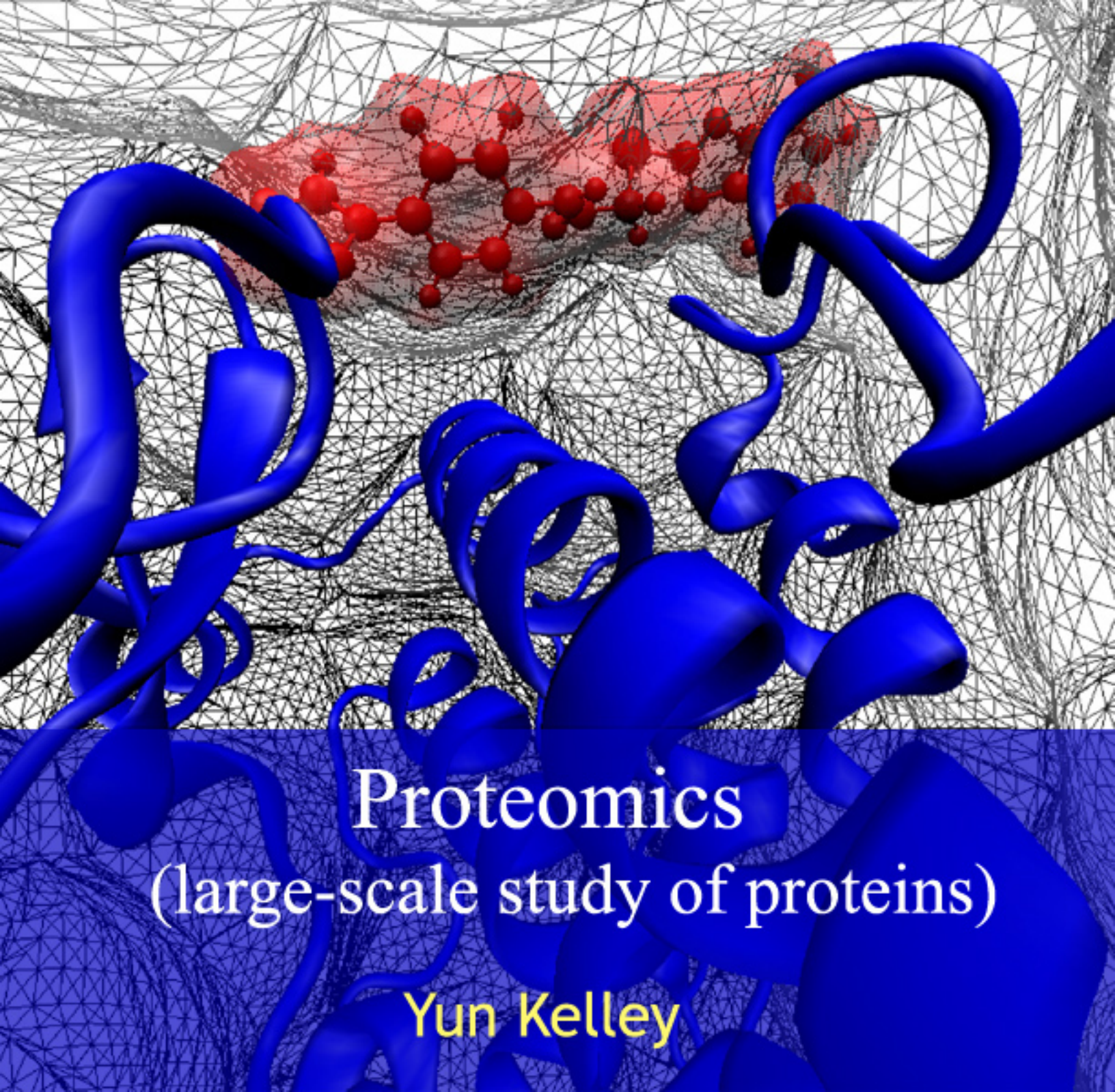


Proteomics
(large-scale study of proteins)

Yun Kelley

First Edition, 2012

ISBN 978-81-323-4321-9

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

Table of Contents

Introduction

Chapter 1 - Biochip & Cleavable Detergent

Chapter 2 - Intrinsically Unstructured Proteins

Chapter 3 - Protein Mass Spectrometry

Chapter 4 - Protein Sequencing

Chapter 5 - Proteomics Identifications Database, Protomap & Protein–Protein Interaction

Chapter 6 - Protein–Protein Interaction Prediction

Chapter 7 - Neuroproteomics

Chapter 8 - Quantitative Proteomics & SILAC

Chapter 9 - Phosphoproteomics

Chapter 10 - Activity-Based Proteomics

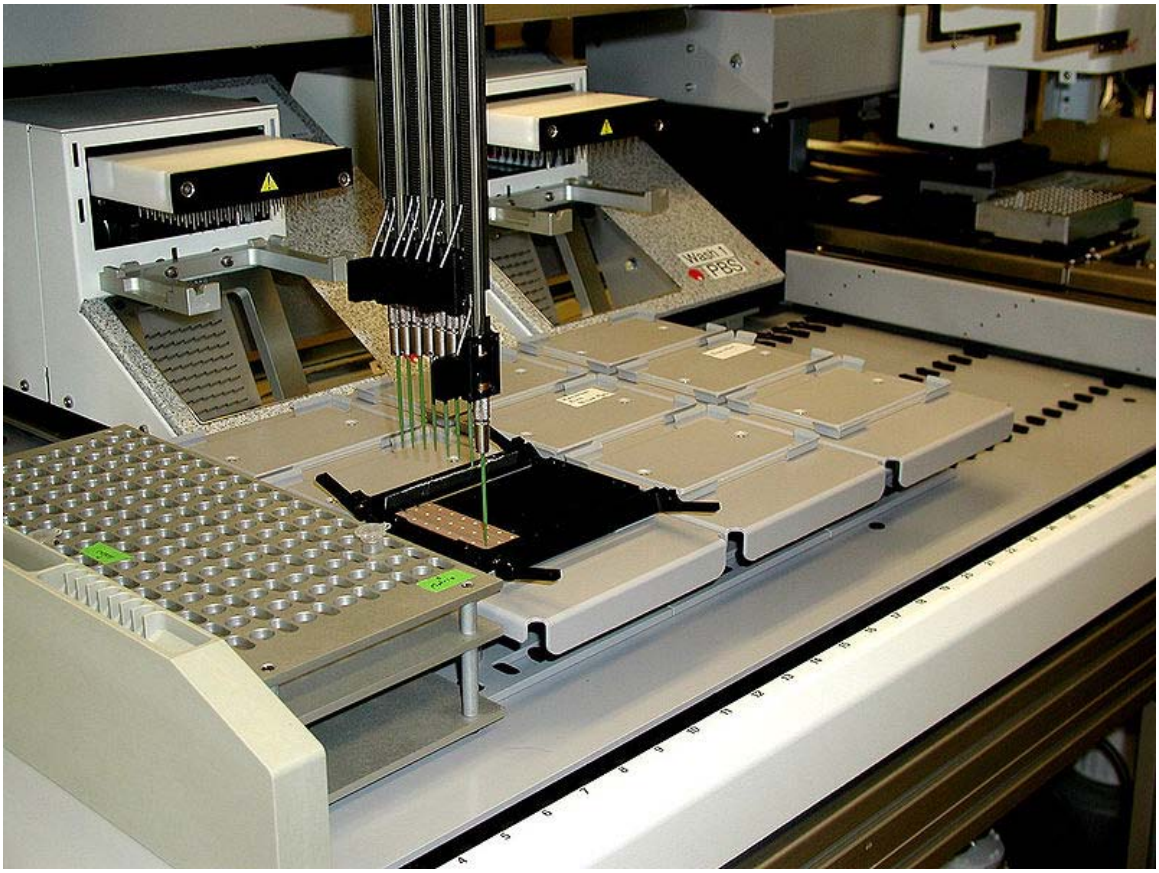
Chapter 11 - Enzyme

Chapter 12 - Receptor (Biochemistry)

Chapter 13 - Chaperone (Protein)

Chapter 14 - Multiprotein Complex

Introduction



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

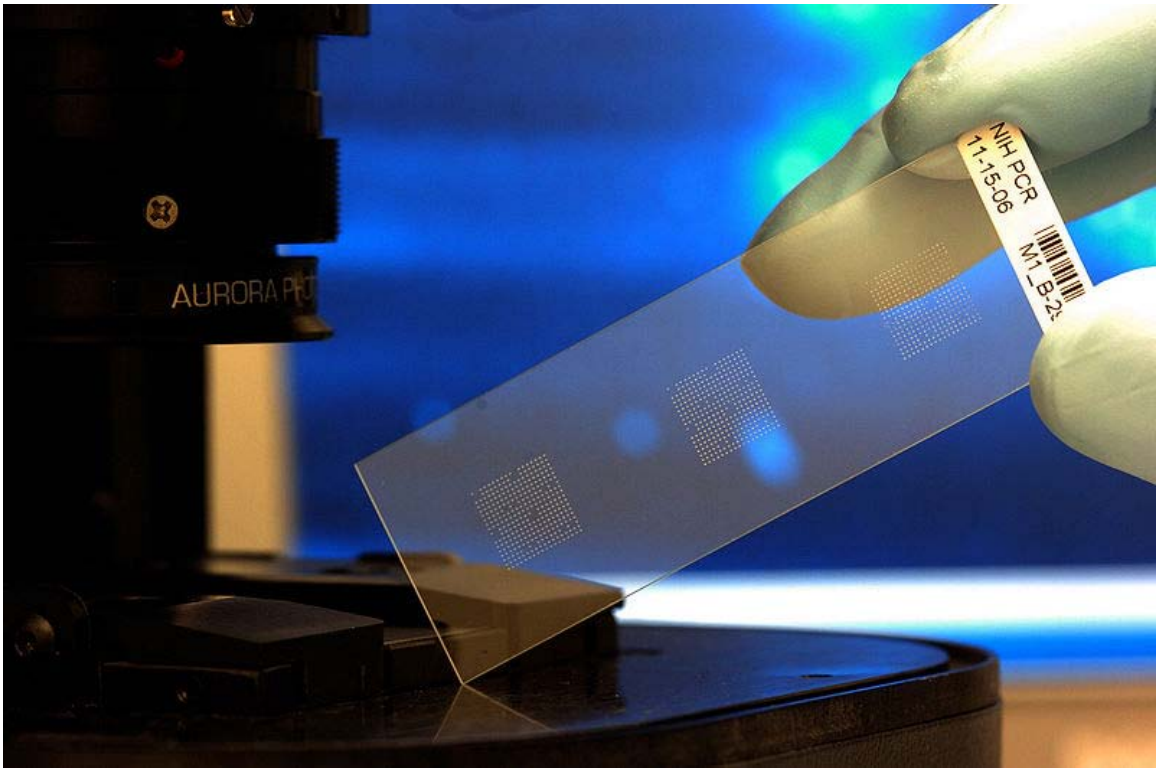
There are many approaches to attempting to characterize the human proteome, which is estimated to exceed 100,000 unique forms, 25,000 genes plus post-translational modifications.

In addition, first promising attempts to decipher the proteom of animal tumors have recently been reported.

Chapter 1

Biochip & Cleavable Detergent

Biochip



Hundreds of gel drops are visible on the biochip

The development of **biochips** is a major thrust of the rapidly growing biotechnology industry, which encompasses a very diverse range of research efforts including genomics,

proteomics, and pharmaceuticals, among other activities. Advances in these areas are giving scientists new methods for unravelling the complex biochemical processes occurring inside cells, with the larger goal of understanding and treating human diseases. At the same time, the semiconductor industry has been steadily perfecting the science of micro-miniaturization. The merging of these two fields in recent years has enabled biotechnologists to begin packing their traditionally bulky sensing tools into smaller and smaller spaces, onto so-called biochips. These chips are essentially miniaturized laboratories that can perform hundreds or thousands of simultaneous biochemical reactions. Biochips enable researchers to quickly screen large numbers of biological analytes for a variety of purposes, from disease diagnosis to detection of bioterrorism agents.

History

The development of biochips has a long history, starting with early work on the underlying sensor technology. One of the first portable, chemistry-based sensors was the glass pH electrode, invented in 1922 by Hughes (Hughes, 1922). Measurement of pH was accomplished by detecting the potential difference developed across a thin glass membrane selective to the permeation of hydrogen ions; this selectivity was achieved by exchanges between H^+ and SiO sites in the glass. The basic concept of using exchange sites to create permselective membranes was used to develop other ion sensors in subsequent years. For example, a K^+ sensor was produced by incorporating valinomycin into a thin membrane (Schultz, 1996). Over thirty years elapsed before the first true biosensor (*i.e.* a sensor utilizing biological molecules) emerged. In 1956, Leland Clark published a paper on an oxygen sensing electrode (Clark, 1956_41). This device became the basis for a glucose sensor developed in 1962 by Clark and colleague Lyons which utilized glucose oxidase molecules embedded in a dialysis membrane (Clark, 1962). The enzyme functioned in the presence of glucose to decrease the amount of oxygen available to the oxygen electrode, thereby relating oxygen levels to glucose concentration. This and similar biosensors became known as enzyme electrodes, and are still in use today.

In 1953, Watson and Crick announced their discovery of the now familiar double helix structure of DNA molecules and set the stage for genetics research that continues to the present day (Nelson, 2000). The development of sequencing techniques in 1977 by Gilbert (Maxam, 1977) and Sanger (Sanger, 1977) (working separately) enabled researchers to directly read the genetic codes that provide instructions for protein synthesis. This research showed how hybridization of complementary single oligonucleotide strands could be used as a basis for DNA sensing. Two additional developments enabled the technology used in modern DNA-based biosensors. First, in 1983 Kary Mullis invented the polymerase chain reaction (PCR) technique (Nelson, 2000), a method for amplifying DNA concentrations. This discovery made possible the detection of extremely small quantities of DNA in samples. Second, in 1986 Hood and co-workers devised a method to label DNA molecules with fluorescent tags instead of radiolabels (Smith, 1986), thus enabling hybridization experiments to be observed optically.

The rapid technological advances of the biochemistry and semiconductor fields in the 1980s led to the large scale development of biochips in the 1990s. At this time, it became clear that biochips were largely a "platform" technology which consisted of several separate, yet integrated components. Figure 1 shows the make up of a typical biochip platform. The actual sensing component (or "chip") is just one piece of a complete analysis system. Transduction must be done to translate the actual sensing event (DNA binding, oxidation/reduction, *etc.*) into a format understandable by a computer (voltage, light intensity, mass, *etc.*), which then enables additional analysis and processing to produce a final, human-readable output. The multiple technologies needed to make a successful biochip — from sensing chemistry, to microarraying, to signal processing — require a true multidisciplinary approach, making the barrier to entry steep. One of the first commercial biochips was introduced by Affymetrix. Their "GeneChip" products contain thousands of individual DNA sensors for use in sensing defects, or single nucleotide polymorphisms (SNPs), in genes such as p53 (a tumor suppressor) and BRCA1 and BRCA2 (related to breast cancer) (Cheng, 2001). The chips are produced using microlithography techniques traditionally used to fabricate integrated circuits.

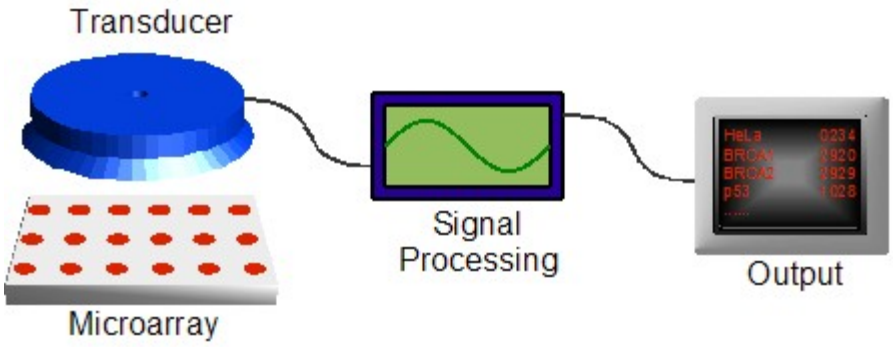
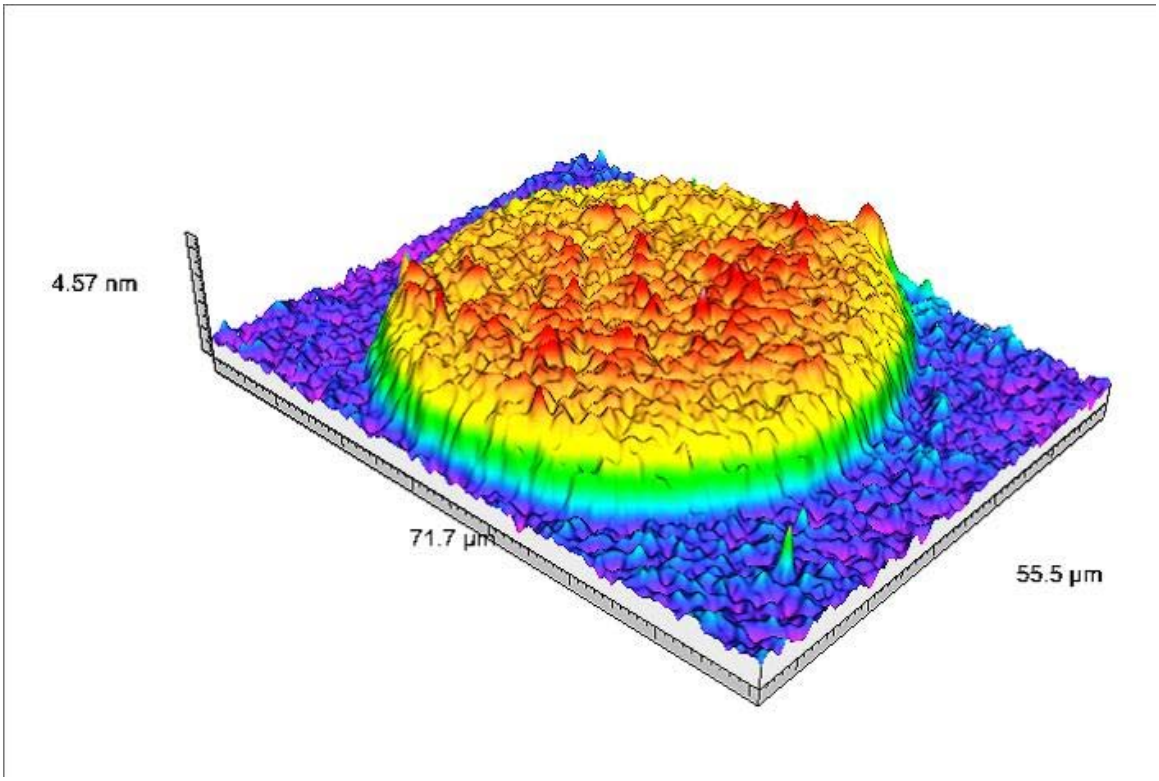


Figure 1. Biochips are a platform that require, in addition to microarray technology, transduction and signal processing technologies to output the results of sensing experiments.

Today, a large variety of biochip technologies are either in development or being commercialized. Numerous advancements continue to be made in sensing research that enable new platforms to be developed for new applications. Cancer diagnosis through DNA typing is just one market opportunity. A variety of industries currently desire the ability to simultaneously screen for a wide range of chemical and biological agents, with purposes ranging from testing public water systems for disease agents to screening airline cargo for explosives. Pharmaceutical companies wish to combinatorially screen drug candidates against target enzymes. To achieve these ends, DNA, RNA, proteins, and even living cells are being employed as sensing mediators on biochips (Potera, 2008). Numerous transduction methods can be employed including surface plasmon resonance, fluorescence, and chemiluminescence. The particular sensing and transduction techniques chosen depend on factors such as price, sensitivity, and reusability.

Microarray fabrication



3D Sarfus image of a DNA biochip.

The microarray — the dense, two-dimensional grid of biosensors — is the critical component of a biochip platform. Typically, the sensors are deposited on a flat substrate, which may either be passive (*e.g.* silicon or glass) or active, the latter consisting of integrated electronics or micromechanical devices that perform or assist signal transduction. Surface chemistry is used to covalently bind the sensor molecules to the substrate medium. The fabrication of microarrays is non-trivial and is a major economic and technological hurdle that may ultimately decide the success of future biochip platforms. The primary manufacturing challenge is the process of placing each sensor at a specific position (typically on a Cartesian grid) on the substrate. Various means exist to achieve the placement, but typically robotic micro-pipetting (Schena, 1995) or micro-printing (MacBeath, 1999) systems are used to place tiny spots of sensor material on the chip surface. Because each sensor is unique, only a few spots can be placed at a time. The low-throughput nature of this process results in high manufacturing costs.

Fodor and colleagues developed a unique fabrication process (later used by Affymetrix) in which a series of microlithography steps is used to combinatorially synthesize hundreds of thousands of unique, single-stranded DNA sensors on a substrate one nucleotide at a time (Fodor, 1991; Pease, 1994). One lithography step is needed per base type; thus, a total of four steps is required per nucleotide level. Although this technique is very powerful in that many sensors can be created simultaneously, it is currently only feasible for creating short DNA strands (15–25 nucleotides). Reliability and cost factors

limit the number of photolithography steps that can be done. Furthermore, light-directed combinatorial synthesis techniques are not currently possible for proteins or other sensing molecules.

As noted above, most microarrays consist of a Cartesian grid of sensors. This approach is used chiefly to map or "encode" the coordinate of each sensor to its function. Sensors in these arrays typically use a universal signalling technique (*e.g.* fluorescence), thus making coordinates their only identifying feature. These arrays must be made using a serial process (*i.e.* requiring multiple, sequential steps) to ensure that each sensor is placed at the correct position.

"Random" fabrication, in which the sensors are placed at arbitrary positions on the chip, is an alternative to the serial method. The tedious and expensive positioning process is not required, enabling the use of parallelized self-assembly techniques. In this approach, large batches of identical sensors can be produced; sensors from each batch are then combined and assembled into an array. A non-coordinate based encoding scheme must be used to identify each sensor. As the figure shows, such a design was first demonstrated (and later commercialized by Illumina) using functionalized beads placed randomly in the wells of an etched fiber optic cable (Steemers, 2000; Michael, 1998) Each bead was uniquely encoded with a fluorescent signature. However, this encoding scheme is limited in the number of unique dye combinations that can be used and successfully differentiated.

Protein biochip array and other microarray technologies

Microarrays are not limited to DNA analysis; protein microarrays, antibody microarray, chemical compound microarray can also be produced using biochips. Randox Laboratories Ltd. launched Evidence, the first protein Biochip Array Technology analyzer in 2003. In protein Biochip Array Technology, the biochip replaces the ELISA plate or cuvette as the reaction platform. The biochip is used to simultaneously analyze a panel of related tests in a single sample, producing a patient profile. The patient profile can be used in disease screening, diagnosis, monitoring disease progression or monitoring treatment. Performing multiple analyses simultaneously, described as multiplexing, allows a significant reduction in processing time and the amount of patient sample required. Biochip Array Technology is a novel application of a familiar methodology, using sandwich, competitive and antibody-capture immunoassays. The difference from conventional immunoassays is that the capture ligands are covalently attached to the surface of the biochip in an ordered array rather than in solution.

In sandwich assays an enzyme-labelled antibody is used; in competitive assays an enzyme-labelled antigen is used. On antibody-antigen binding a chemiluminescence reaction produces light. Detection is by a charge-coupled device (CCD) camera. The CCD camera is a sensitive and high-resolution sensor able to accurately detect and quantify very low levels of light. The test regions are located using a grid pattern then the chemiluminescence signals are analysed by imaging software to rapidly and simultaneously quantify the individual analytes.

Details about other array technologies can be found in the following page: Antibody microarray

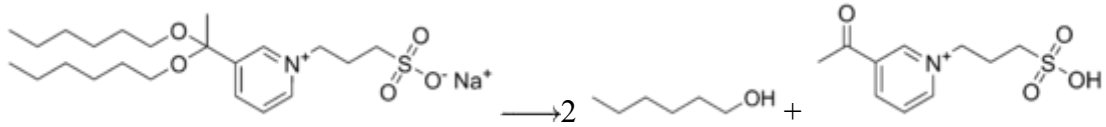
Cleavable Detergent

Cleavable detergents, also known as **cleavable surfactants**, are special surfactants (detergents) that are used in biochemistry and especially in proteomics to enhance protein denaturation and solubility. The detergent is rendered inactive by cleavage, usually under acidic conditions, in order to make the sample compatible with a following procedure or in order to selectively remove the cleavage products.

