SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

**SCHOOL OF BIO AND CHEMICAL ENGINEERING**

**DEPARTMENT OF BIOINFORMATICS**

# UNIT – I - Omics in Biology – SBIA5204

# UNIT I GENOMICS

Genomes – Prokaryotic and Eukaryotic. Genome sequencing methods and mapping methods. Human Genome Project. Epigenomics

## Genomes

A gene is a stretch of nucleotides that ultimately codes for a polypeptide, which in turn are of structural or functional significance to the cell or the organism. The entire sequence of an organism's hereditary information, including coding and non-coding regions, encoded in DNA is known as "genome". Studying genome, including function and interaction of all the genes of an organism is known as genomics. Success of genome sequencing projects has been remarkable; however, in spite of availability of the entire genome being sequenced, the complex biological processes cannot be unravelled until the role of each of the gene expression products is understood. Proteomics is study of the entire protein complement of an organism, under a given set of conditions. Proteomic studies rely on tools that entail understanding the biochemistry of proteins and the pathways in which these proteins participate in order to bring about a well-orchestrated and harmonious functioning of a given cell or organism in consideration.

## THE CENTRAL DOGMA

The central dogma of molecular biology is the concept which governs the flow of biological information from gene to protein. It includes the major steps of DNA Replication, Transcription and Translation. Replication of DNA allows DNA duplication and ensures that a copy of entire chromosome complement is delivered to each of the resulting daughter cells after cell division. This is followed by a complexly controlled, enzyme-dependent step wherein the DNA is transcribed into mRNA. The mRNA strand thus generated is complementary to the DNA strand. The major difference between the two nucleic acids is that the backbone of RNA contains a Ribose sugar instead of the Deoxyribose sugar present in DNA. Besides, the nitrogenous base Thymine is replaced by Uracil in case of RNA. The RNA thus formed acts as a messenger of genetic information and is translocated from the nucleus (the site of transcription) to the cytoplasm. It is the cytoplasm, where the actual process of Translation or converting the genetic information into polypeptide takes place. The processes of transcription and translation are highly complex involving multiple components and are highly regulated. A departure from the normal flow of these steps, consequently affects further downstream processes.

## Prokaryotic and Eukaryotic Genomes

Prokaryotic and eukaryotic DNA carry genetic information for the development, functioning and reproduction of prokaryotes and eukaryotes respectively. Eukaryotes consist of membrane- bound nucleus whereas prokaryotes lack a membrane-bound nucleus. Prokaryotic DNA is double-stranded and circular. But, eukaryotic DNA is double-strand and linear. The amount of DNA in prokaryotic cells is much less than the amount of DNA in eukaryotic cells. Both prokaryotic and eukaryotic

DNA undergo replication by the enzyme DNA polymerase. The **main difference** between prokaryotic and eukaryotic DNA is that **prokaryotic DNA is found in the cytoplasm whereas eukaryotic DNA is packed into the nucleus of the cell.**

# PROKARYOTIC DNA
VERSUS
# EUKARYOTIC DNA

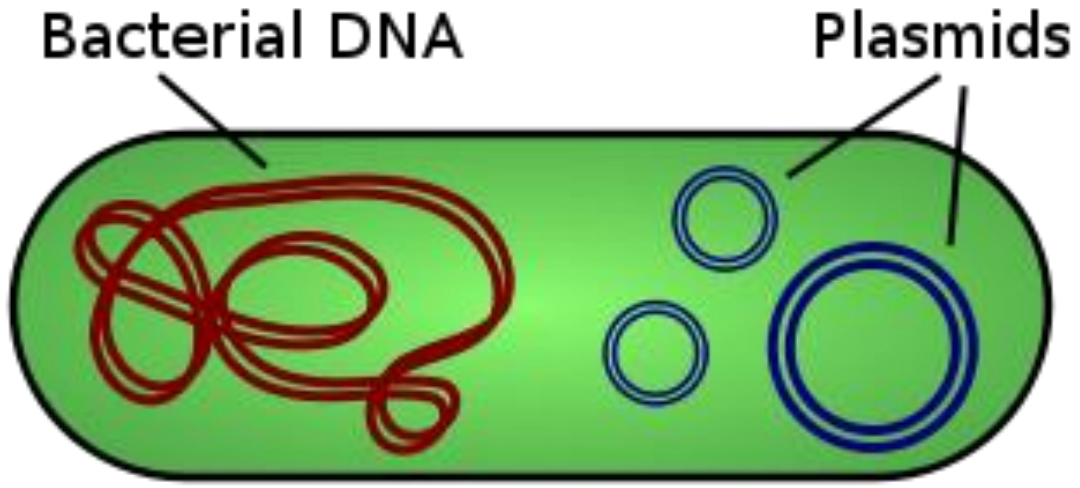
| PROKARYOTIC DNA | EUKARYOTIC DNA |
| --- | --- |
| Prokaryotic DNA is found in the cytoplasm of prokaryotic cells and circular plasmids | Eukaryotic DNA is found in the nucleus of the cell, inside the chloroplast and mitochondria |
| Not found inside organelles | Some DNA is found inside chloroplast and mitochondria |
| Consists of one copy of the genome | Consists of more than one copy of the genome |
| Contains a small number of genes | Contains a large number of genes |
| Organized into a single chromosome | Organized into many chromosomes |
| Not packed with histones | Packed with histones |
| Circular | Linear, with two ends |
| Introns are absent | Introns are present |
| DNA replication occurs in the cytoplasm | DNA replication occurs in the nucleus |
| Chromosome contains a single origin of replication | Chromosome contains many origins of replication |
| DNA replication is rapid | DNA replication is slow |
| Consists of transposons | Do not consist of transposons |
| GC content is more than | T content is more than |

## What is Prokaryotic DNA

The DNA which is carried by prokaryotes is called prokaryotic DNA. Prokaryotic DNA is found in the cytoplasm of bacteria. Some prokaryotic DNA is found as the circular plasmids, carrying additional information. That means prokaryotic DNA does not contain an enclosing nuclear membrane. Prokaryotic DNA is packed into a single circular chromosome. It resides in the region called nucleoid in the cytoplasm. Nucleoid-associated proteins are involved in the packaging of the prokaryotic chromosome in the nucleoid. They help prokaryotic DNA to form a looped structure.

The size of the prokaryotic DNA is around 160,000 to 12.2 million base pairs, depending on the species. Prokaryotic DNA contains a small number of genes. Functionally related genes are organized into operons. Since prokaryotic DNA is rich with genes, the amount of nonfunctional DNA is less. Prokaryotic DNA replication is relatively simple. Prokaryotic chromosome contains a single origin of replication where the initiation of DNA replication occurs. Therefore, a single replication folk and bubble is formed during the replication. The speed of the replication is relatively high in prokaryotes, 2000 nucleotides per second.

DNA polymerase is the enzyme involved in the DNA replication; this contains seven different enzyme families. Out of the seven DNA polymerase families, both prokaryotes and eukaryotes share three families of DNA polymerases: DNA polymerase A, B and Y. DNA polymerase C family is only contained by prokaryotes. Pol III is a replicative DNA polymerase, which belongs to the DNA polymerase family C.

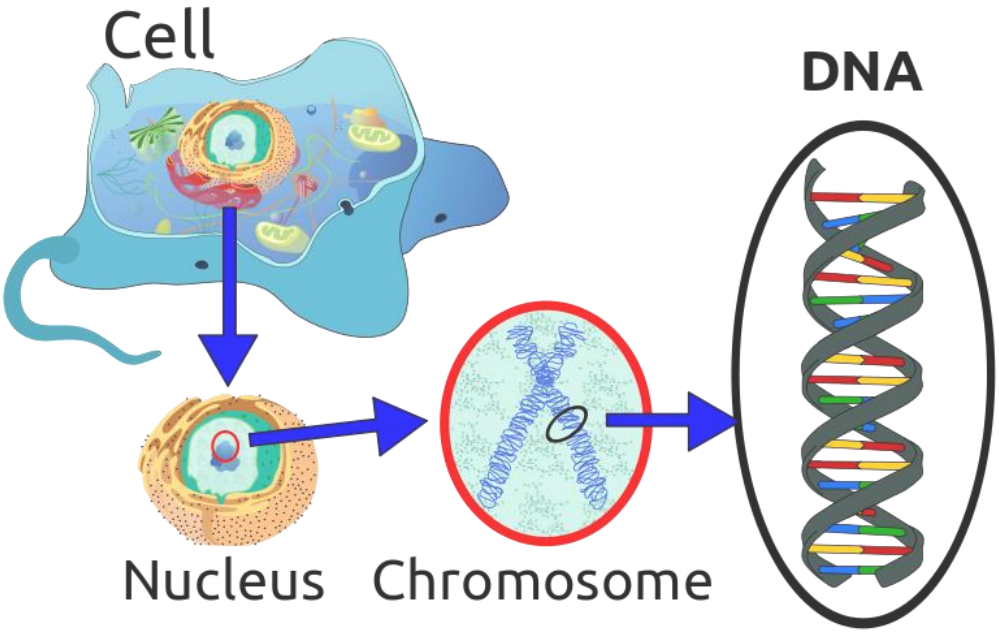



## What is Eukaryotic DNA

The DNA which is contained by eukaryotes is called eukaryotic DNA. Eukaryotic DNA is found in the nucleus of eukaryotic cells. Some eukaryotic DNA is found in organelles like chloroplasts and mitochondria as well. Eukaryotic DNA is enclosed by a nuclear membrane. Eukaryotic DNA is organized into several linear chromosomes. Histones are the proteins, involved in the packaging of eukaryotic chromosomes inside the nucleus. Tight coiling and dense packing are the features of the packing of eukaryotic chromosomes.

Eukaryotes consist of a large number of base pairs in their chromosomes. Most of the eukaryotic DNA consist of several copies of the genome. The size of the human genome is around 2.9 billion base pairs, arranged into 23 homologous chromosome pairs. Eukaryotic genes are encoded for a single protein. Multiple proteins can be achieved by alternative splicing of exons during post transcriptional modifications. The gene density of eukaryotic DNA is low. Hence, the amount of nonfunctional DNA is high in eukaryotic DNA. Eukaryotic DNA replication occurs through multiple origins of replication. The speed of the replication is low in eukaryotes, 100 nucleotides per second.

Multiple protein subunits are involved in the DNA replication of eukaryotes. DNA polymerase families, X like Pol  Pol β, Pol σ, Pol λ, Pol μ and terminal transferases and RT like telomerase are specially contained by  eukaryotes.




Prokaryotic and eukaryotic DNA are the carriers of genetic information required for the development, functioning and reproduction of prokaryotes and eukaryotes. The main difference between prokaryotic and eukaryotic DNA is their quantity, information content, packing and replication. Prokaryotic DNA can be found in the cytoplasm whereas eukaryotic DNA is found in the nucleus, enclosed by the nuclear membrane. Prokaryotic DNA is organized into a single circular chromosome and eukaryotic DNA is organized into several linear chromosomes. The amount of eukaryotic DNA is higher than prokaryotic DNA. Several copies of the genome are found in the eukaryotic DNA as well. The replication of both prokaryotic and eukaryotic DNA occurs in the same way, but prokaryotic DNA replication is relatively simple. Both prokaryotes and eukaryotes contain DNA polymerases, which are capable of replicating and repairing the DNA. Eukaryotes contain telomerase and terminal transferases as well.

Reference:

Bank, E., Leaf Group. “List Ways in which Prokaryotic and Eukaryotic DNA Differ.” List Ways in which Prokaryotic and Eukaryotic DNA Differ | Education – Seattle PI. Seattle PI, 21 Jan. 2014. Web. 23 Apr. 2017.

**Genome sequencing methods and mapping methods**

**Maxam-Gilbert Sequencing:** This method brings about chemical modification of the bases and also requires that the DNA sequence be labeled radioactively at the 5’end. Chemical modification is brought about in such a way that four different types of modified products are generated. The final sequence is deduced by running a SDS-Polyacrylamide gel with the four products in four parallel lanes, which is followed by exposing the gel to X-ray film for autoradiography. Technique is limiting with respect to the use of radioactive labels.

**Sanger Sequencing:** This is also known as the chain termination method, which relies on the fact that; a modified base (for instance dideoxynucleotides instead of the deoxynucleotides) when incorporated during *in vitro* replication causes the process to cease at that point. Each of the four modified bases is labeled with four different fluorescent labels, which are further detected in automated sequencers.

The advent of numerous advancements and improvements in the genomics field has led to the development of high-throughput sequencing methods, which have made it possible that thousands of genes be sequenced at a time. These techniques mainly include: Pyrosequencing, Shotgun Sequencing, Sequencing by synthesis etc.

Currently used sequencing techniques also employ Bacterial Artificial Chromosomes (BACs), which are DNA constructs that are useful for cloning purposes. These cloning vectors can carry DNA inserts of around 150-350 kbp and have been extremely useful in various genome-sequencing projects carried out. The genomic DNA is cleaved using suitable restriction endonuclease and inserted into the bacterial artificial chromosome. The amplified sequences are sequenced using an automated sequencer and then mapped by aligning the overlapping fragments to obtain the original DNA sequence. The information obtained through these sequencing projects is documented in genome databases and are extremely useful in correlating gene and protein sequences.

The use of Next-Gen sequencing platforms has facilitated genomic studies to be carried out on a large-scale and has further expanded the scope of generating information about the structural and functional aspects of various genes. The plethora of attributes associated with the different coding genes can be explored only

when the genomic data is supplemented with data obtained from proteomic studies.

Third Generation Sequencing Technology relies on Scanning Tunnel Electron Microscopy, Fluorescence Resonance Energy Transfer, Single Molecule Real Time Sequencing and Protein nanopores.

**Human Genome Project**

**What is a genome?**

A genome is an organism's complete set of deoxyribonucleic acid (DNA), a chemical compound that contains the genetic instructions needed to develop and direct the activities of every organism. DNA molecules are made of two twisting, paired strands. Each strand is made of four chemical units, called nucleotide bases. The bases are adenine (A), thymine (T), guanine (G) and cytosine (C). Bases on opposite strands pair specifically; an A always pairs with a T, and a C always with a G.

The human genome contains approximately 3 billion of these base pairs, which reside in the 23 pairs of chromosomes within the nucleus of all our cells. Each chromosome contains hundreds to thousands of genes, which carry the instructions for making proteins. Each of the estimated 30,000 genes in the human genome makes an average of three proteins.

**What is DNA sequencing?**

Sequencing means determining the exact order of the base pairs in a segment of DNA. Human chromosomes range in size from about 50,000,000 to 300,000,000 base pairs. Because the bases exist as pairs, and the identity of one of the bases in the pair determines the other member of the pair, scientists do not have to report both bases of the pair.

The primary method used by the HGP to produce the finished version of the human genetic code was map-based, or BAC-based, sequencing. BAC is the acronym for "bacterial artificial chromosome." Human DNA is fragmented into pieces that are relatively large but still manageable in size (between 150,000 and 200,000 base pairs). The fragments are cloned in bacteria, which store and replicate the human DNA so that it can be prepared in quantities large enough for sequencing. If carefully chosen to minimize overlap, it takes about 20,000 different BAC clones to contain the 3 billion pairs of bases of the human genome. A collection of BAC clones containing the entire human genome is called a "BAC library."

In the BAC-based method, each BAC clone is "mapped" to determine where the DNA in BAC clones comes from in the human genome. Using this approach ensures that scientists know both the precise location of the DNA letters that are sequenced from each clone and their spatial relation to sequenced human DNA in other BAC clones.

For sequencing, each BAC clone is cut into still smaller fragments that are about 2,000 bases in length. These pieces are called "subclones." A "sequencing reaction" is carried out on these subclones. The products of the sequencing reaction are then loaded into the sequencing machine (sequencer). The sequencer generates about 500 to 800 base pairs of A, T, C and G from each sequencing reaction, so that each base is sequenced about 10 times. A computer then assembles these short sequences into contiguous stretches of sequence representing the human DNA in the BAC clone.

**Whose DNA was sequenced?**

This was intentionally not known to protect the volunteers who provided DNA samples for sequencing. The sequence is derived from the DNA of several volunteers. To ensure that the identities of the volunteers cannot be revealed, a careful process was developed to recruit the volunteers and to collect and maintain the blood samples that were the source of the DNA.

The volunteers responded to local public advertisements near the laboratories where the DNA "libraries" were prepared. Candidates were recruited from a diverse population. The volunteers provided blood samples after being extensively counseled and then giving their informed consent. About 5 to 10 times as many volunteers donated blood as were eventually used, so that not even the volunteers would know whether their sample was used. All labels were removed before the actual samples were chosen.

