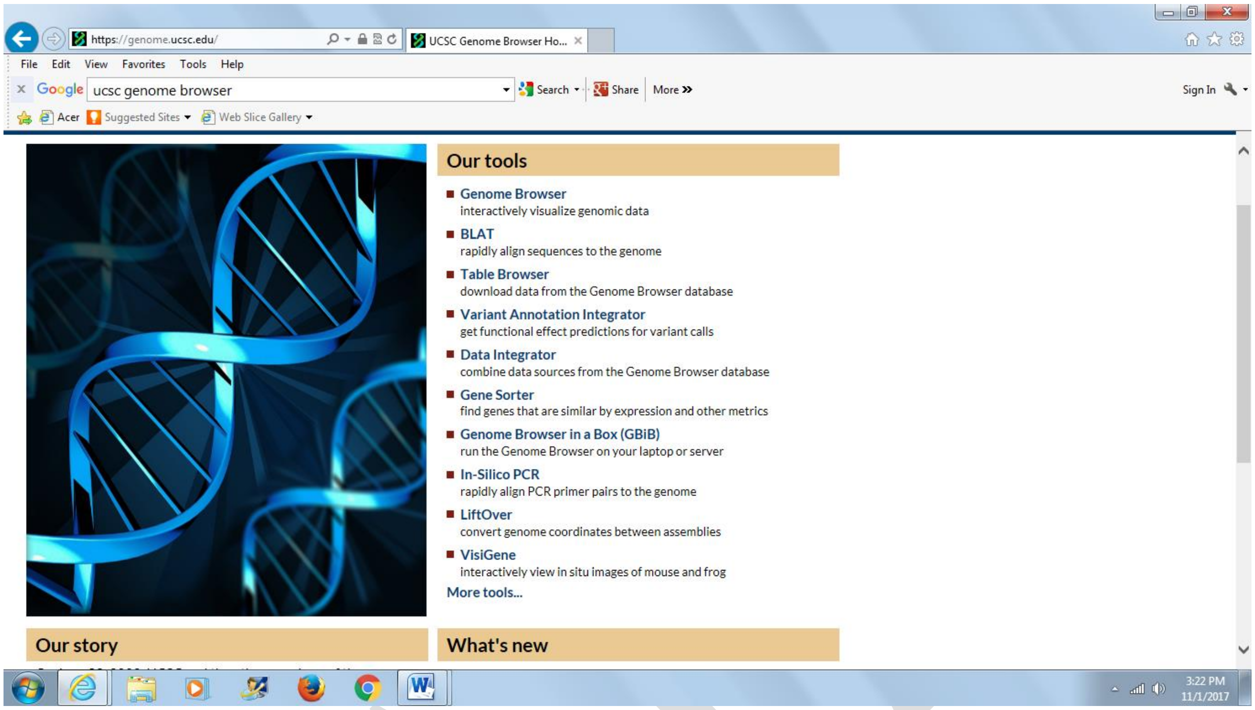


			hg18	hg19
* chrom	start	end	links	links
3	12000000	15000000	ucsc	ucsc
chr3	12000000	15000000	ucsc	ucsc
*NOTE: Different chromNames require different excel link				
chrN is standard ucsc format				
		gene		
		FGFR1	ucsc	ucsc
		EGFR	ucsc	ucsc
		position		
		15q11	ucsc	ucsc
		chr3:12000000-15000000	ucsc	ucsc

Click to download the spreadsheet: ucscLinks.xls

Careful use of Excel's "copy" and "move" functions should allow the links on this sheet to be used without modification.

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it contains. A few weeks later, on July 7, 2000, the newly assembled genome was released on the web at http://genome.ucsc.edu, along with the initial prototype of a graphical viewing tool, the UCSC Genome Browser. In the ensuing years, the website has grown to include a broad collection of vertebrate and model organism assemblies and annotations, along with a large suite of tools for viewing, analyzing and downloading data.



Our tools

- Genome Browser
  interactively visualize genomic data
- BLAT
  rapidly align sequences to the genome
- Table Browser
  download data from the Genome Browser database
- Variant Annotation Integrator
  get functional effect predictions for variant calls
- Data Integrator
  combine data sources from the Genome Browser database
- Gene Sorter
  find genes that are similar by expression and other metrics
- Genome Browser in a Box (GBiB)
  run the Genome Browser on your laptop or server
- In-Silico PCR
  rapidly align PCR primer pairs to the genome
- LiftOver
  convert genome coordinates between assemblies
- VisiGene
  interactively view in situ images of mouse and frog

More tools...

Our story

What's new

National Center for Biotechnology Information

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH). The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper. The NCBI houses a series of databases relevant to biotechnology and biomedicine and is an important resource for bioinformatics tools and services. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. Other databases include the NCBI Epigenomics database. All these databases are available online through the Entrez search engine.

NCBI was directed by David Lipman, one of the original authors of the BLAST sequence alignment program and a widely respected figure in bioinformatics. He also leads an intramural research program, including groups led by Stephen Altschul (another BLAST co-author), David

Landsman, Eugene Koonin (a prolific author on comparative genomics), John Wilbur, Teresa Przytycka, and Zhiyong Lu. David Lipman stood down from his post in May 2017.

NCBI is listed in the Registry of Research Data Repositories re3data.org.

GenBank

NCBI has had responsibility for making available the GenBank DNA sequence database since 1992. GenBank coordinates with individual laboratories and other sequence databases such as those of the European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ).

Since 1992, NCBI has grown to provide other databases in addition to GenBank. NCBI provides Gene, Online Mendelian Inheritance in Man, the Molecular Modeling Database (3D protein structures), dbSNP (a database of single-nucleotide polymorphisms), the Reference Sequence Collection, a map of the human genome, and a taxonomy browser, and coordinates with the National Cancer Institute to provide the Cancer Genome Anatomy Project. The NCBI assigns a unique identifier (taxonomy ID number) to each species of organism.

The NCBI has software tools that are available by WWW browsing or by FTP. For example, BLAST is a sequence similarity searching program. BLAST can do sequence comparisons against the GenBank DNA database in less than 15 seconds.

NCBI Bookshelf

The "NCBI Bookshelf is a collection of freely accessible, downloadable, on-line versions of selected biomedical books. The Bookshelf covers a wide range of topics including molecular biology, biochemistry, cell biology, genetics, microbiology, disease states from a molecular and cellular point of view, research methods, and virology. Some of the books are online versions of previously published books, while others, such as Coffee Break, are written and edited by NCBI staff. The Bookshelf is a complement to the Entrez PubMed repository of peer-reviewed publication abstracts in that Bookshelf contents provide established perspectives on evolving areas of study and a context in which many disparate individual pieces of reported research can be organized.

Basic Local Alignment Search Tool (BLAST)

BLAST is an algorithm used for calculating sequence similarity between biological sequences such as nucleotide sequences of DNA and amino acid sequences of proteins. BLAST is a powerful tool for finding sequences similar to the query sequence within the same organism or in different organisms. It searches the query sequence on NCBI databases and servers and post the results back to the person's browser in chosen format. Input sequences to the BLAST are mostly in FASTA or Genbank format while output could be delivered in variety of formats such as HTML, XML formatting and plain text. HTML is the default output format for NCBI's web-page. Results for NCBI-BLAST are presented in graphical format with all the hits found, a table with sequence identifiers for the hits having scoring related data, along with the alignments for the sequence of interest and the hits received with analogous BLAST scores for these.

Entrez

The Entrez Global Query Cross-Database Search System is used at NCBI for all the major databases such as Nucleotide and Protein Sequences, Protein Structures, PubMed, Taxonomy, Complete Genomes, OMIM, and several others. Entrez is both indexing and retrieval system having data from various sources for biomedical research. NCBI distributed the first version of Entrez in 1991, composed of nucleotide sequences from PDB and GenBank, protein sequences from SWISS-PROT, translated GenBank, PIR, PRF and PDB and associated abstracts and citations from PubMed. Entrez is specially designed to integrate the data from several different sources, databases and formats into a uniform information model and retrieval system which can efficiently retrieve that relevant references, sequences and structures.

Gene

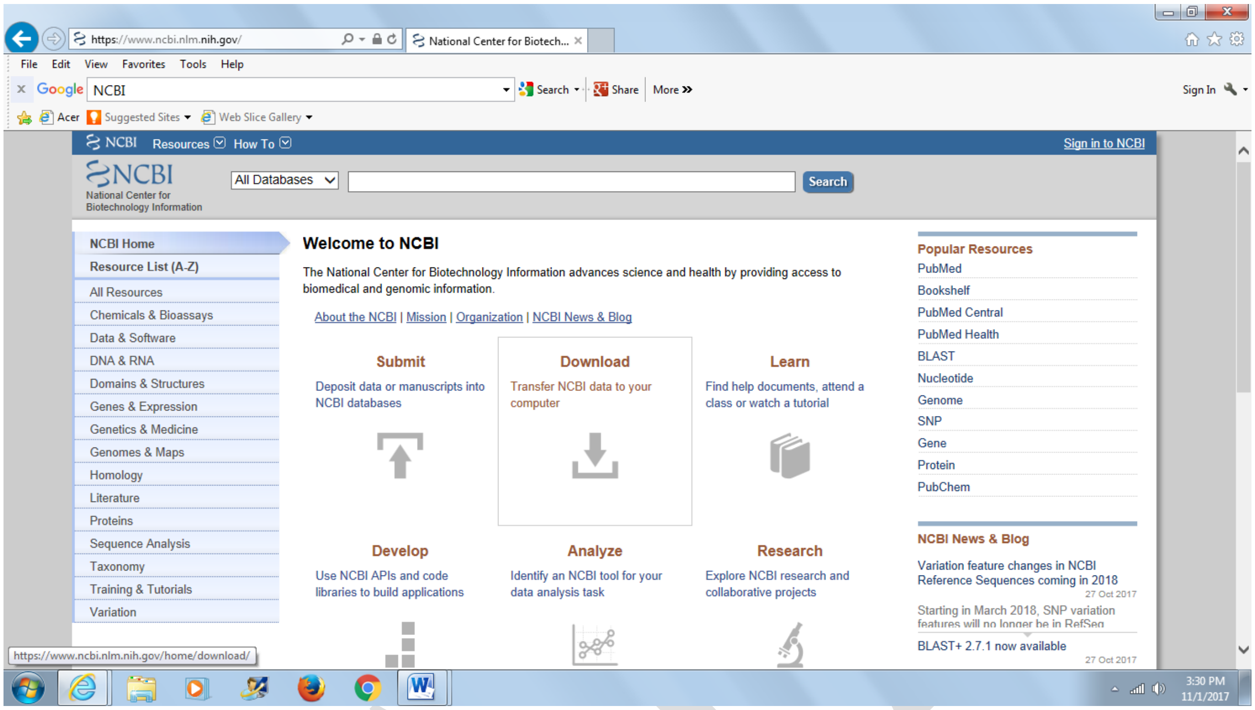
Gene has been implemented at NCBI to characterize and organize the information about genes. It serves as a major node in the nexus of genomic map, expression, sequence, protein function, structure and homology data. A unique GeneID is assigned to each gene record that can be followed through revision cycles. Gene records for known or predicted genes are established here and are demarcated by map positions or nucleotide sequence. Gene has several advantages over its predecessor, LocusLink, including, better integration with other databases in NCBI, broader taxonomic scope, and enhanced options for query and retrieval provided by Entrez system.

Protein

Protein database is an important protein resource at NCBI. It maintains the text record for individual protein sequences, derived from many different resources such as NCBI Reference Sequence (RefSeq) project, GenbBank, PDB and UniProtKB/SWISS-Prot. Protein records are present in different formats including FASTA and XML and are linked to other NCBI resources. Protein provides the relevant data to the users such as genes, DNA/RNA sequences, biological pathways, expression and variation data and literature. It also provides the pre-determined sets of similar and identical proteins for each sequence as computed by the BLAST. The Structure database of NCBI contains 3D coordinate sets for experimentally-determined structures in PDB that are imported by NCBI. The Conserved Domain database (CDD) of protein contains sequence profiles that characterize highly conserved domains within protein sequences. It also has records from external resources like SMART and Pfam. There is another database in protein known as Protein Clusters database which contains sets of proteins sequences that are clustered according to the maximum alignments between the individual sequences as calculated by BLAST.

Pubchem BioAssay database

PubChem BioAssay database of NCBI is a public resource for biological tests of small molecules and siRNA reagents. The major purpose of PubChem repository is to provide easy and free of cost access to all deposited data, and to provide intuitive data analysis tools. It is structured as a set of relational databases organized on Microsoft SQL servers. PubChem’s BioAssay data is searchable and accessible by Entrez information retrieval system. PubChem database provides programmatic and Web-based tools for users to search, review, and download a publications, bioactivity data for a compound, a BioAssay record, a molecular target.



https://www.ncbi.nlm.nih.gov/
National Center for Biotech...
File Edit View Favorites Tools Help
Google NCBI Search Share More »
Sign In
Acer Suggested Sites Web Slice Gallery
NCBI Resources How To
Sign in to NCBI
NCBI
National Center for Biotechnology Information
All Databases
Search

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog

Variation feature changes in NCBI Reference Sequences coming in 2018
27 Oct 2017

Starting in March 2018, SNP variation features will no longer be in RefSeq

BLAST+ 2.7.1 now available
27 Oct 2017

Genome and Organism – Specific Databases

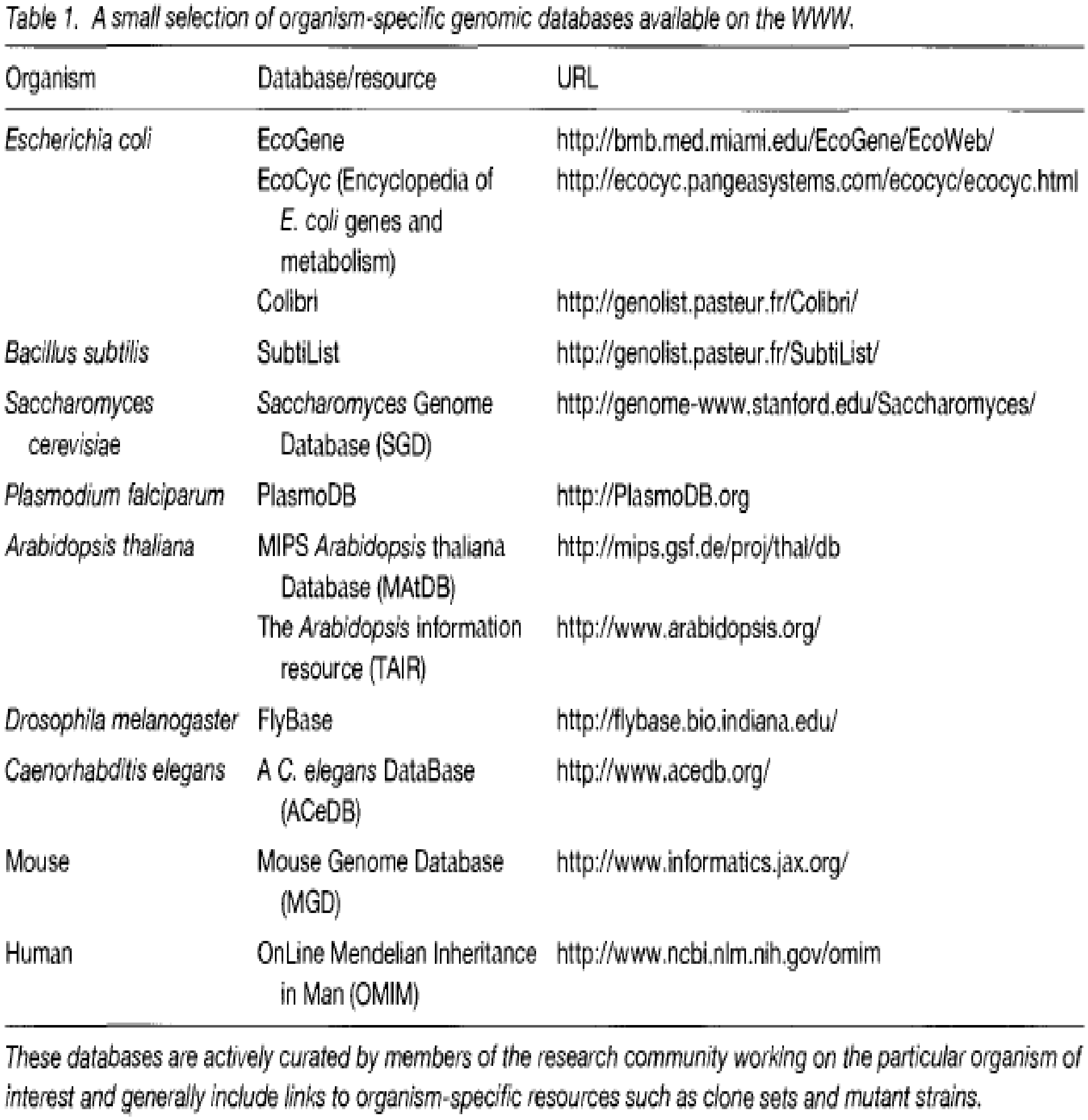


Table 1. A small selection of organism-specific genomic databases available on the WWW.

Organism	Database/resource	URL
Escherichia coli	EcoGene	http://bmb.med.miami.edu/EcoGene/EcoWeb/
	EcoCyc (Encyclopedia of *E. coli* genes and metabolism)	http://ecocyc.pangeasystems.com/ecocyc/ecocyc.html
	Colibri	http://genolist.pasteur.fr/Colibri/
Bacillus subtilis	SubtiList	http://genolist.pasteur.fr/SubtiList/
Saccharomyces cerevisiae	*Saccharomyces* Genome Database (SGD)	http://genome-www.stanford.edu/Saccharomyces/
Plasmodium falciparum	PlasmoDB	http://PlasmoDB.org
Arabidopsis thaliana	MIPS *Arabidopsis* thaliana Database (MAtDB)	http://mips.gsf.de/proj/thal/db
	The *Arabidopsis* information resource (TAIR)	http://www.arabidopsis.org/
Drosophila melanogaster	FlyBase	http://flybase.bio.indiana.edu/
Caenorhabditis elegans	A *C. elegans* DataBase (ACeDB)	http://www.acedb.org/
Mouse	Mouse Genome Database (MGD)	http://www.informatics.jax.org/
Human	OnLine Mendelian Inheritance in Man (OMIM)	http://www.ncbi.nlm.nih.gov/omim

These databases are actively curated by members of the research community working on the particular organism of interest and generally include links to organism-specific resources such as clone sets and mutant strains.

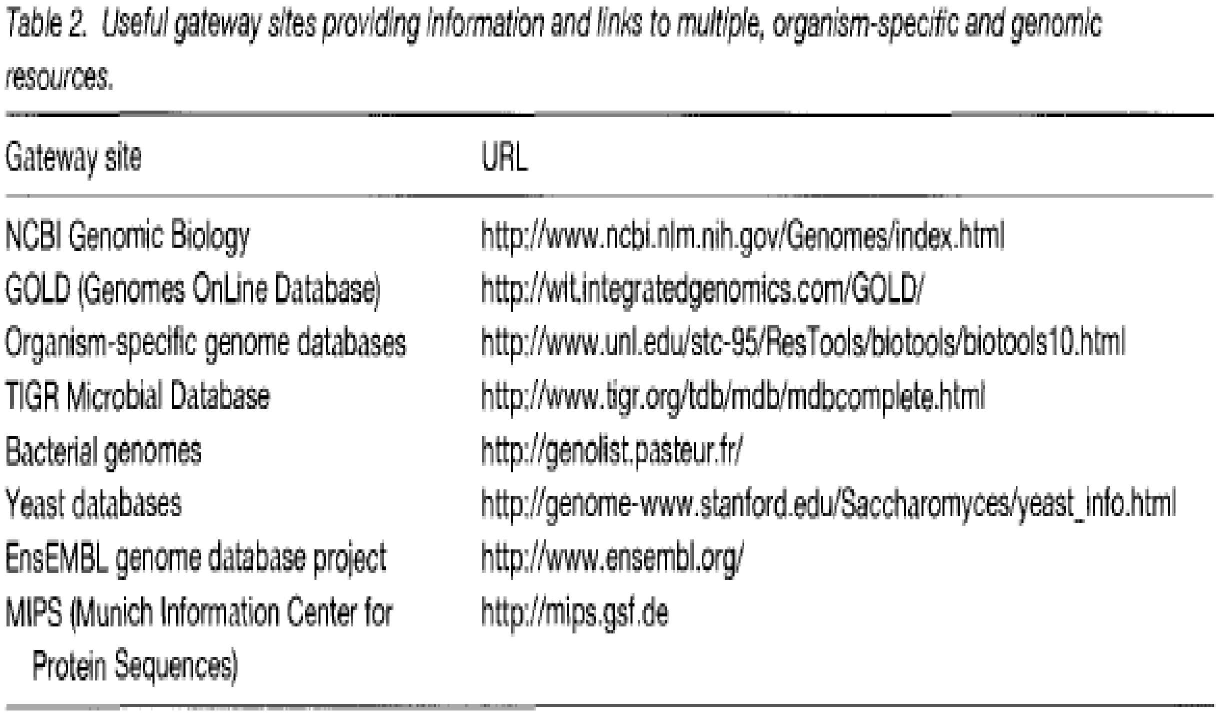


Table 2. Useful gateway sites providing information and links to multiple, organism-specific and genomic resources.

