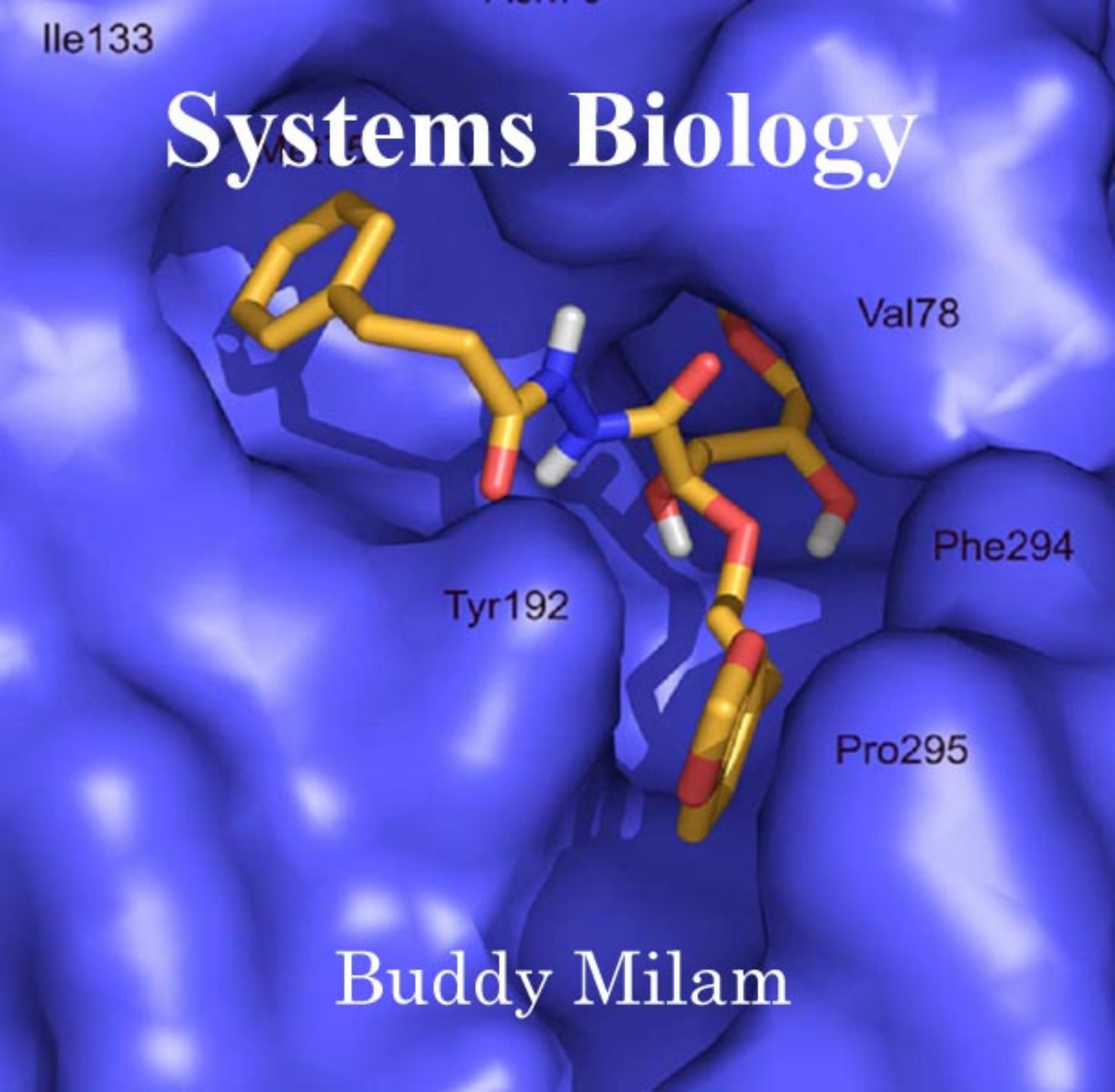


Systems Biology

Buddy Milam

First Edition, 2012

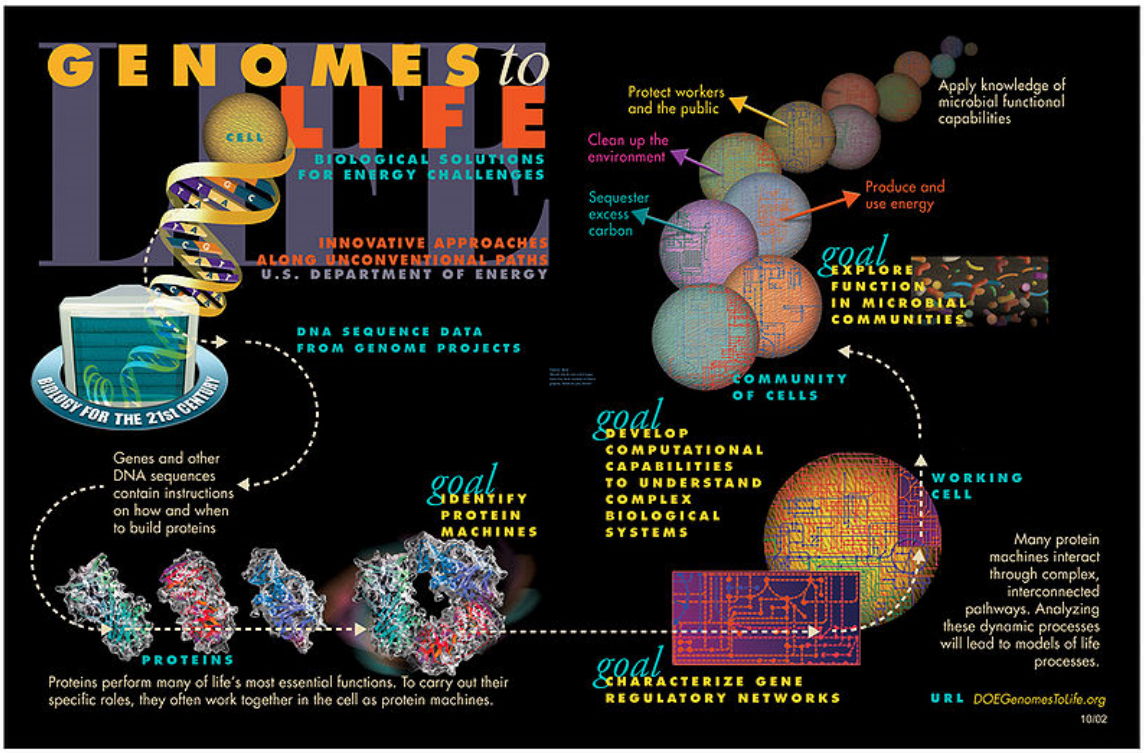
ISBN 978-81-323-3159-9

Published by:
Research World
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Systems Biology



Example of systems biology research.

Systems biology is a term used to describe a number of trends in bioscience research, and a movement which draws on those trends. Proponents describe systems biology as a biology-based inter-disciplinary study field that focuses on complex interactions in biological systems, claiming that it uses a new perspective (holism instead of reduction). Particularly from year 2000 onwards, the term is used widely in the biosciences, and in a variety of contexts. An often stated ambition of systems biology is the modeling and discovery of emergent properties, properties of a system whose theoretical description is only possible using techniques which fall under the remit of systems biology.

Overview

Systems biology can be considered from a number of different aspects:

- As a **field of study**, particularly, the study of the interactions between the components of *biological systems*, and how these interactions give rise to the function and behavior of that system (for example, the enzymes and metabolites in a metabolic pathway).

- As a **paradigm**, usually defined in antithesis to the so-called reductionist paradigm (biological organisation), although fully consistent with the scientific method. The distinction between the two paradigms is referred to in these quotations:

  "The reductionist approach has successfully identified most of the components and many of the interactions but, unfortunately, offers no convincing concepts or methods to understand how system properties emerge...the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously and by rigorous data integration with mathematical models" Science
  "Systems biology...is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different....It means changing our philosophy, in the full sense of the term" Denis Noble

- As a series of **operational protocols used for performing research**, namely a cycle composed of theory, analytic or computational modelling to propose specific testable hypotheses about a biological system, experimental validation, and then using the newly acquired quantitative description of cells or cell processes to refine the computational model or theory. Since the objective is a model of the interactions in a system, the experimental techniques that most suit systems biology are those that are system-wide and attempt to be as complete as possible. Therefore, transcriptomics, metabolomics, proteomics and high-throughput techniques are used to collect quantitative data for the construction and validation of models.

- As the application of dynamical systems theory to molecular biology.

- As a **socioscientific phenomenon** defined by the strategy of pursuing integration of complex data about the interactions in biological systems from diverse experimental sources using interdisciplinary tools and personnel.

This variety of viewpoints is illustrative of the fact that systems biology refers to a cluster of peripherally overlapping concepts rather than a single well-delineated field. However the term has widespread currency and popularity as of 2007, with chairs and institutes of systems biology proliferating worldwide.

History

Systems biology finds its roots in:

- the quantitative modeling of enzyme kinetics, a discipline that flourished between 1900 and 1970,
- the mathematical modeling of population growth,
- the simulations developed to study neurophysiology, and
- control theory and cybernetics.

One of the theorists who can be seen as one of the precursors of systems biology is Ludwig von Bertalanffy with his general systems theory . One of the first numerical simulations in biology was published in 1952 by the British neurophysiologists and Nobel prize winners Alan Lloyd Hodgkin and Andrew Fielding Huxley, who constructed a mathematical model that explained the action potential propagating along the axon of a neuronal cell. Their model described a cellular function emerging from the interaction between two different molecular components, a potassium and a sodium channels, and can therefore be seen as the beginning of computational systems biology. In 1960, Denis Noble developed the first computer model of the heart pacemaker.

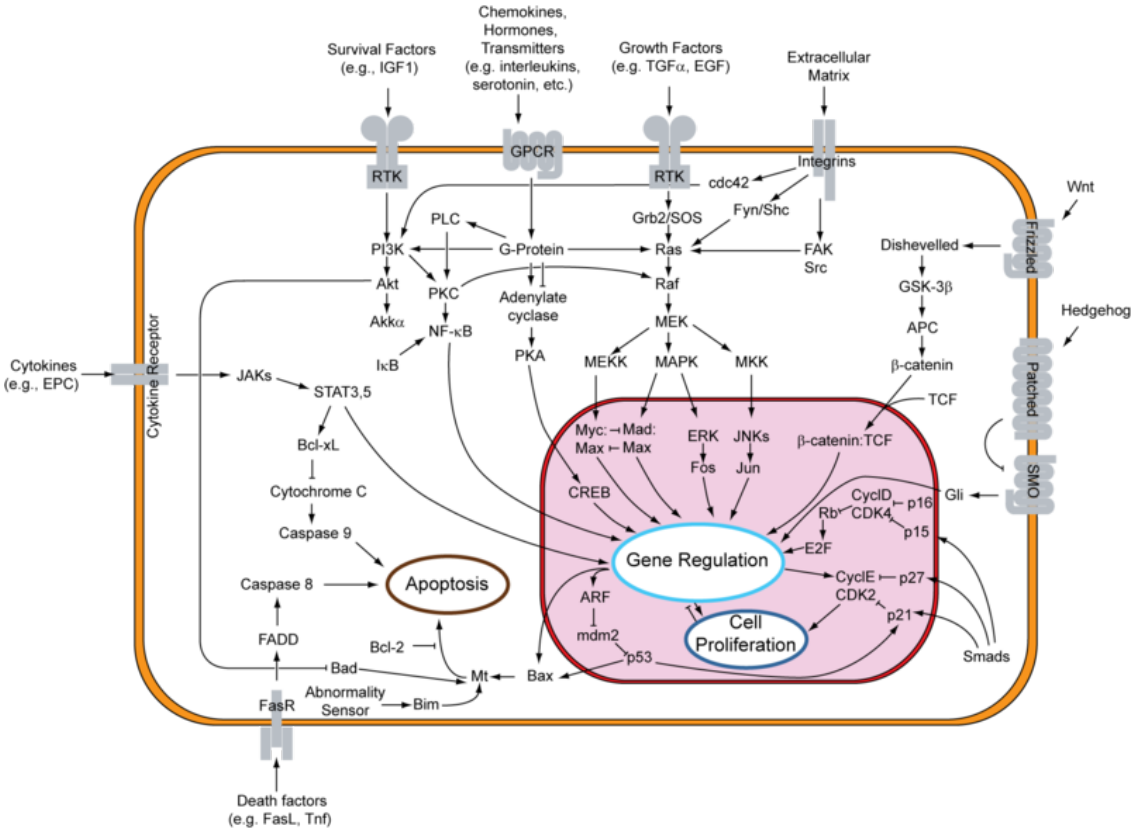
The formal study of systems biology, as a distinct discipline, was launched by systems theorist Mihajlo Mesarovic in 1966 with an international symposium at the Case Institute of Technology in Cleveland, Ohio entitled "Systems Theory and Biology."

The 1960s and 1970s saw the development of several approaches to study complex molecular systems, such as the Metabolic Control Analysis and the biochemical systems theory. The successes of molecular biology throughout the 1980s, coupled with a skepticism toward theoretical biology, that then promised more than it achieved, caused the quantitative modelling of biological processes to become a somewhat minor field.

However the birth of functional genomics in the 1990s meant that large quantities of high quality data became available, while the computing power exploded, making more realistic models possible. In 1997, the group of Masaru Tomita published the first quantitative model of the metabolism of a whole (hypothetical) cell.

Around the year 2000, after Institutes of Systems Biology were established in Seattle and Tokyo, systems biology emerged as a movement in its own right, spurred on by the completion of various genome projects, the large increase in data from the omics (e.g. genomics and proteomics) and the accompanying advances in high-throughput experiments and bioinformatics. Since then, various research institutes dedicated to systems biology have been developed. As of summer 2006, due to a shortage of people in systems biology several doctoral training centres in systems biology have been established in many parts of the world.

Disciplines associated with systems biology



Overview of signal transduction pathways

According to the interpretation of Systems Biology as the ability to obtain, integrate and analyze complex data from multiple experimental sources using interdisciplinary tools, some typical technology platforms are:

- Phenomics: Organismal variation in phenotype as it changes during its life span.
- Genomics: Organismal deoxyribonucleic acid (DNA) sequence, including intra-organisamal cell specific variation. (i.e. Telomere length variation etc.).
- Epigenomics / Epigenetics: Organismal and corresponding cell specific transcriptomic regulating factors not empirically coded in the genomic sequence. (i.e. DNA methylation, Histone Acetelation etc.).
- Transcriptomics: Organismal, tissue or whole cell gene expression measurements by DNA microarrays or serial analysis of gene expression
- Interferomics: Organismal, tissue, or cell level transcript correcting factors (i.e. RNA interference)
- Translatomics / Proteomics: Organismal, tissue, or cell level measurements of proteins and peptides via two-dimensional gel electrophoresis, mass spectrometry or multi-dimensional protein identification techniques (advanced HPLC systems coupled with mass spectrometry). Sub disciplines include phosphoproteomics, glycoproteomics and other methods to detect chemically modified proteins.

- Metabolomics: Organismal, tissue, or cell level measurements of all small-molecules known as metabolites.
- Glycomics: Organismal, tissue, or cell level measurements of carbohydrates.
- Lipidomics: Organismal, tissue, or cell level measurements of lipids.

In addition to the identification and quantification of the above given molecules further techniques analyze the dynamics and interactions within a cell. This includes:

- Interactomics: Organismal, tissue, or cell level study of interactions between molecules. Currently the authoritative molecular discipline in this field of study is protein-protein interactions (PPI), although the working definition does not preclude inclusion of other molecular disciplines such as those defined here.
- Fluxomics: Organismal, tissue, or cell level measurements of molecular dynamic changes over time.
- Biomics: systems analysis of the biome.

The investigations are frequently combined with large scale perturbation methods, including gene-based (RNAi, mis-expression of wild type and mutant genes) and chemical approaches using small molecule libraries. Robots and automated sensors enable such large-scale experimentation and data acquisition. These technologies are still emerging and many face problems that the larger the quantity of data produced, the lower the quality. A wide variety of quantitative scientists (computational biologists, statisticians, mathematicians, computer scientists, engineers, and physicists) are working to improve the quality of these approaches and to create, refine, and retest the models to accurately reflect observations.

The systems biology approach often involves the development of mechanistic models, such as the reconstruction of dynamic systems from the quantitative properties of their elementary building blocks. For instance, a cellular network can be modelled mathematically using methods coming from chemical kinetics and control theory. Due to the large number of parameters, variables and constraints in cellular networks, numerical and computational techniques are often used.

Other aspects of computer science and informatics are also used in systems biology. These include:

- New forms of computational model, such as the use of process calculi to model biological processes (notable approaches include stochastic π-calculus, BioAmbients, Beta Binders, BioPEPA and Brane calculus) and constraint-based modeling.
- Integration of information from the literature, using techniques of information extraction and text mining.
- Development of online databases and repositories for sharing data and models, approaches to database integration and software interoperability via loose coupling of software, websites and databases, or commercial suits.

- Development of syntactically and semantically sound ways of representing biological models.

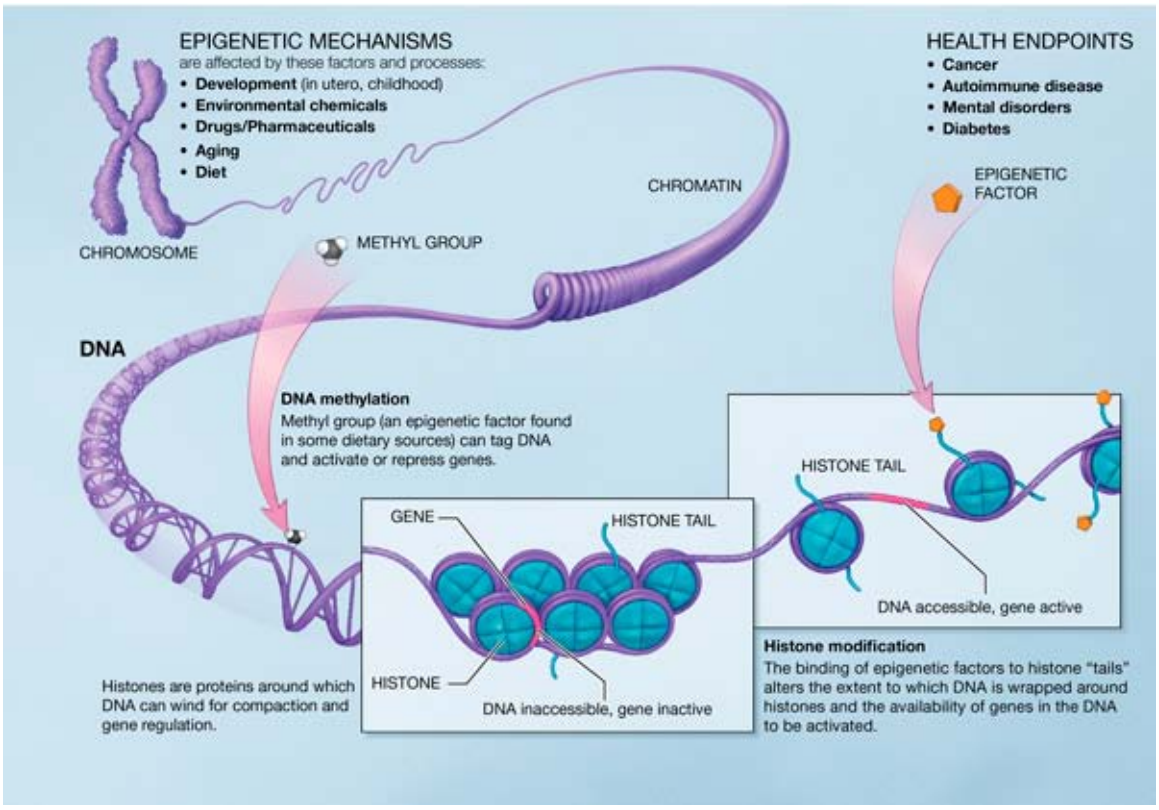
Chapter- 2

Epigenetics

In biology, and specifically genetics, **epigenetics** is the study of heritable changes in phenotype (appearance) or gene expression caused by mechanisms other than changes in the underlying DNA sequence, hence the name *epi-* (Greek: *επί-* over, above) *-genetics*. These changes may remain through cell divisions for the remainder of the cell's life and may also last for multiple generations. However, there is no change in the underlying DNA sequence of the organism; instead, non-genetic factors cause the organism's genes to behave (or "express themselves") differently.

The best example of epigenetic changes 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, a single fertilized egg cell – the zygote – changes into the many cell types including neurons, muscle cells, epithelium, blood vessels etc. as it continues to divide. It does so by activating some genes while inhibiting others.

Etymology and definitions



epigenetic mechanisms

Epigenetics (as in "epigenetic landscape") was coined by C. H. Waddington in 1942 as a portmanteau of the words *genetics* and *epigenesis*. *Epigenesis* is an old word which has more recently been used to describe the differentiation of cells from their initial totipotent state in embryonic development. When Waddington coined the term the physical nature of genes and their role in heredity was not known; he used it as a conceptual model of how genes might interact with their surroundings to produce a phenotype.

Robin Holliday defined epigenetics as "the study of the mechanisms of temporal and spatial control of gene activity during the development of complex organisms." Thus *epigenetic* can be used to describe anything other than DNA sequence that influences the development of an organism.

The modern usage of the word in scientific discourse is more narrow, referring to heritable traits (over rounds of cell division and sometimes transgenerationally) that do not involve changes to the underlying DNA sequence. The Greek prefix *epi-* in *epigenetics* implies features that are "on top of" or "in addition to" genetics; thus *epigenetic* traits exist on top of or in addition to the traditional molecular basis for inheritance.

The similarity of the word to "genetics" has generated many parallel usages. The "epigenome" is a parallel to the word "genome", and refers to the overall epigenetic state of a cell. The phrase "genetic code" has also been adapted—the "epigenetic code" has been used to describe the set of epigenetic features that create different phenotypes in different cells. Taken to its extreme, the "epigenetic code" could represent the total state of the cell, with the position of each molecule accounted for in an *epigenomic map*, a diagrammatic representation of the gene expression, DNA methylation and histone modification status of a particular genomic region. More typically, the term is used in reference to systematic efforts to measure specific, relevant forms of epigenetic information such as the histone code or DNA methylation patterns.