**What were the goals?**

The main goals of the Human Genome Project were first articulated in 1988 by a special committee of the U.S. National Academy of Sciences, and later adopted through a detailed series of five-year plans jointly written by the National Institutes of Health and the Department of Energy. The principal goals laid out by the National Academy of Sciences were achieved, including the essential completion of a high-quality version of the human sequence. Other goals included the creation of physical and genetic maps of the human genome, which were accomplished in the mid-1990s, as well as the mapping and sequencing of a set of five model organisms, including the mouse. All of these goals were achieved within the time frame and budget first estimated by the NAS committee.

Notably, quite a number of additional goals not considered possible in 1988 have been added along the way and successfully achieved. Examples include advanced drafts of the sequences of the mouse and rat genomes, as well as a catalog of variable bases in the human genome.

**What is a draft vs. finished genome sequence?**

On June 26, 2000, the International Human Genome Sequencing Consortium announced the production of a rough draft of the human genome sequence. In April, 2003, the International Human Genome Sequencing Consortium is announcing an essentially finished version of the human genome sequence. This version, which is available to the public, provides nearly all the information needed to do research using the whole genome.

The difference between the draft and finished versions is defined by coverage, the number of gaps and the error rate. The draft sequence covered 90 percent of the genome at an error rate of one in 1,000 base pairs, but there were more than 150,000 gaps and only 28 percent of the genome had reached the finished standard. In the April 2003 version, there are less than 400 gaps and 99 percent of the genome is finished with an accuracy rate of less than one error every 10,000 base pairs. The differences between the two versions are significant for scientists using the sequence to conduct research.

**Who owns the human genome?**

Every part of the genome sequenced by the Human Genome Project was made public immediately, and new information about the genome is posted almost every day in freely accessible databases or published in scientific journals (which may or may not be freely available to the public).

The Supreme Court ruled in 2013 that naturally occurring human genes are not an invention and therefore cannot be patented. However, private companies can apply for patents on edited

or synthetic genes, which have been altered significantly from their natural versions to count as a new, patentable, product.

**Who participated?**

The Human Genome Project could not have been completed s quickly and as effectively without the strong participation of international institutions. In the United States, contributors to the effort include the National Institutes of Health (NIH), which began participation in 1988 when it created the Office for Human Genome Research, later upgraded to the National Center for Human Genome Research in 1990 and then the National Human Genome Research Institute (NHGRI) in 1997; and the U.S. Department of Energy (DOE), where HGP discussions began as early as 1984. However, almost all of the actual sequencing of the genome was conducted at numerous universities and research centers throughout the United States, the United Kingdom, France, Germany, Japan and China.

The International Human Genome Sequencing Consortium included:

- The Whitehead Institute/MIT Center for Genome Research, Cambridge, Mass., U.S.
- The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, U. K.
- Washington University School of Medicine Genome Sequencing Center, St. Louis, Mo., U.S.
- United States DOE Joint Genome Institute, Walnut Creek, Calif., U.S.
- Baylor College of Medicine Human Genome Sequencing Center, Department of Molecular and Human Genetics, Houston, Tex., U.S.
- RIKEN Genomic Sciences Center, Yokohama, Japan
- Genoscope and CNRS UMR-8030, Evry, France
- GTC Sequencing Center, Genome Therapeutics Corporation, Waltham, Mass., USA
- Department of Genome Analysis, Institute of Molecular Biotechnology, Jena, Germany
- Beijing Genomics Institute/Human Genome Center, Institute of Genetics, Chinese Academy of Sciences, Beijing, China
- Multimegabase Sequencing Center, The Institute for Systems Biology, Seattle, Wash.
- Stanford Genome Technology Center, Stanford, Calif., U.S.
- Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, Calif., U.S.
- University of Washington Genome Center, Seattle, Wash., U.S.
- Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan
- University of Texas Southwestern Medical Center at Dallas, Dallas, Tex., U.S.
- University of Oklahoma's Advanced Center for Genome Technology, Dept. of Chemistry and Biochemistry, University of Oklahoma, Norman, Okla., U.S.
- Max Planck Institute for Molecular Genetics, Berlin, Germany

- Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center, Cold Spring Harbor, N.Y., U.S.
- GBF - German Research Centre for Biotechnology, Braunschweig, Germany

**Why does NHGRI study ethical issues?**

Since the beginning of the Human Genome Project, it has been clear that expanding our knowledge of the genome would have a profound impact on individuals and society. The leaders of the Human Genome Project recognized that it would be important to address a wide range of ethical and social issues related to the acquisition and use of genomic information, in order to balance the potential risks and benefits of incorporating this new knowledge into research and clinical care. The Ethical, Legal, and Social Implications (ELSI) program at NHGRI was established in 1990 to oversee research in these areas.

The United States Congress mandates that no less than five percent of the annual NHGRI budget is dedicated to studying the ethical, legal and social implications of human genome research, as well as recommending policy solutions and stimulating public discussion. The ELSI program at NHGRI, which is unprecedented in biomedical science in terms of scope and level of priority, provides an effective basis from which to assess the implications of genome research.

Since its inception the ELSI program at NHGRI has made several notable contributions to the genomics field. Among these are major changes to the way investigators and institutional review boards handle the consent process for genomics studies. Another is key guidance on the NIH's genomic data sharing policy, notably the need to balance open science with personal privacy and autonomy. The ELSI program has been effective in promoting dialogue about the implications of genomics, and shaping the culture around the approach to genomics in research, medical, and community settings.

**What is the future of medical science?**

Having the essentially complete sequence of the human genome is similar to having all the pages of a manual needed to make the human body. The challenge to researchers and scientists now is to determine how to read the contents of all these pages and then understand how the parts work together and to discover the genetic basis for health and the pathology of human disease. In this respect, genome-based research will eventually enable medical science to develop highly effective diagnostic tools, to better understand the health needs of people based on their individual genetic make-ups, and to design new and highly effective treatments for disease.

Individualized analysis based on each person's genome will lead to a very powerful form of preventive medicine. We'll be able to learn about risks of future illness based on DNA analysis. Physicians, nurses, genetic counselors and other health-care professionals will be able to work with individuals to focus efforts on the things that are most likely to maintain health for a particular individual. That might mean diet or lifestyle changes, or it might mean medical surveillance. But there will be a personalized aspect to what we do to keep ourselves healthy. Then, through our understanding at the molecular level of how things like diabetes or heart disease or schizophrenia come about, we should see a whole new generation of interventions,

many of which will be drugs that are much more effective and precise than those available today.

## How did it impact research?

Biological research has traditionally been a very individualistic enterprise, with researchers pursuing medical investigations more or less independently. The magnitude of both the technological challenge and the necessary financial investment prompted the Human Genome Project to assemble interdisciplinary teams, encompassing engineering and informatics as well as biology; automate procedures wherever possible; and concentrate research in major centers to maximize economies of scale.

As a result, research involving other genome-related projects (e.g., the International HapMap Project to study human genetic variation and the Encyclopedia of DNA Elements, or ENCODE, project) is now characterized by large-scale, cooperative efforts involving many institutions, often from many different nations, working collaboratively. The era of team-oriented research in biology is here.

In addition to introducing large-scale approaches to biology, the Human Genome Project has produced all sorts of new tools and technologies that can be used by individual scientists to carry out smaller scale research in a much more effective manner.

### Genome Assemblies

A **reference genome** (also known as a **reference assembly**) is a digital nucleic acid sequence database, assembled by scientists as a representative example of the set of genes in one idealized individual organism of a species. As they are assembled from the sequencing of DNA from a number of individual donors, reference genomes do not accurately represent the set of genes of any single individual organism. Instead a reference provides a haploid mosaic of different DNA sequences from each donor. There are reference genomes for multiple species of viruses, bacteria, fungus, plants, and animals.

For example, the human reference genome, *GRCh38*, from the Genome Reference Consortium is derived from thirteen anonymous volunteers.

As the cost of DNA sequencing falls, and new full genome sequencing technologies emerge, more genome sequences continue to be generated. Reference genomes are typically used as a guide on which new genomes are built, enabling them to be assembled much more quickly and cheaply than the initial Human Genome Project. Most individuals with their entire genome sequenced, such as James D. Watson, had their genome assembled in this manner. For much of a genome, the reference provides a good approximation of the DNA of any single individual. But in regions with high allelic diversity, such as the major histocompatibility complex in humans and the major urinary proteins of mice, the reference genome may differ significantly from other individuals. Comparison between the reference (build 36) and Watson's genome revealed 3.3 million single nucleotide polymorphism differences, while about 1.4 percent of his DNA could not be matched to the reference genome at all. For regions where there is known to be large scale variation, sets of alternate loci are assembled alongside the reference locus.

Reference genomes can be accessed online at several locations, using dedicated browsers such as Ensembl or UCSC Genome Browser.

### Epigenetics and Epigenomics

In biology, **epigenetics** is the study of heritable phenotype changes that do not involve alterations in the DNA sequence. The Greek prefix *epi-* (ἐπι- "over, outside of, around") in *epigenetics* implies features that are "on top of" or "in addition to" the traditional genetic basis for inheritance. Epigenetics most often

involves changes that affect gene activity and expression, but the term can also be used to describe any heritable phenotypic change. Such effects on cellular and physiological phenotypic traits may result from external or environmental factors, or be part of normal development. The standard definition of epigenetics requires these alterations to be heritable in the progeny of either cells or organisms.

The term also refers to the changes themselves: functionally relevant changes to the genome that do not involve a change in the nucleotide sequence. Examples of mechanisms that produce such changes are DNA methylation and histone modification, each of which alters how genes are expressed without altering the underlying DNA sequence. Gene expression can be controlled through the action of repressor proteins that attach to silencer regions of the DNA. These epigenetic changes may last through cell divisions for the duration of the cell's life, and may also last for multiple generations, even though they do not involve changes in the underlying DNA sequence of the organism; instead, non-genetic factors cause the organism's genes to behave (or "express themselves") differently.

One example of an epigenetic change in eukaryotic biology is the process of cellular differentiation. During morphogenesis, totipotent stem cells become the various pluripotent cell lines of the embryo, which in turn become fully differentiated cells. In other words, as a single fertilized egg cell – the zygote – continues to divide, the resulting daughter cells change into all the different cell types in an organism, including neurons, muscle cells, epithelium, endothelium of blood vessels, etc., by activating some genes while inhibiting the expression of others.

Historically, some phenomena not necessarily heritable have also been described as epigenetic. For example, the term "epigenetic" has been used to describe any modification of chromosomal regions, especially histone modifications, whether or not these changes are heritable or associated with a phenotype. The consensus definition now requires a trait to be heritable for it to be considered epigenetic.

**Epigenomics** is the study of the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. The field is analogous to genomics and proteomics, which are the study of the genome and proteome of a cell. Epigenetic modifications are reversible modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence. Epigenomic maintenance is a continuous process and plays an important role in stability of eukaryotic genomes by taking part in crucial biological mechanisms like DNA repair. Plant flavones are said to be inhibiting epigenomic marks that cause cancers. Two of the most characterized epigenetic modifications are DNA methylation and histone modification. Epigenetic modifications play an important role in gene expression and regulation, and are involved in numerous cellular processes such as in differentiation/development and tumorigenesis. The study of epigenetics on a global level has been made possible only recently through the adaptation of genomic high-throughput assays.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

**SCHOOL OF BIO AND CHEMICAL ENGINEERING**

**DEPARTMENT OF BIOINFORMATICS**

# UNIT – II - Omics in Biology – SBIA5204

UNIT II PROTEOMICS AND PEPTIDOMICS

Proteome characterisation, purification, separation – 2D gel electrophoresis, PAGE and affinity chromatography. Protein sequencing – Edman degradation. Interaction Proteomics – protein-protein and protein-DNA interactions. Protein microarrays. Peptidomics

**Proteome and Proteomics**

The field of proteomics originated from the research and development of the Human Genome Project to understand the proteome for the composition, structure, specific activity patterns and unique properties of proteins essential to provide data that complements the genomic information. Genomics deals only with the studies of the entire gene compendium of a particular cell or organism at a given point of time under a defined set of conditions. It is rightly perceived that Genomics is just the starting page in the book of understanding biological functions and the mechanisms that are responsible for maintaining a myriad of biological processes to run smoothly. The delineated role and function of a particular component cannot be completely understood until the gene products are studied in details. Genomics does not take into consideration the fact that a single stretch of nucleotides could give rise to different protein products when processed in different ways, namely in the intermediate steps that occur during the post- transcription and post-translation. This makes it necessary that the proteome of the cell or organism be studied in order to gain deeper insights into the structural and functional significance of the gene products. It is important to note that the proteome when studied, would include the entire protein complement of the cell or organism, and would involve proteins that have already undergone all the modifications that occur intermittently. Besides that, proteomic studies not only provide information about individual proteins but also about the interactions that occur between different proteins, which are in turn responsible for essential biological processes.

Proteomics

Advantages

Gives a real picture about the cellular activity, as all the data that is procured is about proteins that have either structural or functional significance to the cell.

Availability of protein sequencing techniques like Edman Degradation or Mass Spectrometry has enabled the elucidation of deeper insights into proteomic studies.

Limitations

- There is practically no method available for protein amplification as of now.
- For synthesizing a protein right from the beginning, it is essential to have the DNA sequence that encodes for the protein in question.

- Availability of different Bioinformatic- based algorithms like HMM, which also takes into consideration the different conformations that would be a result of protein folding.

Proteomics studies have limitation with respect to reproducibility.

**Proteome Characterisation**

2D gel electrophoresis & PAGE

Electrophoresis (*Electro* refers to the energy of electricity and *Phoresis*, from the Greek verb *phoros*, means to carry across) is a technique for separating or resolving charged molecules (such as amino acids, peptides, proteins, nucleotides, and nucleic acids) in a mixture under the influence of an applied electric field. Charged molecules in an electric field move or mi- grate, at a speed determined by their *charge to mass* ratio.

**Gel electrophoresis**

In gel electrophoresis, gel serves as molecular sieve. There are two basic types of materials used to make gels: agarose and polyacrylamide. Agarose is a natural colloid extracted from seaweed. Agarose gels have very large pore size and are used primarily to separate very large molecules with a molecular mass greater than 200 kDa. Agarose is a linear polysaccharide made up of the basic repeat unit agarobiose, which comprises alternating units of galactose and 3,6-anhydrogalactose. Agarose is usually used at concentrations between 1% and 3%. Agarose gels are used for the electrophoresis of both proteins and nucleic acids.

A polyacrylamide gel consists of chains of acrylamide monomers (CH2 =CH–CO–NH2 ) cross-linked with N, N'-methylenebisacrylamide units (CH2 =CH–CO–NH–CH2 –NH–CO–CH=CH2 ), the latter commonly called bis. The pore size of the gel is determined by both the total concentration of monomers (acrylamide + bis) and the ratio of acrylamide to bis. Polymerization of the acrylamide : bis solution is initiated by ammonium persulfate and catalyzed by TEMED (N, N, N', N'-tetramethylethylenediamine).

**SDS-PAGE**

The separation of macromolecules in an electric field is called *electrophoresis*. A very common method for separating proteins by electrophoresis uses a discontinuous polyacrylamide gel as a support medium and sodium dodecyl sulfate (SDS) to denature the proteins. The method is called sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). The most commonly used system is also called the Laemmli method after U.K. Laemmli, who was the first to publish a paper employing SDS-PAGE in a scientific study.

SDS (also called lauryl sulfate) is an anionic detergent, meaning that when dissolved its molecules have a net negative charge within a wide pH range. A polypeptide chain binds amounts of SDS in proportion to its relative molecuar mass. The negative charges on SDS

destroy most of the complex structure of proteins, and are strongly attracted toward an anode (positively-charged electrode) in an electric field.

Polyacrylamide gels restrain larger molecules from migrating as fast as smaller molecules. Because the charge-to-mass ratio is nearly the same among SDS-denatured polypeptides, the final separation of proteins is dependent almost entirely on the differences in relative molecular mass of polypeptides. In a gel of uniform density the relative migration distance of a protein (Rf, the f as a subscript) is negatively proportional to the log of its mass. If proteins of known mass are run simultaneously with the unknowns, the relationship between Rf and mass can be plotted, and the masses of unknown proteins estimated.

Protein separation by SDS-PAGE can be used to estimate relative molecular mass, to determine the relative abundance of major proteins in a sample, and to determine the distribution of proteins among fractions. The purity of protein samples can be assessed and the progress of a fractionation or purification procedure can be followed. Different staining methods can be used to detect rare proteins and to learn something about their biochemical properties. Specialized techniques such as Western blotting, two-dimensional electrophoresis, and peptide mapping can be used to detect extremely scarce gene products, to find similarities among them, and to detect and separate isoenzymes of proteins.

