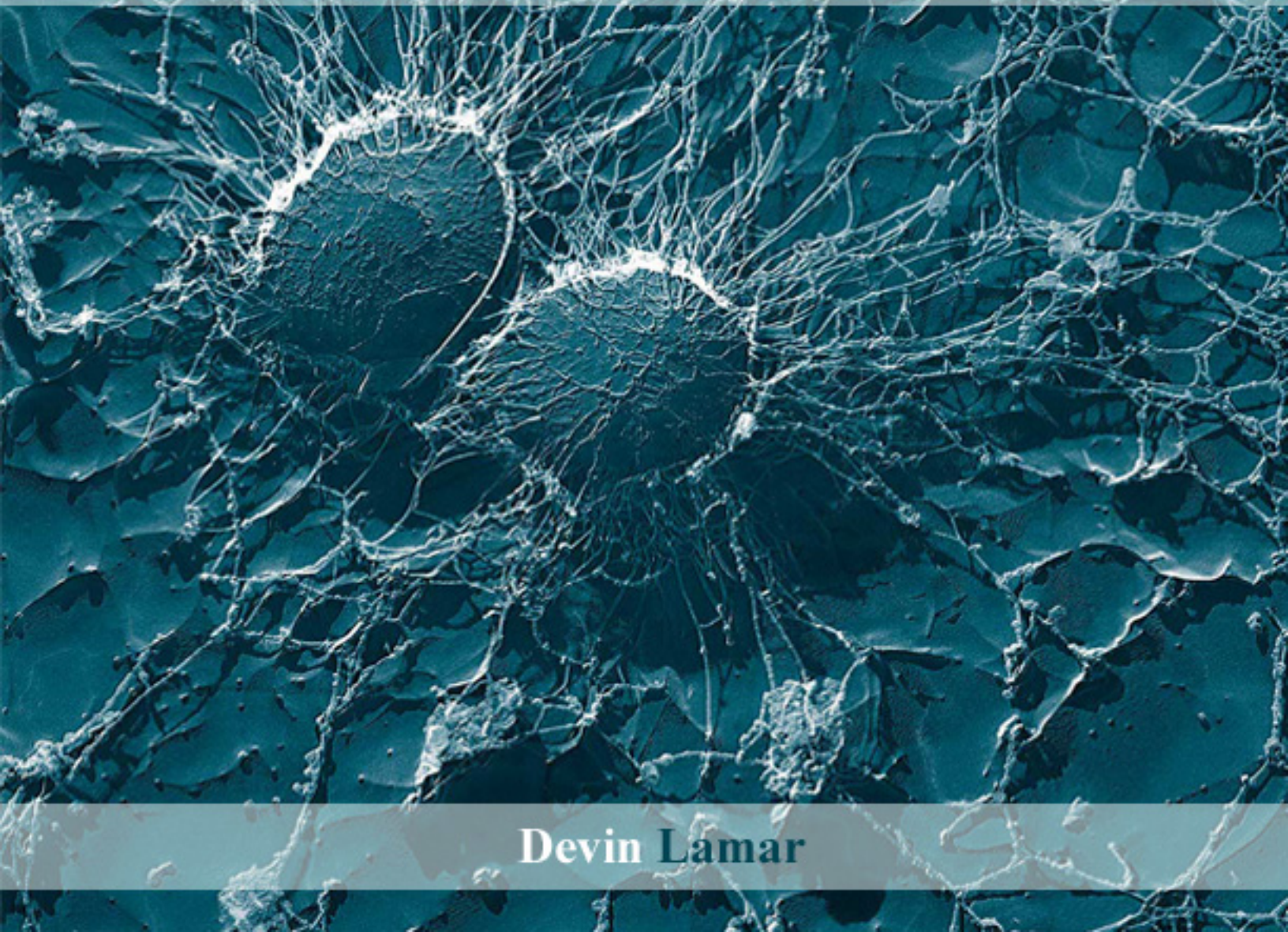
Chemical Biology



Devin Lamar

First Edition, 2012

ISBN 978-81-323-3141-4

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

Table of Contents

Chapter- 1

Chemical Biology

Chemical biology is a scientific discipline spanning the fields of chemistry and biology that involves the application of chemical techniques and tools, often compounds produced through synthetic chemistry, to the study and manipulation of biological systems. This is a subtle difference from biochemistry, which is classically defined as the study of the chemistry of biomolecules. For example, a biochemist would seek to understand the three-dimensional structure of a protein and how that structure relates to the chemistry of the protein. Chemical biologists attempt to utilize chemical principles to modulate systems to either investigate the underlying biology or create new function. In this way, the research done by chemical biologists is often closer related to that of cell biology than biochemistry. In short, biochemists deal with the chemistry *of* biology, chemical biologists deal with chemistry *applied to* biology.

Introduction

Some forms of chemical biology attempt to answer biological questions by directly probing living systems at the chemical level. In contrast to research using biochemistry, genetics, or molecular biology, where mutagenesis can provide a new version of the organism or cell of interest, chemical biology studies sometime probe systems *in vitro* and *in vivo* with small molecules that have been designed for a specific purpose or identified on the basis of biochemical or cell-based screening.

Chemical biology is one of many interfacial sciences that are characteristic of a general trend away from older, reductionist fields toward those whose goals are to achieve a description of scientific holism. In this sense, it is related to other fields such as proteomics. Chemical biology has historical and philosophical roots in medicinal chemistry, supramolecular chemistry (particularly host-guest chemistry), bioorganic chemistry, pharmacology, genetics, biochemistry, and metabolic engineering.

Systems of interest

Proteomics

Proteomics investigates the proteome, the set of expressed proteins at a given time under defined conditions. As a discipline, Proteomics has moved past rapid protein identification and has developed into a biological assay for quantitative analysis of complex protein samples by comparing protein changes in differently perturbed systems. Current goals in proteomics include determining protein sequences, abundance and any post-translational modifications. Also of interest are protein-protein interactions, cellular distribution of proteins and understanding protein activity. Another important aspect of proteomics is the advancement of technology to achieve these goals.