The psychologist Erik Erikson used the term *epigenetic* in his theory of psychosocial development. That usage, however, is of primarily historical interest.

Molecular basis of epigenetics

The molecular basis of epigenetics is complex. It involves modifications of the activation of certain genes, but not the basic structure of DNA. Additionally, the chromatin proteins associated with DNA may be activated or silenced. This accounts for why the differentiated cells in a multi-cellular organism express only the genes that are necessary for their own activity. Epigenetic changes are preserved when cells divide. Most epigenetic changes only occur within the course of one individual organism's lifetime, but, if a mutation in the DNA has been caused in sperm or egg cell that results in fertilization, then some epigenetic changes are inherited from one generation to the next. This raises the question of whether or not epigenetic changes in an organism can alter the basic structure of its DNA, a form of Lamarckism.

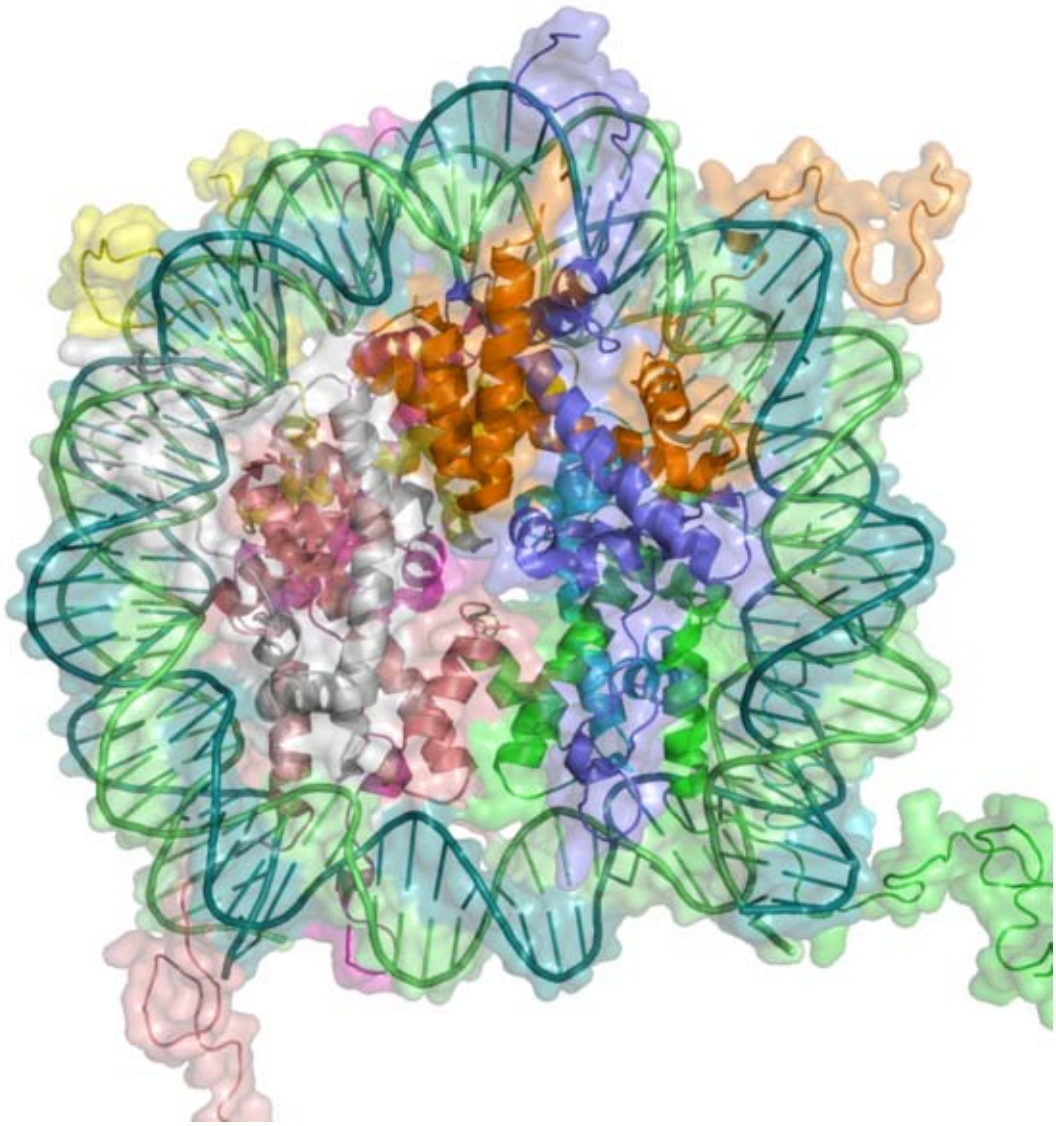
Specific epigenetic processes include paramutation, bookmarking, imprinting, gene silencing, X chromosome inactivation, position effect, reprogramming, transvection, maternal effects, the progress of carcinogenesis, many effects of teratogens, regulation of histone modifications and heterochromatin, and technical limitations affecting parthenogenesis and cloning.

Epigenetic research uses a wide range of molecular biologic techniques to further our understanding of epigenetic phenomena, including chromatin immunoprecipitation (together with its large-scale variants ChIP-on-chip and ChIP-seq), fluorescent in situ hybridization, methylation-sensitive restriction enzymes, DNA adenine methyltransferase identification (DamID) and bisulfite sequencing. Furthermore, the use of bioinformatic methods is playing an increasing role (computational epigenetics).

Mechanisms

Several types of epigenetic inheritance systems may play a role in what has become known as cell memory:

DNA methylation and chromatin remodeling



DNA associates with histone proteins to form chromatin.

Because the phenotype of a cell or individual is affected by which of its genes are transcribed, heritable transcription states can give rise to epigenetic effects. There are several layers of regulation of gene expression. One way that genes are regulated is through the remodeling of chromatin. Chromatin is the complex of DNA and the histone proteins with which it associates. Histone proteins are little spheres that DNA wraps around. If the way that DNA is wrapped around the histones changes, gene expression can change as well. Chromatin remodeling is accomplished through two main mechanisms:

1. The first way is post translational modification of the amino acids that make up histone proteins. Histone proteins are made up of long chains of amino acids. If

the amino acids that are in the chain are changed, the shape of the histone sphere might be modified. DNA is not completely unwound during replication. It is possible, then, that the modified histones may be carried into each new copy of the DNA. Once there, these histones may act as templates, initiating the surrounding new histones to be shaped in the new manner. By altering the shape of the histones around it, these modified histones would ensure that a differentiated cell would stay differentiated, and not convert back into being a stem cell.

2. The second way is the addition of methyl groups to the DNA, mostly at CpG sites, to convert cytosine to 5-methylcytosine. 5-Methylcytosine performs much like a regular cytosine, pairing up with a guanine. However, some areas of genome are methylated more heavily than others and highly methylated areas tend to be less transcriptionally active, through a mechanism not fully understood. Methylation of cytosines can also persist from the germ line of one of the parents into the zygote, marking the chromosome as being inherited from this parent (genetic imprinting).

The way that the cells stay differentiated in the case of DNA methylation is clearer to us than it is in the case of histone shape. Basically, certain enzymes (such as DNMT1) have a higher affinity for the methylated cytosine. If this enzyme reaches a "hemimethylated" portion of DNA (where methylcytosine is in only one of the two DNA strands) the enzyme will methylate the other half.

Although histone modifications occur throughout the entire sequence, the unstructured N-termini of histones (called histone tails) are particularly highly modified. These modifications include acetylation, methylation, ubiquitylation, phosphorylation and sumoylation. Acetylation is the most highly studied of these modifications. For example, acetylation of the K14 and K9 lysines of the tail of histone H3 by histone acetyltransferase enzymes (HATs) is generally correlated with transcriptional competence.

One mode of thinking is that this tendency of acetylation to be associated with "active" transcription is biophysical in nature. Because it normally has a positively charged nitrogen at its end, lysine can bind the negatively charged phosphates of the DNA backbone. The acetylation event converts the positively charged amine group on the side chain into a neutral amide linkage. This removes the positive charge, thus loosening the DNA from the histone. When this occurs, complexes like SWI/SNF and other transcriptional factors can bind to the DNA and allow transcription to occur. This is the "cis" model of epigenetic function. In other words, changes to the histone tails have a direct affect on the DNA itself.

Another model of epigenetic function is the "trans" model. In this model changes to the histone tails act indirectly on the DNA. For example, lysine acetylation may create a binding site for chromatin modifying enzymes (and basal transcription machinery as well). This Chromatin Remodeler can then cause changes to the state of the chromatin. Indeed, the bromodomain — a protein segment (domain) that specifically binds acetyl-

lysine — is found in many enzymes that help activate transcription, including the SWI/SNF complex (on the protein polybromo). It may be that acetylation acts in this and the previous way to aid in transcriptional activation.

The idea that modifications act as docking modules for related factors is borne out by histone methylation as well. Methylation of lysine 9 of histone H3 has long been associated with constitutively transcriptionally silent chromatin (constitutive heterochromatin). It has been determined that a chromodomain (a domain that specifically binds methyl-lysine) in the transcriptionally repressive protein HP1 recruits HP1 to K9 methylated regions. One example that seems to refute this biophysical model for acetylation is that tri-methylation of histone H3 at lysine 4 is strongly associated with (and required for full) transcriptional activation. Tri-methylation in this case would introduce a fixed positive charge on the tail.

It has been shown that the histone lysine methyltransferase (KMT) is responsible for this methylation activity in the pattern of histones H3 & H4. This enzyme utilizes a catalytically active site called the SET domain (Supressor of variegation, Enhancer of zeste, Trithorax). The SET domain is a 130-amino acid sequence involved in modulating gene activities. This domain has been demonstrated to bind to the histone tail and causes the methylation of the histone.

Differing histone modifications are likely to function in differing ways; acetylation at one position is likely to function differently than acetylation at another position. Also, multiple modifications may occur at the same time, and these modifications may work together to change the behavior of the nucleosome. The idea that multiple dynamic modifications regulate gene transcription in a systematic and reproducible way is called the histone code.

DNA methylation frequently occurs in repeated sequences, and helps to suppress the expression and mobility of 'transposable elements': Because 5-methylcytosine is chemically very similar to thymidine, CpG sites are frequently mutated and become rare in the genome, except at CpG islands where they remain unmethylated. Epigenetic changes of this type thus have the potential to direct increased frequencies of permanent genetic mutation. DNA methylation patterns are known to be established and modified in response to environmental factors by a complex interplay of at least three independent DNA methyltransferases, DNMT1, DNMT3A and DNMT3B, the loss of any of which is lethal in mice. DNMT1 is the most abundant methyltransferase in somatic cells, localizes to replication foci, has a 10–40-fold preference for hemimethylated DNA and interacts with the proliferating cell nuclear antigen (PCNA). By preferentially modifying hemimethylated DNA, DNMT1 transfers patterns of methylation to a newly synthesized strand after DNA replication, and therefore is often referred to as the ‘maintenance' methyltransferase. DNMT1 is essential for proper embryonic development, imprinting and X-inactivation.

Histones H3 and H4 can also be manipulated through demethylation using histone lysine demethylase (KDM). This recently identifited enzyme has a catalytically active site

called the Jumonji domain (JmjC). The demethylation occurs when JmjC utilizes multiple cofactors to hydroxylate the methyl group, thereby removing it. JmjC is capable of demethylating mono-, di-, and tri-methylated substrates. .

Chromosomal regions can adopt stable and heritable alternative states resulting in bistable gene expression without changes to the DNA sequence. Epigenetic control is often associated with alternative covalent modifications of histones. The stability and heritability of states of larger chromosomal regions are often thought to involve positive feedback where modified nucleosomes recruit enzymes that similarly modify nearby nucleosomes. A simplified stochastic model for this type of epigenetics is found here .

Because DNA methylation and chromatin remodeling play such a central role in many types of epigenic inheritance, the word "epigenetics" is sometimes used as a synonym for these processes. However, this can be misleading. Chromatin remodeling is not always inherited, and not all epigenetic inheritance involves chromatin remodeling.

It has been suggested that the histone code could be mediated by the effect of small RNAs. The recent discovery and characterization of a vast array of small (21- to 26-nt), non-coding RNAs suggests that there is an RNA component, possibly involved in epigenetic gene regulation. Small interfering RNAs can modulate transcriptional gene expression via epigenetic modulation of targeted promoters.

RNA transcripts and their encoded proteins

Sometimes a gene, after being turned on, transcribes a product that (either directly or indirectly) maintains the activity of that gene. For example, Hnf4 and MyoD enhance the transcription of many liver- and muscle-specific genes, respectively, including their own, through the transcription factor activity of the proteins they encode. RNA signalling includes differential recruitment of a hierarchy of generic chromatin modifying complexes and DNA methyltransferases to specific loci by RNAs during differentiation and development. Other epigenetic changes are mediated by the production of different splice forms of RNA, or by formation of double-stranded RNA (RNAi). Descendants of the cell in which the gene was turned on will inherit this activity, even if the original stimulus for gene-activation is no longer present. These genes are most often turned on or off by signal transduction, although in some systems where syncytia or gap junctions are important, RNA may spread directly to other cells or nuclei by diffusion. A large amount of RNA and protein is contributed to the zygote by the mother during oogenesis or via nurse cells, resulting in maternal effect phenotypes. A smaller quantity of sperm RNA is transmitted from the father, but there is recent evidence that this epigenetic information can lead to visible changes in several generations of offspring.

Prions

Prions are infectious forms of proteins. Proteins generally fold into discrete units which perform distinct cellular functions, but some proteins are also capable of forming an infectious conformational state known as a prion. Although often viewed in the context of

infectious disease, prions are more loosely defined by their ability to catalytically convert other native state versions of the same protein to an infectious conformational state. It is in this latter sense that they can be viewed as epigenetic agents capable of inducing a phenotypic change without a modification of the genome.

Fungal prions are considered epigenetic because the infectious phenotype caused by the prion can be inherited without modification of the genome. PSI+ and URE3, discovered in yeast in 1965 and 1971, are the two best studied of this type of prion. Prions can have a phenotypic effect through the sequestration of protein in aggregates, thereby reducing that protein's activity. In PSI+ cells, the loss of the Sup35 protein (which is involved in termination of translation) causes ribosomes to have a higher rate of read-through of stop codons, an effect which results in suppression of nonsense mutations in other genes. The ability of Sup35 to form prions may be a conserved trait. It could confer an adaptive advantage by giving cells the ability to switch into a PSI+ state and express dormant genetic features normally terminated by premature stop codon mutations.

Structural inheritance systems

In ciliates such as *Tetrahymena* and *Paramecium*, genetically identical cells show heritable differences in the patterns of ciliary rows on their cell surface. Experimentally altered patterns can be transmitted to daughter cells. It seems existing structures act as templates for new structures. The mechanisms of such inheritance are unclear, but reasons exist to assume that multicellular organisms also use existing cell structures to assemble new ones.

Functions and consequences

Development

Somatic epigenetic inheritance, particularly through DNA methylation and chromatin remodeling, is very important in the development of multicellular eukaryotic organisms. The genome sequence is static (with some notable exceptions), but cells differentiate into many different types, which perform different functions, and respond differently to the environment and intercellular signalling. Thus, as individuals develop, morphogens activate or silence genes in an epigenetically heritable fashion, giving cells a "memory". In mammals, most cells terminally differentiate, with only stem cells retaining the ability to differentiate into several cell types ("totipotency" and "multipotency"). In mammals, some stem cells continue producing new differentiated cells throughout life, but mammals are not able to respond to loss of some tissues, for example, the inability to regenerate limbs, which some other animals are capable of. Unlike animals, plant cells do not terminally differentiate, remaining totipotent with the ability to give rise to a new individual plant. While plants do utilise many of the same epigenetic mechanisms as animals, such as chromatin remodeling, it has been hypothesised that plant cells do not have "memories", resetting their gene expression patterns at each cell division using positional information from the environment and surrounding cells to determine their fate.

Medicine

Epigenetics has many and varied potential medical applications. Congenital genetic disease is well understood, and it is also clear that epigenetics can play a role, for example, in the case of Angelman syndrome and Prader-Willi syndrome. These are normal genetic diseases caused by gene deletions or inactivation of the genes, but are unusually common because individuals are essentially hemizygous because of genomic imprinting, and therefore a single gene knock out is sufficient to cause the disease, where most cases would require both copies to be knocked out.

Evolution

Although epigenetics in multicellular organisms is generally thought to be a mechanism involved in differentiation, with epigenetic patterns "reset" when organisms reproduce, there have been some observations of transgenerational epigenetic inheritance (e.g., the phenomenon of paramutation observed in maize). Although most of these multigenerational epigenetic traits are gradually lost over several generations, the possibility remains that multigenerational epigenetics could be another aspect to evolution and adaptation. A sequestered germ line or Weismann barrier is specific to animals, and epigenetic inheritance is expected to be far more common in plants and microbes. These effects may require enhancements to the standard conceptual framework of the modern evolutionary synthesis.

Epigenetic features may play a role in short-term adaptation of species by allowing for reversible phenotype variability. The modification of epigenetic features associated with a region of DNA allows organisms, on a multigenerational time scale, to switch between phenotypes that express and repress that particular gene. When the DNA sequence of the region is not mutated, this change is reversible. It has also been speculated that organisms may take advantage of differential mutation rates associated with epigenetic features to control the mutation rates of particular genes. Interestingly, recent analysis have suggested that members of the APOBEC family of cytosine deaminases are capable of simultaneously mediating genetic and epigenetic inheritance using similar molecular mechanisms.

Epigenetic changes have also been observed to occur in response to environmental exposure—for example, mice given some dietary supplements have epigenetic changes affecting expression of the agouti gene, which affects their fur color, weight, and propensity to develop cancer.

More than 100 cases of transgenerational epigenetic inheritance phenomena have been reported in a wide range of organisms, including prokaryotes, plants, and animals.

Epigenetic effects in humans

Genomic imprinting and related disorders

Some human disorders are associated with genomic imprinting, a phenomenon in mammals where the father and mother contribute different epigenetic patterns for specific genomic loci in their germ cells. The best-known case of imprinting in human disorders is that of Angelman syndrome and Prader-Willi syndrome—both can be produced by the same genetic mutation, chromosome 15q partial deletion, and the particular syndrome that will develop depends on whether the mutation is inherited from the child's mother or from their father. This is due to the presence of genomic imprinting in the region. Beckwith-Wiedemann syndrome is also associated with genomic imprinting, often caused by abnormalities in maternal genomic imprinting of a region on chromosome 11.

Transgenerational epigenetic observations

Marcus Pembrey and colleagues also observed in the Överkalix study that the paternal (but not maternal) grandsons of Swedish boys who were exposed during preadolescence to famine in the 19th century were less likely to die of cardiovascular disease; if food was plentiful then diabetes mortality in the grandchildren increased, suggesting that this was a transgenerational epigenetic inheritance. The opposite effect was observed for females—the paternal (but not maternal) granddaughters of women who experienced famine while in the womb (and their eggs were being formed) lived shorter lives on average.

Cancer and developmental abnormalities

A variety of compounds are considered as epigenetic carcinogens—they result in an increased incidence of tumors, but they do not show mutagen activity (toxic compounds or pathogens that cause tumors incident to increased regeneration should also be excluded). Examples include diethylstilbestrol, arsenite, hexachlorobenzene, and nickel compounds.

Many teratogens exert specific effects on the fetus by epigenetic mechanisms. While epigenetic effects may preserve the effect of a teratogen such as diethylstilbestrol throughout the life of an affected child, the possibility of birth defects resulting from exposure of fathers or in second and succeeding generations of offspring has generally been rejected on theoretical grounds and for lack of evidence. However, a range of male-mediated abnormalities have been demonstrated, and more are likely to exist. FDA label information for Vidaza(tm), a formulation of 5-azacitidine (an unmethylatable analog of cytidine that causes hypomethylation when incorporated into DNA) states that "men should be advised not to father a child" while using the drug, citing evidence in treated male mice of reduced fertility, increased embryo loss, and abnormal embryo development. In rats, endocrine differences were observed in offspring of males exposed to morphine. In mice, second generation effects of diethylstilbesterol have been described occurring by epigenetic mechanisms.

Recent studies have shown that the Mixed Lineage Leukemia (MLL) gene causes leukemia by rearranging and fusing with other genes in different chromosomes, which is a process under epigenetic control.

Other investigations have concluded that alterations in histone acetylation and DNA methylation occur in various genes influencing prostate cancer.

In 2008, the National Institutes of Health announced that $190 million had been earmarked for epigenetics research over the next five years. In announcing the funding, government officials noted that epigenetics has the potential to explain mechanisms of aging, human development, and the origins of cancer, heart disease, mental illness, as well as several other conditions. Some investigators, like Randy Jirtle, PhD, of Duke University Medical Center, think epigenetics may ultimately turn out to have a greater role in disease than genetics.

Cancer Treatment

Current research has shown that epigenetic pharmaceuticals could be a putative replacement or adjuvant therapy for currently accepted treatment methods such as radiation and chemotherapy, or could enhance the effects of these current treatments. It has been shown that the epigenetic control of the proto-onco regions and the tumor suppressor sequences by conformational changes in histones directly affects the formation and progression of cancer Epigenetics also has the factor of reversibility, a characteristic that other cancer treatments do not offer.

Drug development has mainly focused on Histone Acetyltransferase (HAT) and Histone Deactylase (HDAC), including the introduction of the new pharmaceutical Vorinostat, a HDAC inhibitor, to the market. HDAC specifically has been shown to play an integral role in the progression of oral squamous cancer

Current front-runner candidates for new drug targets are Histone Lysine Methyltransferases (KMT) and Protein Arginine Methyltransferases (PRMT).

Twin studies

Recent studies involving both dizygotic and monozygotic twins have produced some evidence of epigenetic influence in humans.