**Polyacrylamide gels for SDS-PAGE**

SDS-PAGE (sodium dodecyl sulphate-polyacrylamide gel electrophoresis) is commonly used in the lab for the separation of proteins based on their molecular weight. It's one of those techniques that is commonly used but not frequently fully understood. So let's try and fix that.

SDS-PAGE separates proteins according to their molecular weight, based on their differential rates of migration through a sieving matrix (a gel) under the influence of an applied electrical field.

**Making the Rate of Protein Migration Proportional to Molecular Weight**

The movement of any charged species through an electric field is determined by its net charge, its molecular radius and the magnitude of the applied field. But the problem with natively folded proteins is that neither their net charge nor their molecular radius is molecular weight dependent. Instead, their net charge is determined by amino acid composition i.e. the sum of the positive and negative amino acids in the protein and molecular radius by the protein's tertiary structure.

So in their native state, different proteins with the same molecular weight would migrate at different speeds in an electrical field depending on their charge and 3D shape.

To separate proteins in an electrical field based on their molecular weight only, we need to destroy the tertiary structure by reducing the protein to a linear molecule, and somehow mask the intrinsic net charge of the protein. That's where SDS comes in.

**The Role of SDS (et al)**

SDS is a detergent that is present in the SDS-PAGE sample buffer where, along with a bit of boiling, and a reducing agent (normally DTT or B-ME to break down protein-protein disulphide bonds), it disrupts the tertiary structure of proteins. This brings the folded proteins down to linear molecules.

SDS also coats the protein with a uniform negative charge, which masks the intrinsic charges on the R-groups. SDS binds fairly uniformly to the linear proteins (around 1.4g SDS/ 1g protein), meaning that the charge of the protein is now approximately proportional to its molecular weight.

SDS is also present in the gel to make sure that once the proteins are linearized and their charges masked, they stay that way throughout the run.

The dominant factor in determining an SDS-coated protein is it's molecular radius. SDS-coated proteins have been shown to be linear molecules, 18 Angstroms wide and with length proportional to their molecular weight, so the molecular radius (and hence their mobility in the gel) is determined by the molecular weight of the protein. Since the SDS-coated proteins have the same charge to mass ratio, there will be no differential migration based on charge.

**The Gel Matrix**

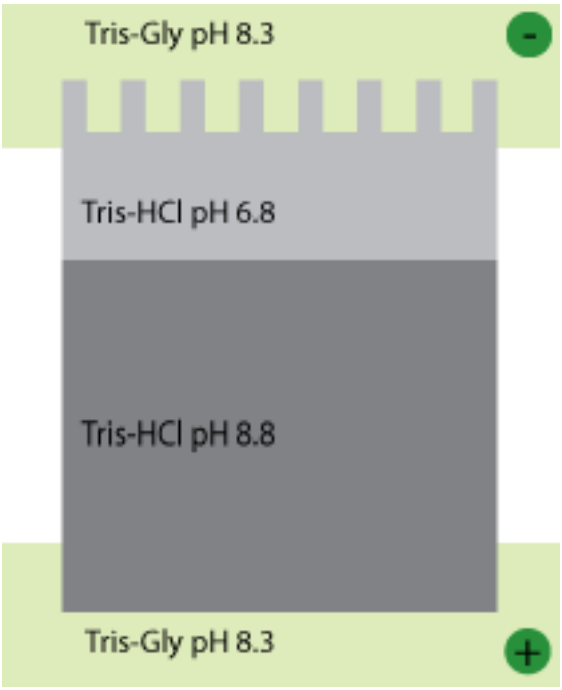
In an applied electrical field, the SDS-treated proteins will now move toward the positive anode at different rates depending on their molecular weight. These different mobilities will be exaggerated due to the high-friction environment of a gel matrix.

As the name suggests, the gel matrix used for SDS-PAGE is polyacrylamide, which is a good choice because it is chemically inert and, crucially, can easily be made up at a variety concentrations to produce different pore sizes giving a variety of separating conditions that can be changed depending on your needs. You may remember that I previously wrote an article about the mechanism of acrylamide polymerization.

The Discontinuous Buffer System and the Stacking Gel – Lining Them Up at the Starting Line

To conduct the current from the cathode (negative) to the anode (positive) through the gel, a buffer is obviously needed. Mostly we use the discontinuous Laemmli buffer system. "Discontinuous" simply means that the buffer in the gel and the tank are different.

Typically, the system is set up with a stacking gel at pH 6.8, buffered by Tris-HCl, a running gel buffered to pH 8.8 by Tris-HCl and an electrode buffer at pH 8.3. The stacking gel has a low concentration of acrylamide and the running gel a higher concentration capable of retarding the movement of the proteins.




So what's with all of those different pH's?

Well, glycine can exist in three different charge states, positive, neutral or negative, depending on the pH. This is shown in the diagram below. Control of the charge state of the glycine by the different buffers is the key to the whole stacking gel thing.

So here's how the stacking gel works. When the power is turned on, the negatively-charged glycine ions in the pH 8.3 electrode buffer are forced to enter the stacking gel, where the pH is 6.8. In this environment, glycine switches predominantly to the zwitterionic (neutrally charged) state. This loss of charge causes them to move very slowly in the electric field.

The Cl- ions (from Tris-HCl) on the other hand, move much more quickly in the electric field and they form an ion front that migrates ahead of the glycine. The separation of Cl- from the Tris counter-ion (which is now moving towards the anode) creates a narrow zone with a steep voltage gradient that pulls the glycine along behind it, resulting in two narrowly separated fronts of migrating ions; the highly mobile Cl- front, followed by the slower, mostly neutral glycine front.

All of the proteins in the gel sample have an electrophoretic mobility that is intermediate between the extreme of the mobility of the glycine and Cl-, so when the two fronts sweep through the sample well, the proteins are concentrated into the narrow zone between the Cl- and glycine fronts.

This procession carries on until it hits the running gel, where the pH switches to 8.8. At this pH the glycine molecules are mostly negatively charged and can migrate much faster than the proteins. So the glycine front accelerates past the proteins, leaving them in the dust.

The result is that the proteins are dumped in a very narrow band at the interface of the stacking and running gels and since the running gel has an increased acrylamide concentration, which slows the the movement of the proteins according to their size, the separation begins.

If you are still wondering why the stacking gel is needed, think of what would happen if you didn't use one.

Gel wells are around 1cm deep and you generally need to substantially fill them to get enough protein onto the gel. So in the absence of a stacking gel, your sample would sit on top of the running gel, as a band of up to 1cm deep.

Rather than being lined up together and hitting the running gel together, this would mean that the proteins in your sample would all enter the running gel at different times, resulting in very smeared bands.

So the stacking gel ensures that all of the proteins arrive at the running gel at the same time so proteins of the same molecular weight will migrate as tight bands.

**Separation**

Once the proteins are in the running gel, they are separated because higher molecular weight proteins move more slowly through the porous acrylamide gel than lower molecular weight proteins. The size of the pores in the gel can be altered depending on the size of the proteins you want to separate by changing the acrylamide concentration. Typical values are shown below.

| % Acrylamide | MW Range (kDa) |
|---|---|
| 7 | 50 – 500 |
| 10 | 20 – 300 |
| 12 | 10 – 200 |
| 15 | 3 - 100 |

For a broader separation range, or for proteins that are hard to separate, a gradient gel, which has layers of increasing acrylamide concentration, can be used.

I think that's about it for Laemmli SDS-PAGE. If you have any questions, corrections or anything further to add, please do get involved in the comments section!

**Affinity chromatography**

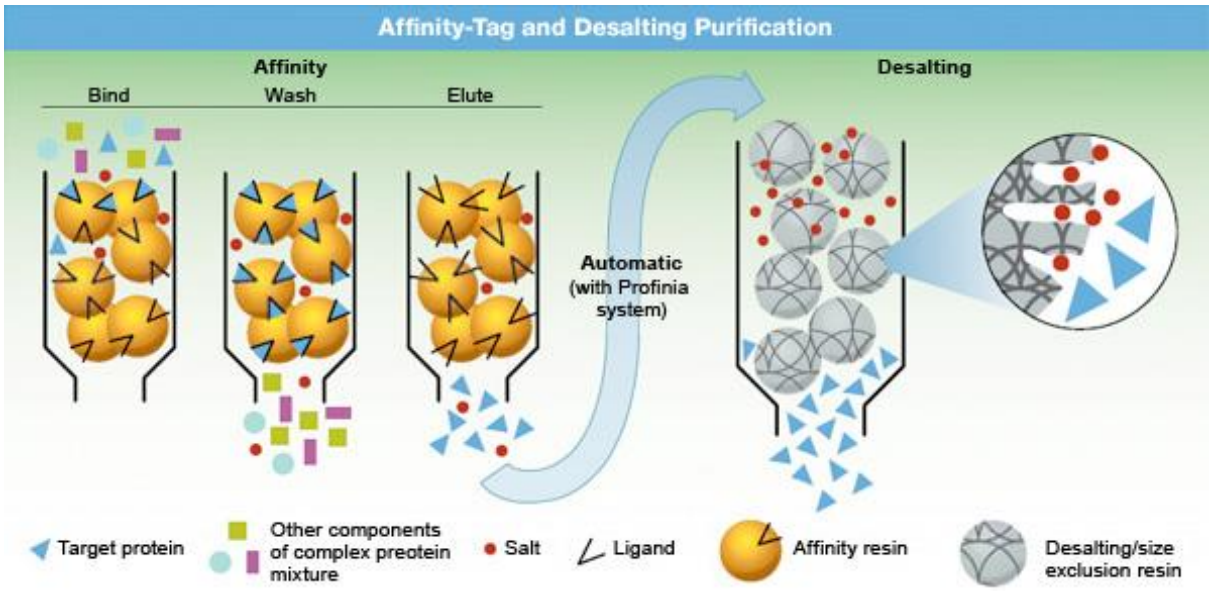
Affinity chromatography is a separation method based on a specific binding interaction between an immobilized ligand and its binding partner. Examples include antibody/antigen, enzyme/substrate, and enzyme/inhibitor interactions. The degree of purification can be quite high depending on the specificity of the interaction and, consequently, it is generally the first step, if not the only step, in a purification strategy.

**Why Use Affinity Chromatography?**

Affinity chromatography offers high selectivity, resolution, and capacity in most protein purification schemes. It has the advantage of utilizing a protein's biological structure or function for purification. As a result, purifications that would otherwise be time consuming and complicated, can often be easily achieved with affinity chromatography.

## Mechanism of Affinity Binding

A commonly used metaphor to illustrate affinity binding is the lock and key analogy. A unique structure present on the surface of a protein is the key that will only bind to the corresponding lock, a specific ligand on a chromatographic support.




## Affinity-tagged purification.

In two-step affinity-tagged protein purification, a protein is first purified by affinity chromatography, then desalted. In some medium pressure chromatography systems, such as the NGC medium pressure chromatography systems, these two steps can be automated. In the first step, a recombinant protein mixture is passed over a chromatography support containing a ligand that selectively binds proteins that contain an affinity-tag sequence (typically His or GST). Contaminants are washed away, and the bound protein is then eluted in pure form.

Affinity tags have different advantages. In immobilized metal affinity chromatography (IMAC), His binds with good selectivity to Ni or other transition metals immobilized to the ligand; the tagged protein can be selectively eluted with imidazole. proteins tagged with GST bind to glutathione as the ligand, and are eluted with solutions of glutathione. Proteins with an enzymatically active GST fusion tag can only be purified under native conditions. In contrast, polyhistidine-tagged proteins may be purified under native or denaturing conditions.

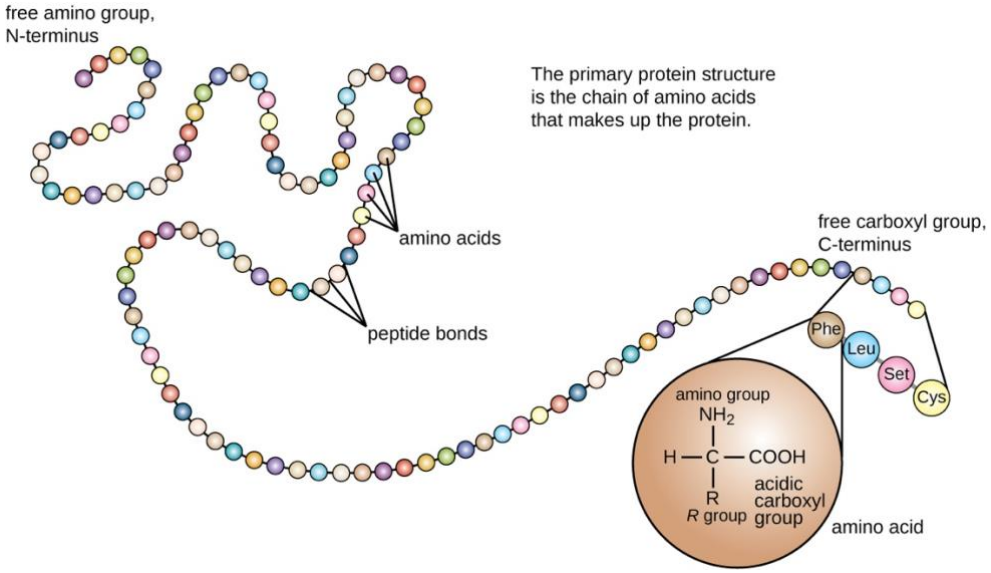
During the second step of desalting, affinity-purified samples can simultaneously undergo buffer exchange to remove salts in preparation for downstream applications.

A number of desalting techniques, including size exclusion chromatography, dialysis, and ultrafiltration, also allow buffer exchange. Desalting often includes the removal not only of salt, but also of other foreign substances, such as detergents, nucleotides, and lipids.

Affinity chromatography can be broadly divided into two method types:

- The first method uses a naturally occurring structure or sequence of amino acids on the protein as the binding site. Examples include the affinity of Affi-Gel Blue support binding for albumin's bilirubin-binding site and the binding of protein A in the Affi-Gel and Affi-Prep protein A supports to the Fc region of IgG. An important consideration for antibody purification is to determine the affinity of your target antibody for protein A/G chromatography media, which varies widely.
- The second method involves binding to a special amino acid sequence engineered into the protein of interest, commonly referred to as a "tag". A number of different tags are available. Two of the most commonly used protein tags are the polyhistidine tag, which binds to certain metal-containing complexes such as those in Profinity™ IMAC resins, and the glutathione s-transferase (GST) sequence, which binds to glutathione, found in Bio-Scale™ Mini Profinity™ GST media. Theoretically, any protein can be purified using the tagging method; however, many factors must be considered to design a process to purify tagged recombinant proteins.

**Protein sequencing – Edman degradation**




The sequence of amino acids in a protein or peptide can be identified by Edman degradation, which was developed by Pehr Edman. This method can label and cleave the peptide from N-terminal without disrupting the peptide bonds between other amino acid residues. The Edman degradation reaction was automated in 1967 by Edman and Beggs. Nowadays, the automated Edman degradation (the protein sequenator) is used widely, and it can sequence peptides up to 50 amino acids.