Gateway site	URL
NCBI Genomic Biology	http://www.ncbi.nlm.nih.gov/Genomes/index.html
GOLD (Genomes OnLine Database)	http://wit.integratedgenomics.com/GOLD/
Organism-specific genome databases	http://www.unl.edu/stc-95/ResTools/biotools/biotools10.html
TIGR Microbial Database	http://www.tigr.org/tdb/mdb/mdbcomplete.html
Bacterial genomes	http://genolist.pasteur.fr/
Yeast databases	http://genome-www.stanford.edu/Saccharomyces/yeast_info.html
EnsEMBL genome database project	http://www.ensembl.org/
MIPS (Munich Information Center for Protein Sequences)	http://mips.gsf.de

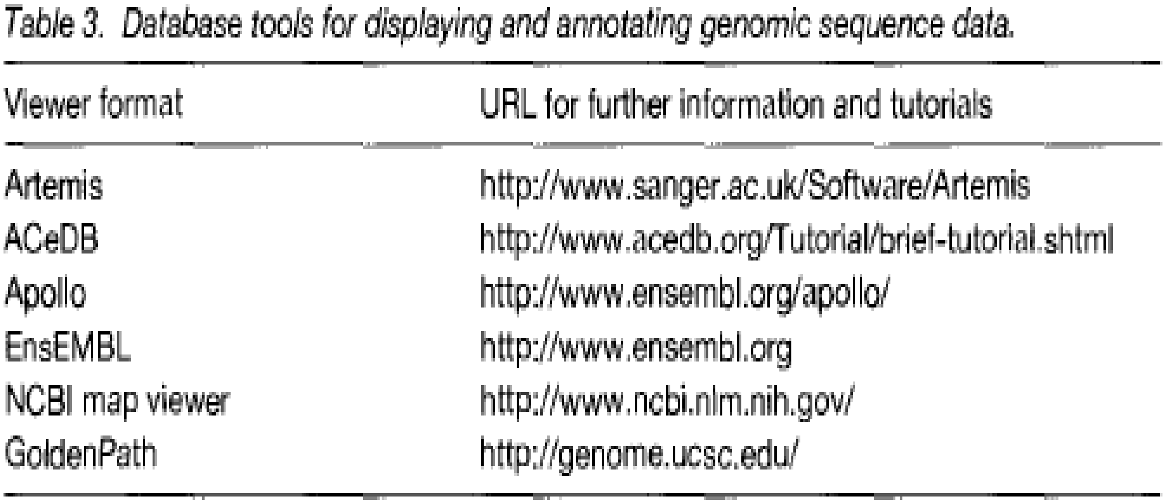


Table 3. Database tools for displaying and annotating genomic sequence data.

Viewer format	URL for further information and tutorials
Artemis	http://www.sanger.ac.uk/Software/Artemis
ACeDB	http://www.acedb.org/Tutorial/brief-tutorial.shtml
Apollo	http://www.ensembl.org/apollo/
EnsEMBL	http://www.ensembl.org
NCBI map viewer	http://www.ncbi.nlm.nih.gov/
GoldenPath	http://genome.ucsc.edu/

FlyBase

FlyBase is an online bioinformatics database and the primary repository of genetic and molecular data for the insect family *Drosophilidae*. For the most extensively studied species and model organism, *Drosophila melanogaster*, a wide range of data are presented in different formats. Information in FlyBase originates from a variety of sources ranging from large-scale genome projects to the primary research literature. These data types include mutant phenotypes, molecular characterization of mutant alleles and other deviations, cytological maps, wild-type expression patterns, anatomical images, transgenic constructs and insertions, sequence-level gene models and molecular classification of gene product functions. Query tools allow navigation of FlyBase through DNA or protein sequence, by gene or mutant name, or through terms from the several ontologies used to capture functional, phenotypic, and anatomical data. The database offers several different query tools in order to provide efficient access to the data available and facilitate the discovery of significant relationships within the database. Links between FlyBase and external databases, such as BDGP or modENCODE, provide opportunity for further exploration into other model organism databases and other resources of biological and molecular information.[3] The FlyBase project is carried out by a consortium of *Drosophila* researchers and computer scientists at Harvard University and Indiana University in the United States, and University of Cambridge in the United Kingdom.

Drosophila melanogaster has been an experimental organism since the early 1900s, and has since been placed at the forefront of many areas of research. As this field of research spread and became global, researchers working on the same problems needed a way to communicate and monitor progress in the field. This niche was initially filled community newsletters such as the Drosophila Information Service (DIS), which dates back to 1934 when the field was starting to spread from Thomas Hunt Morgan's lab. Material in these presented regular 'catalogs' of mutations bibliographies of the Drosophila literature. As computer infrastructure developed in the 80's and 90's, these newsletters gave way and merged with internet mailing lists, and these eventually became online resources and data. In 1992, data on the genetics and genomics of *D.*

melanogaster and related species were electronically available over the Internet through the funded FlyBase, BDGP (Berkeley Drosophila Genome Project) and EDGP (European Drosophila Genome Project) informatics groups. These groups recognized that most genome project and community data types overlapped. They decided it would be of value to present the scientific community with an integrated view of the data. In October 1992, the National Center for Human Genome Research of the NIH funded the FlyBase project with the objective of designing, building and releasing a database of genetic and molecular information concerning *Drosophila melanogaster*. FlyBase also receives support from the Medical Research Council, London. In 1998, the FlyBase consortium integrated the information into a single Drosophila genomics server.

FlyBase contains a complete annotation of the *Drosophila melanogaster* genome that is updated several times per year. It also includes a searchable bibliography of research on *Drosophila* genetics in the last century. Information on current researchers, and a partial pedigree of relationships between current researchers, is searchable, based on registration of the participating scientist (Find a Person). The site also provides a large database of images illustrating the full genome, and several movies detailing embryogenesis (ImageBrowser).

Search Strategies - Gene reports for genes from all twelve sequenced Drosophila genomes are available in FlyBase. There are four main ways this data can be browsed: Precomputed Files, BLAST, Gbrowse, and Gene Report Pages. Gbrowse and precomputed files are for genome-wide analysis, bioinformatics, and comparative genomics. BLAST and gene report pages are for a specific gene, protein, or region across the species. When looking for cytology there are two main tools available. Use Cytosearch when looking for cytologically-mapped genes or deficiencies, that haven’t been molecularly mapped to the sequence. Use Gbrowse when looking for molecularly mapped sequences, insertions, or Affymetrix probes.

There are two main query tools in FlyBase. The first main query tool is called Jump to Gene (J2G). This is found in the top right of the blue navigation bar on every page of FlyBase. This tool is useful when you know exactly what you are looking for and want to go to the report page with that data. The second main query tool is called QuickSearch. This is located on the FlyBase homepage. This tool is most useful when you want to look up something quickly that you may only know a little about. Searching can be performed within D. melanogaster only or within all species. Data other than genes can be searched using the ‘data class’ menu.

Online Mendelian Inheritance in Man

Online Mendelian Inheritance in Man (OMIM) is a continuously updated catalog of human genes and genetic disorders and traits, with a particular focus on the gene-phenotype relationship. As of 12 February 2017, approximately 8,425 of the over 23,000 entries in OMIM represented phenotypes; the rest represented genes, many of which were related to known phenotypes. OMIM is the online continuation of Dr. Victor McKusick's *Mendelian Inheritance in Man* (MIM), which was published in 12 editions between 1966 and 1998. Nearly all of the 1,486 entries in the first edition of MIM discussed phenotypes.

MIM/OMIM is produced and curated at the Johns Hopkins University School of Medicine (JHUSOM). OMIM became available on the internet in 1987 under the direction of the Welch Medical Library at JHUSOM with financial support from the Howard Hughes Medical Institute. From 1995 to 2010, OMIM was available on the World Wide Web with informatics and financial support from the National Center for Biotechnology Information. The current OMIM website (OMIM.org), which was developed with funding from JHUSOM, is maintained by Johns Hopkins University with financial support from the National Human Genome Research Institute. The content of MIM/OMIM is based on selection and review of the published peer-reviewed biomedical literature. Updating of content is performed by a team of science writers and curators under the direction of Dr. Ada Hamosh at the McKusick-Nathans Institute of Genetic Medicine of Johns Hopkins University. While OMIM is freely available to the public, it is designed for use primarily by physicians and other health care professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine.

The database may be used as a resource for locating literature relevant to inherited conditions, and its numbering system is widely used in the medical literature to provide a unified index for genetic diseases.

MIM classification system - MIM numbers

Each OMIM entry is given a unique six-digit identifier as summarized below:

- 100000–299999: Autosomal loci or phenotypes (entries created before May 15, 1994)
- 300000–399999: X-linked loci or phenotypes
- 400000–499999: Y-linked loci or phenotypes
- 500000–599999: Mitochondrial loci or phenotypes

600000 and above: Autosomal loci or phenotypes (entries created after May 15, 1994)

In cases of allelic heterogeneity, the MIM number of the entry is followed by a decimal point and a unique 4-digit number specifying the variant. For example, allelic variants in the HBB gene (141900) are numbered 141900.0001 through 141900.0538.

Because OMIM has responsibility for the classification and naming of genetic disorders, these numbers are stable identifiers of the disorders.

Symbols preceding MIM numbers

Symbols preceding MIM numbers indicate the entry category:

- An asterisk (*) before an entry number indicates a gene.
- A number symbol (#) before an entry number indicates that it is a descriptive entry, usually of a phenotype, and does not represent a unique locus. The reason for the use of the number symbol is given in the first paragraph of the entry. Discussion of any gene(s) related to the phenotype resides in another entry (or entries) as described in the first paragraph.
- A plus sign (+) before an entry number indicates that the entry contains the description of a gene of known sequence and a phenotype.

- A percent sign (%) before an entry number indicates that the entry describes a confirmed Mendelian phenotype or phenotypic locus for which the underlying molecular basis is not known.
- No symbol before an entry number generally indicates a description of a phenotype for which the Mendelian basis, although suspected, has not been clearly established or that the separateness of this phenotype from that in another entry is unclear.
- A caret (^) before an entry number means the entry no longer exists because it was removed from the database or moved to another entry as indicated.

Possible questions

1 Mark questions

1. The human genome consists of _____________ base pairs.
a) two billion b) ten billion c) one billion d) three billion

2. Large datasets can be retrieved using the ___________ tool.
a) BioMart b) SQL c) NCBI d) FASTA

3. _________ can be used to combines transcription factor binding sites database search with a comparative sequence analysis.
a) rVISTA b) mVISTA c) wgVISTA d) phylo-VISTA

4. ENSG### is _______
a) Ensembl Exon ID b) Ensembl Gene ID c) Ensembl Transcript ID d) Ensembl Peptide ID

5. ENSE### is _______
a) Ensembl Exon ID b) Ensembl Gene ID c) Ensembl Transcript ID d) Ensembl Peptide ID

6. ___is an algorithm used for calculating sequence similarity between biological sequences.
a) VISTA b) NCBI c) BLAST d) FASTA

7. Protein-coding genes can be identified by
a) Transposon taggingb) ORF scanning c) Zoo-blotting d)Nuclease S1 mapping

8. ORF scanning is used
a) to find exons b) to find intergenic sequences c) to find gene homologies
d) to find protein-coding genes

9. Human Genome Project was first initiated in the year ------
a) 1965 b) 1970 c) 1975 d) 1980

10. A ________ transplant was used to overcome this genetic disorder.
a) Liver b) Kidney c) Bone marrow d) none of the above

11. Ensembl database focuses on _____.
a) Human b) Invertebrates c) vertebrates d) all the above

2 Marks questions

1. Write a short note on 'UCSC webserver'.

2. Write any two applications of NCBI database.

3. What are databases? Give an example.

4. What are genome-specific databases? Give an example.

6/8 Marks questions

1. Explain various applications of the 'ENSEMBLE' database on analysing genome data.

2. Describe the use of 'VISTA' webserver in the analyses of genome data.

3. Enumerate unique features of 'NCBI' database for retrieving genome data.

4. Describe any two organisms – specific genome database in a detailed manner.

UNIT - III

SYLLABUS

Genomic mapping: Genetic markers **-** VNTR, mini and micro satellites, STS, SNP, ESTs. Types of genome maps, Mapping techniques – Physical and genetic mapping, Map resources, Practical uses genome maps.

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 markersν

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/

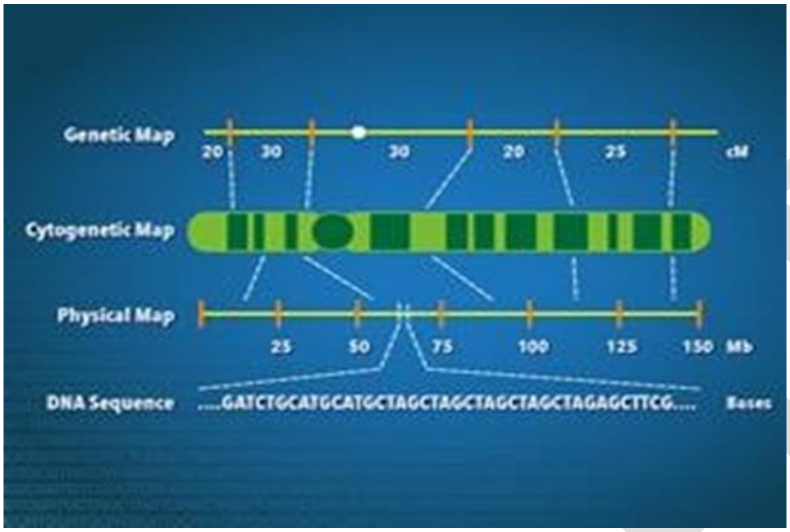
Biochemical markers in Human

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

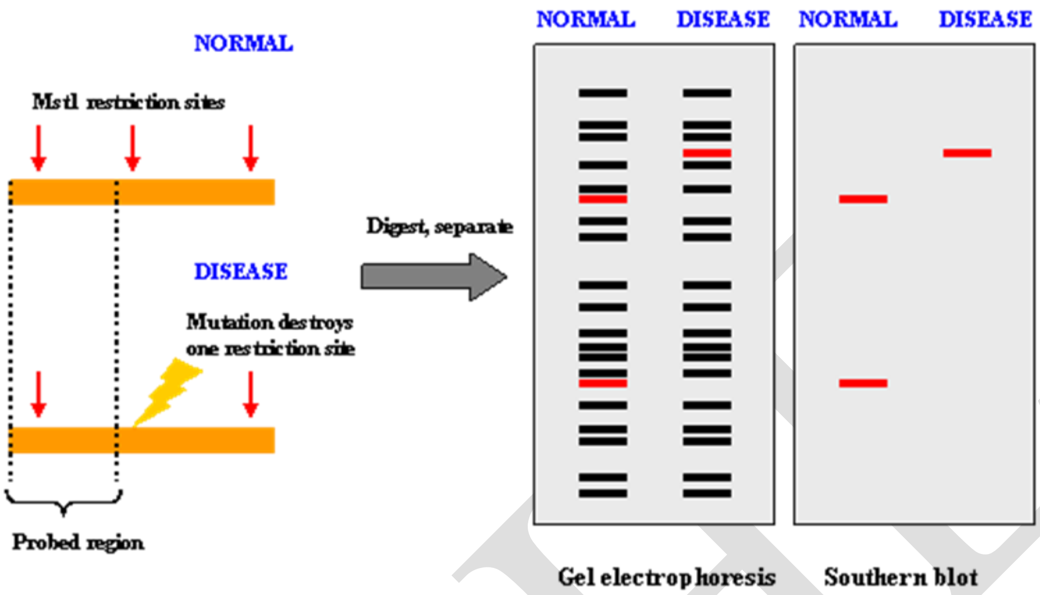
Total DNA is digested with a methylation-sensitive enzyme (for example, *Pst*I), thereby enriching the library for single- or low-copy expressed sequences (*Pst*I clones are based on the suggestion that expressed genes are not methylated).

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 tetra-nucleotide 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:

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

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

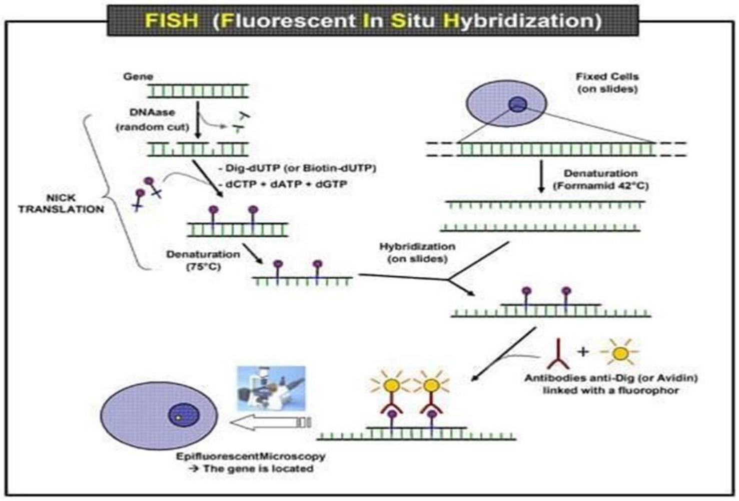
FISH is useful for example, to help a researcher identify where a particular gene falls within an individual's chromosomes. Here's how it works:

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.



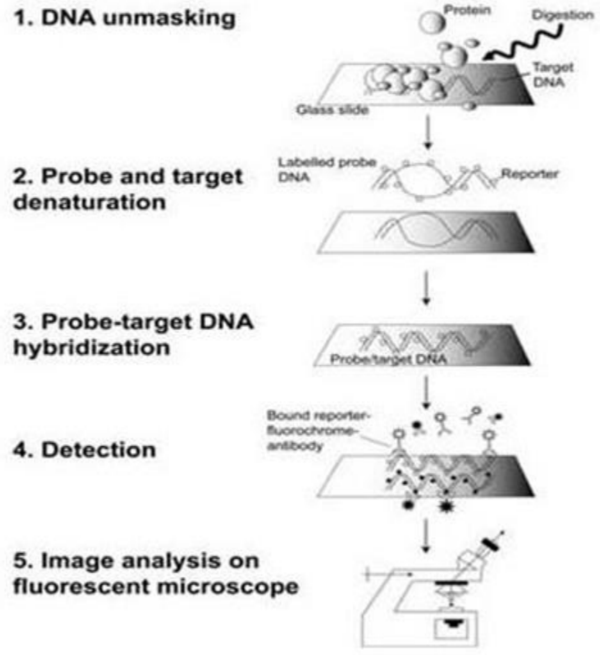

1. DNA unmasking
Protein
Digestion
Target DNA
Glass slide
2. Probe and target denaturation
Labelled probe DNA
Reporter
3. Probe-target DNA hybridization
Probe/target DNA
4. Detection
Bound reporter-fluorochrome-antibody
5. Image analysis on fluorescent microscope

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:

Locus specific probes:

It binds 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.

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.

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)

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. Inaddition 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.

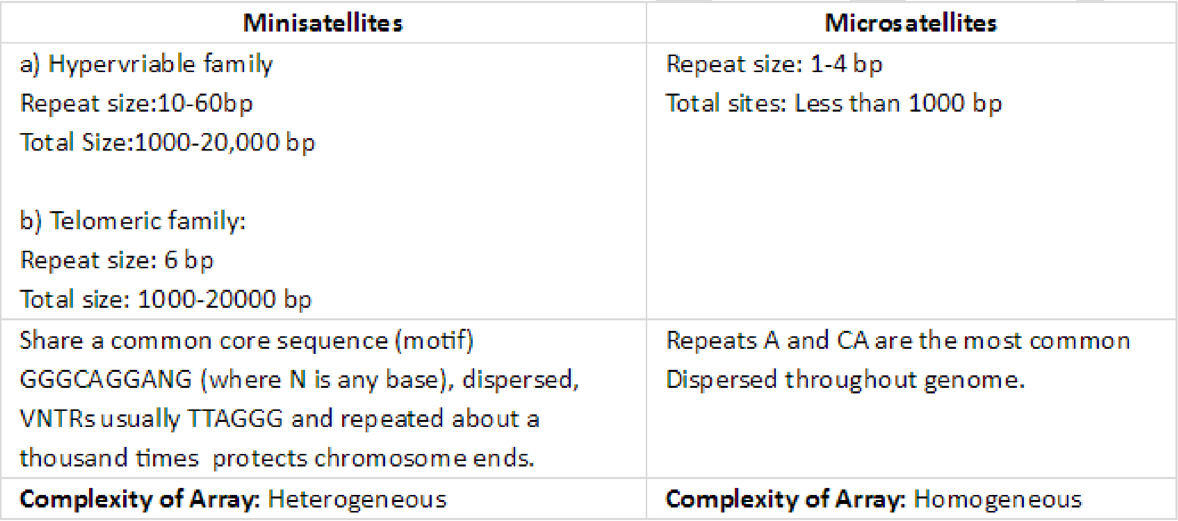
The methodologies for the development of SSR loci, genotyping and their applications as molecular markers are also reviewed.