Epigenetics in microorganisms

Bacteria make widespread use of postreplicative DNA methylation for the epigenetic control of DNA-protein interactions. Bacteria make use of DNA adenine methylation (rather than DNA cytosine methylation) as an epigenetic signal. DNA adenine methylation is important in bacteria virulence in organisms such as *Escherichia coli*, *Salmonella, Vibrio, Yersinia, Haemophilus*, and *Brucella*. In *Alphaproteobacteria*, methylation of adenine regulates the cell cycle and couples gene transcription to DNA

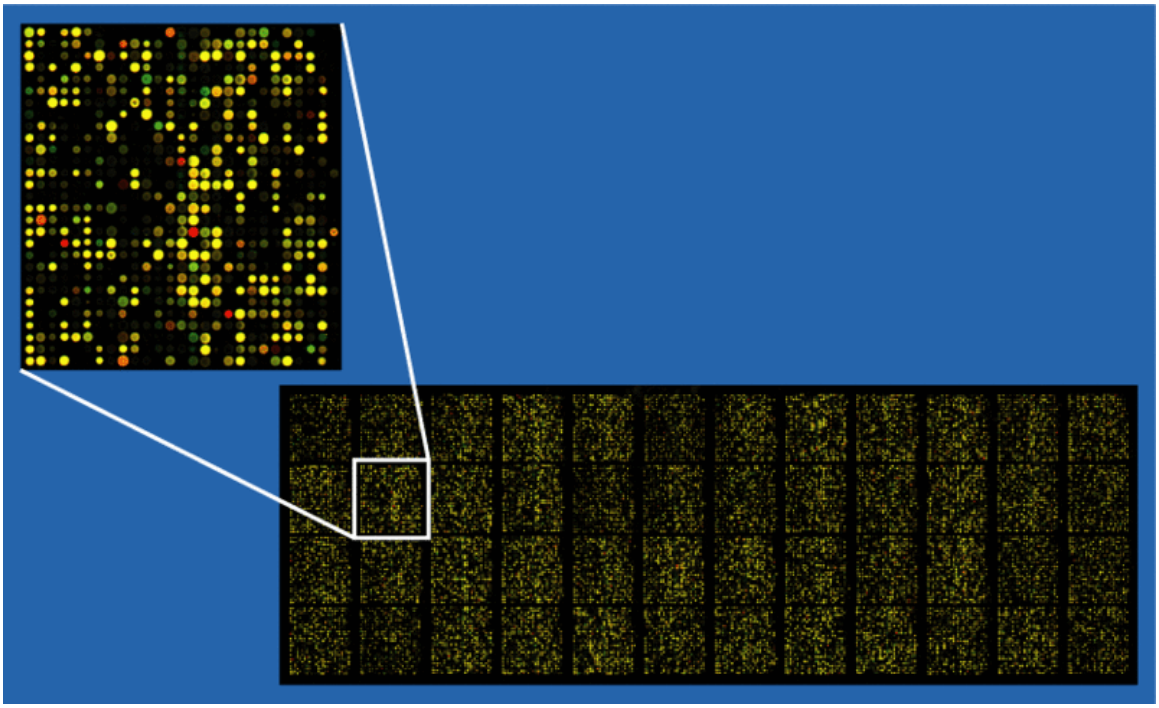
replication. In *Gammaproteobacteria*, adenine methylation provides signals for DNA replication, chromosome segregation, mismatch repair, packaging of bacteriophage, transposase activity and regulation of gene expression.

The filamentous fungus *Neurospora crassa* is a prominent model system for understanding the control and function of cytosine methylation. In this organisms, DNA methylation is associated with relics of a genome defense system called RIP (repeat-induced point mutation) and silences gene expression by inhibiting transcription elongation.

The yeast prion PSI is generated by a conformational change of a translation termination factor, which is then inherited by daughter cells. This can provide a survival advantage under adverse conditions. This is an example of epigenetic regulation enabling unicellular organisms to respond rapidly to environmental stress. Prions can be viewed as epigenetic agents capable of inducing a phenotypic change without modification of the genome.

Chapter- 3

DNA Microarray



Example of an approximately 40,000 probe spotted oligo microarray with enlarged inset to show detail.

A **DNA microarray** is a multiplex technology used in molecular biology. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles (10^{-12} moles) of a specific DNA sequence, known as *probes* (or *reporters*). These can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample (called *target*) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, silver-, or chemiluminescence-labeled targets to determine relative abundance of nucleic acid sequences in the target. Since an array can contain tens of

thousands of probes, a microarray experiment can accomplish many genetic tests in parallel. Therefore arrays have dramatically accelerated many types of investigation.

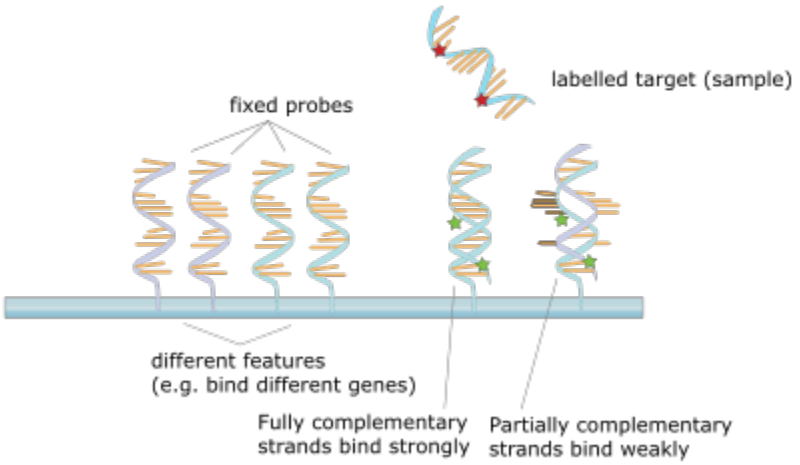
In standard microarrays, the probes are attached via surface engineering to a solid surface by a covalent bond to a chemical matrix (via epoxy-silane, amino-silane, lysine, polyacrylamide or others). The solid surface can be glass or a silicon chip, in which case they are colloquially known as an *Affy chip* when an Affymetrix chip is used. Other microarray platforms, such as Illumina, use microscopic beads, instead of the large solid support. DNA arrays are different from other types of microarray only in that they either measure DNA or use DNA as part of its detection system.

DNA microarrays can be used to measure changes in expression levels, to detect single nucleotide polymorphisms (SNPs), or to genotype or resequence mutant genomes. Microarrays also differ in fabrication, workings, accuracy, efficiency, and cost. Additional factors for microarray experiments are the experimental design and the methods of analyzing the data.

History

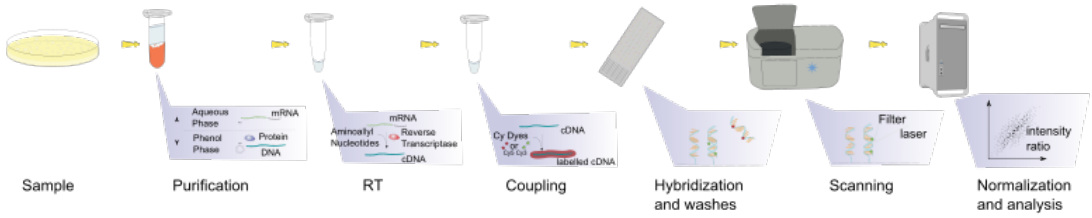
Microarray technology evolved from Southern blotting, where fragmented DNA is attached to a substrate and then probed with a known gene or fragment. Nucleic Acids Res. 1992 Apr 11;20(7):1679-84. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. Maskos U, Southern EM. The first reported use of this approach was the analysis of 378 arrayed lysed bacterial colonies each harboring a different sequence which were assayed in multiple replicas for expression of the genes in multiple normal and tumor tissue (Augenlicht and Kobrin, Cancer Research, 42, 1088–1093, 1982). This was expanded to analysis of more than 4000 human sequences with computer driven scanning and image processing for quantitative analysis of the sequences in human colonic tumors and normal tissue (Augenlicht *et al.*, Cancer Research, 47, 6017-6021, 1987) and then to comparison of colonic tissues at different genetic risk (Augenlicht *et al.*, Proceedings National Academy of Sciences, USA, 88, 3286-3289, 1991). The use of a collection of distinct DNAs in arrays for expression profiling was also described in 1987, and the arrayed DNAs were used to identify genes whose expression is modulated by interferon. These early gene arrays were made by spotting cDNAs onto filter paper with a pin-spotting device. The use of miniaturized microarrays for gene expression profiling was first reported in 1995, and a complete eukaryotic genome (*Saccharomyces cerevisiae*) on a microarray was published in 1997.

Principle



Hybridization of the target to the probe

The core principle behind microarrays is hybridization between two DNA strands, the property of complementary nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. A high number of complementary base pairs in a nucleotide sequence means tighter non-covalent bonding between the two strands. After washing off of non-specific bonding sequences, only strongly paired strands will remain hybridized. So fluorescently labeled target sequences that bind to a probe sequence generate a signal that depends on the strength of the hybridization determined by the number of paired bases, the hybridization conditions (such as temperature), and washing after hybridization. Total strength of the signal, from a spot (feature), depends upon the amount of target sample binding to the probes present on that spot. Microarrays use relative quantization in which the intensity of a feature is compared to the intensity of the same feature under a different condition, and the identity of the feature is known by its position. An alternative to microarrays is serial analysis of gene expression, where the transcriptome is sequenced allowing an absolute measurement.



The step required in a microarray experiment

Uses and types

Two Affymetrix chips

Many types of array exist and the broadest distinction is whether they are spatially arranged on a surface or on coded beads:

- The traditional solid-phase array is a collection of orderly microscopic "spots", called features, each with a specific probe attached to a solid surface, such as glass, plastic or silicon biochip (commonly known as a *genome chip*, *DNA chip* or *gene array*). Thousands of them can be placed in known locations on a single DNA microarray.
- The alternative bead array is a collection of microscopic polystyrene beads, each with a specific probe and a ratio of two or more dyes, which do not interfere with the fluorescent dyes used on the target sequence.

DNA microarrays can be used to detect DNA (as in comparative genomic hybridization), or detect RNA (most commonly as cDNA after reverse transcription) that may or may not

be translated into proteins. The process of measuring gene expression via cDNA is called expression analysis or expression profiling.
Applications include:

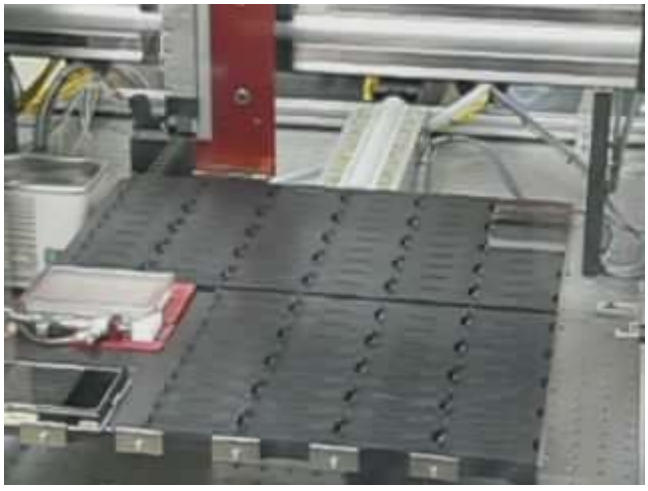
Application or technology	Synopsis
Gene expression profiling	In an mRNA or gene expression profiling experiment the expression levels of thousands of genes are simultaneously monitored to study the effects of certain treatments, diseases, and developmental stages on gene expression. For example, microarray-based gene expression profiling can be used to identify genes whose expression is changed in response to pathogens or other organisms by comparing gene expression in infected to that in uninfected cells or tissues.
Comparative genomic hybridization	Assessing genome content in different cells or closely related organisms.
GeneID	Small microarrays to check IDs of organisms in food and feed (like GMO), mycoplasms in cell culture, or pathogens for disease detection, mostly combining PCR and microarray technology.
Chromatin immunoprecipitation on Chip	DNA sequences bound to a particular protein can be isolated by immunoprecipitating that protein (ChIP), these fragments can be then hybridized to a microarray (such as a tiling array) allowing the determination of protein binding site occupancy throughout the genome. Example protein to immunoprecipitate are histone modifications (H3K27me3, H3K4me2, H3K9me3, etc.), Polycomb-group protein (PRC2:Suz12, PRC1:YY1) and trithorax-group protein (Ash1) to study the epigenetic landscape or RNA Polymerase II to study the transcription landscape.
DamID	Analogously to ChIP, genomic regions bound by a protein of interest can be isolated and used to probe a microarray to determine binding site occupancy. Unlike ChIP, DamID does not require antibodies but makes use of adenine methylation near the protein's binding sites to selectively amplify those regions, introduced by expressing minute amounts of protein of interest fused to bacterial DNA adenine methyltransferase.
SNP detection	Identifying single nucleotide polymorphism among alleles within or between populations. Several applications of microarrays make use of SNP detection, including Genotyping, forensic analysis, measuring predisposition to disease, identifying drug-candidates, evaluating germline mutations in individuals or somatic mutations in cancers, assessing loss of heterozygosity, or genetic linkage analysis.
Alternative splicing detection	An '*exon junction array* design uses probes specific to the expected or potential splice sites of predicted exons for a gene. It

	is of intermediate density, or coverage, to a typical gene expression array (with 1-3 probes per gene) and a genomic tiling array (with hundreds or thousands of probes per gene). It is used to assay the expression of alternative splice forms of a gene. Exon arrays have a different design, employing probes designed to detect each individual exon for known or predicted genes, and can be used for detecting different splicing isoforms.
Fusion genes microarray	A Fusion gene microarray can detect fusion transcripts, *e.g.* from cancer specimens. The principle behind this is building on the alternative splicing microarrays. The oligo design strategy enables combined measurements of chimeric transcript junctions with exon-wise measurements of individual fusion partners.
Tiling array	Genome tiling arrays consist of overlapping probes designed to densely represent a genomic region of interest, sometimes as large as an entire human chromosome. The purpose is to empirically detect expression of transcripts or alternatively splice forms which may not have been previously known or predicted.

Fabrication

Microarrays can be manufactured in different ways, depending on the number of probes under examination, costs, customization requirements, and the type of scientific question being asked. Arrays may have as few as 10 probes or up to 2.1 million micrometre-scale probes from commercial vendors.

Spotted vs. in situ synthesised arrays



A DNA microarray being printed by a robot at the University of Delaware

Microarrays can be fabricated using a variety of technologies, including printing with fine-pointed pins onto glass slides, photolithography using pre-made masks,

photolithography using dynamic micromirror devices, ink-jet printing, or electrochemistry on microelectrode arrays.

In *spotted microarrays*, the probes are oligonucleotides, cDNA or small fragments of PCR products that correspond to mRNAs. The probes are synthesized prior to deposition on the array surface and are then "spotted" onto glass. A common approach utilizes an array of fine pins or needles controlled by a robotic arm that is dipped into wells containing DNA probes and then depositing each probe at designated locations on the array surface. The resulting "grid" of probes represents the nucleic acid profiles of the prepared probes and is ready to receive complementary cDNA or cRNA "targets" derived from experimental or clinical samples. This technique is used by research scientists around the world to produce "in-house" printed microarrays from their own labs. These arrays may be easily customized for each experiment, because researchers can choose the probes and printing locations on the arrays, synthesize the probes in their own lab (or collaborating facility), and spot the arrays. They can then generate their own labeled samples for hybridization, hybridize the samples to the array, and finally scan the arrays with their own equipment. This provides a relatively low-cost microarray that may be customized for each study, and avoids the costs of purchasing often more expensive commercial arrays that may represent vast numbers of genes that are not of interest to the investigator. Publications exist which indicate in-house spotted microarrays may not provide the same level of sensitivity compared to commercial oligonucleotide arrays, possibly owing to the small batch sizes and reduced printing efficiencies when compared to industrial manufactures of oligo arrays.

In *oligonucleotide microarrays*, the probes are short sequences designed to match parts of the sequence of known or predicted open reading frames. Although oligonucleotide probes are often used in "spotted" microarrays, the term "oligonucleotide array" most often refers to a specific technique of manufacturing. Oligonucleotide arrays are produced by printing short oligonucleotide sequences designed to represent a single gene or family of gene splice-variants by synthesizing this sequence directly onto the array surface instead of depositing intact sequences. Sequences may be longer (60-mer probes such as the Agilent design) or shorter (25-mer probes produced by Affymetrix) depending on the desired purpose; longer probes are more specific to individual target genes, shorter probes may be spotted in higher density across the array and are cheaper to manufacture. One technique used to produce oligonucleotide arrays include photolithographic synthesis (Agilent and Affymetrix) on a silica substrate where light and light-sensitive masking agents are used to "build" a sequence one nucleotide at a time across the entire array. Each applicable probe is selectively "unmasked" prior to bathing the array in a solution of a single nucleotide, then a masking reaction takes place and the next set of probes are unmasked in preparation for a different nucleotide exposure. After many repetitions, the sequences of every probe become fully constructed. More recently, Maskless Array Synthesis from NimbleGen Systems has combined flexibility with large numbers of probes.

Two-channel vs. one-channel detection

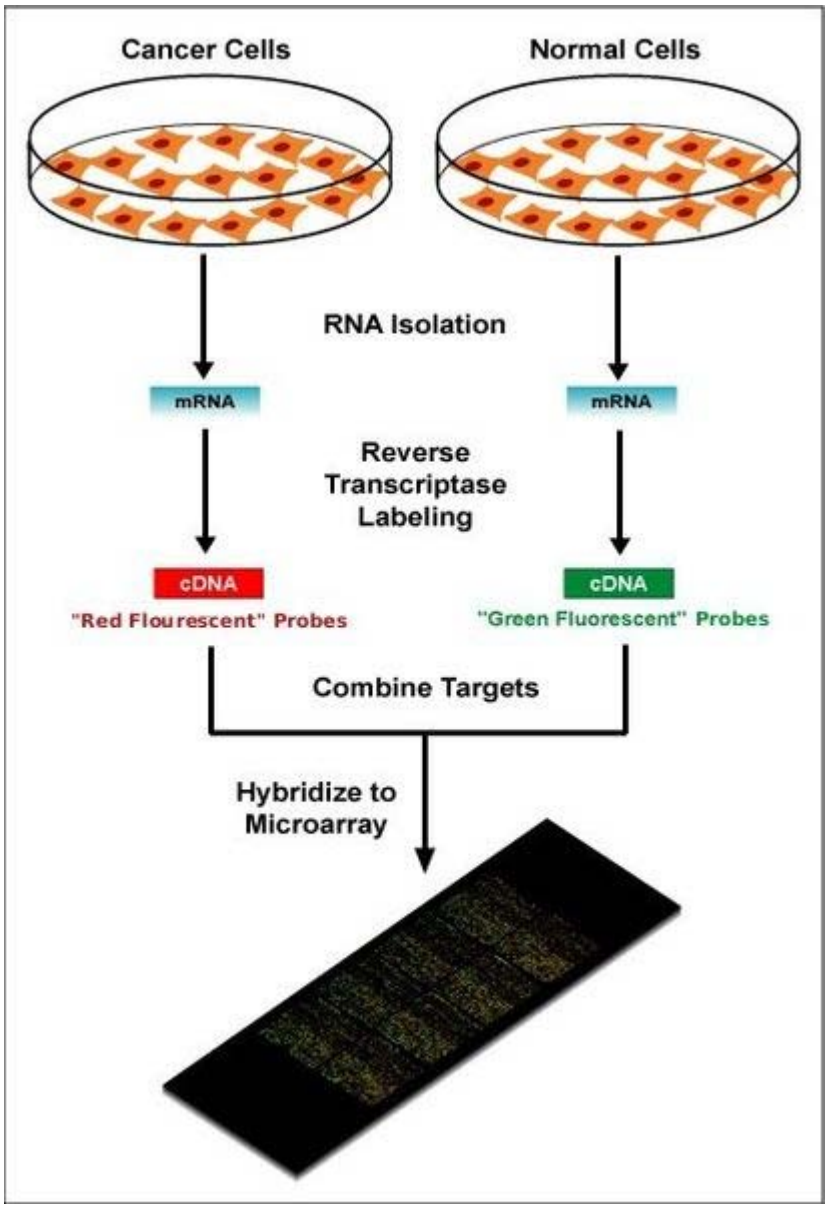


Diagram of typical dual-colour microarray experiment.

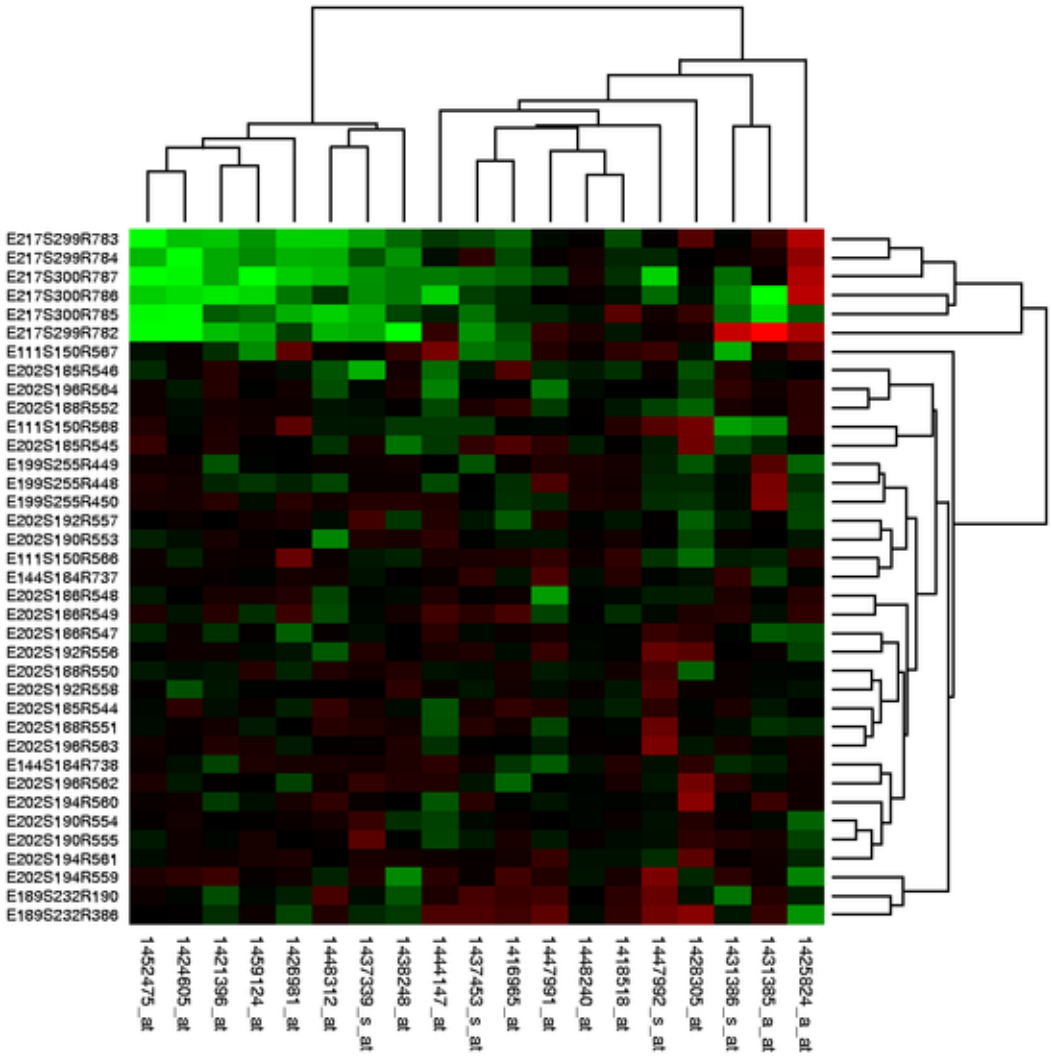
Two-color microarrays or *two-channel microarrays* are typically hybridized with cDNA prepared from two samples to be compared (e.g. diseased tissue versus healthy tissue) and that are labeled with two different fluorophores. Fluorescent dyes commonly used for cDNA labeling include Cy3, which has a fluorescence emission wavelength of 570 nm (corresponding to the green part of the light spectrum), and Cy5 with a fluorescence emission wavelength of 670 nm (corresponding to the red part of the light spectrum). The two Cy-labeled cDNA samples are mixed and hybridized to a single microarray that is then scanned in a microarray scanner to visualize fluorescence of the two fluorophores after excitation with a laser beam of a defined wavelength. Relative intensities of each

fluorophore may then be used in ratio-based analysis to identify up-regulated and down-regulated genes.