Protein levels, modifications, locations and interactions are complex and dynamic properties. With this complexity in mind, experiments need to be carefully designed to answer specific questions especially in the face of the massive amounts of data that are generated by these analyses. The most valuable information comes from proteins that are expressed differently in a system being studied. These proteins can be compared relative to each other using quantitative proteomics which allows a protein to be labeled with a mass tag. Proteomic technologies must be sensitive and robust, it is for these reasons, the mass spectrometer has been the workhorse of protein analysis. The high precision of mass spectrometry can distinguish between closely related species and species of interest can be isolated and fragmented within the instrument. Its applications to protein analysis was only possible in the late 1980s with the development of protein and peptide ionization with minimal fragmentation. These breakthroughs were ESI and MALDI. Mass spectrometry technologies are modular and can be chosen or optimized to the system of interest.

Chemical biologists are poised to impact proteomics through the development of techniques, probes and assays with synthetic chemistry for the characterization of protein samples of high complexity. These approaches include the development of enrichment strategies, chemical affinity tags and probes.

Enrichment techniques

Samples for Proteomics contain a myriad of peptide sequences, the sequence of interest may be highly represented or of low abundance. However, for successful MS analysis the peptide should be enriched within the sample. Reduction of sample complexity is achieved through selective enrichment using affinity chromatography techniques. This involves targeting a peptide with a distinguishing feature like a biotin label or a post translational modification. Interesting methods have been developed that include the use of antibodies, lectins to capture glycoproteins, immobilized metal ions to capture phosphorylated peptides and suicide enzyme substrates to capture specific enzymes. Here, chemical biologists can develop reagents to interact with substrates, specifically and tightly, to profile a targeted functional group on a proteome scale. Development of new enrichment strategies is needed in areas like non ser/thr/tyr phosphorylation sites and

other post translational modifications. Other methods of decomplexing samples relies on upstream chromatographic separations.

Affinity tags

Chemical synthesis of affinity tags has been crucial to the maturation of quantitative proteomics. iTRAQ, Tandem mass tags (TMT) and Isotope-coded affinity tag (ICAT) are protein mass-tags that consist of a covalently attaching group, a mass (isobaric or isotopic) encoded linker and a handle for isolation. Varying mass-tags bind to different proteins as a sort of footprint such that when analyzing cells of differing perturbations, the levels of each protein can be compared relatively after enrichment by the introduced handle. Other methods include SILAC and heavy isotope labeling. These methods have been adapted to identify complexing proteins by labeling a bait protein, pulling it down and analyzing the proteins it has complexed. Another method creates an internal tag by introducing novel amino acids that are genetically encoded in prokaryotic and eukaryotic organisms. These modifications create a new level of control and can facilitate photocrosslinking to probe protein-protein interactions. Additionally, keto, acetylene, azide, thioester, boronate and dehydroalanine containing amino acids can be used to selectively introduce tags, and novel chemical functional groups into proteins.

Enzyme probes

To investigate enzymatic activity as opposed to total protein, activity-based reagents have been developed to label the enzymatically active form of proteins. For example, serine hydrolases and cysteine proteases have been converted to suicide inhibitors. This strategy enhances the ability to selectively analyze low abundance constituents through direct targeting. Structures that mimic these inhibitors could be introduced with modifications that will aid proteomic analysis- like an identification handle or mass tag. Enzyme activity can also be monitored through converted substrate . This strategy relies on using synthetic substrate conjugates that contain moieties that are acted upon by specific enzymes. The product conjugates are then captured by an affinity reagent and analyzed. The measured concentration of product conjugate allow the determination of the enzyme velocity. Identification of enzyme substrates (of which there may be hundreds or thousands, many of which are unknown) is a problem of significant difficulty in proteomics and is vital to the understanding of signal transduction pathways in cells; techniques for labelling cellular substrates of enzymes is an area chemical biologists can address. A method that has been developed uses "analog-sensitive" kinases to label substrates using an unnatural ATP analog, facilitating visualization and identification through a unique handle.

Glycobiology

While DNA, RNA and proteins are all encoded at the genetic level, there exists a separate system of trafficked molecules in the cell that are not encoded directly at any direct level: sugars. Thus, glycobiology is an area of dense research for chemical biologists. For instance, live cells can be supplied with synthetic variants of natural sugars in order to

probe the function of the sugars in vivo. Carolyn Bertozzi at University of California, Berkeley has developed a method for site-specifically reacting molecules the surface of cells that have been labeled with synthetic sugars.

Combinatorial chemistry

Some chemical biologists use automated synthesis of many diverse compounds in order to experiment with effects of small molecules on biological processes. More specifically, they observe changes in the behaviors of proteins when small molecules bind to them. Such experiments may supposedly lead to discovery of small molecules with antibiotic or chemotherapeutic properties. These approaches are identical to those employed in the discipline of Pharmacology.

Molecular sensing

Chemical biologists are also interested in developing new small-molecule and biomolecule-based tools to study biological processes, often by molecular imaging techniques. The field of molecular sensing was popularized by Roger Tsien's work developing calcium-sensing fluorescent compounds as well as pioneering the use of GFP, for which he was awarded the 2008 Nobel Prize in Chemistry. Today, researchers continue to utilize basic chemical principles to develop new compounds for the study of biological metabolites and processes.

siRNA-A tool in chemical biology

siRNA or small interfering RNAs owe their origins to the difficulties the scientific community faced utilizing classical and reverse genetics methods in studying gene expression. Disrupting genes to study their functions is not always optimal; neither is mapping mutations back to their genes easy. The whole process is expensive as well as time-consuming which is why a lot of effort has been devoted to develop methods to silence gene expression in sequence specific manner using nucleic acids. They have the potential to be powerful tools in the field of chemical biology to study the chemistry of gene expression in therapeutic targets of bacteria and viruses.

A number of different types of nucleic acid molecules have already gained prominence because of their potential as therapeutics. They target mRNAs to silence the genes in a sequence specific manner. Oligodeoxyribonucleic acids, ODNs utilize steric interaction to silence gene expression. They can also form triple helices in conjunction with the DNA duplex. Whereas ribozymes can be chemically designed to target specific genes and cleave them in a sequence specific manner. The most promising of these methods however is utilization of short interfering RNA or siRNA to silence gene expression.

siRNA

siRNA or short interfering RNAs exist in nature as a means for the express purpose of controlling gene expression.