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 Mark questions

1. Microsatellites are
a) frequently found in bacterial genomes b) always smaller than 10 bp
c) used as DNA markers d) movable DNA elements

2. Molecular markers are used to construct
a) Chromosome maps b) cytogenetic maps c) physical maps d) geographic maps

3. The variation in number of tandem repeats between two or more individuals is called
a) VNTRs b) RFLP c) SSRs d) AFLP

4. The variant fragment that distinguish one individual from another one is called
a) variant fragment b) marking fragment c) differing fragment d) variable repeats

5. Which of these is a key characterisitic of a molecular marker?
a) It is a known gene b) It is located at a known site on the chromosome
c) It is only useful for linkage and physical mapping studies d) positional analysis

6. A polymorphism is -----
a) Any change in the DNA sequence b) The most common variation of a gene or marker sequence
c) The least common vaariation of a gene or marker sequence
d) Variation of gene or marker sequence present in > 1% of the population

7. The shotgun method ------
a) is used in analyzing transcriptomes b) requires computers
c) is normally used with large genomes d) is more accurate than clone contig method

8. A monomorphic DNA segment is
a) A segment of DNA that exists in many forms in the population
b) A segment of DNA that controls a single gene function
c) A segment of DNA inherited in a dominant fashion
d) A segment of DNA shared by over 99 % of the population

9. Which of these describes a contig?
a) A complete genomci library including overlapping clones
b) A complete mRNA library
c) A chromosome specific library of overlapping clones
d) An ordered genomci library

10. The variation in the restriction DNA fragment lengths between individuals of a species is called
a) Retriction fragment length polymorphism b) RAPD
c) AFLP d) simple sequence repeats

11. All the following statements are true regarding RFLP and RAPD except
a) RAPD is a quick method compared to RFLP b) RFLP is more relible than RAPD
c) Species specific primers are required for RAPD d) Radioactive probes are not required in RAPD

2 Marks question

1. Write a short note on ‘mini satellite.

2. What are ESTs?

3. What are the types of genome maps?

4. What is RFLP?

5. What is SNP?

6/8 Marks questions

1. What are genetic markers? Explain with an example.

2. Describe genetic mapping techniques and their applications in detail.

3. Describe physical mapping techniques and their applications in detail.

4. Briefly explain uses of SNP.

5. How do ‘mini satellites’ differ from ‘micro satellites’?

UNIT - IV

SYLLABUS

Protein and structure determination: Introduction to protein structure, Chemical properties of proteins. Physical interactions that determine the property of proteins. Short-range interactions, electrostatic forces, van der Waal interactions, hydrogen bonds, Hydrophobic interactions. Determination of sizes - Sedimentation analysis, gel filteration, Native PAGE, SDS-PAGE. Determination of covalent structures – Edman degradation.

Three-dimensional Structures of proteins

The properties of a protein are largely determined by its three-dimensional structure. One might naively suppose that since proteins are all composed of the same 20 types of amino acid residues, they would be more or less alike in their properties. Indeed, denatured (unfolded) proteins have rather similar characteristics, a kind of homogeneous "average" of their randomly dangling side chains. How- ever, the three-dimensional structure of a native (physio- logically folded) protein is specified by its primary structure, so that it has a unique set of characteristics. we shall discuss the structural features of proteins, the forces that hold them together, and their hierarchical organization to form complex structures. This will form the basis for understanding the structure–func- tion relationships necessary to comprehend the biochemical roles of proteins.

Amino acids

Amino acids contain both amino and carboxylic acid functional groups. (In biochemistry, the term amino acid is used when referring to those amino acids in which the amino and carboxylate functionalities are attached to the same carbon, plus proline which is not actually an amino acid). Modified amino acids are sometimes observed in proteins; this is usually the result of enzymatic modification after translation (protein synthesis). For example, phosphorylation of serine by kinases and dephosphorylation by phosphatases is an important control mechanism in the cell cycle. Only two amino acids other than the standard twenty are known to be incorporated into proteins during translation, in certain organisms:

- Selenocysteine is incorporated into some proteins at a UGA codon, which is normally a stop codon.
- Pyrrolysine is incorporated into some proteins at a UAG codon. For instance, in some methanogens in enzymes that are used to produce methane.

Besides those used in protein synthesis, other biologically important amino acids include carnitine (used in lipid transport within a cell), ornithine, GABA and taurine.

Protein structure

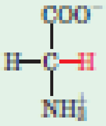
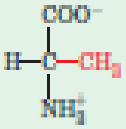
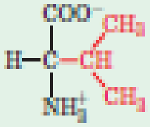
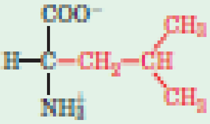
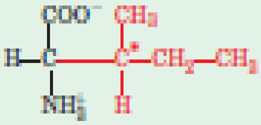
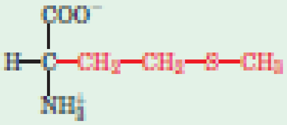
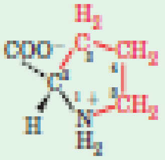
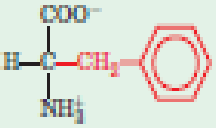
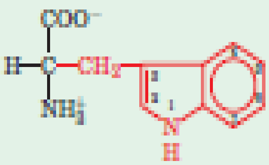
The particular series of amino acids that form a protein is known as that protein's primary structure. This sequence is determined by the genetic makeup of the individual. It specifies the order of side-chain groups along the linear polypeptide "backbone".

Proteins have two types of well-classified, frequently occurring elements of local structure defined by a particular pattern of hydrogen bonds along the backbone: alpha helix and beta sheet. Their number and arrangement is called the secondary structure of the protein. Alpha helices are regular spirals stabilized by hydrogen bonds between the backbone CO group (carbonyl) of one amino acid residue and the backbone NH group (amide) of the i+4 residue. The spiral has about 3.6 amino acids per turn, and the amino acid side chains stick out from the cylinder of the helix. Beta pleated sheets are formed by backbone hydrogen bonds between individual beta strands each of which is in an "extended", or fully stretched-out, conformation. The strands may lie parallel or antiparallel to each other, and the side-chain direction alternates above and below the sheet. Hemoglobin contains only helices, natural silk is formed of beta pleated sheets, and many enzymes have a pattern of alternating helices and beta-strands. The secondary-structure elements are connected by "loop" or "coil" regions of non-repetitive conformation, which are sometimes quite mobile or disordered but usually adopt a well-defined, stable arrangement. The overall, compact, 3D structure of a protein is termed its tertiary structure or its "fold". It is formed as result of various attractive forces like hydrogen bonding, disulfide bridges, hydrophobic interactions, hydrophilic interactions, van der Waals force etc. When two or more polypeptide chains (either of identical or of different sequence) cluster to form a protein, quaternary structure of protein is formed. Quaternary structure is an attribute of polymeric (same-sequence chains) or heteromeric (different-sequence chains) proteins like hemoglobin, which consists of two "alpha" and two "beta" polypeptide chains.

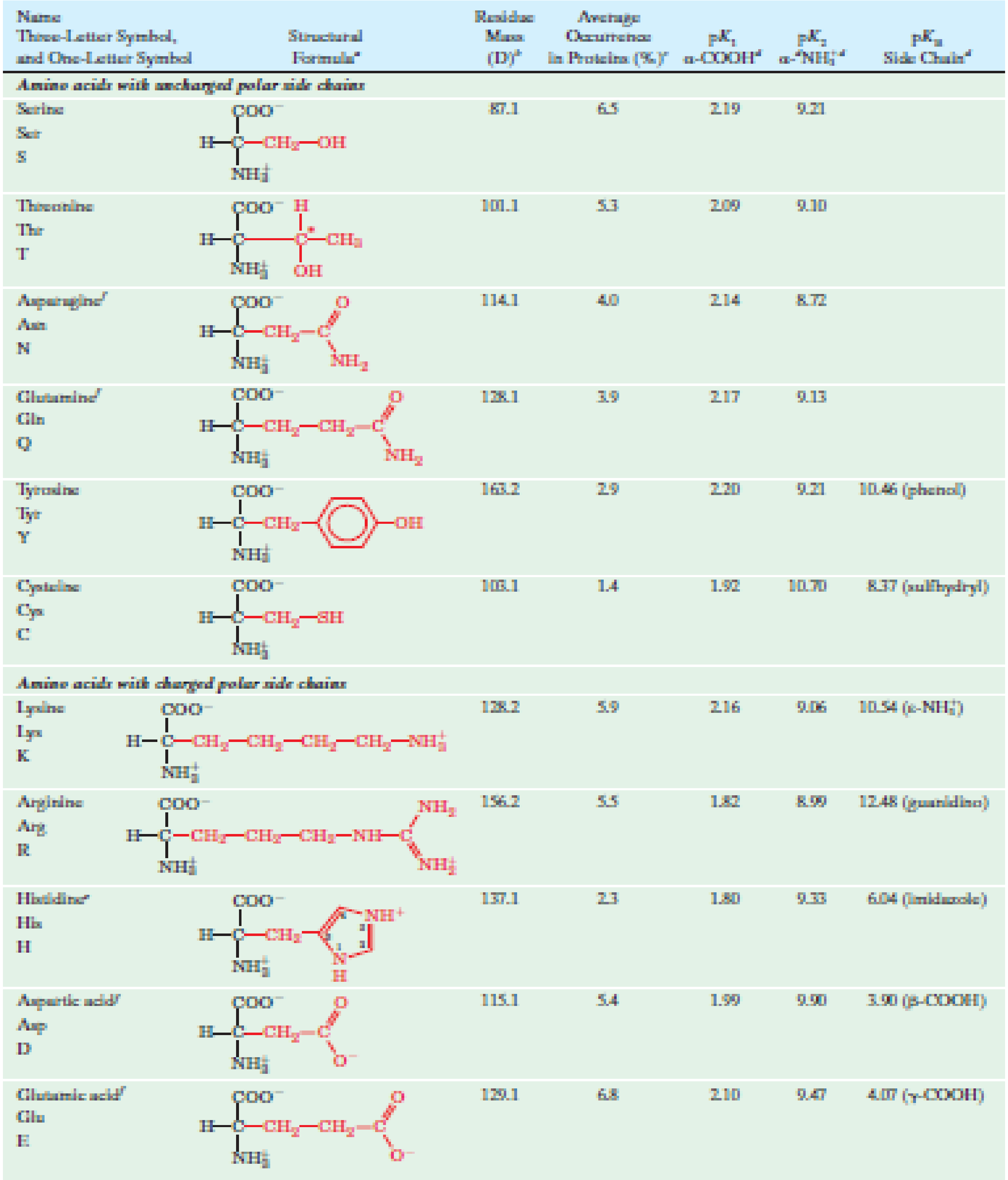
Apoenzymes

An apoenzyme (or, generally, an apoprotein) is the protein without any small-molecule cofactors, substrates, or inhibitors bound. It is often important as an inactive storage, transport, or secretory form of a protein. This is required, for instance, to protect the secretory cell from the activity of that protein. Apoenzymes becomes active enzymes on addition of a cofactor. Cofactors can be either inorganic (e.g., metal ions and iron-sulfur clusters) or organic compounds, (e.g., flavin and heme). Organic cofactors can be either prosthetic groups, which are tightly bound to an enzyme, or coenzymes, which are released from the enzyme's active site during the reaction.

values of their ionizable groups

Name, Three-Letter Symbol, and One-Letter Symbol	Structural Formula^a	Residue Mass (D)^b	Average Occurrence in Proteins (%)^c	pK_1 α-COOH^d	pK_2 α-NH_3^+^{d}	pK_R Side Chain^d
Amino acids with nonpolar side chains						
Glycine Gly G		57.0	7.1	2.35	9.78	
Alanine Ala A		71.1	8.3	2.35	9.87	
Valine Val V		99.1	6.9	2.29	9.74	
Leucine Leu L		113.2	9.7	2.33	9.74	
Isoleucine Ile I		113.2	6.0	2.32	9.76	
Methionine Met M		131.2	2.4	2.13	9.28	
Proline Pro P		97.1	4.7	1.95	10.64	
Phenylalanine Phe F		147.2	3.9	2.20	9.31	
Tryptophan Trp W		186.2	1.1	2.46	9.41	

(continued)



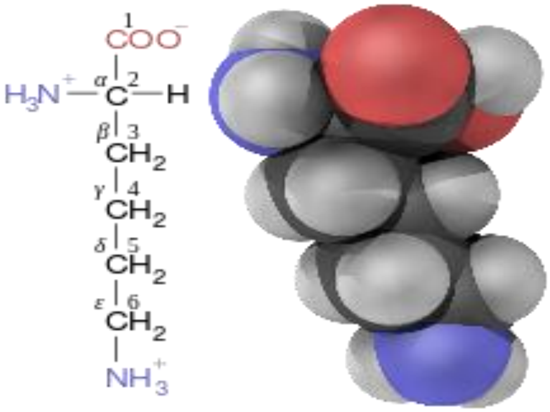
Name Three-Letter Symbol, and One-Letter Symbol	Structural Formula[a]	Residue Mass (D)[b]	Average Occurrence in Proteins (%)[c]	pK_1 α-COOH[d]	pK_2 α-NH_3^+[d]	pK_R Side Chain[d]
Amino acids with uncharged polar side chains						
Serine Ser S		87.1	6.5	2.19	9.21	
Threonine Thr T		101.1	5.3	2.09	9.10	
Asparagine[f] Asn N		114.1	4.0	2.14	8.72	
Glutamine[f] Gln Q		128.1	3.9	2.17	9.13	
Tyrosine Tyr Y		163.2	2.9	2.20	9.21	10.46 (phenol)
Cysteine Cys C		103.1	1.4	1.92	10.70	8.37 (sulfhydryl)
Amino acids with charged polar side chains						
Lysine Lys K		128.2	5.9	2.16	9.06	10.54 (ε-NH_3^+)
Arginine Arg R		156.2	5.5	1.82	8.99	12.48 (guanidino)
Histidine[e] His H		137.1	2.3	1.80	9.33	6.04 (imidazole)
Aspartic acid[f] Asp D		115.1	5.4	1.99	9.90	3.90 (β-COOH)
Glutamic acid[f] Glu E		129.1	6.8	2.10	9.47	4.07 (γ-COOH)

In the structure shown at the top of the page, R represents a side chain specific to each amino acid. The carbon atom next to the carboxyl group (which is therefore numbered 2 in the carbon chain starting from that functional group) is called the α–carbon. Amino acids containing an amino group bonded directly to the alpha carbon are referred to as *alpha amino acids*. These include amino acids such as proline which contain secondary amines, which used to be often referred to as "imino acids".

Isomerism

The alpha amino acids are the most common form found in nature, but only when occurring in the L-isomer. The alpha carbon is a chiral carbon atom, with the exception of glycine which has two indistinguishable hydrogen atoms on the alpha carbon. Therefore, all alpha amino acids but glycine can exist in either of two enantiomers, called L or D amino acids, which are mirror images of each other (*see also Chirality*). While L-amino acids represent all of the amino acids found in proteins during translation in the ribosome, D-amino acids are found in some proteins produced by enzyme posttranslational modifications after translation and translocation to the endoplasmic reticulum, as in exotic sea-dwelling organisms such as cone snails. They are also abundant components of the peptidoglycan cell walls of bacteria,[36] and D-serine may act as a neurotransmitter in the brain. D-amino acids are used in racemic crystallography to create centrosymmetric crystals, which (depending on the protein) may allow for easier and more robust protein structure determination. The L and D convention for amino acid configuration refers not to the optical activity of the amino acid itself but rather to the optical activity of the isomer of glyceraldehyde from which that amino acid can, in theory, be synthesized (D-glyceraldehyde is dextrorotatory; L-glyceraldehyde is levorotatory). In alternative fashion, the *(S)* and *(R)* designators are used to indicate the absolute stereochemistry. Almost all of the amino acids in proteins are *(S)* at the α carbon, with cysteine being *(R)* and glycine non-chiral.[39] Cysteine has its side chain in the same geometric position as the other amino acids, but the *R/S* terminology is reversed because of the higher atomic number of sulfur compared to the carboxyl oxygen gives the side chain a higher priority, whereas the atoms in most other side chains give them lower priority.

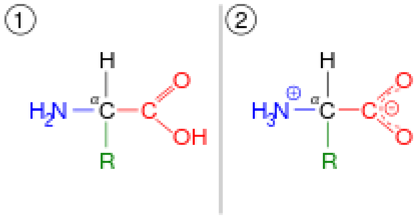
Side chains



Lysine with carbon atoms labeled by position

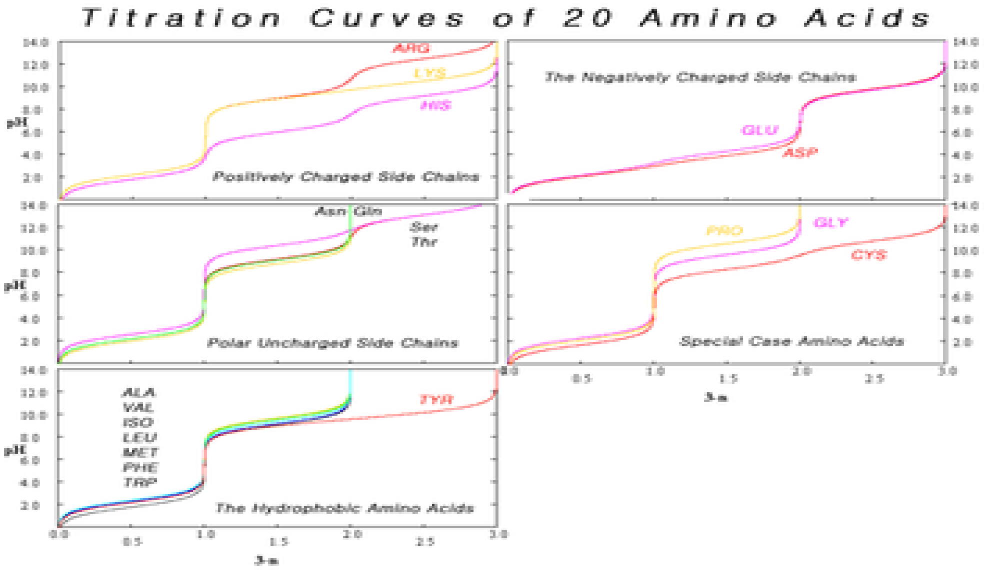
In amino acids that have a carbon chain attached to the α–carbon (such as lysine, shown to the right) the carbons are labeled in order as α, β, γ, δ, and so on. In some amino acids, the amine group is attached to the β or γ-carbon, and these are therefore referred to as *beta* or *gamma amino acids*. Amino acids are usually classified by the properties of their side chain into four groups. The side chain can make an amino acid a weak acid or a weak base, and a hydrophile if the side chain is polar or a hydrophobe if it is nonpolar. The chemical structures of the 22 standard amino acids, along with their chemical properties, are described more fully in the article on these proteinogenic amino acids. The phrase "branched-chain amino acids" or BCAA refers to the amino acids having aliphatic side chains that are non-linear; these are leucine, isoleucine, and valine. Proline is the only proteinogenic amino acid whose side-group links to the α-amino group and, thus, is also the only proteinogenic amino acid containing a secondary amine at this position.[34] In chemical terms, proline is, therefore, an imino acid, since it lacks a primary amino group,[41] although it is still classed as an amino acid in the current biochemical nomenclature,[42] and may also be called an "N-alkylated alpha-amino acid".