Oligonucleotide microarrays often carry control probes designed to hybridize with RNA spike-ins. The degree of hybridization between the spike-ins and the control probes is used to normalize the hybridization measurements for the target probes. Although absolute levels of gene expression may be determined in the two-color array in rare instances, the relative differences in expression among different spots within a sample and between samples is the preferred method of data analysis for the two-color system. Examples of providers for such microarrays includes Agilent with their Dual-Mode platform, Eppendorf with their DualChip platform for colorimetric Silverquant labeling, and TeleChem International with Arrayit.

In *single-channel microarrays* or *one-color microarrays*, the arrays provide intensity data for each probe or probe set indicating a relative level of hybridization with the labeled target. However, they do not truly indicate abundance levels of a gene but rather relative abundance when compared to other samples or conditions when processed in the same experiment. Each RNA molecule encounters protocol and batch-specific bias during amplification, labeling, and hybridization phases of the experiment making comparisons between genes for the same microarray uninformative. The comparison of two conditions for the same gene requires two separate single-dye hybridizations. Several popular single-channel systems are the Affymetrix "Gene Chip", Illumina "Bead Chip", Agilent single-channel arrays, the Applied Microarrays "CodeLink" arrays, and the Eppendorf "DualChip & Silverquant". One strength of the single-dye system lies in the fact that an aberrant sample cannot affect the raw data derived from other samples, because each array chip is exposed to only one sample (as opposed to a two-color system in which a single low-quality sample may drastically impinge on overall data precision even if the other sample was of high quality). Another benefit is that data are more easily compared to arrays from different experiments so long as batch effects have been accounted for. A drawback to the one-color system is that, when compared to the two-color system, twice as many microarrays are needed to compare samples within an experiment.

Microarrays and bioinformatics



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

The advent of inexpensive microarray experiments created several specific bioinformatics challenges:

- the multiple levels of replication in experimental design (Experimental design)
- the number of platforms and independent groups and data format (Standardization)
- the treatment of the data (Statistical analysis)
- accuracy and precision (Relation between probe and gene)
- the sheer volume of data and the ability to share it (Data warehousing)

Experimental design

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

There are three main elements to consider when designing a microarray experiment. First, replication of the biological samples is essential for drawing conclusions from the experiment. Second, technical replicates (two RNA samples obtained from each experimental unit) help to ensure precision and allow for testing differences within treatment groups. The technical replicates may be two independent RNA extractions or two aliquots of the same extraction. Third, spots of each cDNA clone or oligonucleotide are present as replicates (at least duplicates) on the microarray slide, to provide a measure of technical precision in each hybridization. It is critical that information about the sample preparation and handling is discussed, in order to help identify the independent units in the experiment and to avoid inflated estimates of statistical significance.

Standardization

Microarray data is difficult to exchange due to the lack of standardization in platform fabrication, assay protocols, and analysis methods. This presents an interoperability problem in bioinformatics. Various grass-roots open-source projects are trying to ease the exchange and analysis of data produced with non-proprietary chips:

- For example, the "Minimum Information About a Microarray Experiment" (MIAME) checklist helps define the level of detail that should exist and is being adopted by many journals as a requirement for the submission of papers incorporating microarray results. But MIAME does not describe the format for the information, so while many formats can support the MIAME requirements, as of 2007 no format permits verification of complete semantic compliance.
- The "MicroArray Quality Control (MAQC) Project" is being conducted by the US Food and Drug Administration (FDA) to develop standards and quality control metrics which will eventually allow the use of MicroArray data in drug discovery, clinical practice and regulatory decision-making.
- The MGED Society has developed standards for the representation of gene expression experiment results and relevant annotations.

Statistical analysis

Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. Statistical challenges include taking into account effects of background noise and appropriate normalization of the data. Normalization methods may be suited to specific platforms and, in the case of commercial platforms, the analysis may be proprietary. Algorithms that affect statistical analysis include:

- Image analysis: gridding, spot recognition of the scanned image (segmentation algorithm), removal or marking of poor-quality and low-intensity features (called *flagging*).
- Data processing: background subtraction (based on global or local background), determination of spot intensities and intensity ratios, visualisation of data, and log-transformation of ratios, global or local normalization of intensity ratios.
- Identification of statistically significant changes: t-test, ANOVA, Bayesian method Mann–Whitney test methods tailored to microarray data sets, which take into account multiple comparisons or Cluster Analysis. These methods assess statistical power based on the variation present in the data and the number of experimental replicates, and can help minimize Type I and type II errors in the analyses.
- Network-based methods: Statistical methods that take the underlying structure of gene networks into account, representing either associative or causative interactions or dependencies among gene products.

Microarray data may require further processing aimed at reducing the dimensionality of the data to aid comprehension and more focused analysis. Other methods permit analysis of data consisting of a low number of biological or technical replicates; for example, the Local Pooled Error (LPE) test pools standard deviations of genes with similar expression levels in an effort to compensate for insufficient replication.

Relation between probe and gene

The relation between a probe and the mRNA that it is expected to detect is problematic. On the one hand, some mRNAs may cross-hybridize probes in the array that are supposed to detect another mRNA. In addition, mRNAs may experience amplification bias that is sequence or molecule-specific. On the other hand, probes that are designed to detect the mRNA of a particular gene may be relying on genomic EST information that is incorrectly associated with that gene.

Data warehousing

Microarray data was found to be more useful when compared to other similar datasets. The sheer volume (in bytes), specialized formats (such as MIAME), and curation efforts associated with the datasets require specialized databases to store the data.

Chapter- 4

Proteomics

Robotic preparation of MALDI mass spectrometry samples on a sample carrier.

Proteomics is the large-scale study of proteins, particularly their structures and functions. Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. The term "proteomics" was first coined in 1997 to make an analogy with genomics, the study of the genes. The word "proteome" is a blend of "**prote**in" and "gen**ome**", and was coined by Marc Wilkins in 1994 while working on the concept as a PhD student. The proteome is the entire complement of

proteins, including the modifications made to a particular set of proteins, produced by an organism or system. This will vary with time and distinct requirements, or stresses, that a cell or organism undergoes.

Complexity of the problem

After genomics, proteomics is considered the next step in the study of biological systems. It is much more complicated than genomics mostly because while an organism's genome is more or less constant, the proteome differs from cell to cell and from time to time. This is because distinct genes are expressed in distinct cell types. This means that even the basic set of proteins which are produced in a cell needs to be determined.

In the past this was done by mRNA analysis, but this was found not to correlate with protein content. It is now known that mRNA is not always translated into protein, and the amount of protein produced for a given amount of mRNA depends on the gene it is transcribed from and on the current physiological state of the cell. Proteomics confirms the presence of the protein and provides a direct measure of the quantity present.

Post-translational modifications

Not only does the translation from mRNA cause differences, many proteins are also subjected to a wide variety of chemical modifications after translation. Many of these post-translational modifications are critical to the protein's function.

Phosphorylation

One such modification is phosphorylation, which happens to many enzymes and structural proteins in the process of cell signaling. The addition of a phosphate to particular amino acids—most commonly serine and threonine mediated by serine/threonine kinases, or more rarely tyrosine mediated by tyrosine kinases—causes a protein to become a target for binding or interacting with a distinct set of other proteins that recognize the phosphorylated domain.

Because protein phosphorylation is one of the most-studied protein modifications, many "proteomic" efforts are geared to determining the set of phosphorylated proteins in a particular cell or tissue-type under particular circumstances. This alerts the scientist to the signaling pathways that may be active in that instance.

Ubiquitination

Ubiquitin is a small protein that can be affixed to certain protein substrates by enzymes called E3 ubiquitin ligases. Determining which proteins are poly-ubiquitinated can be helpful in understanding how protein pathways are regulated. This is therefore an additional legitimate "proteomic" study. Similarly, once it is determined what substrates are ubiquitinated by each ligase, determining the set of ligases expressed in a particular cell type will be helpful.

Additional modifications

Listing all the protein modifications that might be studied in a "Proteomics" project would require a discussion of most of biochemistry; therefore, a short list will serve here to illustrate the complexity of the problem. In addition to phosphorylation and ubiquitination, proteins can be subjected to (among others) methylation, acetylation, glycosylation, oxidation and nitrosylation. Some proteins undergo ALL of these modifications, often in time-dependent combinations, aptly illustrating the potential complexity one has to deal with when studying protein structure and function.

Distinct proteins are made under distinct settings

Even if one is studying a particular cell type, that cell may make different sets of proteins at different times, or under different conditions. Furthermore, as mentioned, any one protein can undergo a wide range of post-translational modifications.

Therefore a "proteomics" study can become quite complex very quickly, even if the object of the study is very restricted. In more ambitious settings, such as when a biomarker for a tumor is sought - when the proteomics scientist is obliged to study sera samples from multiple cancer patients - the amount of complexity that must be dealt with is as great as in any modern biological project.

Limitations to genomic study

Scientists are very interested in proteomics because it gives a much better understanding of an organism than genomics. First, the level of transcription of a gene gives only a rough estimate of its level of expression into a protein. An mRNA produced in abundance may be degraded rapidly or translated inefficiently, resulting in a small amount of protein. Second, as mentioned above many proteins experience post-translational modifications that profoundly affect their activities; for example some proteins are not active until they become phosphorylated. Methods such as phosphoproteomics and glycoproteomics are used to study post-translational modifications. Third, many transcripts give rise to more than one protein, through alternative splicing or alternative post-translational modifications. Fourth, many proteins form complexes with other proteins or RNA molecules, and only function in the presence of these other molecules. Finally, protein degradation rate plays an important role in protein content.

Methods of studying proteins

Determining proteins which are post-translationally modified

One way in which a particular protein can be studied is to develop an antibody which is specific to that modification. For example, there are antibodies which only recognize certain proteins when they are tyrosine-phosphorylated, known as phospho-specific antibodies; also, there are antibodies specific to other modifications. These can be used to determine the set of proteins that have undergone the modification of interest.

For sugar modifications, such as glycosylation of proteins, certain lectins have been discovered which bind sugars. These too can be used.

A more common way to determine post-translational modification of interest is to subject a complex mixture of proteins to electrophoresis in "two-dimensions", which simply means that the proteins are electrophoresed first in one direction, and then in another... this allows small differences in a protein to be visualized by separating a modified protein from its unmodified form. This methodology is known as "two-dimensional gel electrophoresis".

Recently, another approach has been developed called PROTOMAP which combines SDS-PAGE with shotgun proteomics to enable detection of changes in gel-migration such as those caused by proteolysis or post translational modification.

Determining the existence of proteins in complex mixtures

Classically, antibodies to particular proteins or to their modified forms have been used in biochemistry and cell biology studies. These are among the most common tools used by practicing biologists today.

For more quantitative determinations of protein amounts, techniques such as ELISAs can be used.

For proteomic study, more recent techniques such as matrix-assisted laser desorption/ionization (MALDI) have been employed for rapid determination of proteins in particular mixtures and increasingly electrospray ionization (ESI).

Computational methods in studying protein biomarkers

Computational predictive models have shown that extensive and diverse feto-maternal protein trafficking occurs during pregnancy and can be readily detected non-invasively in maternal whole blood. This computational approach circumvented a major limitation, the abundance of maternal proteins interfering with the detection of fetal proteins, to fetal proteomic analysis of maternal blood. Computational models can use fetal gene transcripts previously identified in maternal whole blood to create a comprehensive proteomic network of the term neonate. Such work shows that the fetal proteins detected in pregnant woman’s blood originate from a diverse group of tissues and organs from the developing fetus. The proteomic networks contain many biomarkers that are proxies for development and illustrate the potential clinical application of this technology as a way to monitor normal and abnormal fetal development.

An information theoretic framework has also been introduced for biomarker discovery, integrating biofluid and tissue information. This new approach takes advantage of functional synergy between certain biofluids and tissues with the potential for clinically significant findings not possible if tissues and biofluids were considered individually. By conceptualizing tissue-biofluid as information channels, significant biofluid proxies can

be identified and then used for guided development of clinical diagnostics. Candidate biomarkers are then predicted based on information transfer criteria across the tissue-biofluid channels. Significant biofluid-tissue relationships can be used to prioritize clinical validation of biomarkers.

Establishing protein-protein interactions

Most proteins function in collaboration with other proteins, and one goal of proteomics is to identify which proteins interact. This is especially useful in determining potential partners in cell signaling cascades.

Several methods are available to probe protein-protein interactions. The traditional method is yeast two-hybrid analysis. New methods include protein microarrays, immunoaffinity chromatography followed by mass spectrometry, dual polarisation interferometry, Microscale Thermophoresis and experimental methods such as phage display and computational methods

Practical applications of proteomics

One of the most promising developments to come from the study of human genes and proteins has been the identification of potential new drugs for the treatment of disease. This relies on genome and proteome information to identify proteins associated with a disease, which computer software can then use as targets for new drugs. For example, if a certain protein is implicated in a disease, its 3D structure provides the information to design drugs to interfere with the action of the protein. A molecule that fits the active site of an enzyme, but cannot be released by the enzyme, will inactivate the enzyme. This is the basis of new drug-discovery tools, which aim to find new drugs to inactivate proteins involved in disease. As genetic differences among individuals are found, researchers expect to use these techniques to develop personalized drugs that are more effective for the individual.

Biomarkers

The FDA defines a biomarker as, “A characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention”.

Understanding the proteome, the structure and function of each protein and the complexities of protein-protein interactions will be critical for developing the most effective diagnostic techniques and disease treatments in the future.

An interesting use of proteomics is using specific protein biomarkers to diagnose disease. A number of techniques allow to test for proteins produced during a particular disease, which helps to diagnose the disease quickly. Techniques include western blot, immunohistochemical staining, enzyme linked immunosorbent assay (ELISA) or mass spectrometry.

Current research methodologies

There are many approaches to attempting to characterize the human proteome, which is estimated to exceed 100,000 unique forms, 25,000 genes plus post-translational modifications.

Chapter- 5

Metabolomics

Metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind" - specifically, the study of their small-molecule metabolite profiles. The metabolome represents the collection of all metabolites in a biological cell, tissue, organ or organism, which are the end products of cellular processes. Thus, while mRNA gene expression data and proteomic analyses do not tell the whole story of what might be happening in a cell, metabolic profiling can give an instantaneous snapshot of the physiology of that cell. One of the challenges of systems biology and functional genomics is to integrate proteomic, transcriptomic, and metabolomic information to give a more complete picture of living organisms.

Metabolomics is the scientific study of chemical processes involving metabolites.

Origins

The idea that biological fluids reflect the health of an individual has existed for a long time. Ancient Chinese doctors used ants for the evaluation of urine of patients to detect whether the urine contained high levels of glucose, and hence detect diabetes. In the middle ages, “urine charts” were used to link the colours, tastes and smells of urine to various medical conditions, which are metabolic in origin.

The concept that individuals might have a “metabolic profile” that could be reflected in the makeup of their biological fluids was introduced by Roger Williams in the late 1940s, who used paper chromatography to suggest characteristic metabolic patterns in urine and saliva were associated with diseases such as schizophrenia. However, it was only through technological advancements in the 1960s and 1970s that it became feasible to quantitatively (as opposed to qualitatively) measure metabolic profiles. The term “metabolic profile” was introduced by Horning, et. al. in 1971 after they demonstrated that GC-MS could be used to measure compounds present in human urine and tissue extracts. The Horning group, along with that of Linus Pauling and Arthur Robinson led the development of GC-MS methods to monitor the metabolites present in urine through the 1970s.

Concurrently, NMR spectroscopy, which was discovered in the 1940s, was also undergoing rapid advances. In 1974, Seeley et al. demonstrated the utility of using NMR to detect metabolites in unmodified biological samples. This first study on muscle highlighted the value of NMR in that it was determined that 90% of cellular ATP is complexed with magnesium. As sensitivity has improved with the evolution of higher magnetic field strengths and magic-angle spinning, NMR continues to be a leading analytical tool to investigate metabolism. Recent efforts to utilize NMR for metabolomics have been largely driven by the laboratory of Dr. Jeremy Nicholson at Birkbeck College, University of London and later at Imperial College London. In 1984, Nicholson showed ^{1}H NMR spectroscopy could potentially be used to diagnose and treat diabetes mellitus, and later pioneered the application of pattern recognition methods to NMR spectroscopic data.

In 2005, the first metabolomics web database, METLIN, for characterizing human metabolites was developed in the Siuzdak laboratory at The Scripps Research Institute and contained over 5,000 metabolites and tandem mass spectral data. As of 2011, METLIN contains over 40,000 metabolites as well as the largest repository of tandem mass spectrometry data in metabolomics.

On 23 January 2007, the Human Metabolome Project, led by Dr. David Wishart of the University of Alberta, Canada, completed the first draft of the human metabolome, consisting of a database of approximately 2500 metabolites, 1200 drugs and 3500 food components. Similar projects have been underway in several plant species, most notably *Medicago truncatula* and *Arabidopsis thaliana* for several years.

As late as mid-2010, metabolomics was still considered an "emerging field". Further, it was noted that further progress in the field depended in large part, through addressing otherwise "irresolvable technical challenges", by technical evolution of mass spectrometry instrumentation.

Metabolome

Metabolome refers to the complete set of small-molecule metabolites (such as metabolic intermediates, hormones and other signaling molecules, and secondary metabolites) to be found within a biological sample, such as a single organism. The word was coined in analogy with transcriptomics and proteomics; like the transcriptome and the proteome, the metabolome is dynamic, changing from second to second. Although the metabolome can be defined readily enough, it is not currently possible to analyse the entire range of metabolites by a single analytical method. The first metabolite database(called METLIN) for searching m/z values from mass spectrometry data was developed by scientists at The Scripps Research Institute in 2005. In January 2007, scientists at the University of Alberta and the University of Calgary completed the first draft of the human metabolome. They catalogued approximately 2500 metabolites, 1200 drugs and 3500 food components that can be found in the human body, as reported in the literature. This information, available at the Human Metabolome Database and based on analysis of information available in the current scientific literature, is far from complete. In contrast, much more

is known about the metabolomes of other organisms. For example, over 50,000 metabolites have been characterized from the plant kingdom, and many thousands of metabolites have been identified and/or characterized from single plants..

Metabolites

Metabolites are the intermediates and products of metabolism. Within the context of metabolomics, a metabolite is usually defined as any molecule less than 1 kDa in size. However, there are exceptions to this depending on the sample and detection method. For example, macromolecules such as lipoproteins and albumin are reliably detected in NMR-based metabolomics studies of blood plasma. In plant-based metabolomics, it is common to refer to "primary" and "secondary" metabolites. A primary metabolite is directly involved in the normal growth, development, and reproduction. A secondary metabolite is not directly involved in those processes, but usually has important ecological function. Examples include antibiotics and pigments. By contrast, in human-based metabolomics, it is more common to describe metabolites as being either endogenous (produced by the host organism) or exogenous. Metabolites of foreign substances such as drugs are termed xenometabolites.

The metabolome forms a large network of metabolic reactions, where outputs from one enzymatic chemical reaction are inputs to other chemical reactions. Such systems have been described as hypercycles.

Metabonomics

Metabonomics is defined as "the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification". The word origin is from the Greek *meta* meaning change and *nomos* meaning a rule set or set of laws. This approach was pioneered by Jeremy Nicholson at Imperial College London and has been used in toxicology, disease diagnosis and a number of other fields. Historically, the metabonomics approach was one of the first methods to apply the scope of systems biology to studies of metabolism.

There has been some disagreement over the exact differences between 'metabolomics' and 'metabonomics'. The difference between the two terms is not related to choice of analytical platform: although metabonomics is more associated with NMR spectroscopy and metabolomics with mass spectrometry-based techniques, this is simply because of usages amongst different groups that have popularized the different terms. While there is still no absolute agreement, there is a growing consensus that 'metabolomics' places a greater emphasis on metabolic profiling at a cellular or organ level and is primarily concerned with normal endogenous metabolism. 'Metabonomics' extends metabolic profiling to include information about perturbations of metabolism caused by environmental factors (including diet and toxins), disease processes, and the involvement of extragenomic influences, such as gut microflora. This is not a trivial difference; metabolomic studies should, by definition, exclude metabolic contributions from extragenomic sources, because these are external to the system being studied. However,

in practice, within the field of human disease research there is still a large degree of overlap in the way both terms are used, and they are often in effect synonymous.