It was discovered in petunia as a post-transcriptional gene silencing measure. It is the resultant product when a long double stranded RNA of 20 -25 nucleotides length was processed in the cells by the enzyme DICER. The newly synthesized siRNA assemble into endoribonuclease-containing complexes known as RNA-induced silencing complexes (RISCs), unwinding in the process. The activated RISC then binds to the complementary RNA molecules by base pairing interactions between the siRNA strand and the mRNA, which is then cleaved. This mechanism is known as RNA interference or RNAi.

Designing and synthesizing siRNAs

It is now possible to order siRNAs designed and synthesized with the express purpose of targeting a particular sequence.

The ambion website has a lot of information on the optimal design of siRNAs.

siRNAs can be synthesized chemically, or enzymatically. RNase III or DICER can be used to cleave the long dsRNAs to produce siRNAs. However the most expedient method is the use of plasmids to express them in vivo by delivering them into the target cell using vectors. This method allows the siRNAs to be expressed in the target cell stably, over a period of time and overcomes the drawbacks of the transience of their effect. Numerous strategies have been developed in order to deliver the siRNA into the cell efficiently:

1. Electroporation
2. Local and systemic injection: This method was the first success scientists had in silencing genes using siRNAs. They were successfully delivered into highly vascularized tissue in mice through using high-pressure tail vein injection. Greater than 90% loss in gene expression was observed in the targets.
3. siRNA producing viruses: This method shows great promise in gene therapy and research is progressing in order to generate recombinant viruses which can produce siRNA in target cells.
4. Small molecules which enhance transdermal penetration: Research in this field is moving at a fast pace in order to synthesize small organic molecules which if injected in conjunction with siRNAs can help them penetrate into the target cells.

===Biological uses of the RNAi approach===

The principle purpose of studying siRNA mediated RNA interference is probably to investigate gene function. It is so much easier to make genetic knock-outs by simply introducing sequence specific siRNAs into cells, multicopy genes can be silenced in one fell swoop by this method. Creation of double-knockout mutants is also easier and consumes much less time. Using local injections in specific regions of the model organisms also help in creating spatially separated and restricted knockout. siRNAs are also being successfully used to screen whole genomes in organisms such as C. elegans & Drosophilla melanogaster. Even in mammalian systems such as Danio rerio (zebrafish) that usually prove intractable to all gene silencing methods, even dsRNA injection,

siRNA can do the job. It is paving a new way in development of therapeutics by identifying human gene orthologs in other species in a remarkably short period of time.

```
Numerous high throughput screening approaches are being developed to
screen large libraries of cells rapidly in order to identify drug
targets.
```

A brief description of few of the screening techniques:

1. Pooled Format Screening: A reagent library of RNAi has to be introduced to the cells so that a particular cell is in one particular reagent. The primary hits are then identified and their identity elucidate by sequencing techniques.
2. Arrayed Format Screening: Each RNAi reagent is placed in separate wells in a plate and multiple manipulations can be done to identify their targets, which are then detected by fluorescence readouts, imaging techniques and other methods as well. Thus the identity of the target cell can be determined through the identity of the reagent in the database.
3. Multiplexed methods: A combination of various assays can be used for high-throuput screening of candidate drug targets. For example, candidate genes can be identified through informatics based methods and then screened against a library of reagents. Many other such methods are being developed in order to make the job of screening therapeutic targets easier.

====siRNA based therapeutics:====

The future in this field rests in the development of siRNA based drugs. This could prove to be a powerful tool in gene based therapy. Research is now concentrated on developing strategies to design siRNA therapeutics for clinical use. A brief description of some novel strategies for siRNA drug development is provided here:

1. Direct Mutation Targeting: The siRNAs are designed to perfectly match mutant alleles but contain one or more mismatches with wild-type alleles, leading to specific degradation of the matching, mutant transcripts.
2. Indirect Mutation Targeting: The siRNA approach will not work if the mutant alleles are too similar to wild type. So an indirect approach is taken in which siRNAs are designed against disease linked markers such as SNP variations. The ones that are screened as positive are targeted for degradation.
3. Exon specific targeting: siRNAs are designed to target expressed regions (exons) of the gene.
4. Targeting exon skipped transcripts: If the problem in the gene lies in aberrant splicing post-transcription, siRNA can be designed to target the unnatural exon-exon interface arising as a result of such alternative splicing.

Conclusion

It is remarkable how much progress this field has made in a remarkably short time. Considering the fact that siRNA was first discovered only in the early 90s by Dr. David

Baulcombe in co suppression of purple color in petunias, the field has risen to the limelight in a meteoric pace particularly owing to the ward of the Nobel Prize in Medicine/Physiology to Dr. Andrew Fire and Craig Mello in 2006. With two siRNA based drugs in clinical trial stages this field shows remarkable promise for the future.

Employing biology

Many research programs are also focused on employing natural biomolecules to perform a task or act as support for a new chemical method or material. In this regard, researchers have shown that DNA can serve as a template for synthetic chemistry, self-assembling proteins can serve as a structural scaffold for new materials, and RNA can be evolved *in vitro* to produce new catalytic function.

Protein misfolding and aggregation as a cause of disease

A common form of aggregation is long, ordered spindles called amyloid fibrils which are implicated in Alzheimer's disease which have been shown to consist of cross-linked beta sheet regions perpendicular to the backbone of the polypeptide. Another form of aggregation occurs with prion proteins, the glycoproteins found with Creutzfeldt-Jakob disease and bovine spongiform encephalopathy. In both structures, aggregation occurs through hydrophobic interactions and water must be excluded from the binding surface before aggregation can occur. A movie of this process can be seen in "Chemical and Engineering News". The diseases associated with misfolded proteins are life-threatening and extremely debilitating which makes them an important target for chemical biology research.

Through the transcription and translation process, DNA encodes for specific sequences of amino acids. The resulting polypeptides fold into more complex secondary, tertiary, and quaternary structures to form proteins. Based on both the sequence and the structure, a particular protein is conferred its cellular function. However, sometimes the folding process fails due to mutations in the genetic code and thus the amino acid sequence or due to changes in the cell environment (e.g. pH, temperature, reduction potential, etc.). Misfolding occurs more often in aged individuals or in cells exposed to a high degree of oxidative stress, but a fraction of all proteins misfold at some point even in the healthiest of cells.