Applications for cleavable detergents include protease digestion of proteins such as in-gel digestion with trypsin after SDS PAGE and peptide extractions from electrophoresis gels. Cleavable detergents are mainly used in sample preparations for mass spectrometry.

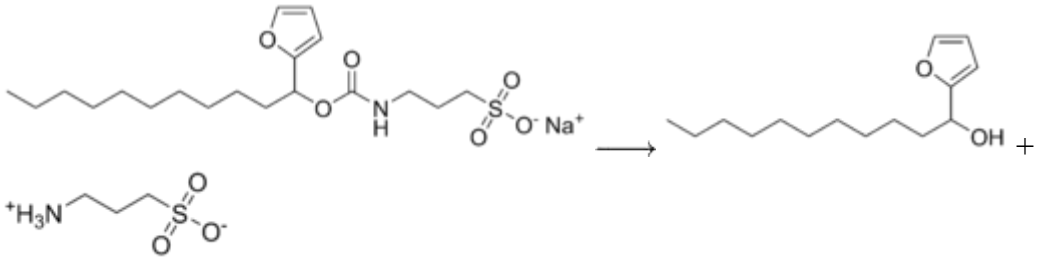
PPS

PPS, available as **PPS Silent Surfactant** from Protein Discovery, is the abbreviation for sodium 3-(4-(1,1-bis(hexyloxy)ethyl)pyridinium-1-yl)propane-1-sulfonate. This acetalic detergent is split under acidic conditions into hexanol and the zwitterionic 3-acetyl-1-(3-sulfopropyl)pyridinium.



ProteaseMAX

ProteaseMAX is the brandname of Promega for sodium 3-((1-(furan-2-yl)undecyloxy)carbonylamino)propane-1-sulfonate. This cleavable detergent is sensitive to heat and acid and is degraded during a typical trypsin digestion into the uncharged lipophilic compound 1-(furan-2-yl)undecan-1-ol and the zwitterionic 3-aminopropane-1-sulfonic acid (homotaurine), which can be removed by C18 solid phase extraction during sample work-up.

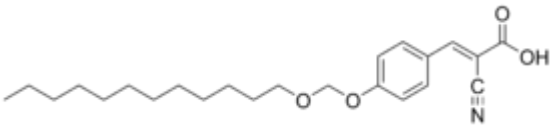


RapiGest SF

RapiGest SF is an acid-cleavable anionic detergent marketed by Waters Corporation. No information about its chemical structure has been published.

Others

MALDI matrix compounds such as α-cyano-4-hydroxycinnamic acid have been linked through a linker consisting of an unsymmetric formaldehyde acetals to dodecanol. This type of cleavable detergent is inherently compatible with MALDI and does not have to be removed prior to analysis:



UV light- or fluoride-cleavable surfactants have also been developed but are not in current use.

Chapter 2

Intrinsically Unstructured Proteins

Intrinsically unstructured proteins, often referred to as *naturally unfolded proteins* or *disordered proteins*, are proteins characterized by lack of stable tertiary structure when the protein exists as an isolated polypeptide chain (a subunit) under physiological conditions in vitro. The discovery of intrinsically unfolded proteins challenged the traditional protein structure paradigm, which states that a specific well-defined structure was required for the correct function of a protein and that the structure defines the function of the protein. This is clearly not the case for intrinsically unfolded proteins that remain functional despite the lack of a well-defined structure. Such proteins, in some cases, can adopt a fixed three dimensional structure after binding to other macromolecules.

Biological role of intrinsic disorder

Many disordered proteins have the binding affinity with their receptors regulated by post-translational modification, thus it has been proposed that the flexibility of disordered proteins facilitates the different conformational requirements for binding the modifying enzymes as well as their receptors. Intrinsic disorder is particularly enriched in proteins implicated in cell signaling, transcription and chromatin remodeling functions.

Flexible linkers

Disordered regions are often found as flexible linkers connecting two globular domains. Linker sequences vary greatly in length and amino acid sequence, but are similar in amino acid composition (rich in polar uncharged amino acids). Flexible linkers allow the connecting domains to freely twist and rotate through space to recruit their binding partners.

Coupled folding and binding

Many unstructured proteins undergo transitions to more ordered states upon binding to their targets. The coupled folding and binding may be local, involving only a few interacting residues, or it might involve an entire protein domain. It was recently shown that the coupled folding and binding allows the burial of a large surface area that would only be possible for fully structured proteins if they were much larger. Moreover, certain disordered regions might serve as "molecular switches" in regulating certain biological function by switching to ordered conformation upon molecular recognition like small molecule-binding, DNA/RNA binding, ion interactions etc.

The ability of disordered proteins to bind, and thus to exert a function, shows that stability is not a required condition.

Sequence signatures of disorder

Intrinsically unstructured proteins are characterized by a low content of bulky hydrophobic amino acids and a high proportion of polar and charged amino acids. Thus disordered sequences cannot bury sufficient hydrophobic core to fold like stable globular proteins. In some cases, hydrophobic clusters in disordered sequences provide the clues for identifying the regions that undergo coupled folding and binding.

Many disordered proteins also reveal low complexity sequences, i.e. sequences with overrepresentation of a few residues. While low complexity sequences are a strong indication of disorder, the reverse is not necessarily true, that is, not all disordered proteins have low complexity sequences.

Disordered proteins have a low content of predicted secondary structure.

Identification of intrinsically unstructured proteins

Intrinsically unfolded proteins, once purified, can be identified by various experimental methods. Folded proteins have a high density (partial specific volume of 0.72-0.74 mL/g) and commensurately small radius of gyration. Hence, unfolded proteins can be detected by methods that are sensitive to molecular size, density or hydrodynamic drag, such as size exclusion chromatography, analytical ultracentrifugation, Small angle X-ray scattering (SAXS), and measurements of the diffusion constant. Unfolded proteins are also characterized by their lack of secondary structure, as assessed by far-UV (170-250 nm) circular dichroism (esp. a pronounced minimum at ~200 nm) or infrared spectroscopy.

Unfolded proteins have exposed backbone peptide groups exposed to solvent, so that they are readily cleaved by proteases, undergo rapid hydrogen-deuterium exchange and exhibit a small dispersion (<1 ppm) in their 1H amide chemical shifts as measured by NMR. (Folded proteins typically show dispersions as large as 5 ppm for the amide protons.)

The primary method to obtain information on disordered regions of a protein is NMR spectroscopy. The lack of electron density in X-ray crystallographic studies may also be a sign of disorder.

De novo *prediction of intrinsically unstructured proteins*

Computational methods exploit the sequence signatures of disorder to predict whether a protein is disordered given its amino acid sequence. The table below, which was originally adapted from and has been recently updated, shows the main features of software for disorder prediction. Note that different software use different definitions of disorder.

Predictor	What is predicted	Based on	Generates and uses multiple sequence alignment?
PONDR	All regions that are not rigid including random coils, partially unstructured regions, and molten globules	Local aa composition, flexibility, hydropathy, etc	No
SEG	Low-complexity segments that is, "simple sequences" or "compositionally biased regions".	Locally optimized low-complexity segments are produced at defined levels of stringency and then refined according to the equations of Wootton and Federhen	No
Disopred2	Regions devoid of ordered regular secondary structure	Cascaded support vector machine classifiers trained on PSI-BLAST profiles	Yes
Globplot	Regions with high propensity for globularity on the Russell/Linding scale (propensities for secondary structures and random coils)	Russell/Linding scale of disorder	No
Disembl	LOOPS (regions devoid of regular secondary structure); HOT LOOPS	Neural networks trained on X-ray structure data	No

	(highly mobile loops); REMARK465 (regions lacking electron density in crystal structure)		
NORSp	Regions with No Ordered Regular Secondary Structure (NORS). Most, but not all, are highly flexible.	Secondary structure and solvent accessibility	Yes
FoldIndex	Regions that have a low hydrophobicity and high net charge (either loops or unstructured regions)	Charge/hydrophaty analyzed locally using a sliding window	No
Charge/hydropathy method.	Fully unstructured domains (random coils)	Global sequence composition	No
HCA (Hydrophobic Cluster Analysis)	Hydrophobic clusters, which tend to form secondary structure elements	Helical visualization of amino acid sequence	No
PreLink	Regions that are expected to be unstructured in all conditions, regardless of the presence of a binding partner	Compositional bias and low hydrophobic cluster content.	No
IUPred	Regions that lack a well-defined 3D-structure under native conditions	Energy resulting from inter-residue interactions, estimated from local amino acid composition	No
RONN	Regions that lack a well-defined 3D structure under native conditions	Bio-basis function neural network trained on disordered proteins	No
MD (Meta-Disorder predictor)	Regions of different "types"; for example, unstructured loops and regions containing few	A neural-network based meta-predictor that uses different sources of information predominantly obtained from orthogonal	Yes

	stable intra-chain contacts	approaches	
GeneSilico Metadisorder	Regions that lack a well-defined 3D structure under native conditions (REMARK-465)	Meta method which uses other disorder predictors (like RONN, IUPred, POODLE and many more). Based on them the consensus is calculated according method accuracy (optimized using ANN, filtering and other techniques). Currently the best available method (first 2 places in last CASP experiment (blind test))	Yes
IUPforest-L	Long disordered regions in a set of proteins	Moreau-Broto auto-correlation function of amino acid indices (AAIs)	No
MFDp	Different types of disorder including random coils, unstructured regions, molten globules, and REMARK-465-based regions.	An ensemble of 3 SVMs specialized for the prediction of short, long and generic disordered regions, which combines three complementary disorder predictors, sequence, sequence profiles, predicted secondary structure, solvent accessibility, backbone dihedral torsion angles, residue flexibility and B-factors. MFDp (unofficially) secured 3rd place in last CASP experiment)	Yes

Since the methods above use different definitions of disorder and they were trained on different datasets, it is difficult to estimate their relative accuracy, but disorder prediction category is a part of biannual CASP experiment which is designed to test methods according accuracy in finding regions with missing 3D structure.

Chapter 3

Protein Mass Spectrometry

A mass spectrometer used for high throughput protein analysis.

Protein mass spectrometry refers to the application of mass spectrometry to the study of proteins. Mass spectrometry is an important emerging method for the characterization of

proteins. The two primary methods for ionization of whole proteins are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). In keeping with the performance and mass range of available mass spectrometers, two approaches are used for characterizing proteins. In the first, intact proteins are ionized by either of the two techniques described above, and then introduced to a mass analyzer. This approach is referred to as "top-down" strategy of protein analysis. In the second, proteins are enzymatically digested into smaller peptides using a protease such as trypsin. Subsequently these peptides are introduced into the mass spectrometer and identified by peptide mass fingerprinting or tandem mass spectrometry. Hence, this latter approach (also called "bottom-up" proteomics) uses identification at the peptide level to infer the existence of proteins.

Whole protein mass analysis is primarily conducted using either time-of-flight (TOF) MS, or Fourier transform ion cyclotron resonance (FT-ICR). These two types of instrument are preferable here because of their wide mass range, and in the case of FT-ICR, its high mass accuracy. Mass analysis of proteolytic peptides is a much more popular method of protein characterization, as cheaper instrument designs can be used for characterization. Additionally, sample preparation is easier once whole proteins have been digested into smaller peptide fragments. The most widely used instrument for peptide mass analysis are the MALDI time-of-flight instruments as they permit the acquisition of peptide mass fingerprints (PMFs) at high pace (1 PMF can be analyzed in approx. 10 sec). Multiple stage quadrupole-time-of-flight and the quadrupole ion trap also find use in this application.

Protein and peptide fractionation coupled with mass spectrometry

Proteins of interest to biological researchers are usually part of a very complex mixture of other proteins and molecules that co-exist in the biological medium. This presents two significant problems. First, the two ionization techniques used for large molecules only work well when the mixture contains roughly equal amounts of constituents, while in biological samples, different proteins tend to be present in widely differing amounts. If such a mixture is ionized using electrospray or MALDI, the more abundant species have a tendency to "drown" or suppress signals from less abundant ones. The second problem is that the mass spectrum from a complex mixture is very difficult to interpret because of the overwhelming number of mixture components. This is exacerbated by the fact that enzymatic digestion of a protein gives rise to a large number of peptide products.

To contend with this problem, two methods are widely used to fractionate proteins, or their peptide products from an enzymatic digestion. The first method fractionates whole proteins and is called two-dimensional gel electrophoresis. The second method, high performance liquid chromatography is used to fractionate peptides after enzymatic digestion. In some situations, it may be necessary to combine both of these techniques.

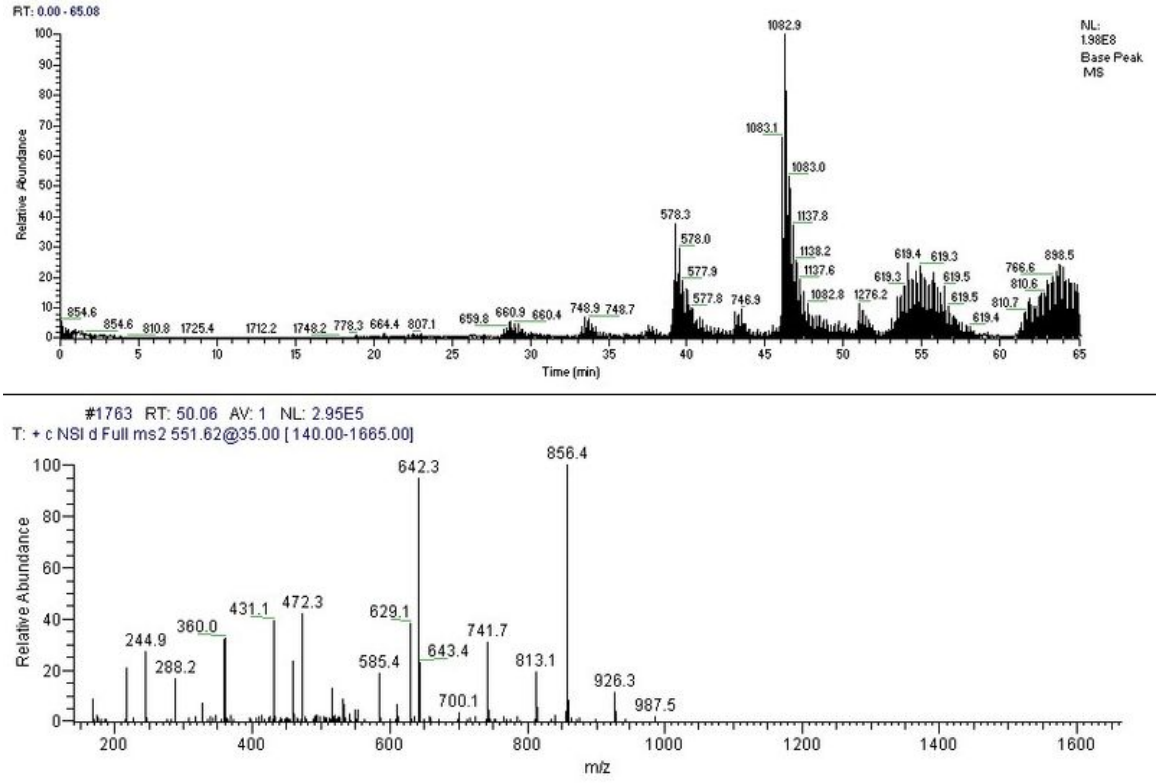
Gel spots identified on a 2D Gel are usually attributable to one protein. If the identity of the protein is desired, usually the method of in-gel digestion is applied, where the protein

spot of interest is excised, and digested proteolytically. The peptide masses resulting from the digestion can be determined by mass spectrometry using peptide mass fingerprinting. If this information does not allow unequivocal identification of the protein, its peptides can be subject to tandem mass spectrometry for de novo sequencing.

Characterization of protein mixtures using HPLC/MS is also called *shotgun proteomics* and *mudpit*. A peptide mixture that results from digestion of a protein mixture is fractionated by one or two steps of liquid chromatography. The eluent from the chromatography stage can be either directly introduced to the mass spectrometer through electrospray ionization, or laid down on a series of small spots for later mass analysis using MALDI.

Protein identification

There are two main ways MS is used to identify proteins. Peptide mass fingerprinting (mentioned in the previous section) uses the masses of proteolytic peptides as input to a search of a database of predicted masses that would arise from digestion of a list of known proteins. If a protein sequence in the reference list gives rise to a significant number of predicted masses that match the experimental values, there is some evidence that this protein was present in the original sample.



Chromatography trace and MS^2 spectra of a peptide.

Tandem MS is becoming a more popular experimental method for identifying proteins. Collision-induced dissociation is used in mainstream applications to generate a set of fragments from a specific peptide ion. The fragmentation process primarily gives rise to cleavage products that break along peptide bonds. Because of this simplicity in fragmentation, it is possible to use the observed fragment masses to match with a database of predicted masses for one of many given peptide sequences. Tandem MS of whole protein ions has been investigated recently using electron capture dissociation and has demonstrated extensive sequence information in principle but is not in common practice. This is sometimes referred to as the "top-down" approach in that it involves starting with the whole mass and then pulling it apart rather than starting with pieces (proteolytic fragments) and piecing the protein back together using de novo repeat detection (bottom-up).

De novo (peptide) sequencing

De novo (peptide) sequencing for mass spectrometry is typically performed without prior knowledge of the amino acid sequence. It is the process of assigning amino acids from peptide fragment masses of a protein. De novo sequencing has proven successful for confirming and expanding upon results from database searches.

As de novo sequencing is based on mass and some amino acids have identical masses (e.g. leucine and isoleucine), accurate manual sequencing can be difficult. Therefore it may be necessary to utilize a sequence homology search application to work in tandem between a database search and de novo sequencing to address this inherent limitation.

Database searching has the advantage of quickly identifying sequences, provided they have already been documented in a database. Other inherent limitations of database searching include:

- Sequence modifications/mutations: some database searches do not adequately account for alterations to the 'documented' sequence, thus can miss valuable information.
- The unknown: if a sequence is not documented, it will not be found
- False positives
- Incomplete and corrupted data: a common, unnoticed problem

Software

A number of different algorithmic approaches have been described to identify peptides and proteins from tandem mass spectrometry (MS/MS), peptide de novo sequencing and sequence tag-based searching.

Protein quantitation

Several recent methods allow for the quantitation of proteins by mass spectrometry (quantitative proteomics). Typically, stable (e.g. non-radioactive) heavier isotopes of

carbon (^{13}C) or nitrogen (^{15}N) are incorporated into one sample while the other one is labeled with corresponding light isotopes (e.g. ^{12}C and ^{14}N). The two samples are mixed before the analysis. Peptides derived from the different samples can be distinguished due to their mass difference. The ratio of their peak intensities corresponds to the relative abundance ratio of the peptides (and proteins). The most popular methods for isotope labeling are SILAC (stable isotope labeling by amino acids in cell culture), trypsin-catalyzed ^{18}O labeling, ICAT (isotope coded affinity tagging), iTRAQ (isobaric tags for relative and absolute quantitation). “Semi-quantitative” mass spectrometry can be performed without labeling of samples. Typically, this is done with MALDI analysis (in linear mode). The peak intensity, or the peak area, from individual molecules (typically proteins) is here correlated to the amount of protein in the sample. However, the individual signal depends on the primary structure of the protein, on the complexity of the sample, and on the settings of the instrument. Other types of "label-free" quantitative mass spectrometry, uses the spectral counts (or peptide counts) of digested proteins as a means for determining relative protein amounts.

Protein structure

Characteristics indicative of the 3-dimensional structure of proteins can be probed with mass spectrometry in various ways. By using chemical crosslinking to couple parts of the protein that are close in space, but far apart in sequence, information about the overall structure can be inferred. By following the exchange of amide protons with deuterium from the solvent, it is possible to probe the solvent accessibility of various parts of the protein.

Chapter 4

Protein Sequencing

Protein sequencing is a technique to determine the amino acid sequence of a protein, as well as which conformation the protein adopts and the extent to which it is complexed with any non-peptide molecules. Discovering the structures and functions of proteins in living organisms is an important tool for understanding cellular processes, and allows drugs that target specific metabolic pathways to be invented more easily.

The two major direct methods of protein sequencing are mass spectrometry and the Edman degradation reaction. It is also possible to generate an amino acid sequence from the DNA or mRNA sequence encoding the protein, if this is known. However, there are a number of other reactions which can be used to gain more limited information about protein sequences and can be used as preliminaries to the aforementioned methods of sequencing or to overcome specific inadequacies within them.

Determining amino acid composition

It is often desirable to know the unordered amino acid composition of a protein prior to attempting to find the ordered sequence, as this knowledge can be used to facilitate the discovery of errors in the sequencing process or to distinguish between ambiguous results. Knowledge of the frequency of certain amino acids may also be used to choose which protease to use for digestion of the protein. A generalised method for doing this is as follows:

1. Hydrolyse a known quantity of protein into its constituent amino acids.
2. Separate the amino acids in some way.

Hydrolysis

Hydrolysis is done by heating a sample of the protein in 6 Molar hydrochloric acid to 100-110 degrees Celsius for 24 hours or longer. Proteins with many bulky hydrophobic groups may require longer heating periods. However, these conditions are so vigorous that some amino acids (serine, threonine, tyrosine, tryptophan, glutamine and cystine) are degraded. To circumvent this problem, Biochemistry Online suggests heating separate samples for different times, analysing each resulting solution, and extrapolating back to zero hydrolysis time. Rastall suggests a variety of reagents to prevent or reduce degradation - thiol reagents or phenol to protect tryptophan and tyrosine from attack by chlorine, and pre-oxidising cysteine. He also suggests measuring the quantity of ammonia evolved to determine the extent of amide hydrolysis.

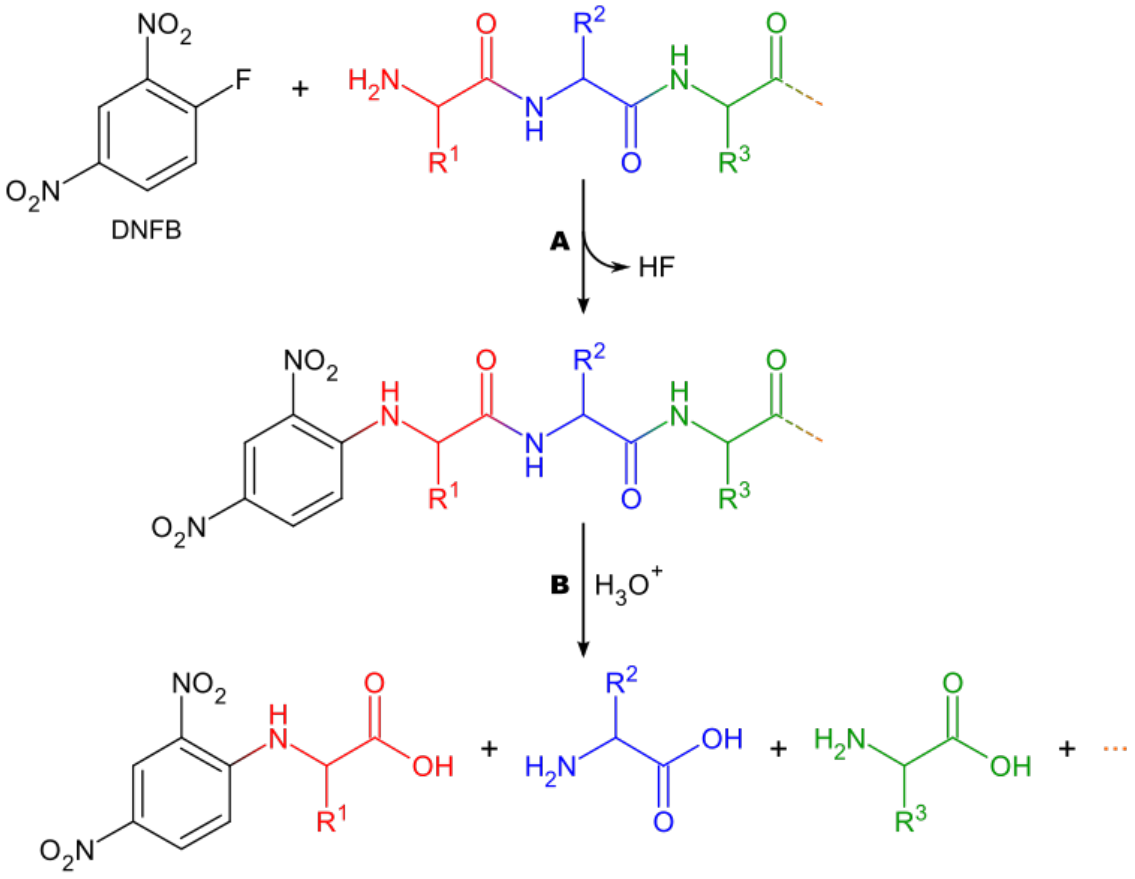
Separation

The amino acids can be separated by ion-exchange chromatography or hydrophobic interaction chromatography. An example of the former is given by the NTRC using sulfonated polystyrene as a matrix, adding the amino acids in acid solution and passing a buffer of steadily increasing pH through the column. Amino acids will be eluted when the pH reaches their respective isoelectric points. The latter technique may be employed through the use of reversed phase chromatography. Many commercially available C8 and C18 silica columns have demonstrated successful separation of amino acids in solution in less than 40 minutes through the use of an optimised elution gradient.