Analytical technologies

Separation methods

- Gas chromatography, especially when interfaced with mass spectrometry (GC-MS), is one of the most widely used and powerful methods. It offers very high chromatographic resolution, but requires chemical derivatization for many biomolecules: only volatile chemicals can be analysed without derivatization. (Some modern instruments allow '2D' chromatography, using a short polar column after the main analytical column, which increases the resolution still further.) Some large and polar metabolites cannot be analysed by GC.

- High performance liquid chromatography (HPLC). Compared to GC, HPLC has lower chromatographic resolution, but it does have the advantage that a much wider range of analytes can potentially be measured.

- Capillary electrophoresis (CE). CE has a higher theoretical separation efficiency than HPLC, and is suitable for use with a wider range of metabolite classes than is GC. As for all electrophoretic techniques, it is most appropriate for charged analytes.

Detection methods

- Mass spectrometry (MS) is used to identify and to quantify metabolites after separation by GC, HPLC (LC-MS), or CE. GC-MS is the most 'natural' combination of the three, and was the first to be developed. In addition, mass spectral fingerprint libraries exist or can be developed that allow identification of a metabolite according to its fragmentation pattern. MS is both sensitive (although, particularly for HPLC-MS, sensitivity is more of an issue as it is affected by the charge on the metabolite, and can be subject to ion suppression artifacts) and can be very specific. There are also a number of studies which use MS as a stand-alone technology: the sample is infused directly into the mass spectrometer with no prior separation, and the MS serves to both separate and to detect metabolites.

- Surface-based mass analysis has seen a resurgence in the past decade, with new MS technologies focused on increasing sensitivity, minimizing background, and reducing sample preparation. The ability to analyze metabolites directly from biofluids and tissues continues to challenge current MS technology, largely because of the limits imposed by the complexity of these samples, which contain thousands to tens of thousands of metabolites. Among the technologies being developed to address this challenge is Nanostructure-Initiator MS (NIMS), a desorption/ ionization approach that does not require the application of matrix and

thereby facilitates small-molecule (i.e., metabolite) identification. MALDI is also used however, the application of a MALDI matrix can add significant background at <1000 Da that complicates analysis of the low-mass range (i.e., metabolites). In addition, the size of the resulting matrix crystals limits the spatial resolution that can be achieved in tissue imaging. Because of these limitations, several other matrix-free desorption/ionization approaches have been applied to the analysis of biofluids and tissues. Secondary ion mass spectrometry (SIMS) was one of the first matrix-free desorption/ionization approaches used to analyze metabolites from biological samples. SIMS uses a high-energy primary ion beam to desorb and generate secondary ions from a surface. The primary advantage of SIMS is its high spatial resolution (as small as 50 nm), a powerful characteristic for tissue imaging with MS. However, SIMS has yet to be readily applied to the analysis of biofluids and tissues because of its limited sensitivity at >500 Da and analyte fragmentation generated by the high-energy primary ion beam. Desorption electrospray ionization (DESI) is a matrix-free technique for analyzing biological samples that uses a charged solvent spray to desorb ions from a surface. Advantages of DESI are that no special surface is required and the analysis is performed at ambient pressure with full access to the sample during acquisition. A limitation of DESI is spatial resolution because "focusing" the charged solvent spray is difficult. However, a recent development termed laser ablation ESI (LAESI) is a promising approach to circumvent this limitation.

- Nuclear magnetic resonance (NMR) spectroscopy. NMR is the only detection technique which does not rely on separation of the analytes, and the sample can thus be recovered for further analyses. All kinds of small molecule metabolites can be measured simultaneously - in this sense, NMR is close to being a universal detector. The main advantages of NMR are high analytical reproducibility and simplicity of sample preparation. Practically, however, it is relatively insensitive compared to mass spectrometry-based techniques.

- Although NMR and MS are the most widely used techniques, other methods of detection that have been used include electrochemical detection (coupled to HPLC) and radiolabel (when combined with thin-layer chromatography).

Statistical methods

The data generated in metabolomics usually consist of measurements performed on subjects under various conditions. These measurements may be digitized spectra, or a list of metabolite levels. In its simplest form this generates a matrix with rows corresponding to subjects and columns corresponding to metabolite levels. Several statistical programs are currently available for analysis of both NMR and mass spectrometry data. For mass spectrometry data, software is available that identifies molecules that vary in subject groups on the basis of mass and sometimes retention time depending on the experimental design. The first comprehensive software to analyze global mass spectrometry-based metabolomics datasets was developed by the Siuzdak laboratory at The Scripps Research Institute in 2006. This software, called XCMS, is freely available, has over 20,000

downloads since its inception in 2006, and is one of the most widely cited mass spectrometry-based metabolomics software programs in scientific literature. Other popular metabolomics programs for mass spectral analysis are MZmine, MetAlign, MathDAMP, which also compensate for retention time deviation during sample analysis. LCMStats is another R package for detailed analysis of liquid chromatography mass spectrometry(LCMS)data and is helpful in identification of co-eluting ions especially isotopologues from a complicated metabolic profile. It combines xcms package functions and can be used to apply many statistical functions for correcting detector saturation using coates correction and creating heat plots. Metabolomics data may also be analyzed by statistical projection (chemometrics) methods such as principal components analysis and partial least squares regression.

Key applications

- Toxicity assessment/toxicology. Metabolic profiling (especially of urine or blood plasma samples) can be used to detect the physiological changes caused by toxic insult of a chemical (or mixture of chemicals). In many cases, the observed changes can be related to specific syndromes, e.g. a specific lesion in liver or kidney. This is of particular relevance to pharmaceutical companies wanting to test the toxicity of potential drug candidates: if a compound can be eliminated before it reaches clinical trials on the grounds of adverse toxicity, it saves the enormous expense of the trials.

- Functional genomics. Metabolomics can be an excellent tool for determining the phenotype caused by a genetic manipulation, such as gene deletion or insertion. Sometimes this can be a sufficient goal in itself—for instance, to detect any phenotypic changes in a genetically-modified plant intended for human or animal consumption. More exciting is the prospect of predicting the function of unknown genes by comparison with the metabolic perturbations caused by deletion/insertion of known genes. Such advances are most likely to come from model organisms such as *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. The Cravatt laboratory at The Scripps Research Institute has recently applied this technology to mammalian systems, identifying the *N*-acyltaurines as previously uncharacterized endogenous substrates for the enzyme fatty acid amide hydrolase (FAAH) and the monoalkylglycerol ethers as endogenous substrates for the uncharacterized hydrolase KIAA1363.

- Nutrigenomics is a generalised term which links genomics, transcriptomics, proteomics and metabolomics to human nutrition. In general a metabolome in a given body fluid is influenced by endogenous factors such as age, sex, body composition and genetics as well as underlying pathologies. The large bowel microflora are also a very significant potential confounder of metabolic profiles and could be classified as either an endogenous or exogenous factor. The main exogenous factors are diet and drugs. Diet can then be broken down to nutrients and non- nutrients. Metabolomics is one means to determine a biological

endpoint, or metabolic fingerprint, which reflects the balance of all these forces on an individual's metabolism.

Chapter- 6

Glycomics and Lipidomics

Glycomics

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 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.

Importance

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

- Gylcoproteins found on the cell surface play a critical role in bacterial and viral recognition.
- They are involved in cellular signaling pathways.
- They affect the stability and folding of proteins.
- There are many **glycan-specific diseases**.

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 and O-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.

Fractionated glycans from HPLC instruments can be further analyzed by MALDI-TOF-MS(MS) to get further informations 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. Detection is here done by mass spectrometry, but in contrast to MALDI-MS with an electrospray ionisation (ESI) interface(PGC-LC-ESI-MS or PGCC-MS)

Table 1:Advantages and disadvantages of mass spectrometry in glycan analysis

Advantages	Disadvantages
• Applicable for small sample amounts (lower fmol range) • Useful for complex glycan mixtures (generation of a further analysis dimension). • Attachment sides can be analysed by tandem MS experiments (side specific glycan analysis). • Glycan sequencing by tandem MS experiments.	• Destructive method. • Need of a proper experimental design.

Arrays

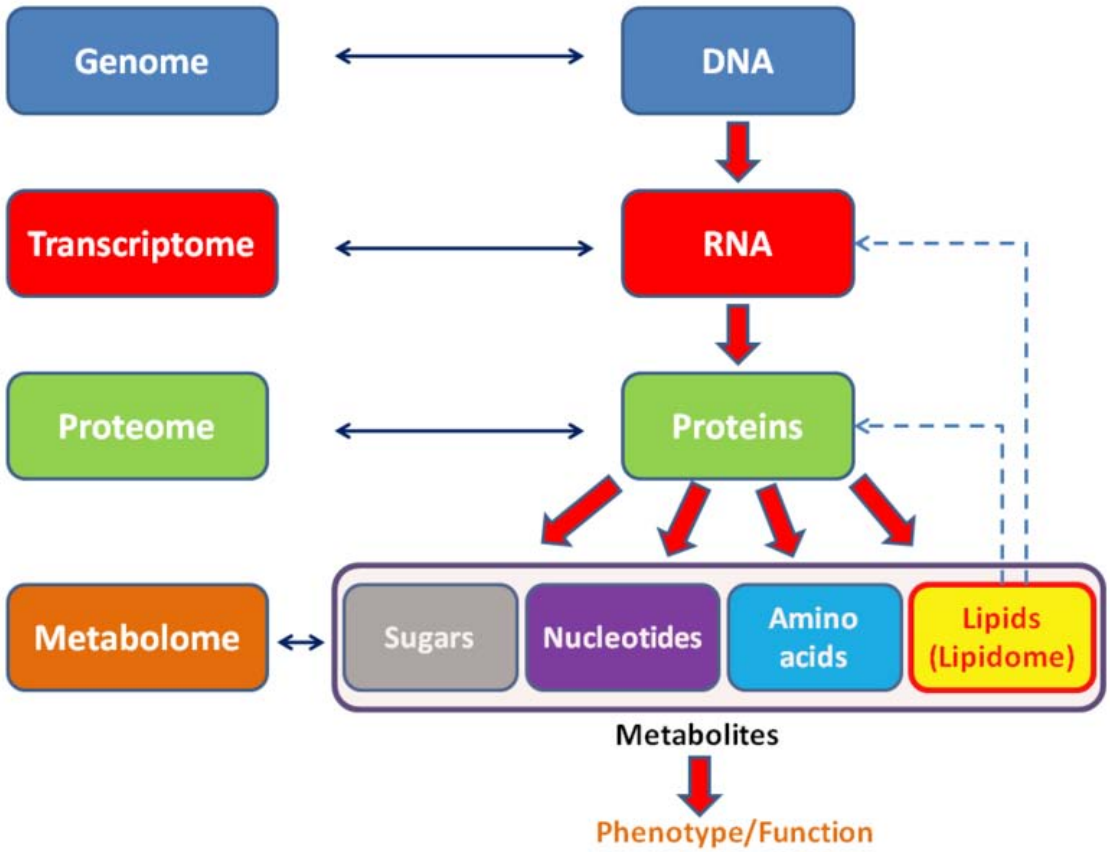
Lectin and antibody arrays provide high-throughput screening of many samples containing glycans. This method uses either naturally occurring lectins or artificial monoclonal antibodies, where both are immobilized on a certain chip and incubated with a fluorescent glycoprotein sample.

Glycan arrays, like that offered by the Consortium for Functional Glycomics, 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.

Lipidomics



General schema showing the relationships of the lipidome to the genome, transcriptome, proteome and metabolome. Lipids also regulate protein function and gene transcription as part of a dynamic "interactome" within the cell.

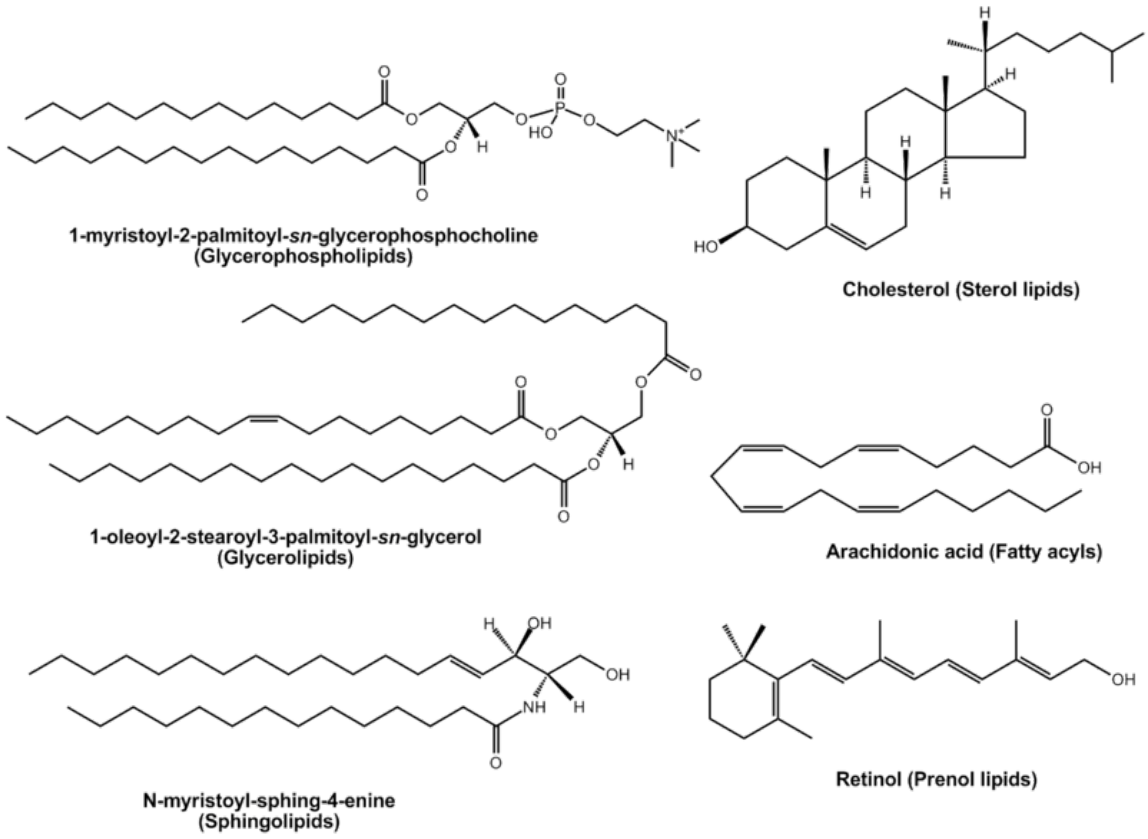
Lipidomics may be defined as the large-scale study of pathways and networks of cellular lipids in biological systems . The word "lipidome" is used to describe the complete lipid profile within a cell, tissue or organism and is a subset of the "metabolome" which also includes the three other major classes 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).



Examples of some lipids from various categories.

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.

Experimental techniques

Lipid extraction

Most methods of lipid extraction and isolation from biological samples exploit the high solubility of hydrocarbon chains in organic solvents. Given the diversity in lipid classes, it is not possible to accommodate all classes with a common extraction method. The traditional Bligh/Dyer procedure uses chloroform/methanol-based protocols that include phase partitioning into the organic layer. These protocols work relatively well for a wide variety of physiologically relevant lipids but they have to be adapted for complex lipid chemistries and low-abundance and labile lipid metabolites. When organic soil was used citrate buffer in the extraction mixture gave higher amounts of lipid phosphate than acetate buffer, Tris, H2O or phosphate buffer.

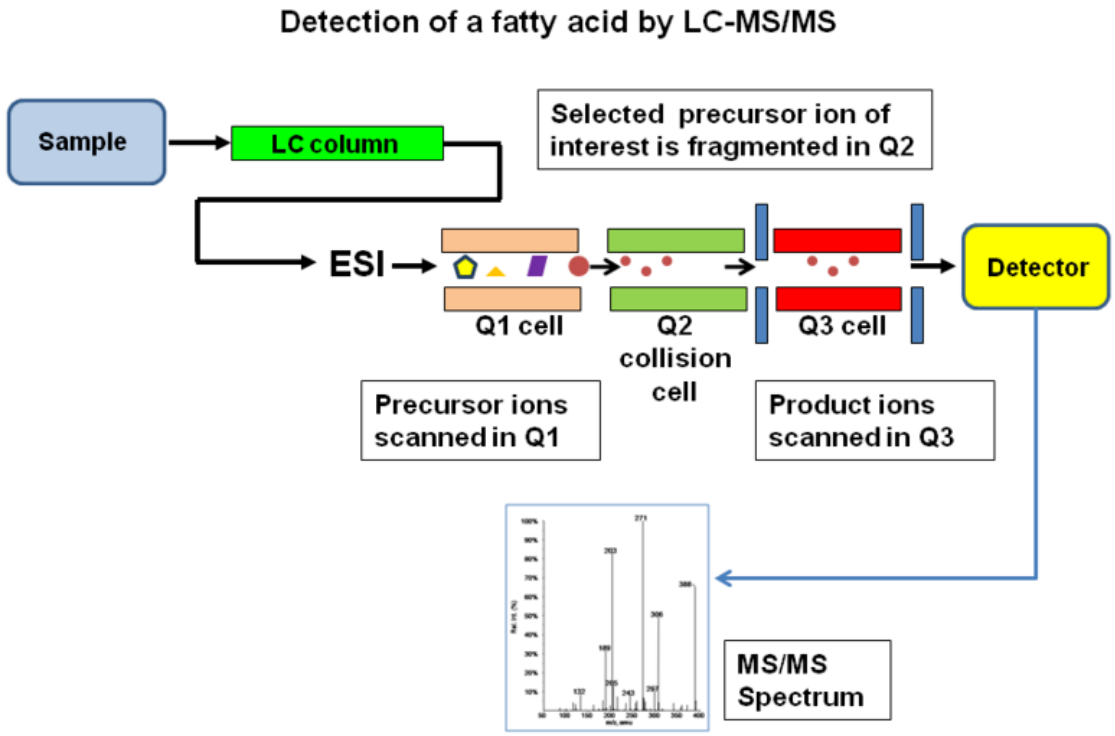
Lipid separation

The simplest method of lipid separation is the use of thin layer chromatography (TLC). Although not as sensitive as other methods of lipid detection, it offers a rapid and comprehensive screening tool prior to more sensitive and sophisticated techniques. Solid-phase extraction (SPE) chromatography is useful for rapid, preparative separation of crude lipid mixtures into different lipid classes. This involves the use of prepacked columns containing silica or other stationary phases to separate glycerophospholipids, fatty acids, cholesteryl esters, glycerolipids, and sterols from crude lipid mixtures. High performance liquid chromatography (HPLC or LC) is extensively used in lipidomic analysis to separate lipids prior to mass analysis. Separation can be achieved by either normal-phase HPLC or reverse-phase HPLC. For example, normal phase HPLC effectively separates glycerophospholipids on the basis of headgroup polarity, whereas reverse-phase HPLC effectively separates fatty acids such as eicosanoids on the basis of chain length, degree of unsaturation and substitution. HPLC of lipids may either be

performed offline or online where the eluate is integrated with the ionization source of a mass spectrometer.

Lipid detection

The progress of modern lipidomics has been greatly accelerated by the development of spectrometric methods in general and soft ionization techniques for mass spectrometry such as electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) in particular. "Soft" ionization does not cause extensive fragmentation, so that comprehensive detection of an entire range of lipids within a complex mixture can be correlated to experimental conditions or disease state. In addition to ESI and MALDI, the technique of atmospheric pressure chemical ionization (APCI) has become increasingly popular for the analysis of nonpolar lipids.



Schema showing detection of a fatty acid by LC-MS/MS using a linear ion-trap instrument and an electrospray (ESI) ion source.

ESI MS

ESI-MS was initially developed by Fenn and colleagues for analysis of biomolecules. It depends on the formation of gaseous ions from polar, thermally labile and mostly non-volatile molecules and thus is completely suitable for a variety of lipids. It is a soft-ionization method that rarely disrupts the chemical nature of the analyte prior to mass analysis. Various ESI-MS methods have been developed for analysis of different classes,

subclasses, and individual lipid species from biological extracts. Comprehensive reviews of the methods and their application have recently been published. The major advantages of ESI-MS are high accuracy, sensitivity, reproducibility, and the applicability of the technique to complex solutions without prior derivatization. Han and coworkers have developed a method known as"shotgun lipidomics" which involves direct infusion of a crude lipid extract into an ESI source optimized for intrasource separation of lipids based on their intrinsic electrical properties.

MALDI MS

MALDI mass spectrometry is a laser-based soft-ionization method often used for analysis of large proteins, but has been used successfully for lipids. The lipid is mixed with a matrix, such as 2,5-dihydroxybenzoic acid, and applied to a sample holder as a small spot. A laser is fired at the spot, and the matrix absorbs the energy, which is then transferred to the analyte, resulting in ionization of the molecule. MALDI-Time-of-flight (MALDI-TOF) MS has become a very promising approach for lipidomics studies, particularly for the imaging of lipids from tissue slides.

APCI MS

The source for APCI is similar to ESI except that ions are formed by the interaction of the heated analyte solvent with a corona discharge needle set at a high electrical potential. Primary ions are formed immediately surrounding the needle, and these interact with the solvent to form secondary ions that ultimately ionize the sample. APCI is particularly useful for the analysis of nonpolar lipids such as triacylglycerols, sterols, and fatty acid esters.

Imaging techniques

Recent developments in MALDI methods have enabled direct detection of lipids in-situ. Abundant lipid-related ions are produced from the direct analysis of thin tissue slices when sequential spectra are acquired across a tissue surface that has been coated with a MALDI matrix. Collisional activation of the molecular ions can be used to determine the lipid family and often structurally define the molecular species. This technique enables detection of phospholipids, sphingolipids and glycerolipids in tissues such as heart, kidney and brain. Furthermore distribution of many different lipid molecular species often define anatomical regions within these tissues.