Normally when a protein does not fold correctly, molecular chaperones in the cell can encourage refolding back into its active form. When refolding is not an option, the cell can also target the protein for degradation back into its component amino acids via proteolytic, lysosomal, or autophagic mechanisms. However, under certain conditions or with certain mutations, the cells can no longer cope with the misfolded protein(s) and a disease state results. Either the protein has a loss-of-function, such as in cystic fibrosis, in which it loses activity or cannot reach its target, or the protein has a gain-of-function, such as with Alzheimer's disease, in which the protein begins to aggregate causing it to become insoluble and non-functional.

Protein misfolding has previously been studied using both computational approaches as well as *in vivo* biological assays in model organisms such as *Drosophila melanogaster* and *C. elegans*. Computational models use a *de novo* process to calculate possible protein structures based on input parameters such as amino acid sequence, solvent effects, and mutations. This method has the shortcoming that the cell environment has been drastically simplified, which limits the factors that influence folding and stability. On the other hand, biological assays can be quite complicated to perform *in vivo* with high-throughput like efficiency and there always remains the question of how well lower organism systems approximate human systems.

Dobson et al. propose combining these two approaches such that computational models based on the organism studies can begin to predict what factors will lead to protein misfolding. Several experiments have already been performed based on this strategy. In experiments on *Drosophila*, different mutations of beta amyloid peptides were evaluated based on the survival rates of the flies as well as their motile ability. The findings from the study show that the more a protein aggregates, the more detrimental the neurological dysfunction . Further studies using transthyretin, a component of cerebrospinal fluid which binds to beta amyloid peptide deterring aggregation but can itself aggregate especially when mutated, indicate that aggregation prone proteins may not aggregate where they are secreted and rather are deposited in specific organs or tissues based on each mutation. Kelly et al. have shown that the more stable, both kinetically and thermodynamically, a misfolded protein is the more likely the cell is to secrete it from the endoplasmic reticulum rather than targeting the protein for degradation. Additionally, the more stress that a cell feels from misfolded proteins, the more probable new proteins will misfold. These experiments as well as others having begun to elucidate both the intrinsic and extrinsic causes of misfolding as well as how the cell recognizes if proteins have folded correctly.

As more information is obtained on how the cell copes with misfolded proteins, new therapeutic strategies begin to emerge. An obvious path would be prevention of misfolding. However, if protein misfolding cannot be avoided, perhaps the cell's natural mechanisms for degradation can be bolstered to better deal with the proteins before they begin to aggregate. Before these ideas can be realized, many more experiments need to be done to understand the folding and degradation machinery as well as what factors lead to misfolding. More information about protein misfolding and how it relates to disease can be found in the recently published book by Dobson, Kelly, and Rameriz-Alvarado entitled Protein Misfolding Diseases Current and Emerging Principles and Therapies.

Chemical synthesis of peptides

In contrast to the traditional biotechnological practice of obtaining peptides or proteins by isolation from cellular hosts through protein expression, advances in chemical techniques for the synthesis and ligation of peptides has allowed for the total synthesis of some peptides and proteins. Chemical synthesis of proteins is a valuable tool in chemical biology as it allows for the introduction of non-natural amino acids as well as residue specific incorporation of "posttranslational modifications" such as phosphorylation,

glycosylation, acetylation, and even ubiquitination. These capabilities are valuable for chemical biologists as non-natural amino acids can be used to probe and alter the functionality of proteins, while post translational modifications are widely known to regulate the structure and activity of proteins. Although strictly biological techniques have been developed to achieve these ends, the chemical synthesis of peptides often has a lower technical and practical barrier to obtaining small amounts of the desired protein. Given the widely recognized importance of proteins as cellular catalysts and recognition elements, the ability to precisely control the composition and connectivity of polypeptides is a valued tool in the chemical biology community and is an area of active research.

While chemists have been making peptides for over 100 years, the ability to efficiently and quickly synthesize short peptides came of age with the development of Bruce Merrifield's solid phase peptide synthesis (SPPS). Prior to the development of SPPS, the concept of step-by-step polymer synthesis on an insoluble support was without chemical precedent. The use of a covalently bound insoluble polymeric support greatly simplified the process of peptide synthesis by reducing purification to a simple "filtration and wash" procedure and facilitated a boom in the field of peptide chemistry. The development and "optimization" of SPPS took peptide synthesis from the hands of the specialized peptide synthesis community and put it into the hands of the broader chemistry, biochemistry, and now chemical biology community. SPPS is still the method of choice for linear synthesis of polypeptides up to 50 residues in length and has been implemented in commercially available automated peptide synthesizers. One inherent shortcoming in any procedure that calls for repeated coupling reactions is the buildup of side products resulting from incomplete couplings and side reactions. This places the upper bound for the synthesis of linear polypeptide lengths at around 50 amino acids, while the "average" protein consists of 250 amino acids. Clearly, there was a need for development of "non-linear" methods to allow synthetic access to the average protein.

Although the shortcomings of linear SPPS were recognized not long after its inception, it took until the early 1990s for effective methodology to be developed to ligate small peptide fragments made by SPPS, into protein sized polypeptide chains. The oldest and best developed of these methods is termed native chemical ligation. Native chemical ligation was unveiled in a 1994 paper from the laboratory of Stephen B. H. Kent. Native chemical ligation involves the coupling of a C-terminal thioester and an N-terminal cysteine residue, ultimately resulting in formation of a "native" amide bond. Further refinements in native chemical ligation have allowed for kinetically controlled coupling of multiple peptide fragments, allowing access to moderately sized peptides such as an HIV-protease dimer and human lysozyme. Even with the successes and attractive features of native chemical ligation, there are still some drawbacks in the utilization of this technique. Some of these drawbacks include the installation and preservation of a reactive C-terminal thioester, the requirement of an N-terminal cysteine residue (which is the second least common amino acid in proteins), and the requirement for a sterically unincumbering C-terminal residue.