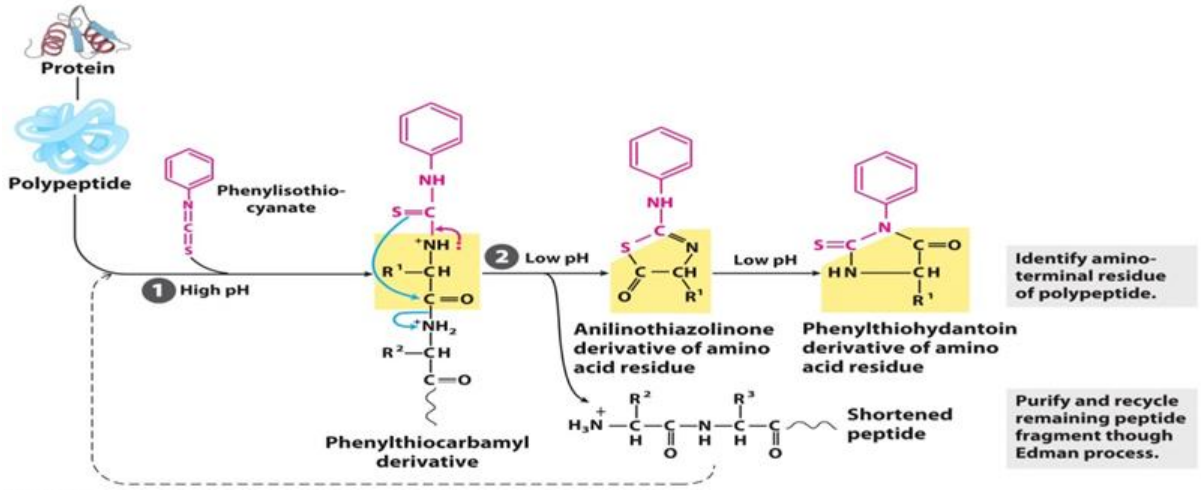



**Figure 3-27**
*Lehninger Principles of Biochemistry*, Sixth Edition


Cyclic degradation of peptides based on the reaction of phenylisothiocyanate with the free amino group of the N-terminal residue such that amino acids are removed one at a time and identified as their phenylthiohydantoin derivatives. Speaking to the specific process, an uncharged peptide is reacted with phenylisothiocyanate (PITC) at the amino terminus under mildly alkaline conditions to give a phenylthiocarbamoyl derivative (PTC-peptide). Then, under acidic conditions, the thiocarbonyl sulfur of the derivative attacks the carbonyl carbon of the N-terminal amino acid. The first amino acid is cleaved as anilinothiazolinone derivative (ATZ-amino acid) and the remainder of the peptide can be isolated and subjected to the next degradation cycle. Once formed, this thiazolone derivative is more stable than phenylthiocarbamyl derivative. The ATZ amino acid is then removed by extraction with ethyl acerate and converted to a phenylthiohydantoin derivative (PTH-amino acid). And the chromatography can be used to identify the PTH residue generated by each cycle.

As to the automated Edman degradation, proteins can be analyzed by applying them in-solution onto a TFA filter and then loaded onto the Edman sequencing instrument. Proteins in mixtures are first separated by 1D or 2D gels and then blotted onto a PVDF membrane. The proteins are detected by Coomassie blue, Amido black or Poncau S staining and the proteins of interest cut out and the PVDF membrane piece loaded onto the Edman sequencer.

With mass spectrometry was developed, the use of Edman degradation sequencing began to decrease. However, it stills remains the methods for several types of protein structural analysis applications. It can be used to verify the N-terminal boundary of recombinant proteins or determining the N-terminus of protease-resistant domains, particularly when the protein or domain is >40 to 80 kDa or cannot be readily purified. It also can be used to identify the new N-terminal and proteolytic cleavage site in the protein fragments. In addition, as to some novel proteins and peptides where sequence databases are not available for MS/MS database searching, Edman degradation can be used for analysis.

The N-terminal amino acid of the protein can be cleaved off. Thus, in the process, the first cycle thus identifies the exact N-terminal amino acid. In addition, because the released amino acids are identified and quantified by chromatography, the amino acids with identical molecular weight can be identified. For example, isoleucine and leucine have a mass of 113 Da, but they have different retention time. Moreover, Edman sequencing can be performed on PVDF blots from 1D and 2D gels, which enables N-terminal sequencing of proteins in the mixture. However, Edman degradation sequencing will not be available when the peptide whose N-terminus has been chemically modified, such as acetylation. And as the PITC cannot reactive with non-α-amino acid, Sequencing will stop if a non-α-amino acid is encountered like isoaspartic acid. Moreover, larger proteins cannot be sequenced by the Edman sequencing.

**Interaction Proteomics – Protein-protein and Protein-DNA interactions.**

Decades of research into cell biology, molecular biology, biochemistry, structural biology, and biophysics have produced a remarkable compendium of knowledge on the function and molecular properties of individual proteins. This knowledge is well recorded and manually curated into major protein databases like UniProt. However, proteins rarely act alone. Many times they team up into "molecular machines" and have intricate physicochemical dynamic connections to undertake biological functions at both cellular and systems levels. A critical step towards unraveling the complex molecular relationships in living systems is the mapping of protein-to-protein physical "interactions". The complete map of protein interactions that can occur in a living organism is called the interactome. Interactome mapping has become one of the main scopes of current biological research, similar to the way "genome" projects were a driving force of molecular biology 20 years ago.

Efficient large-scale technologies that measure proteome-wide physical connections between protein pairs are essential for accomplishing a comprehensive knowledge of the protein interactomes. In recent years, given an explosive development of high-throughput experimental technologies, the number of reported protein–protein interactions (PPIs) has increased substantially. Large collections of PPIs produce "omic" scale views of protein partners and protein memberships in complexes and assemblies. Over the same period as the development of large-scale technologies, efficient collection of a lot of small-scale experimental data published in relevant scientific journals is also taking place. This data compilation work is just as essential to achieving comprehensive knowledge of the interactome. Important efforts have been made to build public repositories that integrate information from large- and small-scale PPI experiments reported in the scientific literature. A compendium of PPI databases can be found in http://www.pathguide.org/.

The first step needed is to define precisely what protein–protein interactions are. Commonly they are understood as physical contacts with molecular docking between proteins that occur in a cell or in a living organism in vivo. As discussed previously the issue of whether two proteins share a "functional contact" is quite distinct from the question of whether the same two proteins interact directly with each other. Any protein in the ribosome or in the basal transcriptional apparatus shares a functional contact with the other proteins in the complex, but certainly not all the proteins in the particular complex interact. Indubitably, the existence of

many other types of functional links between biomolecular entities (genes, proteins, metabolites, etc.) in living organisms should not be confused with protein physical interactions. Investigating these functional links requires different experimental techniques designed to find such specific types of relationships, for example, double mutant synthetic lethality to find genetic interactions or transcriptome expression profiling to find gene co-expression. Identification of other types of protein interactions (protein–DNA, protein–RNA, protein–cofactor, or protein–ligand) is also important for a comprehensive study of the interactome, but again these types of data should not be mixed or confused with PPI data.

The physical contact considered in PPIs should be specific, not just all proteins that bump into each other by chance. It also should exclude interactions that a protein experiences when it is being made, folded, quality checked, or degraded. For example, all proteins at one point "touch" the ribosome, many touch chaperones, and most make contact with the degradation machinery. In many experimental assays, such generic interactions are rightfully filtered out. Therefore, the definition of PPI has to consider (1st) the interaction interface should be intentional and not accidental, i.e., the result of specific selected biomolecular events/forces; and (2nd) the interaction interface should be non-generic, i.e., evolved for a specific purpose distinct from totally generic functions such as protein production, degradation, and others.

That PPIs imply physical contact between proteins does not mean that such contacts are static or permanent. The cell machinery undergoes continuous turnover and reassembly. Some protein assemblies are stable because they constitute macromolecular protein complexes and cellular machines, for example ATP synthase (eight different proteins in mammals) or cytochrome oxidase (13 proteins in mammals). These proteins included in complexes are called "subunits". Other protein assemblies are only built to carry out transient actions, for example, the activation of gene expression by the binding of transcription factors and activators on the DNA promoter region of a gene.

Another essential element for defining PPIs is the biological context. Not all possible interactions will occur in any cell at any time. Instead, interactions depend on cell type, cell cycle phase and state, developmental stage, environmental conditions, protein modifications (e.g., phosphorylation), presence of cofactors, and presence of other binding partners.

**Protein microarrays.**

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

## Types of arrays

There are three types of protein microarrays that are currently used to study the biochemical activities of proteins.

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

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

## Peptidomics

Peptidomics, the comprehensive qualitative and quantitative analysis of all peptides in a biological sample, is an emerging field derived from proteomics and enabled by modern separation, analytical and computation technologies. The complex biological matrices typically examined in peptidomics experiments require systematic peptide extraction to achieve successful analysis. Peptidomic analysis employs many proteomics techniques but with a different target. Rather than examining a sample for which intact proteins are present, peptidomics examines which endogenous protein fragments are present. This review describes applications of peptidomics and modern approaches for peptide extraction, fractionation, detection, quantification, functional annotation, and structural prediction.

## Applications of peptidomics

### Peptidomics for biomarker search

Perhaps the most frequent use of peptidomics thus far has been in search of biomarkers of disease. Peptidomics is appealing for biomarker studies because the knowledge that is generated can present a dynamic view of health status: peptides are created by a complex and fluid interaction of proteases, activators, inhibitors, and protein substrates. A variety of peptide biomarkers have been identified. For example, levels of a fragment of ——————————————————————————————— -amyloid ( ——————————————————————————————— -amyloid 1–42) and tau protein in cerebrospinal fluid can predict which patients with mild cognitive

impairment will progress to Alzheimer's disease. Combinations of urine peptides have been shown to serve as biomarkers (reviewed in depth in) for diabetic nephropathy, chronic kidney disease, acute kidney injury, acute renal allograft rejection, prostate cancer, and coronary artery disease. The application of peptidomic analysis to identify biomarkers of disease has been thoroughly reviewed in a number of articles.

Many biological systems (including blood and digestive samples) contain proteases or contain organisms (e.g., bacteria) that can produce proteases. In order to use peptides for biomarkers, postsample collection proteolysis should be eliminated (protease inhibition) or adequately accounted for in these sample types, as discussed by Diamandis. However, some biological samples, such as urine, are more stable and thus do not require additional treatment to prevent proteolysis. Therefore, these samples can be used without protease inhibitory treatment.

Peptidomics is an expanding new field with a variety of applications including monitoring digestion, annotating food hydrolysates, characterizing hormone levels and identifying disease biomarkers. Innovations in peptide extraction, detection, and analysis are improving peptidomics throughput, accuracy, and utility.

Peptidomics will continue to advance with faster instrument electronics to facilitate the isolation, fragmentation, and detection of more peptides in less time, as well as more sensitive detectors that will allow the detection of less abundant peptides and, after fragmentation, the detection of less abundant fragment ions. sets. However, methods for accurate analytical structure determination (e.g., x-ray crystallography, NMR, and circular dichroism) are typically applied to isolated peptides and have not yet been applied to measure complex peptide mixtures in high-throughput. Strategies for structural determination in-line with LC-MS would be promising.

Traditionally, intact proteins have been considered as the functional units in vivo. However, most proteins undergo proteolytic processing such as auto-activation or degradation by enzymes [166]. Therefore, often, protein fragments are produced that can interact directly with the cellular targets, producing a functional effect. Even if a peptide occurs for only a short time before further degradation, it may still transmit a signal. Therefore, mapping even transitory peptides can be important for understanding complex protein/peptide–health interactions. With the advance of peptidomics, we can now monitor peptide release across time and physiological/sub-cellular location to reveal their roles in complex biological interaction networks.

With the recent realization that peptides can have important functions—both beneficial and detrimentalthroughout the body, peptidomics will become increasingly important in monitoring how dietary proteins are digested. Food producers will soon need to employ peptidomics to characterize not only what peptides and proteins are in their food products, but also what they become in the digestive tract of the consumer.

There is a variety of issues in peptidomic analysis that still need to be addressed. Improved software for identification of peptides with complex modifications is necessary. For example, software that adequately identifies endogenous glycopeptides from complex biological mixtures remains lacking.

Software for peptide functional prediction remains in its infancy, yet will become increasingly important with the large peptide data sets now being produced. Several programs now provide estimates of peptide structure based on sequence, and these can be applied to large peptidomics data

**References**


1.https://www.bio-rad.com/en-in/applications-technologies/introduction-affinity-chromatography?ID=MWHAVG4VY

2. https://bitesizebio.com/580/how-sds-page-works/

3. https://www.creative-proteomics.com/blog/index.php/protein-sequencing-of-edman-degradation

4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2891586/

5.https://lebrilla.faculty.ucdavis.edu/wp-content/uploads/sites/301/2014/01/Current-peptidomics-Applications-purification-identification-quantification-and-functional-analysis.pdf

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

**SCHOOL OF BIO AND CHEMICAL ENGINEERING**
**DEPARTMENT OF BIOINFORMATICS**

# UNIT – III - Omics in Biology – SBIA5204

## UNIT III
## Glycomics and Lipidomics

**Glycomics – Challenges, Importance, Tools used to analyse glycans, softwares and databases. Lipidomics – Structural diversity of lipids, extraction, separation, detection and imaging. Challenges and applications.**

**Glycomics** is the comprehensive study of glycomes (the entire complement of sugars, whether free or present in more complex molecules of an organism), including genetic, physiologic, pathologic, and other aspects. Glycomics "is the systematic study of all glycan structures of a given cell type or organism" and is a subset of glycobiology. The term glycomics is derived from the chemical prefix for sweetness or a sugar, "glyco-", and was formed to follow the *omics* naming convention established by genomics (which deals with genes) and proteomics (which deals with proteins).

### Challenges

- The complexity of sugars: regarding their structures, they are not linear instead they are highly branched. Moreover, glycans can be modified (modified sugars), this increases its complexity.
- Complex biosynthetic pathways for glycans.
- Usually glycans are found either bound to protein (glycoprotein) or conjugated with lipids (glycolipids).
- Unlike genomes, glycans are highly dynamic.

This area of research has to deal with an inherent level of complexity not seen in other areas of applied biology. 68 building blocks (molecules for DNA, RNA and proteins; categories for lipids; types of sugar linkages for saccharides) provide the structural basis for the molecular choreography that constitutes the entire life of a cell. DNA and RNA have four building blocks each (the nucleosides or nucleotides). Lipids are divided into eight categories based on ketoacyl and isoprene. Proteins have 20 (the amino acids). Saccharides have 32 types of sugar linkages. While these building blocks can be attached only linearly for proteins and genes, they can be arranged in a branched array for saccharides, further increasing the degree of complexity.

Add to this the complexity of the numerous proteins involved, not only as carriers of carbohydrate, the glycoproteins, but proteins specifically involved in binding and reacting with carbohydrate:

- Carbohydrate-specific enzymes for synthesis, modulation, and degradation
- Lectins, carbohydrate-binding proteins of all sorts
- Receptors, circulating or membrane-bound carbohydrate-binding receptors

### Importance

To answer this question one should know the different and important functions of glycans. The following are some of those functions:

- Glycoproteins and Glycolipids found on the cell surface play a critical role in bacterial and viral recognition.
- They are involved in cellular signaling pathways and modulate cell function.
- They are important in innate immunity.
- They determine cancer development.
- They orchestrate the cellular fate, inhibit proliferation, regulate circulation and invasion.
- They affect the stability and folding of proteins.
- They affect the pathway and fate of glycoproteins.
- There are many glycan-specific diseases, often hereditary diseases.

There are important medical applications of aspects of glycomics:

- Lectins fractionate cells to avoid graft-versus-host disease in hematopoietic stem cell transplantation.
- Activation and expansion of cytolytic CD8 T cells in cancer treatment.

Glycomics is particularly important in microbiology because glycans play diverse roles in bacterial physiology. Research in bacterial glycomics could lead to the development of:

- novel drugs
- bioactive glycans
- glycoconjugate vaccines

**Tools used**

The following are examples of the commonly used techniques in glycan analysis

**High-resolution mass spectrometry (MS) and high-performance liquid chromatography (HPLC)**

The most commonly applied methods are MS and HPLC, in which the glycan part is cleaved either enzymatically or chemically from the target and subjected to analysis. In case of glycolipids, they can be analyzed directly without separation of the lipid component.

N-glycans from glycoproteins are analyzed routinely by high-performance-liquid-chromatography (reversed phase, normal phase and ion exchange HPLC) after tagging the reducing end of the sugars with a fluorescent compound (reductive labeling). A large variety of different labels were introduced in the recent years, where 2-aminobenzamide (AB), anthranilic acid (AA), 2-aminopyridin (PA), 2-aminoacridone (AMAC) and 3-(acetylamino)-6-aminoacridine (AA-Ac) are just a few of them.

O-glycans are usually analysed without any tags, due to the chemical release conditions preventing them to be labeled.

Fractionated glycans from high-performance liquid chromatography (HPLC) instruments can be further analyzed by MALDI-TOF-MS(MS) to get further information about structure and purity. Sometimes glycan pools are analyzed directly by mass spectrometry without prefractionation, although a discrimination between isobaric glycan structures is more challenging or even not always possible. Anyway, direct MALDI-TOF-MS analysis can lead to a fast and straightforward illustration of the glycan pool.

In recent years, high performance liquid chromatography online coupled to mass spectrometry became very popular. By choosing porous graphitic carbon as a stationary phase for liquid chromatography, even non derivatized glycans can be analyzed. Electrospray ionisation (ESI) is frequently used for this application.

**Multiple Reaction Monitoring (MRM)**

Although MRM has been used extensively in metabolomics and proteomics, its high sensitivity and linear response over a wide dynamic range make it especially suited for glycan biomarker research and discovery. MRM is performed on a triple quadrupole (QqQ) instrument, which is set to detect a predetermined precursor ion in the first quadrupole, a fragmented in the collision quadrupole, and a predetermined fragment ion in the third quadrupole. It is a non-scanning technique, wherein each transition is detected individually and the detection of multiple transitions occurs concurrently in duty cycles. This technique is being used to characterize the immune glycome.