Quantitative analysis

Once the amino acids have been separated, their respective quantities are determined by adding a reagent that will form a coloured derivative. If the amounts of amino acids are in excess of 10 nmol, ninhydrin can be used for this - it gives a yellow colour when reacted with proline, and a vivid purple with other amino acids. The concentration of amino acid is proportional to the absorbance of the resulting solution. With very small quantities, down to 10 pmol, fluorescamine can be used as a marker: this forms a fluorescent derivative on reacting with an amino acid.

N-*terminal amino acid analysis*



Sanger's method of peptide end-group analysis: **A** derivatization of *N*-terminal end with Sanger's reagent (DNFB), **B** total acid hydrolysis of the dinitrophenyl peptide

Determining which amino acid forms the *N*-terminus of a peptide chain is useful for two reasons: to aid the ordering of individual peptide fragments' sequences into a whole chain, and because the first round of Edman degradation is often contaminated by impurities and therefore does not give an accurate determination of the *N*-terminal amino acid. A generalised method for *N*-terminal amino acid analysis follows:

1. React the peptide with a reagent which will selectively label the terminal amino acid.
2. Hydrolyse the protein.
3. Determine the amino acid by chromatography and comparison with standards.

There are many different reagents which can be used to label terminal amino acids. They all react with amine groups and will therefore also bind to amine groups in the side chains of amino acids such as lysine - for this reason it is necessary to be careful in interpreting chromatograms to ensure that the right spot is chosen. Two of the more common reagents are **Sanger's reagent** (1-fluoro-2,4-dinitrobenzene) and dansyl derivatives such as dansyl chloride. Phenylisothiocyanate, the reagent for the Edman degradation, can also be used.

The same questions apply here as in the determination of amino acid composition, with the exception that no stain is needed, as the reagents produce coloured derivatives and only qualitative analysis is required, so the amino acid does not have to be eluted from the chromatography column, just compared with a standard. Another consideration to take into account is that, since any amine groups will have reacted with the labelling reagent, ion exchange chromatography cannot be used, and thin layer chromatography or high pressure liquid chromatography should be used instead.

C-terminal amino acid analysis

The number of methods available for C-terminal amino acid analysis is much smaller than the number of available methods of N-terminal analysis. The most common method is to add carboxypeptidases to a solution of the protein, take samples at regular intervals, and determine the terminal amino acid by analysing a plot of amino acid concentrations against time.

Edman degradation

The Edman degradation is a very important reaction for protein sequencing, because it allows the ordered amino acid composition of a protein to be discovered. Automated Edman sequencers are now in widespread use, and are able to sequence peptides up to approximately 50 amino acids long. A reaction scheme for sequencing a protein by the Edman degradation follows - some of the steps are elaborated on subsequently.

1. Break any disulfide bridges in the protein with an oxidising agent like performic acid or reducing agent like 2-mercaptoethanol. A protecting group such as iodoacetic acid may be necessary to prevent the bonds from re-forming.
2. Separate and purify the individual chains of the protein complex, if there are more than one.
3. Determine the amino acid composition of each chain.
4. Determine the terminal amino acids of each chain.
5. Break each chain into fragments under 50 amino acids long.
6. Separate and purify the fragments.
7. Determine the sequence of each fragment.
8. Repeat with a different pattern of cleavage.
9. Construct the sequence of the overall protein.

Digestion into peptide fragments Peptides longer than about 50-70 amino acids long cannot be sequenced reliably by the Edman degradation. Because of this, long protein chains need to be broken up into small fragments which can then be sequenced individually. Digestion is done either by endopeptidases such as trypsin or pepsin or by chemical reagents such as cyanogen bromide. Different enzymes give different cleavage patterns, and the overlap between fragments can be used to construct an overall sequence.

The Edman degradation reaction

The peptide to be sequenced is adsorbed onto a solid surface - one common substrate is glass fibre coated with polybrene, a cationic polymer. The Edman reagent, phenylisothiocyanate (PTC), is added to the adsorbed peptide, together with a mildly basic buffer solution of 12% trimethylamine. This reacts with the amine group of the N-terminal amino acid.

The terminal amino acid can then be selectively detached by the addition of anhydrous acid. The derivative then isomerises to give a substituted phenylthiohydantoin which can be washed off and identified by chromatography, and the cycle can be repeated. The efficiency of each step is about 98%, which allows about 50 amino acids to be reliably determined.

Limitations of the Edman degradation

Because the Edman degradation proceeds from the N-terminus of the protein, it will not work if the N-terminal amino acid has been chemically modified or if it is concealed within the body of the protein. It also requires the use of either guesswork or a separate procedure to determine the positions of disulfide bridges.

Mass spectrometry

The other major direct method by which the sequence of a protein can be determined is mass spectrometry. This method has been gaining popularity in recent years as new techniques and increasing computing power have facilitated it. Mass spectrometry can, in principle, sequence any size of protein, but the problem becomes computationally more difficult as the size increases. Peptides are also easier to prepare for mass spectrometry than whole proteins, because they are more soluble. One method of delivering the peptides to the spectrometer is electrospray ionization, for which John Bennett Fenn won the Nobel Prize in Chemistry in 2002. The protein is digested by an endoprotease, and the resulting solution is passed through a high pressure liquid chromatography column. At the end of this column, the solution is sprayed out of a narrow nozzle charged to a high positive potential into the mass spectrometer. The charge on the droplets causes them to fragment until only single ions remain. The peptides are then fragmented and the mass-to-charge ratios of the fragments measured. (It is possible to detect which peaks correspond to multiply charged fragments, because these will have auxiliary peaks corresponding to other isotopes - the distance between these other peaks is inversely proportional to the charge on the fragment). The mass spectrum is analysed by computer and often compared against a database of previously sequenced proteins in order to determine the sequences of the fragments. This process is then repeated with a different digestion enzyme, and the overlaps in the sequences are used to construct a sequence for the protein.

Predicting protein sequence from DNA/RNA sequences

The amino acid sequence of a protein can also be determined indirectly from the mRNA or, in organisms that do not have introns (e.g. prokaryotes), the DNA that codes for the protein. If the sequence of the gene is already known, then this is all very easy. However, it is rare that the DNA sequence of a newly isolated protein will be known, and so if this method is to be used, it has to be found in some way. One way that this can be done is to sequence a short section, perhaps 15 amino acids long, of the protein by one of the above methods, and then use this sequence to generate a complementary marker for the protein's RNA. This can then be used to isolate the mRNA coding for the protein, which can then be replicated in a polymerase chain reaction to yield a significant amount of DNA, which can then be sequenced relatively easily. The amino acid sequence of the protein can then be deduced from this. However, it is necessary to take into account the possibility of amino acids being removed after the mRNA has been translated.

Chapter 5

Proteomics Identifications Database, Protomap & Protein–Protein Interaction

Proteomics Identifications Database

The **PRIDE** (PRoteomics IDEntifications database) is one of the most prominent public data repositories of mass spectrometry (MS) based proteomics data, and is maintained by the European Bioinformatics Institute as part of the Proteomics Services Team.

PRIDE stores three different kinds of information: peptide and protein identifications derived from MS or MS/MS experiments, MS and MS/MS mass spectra as peak lists, and any and all associated metadata. Peptide sequences should be captured as parts of identifications .

By September 2010, PRIDE contained more than 13,000 experiments, 4 million protein identifications, 20 million peptide identifications and more than 104 million spectra. A typical PRIDE dataset or project contains more than one experiment (accession numbers or MS runs). As mass spectometry is increasingly used for capturing details of posttranslational modification PRIDE contains modification data in case of the peptides which were chemically modified.

PRIDE was established as a production service in 2005. Several other proteomics databases have been established over the past few years like GPMDB, PeptideAtlas, Proteinpedia and the NCBI Peptidome . Together with the NCBI Peptidome, the PRIDE database constitutes an actual structured data repository, storing the original experimental data from the researchers, and does not assume any editorial control over submitted data.

In total, PRIDE contains data from about 60 species, the biggest fraction of it coming from human samples, followed by the fruitfly Drosophila melanogaster and mouse.

Formats & the submission process

Since detailed proteomics data currently cannot be curated from the existing literature the source of PRIDE data is solely submissions by academic researchers.

PRIDE is a standards compliant public repository meaning that its own XML-based data exchange format for submissions, PRIDE XML, was built around the Proteomics Standards Initiative mzData standard for mass spectrometry. PRIDE is committed to implementing relevant new Proteomics Standards Initiative standards as soon as possible.

As there are many types of different mass spectometry instruments and software formats are currently on the market, wet-lab scientists without a strong bioinformatics background or informatics support were having problems converting their data to PRIDE XML. The development of PRIDE Converter helped to tackle this situation. PRIDE Converter is a tool, written in the Java programming language, that converts 15 different input mass spectometry data formats into PRIDE XML via a wizard-like graphical user interface. It is freely available and is open source under the permissive Apache License.

Browsing, searching & data mining PRIDE

Currently, data can be queried from PRIDE via the PRIDE web and BioMart interfaces.

Additionally one can build complex queries with the PRIDE BioMart using BioMart which is a query-oriented data management system. The extensive use of controlled vocabularies (CVs) and ontologies for flexible yet context-sensitive annotation of data, along with the ability to perform intelligent queries by these annotations, are key features of PRIDE .

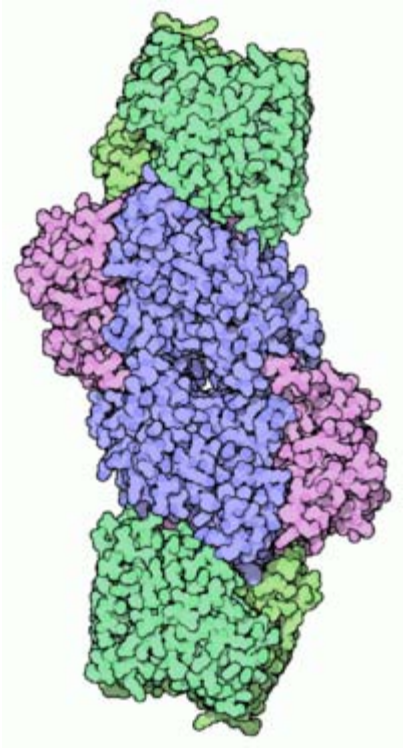
Protomap (Proteomics)

PROTOMAP is a recently developed proteomic technology for identifying changes to proteins that manifest in altered migration by one-dimensional SDS-PAGE. It is similar, conceptually, to two-dimensional gel electrophoresis and difference gel electrophoresis in that it enables global identification of proteins that undergo altered electrophoretic migration resulting from, for example, proteolysis or post-translational modification. However, it is unique in that all proteins are sequenced using mass spectrometry which provides information on the sequence coverage detected in each isoform of each protein thereby facilitating interpretation of proteolytic events.

PROTOMAP is performed by resolving control and experimental samples in separate lanes of a 1D SDS-PAGE gel. Each lane is cut into evenly spaced bands (usually 15-30 bands) and proteins in these bands are sequenced using shotgun proteomics. Sequence information from all of these bands are bioinformatically integrated into a visual format called a **peptograph** which plots gel-migration in the vertical dimension (high- to low-molecular weight, top to bottom) and sequence coverage in the horizontal dimension (N- to C-terminus, left to right). A peptograph is generated for each protein the sample (thousands of peptographs are generated from a single experiment) and this data format enables rapid identification of proteins undergoing proteolytic cleavage by making evident changes in gel-migration that are accompanied by altered topography.

PROTOMAP stands for **PRotein TOpography and Migration Analysis Platform** and was invented and developed by Ben Cravatt and colleagues at The Scripps Research Institute.

Protein–Protein Interaction



The bacterial nitrogenase enzyme is formed by a protein-protein interaction between two copies of two different proteins. One protein is shown in shades of green, the other in shades of blue and purple.

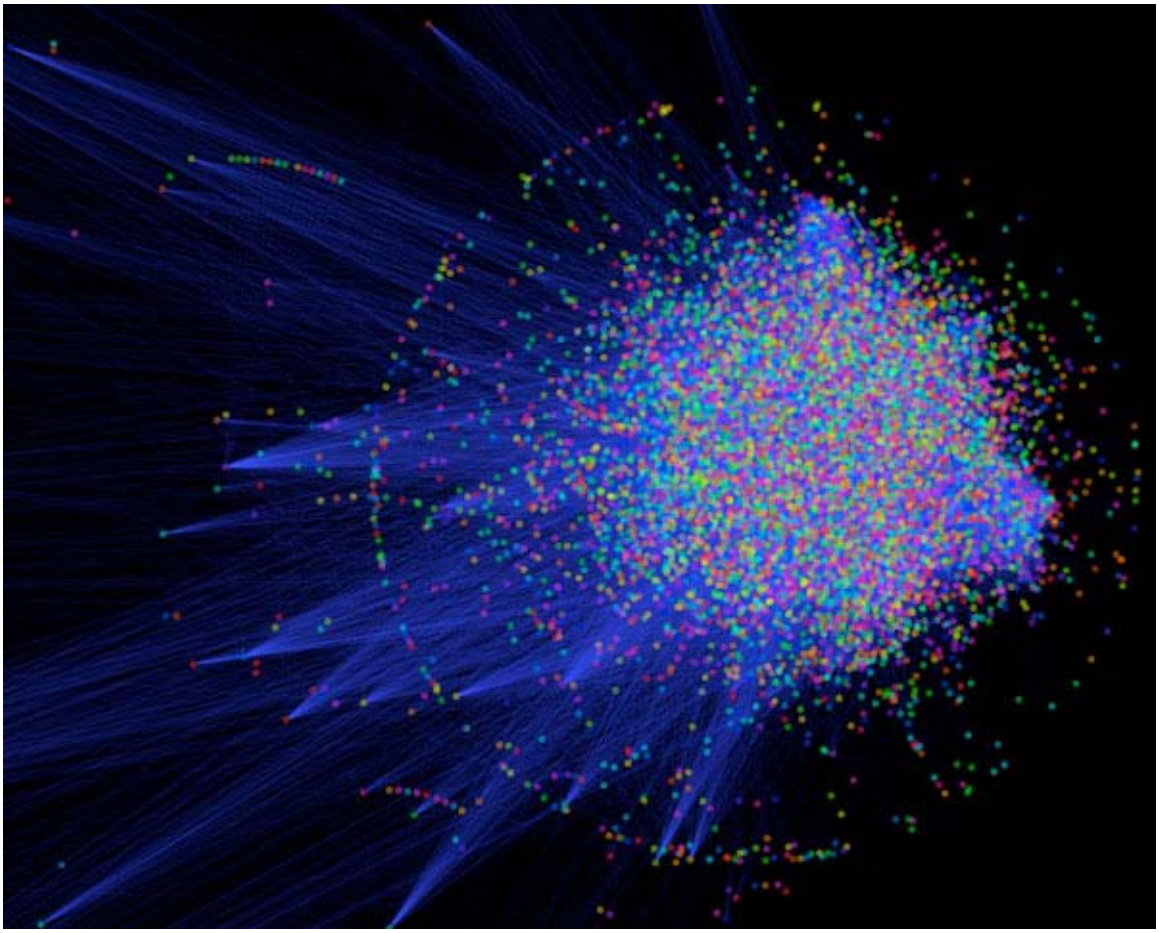
Protein–protein interactions occur when two or more proteins bind together, often to carry out their biological function. Many of the most important molecular processes in the cell such as DNA replication are carried out by large molecular machines that are built from a large number of protein components organised by their protein-protein interactions. Protein interactions have been studied from the perspectives of biochemistry, quantum chemistry, molecular dynamics, signal transduction and other metabolic or genetic/epigenetic networks. Indeed, protein–protein interactions are at the core of the entire interactomics system of any living cell.

Interactions between proteins are important for the majority of biological functions. For example, signals from the exterior of a cell are mediated to the inside of that cell by protein–protein interactions of the signaling molecules. This process, called signal transduction, plays a fundamental role in many biological processes and in many diseases (e.g. cancers). Proteins might interact for a long time to form part of a protein complex, a protein may be carrying another protein (for example, from cytoplasm to nucleus or vice versa in the case of the nuclear pore importins), or a protein may interact briefly with another protein just to modify it (for example, a protein kinase will add a phosphate to a target protein). This modification of proteins can itself change protein–protein interactions. For example, some proteins with SH2 domains only bind to other proteins when they are phosphorylated on the amino acid tyrosine while bromodomains specifically recognise acetylated lysines. In conclusion, protein–protein interactions are of central importance for virtually every process in a living cell. Information about these interactions improves our understanding of diseases and can provide the basis for new therapeutic approaches.

Methods to investigate protein–protein interactions

As protein–protein interactions are so important there are a multitude of methods to detect them. Each of the approaches has its own strengths and weaknesses, especially with regard to the sensitivity and specificity of the method. A high sensitivity means that many of the interactions that occur in reality are detected by the screen. A high specificity indicates that most of the interactions detected by the screen are also occurring in reality. Methods such as yeast two-hybrid screening can be used to detect novel protein-protein interactions. There are also many biophysical methods for investigating the nature and properties of interactions.

Visualization of networks



Network visualisation of the human interactome where each point represents a protein and each blue line between them is an interaction.

Visualization of protein–protein interaction networks is a popular application of scientific visualization techniques. Although protein interaction diagrams are common in textbooks, diagrams of whole cell protein interaction networks were not as common since the level of complexity made them difficult to generate. One example of a manually produced molecular interaction map is Kurt Kohn's 1999 map of cell cycle control. Drawing on Kohn's map, in 2000 Schwikowski, Uetz, and Fields published a paper on protein–protein interactions in yeast, linking together 1,548 interacting proteins determined by two-hybrid testing. They used a force-directed (Sugiyama) graph drawing algorithm to automatically generate an image of their network..

Database collections

Methods for identifying interacting proteins have defined hundreds of thousands of interactions. These interactions are collected together in specialised biological databases that allow the interactions to be assembled and studied further. The first of these databases was DIP, the database of interacting proteins. Since that time a large number of

further database collections have been created such as BioGRID, STRING, and Database of Interacting Proteins.

Chapter 6

Protein–Protein Interaction Prediction

Protein–protein interaction prediction is a field combining bioinformatics and structural biology in an attempt to identify and catalog physical interactions between pairs or groups of proteins. Understanding protein–protein interactions is important for the investigation of intracellular signaling pathways, modelling of protein complex structures and for gaining insights into various biochemical processes. Experimentally, physical interactions between pairs of proteins can be inferred from a variety of experimental techniques, including yeast two-hybrid systems, protein-fragment complementation assays (PCA), affinity purification/mass spectrometry, protein microarrays, fluorescence resonance energy transfer (FRET) and Microscale Thermophoresis (MST). Efforts to experimentally determine the interactome of numerous species are ongoing, and a number of computational methods for interaction prediction have been developed in recent years.

Methods

Proteins that interact are more likely to co-evolve, therefore it is possible to make inferences about interactions between pairs of proteins based on their phylogenetic distances. It has also been observed in some cases that pairs of interacting proteins have fused orthologues in other organisms. In addition, a number of bound protein complexes have been structurally solved and can be used to identify the residues that mediate the interaction so that similar motifs can be located in other organisms.

Phylogenetic profiling

Phylogenetic profiling finds pairs of protein families with similar patterns of presence or absence across large numbers of species. This method identifies pairs likely to act in the same biological process, but does not necessarily imply physical interaction.

Prediction of co-evolved protein pairs based on similar phylogenetic trees

This method involves using a sequence search tool such as BLAST for finding homologues of a pair of proteins, then building multiple sequence alignments with alignment tools such as Clustal. From these multiple sequence alignments, phylogenetic distance matrices are calculated for each protein in the hypothesized interacting pair. If the matrices are sufficiently similar (as measured by their Pearson correlation coefficient) they are deemed likely to interact.

Identification of homologous interacting pairs

This method consists of searching whether the two sequences have homologues which form a complex in a database of known structures of complexes. The identification of the domains is done by sequence searches against domain databases such as Pfam using BLAST. If more than one complex of Pfam domains is identified, then the query sequences are aligned using a hidden Markov tool called HMMER to the closest identified homologues, whose structures are known. Then the alignments are analysed to check whether the contact residues of the known complex are conserved in the alignment.

Identification of structural patterns

This method builds a library of known protein–protein interfaces from the PDB, where the interfaces are defined as pairs of polypeptide fragments that are below a threshold slightly larger than the Van der Waals radius of the atoms involved. The sequences in the library are then clustered based on structural alignment and redundant sequences are eliminated. The residues that have a high (generally >50%) level of frequency for a given position are considered hotspots. This library is then used to identify potential interactions between pairs of targets, providing that they have a known structure (i.e. present in the PDB).

Bayesian network modelling

Bayesian methods integrate data from a wide variety of sources, including both experimental results and prior computational predictions, and use these features to assess the likelihood that a particular potential protein interaction is a true positive result. These methods are useful because experimental procedures, particularly the yeast two-hybrid experiments, are extremely noisy and produce many false positives, while the previously mentioned computational methods can only provide circumstantial evidence that a particular pair of proteins might interact.

3D template-based protein complex modelling

This method makes use of known protein complex structures to predict as well as structurally model interactions between query protein sequences. The prediction process generally starts by employing a sequence based method (e.g Interolog) to search for protein complex structures that are homologous to the query sequences. These known complex structures are then used as templates to structurally model the interaction between query sequences. This method has the advantage of not only inferring protein interactions but also suggests models of how proteins interact structurally, which can provide some insights into the atomic level mechanism of that interaction. On the other hand, the ability for this method to makes a prediction is limited to a relatively small number of known protein complex structures.

Supervised learning problem

The problem of PPI prediction can be framed as a supervised learning problem. In this paradigm the known protein interactions supervise the estimation of a function that can predict whether an interaction exists or not between two proteins given data about the proteins (e.g., expression levels of each gene in different experimental conditions, location information, phylogenetic profile, etc.).

Relationship to docking methods

The field of protein–protein interaction prediction is closely related to the field of protein–protein docking, which attempts to use geometric and steric considerations to fit two proteins of known structure into a bound complex. This is a useful mode of inquiry in cases where both proteins in the pair have known structures and are known (or at least strongly suspected) to interact, but since so many proteins do not have experimentally determined structures, sequence-based interaction prediction methods are especially useful in conjunction with experimental studies of an organism's interactome.

Chapter 7

Neuroproteomics

Neuroproteomics is the study of the protein complexes and species that make up the nervous system. These proteins interact to make the neurons connect in such a way to create the intricacies that nervous system is known for. Neuroproteomics is a complex field that has a long way to go in terms of profiling the entire neuronal proteome. It is a relatively recent field that has many applications in therapy and science. So far, only small subsets of the neuronal proteome have been mapped, and then only when applied to the proteins involved in the synapse.

History

Origins

The word proteomics was first used in 1994 by Marc Wilkins as the study of "the protein equivalent of a genome" . It is defined as the all of the proteins expressed in a biological system under specific physiologic conditions at a certain point in time. It can change with any biochemical alteration, and so it can only be defined under certain conditions. Neuroproteomics is a subset of this field dealing with the complexities and multi-system origin of neurological disease. Neurological function is based on the interactions of many proteins of different origin, and so requires a systematic study of subsystems within its proteomic structure.