Other strategies that have been used for the ligation of peptide fragments using the acyl transfer chemistry first introduced with native chemical ligation include expressed protein ligation, sulfurization/desulfurization techniques, and use of removable thiol auxiliaries.

Expressed protein ligation allows for the biotechnological installation of a C-terminal thioester using intein biochemistry, thereby allowing the appendage of a synthetic N-terminal peptide to the recombinantly produced C-terminal portion. This technique allows for access to much larger proteins, as only the N-terminal portion of the resulting protein has to be chemically synthesized. Both sulfurization/desulfurization techniques and the use of removable thiol auxiliaries involve the installation of a synthetic thiol moiety to carry out the standard native chemical ligation chemistry, followed by removal of the auxiliary/thiol. These techniques help to overcome the requirement of an N-terminal cysteine needed for standard native chemical ligation, although the steric requirements for the C-terminal residue are still limiting.

A final category of peptide ligation strategies include those methods not based on native chemical ligation type chemistry. Methods that fall in this category include the traceless Staudinger ligation, azide-alkyne dipolar cycloadditions, and imine ligations.

Major contributors in this field today include Stephen B. H. Kent, Philip E. Dawson, and Tom W. Muir, as well as many others involved in methodology development and applications of these strategies to biological problems.

Protein design by directed evolution

One of the primary goals of protein engineering is the design of novel peptides or proteins with a desired structure and chemical activity. Because our knowledge of the relationship between primary sequence, structure, and function of proteins is limited, rational design of new proteins with enzymatic activity is extremely challenging. Directed evolution, repeated cycles of genetic diversification followed by a screening or selection process, can be used to mimic Darwinian evolution in the laboratory to design new proteins with a desired activity.

Several methods exist for creating large libraries of sequence variants. Among the most widely used are subjecting DNA to UV radiation or chemical mutagens, error-prone PCR, degenerate codons, or recombination . Once a large library of variants is created, selection or screening techniques are used to find mutants with a desired attribute. Common selection/screening techniques include fluorescence-activated cell sorting (FACS) , mRNA display , phage display, or *in vitro* compartmentalization . Once useful variants are found, their DNA sequence is amplified and subjected to further rounds of diversification and selection. Since only proteins with the desired activity are selected, multiple rounds of directed evolution lead to proteins with an accumulation beneficial traits.

There are two general strategies for choosing the starting sequence for a directed evolution experiment: *de novo* design and redesign. In a protein design experiment, an

initial sequence is chosen at random and subjected to multiple rounds of directed evolution. For example, this has been employed successfully to create a family of ATP-binding proteins with a new folding pattern not found in nature. Random sequences can also be biased towards specific folds by specifying the characteristics (such as polar vs. nonpolar) but not the specific identity of each amino acid in a sequence. Among other things, this strategy has been used to successfully design four-helix bundle proteins. Because it is often thought that a well-defined structure is required for activity, biasing a designed protein towards adopting a specific folded structure is likely to increase the frequency of desirable variants in constructed libraries.

In a protein redesign experiment, an existing sequence serves as the starting point for directed evolution. In this way, old proteins can be redesigned for increased activity or new functions. Protein redesign has been used for protein simplification, creation of new quaternary structures, and topological redesign of a chorismate mutase . To develop enzymes with new activities, one can take advantage of promiscuous enzymes or enzymes with significant side reactions. In this regard, directed evolution has been used on γ-humulene synthase, an enzyme that creates over 50 different sesquiterpenes, to create enzymes that selectively synthesize individual products . Similarly, completely new functions can be selected for from existing protein scaffolds. In one example of this, an RNA ligase was created from a zinc finger scaffold after 17 rounds of directed evolution. This new enzyme catalyzes a chemical reaction not known to be catalyzed by any natural enzyme .

Computational methods, when combined with experimental approaches, can significantly assist both the design and redesign of new proteins through directed evolution. Computation has been used to design proteins with unnatural folds, such as a right-handed coiled coil . These computational approaches could also be used to redesign proteins to selectively bind specific target molecules. By identifying lead sequences using computational methods, the occurrence of functional proteins in libraries can be dramatically increased before any directed evolution experiments in the laboratory.

Frances Arnold, Donald Hilvert, and Jack W. Szostak are significant researchers in this field.

Biocompatible click cycloladdition reactions in chemical biology

Recent advances in technology have allowed scientists to view substructures of cells at levels of unprecedented detail. Unfortunately these “aerial” pictures offer little information about the mechanics of the biological system in question. To be fully effective, precise imaging systems require a complementary technique that better elucidates the machinery of a cell. By attaching tracking devices (optical probes) to biomolecules *in vivo*, one can learn far more about cell metabolism, molecular transport, cell-cell interactions and many other processes

Bioorthogonal reactions

Successful labeling of a molecule of interest requires specific functionalization of that molecule to react chemospecifically with an optical probe. For a labeling experiment to be considered robust, that functionalization must minimally perturb the system. Unfortunately, these requirements can often be extremely hard to meet. Many of the reactions normally available to organic chemists in the laboratory are unavailable in living systems. Water- and redox- sensitive reactions would not proceed, reagents prone to nucleophilic attack would offer no chemospecificity, and any reactions with large kinetic barriers would not find enough energy in the relatively low-heat environment of a living cell. Thus, chemists have recently developed a panel of "bioorthogonal reactions" that proceed chemospecifically, despite the milieu of distracting reactive materials *in vivo*.

Design of bioorthogonal reagents

The coupling of an optical probe to a molecule of interest must occur within a reasonably short time frame; therefore, the kinetics of the coupling reaction should be highly favorable. Click chemistry is well suited to fill this niche, since click reactions are, by definition, rapid, spontaneous, selective, and high-yielding. Unfortunately, the most famous "click reaction," a [3+2] cycloaddition between an azide and an acyclic alkyne, is copper-catalyzed, posing a serious problem for use *in vivo* due to copper's toxicity.

To bypass the necessity for a catalyst, the lab of Dr. Carolyn Bertozzi introduced inherent strain into the alkyne species by using a cyclic alkyne. In particular, cyclooctyne reacts with azido-molecules with distinctive vigor. Further optimization of the reaction led to the use of difluorinated cyclooctynes (DIFOs), which increased yield and reaction rate. Other coupling partners discovered by separate labs to be analogous to cyclooctynes include trans cyclooctene[q], norbornene, and a cyclobutene-functionalized molecule.