Zwitterions



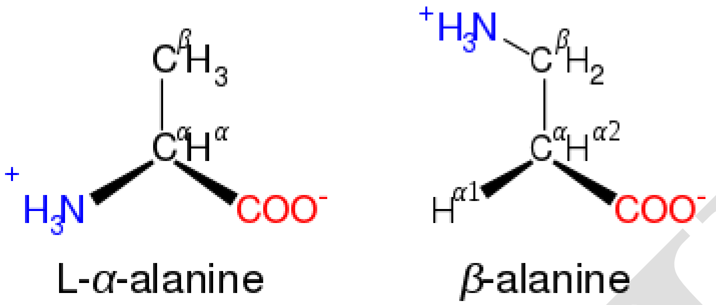
The α-carboxylic acid group of amino acids is a weak acid, meaning that it releases a hydron (such as a proton) at moderate pH values. In other words, carboxylic acid groups (−CO$_2$H) can be deprotonated to become negative carboxylates (−CO$_2^-$). The negatively charged carboxylate ion predominates at pH values greater than the pKa of the carboxylic acid group (mean for the 20 common amino acids is about 2.2, see the table of amino acid structures above). In a complementary fashion, the α-amine of amino acids is a weak base, meaning that it accepts a proton at moderate pH values. In other words, α-amino groups (NH$_2$−) can be protonated to become positive α-ammonium groups ($^+$NH$_3$−). The positively charged α-ammonium group predominates at pH values less than the pKa of the α-ammonium group (mean for the 20 common α-amino acids is about 9.4). Because all amino acids contain amine and carboxylic acid functional groups, they share amphiprotic properties. Below pH 2.2, the predominant form will have a neutral carboxylic acid group and a positive α-ammonium ion (net charge +1), and above pH 9.4, a negative carboxylate and neutral α-amino group (net charge −1). But at pH between 2.2 and 9.4, an amino acid usually contains both a negative carboxylate and a positive α-ammonium group, as shown in structure (2) on the right, so has net zero charge. This molecular state is known as a zwitterion, from the German Zwitter meaning *hermaphrodite* or *hybrid*. The fully neutral form (structure (1) on the left) is a very minor species in aqueous solution throughout the pH range (less than 1 part in 10^7). Amino acids exist as zwitterions also in the solid phase, and crystallize with salt-like properties unlike typical organic acids or amines.

Isoelectric point



Composite of titration curves of twenty proteinogenic amino acids grouped by side chain category.The variation in titration curves when the amino acids can be grouped by category. With the exception of tyrosine, using titration to distinguish among hydrophobic amino acids is problematic. At pH values between the two pKa values, the zwitterion predominates, but coexists in dynamic equilibrium with small amounts of net negative and net positive ions. At the exact midpoint between the two pKa values, the trace amount of net negative and trace of net positive ions exactly balance, so that average net charge of all forms present is zero. This pH is known as the isoelectric point pI, so pI = ½(pKa_1 + pKa_2). The individual amino acids all have slightly different pKa values, so have different isoelectric points. For amino acids with charged side chains, the pKa of the side chain is involved. Thus for Asp, Glu with negative side chains, pI = ½(pKa_1 + pKa_R), where pKa_R is the side chain pKa. Cysteine also has potentially negative side chain with pKa_R = 8.14, so pI should be calculated as for Asp and Glu, even though the side chain is not significantly charged at neutral pH. For His, Lys, and Arg with positive side chains, pI = ½(pKa_R + pKa_2). Amino acids have zero mobility in electrophoresis at their isoelectric point, although this behaviour is more usually exploited for peptides and proteins than single amino acids. Zwitterions have minimum solubility at their isoelectric point and some amino acids (in particular, with non-polar side chains) can be isolated by precipitation from water by adjusting the pH to the required isoelectric point.

Occurrence and functions in biochemistry



β-alanine and its α-alanine isomer

Proteinogenic amino acids

Amino acids are the structural units (monomers) that make up proteins. They join together to form short polymer chains called peptides or longer chains called either polypeptides or proteins. These polymers are linear and unbranched, with each amino acid within the chain attached to two neighboring amino acids. The process of making proteins encoded by DNA/RNA genetic material is called *translation* and involves the step-by-step addition of amino acids to a growing protein chain by a ribozyme that is called a ribosome. The order in which the amino acids are added is read through the genetic code from an mRNA template, which is a RNA copy of one of the organism's genes. Twenty-two amino acids are naturally incorporated into polypeptides and are called proteinogenic or natural amino acids. Of these, 20 are encoded by the universal genetic code. The remaining 2, selenocysteine and pyrrolysine, are incorporated into proteins by unique synthetic mechanisms. Selenocysteine is incorporated when the mRNA being translated includes a SECIS element, which causes the UGA codon to encode selenocysteine instead of a stop codon. Pyrrolysine is used by some methanogenic archaea in enzymes that they use to produce methane. It is coded for with the codon UAG, which is normally a stop codon in other organisms.[48] This UAG codon is followed by a PYLIS downstream sequence.

Non-proteinogenic amino acids

Aside from the 22 proteinogenic amino acids, many *non-proteinogenic* amino acids are known. Those either are not found in proteins (for example carnitine, GABA, Levothyroxine) or are not produced directly and in isolation by standard cellular machinery (for example, hydroxyproline and selenomethionine). Non-proteinogenic amino acids that are found in proteins are formed by post-translational modification, which is modification after translation during protein synthesis. These modifications are often essential for the function or regulation of a protein. For example, the carboxylation of glutamate allows for better binding of calcium cations, and collagen contains hydroxyproline, generated by hydroxylation of proline. Another example is the formation of hypusine in the translation initiation factor EIF5A, through modification of a lysine residue. Such modifications can also determine the localization of the protein, e.g., the addition of long hydrophobic groups can cause a protein to bind to a phospholipid membrane.

Non-standard amino acids

The 20 amino acids that are encoded directly by the codons of the universal genetic code are called *standard* or *canonical* amino acids. A modified form of methionine (*N*-formylmethionine) is often incorporated in place of methionine as the initial amino acid of proteins in bacteria, mitochondria and chloroplasts. Other amino acids are called *non-standard* or *non-canonical*. Most of the non-standard amino acids are also non-proteinogenic (i.e. they cannot be incorporated into proteins during translation), but two of them are proteinogenic, as they can be incorporated translationally into proteins by exploiting information not encoded in the universal genetic code. The two non-standard proteinogenic amino acids are selenocysteine (present in many non-eukaryotes as well as most eukaryotes, but not coded directly by DNA) and pyrrolysine (found only in some archaea and one bacterium). The incorporation of these non-standard amino acids is rare. For example, 25 human proteins include selenocysteine (Sec) in their primary structure, and the structurally characterized enzymes (selenoenzymes) employ Sec as the catalytic moiety in their active sites. Pyrrolysine and selenocysteine are encoded via variant codons. For example, selenocysteine is encoded by stop codon and SECIS element.

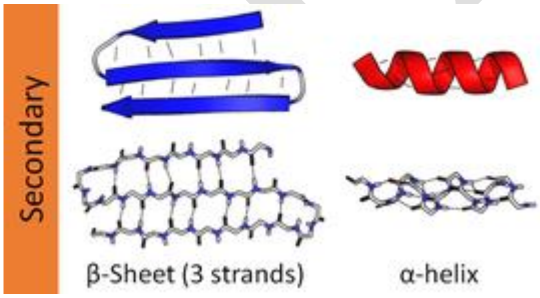
Secondary structure

Protein secondary structure is the three dimensional form of *local segments* of proteins. The two most common secondary structural elements are alpha helices and beta sheets, though beta turns and omega loops occur as well. Secondary structure elements typically spontaneously form as an intermediate before the protein folds into its three dimensional tertiary structure. Secondary structure is formally defined by the pattern of hydrogen bonds between the amino hydrogen and carboxyl oxygen atoms in the peptide backbone. Secondary structure may alternatively be defined based on the regular pattern of backbone dihedral angles in a particular region of the Ramachandran plot regardless of whether it has the correct hydrogen bonds.

The concept of secondary structure was first introduced by Kaj Ulrik Linderstrøm-Lang at Stanford in 1952. Other types of biopolymers such as nucleic acids also possess characteristic secondary structures.

Structural features of the three major forms of protein helices

Geometry attribute	α-helix	3_{10} helix	π-helix
Residues per turn	3.6	3.0	4.4
Translation per residue	1.5 Å (0.15 nm)	2.0 Å (0.20 nm)	1.1 Å (0.11 nm)
Radius of helix	2.3 Å (0.23 nm)	1.9 Å (0.19 nm)	2.8 Å (0.28 nm)
Pitch	5.4 Å (0.54 nm)	6.0 Å (0.60 nm)	4.8 Å (0.48 nm)



The most common secondary structures are alpha helices and beta sheets. Other helices, such as the 3_{10} helix and π helix, are calculated to have energetically favorable hydrogen-bonding patterns but are rarely observed in natural proteins except at the ends of α helices due to unfavorable backbone packing in the center of the helix. Other extended structures such as the polyproline helix and alpha sheet are rare in native state proteins but are often hypothesized as important protein folding intermediates. Tight turns and loose, flexible loops link the more "regular" secondary structure elements. The random coil is not a true secondary structure, but is the class of conformations that indicate an absence of regular secondary structure. Amino acids vary in their ability to form the various secondary structure elements. Proline and glycine are sometimes known as "helix breakers" because they disrupt the regularity of the α helical backbone conformation; however, both have unusual conformational abilities and are commonly found in turns. Amino acids that prefer to adopt helical conformations in proteins include methionine, alanine, leucine, glutamate and lysine ("MALEK" in amino-acid 1-letter codes); by contrast, the large aromatic residues (tryptophan, tyrosine and phenylalanine) and C^{β}-branched amino acids (isoleucine, valine, and threonine) prefer to adopt β-strand conformations. However, these preferences are not strong enough to produce a reliable method of predicting secondary structure from sequence alone. Low frequency collective vibrations are thought to be sensitive to local rigidity within proteins, revealing beta structures to be generically more rigid than alpha or disordered proteins. Neutron scattering measurements have directly connected the spectral feature at ~1 THz to collective motions of the secondary structure of beta-barrel protein GFP. Hydrogen bonding patterns in secondary structures may be significantly distorted, which makes automatic determination of secondary structure difficult. There are several methods for formally defining protein secondary structure (e.g., DSSP, DEFINE, STRIDE, ScrewFit, SST).

Protein tertiary structure is the three dimensional shape of a protein. The tertiary structure will have a single polypeptide chain "backbone" with one or more protein secondary structures, the protein domains. Amino acid side chains may interact and bond in a number of ways. The interactions and bonds of side chains within a particular protein determine its tertiary structure. The protein tertiary structure is defined by its atomic coordinates. These coordinates may refer either to a protein domain or to the entire tertiary structure. A number of tertiary structures may fold into a quaternary structure.

SECONDARY STRUCTURE

A polymer's secondary structure (2° structure) is defined as the local conformation of its backbone. For proteins, this has come to mean the specification of regular polypeptide backbone folding patterns: helices, pleated sheets, and turns. However, before we begin our discussion of these basic structural motifs, let us consider the geometrical properties of the peptide group because its understanding is prerequisite to that of any structure containing it.

The Peptide Group

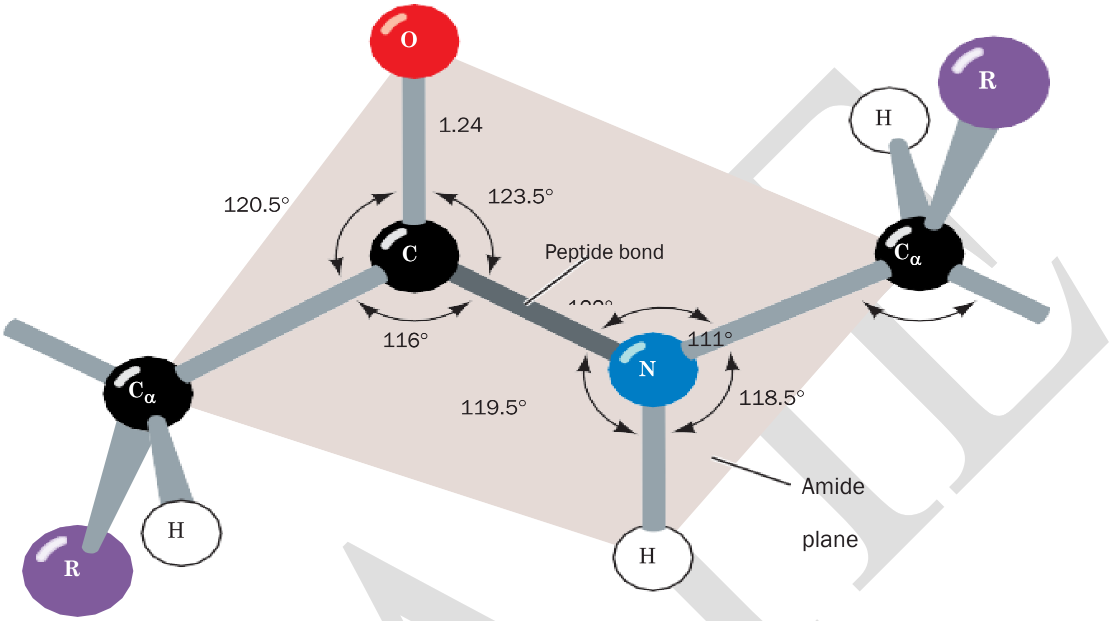
In the 1930s and 1940s, Linus Pauling and Robert Corey determined the X-ray structures of several amino acids and dipeptides in an effort to elucidate the structural con- straints on the conformations of a polypeptide chain. These studies indicated that *the peptide group has a rigid, planar structure, which, Pauling pointed out, is a conse- quence of resonance interactions that give the peptide bond an ~40% double-bond character*.

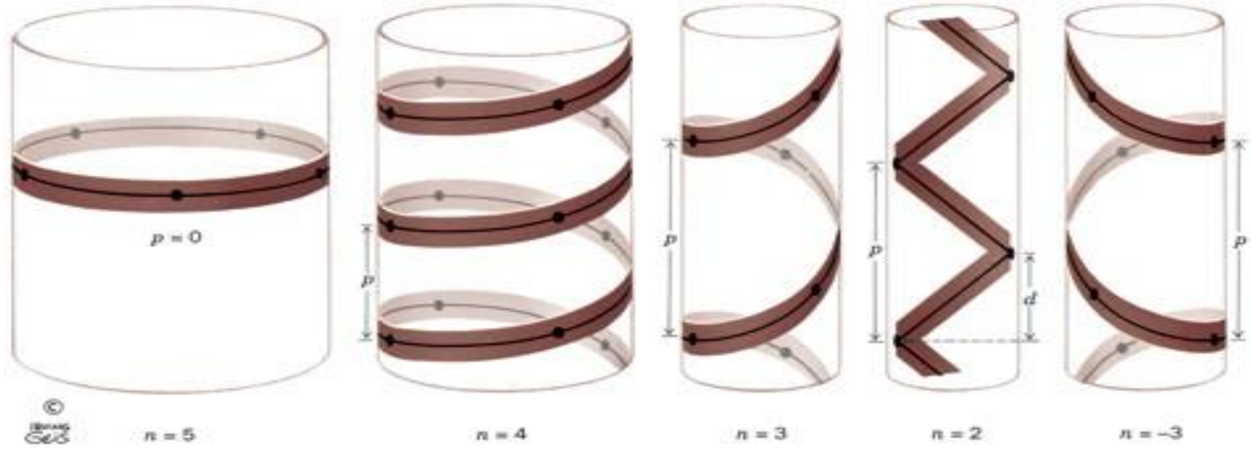
This explanation is supported by the observations that a peptide's C¬N bond is 0.13 Å shorter than its N¬C_a sin- gle bond and that its C˂˂O bond is 0.02 Å longer than that of aldehydes and ketones. The peptide bond's resonance energy has its maximum value, ~85 kJ · mol$^{—1}$, when the peptide group is planar because its g-bonding overlap is maximized in this conformation. This overlap, and thus the resonance energy, falls to zero as the peptide bond is twisted to 90° out of planarity, thereby accounting for the planar peptide group's rigidity. (The positive charge on the above resonance structure should be taken as a formal charge; quantum mechanical calculations indicate that the peptide N atom, in fact, has a partial negative charge aris- ing from the polarization of the C¬N bond.). *Peptide groups, with few exceptions, assume the trans conformation: that in which successive C_a atoms are on op- posite sides of the peptide bond joining them.* This is partly a result of steric interference, which causes the cis conformation to be ~8 kJ ·

mol$^{—1}$ less stable than the trans conformation (this energy difference is somewhat less in peptide bonds followed by a Pro residue and, in fact,~10% of the Pro residues in proteins follow a cis peptide bond, whereas cis peptides are otherwise extremely rare).

Polypeptide Backbone Conformations May Be Described by Their Torsion Angles

The above considerations are important because they indicate that *the backbone of a protein is a linked sequence of rigid planar peptide groups*. We can therefore specify a polypeptide's backbone conformation by the torsion angles (rotation angles or dihedral angles) about the C$_a$¬N bond (\$) and the C$_a$¬C bond (†) of each of its amino acid residues. These angles, \$ and †, are both defined as 180° when the polypeptide chain is in its planar, fully extended (all-trans) conformation and increase for a clockwise rotation when viewed from C$_a$. There are several steric constraints on the torsion an- gles, \$ and †, of a polypeptide backbone that limit its con- formational range. The electronic structure of a single (o) bond, such as a C¬C bond, is cylindrically symmetrical about its bond axis, so that we might expect such a bond to exhibit free rotation. If this were the case, then in ethane, for example, all torsion angles about the C¬C bond would be equally likely. Yet certain conformations in ethane are favored due to quantum mechanical effects arising from the interactions of its molecular orbitals. The staggered conformation (torsion angle = 180°) is ethane's most stable arrangement, whereas the eclipsed conforma- tion (torsion angle = 0°) is least stable. The energy difference between the staggered and eclipsed con- formations in ethane is ~12 kJ · mol$^{—1}$, a quantity that rep- resents an energy barrier to free rotation about the C¬C single bond. Substituents other than hydrogen exhibit greater steric interference; that is, they increase the size of this energy barrier due to their greater bulk. Indeed, with large substituents, some conformations may be sterically forbidden.


O
1.24
120.5°
123.5°
C
Peptide bond
116°
N
111°
119.5°
118.5°
Cα
Cα
H
H
H
R
R
Amide
plane


p = 0
p
p
p
d
p
n = 5
n = 4
n = 3
n = 2
n = −3

Helical Structures

Helices are the most striking elements of protein 2° struc- ture. If a polypeptide chain is twisted by the same amount about each of its C_a atoms, it assumes a helical conforma- tion. As an alternative to specifying its $ and † angles, a helix may be characterized by the number, *n*, of peptide units per helical turn and by its **pitch,** *p*, the distance the helix rises along its axis per turn. Several examples of he- lices are diagrammed in the Fig.. Note that a helix has chi rality; that is, it may be either right handed or left handed (a right-handed helix turns in the direction that the fingers of a right hand curl when its thumb points along the helix axis in the direction that the helix rises). In proteins, more- over, *n* need not be an integer and, in fact, rarely is.

The Helix

Only one helical polypeptide conformation has simulta- neously allowed conformation angles and a favorable hydrogen bonding pattern: the a **helix**, a partic- ularly rigid arrangement of the polypeptide chain. Its dis- covery through model building, by Pauling in 1951, ranks as one of the landmarks of structural biochemistry. For a polypeptide made from $_L$-a-amino acid residues, the a helix is right handed with torsion angles $ = —57° and † = —47°, n = 3.6 residues per turn, and a pitch of 5.4 Å. (An a helix of $_D$-a-amino acid residues is the mir- ror image of that made from $_L$-amino acid residues: It is left handed with conformation angles $ = +57°, † = +47°, and n =—3.6 but with the same value of p.) Figure 8-11 indicates that the hydrogen bonds of an a helix are arranged such that the peptide N¬H bond of the nth residue points along the helix toward the peptide C˛˛O group of the (n — 4)th residue. This result in a strong hydrogen bond that has the nearly optimum N Ᵽ O distance of 2.8 Å. In addition, the core of the a helix is tightly packed; that is, its atoms are in van der Waals con- tact across the helix, thereby maximizing their association energies (Section 8-4A). The R groups, whose positions, as we saw, are not fully dealt with by the Ramachandran diagram, all project backward and outward from the helix so as to avoid steric interference with the polypeptide backbone and with each other. Such an arrangement can also be seen in the fig. Indeed, a ma- jor reason why the left-handed a helix has never been ob- served (its helical parameters are but mildly forbidden; is that its side chains contact its polypeptide backbone too closely. Note, however, that 1 to 2% of the individual non-Gly residues in proteins assume this con- formation. The a helix is a common secondary structural element of both fibrous and globular proteins. In globular proteins, a helices have an average span of ~12 residues, which cor- responds to over three helical turns and a length of 18 Å. However, a helices with as many as 53 residues have been found.