**Arrays**

Lectin and antibody arrays provide high-throughput screening of many samples containing glycans. This method uses either naturally occurring lectins or artificial monoclonal antibodies, where both are immobilized on a certain chip and incubated with a fluorescent glycoprotein sample.

Glycan arrays, like that offered by the Consortium for Functional Glycomics and Z Biotech LLC, contain carbohydrate compounds that can be screened with lectins or antibodies to define carbohydrate specificity and identify ligands.

**Metabolic and covalent labeling of glycans**

Metabolic labeling of glycans can be used as a way to detect glycan structures. A well known strategy involves the use of azide-labeled sugars which can be reacted using the Staudinger ligation. This method has been used for in vitro and in vivo imaging of glycans.

**Tools for glycoproteins**

X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy for complete structural analysis of complex glycans is a difficult and complex field. However, the structure of the binding site of numerous lectins, enzymes and other carbohydrate-binding proteins has revealed a wide variety of the structural basis for glycome function. The purity of test samples have been obtained through chromatography (affinity chromatography etc.) and analytical electrophoresis (PAGE (polyacrylamide electrophoresis), capillary electrophoresis, affinity electrophoresis, etc.).

**Lipidomics**

**Lipidomics** is the large-scale study of pathways and networks of cellular lipids in biological systems The word "lipidome" is used to describe the complete lipid profile within a cell, tissue, organism, or ecosystem and is a subset of the "metabolome" which also includes the three other major classes of biological molecules: proteins/amino-acids, sugars and nucleic acids. Lipidomics is a relatively recent research field that has been driven by rapid advances in technologies such as mass spectrometry (MS), nuclear magnetic resonance (NMR)

spectroscopy, fluorescence spectroscopy, dual polarisation interferometry and computational methods, coupled with the recognition of the role of lipids in many metabolic diseases such as obesity, atherosclerosis, stroke, hypertension and diabetes. This rapidly expanding field complements the huge progress made in genomics and proteomics, all of which constitute the family of systems biology.

Lipidomics research involves the identification and quantification of the thousands of cellular lipid molecular species and their interactions with other lipids, proteins, and other metabolites. Investigators in lipidomics examine the structures, functions, interactions, and dynamics of cellular lipids and the changes that occur during perturbation of the system.

Han and Gross first defined the field of lipidomics through integrating the specific chemical properties inherent in lipid molecular species with a comprehensive mass spectrometric approach. Although lipidomics is under the umbrella of the more general field of "metabolomics", lipidomics is itself a distinct discipline due to the uniqueness and functional specificity of lipids relative to other metabolites.

In lipidomic research, a vast amount of information quantitatively describing the spatial and temporal alterations in the content and composition of different lipid molecular species is accrued after perturbation of a cell through changes in its physiological or pathological state. Information obtained from these studies facilitates mechanistic insights into changes in cellular function. Therefore, lipidomic studies play an essential role in defining the biochemical mechanisms of lipid-related disease processes through identifying alterations in cellular lipid metabolism, trafficking and homeostasis. The growing attention on lipid research is also seen from the initiatives underway of the LIPID Metabolites And Pathways Strategy (LIPID MAPS Consortium). and The European Lipidomics Initiative (ELIfe).

## Structural diversity of lipids

Lipids are a diverse and ubiquitous group of compounds which have many key biological functions, such as acting as structural components of cell membranes, serving as energy storage sources and participating in signaling pathways. Lipids may be broadly defined as hydrophobic or amphipathic small molecules that originate entirely or in part from two distinct types of biochemical subunits or "building blocks": ketoacyl and isoprene groups. The huge structural diversity found in lipids arises from the biosynthesis of various combinations of these building blocks. For example, glycerophospholipids are composed of a glycerol backbone linked to one of approximately 10 possible headgroups and also to 2 fatty acyl/alkyl chains, which in turn may have 30 or more different molecular structures. In practice, not all possible permutations are detected experimentally, due to chain preferences depending on the cell type and also to detection limits - nevertheless several hundred distinct glycerophospholipid molecular species have been detected in mammalian cells.

Plant **chloroplast thylakoid** membranes however, have *unique lipid composition* as they are deficient in phospholipids. Also, their largest constituent, *monogalactosyl diglyceride or MGDG*, does not form aqueous bilayers. Nevertheless, dynamic studies reveal a normal lipid bilayer organisation in thylakoid membranes.

## Lipid profiling

Lipid profiling is a targeted metabolomics platform that provides a comprehensive analysis of lipid species within a cell or tissue. Profiling based on electrospray ionization tandem mass spectrometry (ESI-MS/MS) is capable of providing quantitative data and is adaptable to high

throughput analyses. The powerful approach of transgenics, namely deletion and/or overexpression of a gene product coupled with lipidomics, can give valuable insights into the role of biochemical pathways. Lipid profiling techniques have also been applied to plants and microorganisms such as yeast. A combination of quantitative lipidomic data in conjunction with the corresponding transcriptional data (using gene-array methods) and proteomic data (using tandem MS) enables a systems biology approach to a more in-depth understanding of the metabolic or signaling pathways of interest.

**Informatics**

A major challenge for lipidomics, in particular for MS-based approaches, lies in the computational and bioinformatic demands of handling the large amount of data that arise at various stages along the chain of information acquisition and processing. Chromatographic and MS data collection requires substantial efforts in spectral alignment and statistical evaluation of fluctuations in signal intensities. Such variations have a multitude of origins, including biological variations, sample handling and analytical accuracy. As a consequence several replicates are normally required for reliable determination of lipid levels in complex mixtures. Within the last few years, a number of software packages have been developed by various companies and research groups to analyze data generated by MS profiling of metabolites, including lipids. The data processing for differential profiling usually proceed through several stages, including input file manipulation, spectral filtering, peak detection, chromatographic alignment, normalization, visualization, and data export. An example of metabolic profiling software is the freely-available Java-based Mzmine application. Recently MS-DIAL 4 software was integrated with a comprehensive lipidome atlas with retention time, collision cross-section and tandem mass spectrometry information for 117 lipid subclasses and 8,051 lipids. Some software packages such as Markerview include multivariate statistical analysis (for example, principal component analysis) and these will be helpful for the identification of correlations in lipid metabolites that are associated with a physiological phenotype, in particular for the development of lipid-based biomarkers. Another objective of the information technology side of lipidomics involves the construction of metabolic maps from data on lipid structures and lipid-related protein and genes. Some of these lipid pathways are extremely complex, for example the mammalian glycosphingolipid pathway. The establishment of searchable and interactive databases of lipids and lipid-related genes/proteins is also an extremely important resource as a reference for the lipidomics community. Integration of these databases with MS and other experimental data, as well as with metabolic networks offers an opportunity to devise therapeutic strategies to prevent or reverse these pathological states involving dysfunction of lipid-related processes.

References

1 https://en.wikipedia.org/wiki/Glycomics#:~:text=Glycomics%20is%20the%20comprehensive%20study,%2C%20pathologic%2C%20and%20other%20aspects.
2.https://en.wikipedia.org/wiki/Lipidomics#:~:text=Lipidomics%20is%20the%20large%2Dscale,other%20major%20classes%20of%20biological

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

**SCHOOL OF BIO AND CHEMICAL ENGINEERING**
**DEPARTMENT OF BIOINFORMATICS**

# UNIT – IV - Omics in Biology – SBIA5204

## UNIT IV TRANSCRIPTOMICS

Gene Expression Profiling – DNA microarrays. Transcriptomics: Data Collection- Isolation of RNA, ESTs, SAGE analysis, Microarrays. RNA-seq: Principle and advances. Image processing

The **transcriptome** is the set of all RNA transcripts, including coding and non-coding, in an individual or a population of cells. The term can also sometimes be used to refer to all RNAs, or just mRNA, depending on the particular experiment. The term *transcriptome* is a portmanteau of the words *transcript* and *genome*; it is associated with the process of transcript production during the biological process of transcription.

The early stages of transcriptome annotations began with cDNA libraries published in the 1980s. Subsequently, the advent of high-throughput technology led to faster and more efficient ways of obtaining data about the transcriptome. Two biological techniques are used to study the transcriptome, namely DNA microarray, a hybridization-based technique and RNA-seq, a sequence-based approach. RNA-seq is the preferred method and has been the dominant transcriptomics technique since the 2010s. Single-cell transcriptomics allows tracking of transcript changes over time within individual cells.

Data obtained from the transcriptome is used in research to gain insight into processes such as cellular differentiation, carcinogenesis, transcription regulation and biomarker discovery among others. Transcriptome-obtained data also finds applications in establishing phylogenetic relationships during the process of evolution and in *in vitro* fertilization. The transcriptome is closely related to other -ome based biological fields of study; it is complementary to the proteome and the metabolome and encompasses the translatome, exome, meiome and thanatotranscriptome which can be seen as ome fields studying specific types of RNA transcripts. There are numerous publicly available transcriptome databases.

### Etymology and history

The word *transcriptome* is a portmanteau of the words *transcript* and *genome*. It appeared along other neologisms formed using the suffixes *-ome* and *-omics* to denote all studies conducted on a genome-wide scale in the fields of life sciences and technology. As such, transcriptome and transcriptomics were one of the first words to emerge along with genome and proteome. The first study to present a case of a collection of a cDNA library for silk moth mRNA was published in 1979. The first seminal study to mention and investigate the transcriptome of an organism was published in 1997 and it described 60,633 transcripts expressed in *S. cerevisiae* using serial analysis of gene expression (SAGE). With the rise of high-throughput technologies and bioinformatics and the subsequent increased computational power, it became increasingly efficient and easy to characterize and analyze enormous amount of data. Attempts to characterize the transcriptome became more prominent with the advent of automated DNA sequencing during the 1980s. During the 1990s, expressed sequence tag sequencing was used to identify genes and their fragments. This was followed by techniques such as serial analysis of gene expression (SAGE), cap analysis of gene expression (CAGE), and massively parallel signature sequencing (MPSS).

## Transcription

The transcriptome encompasses all the ribonucleic acid (RNA) transcripts present in a given organism or experimental sample. RNA is the main carrier of genetic information that is responsible for the process of converting DNA into an organism's phenotype. A gene can give rise to a single-stranded messenger RNA (mRNA) through a molecular process known as transcription; this mRNA is complementary to the strand of DNA it originated from. The enzyme RNA polymerase II attaches to the template DNA strand and catalyzes the addition of ribonucleotides to the 3' end of the growing sequence of the mRNA transcript.

In order to initiate its function, RNA polymerase II needs to recognize a promoter sequence, located upstream (5') of the gene. In eukaryotes, this process is mediated by transcription factors, most notably Transcription factor II D (TFIID) which recognizes the TATA box and aids in the positioning of RNA polymerase at the appropriate start site. To finish the production of the RNA transcript, termination takes place usually several hundred nuclecotides away from the termination sequence and cleavage takes place. This process occurs in the nucleus of a cell along with RNA processing by which mRNA molecules are capped, spliced and polyadenylated to increase their stability before being subsequently taken to the cytoplasm. The mRNA gives rise to proteins through the process of translation that takes place in ribosomes.

## Types of RNA transcripts

In accordance with the central dogma of molecular biology, the transcriptome initially encompassed only protein-coding mRNA transcripts. Nevertheless, several RNA subtypes with distinct functions exist. Many RNA transcripts do not code for protein or have different regulatory functions in the process of gene transcription and translation. RNA types which do not fall within the scope of the central dogma of molecular biology are non-coding RNAs which can be divided into two groups of long non-coding RNA and short non-coding RNA.

Long non-coding RNA includes all non-coding RNA transcripts that are more than 200 nucleotides long. Members of this group comprise the largest fraction of the non-coding transcriptome. Short non-coding RNA includes the following members:

- transfer RNA (tRNA)
- micro RNA (miRNA): 19-24 nucleotides (nt) long. Micro RNAs up- or downregulate expression levels of mRNAs by the process of RNA interference at the post-transcriptional level.
- small interfering RNA (siRNA): 20-24 nt
- small nucleolar RNA (snoRNA)
- Piwi-interacting RNA (piRNA): 24-31 nt. They interact with Piwi proteins of the Argonaute family and have a function in targeting and cleaving transposons.
- enhancer RNA (eRNA)

## Scope of study

In the human genome, about 5% of all genes get transcribed into RNA. The transcriptome consists of coding mRNA which comprise around 1-4% of its entirety and non-coding RNAs which comprise the rest of the genome and do not give rise to proteins. The number of non-protein-coding sequences increases in more complex organisms.

Several factors render the content of the transcriptome difficult to establish. These include alternative splicing, RNA editing and alternative transcription among

others. Additionally, transcriptome techniques are capable of capturing transcription occurring in a sample at a specific time point, although the content of the transcriptome can change during differentiation. The main aims of transcriptomics are the following: "catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions".

The term can be applied to the total set of transcripts in a given organism, or to the specific subset of transcripts present in a particular cell type. Unlike the genome, which is roughly fixed for a given cell line (excluding mutations), the transcriptome can vary with external environmental conditions. Because it includes all mRNA transcripts in the cell, the transcriptome reflects the genes that are being actively expressed at any given time, with the exception of mRNA degradation phenomena such as transcriptional attenuation. The study of transcriptomics, (which includes expression profiling, splice variant analysis etc), examines the expression level of RNAs in a given cell population, often focusing on mRNA, but sometimes including others such as tRNAs and sRNAs.

**Methods of construction**

Transcriptomics is the quantitative science that encompasses the assignment of a list of strings ("reads") to the object ("transcripts" in the genome). To calculate the expression strength, the density of reads corresponding to each object is counted. Initially, transcriptomes were analyzed and studied using expressed sequence tags libraries and serial and cap analysis of gene expression (SAGE).

Currently, the two main transcriptomics techniques include DNA microarrays and RNA-Seq. Both techniques require RNA isolation through RNA extraction techniques, followed by its separation from other cellular components and enrichment of mRNA.

There are two general methods of inferring transcriptome sequences. One approach maps sequence reads onto a reference genome, either of the organism itself (whose transcriptome is being studied) or of a closely related species. The other approach, *de novo* transcriptome assembly, uses software to infer transcripts directly from short sequence reads and is used in organisms with genomes that are not sequenced.

DNA microarrays

DNA microarray used to detect gene expression in human (*left*) and mouse (*right*) samples

The first transcriptome studies were based on microarray techniques (also known as DNA chips). Microarrays consist of thin glass layers with spots on which oligonucleotides, known as "probes" are arrayed; each spot contains a known DNA sequence.

When performing microarray analyses, mRNA is collected from a control and an experimental sample, the latter usually representative of a disease. The RNA of interest is converted to cDNA to increase its stability and marked with fluorophores of two colors, usually green and red, for the two groups. The cDNA is spread onto the surface of the microarray where it hybridizes with oligonucleotides on the chip and a laser is used to scan. The fluorescence intensity on each spot of the microarray corresponds to the level of gene expression and based on the color of the fluorophores selected, it can be determined which of the samples exhibits higher levels of the mRNA of interest.

One microarray usually contains enough oligonucleotides to represent all known genes; however, data obtained using microarrays does not provide information about unknown genes. During the 2010s, microarrays were almost completely replaced by next-generation techniques that are based on DNA sequencing.

RNA sequencing

RNA sequencing is a next-generation sequencing technology; as such it requires only a small amount of RNA and no previous knowledge of the genome. It allows for both qualitative and quantitative analysis of RNA transcripts, the former allowing discovery of new transcripts and the latter a measure of relative quantities for transcripts in a sample.

The three main steps of sequencing transcriptomes of any biological samples include RNA purification, the synthesis of an RNA or cDNA library and sequencing the library. The RNA purification process is different for short and long RNAs. This step is usually followed by an assessment of RNA quality, with the purpose of avoiding contaminants such as DNA or technical contaminants related to sample processing. RNA quality is measured using UV spectrometry with an absorbance peak of 260 nm. RNA integrity can also be analyzed quantitatively comparing the ratio and intensity of 28S RNA to 18S RNA reported in the RNA Integrity Number (RIN) score. Since mRNA is the species of interest and it represents only 3% of its total content, the RNA sample should be treated to remove rRNA and tRNA and tissue-specific RNA transcripts.

The step of library preparation with the aim of producing short cDNA fragments, begins with RNA fragmentation to transcripts in length between 50 and 300 base pairs. Fragmentation can be enzymatic (RNA endonucleases), chemical (trismagnesium salt buffer, chemical hydrolysis) or mechanical (sonication, nebulisation). Reverse transcription is used to convert the RNA templates into cDNA and three priming methods can be used to achieve it, including oligo-DT, using random primers or ligating special adaptor oligos.