Modern Times

Neuroproteomics has the difficult task of defining on a molecular level the pathways of consciousness, senses, and self. Neurological disorders are unique in that they do not always exhibit outward symptoms. Defining the disorders becomes difficult and so neuroproteomics is a step in the right direction of identifying biomarkers that can be used to detect diseases. Not only does the field have to map out the different proteins possible from the genome, but there are many modifications that happen after transcription that affect function as well. Because neurons are such dynamic molecules changing with every action potential that travels through them, neuroproteomics offers the most potential for mapping out the molecular template of their function. Genomics offers a static roadmap of the cell, while proteomics can offer a glimpse into structures smaller than the cell because of its specific nature to each moment in time.

Mechanisms of Use

Protein Separation

In order for neuroproteomics to function correctly, proteins must be separated in terms of the proteome from which they came. For example, one set might be under normal conditions, while another might be under diseased conditions. Proteins are commonly separated using two-dimensional polyacrylamide gel electrophoresis (2D PAGE). For this technique, proteins are run across an immobile gel with a pH gradient until they stop at the point where their net charge is neutral. After separating by charge in one direction, sodium dodecyl sulfate is run in the other direction to separate the proteins by size. A two-dimensional map is created using this technique that can be used to match additional proteins later. One can usually match the function of a protein by identifying in an 2D PAGE in simple proteomics because many intracellular somatic pathways are known. In neuroproteomics, however, many proteins combine to give an end result that may be neurological disease or breakdown. It is necessary then to study each protein individually and find a correlation between the different proteins to determine the cause of a neurological disease. New techniques are being developed that can identify proteins once they are separated out using 2D PAGE.

Protein Identification

Protein separate techniques, such as 2D PAGE, are limitd in that they cannot handle very high or low molecular weight protein species. Alternative methods have been developed to deal with such cases. These include liquid chromatography mass spectrometry along with sodium dodecyl sulfate polyacrylamide gel electrophoresis, or liquid chromatography mass spectrometry run in multiple dimensions. Compared to simple 2D page, liquid chromatography mass spectrometry can handle a larger range of protein species size, but it is limited in the amount of protein sample it handle at once. Liquid chromatography mass spectrometry is also limited in its lack of a reference map from which to work with. Complex algorithms are usually used to analyze the fringe results that occur after a procedure is run. The unknown portions of the protein species are

usually not analyzed in favor of familiar proteomes, however. This fact reveals a fault with current technology; new techniques are needed to increase both the specificity and scope of proteome mapping.

Applications

Drug Addiction

It is commonly known that drug addiction involves permanent synaptic plasticity of various neuronal circuits. Neuroproteomics is being applied to study the effect of drug addiction across the synapse. Research is being conducted by isolating distinct regions of the brain in which synaptic transmission takes place and defining the proteome for that particular region. Different stages of drug abuse must be studied, however, in order to map out the progression of protein changes along the course of the drug addiction. These stages include enticement, ingesting, withdrawal, addiction, and removal. It begins with the change in the genome through transcription that occurs due to the abuse of drugs. It continues to identify the most likely proteins to be affected by the drugs and focusing in on that area. For drug addiction, the synapse is the most likely target as it involves communication between neurons. Lack of sensory communication in neurons is often an outward sign of drug abuse, and so neuroproteomics is being applied to find out what proteins are being affected to prevent the transport of neurotransmitters. In particular, the vesicle releasing process is being studied to identify the proteins involved in the synapse during drug abuse. Proteins such as synaptotagmin and synaptobrevin interact to fuse the vesicle into the membrane. Phosphorylation also has its own set of proteins involved that work together to allow the synapse to function properly. Drugs such as morphine change properties such as cell adhesion, neurotransmitter volume, and synaptic traffic. After significant morphine application, tyrosine kinases received less phosphorylation and thus send fewer signals inside the cell. These receptor proteins are unable to initiate the intracellular signaling processes that enable the neuron to live, and necrosis or apoptosis may be the result. With more and more neurons affected along this chain of cell death, permanent loss of sensory or motor function may be the result. By identifying the proteins that are changed with drug abuse, neuroproteomics may give clinicians even earlier biomarkers to test for to prevent permanent neurological damage.

Recently, a novel terminology (Psychoproteomics) has been coined by the University of Florida researchers from Dr. Mark S Gold Lab. Kobeissy et al. defined Psychoproteomics as integral proteomics approach dedicated to studying proteomic changes in the field of psychiatric disorders, particularly substance-and drug-abuse neurotoxicity.

Brain Injury

Traumatic brain injury is defined as a “direct physical impact or trauma to the head followed by a dynamic series of injury and repair events” . Recently, neuroproteomics have been applied to studying the disability that over 5.4 million Americans live with. In

addition to physically injuring the brain tissue, traumatic brain injury induces the release of glutamate that interacts with ionotropic glutamate receptors (iGluRs). These glutamate receptors acidify the surrounding intracranial fluid, causing further injury on the molecular level to nearby neurons. The death of the surrounding neurons is induced through normal apoptosis mechanisms, and it is this cycle that is being studied with neuroproteomics. Three different cysteine protease derivatives are involved in the apoptotic pathway induced by the acidic environment triggered by glutamate. These cysteine proteases include calpain, caspase, and cathepsin. These three proteins are examples of detectable signs of traumatic brain injury that are much more specific than temperature, oxygen level, or intracranial pressure. Proteomics thus also offers a tracking mechanism by which researchers can monitor the progression of traumatic brain injury, or a chronic disease such as Alzheimer's or Parkinson's. Especially in Parkinson's, in which neurotransmitters play a large role, recent proteomic research has involved the study of synaptotagmin. Synaptotagmin is involved in the calcium-induced budding of vesicle containing neurotransmitters from the presynaptic membrane. By studying the intracellular mechanisms involved in neural apoptosis after traumatic brain injury, researchers can create a map that genetic changes can follow later on.

Nerve Growth

One group of researchers applied the field of neuroproteomics to examine how different proteins affect the initial growth of neuritis . The experiment compared the protein activity of control neurons with the activity of neurons treated with nerve growth factor (NGF) and JNJ460, an "immunophilin ligand." JNJ460 is an offspring of another drug that is used to prevent immune attack when organs are transplanted. It is not an immunosuppressant, however, but rather it acts as a shield against microglia. NGF promotes neuron viability and differentiation by binding to TrkA, a tyrosine receptor kinase. This receptor is important in initiating intracellular metabolic pathways, including Ras, Rak, and MAP kinase.

Protein differentiation was measured in each cell sample with and without treatment by NGF and JNJ460. A peptide mixture was made by washing off unbound portions of the amino acid sequence in a reverse column. The resulting mixture was then suspended a peptide mixture in a bath of cation exchange fluid. The proteins were identified by splicing them with trypsin and then searching through the results of passing the product through a mass spectrometer. This applies a form of liquid chromatography mass spectrometry to identify proteins in the mixture

JNJ460 treatment resulted in an increase in "signal transduction" proteins, while NGF resulted in an increase in proteins associated with the ribosome and synthesis of other proteins. JNJ460 also resulted in more structural proteins associated with intercellular growth, such as actin, myosin, and troponin. With NGF treatment, cells increased protein synthesis and creation of ribosomes. This method allows the analysis of all of the protein patterns overall, rather than a single change in an amino acid. Western blots confirmed the results, according to the researchers, though the changes in proteins were not as obvious in their protocol.

The main significance to these findings are that JNJ460 are NGF are distinct processes that both control the protein output of the cell. JNJ460 resulted in increased neuronal size and stability while NGF resulted in increased membrane proteins. When combined together, they significantly increase a neuron's chance of growth. While JNJ460 may "prime" some parts of the cell for NGF treatment, they do not work together. JNJ460 is thought to interact with Schwann cells in regenerating actin and myosin, which are key players in axonal growth. NGF helps the neuron grow as a whole. These two proteins do not play a part in communication with other neurons, however. They merely increase the size of the membrane down which a signal can be sent. Other neurotrophic factor proteomes are needed to guide neurons to each other to create synapses.

Limitations

The broad scope of the available raw neuronal proteins to map requires that initial studies be focused on small areas of the neurons. When taking samples, there are a few places that interest neurologists most. The most important place to start for neurologists is the plasma membrane. This is where most of the communication between neurons takes place. The proteins being mapped here include ion channels, neurotransmitter receptors, and molecule transporters. Along the plasma membrane, the proteins involved in creating cholesterol-rich lipid rafts are being studied because they have been shown to be crucial for glutamate uptake during the initial stages of neuron formation. As mentioned before, vesicle proteins are also being studied closely because they are involved in disease. Collecting samples to study, however, requires special consideration to ensure that the reproducibility of the samples is not compromised. When taking a global sample of one area of the brain for example, proteins that are ubiquitous and relatively unimportant show up very clear in the SDS PAGE. Other unexplored, more specific proteins barely show up and are therefore ignored. It is usually necessary to divide up the plasma membrane proteome, for example, into subproteomes characterized by specific function. This allows these more specific classes of peptides to show up more clearly. In a way, dividing into subproteomes is simply applying a magnifying lens to a specific section of a global proteome's SDS PAGE map. This method seems to be most effective when applied to each cellular organelle separately. Mitochondrial proteins, for example, which are more effective at transporting electrons across its membrane, can be specifically targeted effectively in order to match their electron-transporting ability to their amino acid sequence.

Chapter 8

Quantitative Proteomics & SILAC

Quantitative Proteomics

The aim of **quantitative proteomics** is to obtain quantitative information about all proteins in a sample. Rather than just providing lists of proteins identified in a certain sample, quantitative proteomics yields information about differences between samples. For example, this approach can be used to compare samples from healthy and diseased patients. The methods for protein identification are identical to those used in general (i.e. qualitative) proteomics, but include quantification as an additional dimension.

Mass spectrometry-based proteomics

Due to differences in ionization efficiency and/or detectability the intensity of a peak in a mass spectrum is not a good indicator of the amount of the analyte. However, differences in peak intensity of the *same* analyte accurately reflect differences in its abundance. Therefore, in mass spectrometry-based proteomics *relative* quantification is used to compare protein abundance between samples. This can be achieved by labeling one sample with stable isotopes alone or incorporated into protein crosslinkers like D4-BS3, which leads to a mass shift in the mass spectrum. Differentially labeled samples are combined and analyzed together. The differences in the peak intensities of the isotope pairs accurately reflect difference in the abundance of the corresponding proteins. Current methods include:

- Isotope-coded affinity tags (ICAT)

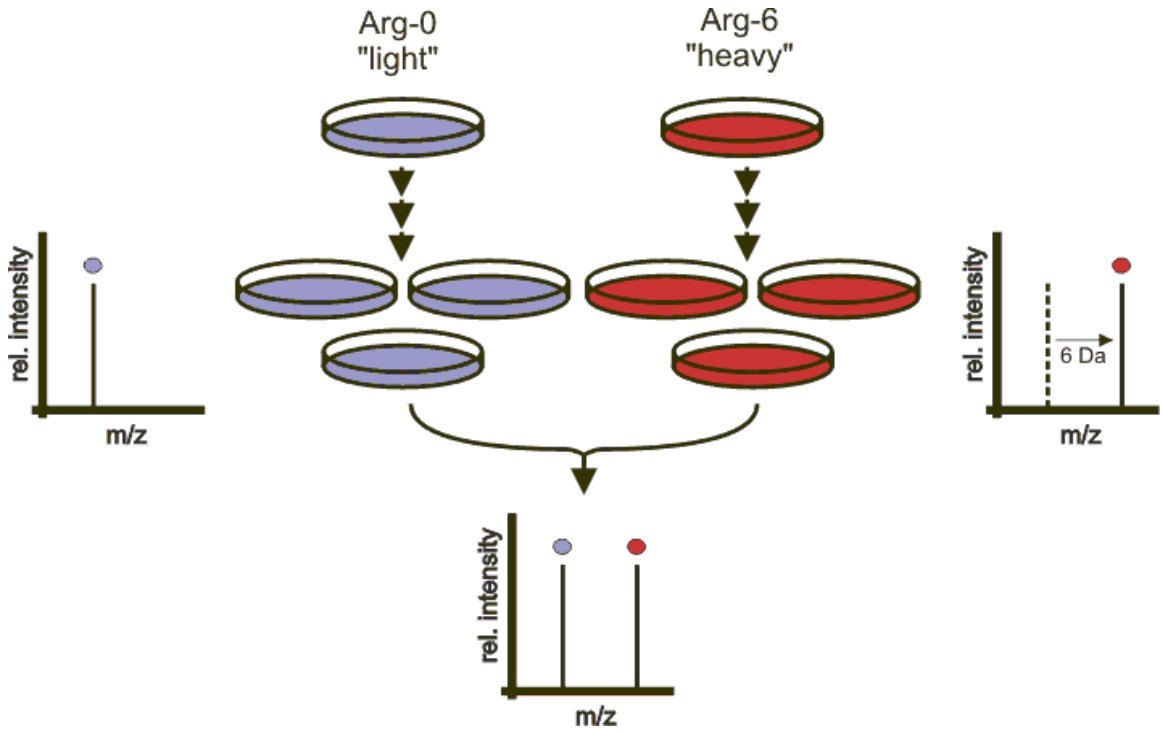
- Isobaric labeling
    - Tandem mass tags (TMT)
    - Isobaric tags for relative and absolute quantitation (iTRAQ)
- Label-free quantification
- Metal-coded tags (MeCATs)
- N-terminal labelling
- Stable isotope labeling with amino acids in cell culture (SILAC)

MeCAT can be used in combination with element mass spectrometry ICP-MS allowing first-time absolute quantification of the metal bound by MeCAT reagent to a protein or biomolecule. Thus it is possible to determine the absolute amount of protein down to attomol range using external calibration by metal standard solution. It is compatible to protein separation by 2D electrophoresis and chromatography in multiplex experiments. Protein identification and relative quantification can be performed by MALDI-MS/MS and ESI-MS/MS.

Two-dimensional gel electrophoresis

Modern day gel electrophoresis research often leverages software-based image analysis tools primarily to analyze bio-markers by quantifying individual, as well as showing the separation between one or more protein "spots" on a scanned image of a 2-DE product. Differential staining of gels with fluorescent dyes (difference gel electrophoresis) can also be used to highlight differences in the spot pattern.

SILAC



The principle of SILAC. Cells are differentially labeled by growing them in light medium with normal arginine (Arg-0, blue colour) or medium with heavy arginine (Arg-6, red colour). Metabolic incorporation of the amino acids into the proteins results in a mass shift of the corresponding peptides. This mass shift can be detected by a mass spectrometer as indicated by the depicted mass spectra. When both samples are combined, the ratio of peak intensities in the mass spectrum reflects the relative protein abundance. In this example, the labeled protein has the same abundance in both samples (ratio 1).

SILAC (stable isotope labeling by/with amino acids in cell culture) is a technique based on mass spectrometry that detects differences in protein abundance among samples using non-radioactive isotopic labeling. It is a popular method for quantitative proteomics.

Procedure

Two populations of cells are cultivated in cell culture. One of the cell populations is fed with growth medium containing normal amino acids. In contrast, the second population is fed with growth medium containing amino acids labeled with stable (non-radioactive) heavy isotopes. For example, the medium can contain arginine labeled with six carbon-13 atoms (^{13}C) instead of the normal carbon-12 (^{12}C). When the cells are growing in this medium, they incorporate the heavy arginine into all of their proteins. Therefore, all of the arginine containing peptides are now 6 Da heavier than their normal counterparts. The trick is that the proteins from both cell populations can be combined and analyzed together by mass spectrometry. Pairs of chemically identical peptides of different stable-isotope composition can be differentiated in a mass spectrometer owing to their mass

difference. The ratio of peak intensities in the mass spectrum for such peptide pairs reflects the abundance ratio for the two proteins.

Applications

A SILAC approach involving incorporation of tyrosine labeled with nine carbon-13 atoms (^{13}C) instead of the normal carbon-12 (^{12}C) has been utilized to study tyrosine kinase substrates in signaling pathways. SILAC has emerged as a very powerful method to study cell signaling, post translation modifications such as phosphorylation, protein-protein interaction and regulation of gene expression. Standardized protocols of SILAC for various application have also been published.

Pulsed SILAC

Pulsed SILAC (pSILAC) is a variation of the SILAC method where the labelled amino acids are added to the growth medium for only a short period of time. This allows monitoring differences in *de novo* protein production rather than raw concentration.

Chapter 9

Phosphoproteomics

Phosphoproteomics is a branch of proteomics that identifies, catalogs, and characterizes proteins containing a phosphate group as a post-translational modification. Phosphorylation is a key reversible modification that regulates protein function, subcellular localization, complex formation, degradation of proteins and therefore cell signalling networks. With all of these modification results, it is assumed that up to 30% of all proteins may be phosphorylated, some multiple times.

Compared to expression analysis, phoshoproteomics provides two additional layers of information. First, it provides clues on what protein or pathway might be activated because a change in phosphorylation status almost always reflects a change in protein activity. Second, it indicates what proteins might be potential drug targets as exemplified by the kinase inhibitor Gleevec. While phosphoproteomics will greatly expand knowledge about the numbers and types of phosphoproteins, its greatest promise is the rapid analysis of entire phosphorylation based signalling networks.

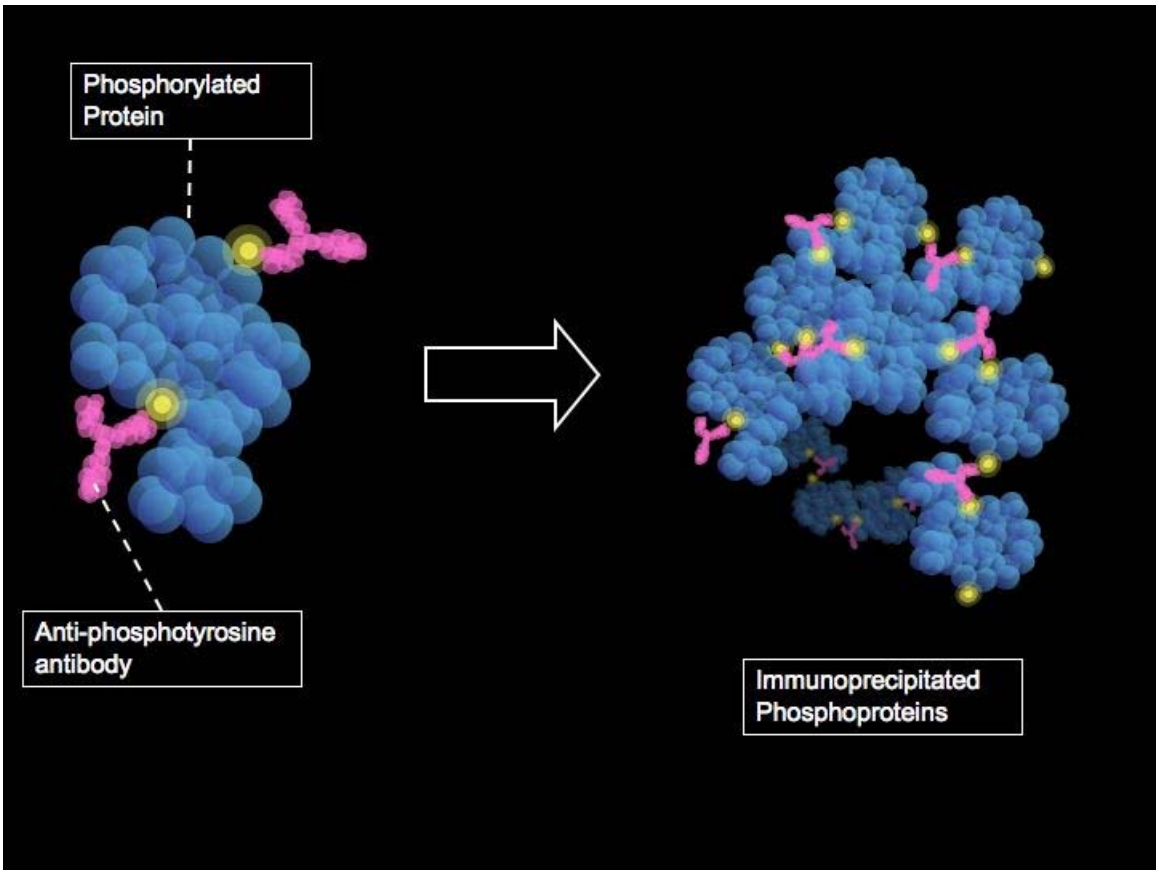
Overview of phosphoproteomic analysis

A sample large-scale phosphoproteomic analysis includes

1. Cultured cells undergo SILAC encoding.
2. Cells are stimulated with factor of interest (eg. growth factor, hormone).
3. Stimulation can occur for various lengths of time for temporal analysis.
4. Cells are lysed and enzymatically digested.
5. Peptides are separated using ion exchange chromatography.
6. Phosphopeptides are enriched using phosphospecific antibodies, immobilized metal affinity chromatography or titanium dioxide (TiO_2) chromatography.

7. Phosphopeptides are analyzed using mass spectrometry.
8. Peptides are sequenced and analyzed.

Tools and methods



Method of phosphoprotein purification by immunoprecipitation with anti-phosohotyrosine antibodies

The analysis of the entire complement of phosphorylated proteins in a cell is certainly a feasible option. This is due to the optimization of enrichment protocols for phosphoproteins and phosphopeptides, better fractionation techniques using chromatography, and improvement of methods to selectively visualize phosphorylated residues using mass spectrometry. Although the current procedures for phosphoproteomic analysis are greatly improved, there is still sample loss and inconsistencies with regards to sample preparation, enrichment , and instrumentation. Bioinformatics tools and biological sequence databases are also necessary for high-throughput phosphoproteomic studies.

Enrichment Strategies

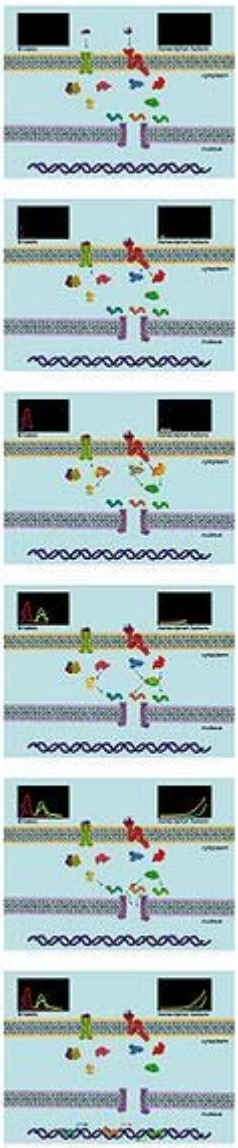
Previous procedures to isolate phosphorylated proteins included radioactive labeling with ^{32}P-labeled ATP followed by SDS polyacrylamide gel electrophoresis or thin layer chromatography. These traditional methods are inefficient because it is impossible to

obtain large amounts of proteins required for phosphorylation analysis. Therefore, the current and simplest methods to enrich phosphoproteins are affinity purification using phosphospecific antibodies, immobilized metal affinity chromatography (IMAC), strong cation exchange (SCX) chromatography, or titanium dioxide chromatography. Antiphosphotyrosine antibodies have been proven very successful in purification, but fewer reports have been published using antibodies against phosphoserine- or phosphothreonine-containing proteins. IMAC enrichment is based on phosphate affinity for immobilized metal chelated to the resin. SCX separates phosphorylated from non-phosphorylated peptides based on the negatively charged phosphate group. Titanium dioxide chromatography is a newer technique that requires significantly less column preparation time. Many phosphoproteomic studies use a combination of these enrichment strategies to obtain the purest sample possible.

Mass Spectrometry Analysis

Mass spectrometry is currently the best method to adequately compare pairs of protein samples. The two main procedures to perform this task are using isotope-coded affinity tags (ICAT) and stable isotopic amino acids in cell culture (SILAC). In the ICAT procedure samples are labeled individually after isolation with mass-coded reagents that modify cysteine residues. In SILAC, cells are cultured separately in the presence of different isotopically labeled amino acids for several cell divisions allowing cellular proteins to incorporate the label. Mass spectrometry is subsequently used to identify phosphoserine, phosphothreonine, and phosphotyrosine-containing peptides.

Phosphoproteomics in the study of signal transduction



Analysis of Signal Transduction Dynamics

Intracellular signal transduction is primarily mediated by the reversible phosphorylation of various signalling molecules by enzymes dubbed kinases. Kinases transfer phosphate groups from ATP to specific serine, threonine or tyrosine residues of target molecules. The resultant phosphorylated protein may have altered activity level, subcellular localization or tertiary structure.

Phosphoproteomic analyses are ideal for the study of the dynamics of signalling networks. In one study design, cells are exposed to SILAC labelling and then stimulated by a specific growth factor. The cells are collected at various timepoints, and the lysates are combined for analysis by tandem MS. This allows experimenters to track the

phosphorylation state of many phosphoproteins in the cell over time. The ability to measure the global phosphorylation state of many proteins at various time points makes this approach much more powerful than traditional biochemical methods for analyzing signalling network behavior.