Lipidomic 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

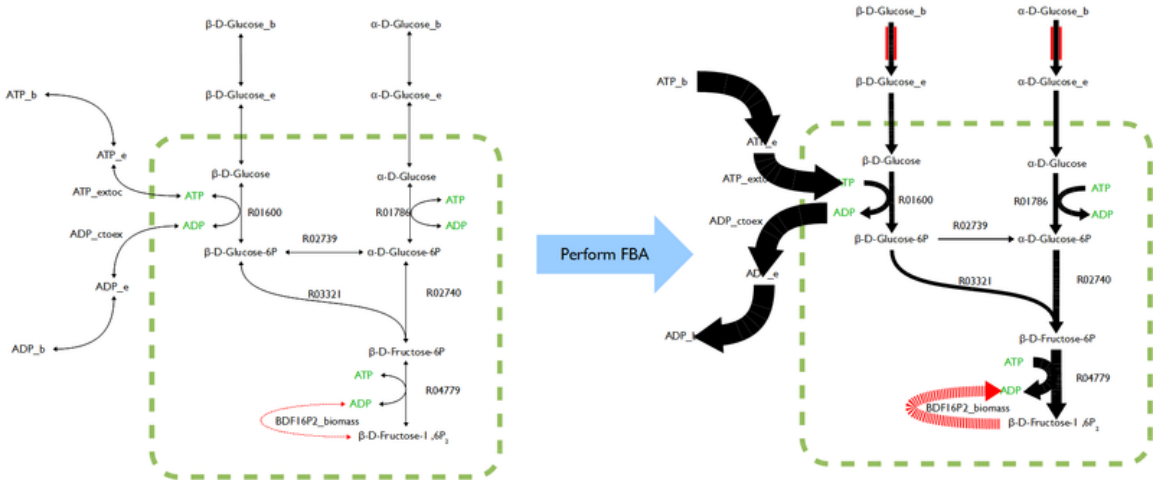
also been applied to plants and microorganisms such as yeast. A combination of quantitative lipidomic data in conjunction with the corresponding transcriptional data (using gene-array methods) and proteomic data (using tandem MS) enables a systems biology approach to a more in-depth understanding of 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. 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.

Chapter- 7

Flux Balance Analysis



The results of FBA on a prepared metabolic network of the top six reactions of glycolysis. The predicted flux through each reaction is proportional to the width of the line. Objective function in red, constraints on alpha-D-Glucose and beta-D-Glucose import represented as red bars.

Flux balance analysis (FBA), which can be considered the computational specialisation of the more general field of **fluxomics**, is a mathematical method for analysing metabolism. It does not require knowledge of metabolite concentration or details of the enzyme kinetics of the system. The assumption is made that the system being studied is homeostatic and the technique then aims to answer the question: given some known available nutrients, which set of metabolic fluxes maximises the growth rate of an organism while preserving the internal concentration of metabolites?

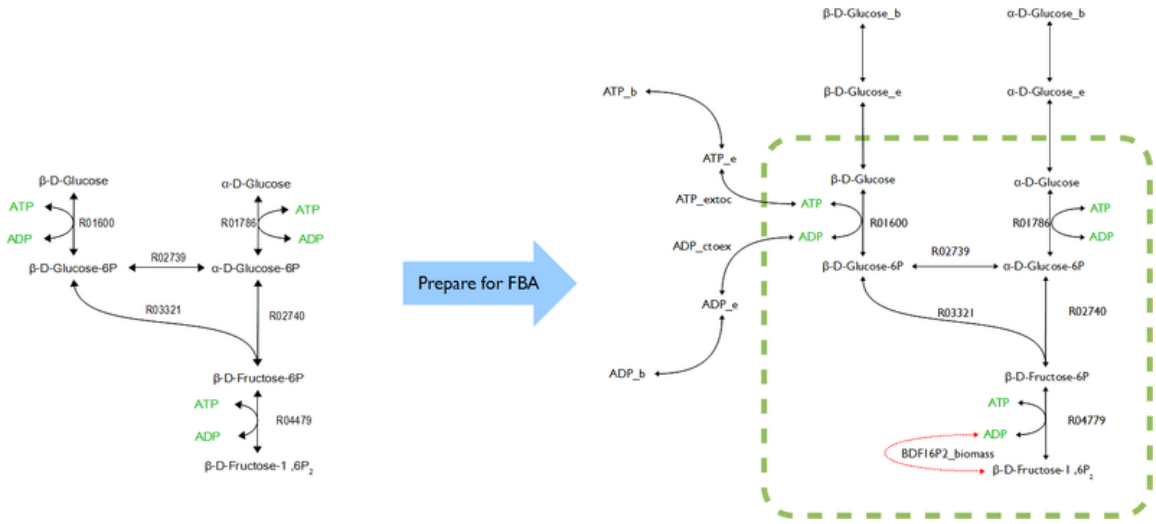
A notable example of the success of FBA is the ability to accurately predict the growth rate of the prokaryote *E. coli* when cultured in different conditions. More generally, suitable organisms can be cultivated in media with defined concentrations of nutrients, and their growth rates measured, so that the predictions of FBA can be compared with experiments and the underlying metabolic model corrected accordingly.

A good description of the basic concepts of FBA can be found in the freely available supplementary material to Edwards et al. 2001 which can be found at the Nature website. Further sources include the book "Systems Biology" by B. Palsson dedicated to the subject and a useful tutorial and paper by J. Orth. Many other sources of information on the technique exist in published scientific literature including Lee et al. 2006 and Feist et al. 2008.

Model preparation

A comprehensive guide to creating, preparing and analysing a metabolic model using FBA, in addition to other techniques, was published by Thiele and Palsson in 2010. The key parts of model preparation are: creating a metabolic network without holes, adding constraints to the model and finally adding an objective function (often called the Biomass function), usually to simulate the growth of the organism being modelled.

The network



The first six reactions in Glycolysis prepared for FBA through the addition of an objective function (red) and the import and export of nutrients (ATP, ADP, BDG, ADG) across the system boundary (dashed green line)

Metabolic networks can vary in scope from those describing the metabolism in a single pathway, up to the cell, tissue or organism. The only requirement of a metabolic network that forms the basis of an FBA-ready network is that it contains no gaps. This typically means that extensive manual curation is required, making the preparation of a metabolic network for flux-balance analysis a process that can take months or years. Software packages such as Simpheny, CellDesigner and MetNetMaker, exist to speed up the creation of new FBA-ready metabolic networks.

Generally models are created in BioPAX or SBML format so that further analysis or visualisation can take place in other software although this not a requirement.

Objective function

In FBA there are a large number of mathematically acceptable solutions to the steady-state problem $(S\vec{v} = 0)$ but the ones that are biologically interesting are those that produce the desired metabolites in the correct proportion. The set of metabolites, in the correct proportions, that an FBA model tries to create is called the objective function. When modelling an organism the objective function is generally the biomass of the organism and simulates growth and reproduction. If the biomass function is defined sensibly, or exactly measured experimentally, it can play an important role in making the results of FBA biologically applicable: by ensuring that the correct proportion of metabolites are produced by metabolism and by predicting exact rates of Biomass production for example.

When modelling smaller networks the objective function can be changed accordingly. An example of this would be in the study of the carbohydrate metabolism pathways where the objective function would probably be defined as a certain proportion of ATP and NADH and thus simulate the production of high energy metabolites by this pathway.

Constraints

A key part of FBA is the ability to add constraints to the flux rates of reactions within networks, forcing them to stay within a range of selected values. This lets the model more accurately simulate real metabolism and can be thought of biologically in two subsets; constraints that limit nutrient uptake and excretion and those that limit the flux through reactions within the organism. FBA-ready metabolic models that have had constraints added can be analysed using software such as the COBRA toolbox.

Growth media

Organisms, and all other metabolic systems, require some input of nutrients. Typically the rate of uptake of nutrients is dictated by their availability (a nutrient that isn't present cannot be absorbed), their concentration and diffusion constants (higher concentrations of quickly-diffusing metabolites are absorbed more quickly) and the method of absorption (such as active transport or facilitated diffusion versus simple diffusion).

If the rate of absorption (and/or excretion) of certain nutrients can be experimentally measured then this information can be added as a constraint on the flux rate at the edges of a metabolic model. This ensures that nutrients that are not present or not absorbed by the organism do not enter its metabolism (the flux rate is constrained to zero) and also means that known nutrient uptake rates are adhered to by the simulation. This provides a secondary method of making sure that the simulated metabolism has experimentally verified properties rather than just mathematically acceptable ones. In mathematical terms, the application of constraints can be considered to reduce the solution space of the FBA model.

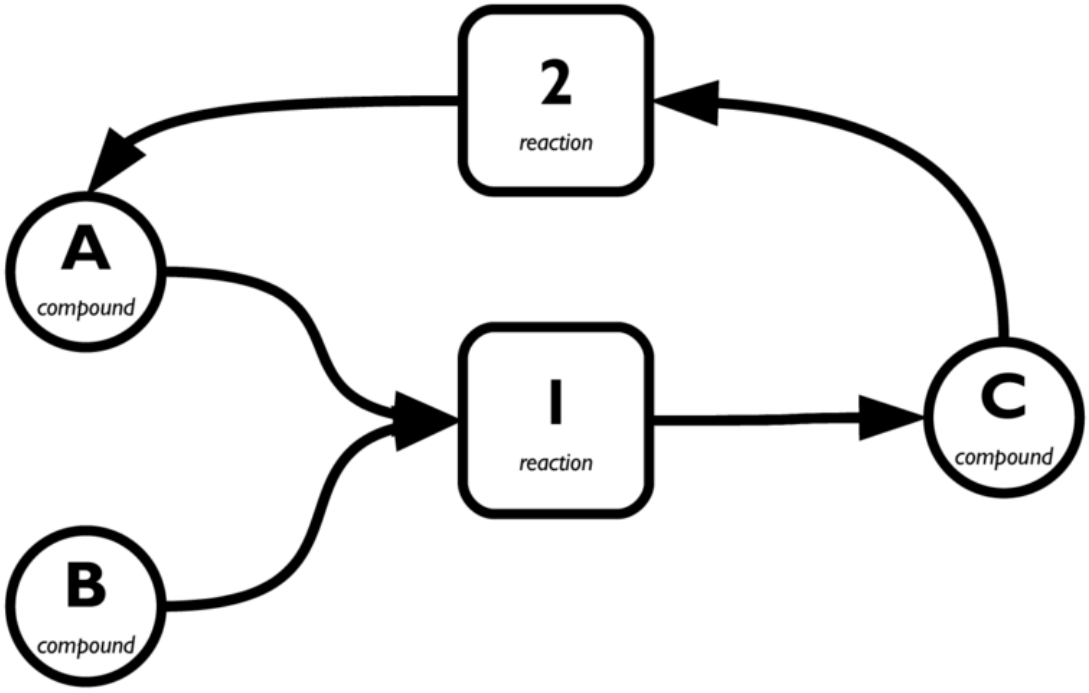
Internal constraints

In addition to constraints applied at the edges of a metabolic network, constraints can be applied to reactions deep within the network. These constraints are normally usually simple; they may constrain the direction of a reaction due to energy considerations or constrain the maximum speed of a reaction due to the finite speed of all reaction in nature.

Mathematical description

A biological network can be thought of as a set of nodes (compounds) connected by directional edges (reactions) and therefore represented as a matrix. The properties of this matrix are well known and thus a biological problem becomes amenable to computational analysis. A real biological system is extremely complex which in turn leads to problems measuring enough parameters to define the system and in some cases requiring a huge amount of computing time to perform simulations. Flux-balance analysis simplifies the representation of the biological system, requiring fewer parameters (such as enzyme kinetic rates, compound concentrations and diffusion constants) and greatly reduces the computer time required for simulations.

A simple example



A simple reaction network with two reactions and three compounds

The concentrations of all the metabolites and the fluxes through all the reactions in this simple system can be represented by the following three differential equations.

$$\frac{d[A]}{dt} = \frac{d[C]_1}{dt} = v_2 - v_1$$
$$\frac{d[B]}{dt} = \frac{d[C]_3}{dt} = -v_1$$
$$\frac{d[C]}{dt} = \frac{d[C]_2}{dt} = v_1 - v_2$$

Solving this system of differential equations is not difficult in this case but quickly becomes computationally expensive as the number of differential equations in the system rises. There is a second obstacle to solving this system; the reaction rates, v_1, $v_2$and $v_3$are themselves dependent on a number of factors generally taken from the Michaelis-Menten kinetic theory, including the kinetic parameters of the enzymes catalysing the reactions and the concentration of the metabolites themselves. Isolating enzymes from living organisms and measuring their kinetic parameters is a difficult task, as is measuring the internal concentrations, and diffusion constants, of metabolites within an organism. For this reason the differential equation approach to modelling metabolism becomes extraordinarily difficult and beyond the current scope of science for all but the most studied organisms (link to Heinemann E. Coli paper with all internal fluxes measured and Manchester yeast paper with internal fluxes measured).

The power of homeostasis

Much of the power of flux-balance analysis comes from applying the principle of homeostasis to the problem. Since the internal concentrations of metabolites within a biological system remain more or less the same over time we can apply the homeostatic condition that,

$$\frac{d[C]_1}{dt} = \frac{d[C]_2}{dt} = \frac{d[C]_3}{dt} = 0$$

Or in the general case,

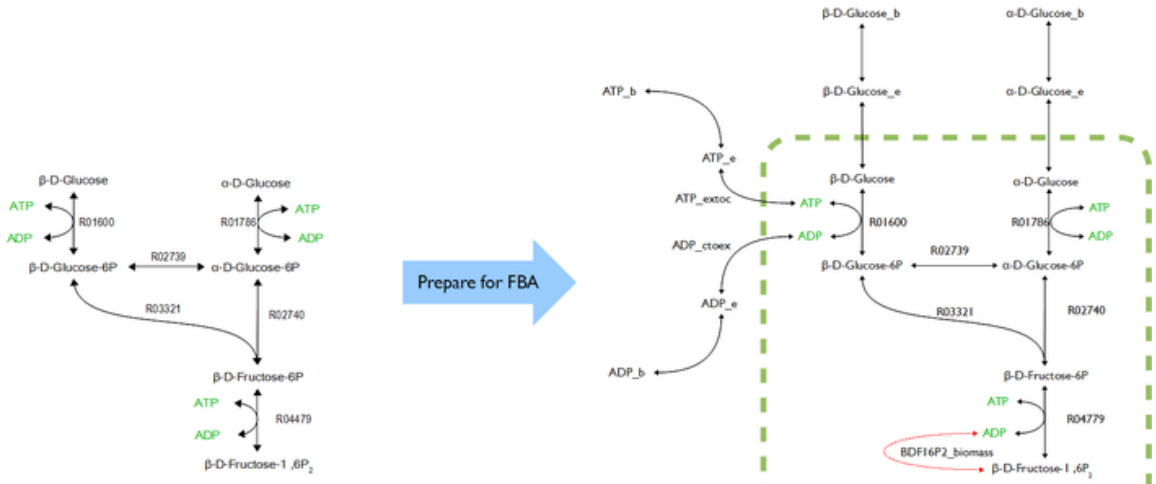
$$\frac{d[C]_i}{dt} = 0$$

And thus simplify the problem to one of simply balancing the fluxes within the system, hence the name flux-balance analysis.

$$v_2 - v_1 = v_1 - v_2 = -v_1$$

This set of equations is now much easier to solve, although in this case the only solution is the null solution $v_1 = v_2 = 0$.

The stoichiometric matrix



	R02739	R03321	R02740	R01600	R01786	R04779
C00002	0	0	0	-1	-1	-1
C00008	0	0	0	1	1	1
C00221	0	0	0	-1	0	0
C00267	0	0	0	0	-1	0
C00668	-1	0	-1	0	1	0
C01172	1	-1	0	1	0	0
C05345	0	1	1	0	0	-1
C05378	0	0	0	0	0	1

	R02739_c	R03321_c	R02740_c	R01600_c	R01786_c	R04779_c	ATP_extoc_tr	ADP_ctoex_tr	BDG_extoc_tr	ADG_extoc_tr	BDF16P2_biomass_tr
C00002[c]	0	0	0	-1	-1	-1	1	0	0	0	0
C00002[e]	0	0	0	0	0	0	-1	0	0	0	0
C00008[c]	0	0	0	1	1	1	0	-1	0	0	1
C00008[e]	0	0	0	0	0	0	0	1	0	0	0
C00221[c]	0	0	0	-1	0	0	0	0	1	0	0
C00221[e]	0	0	0	0	0	0	0	0	-1	0	0
C00267[c]	0	0	0	0	-1	0	0	0	0	1	0
C00267[e]	0	0	0	0	0	0	0	0	0	-1	0
C00668[c]	-1	0	-1	0	1	0	0	0	0	0	0
C01172[c]	1	-1	0	1	0	0	0	0	0	0	0
C05345[c]	0	1	1	0	0	-1	0	0	0	0	0
C05378[c]	0	0	0	0	0	1	0	0	0	0	-1

An example stoichiometric matrix for a network representing the top of glycolysis and that same network after being prepared for FBA.

The representation of the equations above can be generalised to any similar biological network and represented in a more powerful manner by using matrices. The stoichiometric matrix for the simple set of reactions above is,

$$\mathbf{S} = \begin{bmatrix} -1 & 1 \\ -1 & 0 \\ 1 & -1 \end{bmatrix}$$

The stoichiometric matrix is also often referred to in chemistry, metabolic control analysis and dynamical systems with the letter $\mathbf{N}$(meaning number as related to stoichiometry, S is often reserved for species or Entropy) but in FBA it is usually referred to as $\mathbf{S}$. Both letters are exactly equivalent. At this stage it is useful to define a vector $\vec{v}$ where each component of the vector represents the rate (flux through) its respective reaction within the stoichiometric matrix

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

Multiplying this matrix, $\mathbf{S}$, with $\vec{v}$, is completely equivalent to the equations derived directly from the reaction diagram,

> (The following line is erroneous, but it is unclear whether the left or right side of the "=" should be altered.)

$$\begin{bmatrix} -1 & 1 \\ -1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -v_1 + v_2 \\ v_1 - v_2 \\ -v_1 \end{bmatrix} = \begin{bmatrix} \frac{d[C]_1}{dt} \\ \frac{d[C]_2}{dt} \\ \frac{d[C]_3}{dt} \end{bmatrix}$$

Applying the homeostatic condition then gives us,

> (The following line is erroneous, but it is unclear whether the left or right side of the "=" should be altered.)

$$\begin{bmatrix} -1 & 1 \\ -1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} -v_1 + v_2 \\ v_1 - v_2 \\ -v_1 \end{bmatrix} = \begin{bmatrix} \frac{d[C]_1}{dt} \\ \frac{d[C]_2}{dt} \\ \frac{d[C]_3}{dt} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

In the general case we can write,

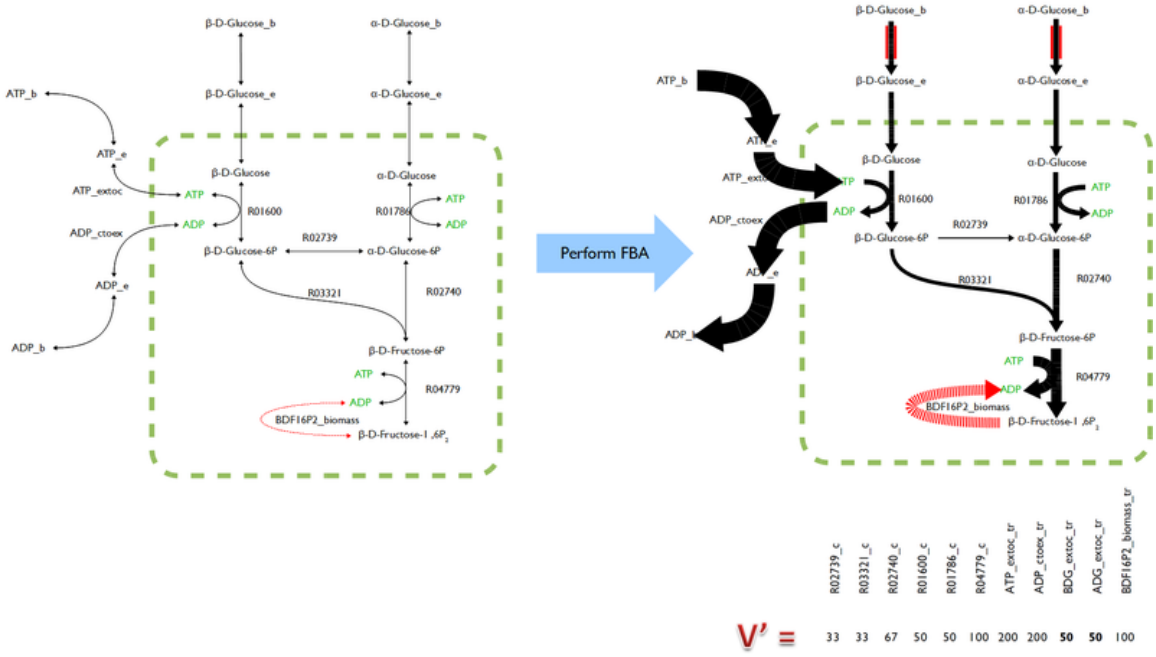
$$\mathbf{S}\,\vec{v} = 0$$

Or often confusingly, given the different nature of the ⋅when referring to the vector dot product, but identically as,

$$\mathbf{S} \cdot \vec{v} = 0$$

With the single 0 representing the null vector,

$$0 = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{bmatrix}.$$



The results of FBA can be represented identically as a vector of fluxes, or by weighting the lines representing the reactions according to the flux they carry.

This general operation is called taking the Null Space of the stoichiometric matrix $\mathbf{S}$and the technique is valid for all stoichiometric matrices, not just the small example here. Since a typical stoichiometric matrix contains many more metabolites than reactions ($m < n$) and the majority of reactions are linearly independent there are many vectors $\vec{v}$ that satisfy the equation and thus span the null space of $\mathbf{S}$.

Application to the biology of the system

The analysis of the null space of matrices is common within linear algebra and many software packages such as Matlab and Octave can help with this process. Nevertheless, knowing the null space of $\mathbf{S}$only tells us all the possible collections of flux vectors (or linear combinations thereof) that balance fluxes within the biological network. Flux-balance analysis has two further aims, to accurately represent the biology limits of the system and to return the flux distribution closest to that naturally occurring within the target system/organism.

Constraints

The stoichiometric matrix is almost always underdetermined meaning that the solution space to $\mathbf{S}\ \vec{v}=0$is very large. The size of the solution space can be reduced, and made more reflective of the biology of the problem through the application of certain constraints on the solutions.