Use in biological systems

As mentioned above, the use of bioorthogonal reactions to tag biomolecules requires that one half of the reactive "click" pair is installed in the target molecule, while the other is attached to an optical probe. When the probe is added to a biological system, it will selectively conjugate with the target molecule.

The most common method of installing bioorthogonal reactivity into a target biomolecule is through metabolic labeling. Cells are immersed in a medium where access to nutrients is limited to synthetically-modified analogues of standard fuels such as sugars. Consequently, these altered biomolecules are incorporated into the cells in the same manner as their wild-type brethren. The optical probe is then incorporated into the system to image the fate of the altered biomolecules. Other methods of functionalization include enzymatically inserting azides into proteins, crosslinking modified peptide domains, and synthesizing phospholipids conjugated to cyclooctynes

Future directions

As these bioorthogonal reactions are further optimized, they will likely be used for increasingly complex interactions involving multiple different classes of biomolecules. More complex interactions have a smaller margin for error, so increased reaction efficiency is paramount to continued success in optically probing cellular machinery. Also, by minimizing side reactions, the experimental design of a minimally perturbed living system is closer to being realized.

Discovery of biomolecules through metagenomics

The advances in modern sequencing technologies in the late 1990s allowed scientists to investigate DNA of communities of organisms in their natural environments, so-called "eDNA", without culturing individual species in the lab. This metagenomic approach enabled scientists to study a wide selection of organisms that were previously not characterized due in part to an incompetent growth condition. These sources of eDNA include, but are not limited to, soils, ocean, subsurface, hot springs, hydrothermal vents, polar ice caps, hypersaline habitats, and extreme pH environments. Of the many applications of metagenomics, chemical biologists and microbiologists such as Jo Handelsman, Jon Clardy, and Robert M. Goodman who are pioneers of metagenomics, explored metagenomic approaches toward the discovery of biologically active molecules such as antibiotics.

Functional or homology screening strategies have been used to identify genes that produce small bioactive molecules. Functional metagenomic studies are designed to search for specific phenotypes that are associated with molecules with specific characteristics. Homology metagenomic studies, on the other hand, are designed to examine genes to identify conserved sequences that are previously associated with the expression of biologically active molecules.

Functional metagenomic studies enable scientists to discover novel genes that encode biologically active molecules. These assays include top agar overlay assays where antibiotics generate zones of growth inhibition against test microbes, and pH assays that can screen for pH change due to newly synthesized molecules using pH indicator on an agar plate. Substrate-induced gene expression screening (SIGEX), a method to screen for the expression of genes that are induced by chemical compounds, has also been used to search genes with specific functions. These led to the discovery and isolation of several novel proteins and small molecules. For example, the Schipper group identified three eDNA derived AHL lactonases that inhibit biofilm formation of Pseudomonas aeruginosa via functional metagenomic assays. However, these functional screening methods require a good design of probes that detect molecules being synthesized and depend on the ability to express metagenomes in a host organism system.

In contrast, homology metagenomic studies led to a faster discovery of genes that have homologous sequences as the previously known genes that are responsible for the biosynthesis of biologically active molecules. As soon as the genes are sequenced,

scientists can compare thousands of bacterial genomes simultaneously. The advantage over functional metagenomic assays is that homology metagenomic studies do not require a host organism system to express the metagenomes, thus this method can potentially save the time spent on analyzing nonfunctional genomes. These also led to the discovery of several novel proteins and small molecules. For example, Banik et al. screened for clones containing genes associated with the synthesis of teicoplanin and vancomycin-like glycopeptide antibiotics and found two new biosynthetic gene clusters. In addition, an *in silico* examination from the Global Ocean Metagenomic Survey found 20 new lantibiotic cyclases.

There are challenges to metagenomic approaches to discover new biologically active molecules. Only 40% of enzymatic activities present in a sample can be expressed in *E. coli.*. In addition, the purification and isolation of eDNA is essential but difficult when the sources of obtained samples are poorly understood. However, collaborative efforts from individuals from diverse fields including bacterial genetics, molecular biology, genomics, bioinformatics, robots, synthetic biology, and chemistry can solve this problem together and potentially lead to the discovery of many important biologically active molecules.

Protein phosphorylation

Posttranslational modification of proteins with phosphate groups has proven to be a key regulatory step throughout all biological systems. Phosphorylation events, either phosphorylation by protein kinases or dephosphorylation by phosphoylases, result in protein activation or deactivation. These events have an immense impact on the regulation of physiological pathways, which makes the ability to dissect and study these pathways integral to understanding the details of cellular processes. There exist a number of challenges—namely the sheer size of the phosphoproteome, the fleeting nature of phosphorylation events and related physical limitations of classical biological and biochemical techniques—that have limited the advancement of knowledge in this area. A recent review provides a detailed examination of the impact of newly developed chemical approaches to dissecting and studying biological systems both in vitro and in vivo.

Through the use of a number of classes of small molecule modulators of protein kinases, chemical biologists have been able to gain a better understanding of the effects of protein phosphorylation. For example, nonselective and selective kinase inhibitors, such as a class of pyridinylimidazole compounds described by Wilson, et al., are potent inhibitors useful in the dissection of MAP kinase signaling pathways. These pyridinylimidazole compounds function by targeting the ATP binding pocket. Although this approach, as well as related approaches, with slight modifications, has proven effective in a number of cases, these compounds lack adequate specificity for more general applications. Another class of compounds, mechanism-based inhibitors, combines detailed knowledge of the chemical mechanism of kinase action with previously utilized inhibition motifs. For example, Parang, et al. describe the development of a “bisubstrate analog” that inhibits kinase action by binding both the conserved ATP binding pocket and an protein/peptide recognition site on the specific kinase. While there is no published in vivo data on

compounds of this type, the structural data acquired from in vitro studies have expanded the current understanding of how a number of important kinases recognize target substrates.