One study was able to simultaneously measure the fold-change in phosphorylation state of 127 proteins between unstimulated and EphrinB1-stimulated cells. Of these 127 proteins, 40 showed increased phosphorylation with stimulation by EphrinB1. The researchers were able to use this information in combination with previously published data to construct a signal transduction network for the proteins downstream of the EphB2 receptor.

Another recent phosphoproteomic study included large-scale identification and quantification of phosphorylation events triggered by the anti-diuretic hormone vasopressin in kidney collecting duct. A total of 714 phosphorylation sites on 223 unique phosphoproteins were identified, including three novel phosphorylation sites in the vasopressin-sensitive water channel aquaporin-2 (AQP2).

Phosphoproteomics in the study of cancer

Since the inception of phosphoproteomics, cancer research has focused on changes to the phosphoproteome during tumor development. Phosphoproteins could be cancer markers useful to cancer diagnostics and therapeutics. In fact, research has shown that there are distinct phosphotyrosine proteomes of breast and liver tumors. There is also evidence of hyperphosphorylation at tyrosine residues in breast tumors but not in normal tissues. Findings like these suggest that it is possible to mine the tumor phosphoproteome for potential biomarkers.

Increasing amounts of data are available suggesting that distinctive phosphoproteins exist in various tumors and that phosphorylation profiling could be used to fingerprint cancers from different origins. In addition, systematic cataloguing of tumor-specific phosphoproteins in individual patients could reveal multiple causative players during cancer formation. By correlating this experimental data to clinical data such as drug response and disease outcome, potential cancer markers could be identified for diagnosis, prognosis, prediction of drug response, and potential drug targets.

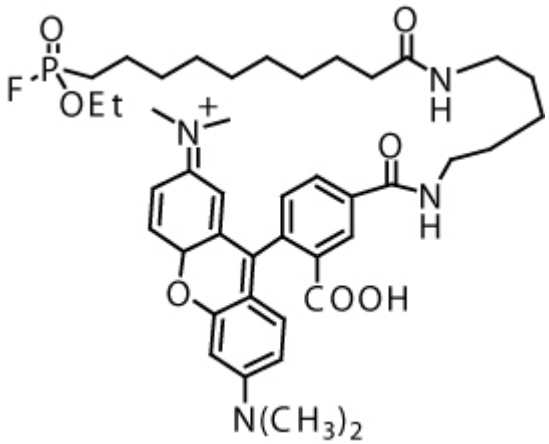
Limitations

While phosphoproteomics has greatly expanded knowledge about the numbers and types of phosphoproteins, along with their role in signaling networks, there are still a few limitations to these techniques. To begin with, isolation methods such as anti-phosphotyrosine antibodies do not distinguish between isolating tyrosine-phosphorylated proteins and proteins associated with tyrosine-phosphorylated proteins. Therefore, even though phosphorylation dependent protein-protein interactions are very important, it is important to remember that a protein detected by this method is not necessarily a direct substrate of any tyrosine kinase. Only by digesting the samples before

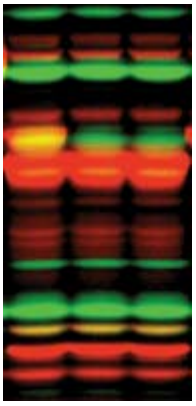
immunoprecipitation can isolation of only phosphoproteins and temporal profiles of individual phosphorylaiton sites be produced. Another limitation is that some relevant proteins will likely be missed since no extraction condition is all encompassing. It is possible that proteins with low stoichiometry of phosphorylation, in very low abundance, or phosphorylated as a target for rapid degradation will be lost.

Chapter 10

Activity-Based Proteomics



Fluorophosphonate-rhodamine (FP-Rhodamine) activity based probe for profiling of the serine hydrolase superfamily. In this probe the fluorophosphonate is the reactive group (RG) as it binds irreversibly to the active-site serine nucleophile of serine hydrolases and the tag is rhodamine, a fluorophore for in-gel visualization.



In-gel ABPP using probes with different fluorophores in the same lane to simultaneously profile differences in enzyme activities

Activity based proteomics, or **activity based protein profiling** (**ABPP**) is a functional proteomic technology that uses specially designed chemical probes that react with mechanistically-related classes of enzymes. The basic unit of ABPP is the probe which typically consists of two elements: a reactive group (RG) and a tag. Additionally, some probes may contain a binding group which enhances selectivity. The reactive group usually contains an electrophile that gets covalently-linked to a nucleophilic residue in the active site of an active enzyme. An enzyme that is inhibited by enzyme inhibitors or post-translational modifications will not react with an activity-based probe. The tag may be either a reporter such as a fluorophore or an affinity label such as biotin or an alkyne or azide for use with the Huisgen 1,3-dipolar cycloaddition.

A major advantage of ABPP is the ability to monitor the availability of the enzyme active site directly, rather than being limited to protein or mRNA abundance. With classes of enzymes such as the serine proteases and metalloproteases that often interact with endogenous inhibitors or that exist as inactive zymogens, this technique offers a valuable advantage over traditional techniques that rely on abundance rather than activity.

Finally, in recent years ABPP has been combined with tandem mass spectrometry enabling the identification of hundreds of active enzymes from a single sample. This technique, known as *ABPP-MudPIT* is especially useful for profiling inhibitor selectivity as the potency of an inhibitor can be tested against hundreds of targets simultaneously.

ABPP was originally invented by James C. Powers (Georgia Tech, Atlanta) in the early 1990s and subsequently adopted and developed by the other labs, for example Benjamin Cravatt III at The Scripps Research Institute, Matthew Bogyo at Stanford, and others.

Chapter 11

Enzyme

Human glyoxalase I. Two zinc ions that are needed for the enzyme to catalyze its reaction are shown as purple spheres, and an enzyme inhibitor called *S*-hexylglutathione is shown as a space-filling model, filling the two active sites.

Enzymes are proteins that catalyze (*i.e.*, increase or decrease the rates of) chemical reactions. In enzymatic reactions, the molecules at the beginning of the process are called substrates, and they are converted into different molecules, called the products. Almost all processes in a biological cell need enzymes to occur at significant rates. Since enzymes are selective for their substrates and speed up only a few reactions from among many possibilities, the set of enzymes made in a cell determines which metabolic pathways occur in that cell.

Like all catalysts, enzymes work by lowering the activation energy ($E_a^{\ddagger}$) for a reaction, thus dramatically increasing the rate of the reaction. As a result, products are formed faster and reactions reach their equilibrium state more rapidly. Most enzyme reaction rates are millions of times faster than those of comparable un-catalyzed reactions. As with all catalysts, enzymes are not consumed by the reactions they catalyze, nor do they alter the equilibrium of these reactions. However, enzymes do differ from most other catalysts by being much more specific. Enzymes are known to catalyze about 4,000 biochemical reactions. A few RNA molecules called ribozymes also catalyze reactions, with an important example being some parts of the ribosome. Synthetic molecules called artificial enzymes also display enzyme-like catalysis.

Enzyme activity can be affected by other molecules. Inhibitors are molecules that decrease enzyme activity; activators are molecules that increase activity. Many drugs and poisons are enzyme inhibitors. Activity is also affected by temperature, chemical environment (*e.g.*, pH), and the concentration of substrate. Some enzymes are used commercially, for example, in the synthesis of antibiotics. In addition, some household products use enzymes to speed up biochemical reactions (*e.g.*, enzymes in biological washing powders break down protein or fat stains on clothes; enzymes in meat tenderizers break down proteins into smaller molecules, making the meat easier to chew).

Etymology and history

As early as the late 17th and early 18th centuries, the digestion of meat by stomach secretions and the conversion of starch to sugars by plant extracts and saliva were known. However, the mechanism by which this occurred had not been identified.

In the 19th century, when studying the fermentation of sugar to alcohol by yeast, Louis Pasteur came to the conclusion that this fermentation was catalyzed by a vital force contained within the yeast cells called "ferments", which were thought to function only within living organisms. He wrote that "alcoholic fermentation is an act correlated with the life and organization of the yeast cells, not with the death or putrefaction of the cells."

In 1877, German physiologist Wilhelm Kühne (1837–1900) first used the term *enzyme*, which comes from Greek *ενζυμον*, "in leaven", to describe this process. The word *enzyme* was used later to refer to nonliving substances such as pepsin, and the word *ferment* was used to refer to chemical activity produced by living organisms.

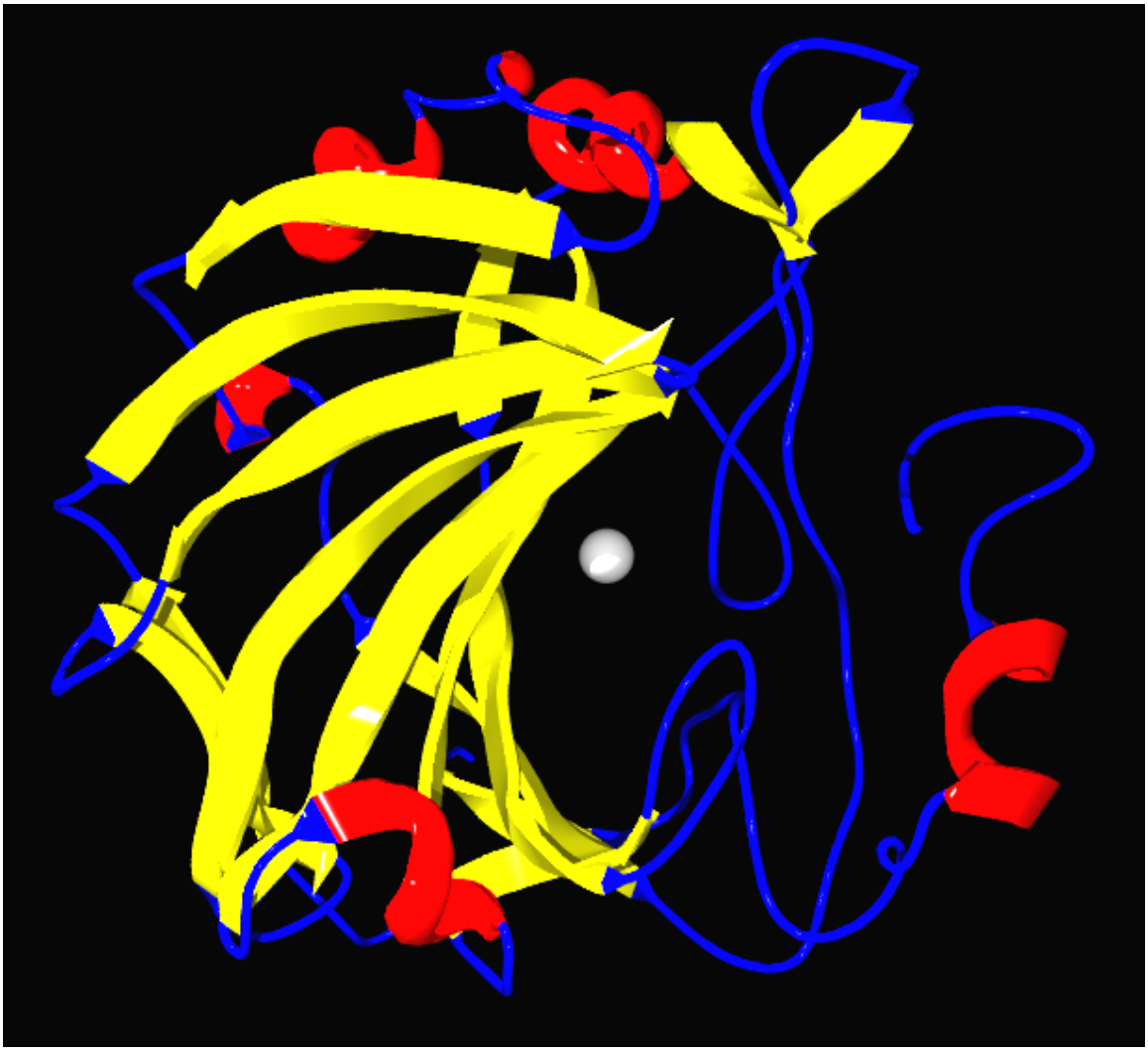
In 1897, Eduard Buchner submitted his first paper on the ability of yeast extracts that lacked any living yeast cells to ferment sugar. In a series of experiments at the University of Berlin, he found that the sugar was fermented even when there were no living yeast cells in the mixture. He named the enzyme that brought about the fermentation of sucrose "zymase". In 1907, he received the Nobel Prize in Chemistry "for his biochemical research and his discovery of cell-free fermentation". Following Buchner's example, enzymes are usually named according to the reaction they carry out. Typically, to generate the name of an enzyme, the suffix *-ase* is added to the name of its substrate (*e.g.*,

lactase is the enzyme that cleaves lactose) or the type of reaction (*e.g.*, DNA polymerase forms DNA polymers).

Having shown that enzymes could function outside a living cell, the next step was to determine their biochemical nature. Many early workers noted that enzymatic activity was associated with proteins, but several scientists (such as Nobel laureate Richard Willstätter) argued that proteins were merely carriers for the true enzymes and that proteins *per se* were incapable of catalysis. However, in 1926, James B. Sumner showed that the enzyme urease was a pure protein and crystallized it; Sumner did likewise for the enzyme catalase in 1937. The conclusion that pure proteins can be enzymes was definitively proved by Northrop and Stanley, who worked on the digestive enzymes pepsin (1930), trypsin and chymotrypsin. These three scientists were awarded the 1946 Nobel Prize in Chemistry.

This discovery that enzymes could be crystallized eventually allowed their structures to be solved by x-ray crystallography. This was first done for lysozyme, an enzyme found in tears, saliva and egg whites that digests the coating of some bacteria; the structure was solved by a group led by David Chilton Phillips and published in 1965. This high-resolution structure of lysozyme marked the beginning of the field of structural biology and the effort to understand how enzymes work at an atomic level of detail.

Structures and mechanisms



Ribbon diagram showing human carbonic anhydrase II. The grey sphere is the zinc cofactor in the active site. Diagram drawn from PDB 1MOO.

Enzymes are generally globular proteins and range from just 62 amino acid residues in size, for the monomer of 4-oxalocrotonate tautomerase, to over 2,500 residues in the animal fatty acid synthase. A small number of RNA-based biological catalysts exist, with the most common being the ribosome; these are referred to as either RNA-enzymes or ribozymes. The activities of enzymes are determined by their three-dimensional structure. However, although structure does determine function, predicting a novel enzyme's activity just from its structure is a very difficult problem that has not yet been solved.

Most enzymes are much larger than the substrates they act on, and only a small portion of the enzyme (around 3–4 amino acids) is directly involved in catalysis. The region that contains these catalytic residues, binds the substrate, and then carries out the reaction is known as the active site. Enzymes can also contain sites that bind cofactors, which are

needed for catalysis. Some enzymes also have binding sites for small molecules, which are often direct or indirect products or substrates of the reaction catalyzed. This binding can serve to increase or decrease the enzyme's activity, providing a means for feedback regulation.

Like all proteins, enzymes are long, linear chains of amino acids that fold to produce a three-dimensional product. Each unique amino acid sequence produces a specific structure, which has unique properties. Individual protein chains may sometimes group together to form a protein complex. Most enzymes can be denatured—that is, unfolded and inactivated—by heating or chemical denaturants, which disrupt the three-dimensional structure of the protein. Depending on the enzyme, denaturation may be reversible or irreversible.

Structures of enzymes in complex with substrates or substrate analogs during a reaction may be obtained using Time resolved crystallography methods.

Specificity

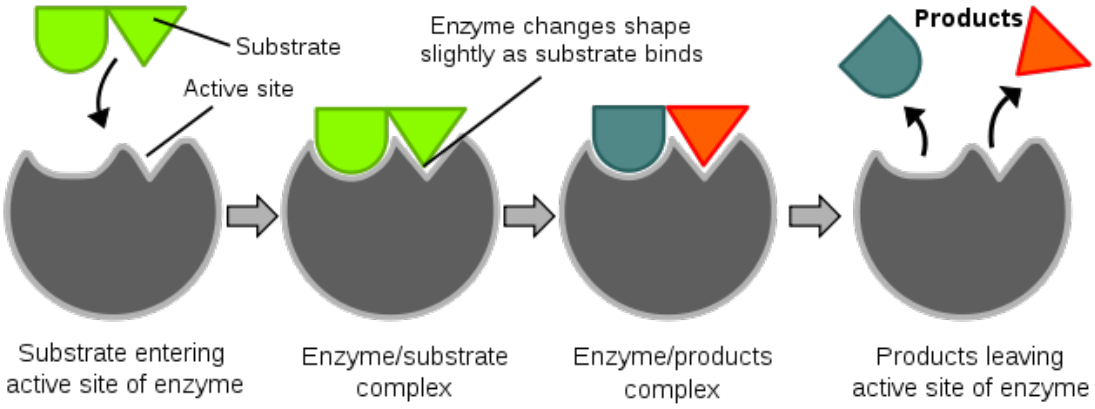
Enzymes are usually very specific as to which reactions they catalyze and the substrates that are involved in these reactions. Complementary shape, charge and hydrophilic/hydrophobic characteristics of enzymes and substrates are responsible for this specificity. Enzymes can also show impressive levels of stereospecificity, regioselectivity and chemoselectivity.

Some of the enzymes showing the highest specificity and accuracy are involved in the copying and expression of the genome. These enzymes have "proof-reading" mechanisms. Here, an enzyme such as DNA polymerase catalyzes a reaction in a first step and then checks that the product is correct in a second step. This two-step process results in average error rates of less than 1 error in 100 million reactions in high-fidelity mammalian polymerases. Similar proofreading mechanisms are also found in RNA polymerase, aminoacyl tRNA synthetases and ribosomes.

Some enzymes that produce secondary metabolites are described as promiscuous, as they can act on a relatively broad range of different substrates. It has been suggested that this broad substrate specificity is important for the evolution of new biosynthetic pathways.

"Lock and key" model

Enzymes are very specific, and it was suggested by the Nobel laureate organic chemist Emil Fischer in 1894 that this was because both the enzyme and the substrate possess specific complementary geometric shapes that fit exactly into one another. This is often referred to as "the lock and key" model. However, while this model explains enzyme specificity, it fails to explain the stabilization of the transition state that enzymes achieve.



Diagrams to show the induced fit hypothesis of enzyme action

In 1958, Daniel Koshland suggested a modification to the lock and key model: since enzymes are rather flexible structures, the active site is continually reshaped by interactions with the substrate as the substrate interacts with the enzyme. As a result, the substrate does not simply bind to a rigid active site; the amino acid side chains which make up the active site are molded into the precise positions that enable the enzyme to perform its catalytic function. In some cases, such as glycosidases, the substrate molecule also changes shape slightly as it enters the active site. The active site continues to change until the substrate is completely bound, at which point the final shape and charge is determined. Induced fit may enhance the fidelity of molecular recognition in the presence of competition and noise via the conformational proofreading mechanism .

Mechanisms

Enzymes can act in several ways, all of which lower $\Delta G^{\ddagger}$:

- Lowering the activation energy by creating an environment in which the transition state is stabilized (e.g. straining the shape of a substrate—by binding the transition-state conformation of the substrate/product molecules, the enzyme distorts the bound substrate(s) into their transition state form, thereby reducing the amount of energy required to complete the transition).
- Lowering the energy of the transition state, but without distorting the substrate, by creating an environment with the opposite charge distribution to that of the transition state.
- Providing an alternative pathway. For example, temporarily reacting with the substrate to form an intermediate ES complex, which would be impossible in the absence of the enzyme.
- Reducing the reaction entropy change by bringing substrates together in the correct orientation to react. Considering $\Delta H^{\ddagger}$ alone overlooks this effect.
- Increases in temperatures speed up reactions. Thus, temperature increases help the enzyme function and develop the end product even faster. However, if heated too

much, the enzyme’s shape deteriorates and the enzyme becomes denatured. Some enzymes like thermolabile enzymes work best at low temperatures.

Interestingly, this entropic effect involves destabilization of the ground state, and its contribution to catalysis is relatively small.

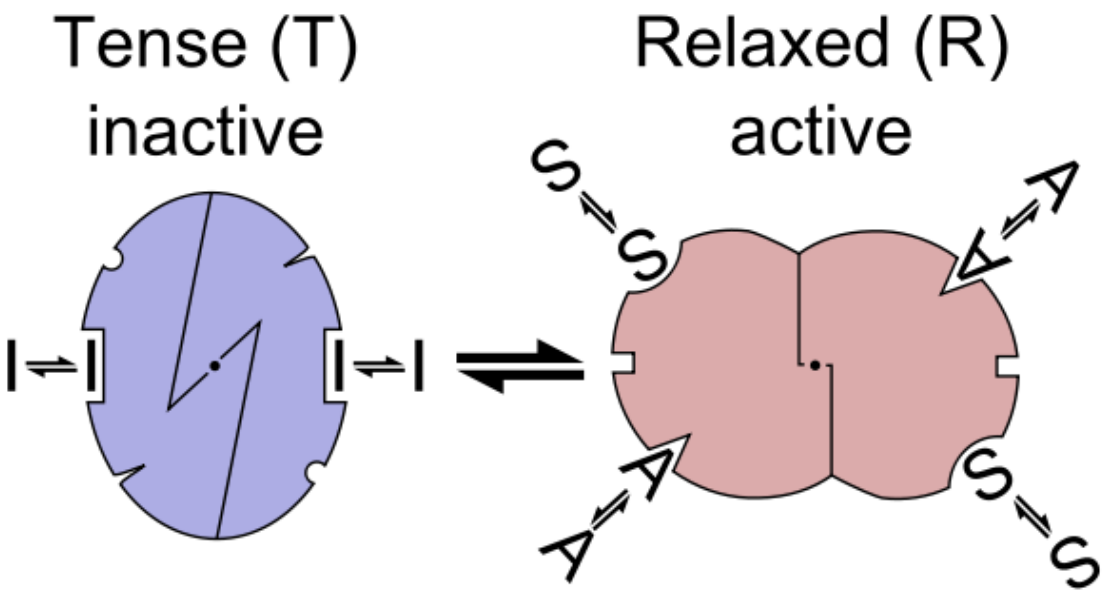
Transition State Stabilization

The understanding of the origin of the reduction of $\Delta G^{\ddagger}$ requires one to find out how the enzymes can stabilize its transition state more than the transition state of the uncatalyzed reaction. Apparently, the most effective way for reaching large stabilization is the use of electrostatic effects, in particular, by having a relatively fixed polar environment that is oriented toward the charge distribution of the transition state. Such an environment does not exist in the uncatalyzed reaction in water.

Dynamics and function

The internal dynamics of enzymes is linked to their mechanism of catalysis. Internal dynamics are the movement of parts of the enzyme's structure, such as individual amino acid residues, a group of amino acids, or even an entire protein domain. These movements occur at various time-scales ranging from femtoseconds to seconds. Networks of protein residues throughout an enzyme's structure can contribute to catalysis through dynamic motions. Protein motions are vital to many enzymes, but whether small and fast vibrations, or larger and slower conformational movements are more important depends on the type of reaction involved. However, although these movements are important in binding and releasing substrates and products, it is not clear if protein movements help to accelerate the chemical steps in enzymatic reactions. These new insights also have implications in understanding allosteric effects and developing new drugs.

Allosteric modulation



Allosteric transition of an enzyme between R and T states, stabilized by an agonist, an inhibitor and a substrate (the MWC model)

Allosteric sites are sites on the enzyme that bind to molecules in the cellular environment. The sites form weak, noncovalent bonds with these molecules, causing a change in the conformation of the enzyme. This change in conformation translates to the active site, which then affects the reaction rate of the enzyme. Allosteric interactions can both inhibit and activate enzymes and are a common way that enzymes are controlled in the body.