Beta Structures

In 1951, the year that they proposed the a helix, Pauling and Corey also postulated the existence of a different polypeptide sec- ondary structure, the þ **pleated sheet.** As with the a helix, the þ pleated sheet's conformation has repeating $ and † angles that fall in the allowed region of the Ramachandran diagram and utilizes the full hydrogen bonding capacity of the polypeptide backbone. *In þ pleated sheets, however, hydrogen bonding occurs between neighboring polypeptide chains* rather than within one as in *a* helices.þ Pleated sheets come in two varieties: The antiparallel þ pleated sheet, in which neighboring hydrogen bonded polypeptide chains run in opposite directions. The parallel þ pleated sheet, in which the hydrogen bonded chains extend in the same direction. The conformations in which these þ structures are opti- mally hydrogen bonded vary somewhat from that of a fully extended polypeptide ($ = † = ±180°), as indicated in the fig. They therefore have a rippled or pleated edge-on appearance, which accounts for the appellation "pleated sheet." In this conformation, successive side chains of a polypeptide chain extend to opposite sides of the pleated sheet with a two-residue repeat distance of 7.0 Å. þ Sheets are common structural motifs in proteins. In globular proteins, they consist of from 2 to as many as 15 polypeptide strands, the average being 6 strands, which have an aggregate width of ~25 Å. The polypeptide chains in a þ sheet are known to be up to 15 residues long, with the average being 6 residues that have a length of ~21 Å. A 6-stranded antiparallel þ sheet, for example, occurs in the jack bean protein concanavalin A. Parallel þ sheets of less than five strands are rare. This observation suggests that parallel þ sheets are less stable than antiparallel þ sheets, possibly because the hydrogen bonds of parallel sheets are distorted in comparison to those of the antiparallel sheets.

PROTEIN STABILITY

Incredible as it may seem, thermodynamic measurements indicate that *native proteins are only marginally stable en- tities under physiological conditions*. The free energy re- quired to denature them is ~0.4 kJ · mol$^{—1}$ of amino acid residues, so that 100-residue proteins are typically stable by only around 40 kJ · mol$^{—1}$. In contrast, the energy re- quired to break a typical hydrogen bond is ~20 kJ · mol$^{—1}$. The various noncovalent influences to which proteins are subject—electrostatic interactions (both attractive and repulsive), hydrogen bonding (both intramolecular and to water), and hydrophobic forces—each have energetic mag- nitudes that may total thousands of kilojoules per mole over an entire protein molecule. Consequently, *a protein structure arises from a delicate balance among powerful countervailing forces*. In this section we discuss the nature of these forces and end by considering protein denatura- tion, that is, how these forces can be disrupted.

Electrostatic Forces

Molecules are collections of electrically charged particles and hence, to a reasonable degree of approximation, their interactions are determined by the laws of classical elec- trostatics (more exact calculations require the application of quantum mechanics). The energy of association, *U*, of two electric charges, q_1 and q_2, that are separated by the distance *r* is found by integrating the expression for Coulomb's law, Eq. [2.1], to determine the work necessary to separate these charges by an infinite distance: Here $k = 9.0 \times 10^9$ J · m · C$^{—2}$ and *D* is the dielectric constant of the medium in which the charges are immersed (recall that *D* = 1 for a vacuum and, for the most part, in- creases with the polarity of the medium; Table 2-1). The dielectric constant of a molecule-sized region is difficult to estimate. For the interior of a protein, it is usually taken to be in the range 3 to 5 in analogy with the measured di- electric constants of substances that have similar polarities, such as benzene and diethyl ether.

Hydrogen Bonding Forces

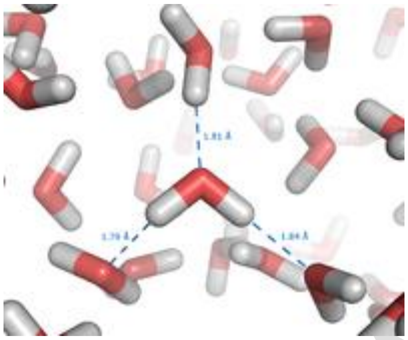
Hydrogen bonds (D¬H Þ A), as we discussed in Section 2-1A, are predominantly electrostatic interactions (but with ~10% covalent character) between a weakly acidic donor group (D¬H) and an acceptor (A) that bears a lone pair of electrons. In biological systems, D and A can both be the highly electronegative N and O atoms and occasionally S atoms. In addition, a relatively acidic C¬H group (e.g., a C$_a$¬H group) can act as a weak hydrogen bond donor, and the polarizable g electron system of an aromatic ring (e.g., that of Trp) can act as a weak acceptor. Hydrogen bonds have association energies that are nor- mally in the range —12 to —40 kJ · mol$^{—1}$ (but only around —8 to —16 kJ · mol$^{—1}$ for C¬H Þ A and D¬H Þ g hydrogen bonds and —2 to —4 kJ · mol$^{—1}$ for C¬H Þ g hydrogen bonds), values which are between those for covalent bonds and van der Waals forces. Hydrogen bonds **(H bonds)** are much more directional than are van der Waals forces but less so than are covalent bonds. The D Þ A distance is normally in the range 2.7 to 3.1 Å, al- though since H atoms are unseen in all but the very high- est resolution macromolecular X-ray structures, a possible D¬H Þ A interaction (where D and A are either N or O) is assumed to be a H bond if its D Þ A distance is sig- nificantly less than the 3.7 Å sum of a D¬H bond length (~1.0 Å) and the van der Waals contact distance between H and A (~2.7 Å). Keep in mind, however, that there is no rigid cutoff distance beyond which H bonds cease to exist because the energy of an H bond, which is mainly electrostatic in character, varies inversely with the distance between the negative and positive centers. H bonds tend to be linear, with the D¬H bond pointing along the acceptor's lone pair orbital hydrogen bonds, roughly perpendicular to the aromatic ring and pointing at its center with the distance from the D atom to the center of the aromatic ring normally in the range 3.2 – 3.8 Å).

Hydrophobic effect

The hydrophobic effect is the observed tendency of nonpolar substances to aggregate in aqueous solution and exclude water molecules. The word hydrophobic literally means "water-fearing", and it describes the segregation of water and nonpolar substances, which maximizes hydrogen bonding between molecules of water and minimizes the area of contact between water and nonpolar molecules. The hydrophobic effect is responsible for the separation of a mixture of oil and water into its two components. It is also responsible for effects related to biology, including: cell membranes and vesicles formation, protein folding, insertion of membrane proteins into the nonpolar lipid environment and protein-small molecule associations. Hence the hydrophobic effect is essential to life.[3][4][5][6] Substances for which this effect is observed are known as hydrophobes. The hydrophilic groups prevent phase separation of the molecules by maintaining the hydrophobic groups in water through formation of strong hydrogen bonds with water molecules. The driving force for this self-assembly is the hydrophobic effect. Amphiphiles are molecules that have both hydrophobic and hydrophilic domains. Detergents are composed of amphiphiles that allow hydrophobic molecules to be solubilized in water by forming micelles and bilayers (as in soap bubbles). They are also important to cell membranes composed of amphiphilic phospholipids that prevent the internal aqueous environment of a cell from mixing with external water.

Folding of macromolecules

In the case of protein folding, the hydrophobic effect is important to understanding the structure of proteins that have hydrophobic amino acids (such as alanine, valine, leucine, isoleucine, phenylalanine, tryptophan and methionine) clustered together within the protein. Structures of water-soluble proteins have a hydrophobic core in which side chains are buried from water, which stabilizes the folded state. Charged and polar side chains are situated on the solvent-exposed surface where they interact with surrounding water molecules. Minimizing the number of hydrophobic side chains exposed to water is the principal driving force behind the folding process,although formation of hydrogen bonds within the protein also stabilizes protein structure. The energetics of DNA tertiary structure assembly were determined to be driven by the hydrophobic effect, in addition to Watson-Crick base pairing, which is responsible for sequence selectivity, and stacking interactions between the aromatic bases. In biochemistry, the hydrophobic effect can be used to separate mixtures of proteins based on their hydrophobicity. Column chromatography with a hydrophobic stationary phase such as phenyl-sepharose will cause more hydrophobic proteins to travel more slowly, while less hydrophobic ones elute from the column sooner. To achieve better separation, a salt may be added (higher concentrations of salt increase the hydrophobic effect) and its concentration decreased as the separation progresses.



Dynamic hydrogen bonds between molecules of liquid water

The origin of the hydrophobic effect is not fully understood. Some argue that the hydrophobic interaction is mostly an entropic effect originating from the disruption of highly dynamic hydrogen bonds between molecules of liquid water by the nonpolar solute. A hydrocarbon chain or a similar nonpolar region of a large molecule is incapable of forming hydrogen bonds with water. Introduction of such a non-hydrogen bonding surface into water causes disruption of the hydrogen bonding network between water molecules. The hydrogen bonds are reoriented tangentially to such surface to minimize disruption of the hydrogen bonded 3D network of water molecules, and this leads to a structured water "cage" around the nonpolar surface. The water molecules that form the "cage" (or solvation shell) have restricted mobility. In the solvation shell of small nonpolar particles, the restriction amounts to some 10%. For example, in the case of dissolved xenon at room temperature a mobility restriction of 30% has been found. In the case of larger nonpolar molecules, the reorientational and translational motion of the water molecules in the solvation shell may be restricted by a factor of two to four; thus, at

25 °C the reorientational correlation time of water increases from 2 to 4-8 picoseconds. Generally, this leads to significant losses in translational and rotational entropy of water molecules and makes the process unfavorable in terms of the free energy in the system. By aggregating together, nonpolar molecules reduce the surface area exposed to water and minimize their disruptive effect.

The hydrophobic effect can be quantified by measuring the partition coefficients of non-polar molecules between water and non-polar solvents. The partition coefficients can be transformed to free energy of transfer which includes enthalpic and entropic components, $\Delta G = \Delta H - T\Delta S$. These components are experimentally determined by calorimetry. The hydrophobic effect was found to be entropy-driven at room temperature because of the reduced mobility of water molecules in the solvation shell of the non-polar solute; however, the enthalpic component of transfer energy was found to be favorable, meaning it strengthened water-water hydrogen bonds in the solvation shell due to the reduced mobility of water molecules. At the higher temperature, when water molecules become more mobile, this energy gain decreases along with the entropic component. The hydrophobic effect depends on the temperature, which leads to "cold denaturation" of proteins.

Van der Waals force

In physical chemistry, the van der Waals forces, named after Dutch scientist Johannes Diderik van der Waals, are distance-dependent interactions between atoms or molecules. Unlike ionic or covalent bonds, these attractions are not a result of any chemical electronic bond, and they are comparatively weak and more susceptible to being perturbed. Van der Waals forces quickly vanish at longer distances between interacting molecules. Van der Waals forces play a fundamental role in fields as diverse as supramolecular chemistry, structural biology, polymer science, nanotechnology, surface science, and condensed matter physics. van der Waals forces also define many properties of organic compounds and molecular solids, including their solubility in polar and non-polar media.

If no other forces are present, the point at which the force becomes repulsive rather than attractive as two atoms near one another is called the van der Waals contact distance. This results from the electron clouds of two atoms unfavorably coming into contact. It can be shown that van der Waals forces are of the same origin as the Casimir effect, arising from quantum interactions with the zero-point field. The resulting van der Waals forces can be attractive or repulsive. It is also sometimes used loosely as a synonym for the totality of intermolecular forces. The term includes the force between permanent dipoles (Keesom force), the force between a permanent dipole and a corresponding induced dipole (Debye force), and the force between instantaneously induced dipoles (London dispersion force).

Being the weakest of the weak chemical forces, with a strength between 0.4 and 4kJ/mol they may still support an integral structural load when multitudes of such interactions are present. Such a force results from a transient shift in electron density. Specifically, as the electrons are in orbit of the protons and neutrons within an atom the electron density may tend to shift more greatly on a side. Thus, this generates a transient charge to which a nearby atom can be either attracted or repelled. When the interatomic distance of two atoms is greater than 0.6 nm the force is not strong enough to be observed. In the same vein, when the interatomic distance is below 0.4 nm the force becomes repulsive.

Intermolecular forces have four major contributions:

1. A repulsive component resulting from the Pauli exclusion principle that prevents the collapse of molecules.
2. Attractive or repulsive electrostatic interactions between permanent charges (in the case of molecular ions), dipoles (in the case of molecules without inversion center), quadrupoles (all molecules with symmetry lower than cubic), and in general between permanent multipoles. The electrostatic interaction is sometimes called the Keesom interaction or Keesom force after Willem Hendrik Keesom.
3. Induction (also known as polarization), which is the attractive interaction between a permanent multipole on one molecule with an induced multipole on another. This interaction is sometimes called Debye force after Peter J.W. Debye.

4. Dispersion (usually named after Fritz London), which is the attractive interaction between any pair of molecules, including non-polar atoms, arising from the interactions of instantaneous multipoles.

Returning to nomenclature, different texts refer to different things using the term "van der Waals force". Some texts describe the van der Waals force as the totality of forces (including repulsion); others mean all the attractive forces (and then sometimes distinguish van der Waals-Keesom, van der Waals-Debye, and van der Waals-London). All intermolecular/van der Waals forces are anisotropic (except those between two noble gas atoms), which means that they depend on the relative orientation of the molecules. The induction and dispersion interactions are always attractive, irrespective of orientation, but the electrostatic interaction changes sign upon rotation of the molecules. That is, the electrostatic force can be attractive or repulsive, depending on the mutual orientation of the molecules. When molecules are in thermal motion, as they are in the gas and liquid phase, the electrostatic force is averaged out to a large extent, because the molecules thermally rotate and thus probe both repulsive and attractive parts of the electrostatic force. Sometimes this effect is expressed by the statement that "random thermal motion around room temperature can usually overcome or disrupt them" (which refers to the electrostatic component of the van der Waals force). Clearly, the thermal averaging effect is much less pronounced for the attractive induction and dispersion forces. The Lennard-Jones potential is often used as an approximate model for the isotropic part of a total (repulsion plus attraction) van der Waals force as a function of distance.

Van der Waals forces are responsible for certain cases of pressure broadening (van der Waals broadening) of spectral lines and the formation of van der Waals molecules. The London-van der Waals forces are related to the Casimir effect for dielectric media, the former being the microscopic description of the latter bulk property. The first detailed calculations of this were done in 1955 by E. M. Lifshitz. A more general theory of van der Waals forces has also been developed.

The main characteristics of van der Waals forces are:

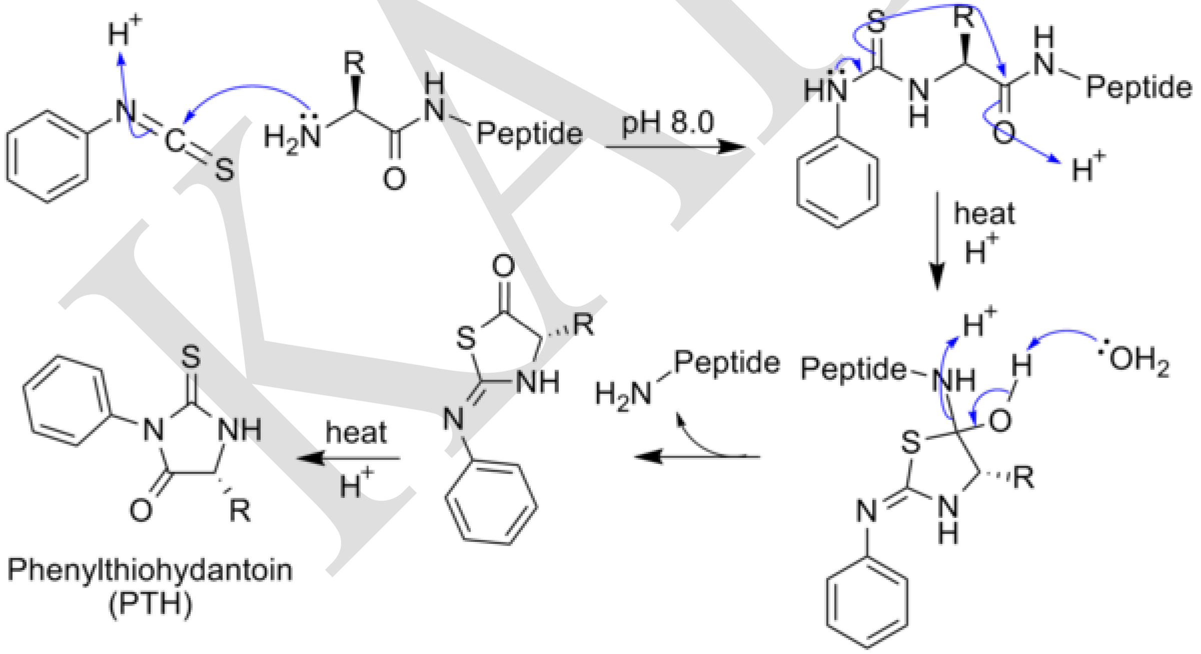
- They are weaker than normal covalent and ionic bonds.

- van der Waals forces are additive and cannot be saturated.
- They have no directional characteristic.
- They are all short-range forces and hence only interactions between the nearest particles need to be considered (instead of all the particles). Van der Waals attraction is greater if the molecules are closer.
- van der Waals forces are independent of temperature except dipole – dipole interactions.

In low molecular weight alcohols, the hydrogen-bonding properties of their polar hydroxyl group dominate other weaker van der Waals interactions. In higher molecular weight alcohols, the properties of the nonpolar hydrocarbon chain(s) dominate and define the solubility.

Protein Sequencing

Edman degradation, developed by Pehr Edman, is a method of sequencing amino acids in a peptide. In this method, the amino-terminal residue is labeled and cleaved from the peptide without disrupting the peptide bonds between other amino acid residues.



Phenyl isothiocyanate is reacted with an uncharged N-terminal amino group, under mildly alkaline conditions, to form a cyclical *phenylthiocarbamoyl* derivative. Then, under acidic conditions, this derivative of the terminal amino acid is cleaved as a thiazolinone derivative. The thiazolinone amino acid is then selectively extracted into an organic solvent and treated with acid to form the more stable phenylthiohydantoin (PTH)- amino acid derivative that can be identified by using chromatography or electrophoresis. This procedure can then be repeated again to identify the next amino acid. A major drawback to this technique is that the peptides being sequenced in this manner cannot have more than 50 to 60 residues (and in practice, under 30). The peptide length is limited due to the cyclical derivatization not always going to completion. The derivatization problem can be resolved by cleaving large peptides into smaller peptides before proceeding with the reaction. It is able to accurately sequence up to 30 amino acids with modern machines capable of over 99% efficiency per amino acid. An advantage of the Edman degradation is that it only uses 10 - 100 pico-moles of peptide for the sequencing process. The Edman degradation reaction was automated in 1967 by Edman and Beggs to speed up the process and 100 automated devices were in use worldwide by 1973.

Because the Edman degradation proceeds from the N-terminus of the protein, it will not work if the N-terminus has been chemically modified (e.g. by acetylation or formation of pyroglutamic acid). Sequencing will stop if a non-α-amino acid is encountered (e.g. isoaspartic acid), since the favored five-membered ring intermediate is unable to be formed. Edman degradation is generally not useful to determine the positions of disulfide bridges. It also requires peptide amounts of 1 picomole or above for discernible results.

Sedimentation

Sedimentation is the tendency for particles in suspension to settle out of the fluid in which they are entrained and come to rest against a barrier. This is due to their motion through the fluid in response to the forces acting on them: these forces can be due to gravity, centrifugal acceleration, or electromagnetism. In geology, sedimentation is often used as the opposite of erosion, i.e., the terminal end of sediment transport. In that sense, it includes the termination of transport by saltation or true bedload transport. Settling is the falling of suspended particles through the liquid, whereas sedimentation is the termination of the settling process. In estuarine environments, settling can be influenced by the presence or absence of vegetation. Trees such as mangroves are crucial to the attenuation of waves or currents, promoting the settlement of suspended particles.

Sedimentation may pertain to objects of various sizes, ranging from large rocks in flowing water to suspensions of dust and pollen particles to cellular suspensions to solutions of single molecules such as proteins and peptides. Even small molecules supply a sufficiently strong force to produce significant sedimentation. The term is typically used in geology to describe the deposition of sediment which results in the formation of sedimentary rock, but it is also used in various chemical and environmental fields to describe the motion of often-smaller particles and molecules. This process is also used in the biotech industry to separate cells from the culture media.

Size-exclusion chromatography

Size-exclusion chromatography (SEC), also known as molecular sieve chromatography, is a chromatographic method in which molecules in solution are separated by their size, and in some cases molecular weight. It is usually applied to large molecules or macromolecular complexes such as proteins and industrial polymers. Typically, when an aqueous solution is used to transport the sample through the column, the technique is known as gel-filtration chromatography, versus the name gel permeation chromatography, which is used when an organic solvent is used as a mobile phase. SEC is a widely used polymer characterization method because of its ability to provide good molar mass distribution (Mw) results for polymers.

The main application of gel-filtration chromatography is the fractionation of proteins and other water-soluble polymers, while gel permeation chromatography is used to analyze the molecular weight distribution of organic-soluble polymers. Either technique should not be confused with gel electrophoresis, where an electric field is used to "pull" or "push" molecules through the gel depending on their electrical charges.