The development of novel chemical means of incorporating phophomimetics into proteins has provided important insight into the effects of phosphorylation events. Historically, phosphorylation events have been studied by mutating an identified phosphorylation site (serine, threonine or tyrosine) to an amino acid, such as alanine, that cannot be phosphorylated. While this approach has been successful in some cases, mutations are permanent in vivo and can have potentially detrimental effects on protein folding and stability. Thus, chemical biologists have developed new ways of investigating protein phosphorylation. By installing phospho-serine, phospho-threonine or analogous phosphonate mimics into native proteins, researchers are able to perform in vivo studies to investigate the effects of phosphorylation by extending the amount of time a phosphorylation event occurs while minimizing the often-unfavorable effects of mutations. Protein semisynthesis, or more specifically expressed protein ligation (EPL), has proven to be successful techniques for synthetically producing proteins that contain phosphomimetic molecules at either the C- or N-terminus. Additionally, researchers have built upon an established technique in which one can insert an unnatural amino acid into a peptide sequence by charging synthetic tRNA that recognizes a nonsense codon with a unnatural amino acid. Recent developments indicate that this technique can also be employed in vivo, although, due to permeability issues, these in vivo experiments using phosphomimetic molecules have not yet been possible.

Advances in chemical biology have also improved upon classical techniques of imaging kinase action. For example, the development of peptide biosensors—peptides containing incorporated fluorophore molecules—allowed for improved temporal resolution in in vitro binding assays. Experimental limitations, however, prevent this technique from being effectively used in vivo. One of the most useful techniques to study kinase action is Fluorescence Resonance Energy Transfer (FRET). To utilize FRET for phosphorylation studies, fluorescent proteins are coupled to both a phosphoamino acid binding domain and a peptide that can by phosphorylated. Upon phosphorylation or dephosphorylation of a substrate peptide, a conformational change occurs that results in a change in fluorescence. FRET has also been used in tandem with Fluorescence Lifetime Imaging Microscopy (FLIM) or fluorescently conjugated antibodies and flow cytometry to provide a detailed, specific, quantitative results with excellent temporal and spatial resolution.

Through the augmentation of classical biochemical methods as well as the development of new tools and techniques, chemical biologists have improved accuracy and precision in the study of protein phosphorylation.

Metal complexes in medicine

Metal complexes have many characteristics that can be advantageous in drug design. In comparison to organic-based medicines, metal complexes have many more coordination

numbers, geometries, and oxidation/reduction states that can be used to make structures that interact with targets in unique ways unavailable to most organic molecules. In addition, the cationic metal is advantageous in complexing with charged targets within biological systems like the phosphate backbone of DNA. Targets of metal-based medicines include DNA, proteins, and enzymes. Each target tupe is described in turn below.

Metal complexes targeting DNA

DNA has been the primary target of metal complexes due to the ability of cationic metal interacting with the anionic backbone of DNA. The anticancer chemotherapy drug cisplatin covalently binds to DNA, which disrupts transcription and leads to programed cell death. Assuming early detection, cisplatin cures almost all cases of testicular cancer. This drug, however, has severe side effects and great effort is being made to improve drug delivery including attachment to single-walled carbon nanotubes, encapsulation in proteins cages, among other clever strategies.

Another major effort for anticancer metal-based drugs centers around stabilization of the G-quadruplex of DNA. These drugs generally have a non-covalent interaction with the G-quadruplex, as well as a planar positively charged structure.

Metal complexes targeting enzymes and proteins

Though DNA has been a primary target for inorganic medicines, enzymes and proteins also can be modulated through interactions with these compounds. Metal complexes can interact with the amino acids with the highest reduction potential (histidine, cysteine, and selenocysteine). Metals used in such complexes include gold, platinum, ruthenium, vanadium, cobalt and others. Several new potential therapeutic complexes are currently in the process of discovery and investigation.

Gold

Some gold complexes are showing potential as medicines. A rheumatoid arthritis drug (auranofin, a gold(I) phosphine complex) has shown value in treating parasitic disease through inhibiting thioredoxin glutathione reductase.

Platinum

Along with cisplatin, many other platinum complexes are potential therapeutics. Like auranofin, terpyridine platinum inhibits thioredoxin reductase with nanomolar IC50. This complex also is an inhibitor of the common target enzyme topoisomerase I. Yet another family of complexes with potential anticancer properties are dichloro(SMP)-platinum(II) complexes. These complexes target the matrix metalloproteinase, where the complex coordinates with amino acids of the enzyme in the coordination sites previously held by chlorides, and through the smp ligand. As seen by these few examples, platinum complexes are a particularly active area of research for metal-based medicines.

Ruthenium

Ruthenium complexes have anticancer activity. A library of glutathione transferase inhibitors were created through a combination of ethacrynic acid (a known inhibitor of the enzyme) and ruthenium complexes.

Vanadium

Vanadium complexes have been used in multiple therapeutic settings. A new area in which vanadium may have a great medicinal impact is through the oxovanadium porphyrin complexes. These complexes have demonstrated HIV-1 reverse transcriptase inhibition in vitro.

Issues and outlook

Though there is currently much excitement in the field of metal-based medicines, many challenges still face researchers. One such challenge is selectivity of complexes in vivo. Many of these complexes can bind to common proteins like serum albumin in addition to other proteins with amino acids that are common in protein-metal complex interactions like histidine, cysteine, and selenocysteine. Along with selectivity issues, much is yet unknown about mechanisms through which metal complexes interact with proteins. How complexing between a given metal complex and target protein or enzyme occurs is often unknown or unclear and requires much more elucidation before truly effective metal complexes can be designed and delivered. Currently, physicians utilize very few metal-based medicines in the clinics. For example, none of the 21 drugs approved by the U.S. Food and Drug Administration (FDA) in 2008 were inorganic. However,with the success of cisplatin in cancer treatment, it is not unreasonable to anticipate more metal complexes will be actively used in the treatment of diseases.

Synthetic biology

Synthetic biology focuses on the manipulation of biological components to form new systems or the generation of living systems with synthetic parts. The canonical idea of synthetic biology is the creation of new life, but recently it has come to include bioengineering in terms of the use of interchangeable components to give novel outputs. In the search for modular parts, it is most facile if the building blocks contribute independently to the function of the whole unit so that the modules can be recombined in predictable ways. It is useful for synthetic biologists to define “life”: in this context, to be alive an organism must be capable of Darwinian evolution – genetic mutation, self-replication and inheritance of mutations.