Cofactors and coenzymes

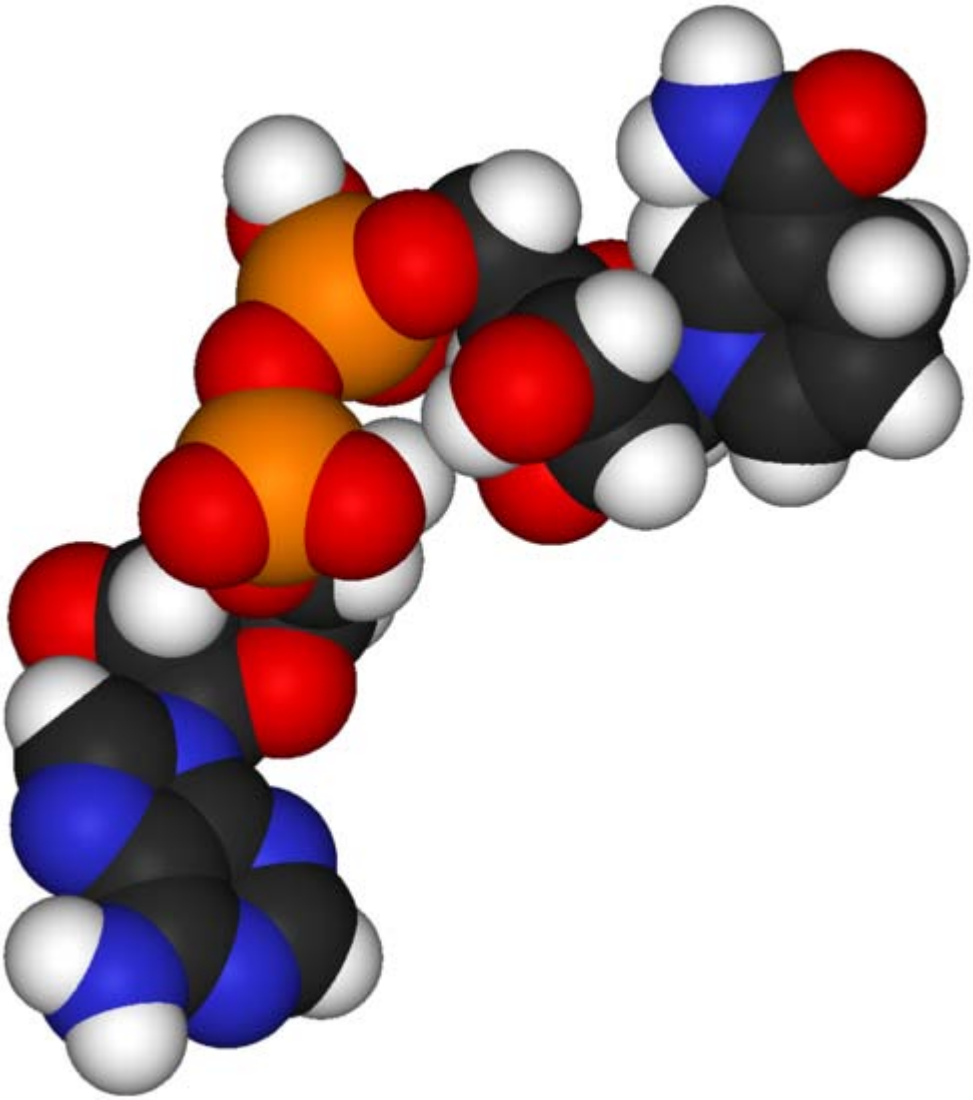
Cofactors

Some enzymes do not need any additional components to show full activity. However, others require non-protein molecules called cofactors to be bound for activity. Cofactors can be either inorganic (*e.g.*, metal ions and iron-sulfur clusters) or organic compounds (e.g., flavin and heme). Organic cofactors can be either prosthetic groups, which are tightly bound to an enzyme, or coenzymes, which are released from the enzyme's active site during the reaction. Coenzymes include NADH, NADPH and adenosine triphosphate. These molecules transfer chemical groups between enzymes.

An example of an enzyme that contains a cofactor is carbonic anhydrase, and is shown in the ribbon diagram above with a zinc cofactor bound as part of its active site. These tightly bound molecules are usually found in the active site and are involved in catalysis. For example, flavin and heme cofactors are often involved in redox reactions.

Enzymes that require a cofactor but do not have one bound are called *apoenzymes* or *apoproteins*. An apoenzyme together with its cofactor(s) is called a *holoenzyme* (this is the active form). Most cofactors are not covalently attached to an enzyme, but are very tightly bound. However, organic prosthetic groups can be covalently bound (*e.g.*, thiamine pyrophosphate in the enzyme pyruvate dehydrogenase). The term "holoenzyme" can also be applied to enzymes that contain multiple protein subunits, such as the DNA polymerases; here the holoenzyme is the complete complex containing all the subunits needed for activity.

Coenzymes



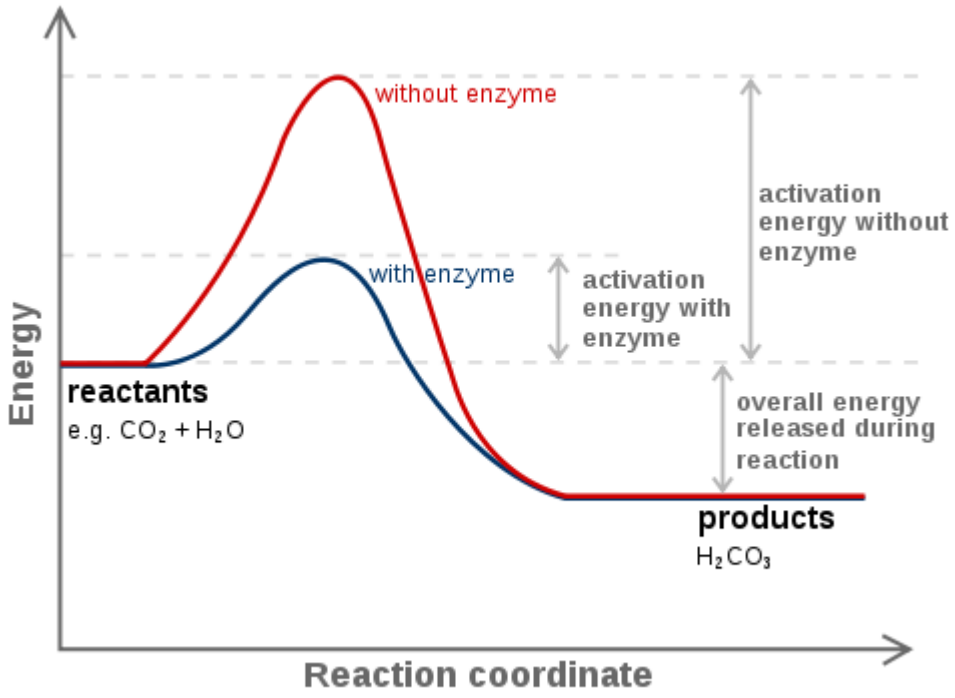
Space-filling model of the coenzyme NADH

Coenzymes are small organic molecules that can be loosely or tightly bound to an enzyme. Tightly bound coenzymes can be called allosteric groups. Coenzymes transport chemical groups from one enzyme to another. Some of these chemicals such as riboflavin, thiamine and folic acid are vitamins (compounds which cannot be synthesized by the body and must be acquired from the diet). The chemical groups carried include the hydride ion (H^-) carried by NAD or $NADP^+$, the phosphate group carried by adenosine triphosphate, the acetyl group carried by coenzyme A, formyl, methenyl or methyl groups carried by folic acid and the methyl group carried by S-adenosylmethionine.

Since coenzymes are chemically changed as a consequence of enzyme action, it is useful to consider coenzymes to be a special class of substrates, or second substrates, which are common to many different enzymes. For example, about 700 enzymes are known to use the coenzyme NADH.

Coenzymes are usually continuously regenerated and their concentrations maintained at a steady level inside the cell: for example, NADPH is regenerated through the pentose phosphate pathway and *S*-adenosylmethionine by methionine adenosyltransferase. This continuous regeneration means that even small amounts of coenzymes are used very intensively. For example, the human body turns over its own weight in ATP each day.

Thermodynamics



The energies of the stages of a chemical reaction. Substrates need a lot of energy to reach a transition state, which then decays into products. The enzyme stabilizes the transition state, reducing the energy needed to form products.

As all catalysts, enzymes do not alter the position of the chemical equilibrium of the reaction. Usually, in the presence of an enzyme, the reaction runs in the same direction as it would without the enzyme, just more quickly. However, in the absence of the enzyme, other possible uncatalyzed, "spontaneous" reactions might lead to different products, because in those conditions this different product is formed faster.

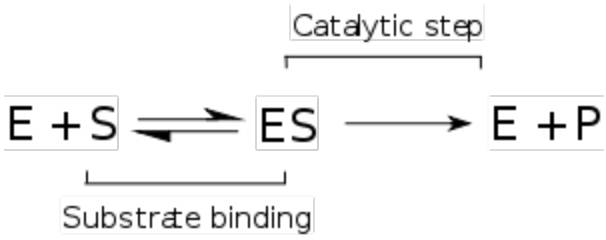
Furthermore, enzymes can couple two or more reactions, so that a thermodynamically favorable reaction can be used to "drive" a thermodynamically unfavorable one. For example, the hydrolysis of ATP is often used to drive other chemical reactions.

Enzymes catalyze the forward and backward reactions equally. They do not alter the equilibrium itself, but only the speed at which it is reached. For example, carbonic anhydrase catalyzes its reaction in either direction depending on the concentration of its reactants.

$$CO_2 + H_2O \xrightarrow{\text{Carbonic anhydrase}} H_2CO_3 \text{(in tissues; high } CO_2 \text{ concentration)}$$
$$H_2CO_3 \xrightarrow{\text{Carbonic anhydrase}} CO_2 + H_2O \text{(in lungs; low } CO_2 \text{ concentration)}$$

Nevertheless, if the equilibrium is greatly displaced in one direction, that is, in a very exergonic reaction, the reaction is *effectively* irreversible. Under these conditions the enzyme will, in fact, only catalyze the reaction in the thermodynamically allowed direction.

Kinetics



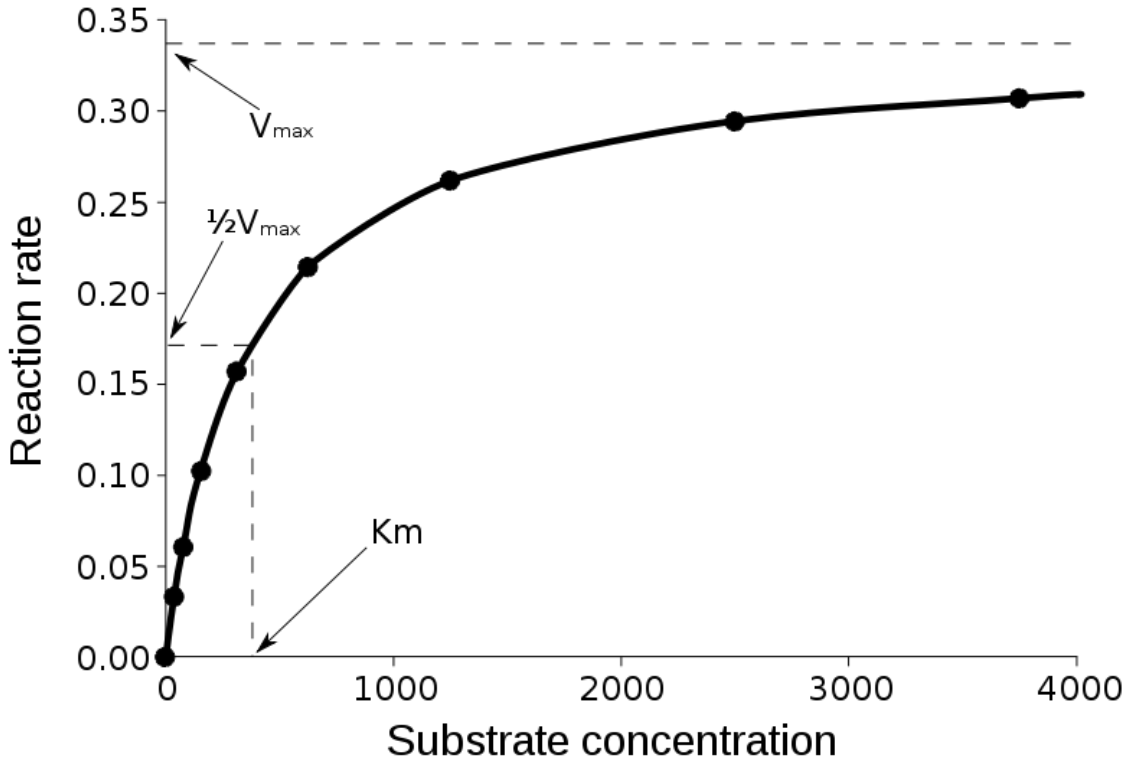
Mechanism for a single substrate enzyme catalyzed reaction. The enzyme (E) binds a substrate (S) and produces a product (P).

Enzyme kinetics is the investigation of how enzymes bind substrates and turn them into products. The rate data used in kinetic analyses are obtained from enzyme assays.

In 1902 Victor Henri proposed a quantitative theory of enzyme kinetics, but his experimental data were not useful because the significance of the hydrogen ion concentration was not yet appreciated. After Peter Lauritz Sørensen had defined the logarithmic pH-scale and introduced the concept of buffering in 1909 the German chemist Leonor Michaelis and his Canadian postdoc Maud Leonora Menten repeated Henri's experiments and confirmed his equation which is referred to as Henri-Michaelis-Menten kinetics (sometimes also Michaelis-Menten kinetics). Their work was further

developed by G. E. Briggs and J. B. S. Haldane, who derived kinetic equations that are still widely used today.

The major contribution of Henri was to think of enzyme reactions in two stages. In the first, the substrate binds reversibly to the enzyme, forming the enzyme-substrate complex. This is sometimes called the Michaelis complex. The enzyme then catalyzes the chemical step in the reaction and releases the product.



Saturation curve for an enzyme reaction showing the relation between the substrate concentration (S) and rate (v)

Enzymes can catalyze up to several million reactions per second. For example, the uncatalyzed decarboxylation of orotidine 5'-monophosphate has a half life of 78 million years. However, when the enzyme orotidine 5'-phosphate decarboxylase is added, the same process takes just 25 milliseconds. Enzyme rates depend on solution conditions and substrate concentration. Conditions that denature the protein abolish enzyme activity, such as high temperatures, extremes of pH or high salt concentrations, while raising substrate concentration tends to increase activity. To find the maximum speed of an enzymatic reaction, the substrate concentration is increased until a constant rate of product formation is seen. This is shown in the saturation curve on the right. Saturation happens because, as substrate concentration increases, more and more of the free enzyme is converted into the substrate-bound ES form. At the maximum reaction rate (V_{max}) of the enzyme, all the enzyme active sites are bound to substrate, and the amount of ES complex is the same as the total amount of enzyme. However, V_{max} is only one kinetic

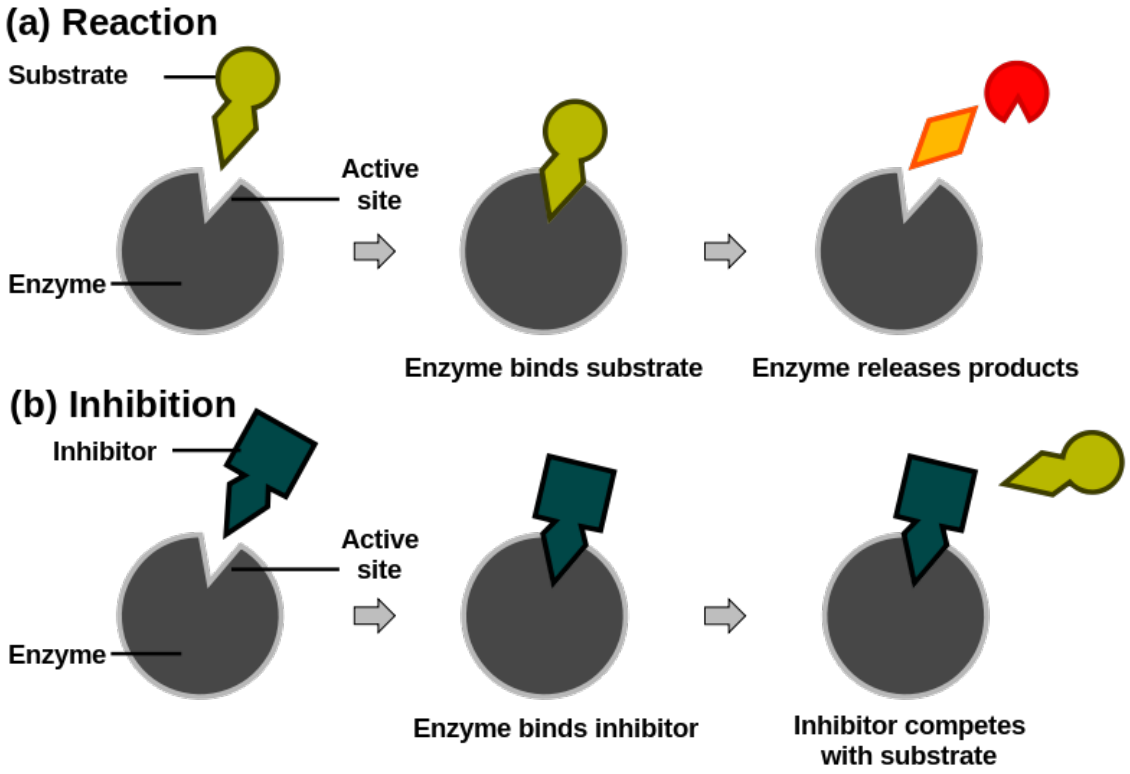
constant of enzymes. The amount of substrate needed to achieve a given rate of reaction is also important. This is given by the Michaelis-Menten constant (K_m), which is the substrate concentration required for an enzyme to reach one-half its maximum reaction rate. Each enzyme has a characteristic K_m for a given substrate, and this can show how tight the binding of the substrate is to the enzyme. Another useful constant is k_{cat}, which is the number of substrate molecules handled by one active site per second.

The efficiency of an enzyme can be expressed in terms of k_{cat}/K_m. This is also called the specificity constant and incorporates the rate constants for all steps in the reaction. Because the specificity constant reflects both affinity and catalytic ability, it is useful for comparing different enzymes against each other, or the same enzyme with different substrates. The theoretical maximum for the specificity constant is called the diffusion limit and is about 10^8 to 10^9 (M^{-1} s^{-1}). At this point every collision of the enzyme with its substrate will result in catalysis, and the rate of product formation is not limited by the reaction rate but by the diffusion rate. Enzymes with this property are called *catalytically perfect* or *kinetically perfect*. Example of such enzymes are triose-phosphate isomerase, carbonic anhydrase, acetylcholinesterase, catalase, fumarase, β-lactamase, and superoxide dismutase.

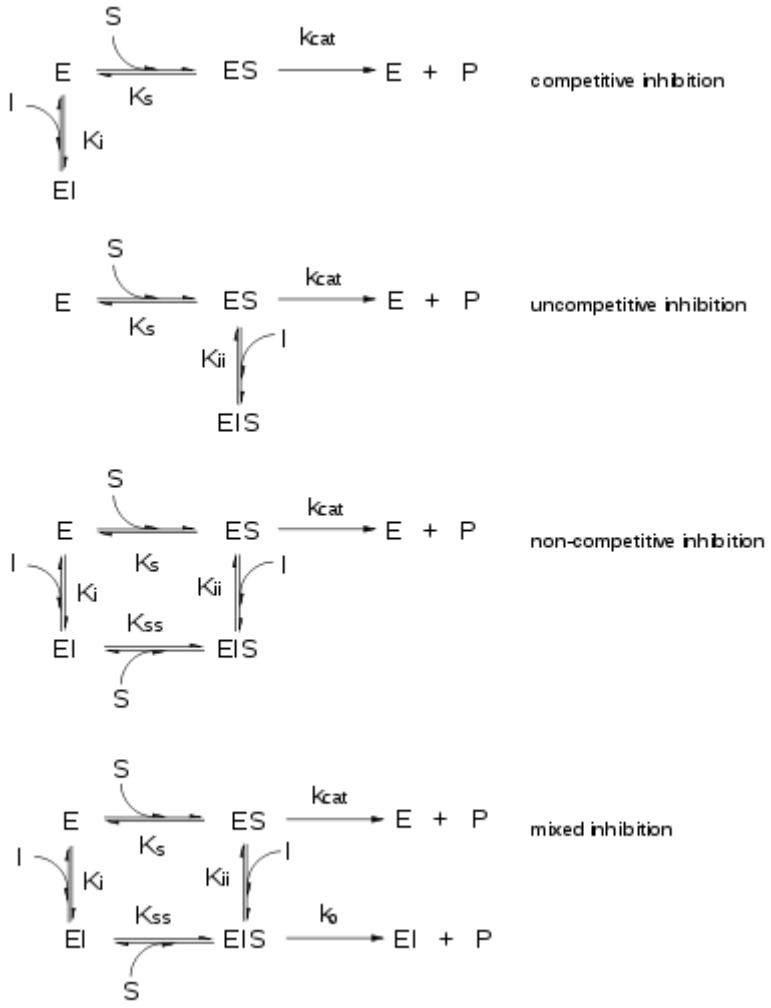
Michaelis-Menten kinetics relies on the law of mass action, which is derived from the assumptions of free diffusion and thermodynamically driven random collision. However, many biochemical or cellular processes deviate significantly from these conditions, because of macromolecular crowding, phase-separation of the enzyme/substrate/product, or one or two-dimensional molecular movement. In these situations, a fractal Michaelis-Menten kinetics may be applied.

Some enzymes operate with kinetics which are faster than diffusion rates, which would seem to be impossible. Several mechanisms have been invoked to explain this phenomenon. Some proteins are believed to accelerate catalysis by drawing their substrate in and pre-orienting them by using dipolar electric fields. Other models invoke a quantum-mechanical tunneling explanation, whereby a proton or an electron can tunnel through activation barriers, although for proton tunneling this model remains somewhat controversial. Quantum tunneling for protons has been observed in tryptamine. This suggests that enzyme catalysis may be more accurately characterized as "through the barrier" rather than the traditional model, which requires substrates to go "over" a lowered energy barrier.

Inhibition



Competitive inhibitors bind reversibly to the enzyme, preventing the binding of substrate. On the other hand, binding of substrate prevents binding of the inhibitor. Substrate and inhibitor compete for the enzyme.



Types of inhibition. This classification was introduced by W.W. Cleland.

Enzyme reaction rates can be decreased by various types of enzyme inhibitors.

Competitive inhibition

In competitive inhibition, the inhibitor and substrate compete for the enzyme (i.e., they can not bind at the same time). Often competitive inhibitors strongly resemble the real substrate of the enzyme. For example, methotrexate is a competitive inhibitor of the enzyme dihydrofolate reductase, which catalyzes the reduction of dihydrofolate to tetrahydrofolate. The similarity between the structures of folic acid and this drug are shown in the figure to the *right* bottom. Note that binding of the inhibitor need *not* be to the substrate binding site (as frequently stated), if binding of the inhibitor changes the conformation of the enzyme to prevent substrate binding and *vice versa*. In competitive inhibition the maximal rate of the reaction is not changed, but higher substrate concentrations are required to reach a given maximum rate, increasing the apparent K_m.

Uncompetitive inhibition

In uncompetitive inhibition the inhibitor can not bind to the free enzyme, but only to the ES-complex. The EIS-complex thus formed is enzymatically inactive. This type of inhibition is rare, but may occur in multimeric enzymes.

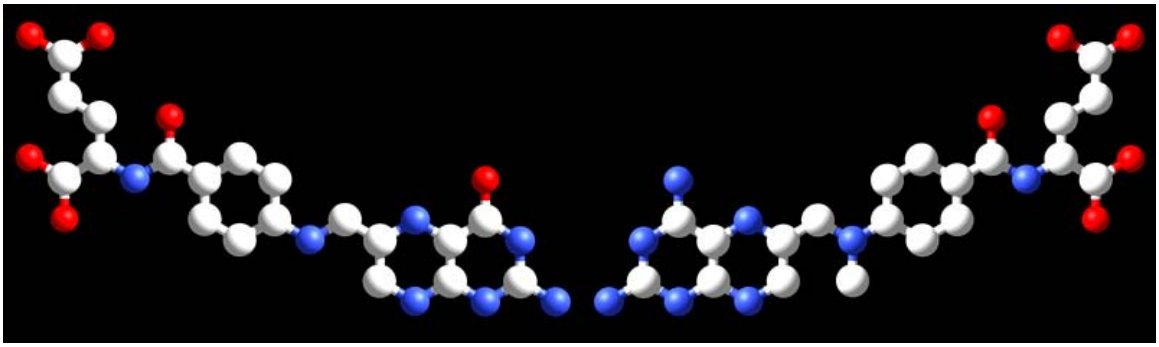
Non-competitive inhibition

Non-competitive inhibitors can bind to the enzyme at the binding site at the same time as the substrate,but not to the active site. Both the EI and EIS complexes are enzymatically inactive. Because the inhibitor can not be driven from the enzyme by higher substrate concentration (in contrast to competitive inhibition), the apparent V_{max} changes. But because the substrate can still bind to the enzyme, the K_m stays the same.