Thermodynamic

In principle all reactions are reversible however in practise many reactions effectively occur in only one direction. This can be because of a significantly higher concentration of reactants compared to the concentration of the products of the reaction but is more often because the products of a reaction have a much lower free energy than the reactants and therefore the forward direction of a reaction is massively favoured. For ideal reactions,

$$-\infty < v_i < \infty$$

For certain reactions a thermodynamic constraint can be applied implying direction (in this case forward)

$$0 < v_i < \infty$$

Realistically the flux through a reaction cannot be infinite which implies that,

$$0 < v_i < v_{\max}$$

Measured flux rates

Certain flux rates can be measured experimentally ($v_{i,m}$) and the fluxes within a metabolic model can be constrained, within some error (ε), to ensure these known flux rates are accurately reproduced in the simulation.

$$v_{i,m} - \varepsilon < v_i < v_{i,m} + \varepsilon$$

Flux rates are most easily measured for nutrient uptake at the edge of the network but measurements of internal fluxes are possible, generally using radioactively labelled or NMR visible metabolites.

Optimisation (the objective/biomass function)

Even after the application of constraints there is usually a large number of possible solutions to the flux-balance problem. If an optimisation goal is defined, linear programming can be used to find a single optimal solution. The most common biological optimisation goal for a whole organism metabolic network would be to choose the flux vector $\vec{v}$ that maximises the flux through a biomass function composed of the constituent metabolites of the organism placed into the stoichiometric matrix and denoted v_{biomass} or simply v_b

$$\max_{\vec{v}} \; v_b \qquad \text{s.t.} \qquad \mathbf{S}\,\vec{v} = 0$$

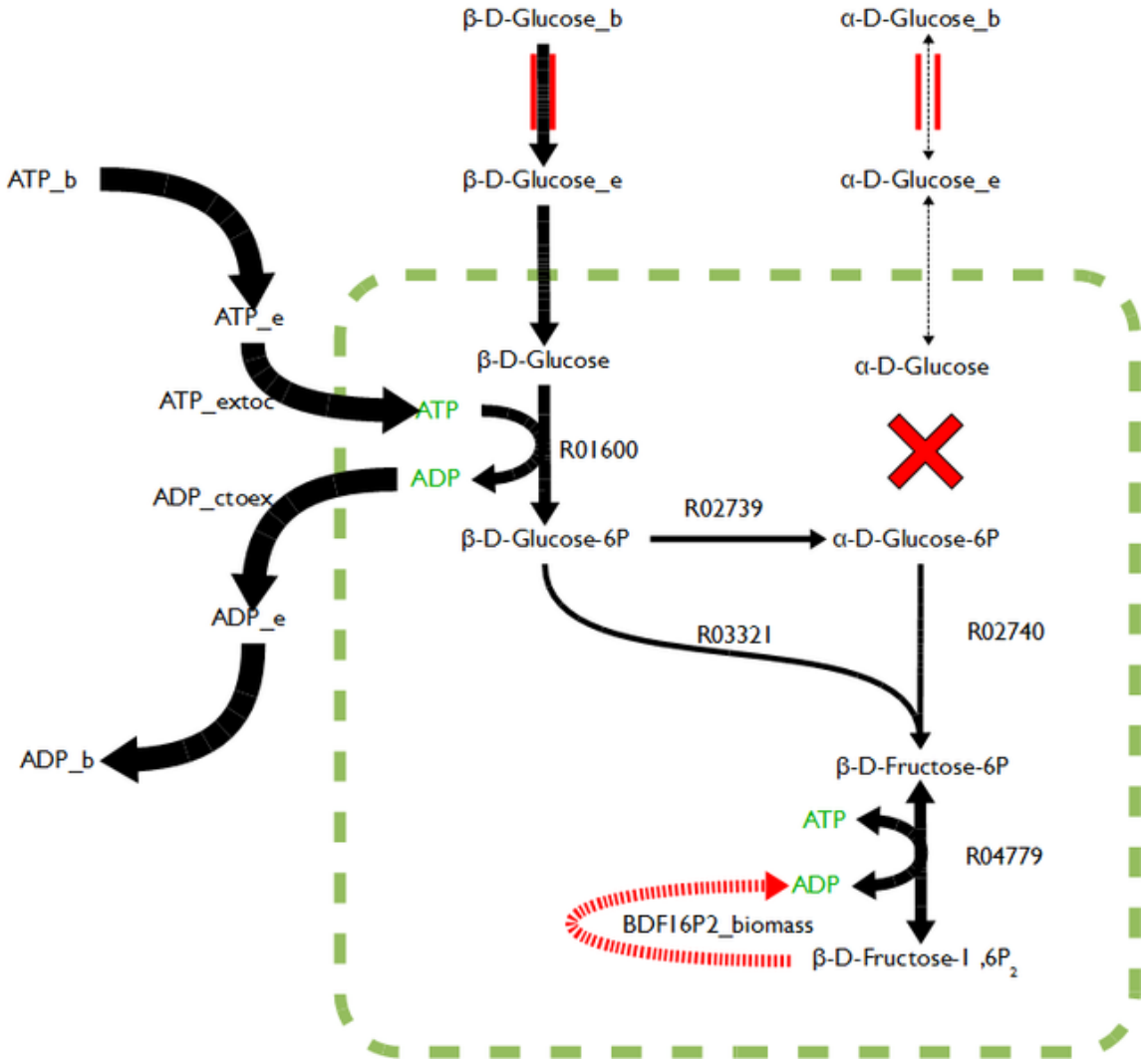
In the more general case any reaction be defined and added defined as a biomass function with either the condition that it be maximised or minimised if a single “optimal” solution is desired. Alternatively, and in the most general case, a vector $\vec{c}$ can be defined which

defines the weighted set of reactions that the linear programming model should aim to maximise or minimise,

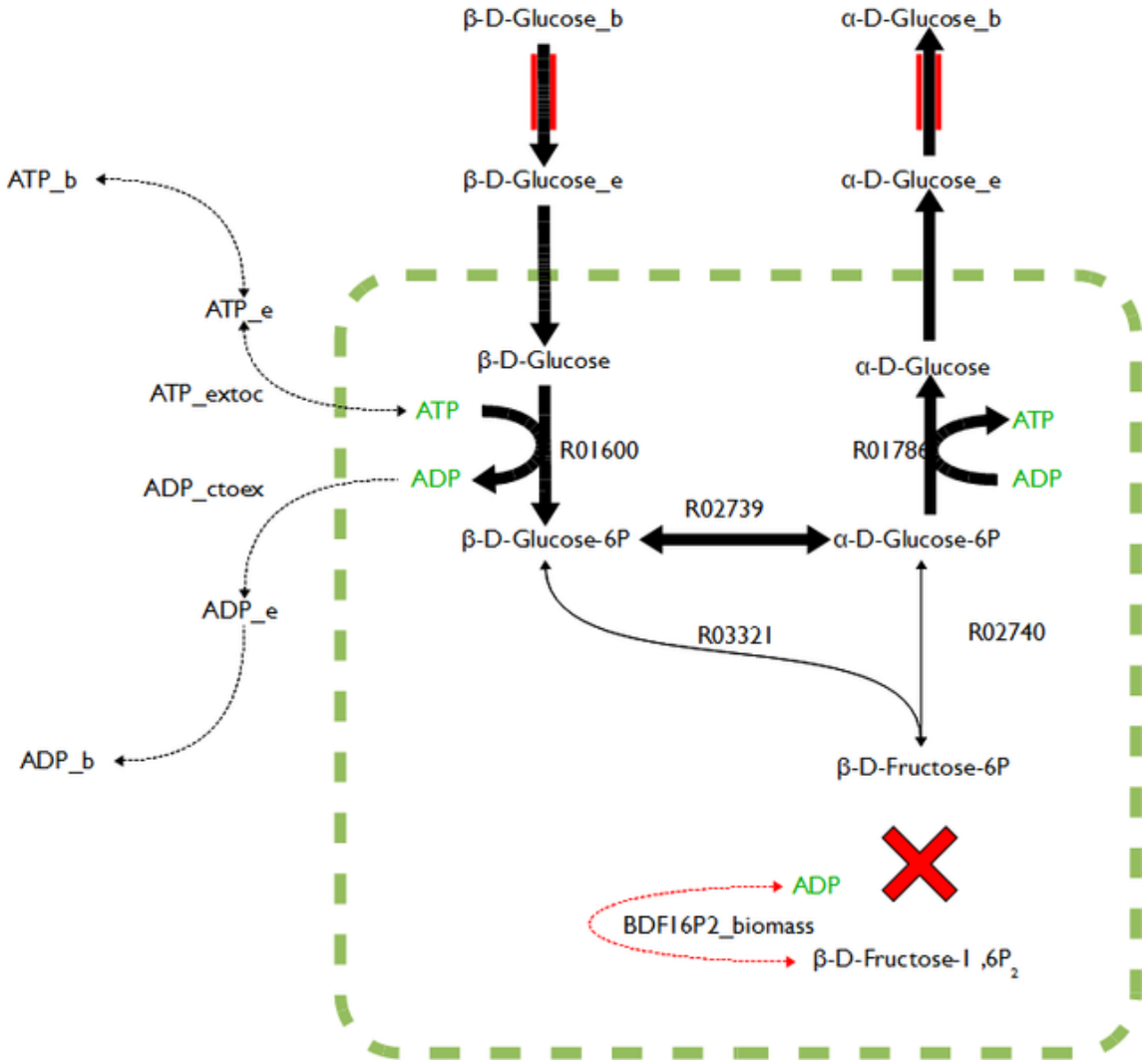
$$\max_{\vec{v}} \ \vec{v} \cdot \vec{c} \qquad \text{s.t.} \qquad \mathbf{S}\,\vec{v} = 0$$

In the case of there being only a single separate biomass function/reaction within the stoichiometric matrix $\vec{c}$ would simplify to all zeroes with a value of 1 (or any non-zero value) in the position corresponding to that biomass function. Where there were multiple separate objective functions $\vec{c}$ would simplify to all zeroes with weighted values in the positions corresponding to all objective functions.

Simulating perturbations



An example of a non lethal gene deletion in a sample metabolic network with fluxes shown by the weight of the reaction lines as calculated by FBA. Here the flux through the objective function is halved but is still present.



An example of a lethal gene deletion in a sample metabolic network with fluxes shown by the weight of the reaction lines as calculated by FBA. Here there is no flux through the objective function, simulating that the pathway is no longer functional.

FBA is not computationally intensive, taking on the order of seconds to calculate optimal fluxes for biomass production for a simple organism (around 1000 reactions). This means that the effect of deleting reactions from the network and/or changing flux constraints can be sensibly modelled on a single computer.

Single reaction deletion

A frequently used technique to search a metabolic network for reactions that are particularly critical to the production of biomass. By removing each reaction in a network in turn and measuring the predicted flux through the biomass function, each reaction can be classified as either essential (if the flux through the biomass function is substantially

reduced) or non-essential (if the flux through the biomass function is unchanged or only slightly reduced).

Reaction inhibition

The effect of inhibiting a reaction, rather than removing it entirely, can be simulated in FBA by restricting the allowed flux through it. The effect of an inhibition can be classified as lethal or non-lethal by applying the same criteria as in the case of a deletion where a suitable threshold is used to distinguish "substantially reduced" from "slightly reduced". Generally the choice of threshold is arbitrary but a reasonable estimate can be obtained from growth experiments where the simulated inhibitions/deletions are actually performed and growth rate is measured.

Interpreting results

The utility of reaction inhibition and deletion analyses is most clear if a gene-protein-reaction matrix has been assembled for the network being studied with FBA. If this has been done then information on which reactions are essential can be converted into information on which genes are essential (and thus what gene defects may cause a certain disease) or which proteins/enzymes are essential (and thus what enzymes are the most promising drug targets in pathogens).

Reaction deletion in pairs

An extension of single reaction deletions are double reaction deletions where all possible pairs of reactions are deleted. This can be useful when looking for drug targets as it allows the simulation of multi-target treatments, either by a single drug with multiple targets or by drug combinations.

Growth media modification

FBA has also been used to simulate the effect on growth rate of changes in the growth media of the metabolic system being studied. In *E. coli* the predicted growth rates of bacteria in varying media have been shown to correlate well with experimental results as well as to define precise minimal media for the culture of *Salmonella typhimurium*.

Comparison with other techniques

FBA provides a less simplistic analysis than Choke Point Analysis while requiring far less information on reaction rates and a much less complete network reconstruction than a full dynamic simulation would require. In filling this niche, FBA has been shown to be a very useful technique for analysis of the metabolic capabilities of cellular systems.

Choke point analysis

Unlike choke point analysis, FBA is a true form of metabolic network modelling because it considers the metabolic network as a single entity (the stoichiometric matrix) at all stages of analysis. This means that network effects, such as chemical reactions in distant pathways affecting each other, can be reproduced in the model. The upside to the inability of choke point analysis to simulate network effects is that it considers each reaction within a network in isolation and thus can suggest important reactions in a network even if a network is highly fragmented and contains many gaps.

Dynamic metabolic simulation

Unlike dynamic metabolic simulation, FBA assumes that the internal concentration of metabolites within a system stays constant over time and thus is unable to provide anything other than steady-state solutions. It is unlikely that FBA could, for example, simulate the functioning of a nerve cell. Since the internal concentration of metabolites is not considered within a model, it is possible that an FBA solution could contain metabolites at a concentration too high to be biologically acceptable. This is a problem that dynamic metabolic simulations would probably avoid. One advantage of the simplicity of FBA over dynamic simulations is that they are far less computationally expensive, allowing the simulation of large numbers of perturbations to the network. A second advantage is that the reconstructed model can be substantially simpler by avoiding the need to consider enzyme rates and the effect of complex interactions on enzyme kinetics.

Chapter- 8

Genomics, Transcriptome and Interactomics

Genomics

Genomics is a discipline in genetics concerning the study of the genomes of organisms. The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic mapping efforts. The field also includes studies of intragenomic phenomena such as heterosis, epistasis, pleiotropy and other interactions between loci and alleles within the genome. In contrast, the investigation of the roles and functions of single genes is a primary focus of molecular biology or genetics and is a common topic of modern medical and biological research. Research of single genes does not fall into the definition of genomics unless the aim of this genetic, pathway, and functional information analysis is to elucidate its effect on, place in, and response to the entire genome's networks.

For the United States Environmental Protection Agency, "the term "genomics" encompasses a broader scope of scientific inquiry associated technologies than when genomics was initially considered. A genome is the sum total of all an individual organism's genes. Thus, genomics is the study of all the genes of a cell, or tissue, at the DNA (genotype), mRNA (transcriptome), or protein (proteome) levels."

History

The first genomes to be sequenced were those of a virus and a mitochondrion, and were done by Fred Sanger. His group established techniques of sequencing, genome mapping, data storage, and bioinformatic analyses in the 1970-1980s. A major branch of genomics is still concerned with sequencing the genomes of various organisms, but the knowledge of full genomes has created the possibility for the field of functional genomics, mainly concerned with patterns of gene expression during various conditions. The most important tools here are microarrays and bioinformatics. Study of the full set of proteins in a cell type or tissue, and the changes during various conditions, is called proteomics. A

related concept is materiomics, which is defined as the study of the material properties of biological materials (e.g. hierarchical protein structures and materials, mineralized biological tissues, etc.) and their effect on the macroscopic function and failure in their biological context, linking processes, structure and properties at multiple scales through a materials science approach. The actual term 'genomics' is thought to have been coined by Dr. Tom Roderick, a geneticist at the Jackson Laboratory (Bar Harbor, ME) over beer at a meeting held in Maryland on the mapping of the human genome in 1986.

In 1972, Walter Fiers and his team at the Laboratory of Molecular Biology of the University of Ghent (Ghent, Belgium) were the first to determine the sequence of a gene: the gene for Bacteriophage MS2 coat protein. In 1976, the team determined the complete nucleotide-sequence of bacteriophage MS2-RNA. The first DNA-based genome to be sequenced in its entirety was that of bacteriophage Φ-X174; (5,368 bp), sequenced by Frederick Sanger in 1977.

The first free-living organism to be sequenced was that of *Haemophilus influenzae* (1.8 Mb) in 1995, and since then genomes are being sequenced at a rapid pace.

As of September 2007, the complete sequence was known of about 1879 viruses , 577 bacterial species and roughly 23 eukaryote organisms, of which about half are fungi. Most of the bacteria whose genomes have been completely sequenced are problematic disease-causing agents, such as *Haemophilus influenzae*. Of the other sequenced species, most were chosen because they were well-studied model organisms or promised to become good models. Yeast (*Saccharomyces cerevisiae*) has long been an important model organism for the eukaryotic cell, while the fruit fly *Drosophila melanogaster* has been a very important tool (notably in early pre-molecular genetics). The worm *Caenorhabditis elegans* is an often used simple model for multicellular organisms. The zebrafish *Brachydanio rerio* is used for many developmental studies on the molecular level and the flower *Arabidopsis thaliana* is a model organism for flowering plants. The Japanese pufferfish (*Takifugu rubripes*) and the spotted green pufferfish (*Tetraodon nigroviridis*) are interesting because of their small and compact genomes, containing very little non-coding DNA compared to most species. The mammals dog (*Canis familiaris*), brown rat (*Rattus norvegicus*), mouse (*Mus musculus*), and chimpanzee (*Pan troglodytes*) are all important model animals in medical research.

Human genomics

A rough draft of the human genome was completed by the Human Genome Project in early 2001, creating much fanfare. By 2007 the human sequence was declared "finished" (less than one error in 20,000 bases and all chromosomes assembled). Display of the results of the project required significant bioinformatics resources. The sequence of the human reference assembly can be explored using the UCSC Genome Browser.

Bacteriophage genomics

Bacteriophages have played and continue to play a key role in bacterial genetics and molecular biology. Historically, they were used to define gene structure and gene regulation. Also the first genome to be sequenced was a bacteriophage. However, bacteriophage research did not lead the genomics revolution, which is clearly dominated by bacterial genomics. Only very recently has the study of bacteriophage genomes become prominent, thereby enabling researchers to understand the mechanisms underlying phage evolution. Bacteriophage genome sequences can be obtained through direct sequencing of isolated bacteriophages, but can also be derived as part of microbial genomes. Analysis of bacterial genomes has shown that a substantial amount of microbial DNA consists of prophage sequences and prophage-like elements. A detailed database mining of these sequences offers insights into the role of prophages in shaping the bacterial genome.

Cyanobacteria genomics

At present there are 24 cyanobacteria for which a total genome sequence is available. 15 of these cyanobacteria come from the marine environment. These are six *Prochlorococcus* strains, seven marine *Synechococcus* strains, *Trichodesmium erythraeum* IMS101 and *Crocosphaera watsonii* WH8501. Several studies have demonstrated how these sequences could be used very successfully to infer important ecological and physiological characteristics of marine cyanobacteria. However, there are many more genome projects currently in progress, amongst those there are further *Prochlorococcus* and marine *Synechococcus* isolates, *Acaryochloris* and *Prochloron*, the N_2-fixing filamentous cyanobacteria *Nodularia spumigena*, *Lyngbya aestuarii* and *Lyngbya majuscula*, as well as bacteriophages infecting marine cyanobaceria. Thus, the growing body of genome information can also be tapped in a more general way to address global problems by applying a comparative approach. Some new and exciting examples of progress in this field are the identification of genes for regulatory RNAs, insights into the evolutionary origin of photosynthesis, or estimation of the contribution of horizontal gene transfer to the genomes that have been analyzed.

Transcriptome

The **transcriptome** is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in one or a population of cells. 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*, also referred to as expression profiling, examines the expression level of mRNAs in a given cell population, often using high-throughput techniques based on

DNA microarray technology. The use of next-generation sequencing technology to study the transcriptome at the nucleotide level is known as RNA-Seq .

Transcriptomics is the branch of molecular biology that deals with the study of messenger RNA molecules produced in an individual or population of a particular cell type.

Applications and analysis

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 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 or subtypes; 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.

mRNA regulation

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. The complete protein complement of a cell or organism is known as the proteome.

A study of 158,807 mouse transcripts revealed that 4520 of these transcripts form antisense partners that are base pair complementary to the exons of genes. These results raise the possibility that significant numbers of "antisense RNA-coding genes" might participate in the regulation of the levels of expression of protein-coding mRNAs.

Interactomics

Interactomics is a discipline at the intersection of bioinformatics and biology that deals with studying both the interactions and the consequences of those interactions between and among proteins, and other molecules within a cell. The network of all such interactions is called the Interactome. Interactomics thus aims to compare such networks of interactions (i.e., interactomes) between and within species in order to find how the

traits of such networks are either preserved or varied. From a mathematical, or mathematical biology viewpoint an interactome network is a graph or a category representing the most important interactions pertinent to the normal physiological functions of a cell or organism.

Interactomics is an example of "top-down" systems biology, which takes an overhead, as well as overall, view of a biosystem or organism. Large sets of genome-wide and proteomic data are collected, and correlations between different molecules are inferred. From the data new hypotheses are formulated about feedbacks between these molecules. These hypotheses can then be tested by new experiments.

Through the study of the interaction of all of the molecules in a cell the field looks to gain a deeper understanding of genome function and evolution than just examining an individual genome in isolation. Interactomics goes beyond cellular proteomics in that it not only attempts to characterize the interaction between proteins, but between all molecules in the cell.