Single-cell transcriptomics

Transcription can also be studied at the level of individual cells by single-cell transcriptomics. Single-cell RNA sequencing (scRNA-seq) is a recently developed technique that allows the analysis of the transcriptome of single cells. With single-cell transcriptomics, subpopulations of cell types that constitute the tissue of interest are also taken into consideration. This approach allows to identify whether changes in experimental samples are due to phenotypic cellular

changes as opposed to proliferation, with which a specific cell type might be overexpressed in the sample. Additionally, when assessing cellular progression through differentiation, average expression profiles are only able to order cells by time rather than their stage of development and are consequently unable to show trends in gene expression levels specific to certain stages. Single-cell trarnscriptomic techniques have been used to characterize rare cell populations such as circulating tumor cells, cancer stem cells in solid tumors, and embryonic stem cells (ESCs) in mammalian blastocysts.

Although there are no standardized techniques for single-cell transcriptomics, several steps need to be undertaken. The first step includes cell isolation, which can be performed using low- and high-throughput techniques. This is followed by a qPCR step and then single-cell RNAseq where the RNA of interest is converted into cDNA. Newer developments in single-cell transcriptomics allow for tissue and sub-cellular localization preservation through cryo-sectioning thin slices of tissues and sequencing the transcriptome in each slice. Another technique allows the visualization of single transcripts under a microscope while preserving the spatial information of each individual cell where they are expressed.

## Analysis

A number of organism-specific transcriptome databases have been constructed and annotated to aid in the identification of genes that are differentially expressed in distinct cell populations.

RNA-seq is emerging (2013) as the method of choice for measuring transcriptomes of organisms, though the older technique of DNA microarrays is still used. RNA-seq measures the transcription of a specific gene by converting long RNAs into a library of cDNA fragments. The cDNA fragments are then sequenced using high-throughput sequencing technology and aligned to a reference genome or transcriptome which is then used to create an expression profile of the genes.

## Applications

Mammals

The transcriptomes of stem cells and cancer cells are of particular interest to researchers who seek to understand the processes of cellular differentiation and carcinogenesis. A pipeline using RNA-seq or gene array data can be used to track genetic changes occurring in stem and precursor cells and requires at least three independent gene expression data from the former cell type and mature cells.

Analysis of the transcriptomes of human oocytes and embryos is used to understand the molecular mechanisms and signaling pathways controlling early embryonic development, and could theoretically be a powerful tool in making proper embryo selection in in vitro fertilisation. Analyses of the transcriptome content of the placenta in the first-trimester of pregnancy in *in vitro* fertilization and embryo transfer (IVT-ET) revealed differences in genetic expression which are associated with higher frequency of adverse perinatal outcomes. Such insight can be used to optimize the practice. Transcriptome analyses can also be used to optimize cryopreservation of oocytes, by lowering injuries associated with the process.

Transcriptomics is an emerging and continually growing field in biomarker discovery for use in assessing the safety of drugs or chemical risk assessment.

Transcriptomes may also be used to infer phylogenetic relationships among individuals or to detect evolutionary patterns of transcriptome conservation.

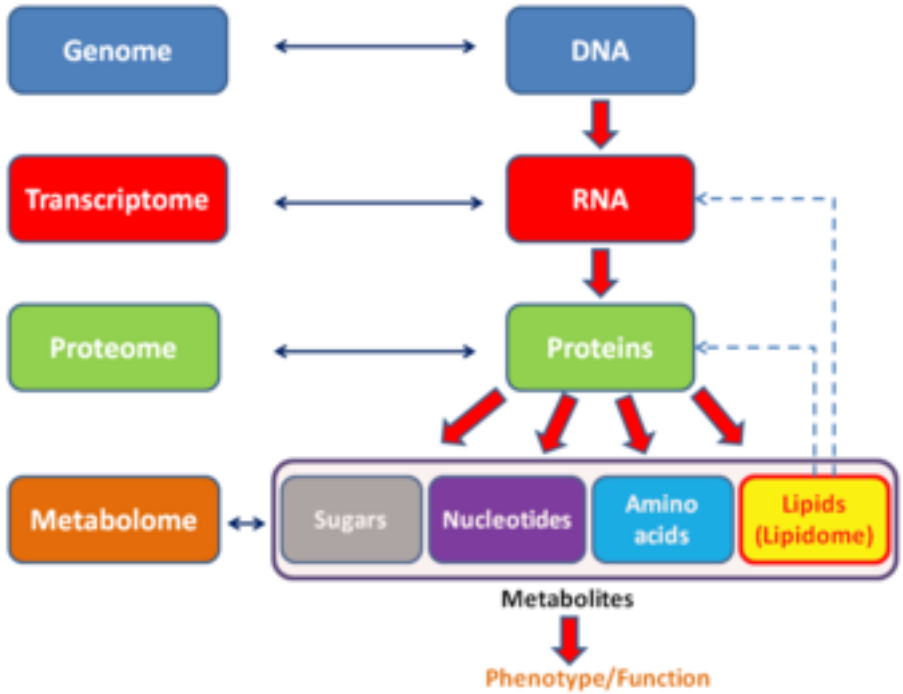
Transcriptome analyses were used to discover the incidence of antisense transcription, their role in gene expression through interaction with surrounding genes and their abundance in

different chromosomes. RNA-seq was also used to show how RNA isoforms, transcripts stemming from the same gene but with different structures, can produce complex phenotypes from limited genomes.

Plants

Transcriptome analysis have been used to study the evolution and diversification process of plant species. In 2014, the 1000 Plant Genomes Project was completed in which the transcriptomes of 1,124 plant species from the families viridiplantae, glaucophyta and rhodophyta were sequenced. The protein coding sequences were subsequently compared to infer phylogenetic relationships between plants and to characterize the time of their diversification in the process of evolution. Transcriptome studies have been used to characterize and quantify gene expression in mature pollen. Genes involved in cell wall metabolism and cytoskeleton were found to be overexpressed. Transcriptome approaches also allowed to track changes in gene expression through different developmental stages of pollen, ranging from microspore to mature pollen grains; additionally such stages could be compared across species of different plants including *Arabidopsis*, rice and tobacco.

**Relation to other ome fields**




General schema showing the relationships of the genome, transcriptome, proteome, and metabolome (lipidome).

Similar to other -ome based technologies, analysis of the transcriptome allows for an unbiased approach when validating hypotheses experimentally. This approach also allows for the discovery of novel mediators in signaling pathways. As with other -omics based technologies, the transcriptome can be analyzed within the scope of a multiomics approach. It is complementary to metabolomics but contrary to proteomics, a direct association between a transcript and metabolite cannot be established.

There are several -ome fields that can be seen as subcategories of the transcriptome. The exome differs from the transcriptome in that it includes only those RNA molecules found in a specified cell population, and usually includes the amount or concentration of each RNA molecule in addition to the molecular identities. Additionally, the transcritpome also differs from the translatome, which is the set of RNAs undergoing translation.

The term meiome is used in functional genomics to describe the meiotic transcriptome or the set of RNA transcripts produced during the process of meiosis. Meiosis is a key feature of sexually reproducing eukaryotes, and involves the pairing of homologous chromosome, synapse and recombination. Since meiosis in most organisms occurs in a short time period, meiotic transcript profiling is difficult due to the challenge of isolation (or enrichment) of meiotic cells (meiocytes). As with transcriptome analyses, the meiome can be studied at a whole-genome level using large-scale transcriptomic techniques. The meiome has been well-characterized in mammal and yeast systems and somewhat less extensively characterized in plants.

The thanatotranscriptome consists of all RNA transcripts that continue to be expressed or that start getting re-expressed in internal organs of a dead body 24–48 hours following death. Some genes include those that are inhibited after fetal development. If the thanatotranscriptome is related to the process of programmed cell death (apoptosis), it can be referred to as the apoptotic thanatotranscriptome. Analyses of the thanatotranscriptome are used in forensic medicine.

eQTL mapping can be used to complement genomics with transcriptomics; genetic variants at DNA level and gene expression measures at RNA level.

Relation to proteome

The transcriptome can be seen as a subset of the proteome, that is, the entire set of proteins expressed by a genome.

However, the analysis of relative mRNA expression levels can be complicated by the fact that relatively small changes in mRNA expression can produce large changes in the total amount of the corresponding protein present in the cell. One analysis method, known as gene set enrichment analysis, identifies coregulated gene networks rather than individual genes that are up- or down-regulated in different cell populations.

Although microarray studies can reveal the relative amounts of different mRNAs in the cell, levels of mRNA are not directly proportional to the expression level of the proteins they code for. The number of protein molecules synthesized using a given mRNA molecule as a template is highly dependent on translation-initiation features of the mRNA sequence; in particular, the ability of the translation initiation sequence is a key determinant in the recruiting of ribosomes for protein translation.

References:
1. https://en.wikipedia.org/wiki/Transcriptome
2. https://en.wikipedia.org/wiki/Transcriptomics_technologies#Transcriptome_databases

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

**SCHOOL OF BIO AND CHEMICAL ENGINEERING**
**DEPARTMENT OF BIOINFORMATICS**

# UNIT – V - Omics in Biology – SBIA5204

# UNIT V OTHER OMICS IN BIOLOGY

Secretomics, Metablolomics, fluxomics, nutrigenomics, Metagenomics, Organomics, Pharmacogenomics, Phytochemomics, Microbiomics

## Secretomics

**Secretomics** is a type of proteomics which involves the analysis of the secretome—all the secreted proteins of a cell, tissue or organism. Secreted proteins are involved in a variety of physiological processes, including cell signaling and matrix remodeling, but are also integral to invasion and metastasis of malignant cells. Secretomics has thus been especially important in the discovery of biomarkers for cancer and understanding molecular basis of pathogenesis. The analysis of the insoluble fraction of the secretome (the extracellular matrix) has been termed matrisomics.

### History of the secretome

In 2000 Tjalsma et al. coined the term 'secretome' in their study of the eubacterium *B. subtilis*. They defined the secretome as all of the secreted proteins and secretory machinery of the bacteria. Using a database of protein sequences in *B. subtilis* and an algorithm that looked at cleavage sites and amino-terminal signal peptides characteristic of secreted proteins they were able to predict what fraction of the proteome is secreted by the cell. In 2001 the same lab set a standard of secretomics – predictions based on amino acid sequence alone are not enough to define the secretome. They used two-dimensional gel electrophoresis and mass spectrometry to identify 82 proteins secreted by *B. subtilis*, only 48 of which had been predicted using the genome-based method of their previous paper. This demonstrates the need for protein verification of predicted findings.

As the complicated nature of secretory pathways was revealed – namely that there are many non-classical pathways of secretion and there are many non-secreted proteins that are a part of the classical secretory pathway – a more in-depth definition of the secretome became necessary. In 2010, Agrawal et al. suggested defining the secretome as "the global group of secreted proteins into the extracellular space by a cell, tissue, organ, or organism at any given time and conditions through known and unknown secretory mechanisms involving constitutive and regulated secretory organelles".

### Methods

### Genome-wide prediction

Many secreted proteins have an N-terminal peptide sequence that signals for the translated protein to move into the endoplasmic reticulum where the processing occurs that will ultimately lead to secretion. The presence of these signal peptides can be used to predict the secretome of a cell. Software such as SignalP can identify signal sequences (and their cleavage sites) to predict proteins that are secreted. Since transmembrane proteins are also processed in the ER, but not secreted, software like the TMHMM server is used to predict

transmembrane domains and therefore eliminate false positives. Some secretory proteins do not have classical signal peptide sequences. These 'leaderless secretory proteins' (LSPs) will be missed by SignalP. SecretomeP is a software that has been developed to try to predict these non-classical secretory proteins from their sequences. Genome-wide secretomes have been predicted for a wide range of organisms, including human, mouse, zebrafish, and hundreds of bacteria.

Genome-wide prediction methods have a variety of problems. There is a high possibility of false positives and false negatives. In addition, gene expression is heavily influenced by environmental conditions, meaning a secretome predicted from the genome or a cDNA library is not likely to match completely with the true secretome. Proteomic approaches are necessary to validate any predicted secreted proteins.

Several genome-wide secretome databases or knowledgebases are available based on both curation and computational prediction. These databases include the fungal secretome database (FSD), the fungal secretome knowledgebase (FunSecKB), and the lactic acid bacterial secretome database. The human and animal protein subcellular location database (MetaSecKB) and the protist subcellular proteome database (ProtSecKB) are also recently released. Though there are some inaccuracies in the computational prediction, these databases provide useful resources for further characterizing the protein subcellular locations.

**Proteomic approaches**[edit]

Mass spectrometry analysis is integral to secretomics. Serum or supernatant containing secreted proteins is digested with a protease and the proteins are separated by 2D gel electrophoresis or chromatographic methods. Each individual protein is then analyzed by mass spectrometry and the peptide-mass fingerprint generated can be run through a database to identify the protein.

Stable isotope labeling by amino acids in cell culture (SILAC) has emerged as an important method in secretomics – it helps to distinguish between secreted proteins and bovine serum contaminants in cell culture. Supernatant from cells grown in normal medium and cells grown in medium with stable-isotope labeled amino acids is mixed in a 1:1 ratio and analyzed by mass spectrometry. Protein contaminants in the serum will only show one peak because they do not have a labeled equivalent. As an example, the SILAC method has been used successfully to distinguish between proteins secreted by human chondrocytes in culture and serum contaminants.

An antibody microarray is a highly sensitive and high-throughput method for protein detection that has recently become part of secretomic analysis. Antibodies, or another type of binder molecule, are fixed onto a solid support and a fluorescently labeled protein mixture is added. Signal intensities are used to identify proteins. Antibody microarrays are extremely versatile – they can be used to analyze the amount of protein in a mixture, different protein isoforms, posttranslational modifications, and the biochemical activity of proteins. In addition, these microarrays are highly sensitive – they can detect single molecules of protein. Antibody microarrays are currently being used mostly to analyze human plasma samples but can also be used for cultured cells and body fluid secretomics, presenting a simple way to look for the presence of many proteins at one time.

**Implications and significance**

**Discovery of cancer biomarkers**[edit]

Besides being important in normal physiological processes, secreted proteins also have an integral role in tumorigenesis through cell growth, migration, invasion, and angiogenesis, making secretomics an excellent method for the discovery of cancer biomarkers. Using a body fluid or full serum proteomic method to identify biomarkers can be extremely difficult – body fluids are complex and highly variable. Secretomic analysis of cancer cell lines or diseased tissue presents a simpler and more specific alternative for biomarker discovery.

The two main biological sources for cancer secretomics are cancer cell line supernatants and proximal biological fluids, the fluids in contact with a tumor. Cancer cell line supernatant is an attractive source of secreted proteins. There are many standardized cell lines available and supernatant is much simpler to analyze than proximal body fluid. But it is unclear whether a cell line secretome is a good representation of an actual tumor in its specific microenvironment and a standardized cell line is not illustrative of the heterogeneity of a real tumor. Analysis of proximal fluids can give a better idea of a human tumor secretome, but this method also has its drawbacks. Procedures for collecting proximal fluids still need to be standardized and non-malignant controls are needed. In addition, environmental and genetic differences between patients can complicate analysis.

# Metabolomics

Metabolomics is the large-scale study of small molecules, commonly known as metabolites, within cells, biofluids, tissues or organisms. Collectively, these small molecules and their interactions within a biological system are known as the metabolome.

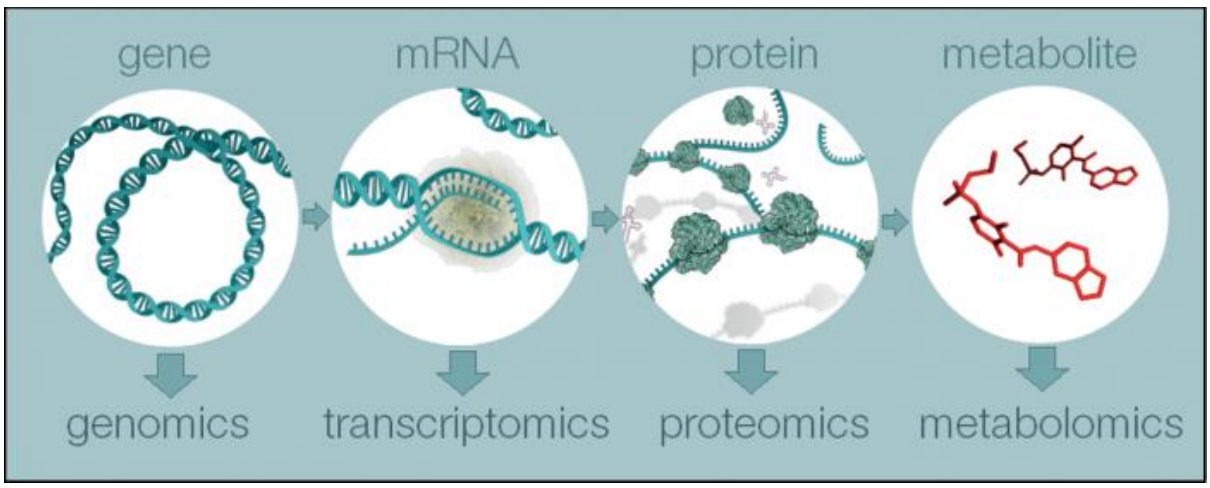



Figure 1 Overview of the four major “omics” fields, from genomics to metabolomics

Just as genomics is the study of DNA and genetic information within a cell, and transcriptomics is the study of RNA and differences in mRNA expression; metabolomics is the study of substrates and products of metabolism, which are influenced by both genetic and environmental factors (Figure 1).