Mixed inhibition

This type of inhibition resembles the non-competitive, except that the EIS-complex has residual enzymatic activity.This type of inhibitor does not follow Michaelis-Menten equation.

In many organisms inhibitors may act as part of a feedback mechanism. If an enzyme produces too much of one substance in the organism, that substance may act as an inhibitor for the enzyme at the beginning of the pathway that produces it, causing production of the substance to slow down or stop when there is sufficient amount. This is a form of negative feedback. Enzymes which are subject to this form of regulation are often multimeric and have allosteric binding sites for regulatory substances. Their substrate/velocity plots are not hyperbolar, but sigmoidal (S-shaped).



The coenzyme folic acid (left) and the anti-cancer drug methotrexate (right) are very similar in structure. As a result, methotrexate is a competitive inhibitor of many enzymes that use folates.

Irreversible inhibitors react with the enzyme and form a covalent adduct with the protein. The inactivation is irreversible. These compounds include eflornithine a drug used to treat the parasitic disease sleeping sickness. Penicillin and Aspirin also act in this manner. With these drugs, the compound is bound in the active site and the enzyme then converts

the inhibitor into an activated form that reacts irreversibly with one or more amino acid residues.

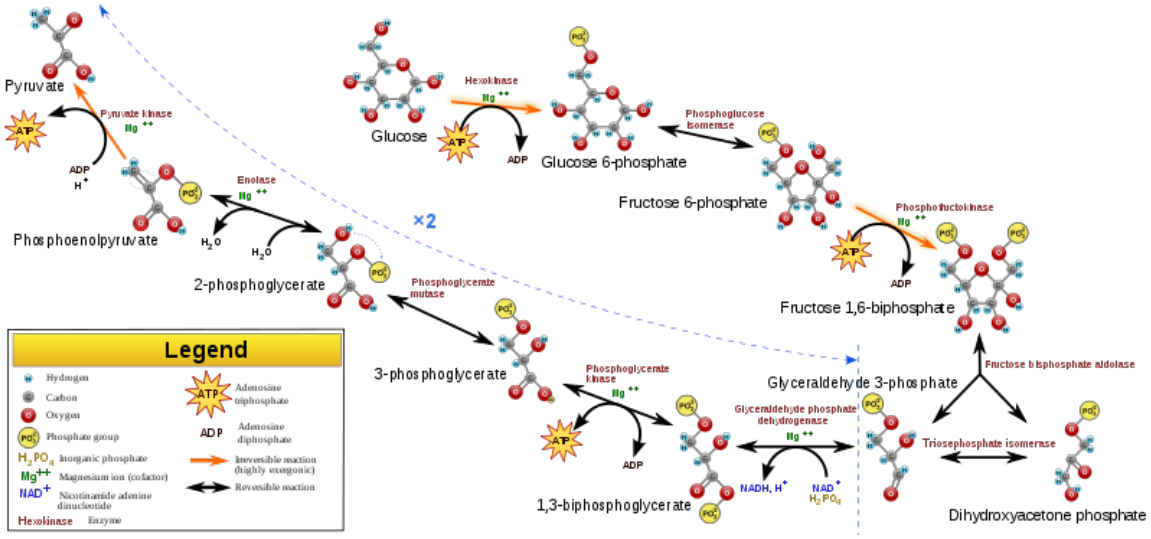
Uses of inhibitors

Since inhibitors modulate the function of enzymes they are often used as drugs. A common example of an inhibitor that is used as a drug is aspirin, which inhibits the COX-1 and COX-2 enzymes that produce the inflammation messenger prostaglandin, thus suppressing pain and inflammation. However, other enzyme inhibitors are poisons. For example, the poison cyanide is an irreversible enzyme inhibitor that combines with the copper and iron in the active site of the enzyme cytochrome c oxidase and blocks cellular respiration.

Biological function

Enzymes serve a wide variety of functions inside living organisms. They are indispensable for signal transduction and cell regulation, often via kinases and phosphatases. They also generate movement, with myosin hydrolysing ATP to generate muscle contraction and also moving cargo around the cell as part of the cytoskeleton. Other ATPases in the cell membrane are ion pumps involved in active transport. Enzymes are also involved in more exotic functions, such as luciferase generating light in fireflies. Viruses can also contain enzymes for infecting cells, such as the HIV integrase and reverse transcriptase, or for viral release from cells, like the influenza virus neuraminidase.

An important function of enzymes is in the digestive systems of animals. Enzymes such as amylases and proteases break down large molecules (starch or proteins, respectively) into smaller ones, so they can be absorbed by the intestines. Starch molecules, for example, are too large to be absorbed from the intestine, but enzymes hydrolyse the starch chains into smaller molecules such as maltose and eventually glucose, which can then be absorbed. Different enzymes digest different food substances. In ruminants which have herbivorous diets, microorganisms in the gut produce another enzyme, cellulase to break down the cellulose cell walls of plant fiber.



Glycolytic enzymes and their functions in the metabolic pathway of glycolysis

Several enzymes can work together in a specific order, creating metabolic pathways. In a metabolic pathway, one enzyme takes the product of another enzyme as a substrate. After the catalytic reaction, the product is then passed on to another enzyme. Sometimes more than one enzyme can catalyze the same reaction in parallel, this can allow more complex regulation: with for example a low constant activity being provided by one enzyme but an inducible high activity from a second enzyme.

Enzymes determine what steps occur in these pathways. Without enzymes, metabolism would neither progress through the same steps, nor be fast enough to serve the needs of the cell. Indeed, a metabolic pathway such as glycolysis could not exist independently of enzymes. Glucose, for example, can react directly with ATP to become phosphorylated at one or more of its carbons. In the absence of enzymes, this occurs so slowly as to be insignificant. However, if hexokinase is added, these slow reactions continue to take place except that phosphorylation at carbon 6 occurs so rapidly that if the mixture is tested a short time later, glucose-6-phosphate is found to be the only significant product. Consequently, the network of metabolic pathways within each cell depends on the set of functional enzymes that are present.

Control of activity

There are five main ways that enzyme activity is controlled in the cell.

1. **Enzyme production** (transcription and translation of enzyme genes) can be enhanced or diminished by a cell in response to changes in the cell's environment. This form of gene regulation is called enzyme induction and inhibition. For example, bacteria may become resistant to antibiotics such as penicillin because enzymes called beta-lactamases are induced that hydrolyse the crucial beta-lactam ring within the penicillin molecule. Another example are enzymes in the liver

called cytochrome P450 oxidases, which are important in drug metabolism. Induction or inhibition of these enzymes can cause drug interactions.

2. Enzymes can be **compartmentalized**, with different metabolic pathways occurring in different cellular compartments. For example, fatty acids are synthesized by one set of enzymes in the cytosol, endoplasmic reticulum and the Golgi apparatus and used by a different set of enzymes as a source of energy in the mitochondrion, through β-oxidation.
3. Enzymes can be regulated by **inhibitors and activators**. For example, the end product(s) of a metabolic pathway are often inhibitors for one of the first enzymes of the pathway (usually the first irreversible step, called *committed step*), thus regulating the amount of end product made by the pathways. Such a regulatory mechanism is called a negative feedback mechanism, because the amount of the end product produced is regulated by its own concentration. Negative feedback mechanism can effectively adjust the rate of synthesis of intermediate metabolites according to the demands of the cells. This helps allocate materials and energy economically, and prevents the manufacture of excess end products. The control of enzymatic action helps to maintain a stable internal environment in living organisms.
4. Enzymes can be regulated through **post-translational modification**. This can include phosphorylation, myristoylation and glycosylation. For example, in the response to insulin, the phosphorylation of multiple enzymes, including glycogen synthase, helps control the synthesis or degradation of glycogen and allows the cell to respond to changes in blood sugar. Another example of post-translational modification is the cleavage of the polypeptide chain. Chymotrypsin, a digestive protease, is produced in inactive form as chymotrypsinogen in the pancreas and transported in this form to the stomach where it is activated. This stops the enzyme from digesting the pancreas or other tissues before it enters the gut. This type of inactive precursor to an enzyme is known as a zymogen.
5. Some enzymes may become **activated when localized to a different environment** (e.g. from a reducing (cytoplasm) to an oxidizing (periplasm) environment, high pH to low pH etc.). For example, hemagglutinin in the influenza virus is activated by a conformational change caused by the acidic conditions, these occur when it is taken up inside its host cell and enters the lysosome.

Involvement in disease

Phenylalanine hydroxylase. Created from PDB 1KW0

Since the tight control of enzyme activity is essential for homeostasis, any malfunction (mutation, overproduction, underproduction or deletion) of a single critical enzyme can lead to a genetic disease. The importance of enzymes is shown by the fact that a lethal illness can be caused by the malfunction of just one type of enzyme out of the thousands of types present in our bodies.

One example is the most common type of phenylketonuria. A mutation of a single amino acid in the enzyme phenylalanine hydroxylase, which catalyzes the first step in the

degradation of phenylalanine, results in build-up of phenylalanine and related products. This can lead to mental retardation if the disease is untreated.

Another example is when germline mutations in genes coding for DNA repair enzymes cause hereditary cancer syndromes such as xeroderma pigmentosum. Defects in these enzymes cause cancer since the body is less able to repair mutations in the genome. This causes a slow accumulation of mutations and results in the development of many types of cancer in the sufferer.

Naming conventions

An enzyme's name is often derived from its substrate or the chemical reaction it catalyzes, with the word ending in ***-ase***. Examples are lactase, alcohol dehydrogenase and DNA polymerase. This may result in different enzymes, called isozymes, with the same function having the same basic name. Isoenzymes have a different amino acid sequence and might be distinguished by their optimal pH, kinetic properties or immunologically. Isoenzyme and isozyme are homologous proteins. Furthermore, the normal physiological reaction an enzyme catalyzes may not be the same as under artificial conditions. This can result in the same enzyme being identified with two different names. *E.g.* Glucose isomerase, used industrially to convert glucose into the sweetener fructose, is a xylose isomerase *in vivo*.

The International Union of Biochemistry and Molecular Biology have developed a nomenclature for enzymes, the **EC numbers**; each enzyme is described by a sequence of four numbers preceded by "EC". The first number broadly classifies the enzyme based on its mechanism.

The top-level classification is

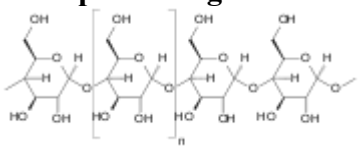
- EC 1 *Oxidoreductases*: catalyze oxidation/reduction reactions
- EC 2 *Transferases*: transfer a functional group (*e.g.* a methyl or phosphate group)
- EC 3 *Hydrolases*: catalyze the hydrolysis of various bonds
- EC 4 *Lyases*: cleave various bonds by means other than hydrolysis and oxidation
- EC 5 *Isomerases*: catalyze isomerization changes within a single molecule
- EC 6 *Ligases*: join two molecules with covalent bonds.

According to the naming conventions, enzymes are generally classified into six main family classes and many sub-family classes. Some web-servers, e.g., EzyPred and bioinformatics tools have been developed to predict which main family class and sub-family class an enzyme molecule belongs to according to its sequence information alone via the pseudo amino acid composition.

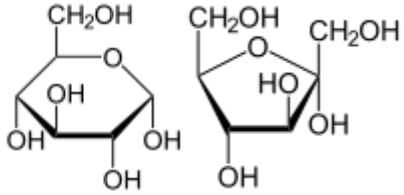
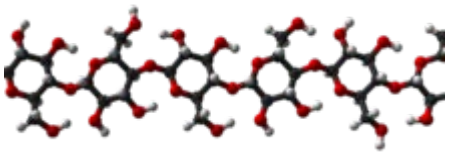
Industrial applications

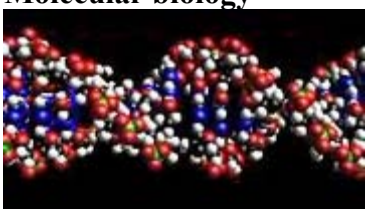
Enzymes are used in the chemical industry and other industrial applications when extremely specific catalysts are required. However, enzymes in general are limited in the

number of reactions they have evolved to catalyze and also by their lack of stability in organic solvents and at high temperatures. Consequently, protein engineering is an active area of research and involves attempts to create new enzymes with novel properties, either through rational design or *in vitro* evolution. These efforts have begun to be successful, and a few enzymes have now been desiged "from scratch" to catalyze reactions that do not occur in nature.

Application	Enzymes used	Uses
Food processing  Amylases catalyze the release of simple sugars from starch.	Amylases from fungi and plants	Production of sugars from starch, such as in making high-fructose corn syrup. In baking, catalyze breakdown of starch in the flour to sugar. Yeast fermentation of sugar produces the carbon dioxide that raises the dough.
	Proteases	Biscuit manufacturers use them to lower the protein level of flour.
Baby foods	Trypsin	To predigest baby foods
Brewing industry  Germinating barley used for malt	Enzymes from barley are released during the mashing stage of beer production.	They degrade starch and proteins to produce simple sugar, amino acids and peptides that are used by yeast for fermentation.
	Industrially produced barley enzymes	Widely used in the brewing process to substitute for the natural enzymes

		found in barley.
	Amylase, glucanases, proteases	Split polysaccharides and proteins in the malt.
	Betaglucanases and arabinoxylanases	Improve the wort and beer filtration characteristics.
	Amyloglucosidase and pullulanases	Low-calorie beer and adjustment of fermentability.
	Proteases	Remove cloudiness produced during storage of beers.
	Acetolactatedecarboxylase (ALDC)	Increases fermentation efficiency by reducing diacetyl formation.
Fruit juices	Cellulases, pectinases	Clarify fruit juices.
Dairy industry	Rennin, derived from the stomachs of young ruminant animals (like calves and lambs)	Manufacture of cheese, used to hydrolyze protein
	Microbially produced enzyme	Now finding increasing use in the dairy industry
 Roquefort cheese	Lipases	Is implemented during the production of Roquefort cheese to enhance the ripening of the blue-mould cheese.
	Lactases	Break down lactose to glucose and galactose.
Meat tenderizers	Papain	To soften meat for cooking

	Amylases, amyloglucosideases and glucoamylases	Converts starch into glucose and various syrups.
Starch industry  Glucose Fructose	Glucose isomerase	Converts glucose into fructose in production of high fructose syrups from starchy materials. These syrups have enhanced sweetening properties and lower calorific values than sucrose for the same level of sweetness.
Paper industry  A paper mill in South Carolina	Amylases, Xylanases, Cellulases and ligninases	Degrade starch to lower viscosity, aiding sizing and coating paper. Xylanases reduce bleach required for decolorising; cellulases smooth fibers, enhance water drainage, and promote ink removal; lipases reduce pitch and lignin-degrading enzymes remove lignin to soften paper.
Biofuel industry  Cellulose in 3D	Cellulases Ligninases	Used to break down cellulose into sugars that can be fermented Use of lignin waste
Biological detergent	Primarily proteases, produced in an	Used for presoak conditions and

	extracellular form from bacteria	direct liquid applications helping with removal of protein stains from clothes
	Amylases	Detergents for machine dish washing to remove resistant starch residues
	Lipases	Used to assist in the removal of fatty and oily stains
	Cellulases	Used in biological fabric conditioners
Contact lens cleaners	Proteases	To remove proteins on contact lens to prevent infections
Rubber industry	Catalase	To generate oxygen from peroxide to convert latex into foam rubber
Photographic industry	Protease (ficin)	Dissolve gelatin off scrap film, allowing recovery of its silver content.
Molecular biology  Part of the DNA double helix	Restriction enzymes, DNA ligase and polymerases	Used to manipulate DNA in genetic engineering, important in pharmacology, agriculture and medicine. Essential for restriction digestion and the

polymerase chain reaction. Molecular biology is also important in forensic science.

Chapter 12

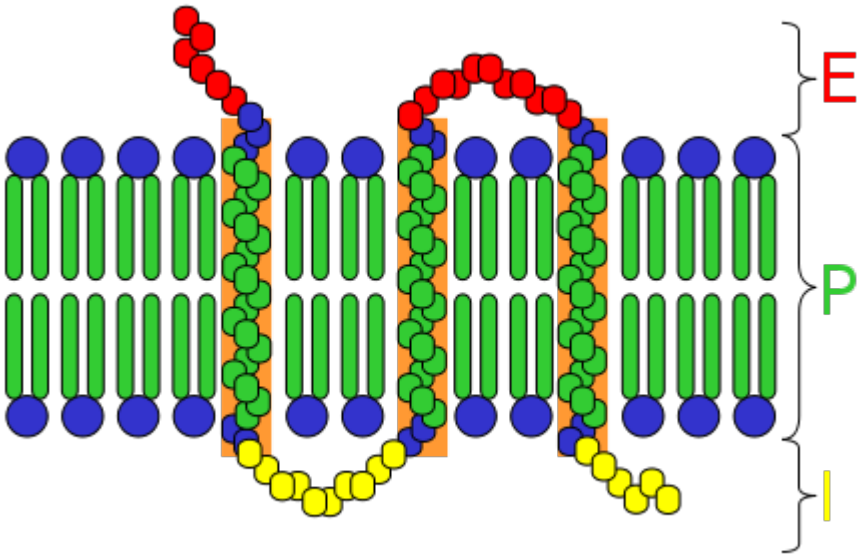
Receptor (Biochemistry)

In biochemistry, a **receptor** is a protein molecule, embedded in either the plasma membrane or the cytoplasm of a cell, to which one or more specific kinds of signaling molecules may attach. A molecule which binds (attaches) to a receptor is called a ligand, and may be a peptide (short protein) or other small molecule, such as a neurotransmitter, a hormone, a pharmaceutical drug, or a toxin. Each kind of receptor can bind only certain ligand shapes. Each cell typically has many receptors, of many different kinds. Simply put, a receptor functions as a keyhole that opens a neural path when the proper ligand is inserted.

Ligand binding stabilizes a certain receptor conformation (the three-dimensional shape of the receptor protein, with no change in sequence). This is often associated with gain of or loss of protein activity, ordinarily leading to some sort of cellular response. However, some ligands (e.g. antagonists) merely block receptors without inducing any response. Ligand-induced changes in receptors result in cellular changes which constitute the biological activity of the ligands. Many functions of the human body are regulated by these receptors responding uniquely to specific molecules like this.

Overview

The shapes and actions of receptors are studied by X-ray crystallography, dual polarisation interferometry, computer modelling, and structure-function studies, which have advanced the understanding of drug action at the binding sites of receptors. Structure activity relationships correlate induced conformational changes with biomolecular activity, and are studied using dynamic techniques such as circular dichroism and dual polarisation interferometry.



Transmembrane receptor:E=extracellular space; I=intracellular space; P=plasma membrane

Depending on their functions and ligands, several types of receptors may be identified:

- Some receptor proteins are peripheral membrane proteins.
- Many hormone and neurotransmitter receptors are transmembrane proteins: transmembrane receptors are embedded in the phospholipid bilayer of cell membranes, that allow the activation of signal transduction pathways in response to the activation by the binding molecule, or ligand.
    - Metabotropic receptors are coupled to G proteins and affect the cell indirectly through enzymes which control ion channels.
    - Ionotropic receptors (also known as ligand-gated ion channels) contain a central pore which opens in response to the binding of ligand.
- Another major class of receptors are intracellular proteins such as those for steroid and intracrine peptide hormone receptors. These receptors often can enter the cell nucleus and modulate gene expression in response to the activation by the ligand.

Membrane receptors are isolated from cell membranes by complex extraction procedures using solvents, detergents, and/or affinity purification.

Binding and activation

Ligand binding is an equilibrium process. Ligands bind to receptors and dissociate from them according to the law of mass action.

$$[\text{Ligand}] \cdot [\text{Receptor}] \overset{K_d}{\rightleftharpoons} [\text{Ligand-receptor complex}]$$

(the brackets stand for concentrations)

One measure of how well a molecule fits a receptor is the binding affinity, which is inversely related to the dissociation constant K_d. A good fit corresponds with high affinity and low K_d. The final biological response (e.g. second messenger cascade, muscle contraction), is only achieved after a significant number of receptors are activated.

The receptor-ligand affinity is greater than enzyme-substrate affinity. Whilst both interactions are specific and reversible, there is no chemical modification of the ligand as seen with the substrate upon binding to its enzyme.

If the receptor exists in two states, then the ligand binding must account for these two receptor states.

Constitutive activity

A receptor which is capable of producing its biological response in the absence of a bound ligand is said to display "constitutive activity". The constitutive activity of receptors may be blocked by inverse agonist binding. Mutations in receptors that result in increased constitutive activity underlie some inherited diseases, such as precocious puberty (due to mutations in luteinizing hormone receptors) and hyperthyroidism (due to mutations in thyroid-stimulating hormone receptors).

Theories of drug receptor interaction

Occupation theory

Drug effect is directly proportional to number of receptors occupied Drug effect ceases as drug-receptor complex dissociate

Ariens & Stephenson theory

introduced Terms of "affinity" & "efficacy" Affinity: ability of the drug to combine with receptor to create drug-receptor complex Efficacy: ability of the drug-receptor complex to initiate a response

Affinity “drug-receptor interaction” is governed by the law of mass action.

In this theory

Agonist: drug with high affinity & high intrinsic activity Partial agonist: drug with high affinity & low intrinsic activity Antagonist: drug with high affinity & low intrinsic activity

Rate theory

The activation of receptors is directly proportional to the total number of encounters of the drug with its receptors per unit time Pharmacological activity is directly proportional to the rate of dissociation & association not number of receptors occupied

Agonist:drug with fast association & fast dissociation Partial agonist:drug with intermediate association & intermediate dissociation Antagonist:drug with fast association & slow dissociation

Induced fit theory

As the drug approaches the receptor the receptors alters the conformation of its binding site to produce drug—receptor complex

Agonists versus antagonists

Not every ligand that binds to a receptor also activates the receptor. The following classes of ligands exist:

- *(Full) agonists* are able to activate the receptor and result in a maximal biological response. Most natural ligands are full agonists.
- *Partial agonists* do not activate receptors thoroughly, causing responses which are partial compared to those of full agonists.
- *Antagonists* bind to receptors but do not activate them. This results in receptor blockage, inhibiting the binding of other agonists.
- *Inverse agonists* reduce the activity of receptors by inhibiting their constitutive activity.

Peripheral membrane protein receptors

These receptors are relatively rare compared to the much more common types of receptors that cross the cell membrane. An example of a receptor that is a peripheral membrane protein is the elastin receptor.

Transmembrane receptors

Metabotropic receptors

G protein-coupled receptors

These receptors are also known as **seven transmembrane receptors** or **7TM** receptors, because they pass through the membrane seven times.

- Muscarinic acetylcholine receptor (Acetylcholine and Muscarine)
- Adenosine receptors (Adenosine)

- Adrenoceptors (also known as Adrenergic receptors, for *adren*aline, and other structurally related hormones and drugs)
- GABA receptors, Type-B (γ-Aminobutyric acid or GABA)
- Angiotensin receptors (Angiotensin)
- Cannabinoid receptors (Cannabinoids)
- Cholecystokinin receptors (Cholecystokinin)
- Dopamine receptors (Dopamine)
- Glucagon receptors (Glucagon)
- Metabotropic glutamate receptors (Glutamate)
- Histamine receptors (Histamine)
- Olfactory receptors (for the sense of smell)
- Opioid receptors (Opioids)
- Protease-activated receptors
- Rhodopsin (a photoreceptor protein)
- Secretin receptors (Secretin)
- Serotonin receptors, except Type-3 (Serotonin, also known as 5-Hydroxytryptamine or 5-HT)
- Somatostatin receptors (Somatostatin)
- Calcium-sensing receptor (Calcium)
- Chemokine receptors (Chemokines)
- *many more ...*

Receptor tyrosine kinases

These receptors detect ligands and propagate signals via the tyrosine kinase of their intracellular domains. This family of receptors includes;

- Erythropoietin receptor (Erythropoietin)
- Insulin receptor (Insulin)
- Eph receptors
- Insulin-like growth factor 1 receptor
- various other growth factor and cytokine receptors
- ....