The advantages of this method include good separation of large molecules from the small molecules with a minimal volume of eluate, and that various solutions can be applied without interfering with the filtration process, all while preserving the biological activity of the particles to separate. The technique is generally combined with others that further separate molecules by other characteristics, such as acidity, basicity, charge, and affinity for certain compounds. With size exclusion chromatography, there are short and well-defined separation times and narrow bands, which lead to good sensitivity. There is also no sample loss because solutes do not interact with the stationary phase. The other advantage to this experimental method is that in certain cases, it is feasible to determine the approximate molecular weight of a compound. The shape and size of the compound (eluent) determine how the compound interacts with the gel (stationary phase). To determine approximate molecular weight, the elution volumes of compounds with their corresponding molecular weights are obtained and then a plot of "K_{av}" vs "log(Mw)" is made, where $K_{av} = (V_e-V_o)/(V_t-V_o)$ and Mw is the molecular mass. This plot acts as a calibration curve, which is used to approximate the desired compound's molecular weight. The V_e component represents the volume at which the intermediate molecules elute such as molecules that have partial access to the beads of the column. In addition, V_t is the sum of the total volume between the beads and the volume within the beads. The V_o component represents the volume at which the larger molecules elute, which elute in the beginning. Disadvantages are, for example, that only a limited number of bands can be accommodated because the time scale of the chromatogram is short, and, in general, there must be a 10% difference in molecular mass to have a good resolution.

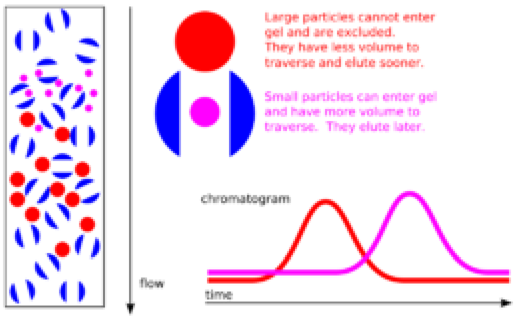
The technique was invented by Grant Henry Lathe and Colin R Ruthven, working at Queen Charlotte's Hospital, London. They later received the John Scott Award for this invention. While Lathe and Ruthven used starch gels as the matrix, Jerker Porath and Per Flodin later introduced dextran gels; other gels with size fractionation properties include agarose and polyacrylamide. A short review of these developments has appeared. There were also attempts to fractionate synthetic high polymers; however, it was not until 1964, when J. C. Moore of the Dow Chemical Company published his work on the preparation of gel permeation chromatography (GPC) columns based on cross-linked polystyrene with controlled pore size, that a rapid increase of research activity in this field began. It was recognized almost immediately that with proper calibration, GPC was capable to provide molar mass and molar mass distribution information for synthetic polymers. Because the latter information was difficult to obtain by other methods, GPC came rapidly into extensive use.

SEC is used primarily for the analysis of large molecules such as proteins or polymers. SEC works by trapping smaller molecules in the pores of the adsorbent materials adsorption ("stationary phases"). This process is usually performed with a column, which consists of a hollow tube tightly packed with extremely small porous polymer beads designed to have pores of different sizes. These pores may be depressions on the surface or channels through the bead. As the solution travels down the column some particles enter into the pores. Larger particles cannot enter into as many pores. The larger the particles, the faster the elution. The larger molecules simply pass by the pores because those molecules are too large to enter the pores. Larger molecules therefore flow through the column more quickly than smaller molecules, that is, the smaller the molecule, the longer the retention time.

One requirement for SEC is that the analyte does not interact with the surface of the stationary phases, with differences in elution time between analytes ideally being based solely on the solute volume the analytes can enter, rather than chemical or electrostatic interactions with the stationary phases. Thus, a small molecule that can penetrate every region of the stationary phase pore system can enter a total volume equal to the sum of the entire pore volume and the interparticle volume. This small molecule elutes late (after the molecule has penetrated all of the pore- and interparticle volume—approximately 80% of the column volume). At the other extreme, a very large molecule that cannot penetrate any the smaller pores can enter only the interparticle volume (~35% of the column volume) and elutes earlier when this volume of mobile phase has passed through the column. The underlying principle of SEC is that particles of different sizes elute (filter) through a stationary phase at different rates. This results in the separation of a solution of particles based on size. Provided that all the particles are loaded simultaneously or near-simultaneously, particles of the same size should elute together.

However, as there are various measures of the size of a macromolecule (for instance, the radius of gyration and the hydrodynamic radius), a fundamental problem in the theory of SEC has been the choice of a proper molecular size parameter by which molecules of different kinds are separated. Experimentally, Benoit and co-workers found an excellent correlation between elution volume and a dynamically based molecular size, the hydrodynamic volume, for several different chain architecture and chemical compositions. The observed correlation based on the hydrodynamic volume became accepted as the basis of universal SEC calibration. Still, the use of the hydrodynamic volume, a size based on dynamical properties, in the interpretation of SEC data is not fully understood. This is because SEC is typically run under low flow rate conditions where hydrodynamic factor should have little effect on the separation. In fact, both theory and computer simulations assume a thermodynamic separation principle: the separation process is determined by the equilibrium distribution (partitioning) of solute macromolecules between two phases --- a dilute bulk solution phase located at the interstitial space and confined solution phases within the pores of column packing material. Based on this theory, it has been shown that the relevant size parameter to the partitioning of polymers in pores is the mean span dimension (mean maximal projection onto a line). Although this issue has not been fully resolved, it is likely that the mean span dimension and the hydrodynamic volume are strongly correlated.

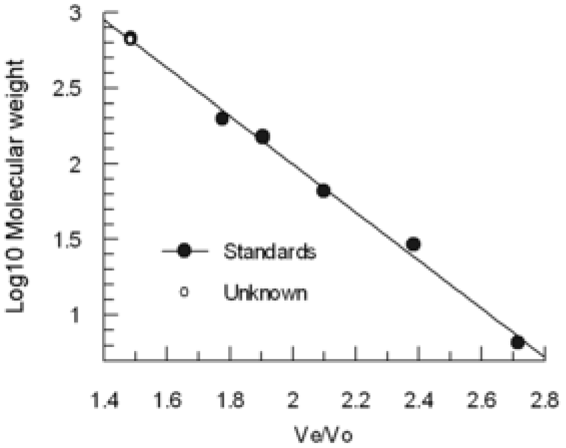
Each size exclusion column has a range of molecular weights that can be separated. The exclusion limit defines the molecular weight at the upper end of the column 'working' range and is where molecules are too large to get trapped in the stationary phase. The lower end of the range is defined by the permeation limit, which defines the molecular weight of a molecule that is small enough to penetrate all pores of the stationary phase. All molecules below this molecular mass are so small that they elute as a single band. The filtered solution that is collected at the end is known as the eluate. The void volume includes any particles too large to enter the medium, and the solvent volume is known as the column volume.


Large particles cannot enter gel and are excluded. They have less volume to traverse and elute sooner.
Small particles can enter gel and have more volume to traverse. They elute later.
chromatogram
flow
time

A cartoon illustrating the theory behind size exclusion chromatography

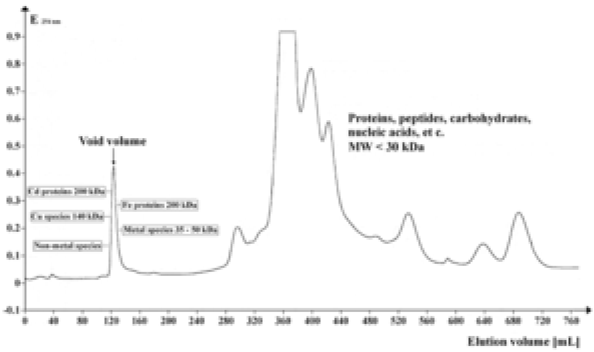
In real-life situations, particles in solution do not have a fixed size, resulting in the probability that a particle that would otherwise be hampered by a pore passing right by it. Also, the stationary-phase particles are not ideally defined; both particles and pores may vary in size. Elution curves, therefore, resemble Gaussian distributions. The stationary phase may also interact in undesirable ways with a particle and influence retention times, though great care is taken by column manufacturers to use stationary phases that are inert and minimize this issue.

Like other forms of chromatography, increasing the column length enhances resolution, and increasing the column diameter increases column capacity. Proper column packing is important for maximum resolution: An over-packed column can collapse the pores in the beads, resulting in a loss of resolution. An under-packed column can reduce the relative surface area of the stationary phase accessible to smaller species, resulting in those species spending less time trapped in pores. Unlike affinity chromatography techniques, a solvent head at the top of the column can drastically diminish resolution as the sample diffuses prior to loading, broadening the downstream elution. In simple manual columns, the eluent is collected in constant volumes, known as fractions. The more similar the particles are in size the more likely they are in the same fraction and not detected separately. More advanced columns overcome this problem by constantly monitoring the eluent.



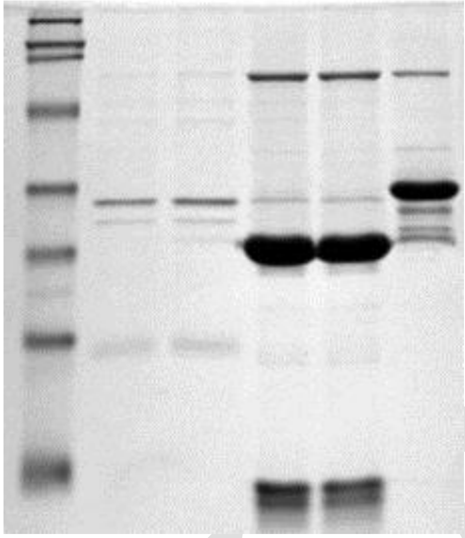
Standardization of a size exclusion column.

The collected fractions are often examined by spectroscopic techniques to determine the concentration of the particles eluted. Common spectroscopy detection techniques are refractive index (RI) and ultraviolet (UV). When eluting spectroscopically similar species (such as during biological purification), other techniques may be necessary to identify the contents of each fraction. It is also possible to analyse the eluent flow continuously with RI, LALLS, Multi-Angle Laser Light Scattering MALS, UV, and/or viscosity measurements.



The elution volume (Ve) decreases roughly linear with the logarithm of the molecular hydrodynamic volume. Columns are often calibrated using 4-5 standard samples (e.g., folded proteins of known molecular weight), and a sample containing a very large molecule such as thyroglobulin to determine the void volume. (Blue dextran is not recommended for Vo determination because it is heterogeneous and may give variable results) The elution volumes of the standards are divided by the elution volume of the thyroglobulin (Ve/Vo) and plotted against the log of the standards.

Polyacrylamide gel electrophoresis



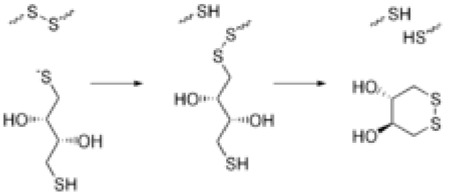
Picture of an SDS-PAGE. The molecular markers (ladder) are in the left lane

Polyacrylamide gel electrophoresis (PAGE), describes a technique widely used in biochemistry, forensics, genetics, molecular biology and biotechnology to separate biological macromolecules, usually proteins or nucleic acids, according to their electrophoretic mobility. Mobility is a function of the length, conformation and charge of the molecule. As with all forms of gel electrophoresis, molecules may be run in their native state, preserving the molecules' higher-order structure. This method is called native-PAGE. Alternatively, a chemical denaturant may be added to remove this structure and turn the molecule into an unstructured molecule whose mobility depends only on its length and mass-to-charge ratio. This procedure is called SDS-PAGE. Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) is a method of separating molecules based on the difference of their molecular weight. At the pH at which gel electrophoresis is carried out the SDS molecules are negatively charged and bind to proteins in a set ratio, approximately one molecule of SDS for every 2 amino acids. In this way, the detergent provides all proteins with a uniform charge-to-mass ratio, independently of their original charge. By binding to the proteins the detergent destroys their secondary, tertiary and/or quaternary structure denaturing them and turning them into negatively charged linear poly peptide chains. When subjected to an electric field in PAGE, the negatively charged poly peptide chains travel toward the anode with different mobility. Their mobility, or the distance traveled by molecules, is inversely proportional to the logarithm of their molecular weight. By comparing the relative ratio of the distance traveled by each protein to the length of the gel (Rf) one can make conclusions about the relative molecular weight of the proteins, where the length of the gel is determined by the distance traveled by a small molecule like a tracking dye.

For nucleic acids, urea is the most commonly used denaturant. For proteins, sodium dodecyl sulfate (SDS) is an anionic detergent applied to protein samples to coat proteins in order to impart two negative charges (from every SDS molecule) to every two amino acids of the denatured protein. 2-Mercaptoethanol may also be used to disrupt the disulfide bonds found between the protein complexes, which helps further denature the protein. In most proteins, the binding of SDS to the polypeptide chain imparts an even distribution of charge per unit mass, thereby resulting in a fractionation by approximate size during electrophoresis. Proteins that have a greater hydrophobic content – for instance, many membrane proteins, and those that interact with surfactants in their native environment – are intrinsically harder to treat accurately using this method, due to the greater variability in the ratio of bound SDS. Procedurally, using both Native and SDS-PAGE together can be used to purify and to separate the various subunits of the protein. Native-PAGE keeps the oligomeric form intact and will show a band on the gel that is representative of the level of activity. SDS-PAGE will denature and separate the oligomeric form into its monomers, showing bands that are representative of their molecular weights. These bands can be used to assess the purity of and identify the protein.

Sample preparation

Samples may be any material containing proteins or nucleic acids. These may be biologically derived, for example from prokaryotic or eukaryotic cells, tissues, viruses, environmental samples, or purified proteins. In the case of solid tissues or cells, these are often first broken down mechanically using a blender (for larger sample volumes), using a homogenizer (smaller volumes), by sonicator or by using cycling of high pressure, and a combination of biochemical and mechanical techniques – including various types of filtration and centrifugation – may be used to separate different cell compartments and organelles prior to electrophoresis. Synthetic biomolecules such as oligonucleotides may also be used as analytes.



Reduction of a typical disulfide bond by DTT via two sequential thiol-disulfide exchange reactions. The sample to analyze is optionally mixed with a chemical denaturant if so desired, usually SDS for proteins or urea for nucleic acids. SDS is an anionic detergent that denatures secondary and non–disulfide–linked tertiary structures, and additionally applies a negative charge to each protein in proportion to its mass. Urea breaks the hydrogen bonds between the base pairs of the nucleic acid, causing the constituent strands to anneal. Heating the samples to at least 60 °C further promotes denaturation.

In addition to SDS, proteins may optionally be briefly heated to near boiling in the presence of a reducing agent, such as dithiothreitol (DTT) or 2-mercaptoethanol (beta-mercaptoethanol/BME), which further denatures the proteins by reducing disulfide linkages, thus overcoming some forms of tertiary protein folding, and breaking up quaternary protein structure (oligomeric subunits). This is known as reducing SDS-PAGE.

A tracking dye may be added to the solution. This typically has a higher electrophoretic mobility than the analytes to allow the experimenter to track the progress of the solution through the gel during the electrophoretic run.

Preparing acrylamide gels

The gels typically consist of acrylamide, bisacrylamide, the optional denaturant (SDS or urea), and a buffer with an adjusted pH. The solution may be degassed under a vacuum to prevent the formation of air bubbles during polymerization. Alternatively, butanol may be added to the resolving gel (for proteins) after it is poured, as butanol removes bubbles and makes the surface smooth. A source of free radicals and a stabilizer, such as ammonium persulfate and TEMED are added to initiate polymerization. The polymerization reaction creates a gel because of the added bisacrylamide, which can form cross-links between two acrylamide molecules. The ratio of bisacrylamide to acrylamide can be varied for special purposes, but is generally about 1 part in 35. The acrylamide concentration of the gel can also be varied, generally in the range from 5% to 25%. Lower percentage gels are better for resolving very high molecular weight molecules, while much higher percentages of acrylamide are needed to resolve smaller proteins. Gels are usually polymerized between two glass plates in a gel caster, with a comb inserted at the top to create the sample wells. After the gel is polymerized the comb can be removed and the gel is ready for electrophoresis.

Various buffer systems are used in PAGE depending on the nature of the sample and the experimental objective. The buffers used at the anode and cathode may be the same or different.
An electric field is applied across the gel, causing the negatively charged proteins or nucleic acids to migrate across the gel away from the negative electrode (which is the cathode being that this is an electrolytic rather than galvanic cell) and towards the positive electrode (the anode). Depending on their size, each biomolecule moves differently through the gel matrix: small molecules more easily fit through the pores in the gel, while larger ones have more difficulty. The gel is run usually for a few hours, though this depends on the voltage applied across the gel; migration occurs more quickly at higher voltages, but these results are typically less accurate than at those at lower voltages. After the set amount of time, the biomolecules have migrated different distances based on their size. Smaller biomolecules travel farther down the gel, while larger ones remain closer to the point of origin. Biomolecules may therefore be separated roughly according to size, which depends mainly on molecular weight under denaturing

conditions, but also depends on higher-order conformation under native conditions. The gel mobility is defined as the rate of migration traveled with a voltage gradient of 1V/cm and has units of cm^2/sec/V. For analytical purposes, the relative mobility of biomolecules, R_f, the ratio of the distance the molecule traveled on the gel to the total travel distance of a tracking dye is plotted versus the molecular weight of the molecule (or sometimes the log of MW, or rather the M_r, molecular radius). Such typically linear plots represent the standard markers or calibration curves that are widely used for the quantitative estimation of a variety of biomolecular sizes.

Certain glycoproteins, however, behave anomalously on SDS gels. Additionally, the analysis of larger proteins ranging from 250,000 to 600,000 Da is also reported to be problematic due to the fact that such polypeptides move improperly in the normally used gel systems.

Polyacrylamide gel (PAG) had been known as a potential embedding medium for sectioning tissues as early as 1964, and two independent groups employed PAG in electrophoresis in 1959. It possesses several electrophoretically desirable features that make it a versatile medium. It is a synthetic, thermo-stable, transparent, strong, chemically relatively inert gel, and can be prepared with a wide range of average pore sizes. The pore size of a gel and the reproducibility in gel pore size are determined by three factors, the total amount of acrylamide present (%T) (T = Total concentration of acrylamide and bisacrylamide monomer), the amount of cross-linker (%C) (C = bisacrylamide concentration), and the time of polymerization of acrylamide (cf. QPNC-PAGE). Pore size decreases with increasing %T; with cross-linking, 5%C gives the smallest pore size. Any increase or decrease in %C from 5% increases the pore size, as pore size with respect to %C is a parabolic function with vertex as 5%C. This appears to be because of non-homogeneous bundling of polymer strands within the gel. This gel material can also withstand high voltage gradients, is amenable to various staining and destaining procedures, and can be digested to extract separated fractions or dried for autoradiography and permanent recording.

Components

Polyacrylamide gels are composed of a stacking gel and separating gel. Stacking gels have a higher porosity relative to the separating gel, and allow for proteins to migrate in a concentrated area. Additionally, stacking gels usually have a pH of 6.8, since the neutral glycine molecules allow for faster protein mobility. Separating gels have a pH of 8.8, where the anionic glycine slows down the mobility of proteins. Separating gels allow for the separation of proteins and have a relatively lower porosity. Here, the proteins are separated based on size (in SDS-PAGE) and size/ charge (Native PAGE).

Chemical buffer stabilizes the pH value to the desired value within the gel itself and in the electrophoresis buffer. The choice of buffer also affects the electrophoretic mobility of the buffer counterions and thereby the resolution of the gel. The buffer should also be unreactive and not modify or react with most proteins. Different buffers may be used as cathode and anode buffers, respectively, depending on the application. Multiple pH values may be used within a single gel, for example in DISC electrophoresis. Common buffers in PAGE include Tris, Bis-Tris, or imidazole.

Counterion balance the intrinsic charge of the buffer ion and also affect the electric field strength during electrophoresis. Highly charged and mobile ions are often avoided in SDS-PAGE cathode buffers, but may be included in the gel itself, where it migrates ahead of the protein. In applications such as DISC SDS-PAGE the pH values within the gel may vary to change the average charge of the counterions during the run to improve resolution. Popular counterions are glycine and tricine. Glycine has been used as the source of trailing ion or slow ion because its pKa is 9.69 and mobility of glycinate are such that the effective mobility can be set at a value below that of the slowest known proteins of net negative charge in the pH range. The minimum pH of this range is approximately 8.0.

Acrylamide (C_3H_5NO; mW: 71.08) when dissolved in water, slow, spontaneous autopolymerization of acrylamide takes place, joining molecules together by head on tail fashion to form long single-chain polymers. The presence of a free radical-generating system greatly accelerates polymerization. This kind of reaction is known as vinyl addition polymerisation. A solution of these polymer chains becomes viscous but does not form a gel, because the chains simply slide over one another. Gel formation requires linking various chains together.

Acrylamide is carcinogenic, a neurotoxin, and a reproductive toxin. It is also essential to store acrylamide in a cool dark and dry place to reduce autopolymerisation and hydrolysis.