Metabolomics is a powerful approach because metabolites and their concentrations, unlike other “omics” measures, directly reflect the underlying biochemical activity and state of cells / tissues. Thus metabolomics best represents the molecular phenotype.

Small molecules

What are small molecules?

A small molecule (or metabolite) is a low molecular weight organic compound, typically involved in a biological process as a substrate or product. Metabolomics usually studies small molecules within a mass range of 50 – 1500 daltons (Da).

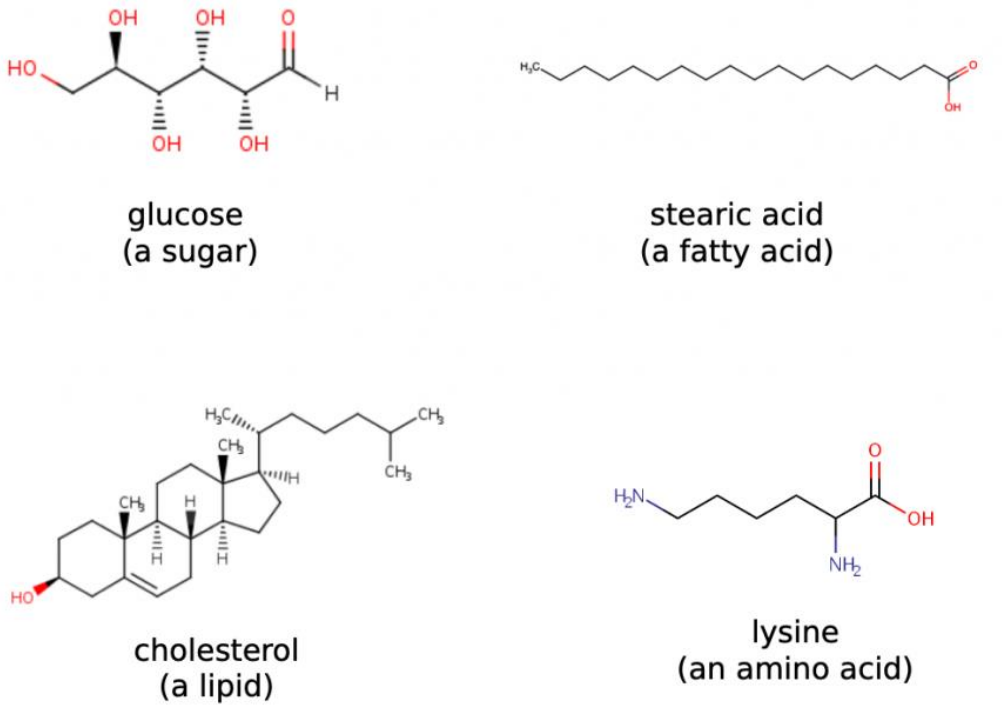



**Figure 2** Examples of small molecules
Examples of small molecules can be seen in Figure 2 and include:

- sugars
- lipids
- amino acids
- fatty acids
- phenolic compounds
- alkaloids

There is a great deal of variation in metabolites between species, it is estimated there are around 200,000 metabolites across the plant kingdom, and somewhere between 7,000 and 15,000 within an individual plant species (1, 2). By contrast, in humans, there are thought to be around 3,000 endogenous or common metabolites (3). These estimates are approximations that are likely to be underestimates because it is difficult to detect low-abundance molecules. Nonetheless, it can be concluded that plants are particularly biochemically rich by comparison with many other species. They also typically contain larger numbers of genes than other eukaryotes.

## The metabolome and metabolic reactions

### The metabolome

The metabolome is the complete set of metabolites within a cell, tissue or biological sample at any given time point. The metabolome is inherently very dynamic: small molecules are continuously absorbed, synthesised, degraded and interact with other molecules, both within and between biological systems, and with the environment. The main metabolic reactions are depicted in Figure 3 below. These cellular reactions are shown as they are represented in the Reactome database.

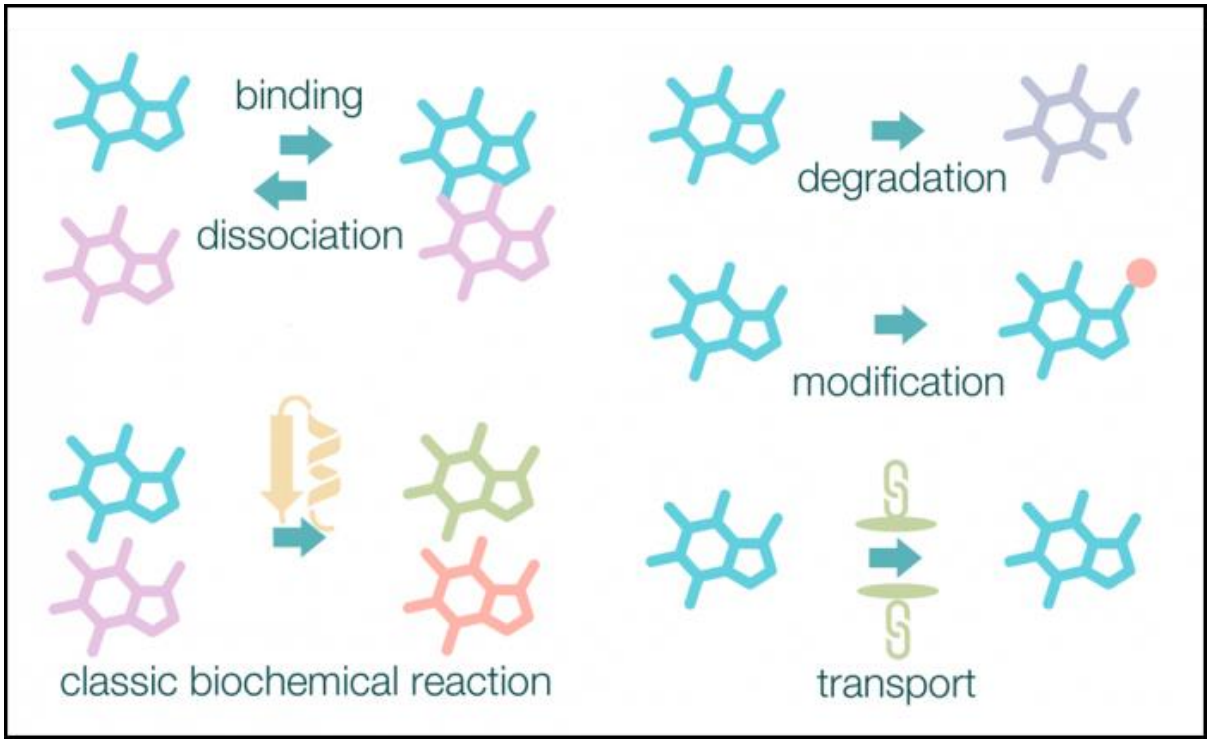



**Figure 3** The main types of metabolic reactions that take place in a cell

*Metabolomics – a 'snapshot' in time*

Many reactions take place continuously within cells, so concentrations of metabolites are considered to be very dynamic, and may change rapidly from one time point to the next. Current analytical techniques used to investigate metabolomics can only take a snapshot in time under a set of defined conditions.

## Metabolic reactions

Metabolic pathways are essentially a series of chemical reactions, catalysed by enzymes, whereby the product of one reaction becomes the substate for the next reaction. These reactions can be divided into anabolic and catabolic.

## The importance of metabolomics

### Why is metabolomics important?

The non-invasive nature of metabolomics and its close link to the phenotype make it an ideal tool for the pharmaceutical, preventive healthcare, and agricultural industries, among others. Biomarker discovery and drug safety screens are two examples where metabolomics has already enabled informed decision making. In the future, with the availablity of personalised metabolomics, we will potentially be able to track the trends of our own metabolome for personalised drugs and improved treatment strategies. Personalised treatment is likely to be more effective than our current medical population-based approaches.

### How is metabolomics used?

We benefit from metabolomics on various levels: from product and stress testing in food industries, e.g. control of pesticides and identification of potentially harmful bacterial strains, to research in agriculture (crop protection and engineering), medical diagnostics in healthcare, and future applications in personalised medicine resulting in personalised treatment strategies.

## Applications of metabolomics

### Agricultural

The development of new pesticides is critical to meet the growing demands on farming. Metabolomics enables us to improve genetically modified plants, and helps us to estimate associated risks by allowing us to get a glimpse of their complex biochemistry viainformative snapshots acquired at different time points during plant development.

Plant metabolomics is particularly interesting because of the range and functions of primary and secondary metabolites in plants. About 300 distinct metabolites could be routinely identified per sample a decade ago, and the number is gradually increasing over time.

### Biomarker discovery

Biomarker discovery is another area where metabolomics informs decision making. Biomarkers are “objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly” (4). In metabolomics, biomarkers are small molecules (metabolites) that can be used to distinguish two groups of samples,

typically a disease and control group. For example, a metabolite reliably present in disease samples, but not in healthy individuals would be classed as a biomarker. Samples of urine, saliva, bile, or seminal fluid contain highly informative metabolites, and can be readily analysed through metabolomics fingerprinting or profiling, for the purpose of biomarker discovery.

Personalised medicine

Personalised medicine, the ultimate customisation of healthcare, requires metabolomics for quick medical diagnosis to identify disease. In healthcare, we currently use classical biochemical tests to measure individual metabolite concentrations to identify disease states (e.g. the blood-glucose level in the case of diabetes). Metabolomics offers the potential for the rapid indentification of hundreds of metabolites, enabling us to identify these disease states much earlier.

https://www.ebi.ac.uk/training/online/courses/metabolomics-introduction/what-is/

# Fluxomics

**Fluxomics** describes the various approaches that seek to determine the rates of metabolic reactions within a biological entity. While metabolomics can provide instantaneous information on the metabolites in a biological sample, metabolism is a dynamic process. The significance of fluxomics is that metabolic fluxes determine the cellular phenotype. It has the added advantage of being based on the metabolome which has fewer components than the genome or proteome.

Fluxomics falls within the field of systems biology which developed with the appearance of high throughput technologies. Systems biology recognizes the complexity of biological systems and has the broader goal of explaining and predicting this complex behavior.

**Metabolic flux**

Metabolic flux refers to the rate of metabolite conversion in a metabolic network.[1][6] For a reaction this rate is a function of both enzyme abundance and enzyme activity.[1] Enzyme concentration is itself a function of transcriptional and translational regulation in addition to the stability of the protein.[1] Enzyme activity is affected by the kinetic parameters of the enzyme, the substrate concentrations, the product concentrations, and the effector molecules concentration.[1] The genomic and environmental effects on metabolic flux are what determine healthy or diseased phenotype.[6]

**Fluxome**

Similar to genome, transcriptome, proteome, and metabolome, the fluxome is defined as the complete set of metabolic fluxes in a cell.[5] However, unlike the others the fluxome is a dynamic representation of the phenotype.[5] This is due to the fluxome resulting from the interactions of the metabolome, genome, transcriptome, proteome, post-translational modifications and the environment.[5]

**Examples of use in research**

One potential application of fluxomic techniques is in drug design. Rama et al. used FBA to study the mycolic acid pathway in *Mycobacterium tuberculosis*. Mycolic acids are known to be important to *M. tuberculosis* survival and as such its pathway has been studied extensively. This allowed the construction of a model of the pathway and for FBA to analyze it. The results of this found multiple possible drug targets for future investigation.

FBA was used to analyze the metabolic networks of multidrug-resistant *Staphylococcus aureus*. By performing in silico single and double gene deletions many enzymes essential to growth were identified.

# Nutrigenomics

Nutrigenomics (also known as nutritional genomics) is broadly defined as the relationship between nutrients, diet, and gene expression [1]. The launch of the Human Genome Project in the 1990s and the subsequent mapping of human DNA sequencing ushered in the 'era of big science', jump-starting the field of nutrigenomics that we know today [2].

Although much of the early 'hype' around nutrigenomics has not yet come to fruition, the field remains nascent and fast-moving, with the potential to lay the foundations of truly 'personalised nutrition' approaches tailored to individuals [3]. It also poses both ethical and regulatory challenges. There is potential for personal data to be misused, in addition to the question of whether it is appropriate to screen for certain genetic phenotypic predispositions where no proven 'treatment' exists. A broad spectrum of stakeholders must therefore engage with the topic, from governments to nutritionists and dietitians, GPs to research scientists.

Such is the hypothetical potential for nutrigenomics to change healthcare, that a 2003 UK Department of Health whitepaper forecast that with increased knowledge of genetics, *"treatment, lifestyle advice, and monitoring aimed at disease prevention could then be tailored appropriately to suit each individual".* The establishment of pan-national organisations such as the European Nutrigenomics Organisation (NUGO) and the International Society for Nutrigenomics & Nutrigenetics have further served to increase the infrastructure and international collaboration around nutrigenomics research. Given the increasing global burden of nutrition-related noncommunicable diseases [4], nutrigenomics could help to develop more sustainable approaches to encouraging dietary change at a

population-level, although a lack of human experimental trials remains a barrier for translating research into policy and practice [5].

*How does nutrigenomics work?*

In addition to the effect of genes on the phenotype (i.e. the physical expression of genetic traits), genes can also respond to environmental influences – of which nutrition is one such influence. Key nutrients of note include those involved in the one-carbon cycle such as folate, choline, and vitamins B2, B6 and B12, and others such as vitamin A, which regulates gene expression. More general dietary patterns such as diets with a high Glycaemic Index (GI) load have also been associated with gene expression, for example the association between a high GI diet and exaggerated polymorphism of the Adiponectin gene, contributing to insulin resistance and diabetes type II.

Nutrigenomics as a research field very much depends on the recent development of advanced technologies that allow us to process a large amount of data relating to gene variants. These so-called '-omic' technologies: genomic, proteomic, metabolomic and transcriptomic, allow us to identify and measure many different types of molecule simultaneously. This is important given that most chronic diseases are not caused by monogenic mutations (as in the case of leptin deficiency), or single genetic effects affected by a single dietary exposure (such as phenylalanine and PKU), but by complex interactions among a very large number of different gene variants [6].

And herein lies one of nutrigenomics' major challenges. The complex biology of human beings makes a mechanistic understanding of exactly how dietary bio-actives react in our bodies difficult to elicit. How to define the optimal intake of individual nutrients for the maintenance of human cells in a 'genomically stable' way remains largely unknown. Diverse genetic backgrounds further complicate the prediction of phenotypes, with some more susceptible to certain conditions than others. The APOE gene for example has three different phenotypes, each with a different probability of CVD risk, and all responding differently to diet and lifestyle factors [7].

*What does the future hold for nutrigenomics?*

Although progress is being made in each of the individual 'omics' fields, effective integration is required in order to provide more comprehensive phenotypic profiles. A recent editorial in *Genes and Nutrition* by NUGO, emphasised the importance of taking a systems approach in future research, with human research studies that incorporate the totality of diet interactions required in order for nutrigenomics to reach its full potential.

Debate remains around the relative impact of genes on the development of chronic disease. In Professor Mathers' 2017 conference talk on the topic (available in *PNS*), he noted that despite some 97 genetic loci (gene variants) identified as contributing to fat accumulation, together the 97 variants explain less than 3% of BMI variance. Neither genes nor our diets alone can therefore entirely explain why some are predisposed to develop certain conditions. Gene expression depends on a complex interplay of genetics with an individual's environment.