Guanylyl cyclase receptors

- GC-A & GC-B: receptors for Atrial-natriuretic peptide (ANP) and other natriuretic peptides
- GC-C: Guanylin receptor

Ionotropic receptors

Ionotropic receptors are heteromeric or homomeric oligomers . They are receptors that respond to extracellular ligands and receptors that respond to intracellular ligands.

Extracellular ligands

Receptor	Ligand	Ion current
Nicotinic acetylcholine receptor	Acetylcholine, Nicotine	Na^+, K^+, Ca^{2+}
Glycine receptor (GlyR)	Glycine, Strychnine	Cl^- > HCO^-_3
GABA receptors: GABA-A, GABA-C	GABA	Cl^- > HCO^-_3
Glutamate receptors: NMDA receptor, AMPA receptor, and Kainate receptor	Glutamate	Na^+, K^+, Ca^{2+}
5-HT_3 receptor	Serotonin	Na^+, K^+
P2X receptors	ATP	Ca^{2+}, Na^+, Mg^{2+}

Intracellular ligands

Receptor	Ligand	Ion current
cyclic nucleotide-gated ion channels	cGMP (vision), cAMP and cGTP (olfaction)	Na^+, K^+
IP_3 receptor	IP_3	Ca^{2+}
Intracellular ATP receptors	ATP (closes channel)	K^+
Ryanodine receptor	Ca^{2+}	Ca^{2+}

The entire repertoire of human plasma membrane receptors is listed at the Human Plasma Membrane Receptome.

Intracellular receptors

Transcription factors

- nuclear receptor:
  - Steroid hormone receptor

Various

- Ionotropic receptors (IP_3 receptor above)
- sigma1 (neurosteroids)
- G protein-coupled receptors

Role in genetic disorders

Many genetic disorders involve hereditary defects in receptor genes. Often, it is hard to determine whether the receptor is nonfunctional or the hormone is produced at decreased level; this gives rise to the "pseudo-hypo-" group of endocrine disorders, where there appears to be a decreased hormonal level while in fact it is the receptor that is not responding sufficiently to the hormone.

Receptor regulation

Cells can increase (upregulate) or decrease (downregulate) the number of receptors to a given hormone or neurotransmitter to alter its sensitivity to this molecule. This is a locally acting feedback mechanism.

Receptor desensitization

Ligand-bound desensitation Vol. 135. No. 5 2130–2136</ref>

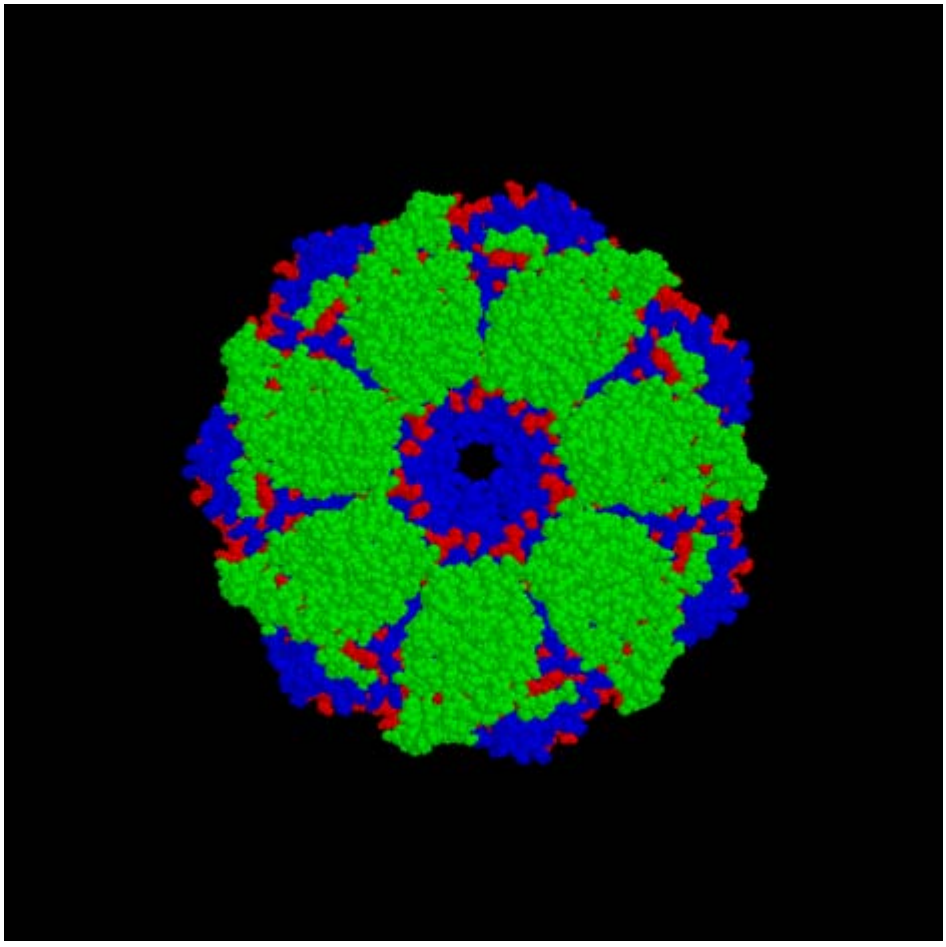
- Uncoupling of receptor effector molecules.
- Receptor sequestration (internalization).

In immune system

The main receptors in the immune system are pattern recognition receptors (PRRs), toll-like receptors (TLRs), killer activated and killer inhibitor receptors (KARs and KIRs), complement receptors, Fc receptors, B cell receptors and T cell receptors.

Chapter 13

Chaperone (Protein)

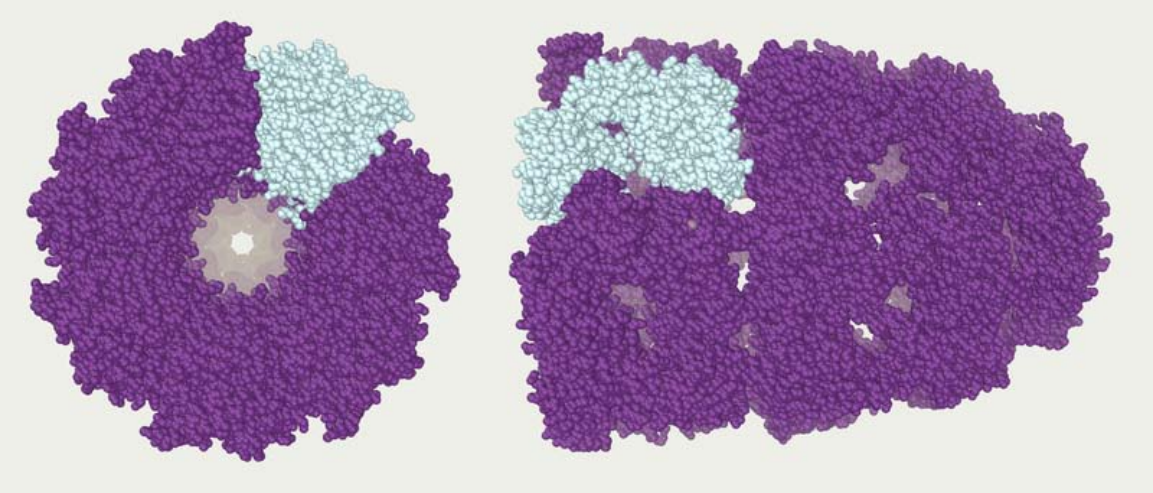


A top-view of the GroES/GroEL bacterial chaperone complex model

In molecular biology, **chaperones** are proteins that assist the non-covalent folding or unfolding and the assembly or disassembly of other macromolecular structures, but do

not occur in these structures when the structures are performing their normal biological functions having completed the processes of folding and/or assembly. The common perception that chaperones are primarily concerned with protein folding is incorrect. The first protein to be called a chaperone assists the assembly of nucleosomes from folded histones and DNA and such assembly chaperones, especially in the nucleus, are concerned with the assembly of folded subunits into oligomeric structures.

Chaperones do not necessarily convey steric information required for proteins to fold: thus statements of the form 'chaperones fold proteins' can be misleading. One major function of chaperones is to prevent both newly synthesised polypeptide chains and assembled subunits from aggregating into nonfunctional structures. It is for this reason that many chaperones, but by no means all, are also heat shock proteins because the tendency to aggregate increases as proteins are denatured by stress. However, 'steric chaperones' directly assist in the folding of specific proteins by providing essential steric information, e.g. prodomains of bacterial proteases, lipase-specific foldases, or chaperones in fimbrial adhesion systems.



The crystal structure of the chaperonin.

Location and functions

Many chaperones are heat shock proteins, that is, proteins expressed in response to elevated temperatures or other cellular stresses. The reason for this behaviour is that protein folding is severely affected by heat and, therefore, some chaperones act to repair the potential damage caused by misfolding. Other chaperones are involved in folding newly made proteins as they are extruded from the ribosome. Although most newly synthesized proteins can fold in absence of chaperones, a minority strictly requires them for the same.

Macromolecular crowding may be important in chaperone function. The crowded environment of the cytosol can accelerate the folding process, since a compact folded protein will occupy less volume than an unfolded protein chain. However, crowding can

reduce the yield of correctly-folded protein by increasing protein aggregation. Crowding may also increase the effectiveness of the chaperone proteins such as GroEL, which could counteract this reduction in folding efficiency.

Other types of chaperones are involved in transport across membranes, for example membranes of the mitochondria and endoplasmic reticulum (ER) in eukaryotes. Bacterial translocation—specific chaperone maintains newly synthesized precursor polypeptide chains in a translocation-competent (generally unfolded) state and guides them to the translocon.

New functions for chaperones continue to be discovered, such as assistance in protein degradation, bacterial adhesin activity, and in responding to diseases linked to protein aggregation.

Human chaperone proteins

Chaperones are found in, for example, the endoplasmic reticulum (ER), since protein synthesis often occurs in this area.

In endoplasmic reticulum

In the endoplasmic reticulum (ER) there are general, lectin- and non-classical molecular chaperones helping to fold proteins.

- General chaperones: BiP, GRP94, GRP170.
- Lectin chaperones: calnexin and calreticulin
- Non-classical molecular chaperones: HSP47 and ERp29
- Folding chaperones:
    - Protein disulfide isomerase (PDI),
    - *Peptidyl prolyl cis-trans-isomerase* (PPI),
    - ERp57

Nomenclature and examples of bacterial and archeal chaperones

There are many different families of chaperones; each family acts to aid protein folding in a different way. In bacteria like *E. coli*, many of these proteins are highly expressed under conditions of high stress, for example, when placed in high temperatures. For this reason, the term "heat shock protein" has historically been used to name these chaperones. The prefix "Hsp" designates that the protein is a heat shock protein.

Hsp60

Hsp60 (GroEL/GroES complex in *E. coli*) is the best characterized large (~ 1 MDa) chaperone complex. GroEL is a double-ring 14mer with a greasy hydrophobic patch at its opening; it is so large it can accommodate native folding of 54-kDa GFP in its lumen.

GroES is a single-ring heptamer that binds to GroEL in the presence of ATP or ADP. GroEL/GroES may not be able to undo previous aggregation, but it does compete in the pathway of misfolding and aggregation. Also acts in mitochondrial matrix as molecular chaperone.

Hsp70

Hsp70 (DnaK in *E. coli*) is perhaps the best characterized small (~ 70 kDa) chaperone.

The Hsp70 proteins are aided by Hsp40 proteins (DnaJ in *E. coli*), which increase the ATP consumption rate and activity of the Hsp70s.

It has been noted that increased expression of Hsp70 proteins in the cell results in a decreased tendency towards apoptosis.

Although a precise mechanistic understanding has yet to be determined, it is known that Hsp70's have a high-affinity bound state to unfolded proteins when bound to ADP, and a low-affinity state when bound to ATP.

It is thought that many Hsp70s crowd around an unfolded substrate, stabilizing it and preventing aggregation until the unfolded molecule folds properly, at which time the Hsp70s lose affinity for the molecule and diffuse away. Hsp70 also acts as a mitochondrial and chloroplastic molecular chaperone in eukaryotes.

Hsp90

Hsp90 (HtpG in *E. coli*) may be the least understood chaperone. Its molecular weight is about 90 kDa, and it is necessary for viability in eukaryotes (possibly for prokaryotes as well).

Heat shock protein 90 (Hsp90) is a molecular chaperone essential for activating many signaling proteins in the eukaryotic cell.

Each Hsp90 has an ATP-binding domain, a middle domain, and a dimerization domain. Originally thought to clamp onto their substrate protein (also known as a client protein) upon binding ATP, the recently published structures by Vaughan *et al.* and Ali *et al.* indicate that client proteins may bind externally to both the N-terminal and middle domains of Hsp90.

Hsp90 may also require co-chaperones like immunophilins, Sti1, p50 (Cdc37) and Aha1 and also cooperates with the Hsp70 chaperone system.

Hsp100

Hsp100 (Clp family in *E. coli*) proteins have been studied *in vivo* and *in vitro* for their ability to target and unfold tagged and misfolded proteins.

Proteins in the Hsp100/Clp family form large hexameric structures with unfoldase activity in the presence of ATP. These proteins are thought to function as chaperones by processively threading client proteins through a small 20 Å (2 nm) pore, thereby giving each client protein a second chance to fold.

Some of these Hsp100 chaperones, like ClpA and ClpX, associate with the double-ringed tetradecameric serine protease ClpP; instead of catalyzing the refolding of client proteins, these complexes are responsible for the targeted destruction of tagged and misfolded proteins.

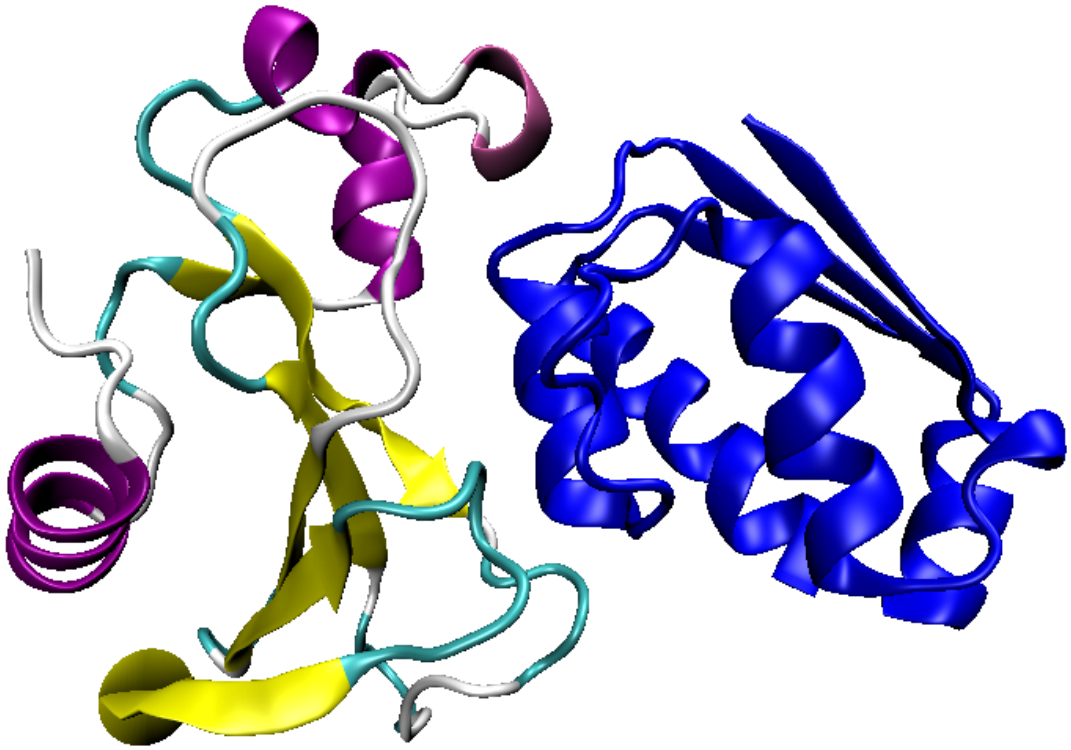
Hsp104, the Hsp100 of Saccharomyces cerevisiae, is essential for the propagation of many yeast prions. Deletion of the HSP104 gene results in cells that are unable to propagate certain prions.

History

The investigation of chaperones has a long history. The term `molecular chaperone` appeared first in the literature in 1978, and was invented by Ron Laskey to describe the ability of a nuclear protein called nucleoplasmin to prevent the aggregation of folded histone proteins with DNA during the assembly of nucleosomes. The term was later extended by R. John Ellis in 1987 to describe proteins that mediated the post-translational assembly of protein complexes. In 1988, it was realised that similar proteins mediated this process in both prokaryotes and eukaryotes. The details of this process were determined in 1989, when the ATP-dependent protein folding was demonstrated *in vitro*.

Chapter 14

Multiprotein Complex



Bacillus amyloliquefaciens proteins in a complex

A **multiprotein complex** (or **protein complex**) is a group of two or more associated polypeptide chains. If the different polypeptide chains contain different protein domain, the resulting multiprotein complex can have multiple catalytic functions. This is distinct from a multienzyme polypeptide, in which multiple catalyic domains are found in a single polypeptide chain.

Protein complexes are a form of quaternary structure. Proteins in a protein complex are linked by non-covalent protein-protein interactions, and different protein complexes have

different degrees of stability over time. These complexes are a cornerstone of many (if not most) biological processes and together they form various types of molecular machinery that perform a vast array of biological functions. Increasingly, scientists view the cell as composed of modular supramolecular complexes, each of which performs an independent, discrete biological function. By existing in proximity, the speed and selectivity of binding interactions between enzymatic complex and substrates can be vastly improved, leading to higher cellular efficiency. Unfortunately, many of the techniques used to break open cells and isolate proteins are inherently disruptive to such large complexes, so their protein complexes within the cell may be even more widespread than can be detected. Examples include the proteasome for molecular degradation, the metabolon for oxidative energy generation, and the ribosome for protein sythesis. In stable complexes, large hydrophobic interfaces between proteins typically bury surface areas larger than 2500 square angstroms.

However, complexes need not be stable. Understanding the functional interactions of proteins is an important research focus in biochemistry and cell biology. Protein complex formation sometimes serves to activate or inhibit one or more of the complex members and in this way, protein complex formation can be similar to phosphorylation. Individual proteins can participate in the formation of a variety of different protein complexes. Different complexes perform different functions, and the same complex can perform very different functions that depend on a variety of factors. Some of these factors are:

- Which cellular compartment the complex exists in when it is contained
- Which stage in the cell cycle the complexes are present
- The nutritional status of the cell
- Others

Many protein complexes are well understood, particularly in the model organism *Saccharomyces cerevisiae* (a strain of yeast). For this relatively simple organism, the study of protein complexes is now being performed genome wide and the elucidation of most protein complexes of the yeast is undergoing.

The molecular structure of protein complexes can be determined by experimental techniques such as X-ray crystallography or nuclear magnetic resonance. Increasingly the theoretical option of protein-protein docking is also becoming available. One method that is commonly used for identifying the members of protein complexes is immunoprecipitation.

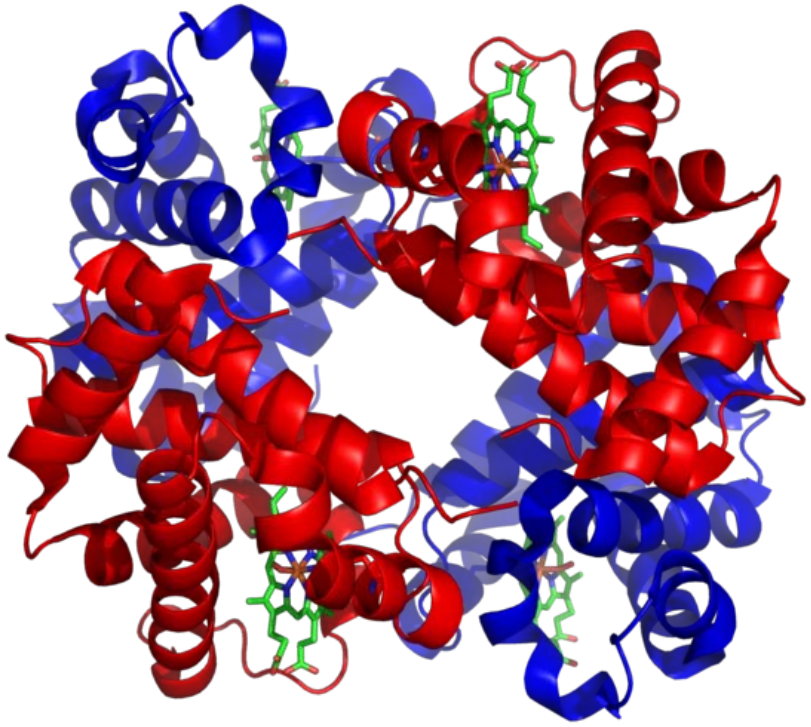
Homomultimeric and heteromultimeric proteins

The subunits of a multimeric protein may be identical as in a homomultimeric protein or different as in a heteromultimeric protein.

The voltage-gated potassium channels in the plasma membrane of a neuron are heteromultimeric proteins composed of four of forty known alpha subunits. Subunits must be of the same subfamily to form the multimeric protein channel. The tertiary

structure of the channel allows ions to flow through the hydrophobic plasma membrane. Connexons are an example of a homomultimeric protein composed of six identical connexins. A cluster of connexons forms the gap-junction in two neurons that transmit signals through an electrical synapse.

Globular Protein



3-dimensional structure of hemoglobin, a globular protein.

Globular proteins, or **spheroproteins** are one of the two main protein classes, comprising "globe"-like proteins that are more or less soluble in aqueous solutions (where they form colloidal solutions). This main characteristic helps distinguishing them from fibrous proteins (the other class), which are practically insoluble.

The term globin can refer more specifically to proteins including the globin fold.

Globular structure and solubility

The term globular protein is quite old (dating probably from the 19th century) and is now somewhat archaic given the hundreds of thousands of proteins and more elegant and descriptive structural motif vocabulary. The globular nature of these proteins can be determined without the means of modern techniques, but only by using ultracentrifuges or dynamic light scattering techniques.

The spherical structure is induced by the protein's tertiary structure. The molecule's apolar (hydrophobic) amino acids are bounded towards the molecule's interior whereas polar (hydrophilic) amino acids are bound outwards, allowing dipole-dipole interactions with the solvent, which explains the molecule's solubility.

Globular protein is only marginally stable because the free energy released when the protein folded into its native conformation is relatively small. This is because protein folding requires entropic cost. As a primary sequence of a polypeptide chain can form numerous conformations, native globular structure restricts its conformation to a few only. It results in a decrease in randomness, although non-covalent interactions such as hydrophobic interactions stabilizes the structure.

Although it is still unknown how proteins fold up naturally, new evidence has helped advance understanding. Part of the protein folding problem is that several noncovalent, weak interactions are formed, such as Hydrogen bonds and Van der Waals interactions. Via several techniques, the mechanism of protein folding is currently being studied. Even in the protein’s denatured state, it can be folded into the correct structure. Globular proteins seem to have two mechanisms for protein folding, either the diffusion-collision model or nucleation condensation model, although recent findings have shown globular proteins, such as PTP-BL PDZ2, which fold with characteristic features of both models. These new findings have shown that the transition states of proteins may affect the way that it folds. The folding of globular proteins has also recently been connected to treatment of diseases and anti-cancer ligands have been developed which bind to the folded but not the natural protein. These studies in return have shown that the folding of globular proteins affects its function.

Recall the second law of thermodynamics, the free energy difference between unfolded and folded state is contributed by enthalpy and entropy changes. As the free energy difference in globular protein results from folding into its native conformation is small, it is marginally stable, thus providing a rapid turnover rate and provide effective control of protein degradation and synthesis.

Role

Unlike fibrous proteins which only play a structural function, globular proteins can act as:

- Enzymes, by catalyzing organic reactions taking place in the organism in mild conditions and with a great specificity. Different esterases fulfill this role.

- Messengers, by transmitting messages to regulate biological processes. This function is done by hormones, i.e. insulin etc.
- Transporters of other molecules through membranes
- Stocks of amino acids.
- Regulatory roles are also performed by globular proteins rather than fibrous proteins. fulfill this role

Members

Among the most known globular proteins is hemoglobin, a member of the globin protein family. Other globular proteins are the immunoglobulins (IgA, IgD, IgE, IgG and IgM), and alpha, beta and gamma globulins.