Methods of interactomics

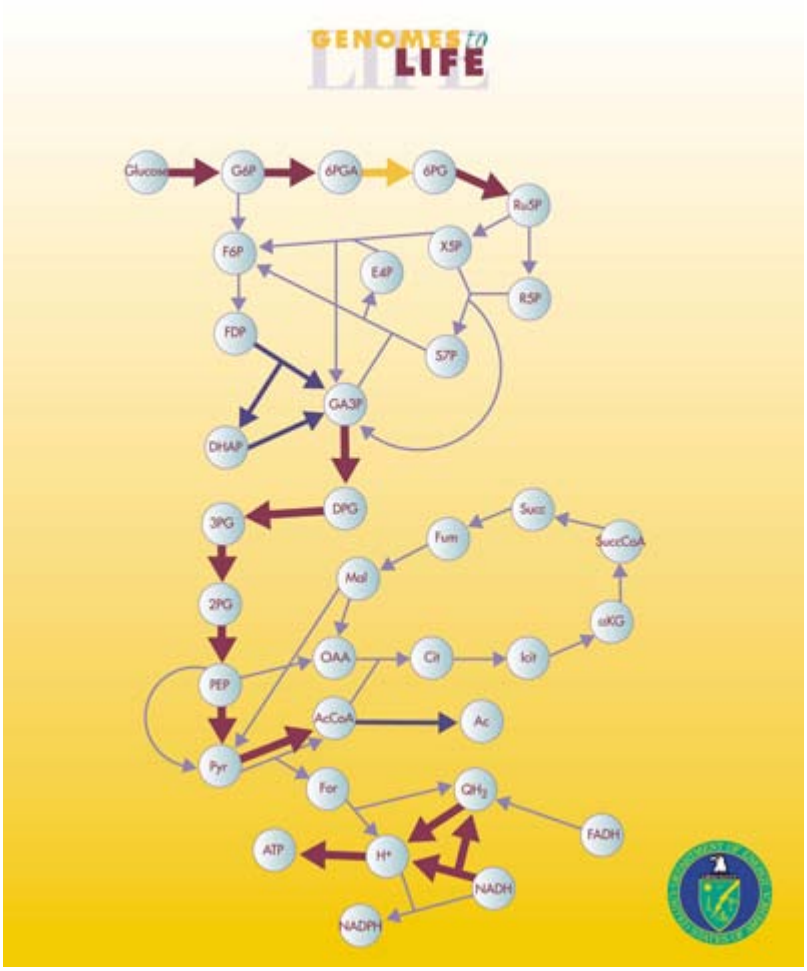
The study of the interactome requires the collection of large amounts of data by way of high throughput experiments. Through these experiments a large number of data points are collected from a single organism under a small number of perturbations These experiments include:

- Two-hybrid screening
- Tandem Affinity Purification
- X-ray tomography
- Optical fluorescence microscopy

Recent developments

The field of interactomics is currently rapidly expanding and developing. While no biological interactomes have been fully characterized. Over 90% of proteins in *Saccharomyces cerevisiae* have been screened and their interactions characterized, making it the first interactome to be nearly fully specified.

Also there have been recent systematic attempts to explore the human interactome and.



Metabolic Network Model for Escherichia coli.

Other species whose interactomes have been studied in some detail include *Caenorhabditis elegans* and *Drosophila melanogaster*.

Criticisms and concerns

Kiemer and Cesareni raise the following concerns with the current state of the field:

- The experimental procedures associated with the field are error prone leading to "noisy results". This leads to 30% of all reported interactions being artifacts. In fact, two groups using the same techniques on the same organism found less than 30% interactions in common.
- Techniques may be biased, i.e. the technique determines which interactions are found.
- Ineractomes are not nearly complete with perhaps the exception of *S. cerivisiae*.
- While genomes are stable, interactomes may vary between tissues and developmental stages.

- Genomics compares amino acids, and nucleotides which are in a sense unchangeable, but interactomics compares proteins and other molecules which are subject to mutation and evolution.
- It is difficult to match evolutionarily related proteins in distantly related species.

Chapter- 9

Cell Signaling

Cell signaling is part of a complex system of communication that governs basic cellular activities and coordinates cell actions. The ability of cells to perceive and correctly respond to their microenvironment is the basis of development, tissue repair, and immunity as well as normal tissue homeostasis. Errors in cellular information processing are responsible for diseases such as cancer, autoimmunity, and diabetes. By understanding cell signaling, diseases may be treated effectively and, theoretically, artificial tissues may be created.

Traditional work in biology has focused on studying individual parts of cell signaling pathways. Systems biology research helps us to understand the underlying structure of cell signaling networks and how changes in these networks may affect the transmission and flow of information. Such networks are complex systems in their organization and may exhibit a number of emergent properties including bistability and ultrasensitivity. Analysis of cell signaling networks requires a combination of experimental and theoretical approaches including the development and analysis of simulations and modelling.

Unicellular and multicellular organism cell signaling

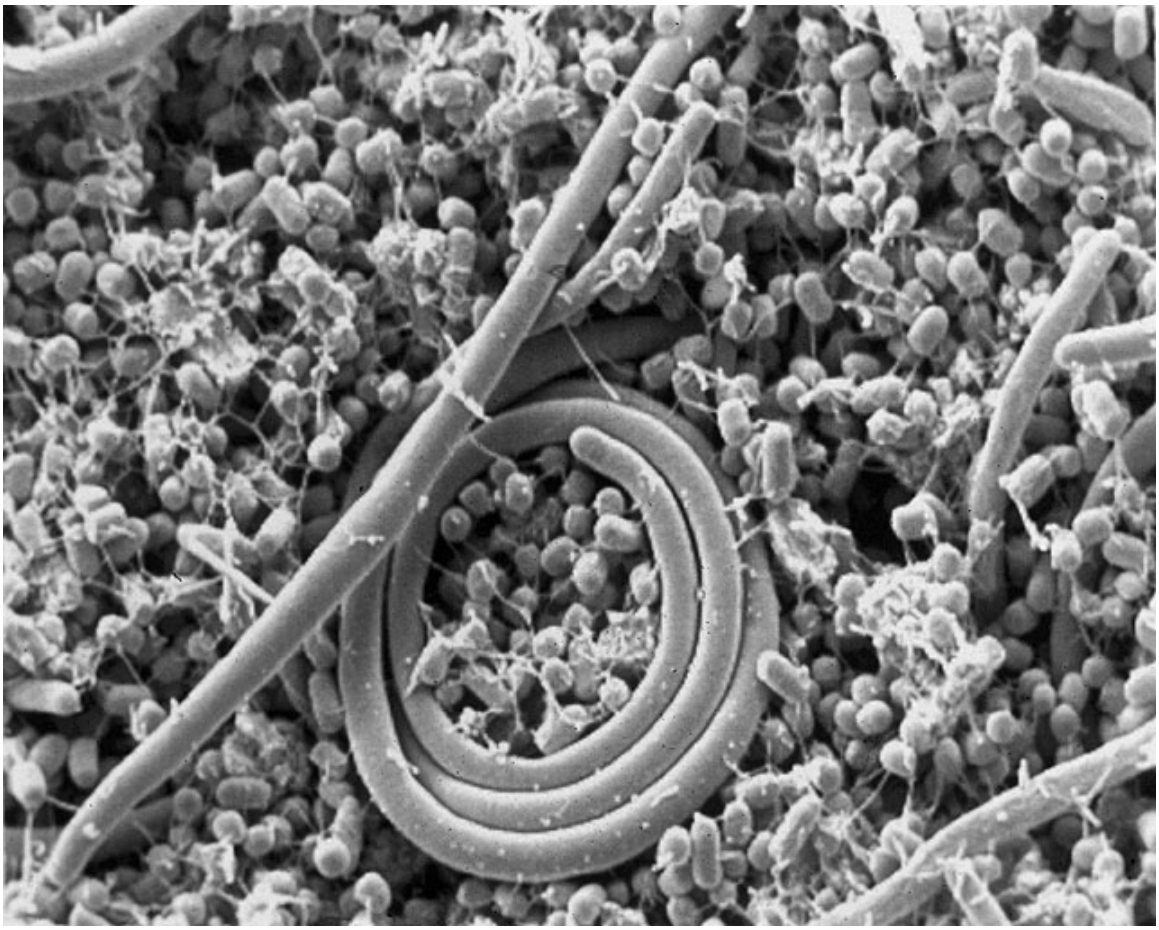


Figure 1. Example of signaling between bacteria. Salmonella enteritidis uses acyl-homoserine lactone for Quorum sensing (see: Inter-Bacterial Communication)

Cell signaling has been most extensively studied in the context of human diseases and signaling between cells of a single organism. However, cell signaling may also occur between the cells of two different organisms. In many mammals, early embryo cells exchange signals with cells of the uterus. In the human gastrointestinal tract, bacteria exchange signals with each other and with human epithelial and immune system cells. For the yeast *Saccharomyces cerevisiae* during mating, some cells send a peptide signal (mating factor *pheromones*) into their environment. The mating factor peptide may bind to a cell surface receptor on other yeast cells and induce them to prepare for mating.

Types of signals

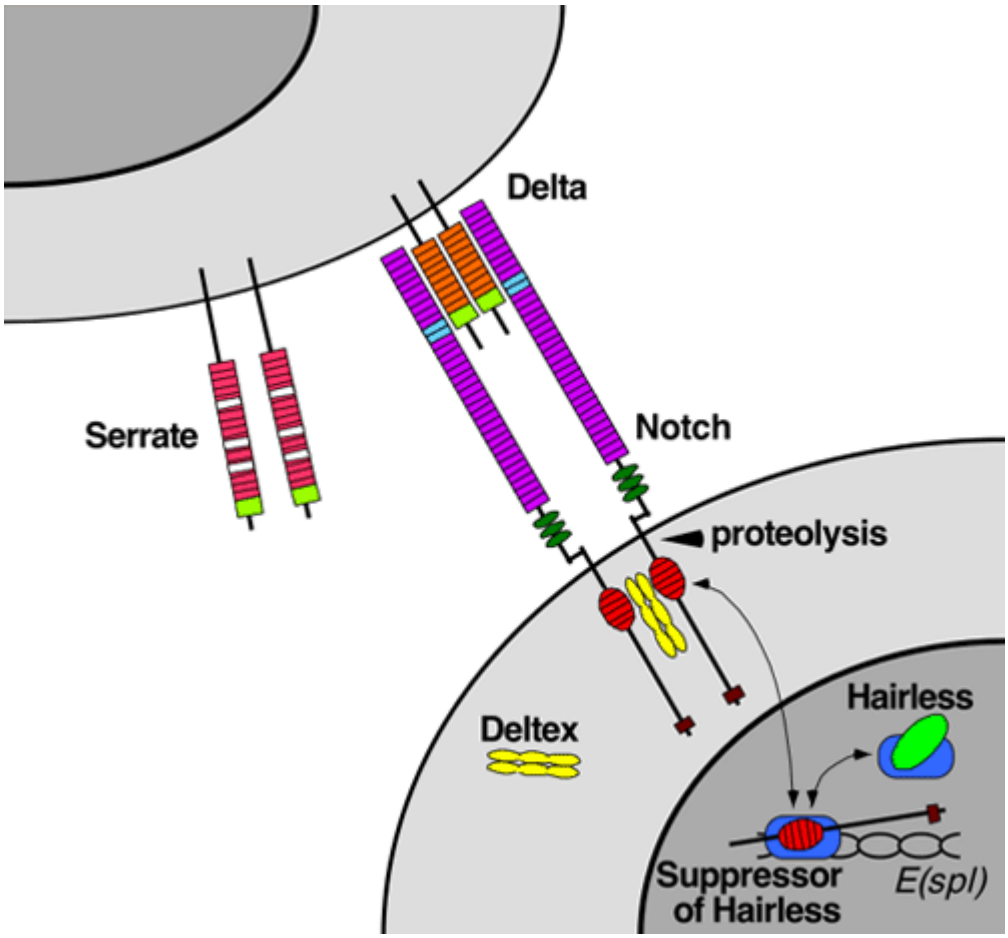


Figure 2. Notch-mediated juxtacrine signal between adjacent cells.

Cells communicate with each other via direct contact (juxtacrine signaling), over short distances (paracrine signaling), or over large distances and/or scales (endocrine signaling).

Some cell-to-cell communication requires direct cell-cell contact. Some cells can form gap junctions that connect their cytoplasm to the cytoplasm of adjacent cells. In cardiac muscle, gap junctions between adjacent cells allows for action potential propagation from the cardiac pacemaker region of the heart to spread and coordinately cause contraction of the heart.

The Notch signaling mechanism is an example of juxtacrine signalling (also known as contact-dependent signaling) in which two adjacent cells must make physical contact in order to communicate. This requirement for direct contact allows for very precise control of cell differentiation during embryonic development. In the worm *Caenorhabditis elegans*, two cells of the developing gonad each have an equal chance of terminally differentiating or becoming a uterine precursor cell that continues to divide. The choice of which cell continues to divide is controlled by competition of cell surface signals. One

cell will happen to produce more of a cell surface protein that activates the Notch receptor on the adjacent cell. This activates a feedback loop or system that reduces Notch expression in the cell that will differentiate and that increases Notch on the surface of the cell that continues as a stem cell.

Many cell signals are carried by molecules that are released by one cell and move to make contact with another cell. *Endocrine* signals are called hormones. Hormones are produced by endocrine cells and they travel through the blood to reach all parts of the body. Specificity of signaling can be controlled if only some cells can respond to a particular hormone. *Paracrine* signals such as retinoic acid target only cells in the vicinity of the emitting cell. Neurotransmitters represent another example of a paracrine signal. Some signaling molecules can function as both a hormone and a neurotransmitter. For example, epinephrine and norepinephrine can function as hormones when released from the adrenal gland and are transported to the heart by way of the blood stream. Norepinephrine can also be produced by neurons to function as a neurotransmitter within the brain. Estrogen can be released by the ovary and function as a hormone or act locally via paracrine or autocrine signaling. Active species of oxygen and nitric oxide can also act as cellular messengers. This process is dubbed redox signaling.

Receptors for cell moves

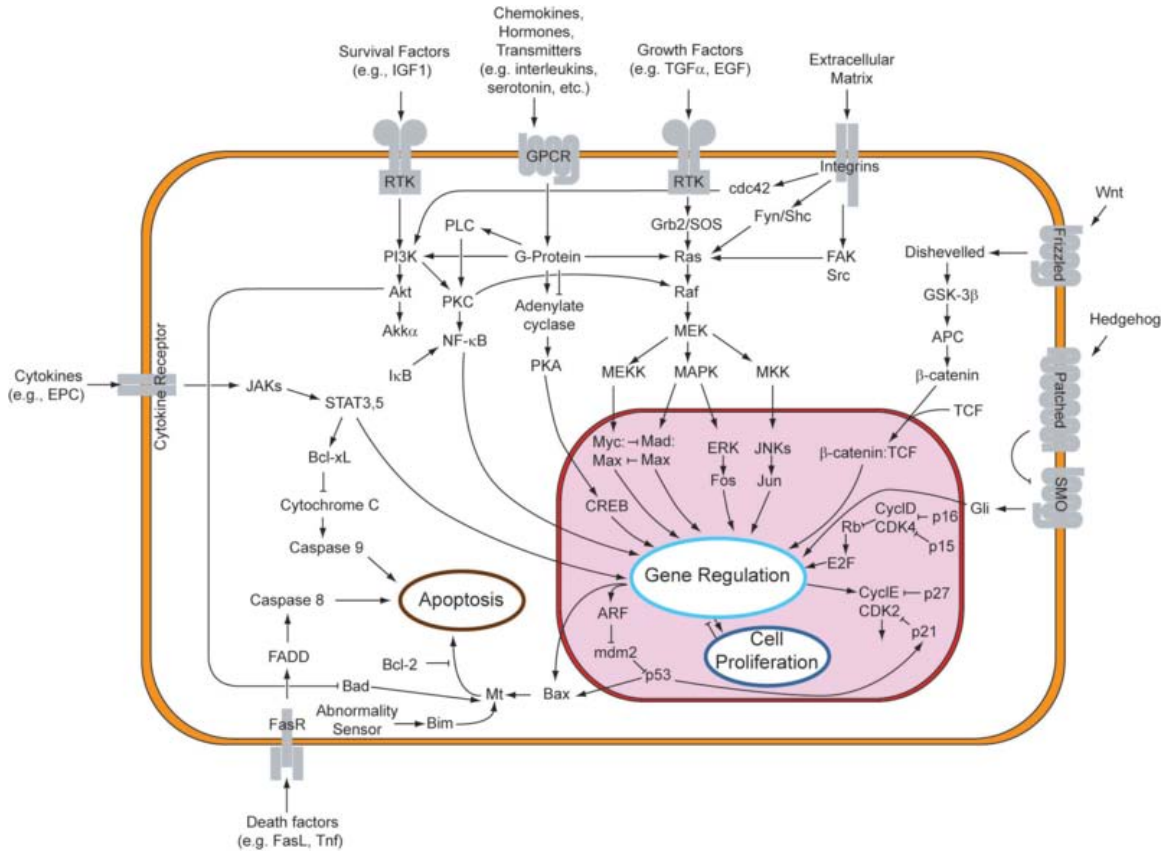
Cells receive information from their environment through a class of proteins known as receptors. Notch is a cell surface protein that functions as a receptor. Animals have a small set of genes that code for signaling proteins that interact specifically with Notch receptors and stimulate a response in cells that express Notch on their surface. Molecules that activate (or, in some cases, inhibit) receptors can be classified as hormones, neurotransmitters, cytokines, growth factors but all of these are called receptor ligands. The details of ligand-receptor interactions are fundamental to cell signaling.

As shown in Figure 2 (above, left), Notch acts as a receptor for ligands that are expressed on adjacent cells. While many receptors are cell surface proteins, some are found inside cells. For example, oestrogen is a hydrophobic molecule that can pass through the lipid bilayer of cell surface membranes. Oestrogen receptors inside cells of the uterus can be activated by oestrogen that comes from the ovaries, enters the target cells, and binds to oestrogen receptors.

A number of transmembrane receptors for molecules that include peptide hormones and of intracellular receptors for steroid hormones exist, giving to a cell the ability to respond to a great number of hormonal and pharmacological stimuli. In diseases, often, proteins that interact with receptors are aberrantly activated, resulting in constitutively activated downstream signals.

For several types of intercellular signaling molecules that are unable to permeate the hydrophobic cell membrane due to their hydrophilic nature, the target receptor is expressed on the membrane. When such signaling molecule activates its receptor, the signal is carried into the cell usually by means of a second messenger such as cAMP.

Signaling pathways



Overview of signal transduction pathways.

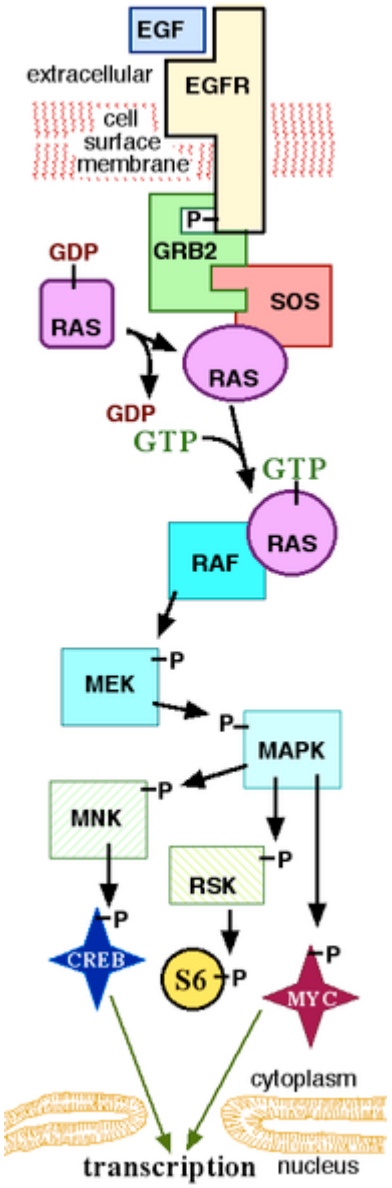


Figure 3. Diagram showing key components of a signal transduction pathway.

In some cases, receptor activation caused by ligand binding to a receptor is directly coupled to the cell's response to the ligand. For example, the neurotransmitter GABA can activate a cell surface receptor that is part of an ion channel. GABA binding to a GABA A receptor on a neuron opens a chloride-selective ion channel that is part of the receptor. GABA A receptor activation allows negatively-charged chloride ions to move into the neuron, which inhibits the ability of the neuron to produce action potentials. However, for many cell surface receptors, ligand-receptor interactions are not directly linked to the cell's response. The activated receptor must first interact with other proteins inside the cell before the ultimate physiological effect of the ligand on the cell's behavior is produced. Often, the behavior of a chain of several interacting cell proteins is altered

following receptor activation. The entire set of cell changes induced by receptor activation is called a signal transduction mechanism or pathway.

In the case of Notch-mediated signaling, the signal transduction mechanism can be relatively simple. As shown in Figure 2 (above, left), activation of Notch can cause the Notch protein to be altered by a protease. Part of the Notch protein is released from the cell surface membrane and can act to change the pattern of gene transcription in the cell nucleus. This causes the responding cell to make different proteins, resulting in an altered pattern of cell behavior. Cell signaling research involves studying the spatial and temporal dynamics of both receptors and the components of signaling pathways that are activated by receptors in various cell types.

A more complex signal transduction pathway is shown in Figure 3. This pathway involves changes of protein-protein interactions inside the cell, induced by an external signal. Many growth factors bind to receptors at the cell surface and stimulate cells to progress through the cell cycle and divide. Several of these receptors are kinases that start to phosphorylate themselves and other proteins when binding to a ligand. This phosphorylation can generate a binding site for a different protein and thus induce protein-protein interaction. In Figure 3, the ligand (called epidermal growth factor (EGF)) binds to the receptor (called EGFR). This activates the receptor to phosphorylate itself. The phosphorylated receptor binds to an adaptor protein (GRB2), which couples the signal to further downstream signaling processes. For example, one of the signal transduction pathways that are activated is called the mitogen-activated protein kinase (MAPK) pathway. The signal transduction component labeled as "MAPK" in the pathway was originally called "ERK," so the pathway is called the MAPK/ERK pathway. The MAPK protein is an enzyme, a protein kinase that can attach phosphate to target proteins such as the transcription factor MYC and, thus, alter gene transcription and, ultimately, cell cycle progression. Many cellular proteins are activated downstream of the growth factor receptors (such as EGFR) that initiate this signal transduction pathway.

Some signaling transduction pathways respond differently depending on the amount of signaling received by the cell. For instance, the hedgehog protein activates different genes, depending on the amount of hedgehog protein present.

Complex multi-component signal transduction pathways provide opportunities for feedback, signal amplification, and interactions inside one cell between multiple signals and signaling pathways.

Classification of intercellular communication

Within endocrinology (the study of intercellular signalling in animals) and the endocrine system, intercellular signalling is subdivided into the following classifications:

- *Intracrine* signals are produced within the target cell.

- *Autocrine* signals target the cell itself. Sometimes autocrine cells can target cells close by if they are the same type of cell as the emitting cell. An example of this are immune cells.
- *Juxtacrine* signals target adjacent (touching) cells. These signals are transmitted along cell membranes via protein or lipid components integral to the membrane and are capable of affecting either the emitting cell or cells immediately adjacent.
- *Paracrine* signals target cells in the vicinity of the emitting cell. Neurotransmitters represent an example.
- *Endocrine* signals target distant cells. Endocrine cells produce hormones that travel through the blood to reach all parts of the body.