Bisacrylamide (*N*,*N*′-Methylenebisacrylamide) ($C_7H_{10}N_2O_2$; mW: 154.17) is the most frequently used cross linking agent for polyacrylamide gels. Chemically it can be thought of as two acrylamide molecules coupled head to head at their non-reactive ends. Bisacrylamide can crosslink two polyacrylamide chains to one another, thereby resulting in a gel.

Sodium dodecyl sulfate (SDS) ($C_{12}H_{25}NaO_4S$; mW: 288.38) (only used in denaturing protein gels) is a strong detergent agent used to denature native proteins to individual polypeptides. This denaturation, which is referred to as reconstructive denaturation, is not accomplished by the total linearization of the protein, but instead, through a conformational change to a combination of random coil and α helix secondary structures. When a protein mixture is heated to 100 °C in presence of SDS, the detergent wraps around the polypeptide backbone. It binds to polypeptides in a constant weight ratio of 1.4 g SDS/g of polypeptide. In this process, the intrinsic charges of polypeptides become negligible when compared to the negative charges contributed by SDS. Thus polypeptides after treatment become rod-like structures possessing a uniform charge density, that is same net negative charge per unit weight. The electrophoretic mobilities of these proteins is a linear function of the logarithms of their molecular weights. Without SDS, different proteins with similar molecular weights would migrate differently due to differences in mass-charge ratio, as each protein has an isoelectric point and molecular weight particular to its primary structure. This is known as native PAGE. Adding SDS solves this problem, as it binds to and unfolds the protein, giving a near uniform negative charge along the length of the polypeptide.

Urea ($CO(NH_2)_2$; mW: 60.06) is a chaotropic agent that increases the entropy of the system by interfering with intramolecular interactions mediated by non-covalent forces such as hydrogen bonds and van der Waals forces. Macromolecular structure is dependent on the net effect of these forces, therefore it follows that an increase in chaotropic solutes denatures macromolecules, Ammonium persulfate (APS) ($N_2H_8S_2O_8$; mW: 228.2) is a source of free radicals and is often used as an initiator for gel formation. An alternative source of free radicals is riboflavin, which generated free radicals in a photochemical reaction.

TEMED (*N*, *N*, *N′*, *N′*-tetramethylethylenediamine) ($C_6H_{16}N_2$; mW: 116.21) stabilizes free radicals and improves polymerization. The rate of polymerisation and the properties of the resulting gel depend on the concentrations of free radicals. Increasing the amount of free radicals results in a decrease in the average polymer chain length, an increase in gel turbidity and a decrease in gel elasticity. Decreasing the amount shows the reverse effect. The lowest catalytic concentrations that allow polymerisation in a reasonable period of time should be used. APS and TEMED are typically used at approximately equimolar concentrations in the range of 1 to 10 mM.

The following chemicals and procedures are used for processing of the gel and the protein samples visualized in it. Tracking dye; as proteins and nucleic acids are mostly colorless, their progress through the gel during electrophoresis cannot be easily followed. Anionic dyes of a known electrophoretic mobility are therefore usually included in the PAGE sample buffer. A very common tracking dye is Bromophenol blue (BPB, 3',3",5',5" tetrabromophenolsulfonphthalein). This dye is coloured at alkali and neutral pH and is a small negatively charged molecule that moves towards the anode. Being a highly mobile molecule it moves ahead of most proteins. As it reaches the anodic end of the electrophoresis medium electrophoresis is stopped. It can weakly bind to some proteins and impart a blue colour. Other common tracking dyes are xylene cyanol, which has lower mobility, and Orange G, which has a higher mobility.

Loading aids; most PAGE systems are loaded from the top into wells within the gel. To ensure that the sample sinks to the bottom of the gel, sample buffer is supplemented with additives that increase the density of the sample. These additives should be non-ionic and non-reactive towards proteins to avoid interfering with electrophoresis. Common additives are glycerol and sucrose.

Coomassie Brilliant Blue R-250 (CBB)($C_{45}H_{44}N_3NaO_7S_2$; mW: 825.97) is the most popular protein stain. It is an anionic dye, which non-specifically binds to proteins. The structure of CBB is predominantly non-polar, and it is usually used in methanolic solution acidified with acetic acid. Proteins in the gel are fixed by acetic acid and simultaneously stained. The excess dye incorporated into the gel can be removed by destaining with the same solution without the dye. The proteins are detected as blue bands on a clear background. As SDS is also anionic, it may interfere with staining process. Therefore, large volume of staining solution is recommended, at least ten times the volume of the gel.

Ethidium bromide (EtBr) is the traditionally most popular nucleic acid stain. Silver staining is used when more sensitive method for detection is needed, as classical Coomassie Brilliant Blue staining can usually detect a 50 ng protein band, Silver staining increases the sensitivity typically 10-100 fold more. This is based on the chemistry of photographic development. The proteins are fixed to the gel with a dilute methanol solution, then incubated with an acidic silver nitrate solution. Silver ions are reduced to their metallic form by formaldehyde at alkaline pH. An acidic solution, such as acetic acid stops development.[21] Silver staining was introduced by Kerenyi and Gallyas as a sensitive procedure to detect trace amounts of proteins in gels.[22] The technique has been extended to the study of other biological macromolecules that have been separated in a variety of supports.[23] Many variables can influence the colour intensity and every protein has its own staining characteristics; clean glassware, pure reagents and water of highest purity are the key points to successful staining.[24] Silver staining was developed in the 14th century for colouring the surface of glass. It has been used extensively for this purpose since the 16th century. The colour produced by the early silver stains ranged between light yellow and an orange-red. Camillo Golgi perfected the silver staining for the study of the nervous system. Golgi's method stains a limited number of cells at random in their entirety.

Autoradiography, also used for protein band detection post gel electrophoresis, uses radioactive isotopes to label proteins, which are then detected by using X-ray film. Western blotting is a process by which proteins separated in the acrylamide gel are electrophoretically transferred to a stable, manipulable membrane such as a nitrocellulose, nylon, or PVDF membrane. It is then possible to apply immunochemical techniques to visualise the transferred proteins, as well as accurately identify relative increases or decreases of the protein of interest.

Possible questions

1 Mark questions

1. Which of the following amino acids is an alpha helix terminator?
a) cysteine b) alanine c) proline d) glycine

2. The secondary structure is primarily maintained by
a) Vander waals force b) Hydrogen bond c) Ionic bond d) covalent bond

3. Which of the following is the most common and stable conformation for a poly peptide chain?
a) Alpha helix b) Beta pleated sheets c) Beta bends d) loops

4. The most common amino acid in beta bend is
a) cystein b) glycine c) serine d) Aspartic acid

5. Alpha-helix structure resembles
a) sheet b) coiled spring c) linear chain d) random coil

6. The alpha helix rises per turn a distance of
a) 0.54nm b) 1.5nm c) 3nm d) 6nm

7. Pulses contain incomplete proteins since they lack
a) Lysine b) Tryptophan c) Phenyl alanine d) Methinine

8. The fact that the core of most globular proteins is tightly packed is due to
a) the hydrophobic effect b) hydrogen bonding. c) electrostatic effects d)van der Waals forces

9. Which of the following compound is not involved in Edman degradation?
a) Phenylisothiocyanate b) CF3 COOH c) FDNB d) Phenylthiocarbonyl

10. Which of the following is the correct order of sequencing?
a) Cleaving, sequencing and ordering b) Sequencing, ordering and cleaving
c) Ordering, cleaving and sequencing d) Ordering, sequencing and cleaving

11. Which of the following is Edman reagent?
a) Phenylisothiocyanate b) CF3 COOH c) FDNB d) Phenylthiocarbonyl

2 Marks questions

1. Draw the structures of Cystine and Cysteine.
2. Draw the structures of L-α-Alanine and D-α-Alanine.
3. How will you differentiate peptides from proteins?
4. What is homomultimeric protein? Give an example.
5. What is heteromultimeric protein? Give an example.
6. Define the quaternary structure of proteins. Give an example.
7. What is Protomer? Give an example.
8. Define the quaternary structure of proteins. Give an example.
9. What is Protomer? Give an example.
10. Why is electrophoresis done in solution having low salt concentration?
11. Write a short note on ‘Isotachophoresis’.
12. Briefly write about the functions of SDS in SDS-PAGE.

6/8 Marks questions

1. How are proteins classified based on their compositions, shapes, functions, nutritional factors, secondary structures and folds?

2. Discuss in detail the sequencing of amino-acids by Edman method and its merits over Sanger method for protein sequencing.

3. Outline various strategies for determining primary structure and disulfide pattern(s) of a polypeptide chain.

4. Discuss on the following topics with suitable examples.
a) Structural architectures of proteins.
b) Protein classifications.

5. Write an essay on various types of regular and irregular secondary structures of proteins with suitable examples.

6. What are essential amino acids? Draw the structures of any two essential amino acids?

7. How are amino acids classified based on their metabolic roles?

8. How will you determine the number of polypeptide chains present in a protein?

9. Discuss in detail the principles, experimental conditions and applications of agarose gel electrophoresis/SDS-PAGE/IEF/PAGE.

10. A protein analyzed by SDS-PAGE and Gel filtration techniques shows molecular weight of 100 kD and 200 kD, respectively. How will you interpret the data?

UNIT - V

SYLLABUS

Proteomics: Introduction to Proteomics, Analysis of proteomes. 2D-PAGE. Sample preparation, solubilization, reduction, resolution. Reproducibility of 2D-PAGE. Mass spectrometry based methods for protein identification. *De novo* sequencing using mass spectrometric data.

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.

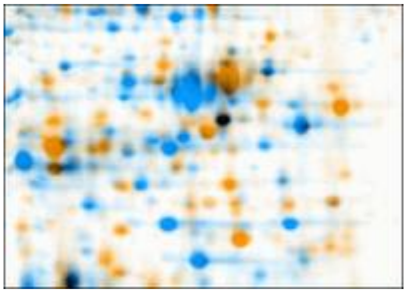
Common techniques

IPG-DALT

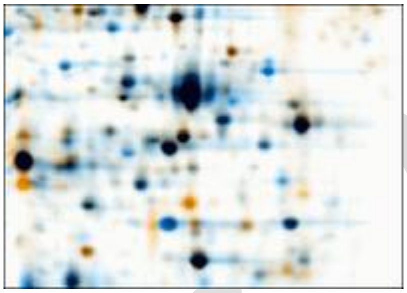
A common technique is to use an Immobilized pH gradient (IPG) in the first dimension. This technique is referred to as **IPG-DALT**. The sample is first separated onto IPG gel (which is commercially available) then the gel is cut into slices for each sample which is then equilibrated in SDS-mercaptoethanol 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.

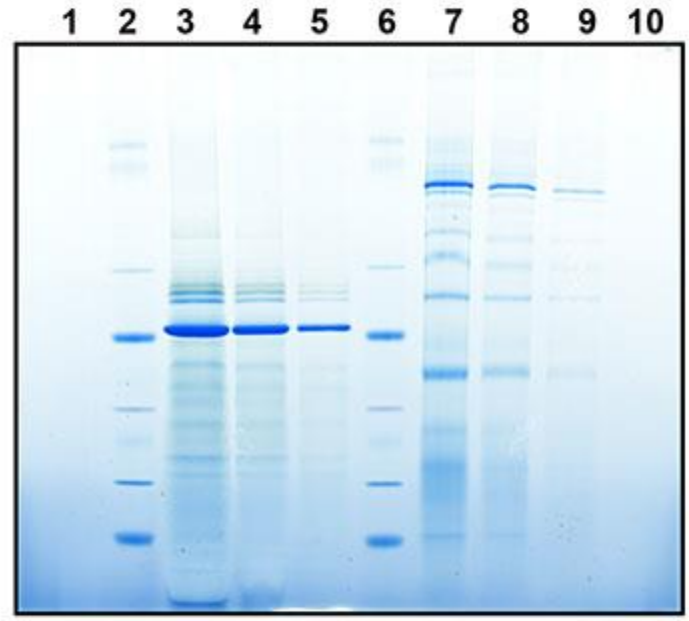
Given the common constraints of this format, most staining methods involve some version of the same general incubation steps:

- 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.

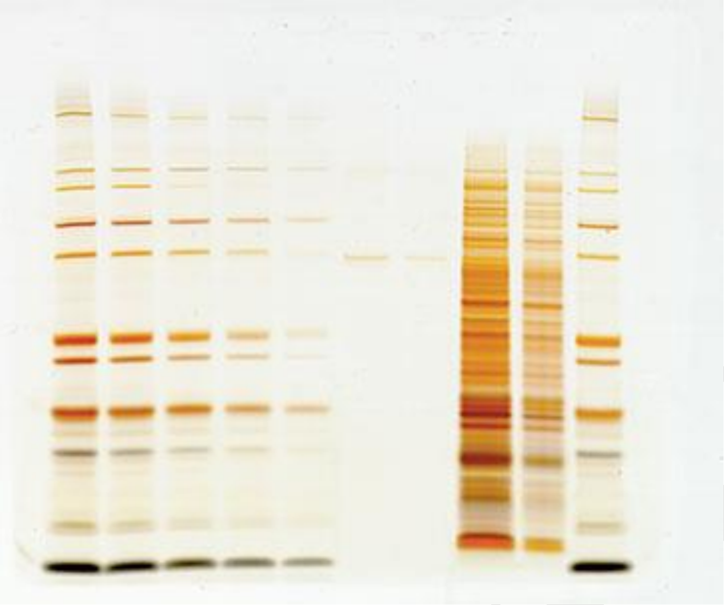
Coomassie dye staining is especially convenient because it involves a single ready-to-use reagent and does not permanently chemically modify the target proteins. An initial water wash step is necessary to remove residual SDS, which interferes with dye binding. Then the staining reagent is added, usually 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.

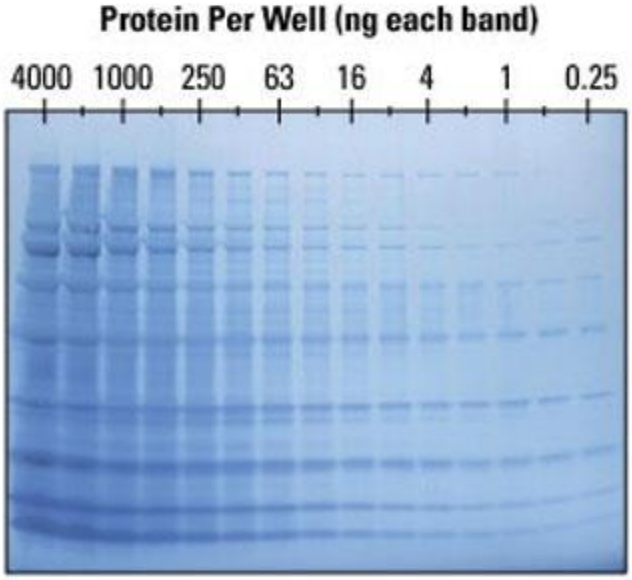


Gel staining with silver stain

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

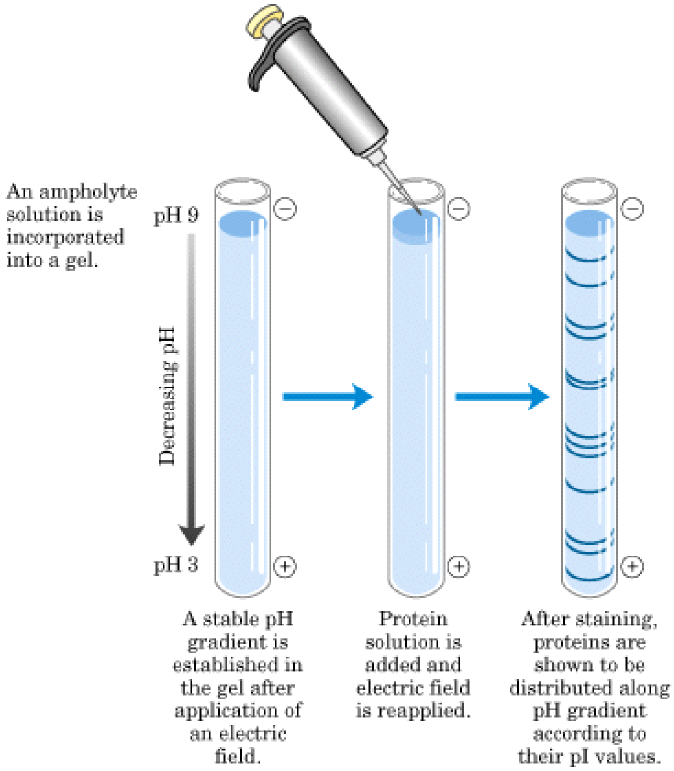
According to some opinions, living eukaryotic cells perform isoelectric focusing of proteins in their interior to overcome a limitation of the rate of metabolic reaction by diffusion of enzymes and their reactants, and to regulate the rate of particular biochemical processes. By concentrating the enzymes 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.

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.

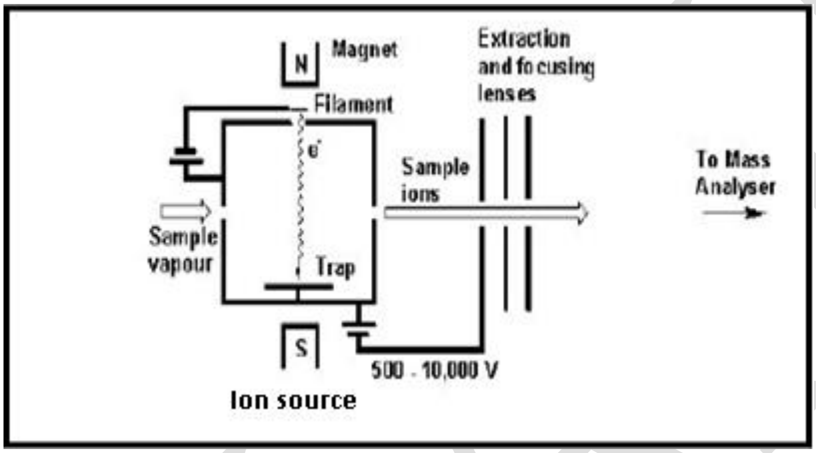


Figure 1. Schematic diagram of EIMS

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).

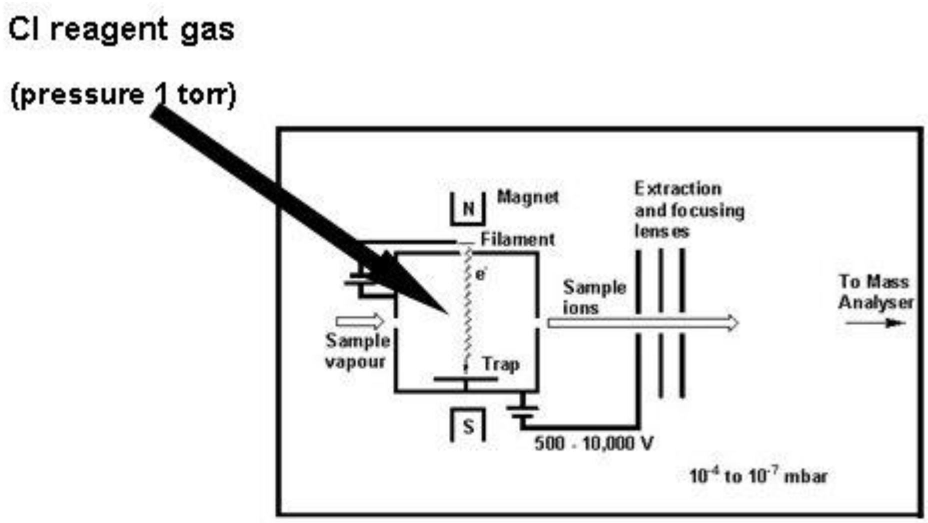


Figure 2. Schematic diagram of CI interface

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

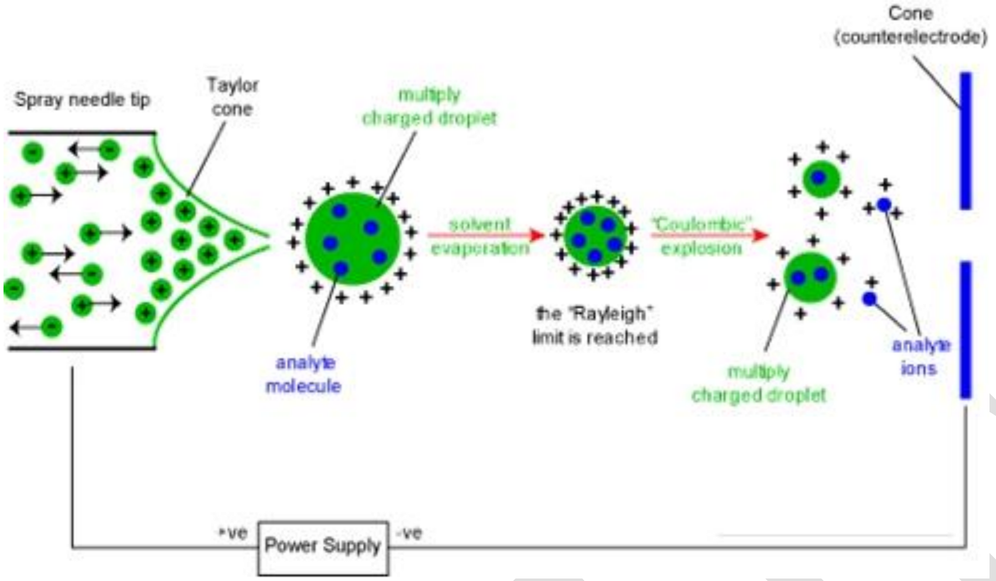
Negative ion mode: [G-H]- + M ------> [M-H]- + G

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.