On the question of personalised nutrition, and whether nutrigenomics can help to effect sustainable individual diet and lifestyle changes, the recent EU-funded multi-centre Food4Me trial attempted to answer some of these questions. Developing algorithms that integrated information on diet, phenotype and genotype, the trial suggested that personalised nutrition approaches can offer larger health gains than adhering to standard dietary guidelines. However, it should be noted that no significant difference was found between a personalised nutrition approach reliant on counselling, and personalised approaches using genotypic and phenotypic information [8].

Indeed, despite trials such as *Food4Me*, we are not yet at the stage where routine public healthcare encompasses either personalised nutrition or nutrigenomics. A survey undertaken in 2012, found that while some 80% of healthcare practitioners in Greece were willing to recommend a nutrigenomic approach to their patients, only 17% had actually done so.

In bringing together the science of bioinformatics, nutrition, epidemiology, molecular biology and genomics, much remains to be both discovered and determined, but future nutrigenomics research will no doubt provide further intriguing insights into both nutritional science and the human genome.

# Metagenomics

Microbes run the world. It's that simple. Although we can't usually see them, microbes are essential for every part of human life—indeed all life on Earth. Every process in the biosphere is touched by the seemingly endless capacity of microbes to transform the world around them. The chemical cycles that convert the key elements of life—carbon, nitrogen, oxygen, and sulfur—into biologically accessible forms are largely directed by and dependent on microbes. All plants and animals have closely associated microbial communities that make necessary nutrients, metals, and vitamins available to their hosts. Through fermentation and other natural processes, microbes create or add value to many foods that are staples of the human diet. We depend on microbes to remediate toxins in the environment—both the ones that are produced naturally and the ones that are the byproducts of human activities, such as oil and chemical spills. The microbes associated with the human body in the intestine and mouth enable us to extract energy from food that we could not digest without them and protect us against disease-causing agents.

These functions are conducted within complex communities—intricate, balanced, and integrated entities that adapt swiftly and flexibly to environmental change. But historically, the study of microbes has focused on single species in pure culture, so understanding of these complex communities lags behind understanding of their individual members. We know

enough, however, to confirm that microbes, as communities, are key players in maintaining environmental stability.

By making microbes visible, the invention of microscopes in the late 18th century made us aware of their existence. The development of laboratory cultivation methods in the middle 1800s taught us how a few microbes make their livings as individuals, and the molecular biology and genomics revolutions of the last half of the 20th century united this physiological knowledge with a thorough understanding of its underlying genetic basis. Thus, almost all knowledge about microbes is largely "laboratory knowledge," attained in the unusual and unnatural circumstances of growing them optimally in artificial media in pure culture without ecological context. The science of metagenomics, only a few years old, will make it possible to investigate microbes in their natural environments, the complex communities in which they normally live. It will bring about a transformation in biology, medicine, ecology, and biotechnology that may be as profound as that initiated by the invention of the microscope.

**WHAT IS METAGENOMICS?**

Like genomics itself, metagenomics is both a set of *research techniques*, comprising many related approaches and methods, and a *research field*. In Greek, *meta* means "transcendent." In its approaches and methods, metagenomics circumvents the unculturability and genomic diversity of most microbes, the biggest roadblocks to advances in clinical and environmental microbiology. *Meta* in the first context recognizes the need to develop computational methods that maximize understanding of the genetic composition and activities of communities so complex that they can only be sampled, never completely characterized. In the second sense, that of a research field, *meta* means that this new science seeks to understand biology at the aggregate level, transcending the individual organism to focus on the genes in the community and how genes might influence each other's activities in serving collective functions. Individual organisms remain the units of community activities, of course, and we anticipate that metagenomics will complement and stimulate research on individuals and their genomes. In the next decades, we expect that the top-down approach of metagenomics, the bottom-up approach of classical microbiology, and organism-level genomics will merge. We will understand communities, and the collection of communities that forms the biosphere, as a nested system of systems of which humans are a part and on which human survival depends. In some situations, it will be possible to apply the new understanding to problems of urgency and importance.

Metagenomics in either sense will probably never be circumscribed tightly by a definition, and it would be undesirable to attempt to so limit it now, but the term includes cultivation-independent genome-level characterization of communities or their members, high-throughput gene-level studies of communities with methods borrowed from genomics, and other "omics" studies, which are aimed at understanding transorganismal behaviors and the biosphere at the genomic level. Although in its current early implementation (and for the purposes of this report) metagenomics focuses on non-eukaryotic microbes, there is no doubt that its concepts and methods will ultimately transform all biology. In just this way has genomics, a science developed to aid the advancement of biomedicine and the understanding of our own species, transformed the science of all organisms and the application of that science in epidemiology, clinical microbiology, virology, agriculture, forestry, fisheries, biotechnology, microbial forensics, and many other fields.

**METAGENOMICS OFFERS A WAY FORWARD**

The pure culture paradigm has not only limited what microbiologists have studied; it has also limited how they think about microbes. Microbes have been studied as sovereign entities and examined only for their responses to the simple chemicals that can be added to their media. We know little about their behavior as partners in the strategic alliances that are metabolic consortia, such as the consortia that decontaminate drinking water or that make up the complex structured biofilms that keep dental hygienists busy. The invisible members of a microbial community can differ vastly in their biochemical activities and interactions, not only between species but also within species. Phylotyping gives some reliable information about "Who is there?" but because of within-species genomic diversity, only imperfect guesses as to "What are they doing?" Metagenomic methods, which will be discussed later, go a long way toward answering the second question. In the end, it may be possible to view ecosystems themselves as biological units with their own genetic repertoires and to sidestep consideration of individual species. Then, both "Who is there?" and "What are they doing?" could be replaced with "What is being done by the community?"

Such understanding can be achieved only with methods that go beyond the pure-culture and single-whole-genome approaches that have dominated microbial genomics. We must move directly to the genes, to defining environments by the potential and realized biochemical and geochemical activities of the genes that are there, and the complex patterns of interactions within and between cells that regulate their responses to changes in their physical and biological surroundings. We must do this while recognizing that—except in restricted environments and specialized consortia with limited numbers of genetically homogeneous constituents—we will be dealing with enormous amounts of data that will represent an incomplete sampling of the genetic diversity present. In short, we must adopt the methods of metagenomics.

# Organomics

Innovative methods designed to recapitulate human organogenesis from pluripotent stem cells provide a means to explore human developmental biology. New technologies to sequence and analyze single-cell transcriptomes can deconstruct these 'organoids' into constituent parts, and reconstruct lineage trajectories during cell differentiation. In this Spotlight article we summarize the different approaches to performing single-cell transcriptomics on organoids, and discuss the opportunities and challenges of applying these techniques to generate organ-level, mechanistic models of human development and disease. Together, these technologies will move past characterization to the prediction of human developmental and disease-related phenomena.

Understanding how multiple different cell types come together to build an organ has been a long-standing fascination in developmental biology. Over the years, we have learned much with regard to the molecular events that instruct cell lineage, the specific growth factors that are required, and the morphological aspects that drive organ development. Most of this knowledge has been gained from studying non-human vertebrate organogenesis; however, the observation that differences exist between how organs are formed across a range of species has led us to question what it is that makes us uniquely human. The revelation that human pluripotent stem cells can self-organize into three-dimensional structures that contain multiple differentiated cell types organized to resemble primary human tissue has revitalized the field of human developmental biology (McCauley and Wells, 2017). In general, these structures are referred to as organoids, and protocols have been developed to generate gut, kidney, liver bud, multiple regions of the human brain, and other tissues (McCauley and Wells, 2017). Conventional strategies to analyze human organoid development often assess cell composition and differentiation using immunohistochemistry of a limited set of marker proteins, or cell tracking via a reporter gene. Because organoids are, by definition, composed of many different cell states and often show large organoid-to-organoid variability, high-throughput single-cell transcriptomics represents an exciting strategy to assess cell composition, lineage relationships, and gene networks in organoids. We focus on single-cell RNA sequencing (scRNA-seq), summarizing some of the commonly used methods and their advantages and limitations. We further discuss how we envisage improvements in single-cell transcriptomic methodology will enhance our comprehension of human developmental biology and disease

Human organoids are manipulable, genetically and otherwise, a feature once reserved for classical model systems such as yeast, worms, flies, fish and mice. As a technology, however, in vitro organogenesis is still in its infancy, and in many cases it is unclear exactly what cell types are present within organoids and whether each cell type can be created in a reproducible manner. scRNA-seq will help to address this uncertainty, providing a greater depth of analysis of cell heterogeneity and reproducibility. As protocols continue to evolve, human organoids are likely to come even closer to recapitulating bona fide human organogenesis in a predictable and reproducible way, making organoids a highly relevant system for understanding human development. We feel that quantitative single-cell transcriptomic approaches will provide impressive resolution of cell composition, lineage relationships, and gene network function within developing organoids, and, together with other genomic approaches, will offer unprecedented insight into the mechanisms that underpin human organogenesis. Methods to analyze DNA, methylation, chromatin accessibility, nonmessenger RNAs and proteins in single cells will further advance the field. The cost-per-cell of many single-cell approaches is rapidly reducing and new methods are emerging that are relatively simple to implement in the lab. Hence, we believe that these technologies applied to human organoids represent a new direction in human developmental biology, and will help pave the way towards a better appreciation of what makes us uniquely human.

# Pharmacogenomics

**Pharmacogenomics** is the study of the role of the genome in drug response. Its name (*pharmaco-* + *genomics*) reflects its combining of pharmacology and genomics. Pharmacogenomics analyzes how the genetic makeup of an individual affects his/her response to drugs. It deals with the influence of acquired and inherited genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with pharmacokinetics (drug absorption, distribution, metabolism, and elimination) and pharmacodynamics (effects mediated through a drug's biological targets). The term *pharmacogenomics* is often used interchangeably with *pharmacogenetics*. Although both terms relate to drug response based on genetic influences, pharmacogenetics focuses on single drug-gene interactions, while pharmacogenomics encompasses a more genome-wide association approach, incorporating genomics and epigenetics while dealing with the effects of multiple genes on drug response.

Pharmacogenomics aims to develop rational means to optimize drug therapy, with respect to the patients' genotype, to ensure maximum efficiency with minimal adverse effects. Through the utilization of pharmacogenomics, it is hoped that pharmaceutical drug treatments can deviate from what is dubbed as the "one-dose-fits-all" approach. Pharmacogenomics also attempts to eliminate the trial-and-error method of prescribing, allowing physicians to take into consideration their patient's genes, the functionality of these genes, and how this may affect the efficacy of the patient's current or future treatments (and where applicable, provide an explanation for the failure of past treatments). Such approaches promise the advent of precision medicine and even personalized medicine, in which drugs and drug combinations are optimized for narrow subsets of patients or even for each individual's unique genetic makeup. Whether used to explain a patient's response or lack thereof to a treatment, or act as a predictive tool, it hopes to achieve better treatment outcomes, greater efficacy, minimization of the occurrence of drug toxicities and adverse drug reactions (ADRs). For patients who have lack of therapeutic response to a treatment, alternative therapies can be prescribed that would best suit their requirements. In order to provide pharmacogenomic recommendations for a given drug, two possible types of input can be used: genotyping or exome or whole genome sequencing. Sequencing provides many more data points, including detection of mutations that prematurely terminate the synthesized protein (early stop codon).

**Applications**

The list below provides a few more commonly known applications of pharmacogenomics:

- Improve drug safety, and reduce ADRs;
- Tailor treatments to meet patients' unique genetic pre-disposition, identifying optimal dosing;
- Improve drug discovery targeted to human disease; and
- Improve proof of principle for efficacy trials.

Pharmacogenomics may be applied to several areas of medicine, including pain management, cardiology, oncology, and psychiatry. A place may also exist in forensic pathology, in which pharmacogenomics can be used to determine the cause of death in drug-related deaths where no findings emerge using autopsy.

In cancer treatment, pharmacogenomics tests are used to identify which patients are most likely to respond to certain cancer drugs. In behavioral health, pharmacogenomic tests provide tools for physicians and care givers to better manage medication selection and side effect amelioration. Pharmacogenomics is also known as companion diagnostics, meaning tests being bundled with drugs. Examples include KRAS test with cetuximab and EGFR test with gefitinib. Beside efficacy, germline pharmacogenetics can help to identify patients likely to undergo severe toxicities when given cytotoxics showing impaired detoxification in relation with genetic polymorphism, such as canonical 5-FU. In particular, genetic deregulations affecting genes coding for DPD, UGT1A1, TPMT, CDA and CYP2D6 are now considered as critical issues for patients treated with 5-FU/capecitabine, irinotecan, mercaptopurine/azathioprine, gemcitabine/capecitabine/AraC and tamoxifen, respectively.

In cardiovascular disorders, the main concern is response to drugs including warfarin, clopidogrel, beta blockers, and statins. In patients with CYP2C19, who take clopidogrel, cardiovascular risk is elevated, leading to medication package insert updates by regulators. In patients with type 2 diabetes, haptoglobin (Hp) genotyping shows an effect on cardiovascular disease, with Hp2-2 at higher risk and supplemental vitamin E reducing risk by affecting HDL.

In psychiatry, as of 2010, research has focused particularly on 5-HTTLPR and DRD2.

# Phytochemomics

Various food components positively affect human health and wellness. Phytochemicals have been proposed as health promoters. Several claimed healthy products including foods, dietary supplements, nutraceutics and cosmetics containing phytochemicals are commercialized worldwide. Products based on phytochemicals are nowadays very popular. Phytochemicals´ health promoting properties are under evaluation by scientists and regulators´ authorities. Phytochemomics is a comprehensive concept aimed to increase the knowledge on phytochemicals´ bioactivity and their impact in health, aging and diseases, which is of growing importance in food, medicine and cosmetic sciences. These achievements are based on up-to-date analytical platforms including, but not limited, to mass spectrometric approaches. Foods are very complexmixtures of bioactive components in different concentrations. Phytochemomics together with other omics are essential for authorizing or rejecting nutrition and health claims made on foods. On the basis of the data collected by using omic approaches a cause-effect relationship may be established between a food category, a food or one of its constituents and the claimed effect.

## Microbiomics

Microbiomics' is a fast-growing field in which all the microorganisms of a given community (a 'microbiota') are investigated together. This could be the microbiota of an environmental sample (e.g. soil or water), a particular body site (e.g. the gut or the mouth) or from a particular organism (e.g. farm or zoo animals).

Investigation of environmental microbial communities can be of particular interest in discovery of 'novel' organisms with exciting properties, such as production of natural products with antimicrobial properties. Research into farm animal gut microbiota has the potential to help the agricultural industry improve growth of animals whilst reducing excess antibiotic use (an important driver of antimicrobial resistance), and research groups investigating the human 'commensal' microbiota (the 'normal' microflora associated with our bodies) are exploring the many potential roles these organisms play in health and disease, and how disruption of these microbial communities can affect our wellbeing.

## Microbiota profiling

At the simplest level, microbiomics looks to investigate the make-up of a particular microbial community and how this might change over time, or with particular pressure. This is typically done using 16S profiling. DNA is extracted from the target sample (soil / water / faeces) and the 16S RNA gene amplified using the polymerase chain reaction (PCR). The resulting amplified DNA fragments are sequenced, and can be matched against sequence databases to achieve identifications of the organisms present. This can give a 'snapshot' of the communities present in a sample at any given time, and can be used to compare the communities in different samples, or follow changes in the communities of a particular sample site over time.

The Healthcare Associated Infection Research Group at Leeds is interested in the gut microbiome, and how the 'normal' communities of the gut interact with potentially harmful bacteria such as *Clostridium difficile* and Multi Drug Resistant Enterobacteriaceae (MDRE). We use an artificial human gut system to model the bacterial populations of the gut and to investigate changes to these populations following interventions such as exposure to antibiotics.

## Meta-Omics

Rapid development of high throughput molecular methods has made it possible to carry out detailed investigations of the microbiota with regards to their genetic and functional diversity, providing a better understanding of what species are present, how they relate to each other, what genes are expressed and what metabolic activities are undergoing. Various platforms enabling these technologies (Illumina, 454 Roche, PacBio, Oxford Nanopores etc) now allow the study of metagenomics (all the genes in a given sample), metatranscriptomics (the gene

expression (mRNA) in a given sample), metaproteomics (all the proteins in a given sample) and metametabolomics (all the metabolites in a given sample), exploring the biological signatures that are associated with specific environments.

The Microbiology and Cell Biology Group (School of Dentistry, Leeds) aims to investigate host-microbiome interactions to understand the role of the microbiome and host responses in health and infectious diseases. The research group employs an interdisciplinary approach to decipher the links between oral health and systemic health or diseases, using various biofilm models and clinical samples, making use of our state of the art Leeds Dental Translational and Clinical Research Unit facility. For example, we are working on various projects exploring links between periodontitis and arthritis and diabetes.