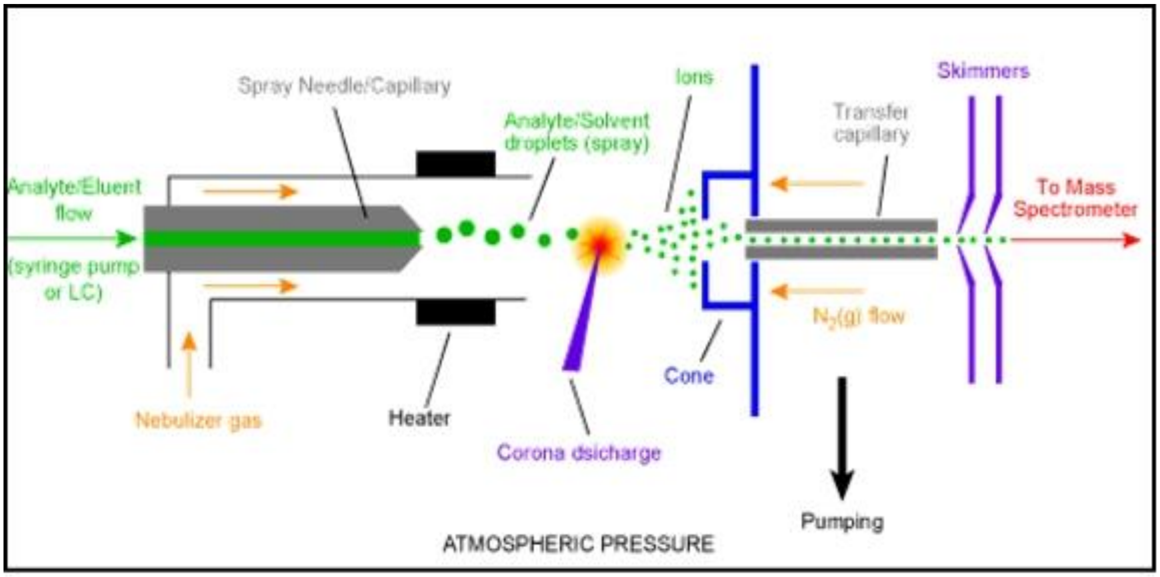
ESI can produce singly or multiply charged ions. The number of charges retained by a particular analyte depends on several factors such as the size, chemical composition, and higher order structure of the analyte molecule, the solvent composition, the presence of co-solutes, and the instrument parameters. 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

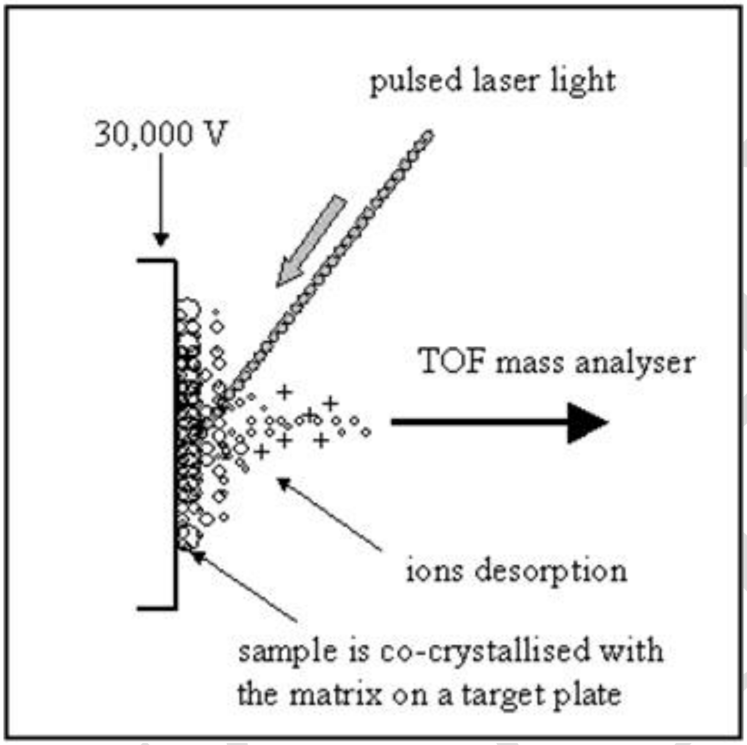
APCI has also become an important ionization source because it generates ions directly from solution and it is capable of analyzing relatively non-polar compounds. Similar to electrospray, the liquid effluent of APCI (Figure 4) is introduced directly into the ionization source. The droplets are not charged 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

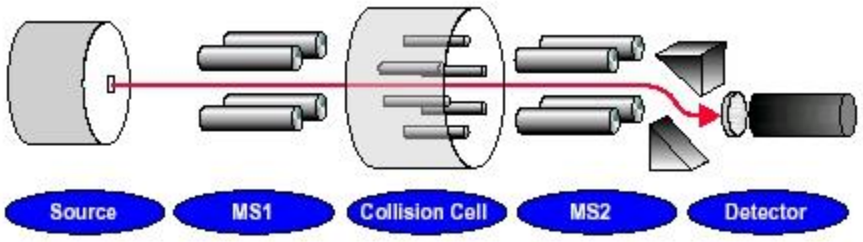
Matrix-assisted laser desorption/ionization (MALDI) is a technique to allows the high molecular weight compounds such as organic macro molecules and labile bimolecular into the gas phase as intact ions. MALDI is one of the recent developments of soft ionization techniques in the field of mass spectrometry. 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.



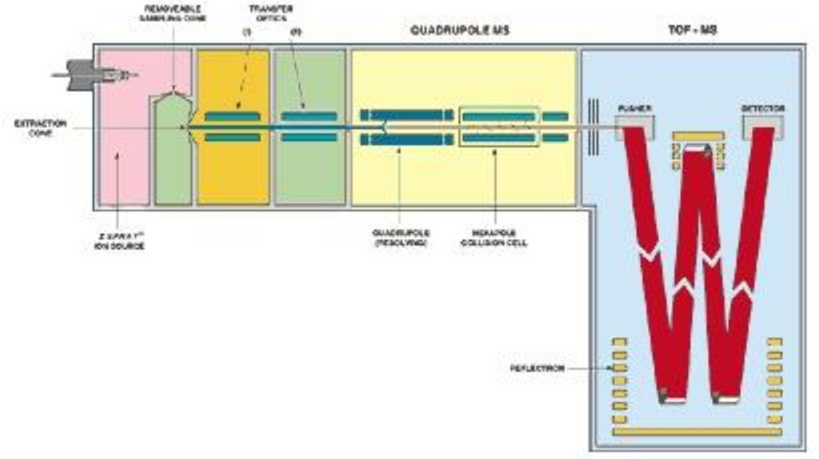
Schematic diagram of MALDI source

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

Possible questions

1 Mark 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.

9. 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

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

11. 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

2/6/8/10 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. Write the unique feature of MALDI-TOF.

6/8 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 an essay on principles, experimental set-up and applications of Isoelectric focussing electrophoresis technique.

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

4. Describe the principles of MALDI-MS/MALDI-MS-TOF and its applications on analyzing protein-protein interactions in a detailed manner.

LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600 034

B.Sc. DEGREE EXAMINATION – COMP.SCIENCE.,COMP.APP.&PYSICS

FOURTH SEMESTER – APRIL 2010

PB 4208 - BIOINFORMATICS - II (GENOMICS & PROTEOMICS)

Date & Time: 19/04/2010 / 9:00 - 12:00 Dept. No. □ Max. : 100 Marks

PART A **(20 marks)**

I Choose the best answer **(5 x 1 = 5)**

1. Current estimate of the mouse transcriptome is _____________ bases.
a) 39,000 b) 47,000 c) 30,000 d) 20,366

2. A prediction program, which offer models for introns and exons is_____________.
a) FGENESH b) GENEMARK c) GENOMESCAN d) TWINSCAN

3. The total number of protein families identified is ____________.
a) 58,000 b) 59,000 c) 60,000 d) 60,050

4. ____________ pathway turns on or off the genes.
a) Metabolic b) Signaling c) Biological d) Gene regulation

5. Which of the following is the third step in phylogenetic analysis?
a) tree building b) alignment c) substitution d) evaluation

II State whether the following statements are True or False **(5 x 1 = 5)**

6. Comparative based method is used in predicting genes.
7. Whole genome shotgun sequencing was developed by J.Craig and V.Smith.
8. Proteomics involves coordinated functioning of cellular proteins.
9. *H.influenzae* was sequenced in the year 1996.
10. Co-immunoprecipitation is a technique that uses antibody.

III Complete the following **(5 x 1 = 5)**

11. Alu elements in the genome organization account for __________.
12. __________ database gives information on the 2D gel electrophoresis of proteins.
13. Discovery of exons and introns occurred in the year __________.
14. __________ method is based on physical and chemical properties of amino acid.
15. An example for phylogenetic analysis database is__________.

IV Answer in one or two sentences, each in about 50 words (5 x 1 = 5)

16. Define reterotransposons.
17. What is promoter?
18. Define trGENE.
19. What is meant by A level in CATH?
20. Expand PAUP.

PART B

V Answer any FIVE of the following, each in about 350 words (5 x 8 = 40)

21. Explain any one method to find genes in larger genomes.
22. Explain whole genome shotgun sequencing method.
23. Write notes on splice site and shotgun libraries.
24. Briefly explain glycosylation, phosphorylation and signal peptides.
25. Explain any two types of intermolecular interactions.
26. Define homology? Explain how it is used in phylogenetic studies.
27. List out the completed genome projects in animals.
28. What are biological pathways? Discuss the role of IT in studying the biological pathways.

PART C

VI Answer the following, each in about 1500 words (2 x 20 = 40)

29. (a) Give a brief account of the genome structure. Add a note on gene prediction program.

OR

(b). Explain the working principle and applications of DNA microarray.

30. (a) Define exons and introns. Explain the various tools used in their prediction.

OR

(b) Draw a dynamic programming for the following sequences

ABCNJRQCLCRPM
AJCJNRCKCRBP

****

LOYOLA COLLEGE (AUTONOMOUS), CHENNAI – 600 034

B.Sc. DEGREE EXAMINATION – **PHYSICS**

FOURTH SEMESTER – NOVEMBER 2016

PB 4208 - BIOINFORMATICS-II (GENOMICS & PROTEOMICS)

Date: 11-11-2016 Dept. No. Max. : 100 Marks
Time: 01:00-04:00

Part –A **(10x2=20 Marks)**

Answer the following, each within 50 words.

1. List any two objectives of learning bioinformatics.
2. Define genome.
3. Give the sequence of 2 stop codons.
4. What are signal peptides?
5. Differentiate orthologs and paralogs.
6. What is a database?
7. Write any two applications of DNA sequencing.
8. Define signal transduction.
9. Expand RCSB. Mention its importance.
10. What is the objective of HGP?

Part - B **(5x7=35 Marks)**

Answer the following, each within 500 words. Draw diagrams and flowcharts wherever necessary.

11. a) Compare prokaryotic and eukaryotic genome.
Or
b) Write short note on: (i) EnSemble Database (ii) *E.coli* Database

12. a) Write about : (i) Splice sites (ii) Introns (iii) Start Codon sequence.
Or
b) Give the importance of codon usage bias

13. a) Illustrate any two post translational modifications of protein.
Or
b) Write the significances of signal peptides and signal peptidases.

14. a) Write short note on Intermolecular interaction.
Or
b) Define SCOP. Explain the classifications of it.

15. a) Write about the potential benefits of HGP.
Or
b) Explain evolutionary tree with graphical representation.

Part – C **(3x15=45 Marks)**

Answer any three of the following, each within 1200 words. Draw diagrams and flowcharts wherever necessary.

16. Describe gene finding in large genomes using different gene finding tools.

17. Describe chain termination DNA sequencing method.

18. Explain the various protein family databases.

19. What is protein-protein interaction? Explain the biochemical methods used to investigate it.

20. Elaborate on the potentials of microarray technology in disease diagnosis.

****

Reg. No:…………

[17BTU403]

KARPAGAM ACADEMY OF HIGHER EDUCATION, COIMBATORE
FIRST INTERNAL EXAMINATION, DECEMBER 2018
DEPARTMENT OF BIOTECHNOLOGY
B.Sc., Biotechnology – Third Semester
GENOMICS AND PROTEOMICS

Time: 2 hour **Maximum: 50 marks**
Date: **Class: II B. Sc**

PART-A

Answer All Questions 20 X 1 = 20 marks

1. What are the repeating units of nucleic acids?
a) phosphate molecules b) nucleotides c) bases d) sugar molecules
2. Which of the following is the correct order of organization of genetic material from largest to smallest?
a) Genome, chromosome, gene, nucleotide
b) Nucleotide, gene, chromosome, genome
c) Gene, nucleotide, chromosome, genome
d) Chromosome, genome, nucleotide, gene
3. Pick one which is suitable for a pseudogene
a) Start codon b) terminal stop codon c) Promoter-less gene d) ORF
4. The enzyme used in Maxam-Gilbert method for P^{32} labelling is
a) polynucelotide kinase b) alkaline phosphatase
c) exonuclease d) terminal nucelotidyl transferase
5. The principle of Snger's sequencing method relies on the use of:
a) chemicals for base specific cleavage b) dNTPs for chain termination
c) ddNTPs for chain termination d) P^{32} Chain termination
6. The sample in Chemical sequencing method after reaction are separated using
a) AGE b) PAGE c) PFGE d) 2-D gel electrophoresis
7. Automated DNA sequencing is an improvement of Sanger's method use
a) ddNTPs b) PCR amplicon
c) Fluorescently labelled dNTPs d) Fluorescently labelled ddNTPs
8. Pseudogenes are important to evolution because they are
a) silent genes that can be activated by mutations b) were not exist until recently c) occurring entirely within genome d) transcribed more frequently
9. Which of these describes a contig?
a) A complete genomic library b) A complete mRNA library
c) An ordered genomic library d) A collection of overlapping sequences
10. The shotgun method----
a) is used in analyze transcriptome b) requires computer analysis
c) meant for small genomes d) a method to introduce gene
11. Guanine specific cleavage in Maxum-Gilbert method is done by
a) Formic acid b) Hydrazene c) Dimethyl sulfoxide d) Piperidine
12. The correct sequence of enzymatic reaction in pyrosequencing—
a) Apyrase- Sulfurylase-luciferrase- DNA polymerase
b) Sulfurylase –Apyrase- DNA polymerase-luciferrase
c) luciferrase- DNA polymerase -Apyrase- Sulfurylase-
d) DNA polymerase- Sulfurylase-luciferrase- Apyrase
13. The promoter is the binding place for -------
a) RNA polymerase b) Repressor c)DNA polymerase d) Inducer
14. Which of the following is a pair wise alignment tool
a) Pfarm b) PSIBLAST c) BLAST d) Prosite
15. The most common output format is
a) FASTA b) PHRED c) Peaks d) spots
16. The most common method for predicting the function of a gene
a) Promoter analysis b) Homology search
c) ORF scan d) SNP analysis
17. Which is NOT a substrate for Klenow fragment in pyrosequencing
a) GTP b) CTP c) ATP d) TTP
18. Reads are the data generated during
a) Pyrosequencing b) Shotgun sequencing
c) Sanger's method d) Maxum Gilbert method
19. Which of the following is NOT a genome assembly software
a) DNASTAR b) Newbler c) HGAP d) Swiss-Prot
20. UCSC is a---------
a) Proteomic tool b) MSA tool
c) Genome browser d) Genetic map resourse

PART-B 3 X 2 = 6 marks

Answer All questions. All questions carry equal marks

21. Write a short note on Pseudogenes
22. Describe Shine-Dalgarno (SD) Sequence?
23. Define genome annotation

PART-C 3 X 8 = 24 marks

Answer all questions choosing either a or b. All questions carry equal marks.

24 **a.** Describe in detail about the similarities and differences between genes and pseudogenes

OR

b. Explain Maxum –Gilbert Method of DNA sequencing

25 **a.** Illustrate the chain termination method of DNA sequencing

OR

b. Elaborate the ideas of shotgun and pyro-sequencing

26 **a.** Discuss various applications of VISTA server

OR

b. Write in detail about the UCSC server and it application in genomics

Reg. No:…………
[17BTU403]

KARPAGAM ACADEMY OF HIGHER EDUCATION, COIMBATORE
SECOND INTERNAL EXAMINATION, FEBRUARY 2019
(For candidates admitted from 2017 and onwards)

DEPARTMENT OF BIOTECHNOLOGY
B.Sc., Biotechnology – Fourth Semester

GENOMICS AND PROTEOMICS

Time: 2 hour **Maximum: 50 marks**
Date: **Class: II B.Sc.,**

PART - A
Answer all the questions **20 X 1 = 20 marks**

1. Which one of the following is *E-coli* specific - genome database?
a) Ecogene b) FlyBase c) OMIM d) PlasmoDB

2. Which one of the following is not a genetic marker?
a) SNP b) ORF c) EST d) STS

3. Which one of the following is *Drosophila melanogaster* – specific genome database?
a) Ecogene b) FlyBase c) OMIM d) PlasmoDB

4. A caret symbol (^) before an entry number of OMIM indicates
a) the entry is fully authenticated b) the entry is not fully annotated
c) the entry needs further verification d) the entry no longer exists

5. Each OMIM entry is given a unique
a) three-digit identifier b) four-digit identifier
c) five-digit identifier d) six-digit identifier

6. The human genome consists of ____________ base pairs.
a) two billion b) ten billion c) one billion d) three billion

7. A ‘codon bias’ is used to
a) genome mapping b) find intergenic sequences
c) identify genes d) intergenic regions

8. Microsatellites are
a) frequently found in bacterial genomes b) always smaller than 10 bp
c) used as DNA markers d) movable DNA elements

9. Minisatellites are
a) Homogeneous b) Heterogeneous c) Homologous d) Orthologous

10. Molecular markers are used to construct
a) Chromosome maps b) Cytogenetic maps c) Physical maps d) Geographic maps

11. The variation in number of tandem repeats between two or more individuals is called
a) VNTRs b) RFLP c) SSRs d) AFLP

12. The variant fragment that distinguish one individual from another one is called
a) variant fragment b) marking fragment c) differing fragment d) variable repeats

13. Which of these is a key characteristic of a molecular marker?
a) It is a known gene b) It is located at a known site on the chromosome
c) It is only useful for linkage and physical mapping studies d) positional analysis

14. A polymorphism is -----
a) Any change in the DNA sequence b) The most common variation of a gene or marker sequence c) The least common variation of a gene or marker sequence
d) Variation of gene or marker sequence present in > 1% of the population

15. A ________ transplant was used to overcome this genetic disorder.
a) liver b) kidney c) bone marrow d) heart

16. A monomorphic DNA segment is
a) A segment of DNA that exists in many forms in the population b) A segment of DNA that controls a single gene function c) A segment of DNA inherited in a dominant fashion
d) A segment of DNA shared by over 99 % of the population

17. Protein-coding genes can be identified by
a) Transposon tagging b) ORF scanning c) Zoo-blotting d) Nuclease S1 mapping

18. The variation in the restriction DNA fragment lengths between individuals of a species is called as
a) Restriction fragment length polymorphism b) RAPD
c) AFLP d) Simple sequence repeats

19. All the following statements are true regarding RFLP and RAPD except
a) RAPD is a quick method compared to RFLP b) RFLP is more reliable than RAPD
c) Species specific primers are required for RAPD d) Radioactive probes are required

20. What type of genome map is the most ideal for understanding the nature of genes?
a) Linkage maps b) Physical maps c) Sequence Maps d) Fingerprint maps

PART – B
3 X 2 = 6 marks
Answer all the questions

21. How are gene-entries classified in OMIM database?
22. What are genetic markers? Give two examples.
23. Write a unique feature of LOD score.

PART - C
3 X 8 = 24 marks
Answer all the questions. Choosing either a or b. All questions carry equal marks.

24. a Briefly write about gateway sites providing links to organism-specific databases and tools for annotating genomic data. (OR)
b. Describe an organism – specific genome database in a detailed manner
25. a. Explain the genetic marker RFLP and its applications in detail. (OR)
b. Explain the microsatellites and minisatellites polymorphisms in detail.
26. a. Describe benefits of genetic mapping with suitable example. (OR)
b. Give a detailed account on a physical mapping technique.