Synthetic cells

J. Craig Venter’s group has created the first “synthetic” cell – the first cells to exist with fully synthetic DNA. Venter was able to manipulate the synthetic genome to dictate the proteins expressed in the organism. Note that these were not fully synthetic cells – but

that the synthetic DNA was able to take over all metabolic processes necessary for cell survival and proliferation.

DNA as interchangeable parts

DNA is composed of repeating modular units consisting of an anion phosphate group that forms the polyanion backbone, and nucleotide base pairs that engage in Watson-Crick base pairing to form the double strand. Because the molecular recognition of DNA is mostly based on the polyanion backbone, the nucleotides can be modified without altering the structural integrity of the DNA. Steven Benner's group has generated an artificial genetic alphabet of eight new base pairs that can be amplified by polymerase chain reaction; this indicates that these base pairs can be used in systems that undergo Darwinian evolution.

Proteins as interchangeable parts

Amino acids

Amino acids are poor modular building blocks because they don't act independently and there is a fundamental lack of understanding about the relationship between linear amino acid sequences and the folding and functionality of proteins. Chemical biologists have been able to create small peptide secondary structures through rational design such as alpha helices based on the manipulation of hydrophobic packing interactions.

Protein secondary structure

Modules consisting of protein secondary structure can be designed to perform specific functions; for example, it has been demonstrated that alpha helices can be used as functional peptide catalysts. The Ghadiri group has created a template peptide that promotes the ligation of two modified helices by bringing the helices into close proximity by specifically designed hydrophobic interactions of the helices with the template.

Folded proteins

Fully folded proteins can be combined in novel ways to generate specific non-natural outcomes. This is highly useful commercially from drug development to the production of polymers – one can imagine the economic benefits if scientists can design systems in which proteins catalyze reactions without the necessity of excessive human intervention to produce commercially relevant materials. For example, the Keasling group has developed a series of proteins that catalyze conversion of acetyl CoA, a common cellular metabolite, into a precursor for the potent antimalarial drug artemisinin.

Modifying molecular switches

Signaling pathways can be modified to be turned on or off by non-natural ligands or inputs to the system. For instance, systems can be modified so that they are autoinhibited

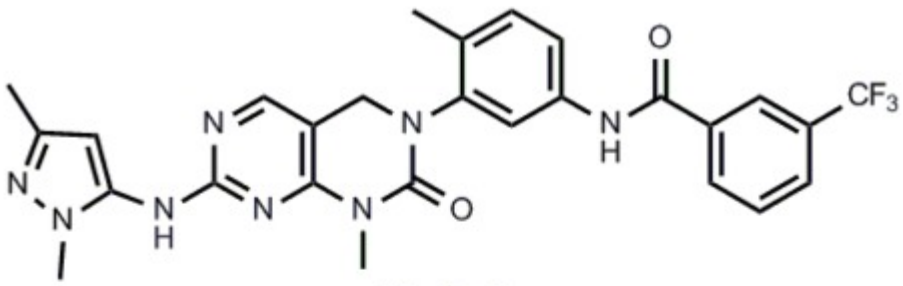
by non-natural proteins that release their inhibition upon binding with a specific molecule that is different from the natural signaling molecule of the path. This allows new approaches to studying signal circuits specifically and with user-designed inputs.

Chemical approaches to stem-cell biology

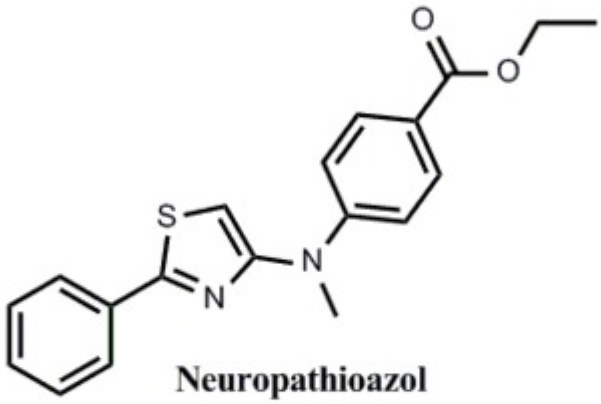
Advances in stem-cell biology have typically been driven by discoveries in molecular biology and genetics. These have included optimization of culture conditions for the maintenance and differentiation of pluripotent and multipotent stem-cells and the deciphering of signaling circuits that control stem-cell fate. However, chemical approaches to stem-cell biology have recently received increased attention due to the identification of several small molecules capable of modulating stem-cell fate in vitro. A small molecule approach offers particular advantages over traditional methods in that it allows a high degree of temporal control, since compounds can be added or removed at will, and tandem inhibition/activation of multiple cellular targets.

Small molecules that modulate stem-cell behavior are commonly identified in high-throughput screens. Libraries of compounds are screened for the induction of a desired phenotypic change in cultured stem-cells. This is usually observed through activation or repression of a fluorescent reporter or by detection of specific cell surface markers by FACS or immunohistochemistry. Hits are then structurally optimized for activity by the synthesis and screening of secondary libraries. The cellular targets of the small molecule can then be identified by affinity chromatography, mass spectrometry, or DNA microarray.

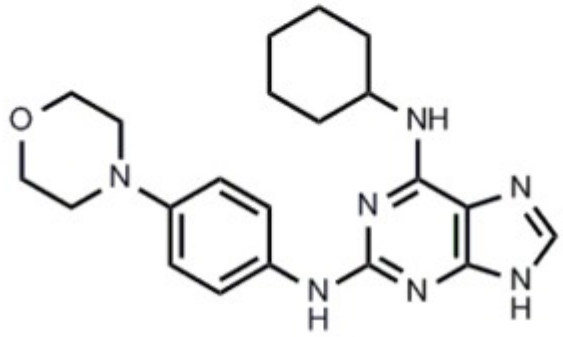
A trademark of pluripotent stem-cells, such as embryonic stem-cells (ESCs), is the ability to self-renew indefinitely. The conventional use of feeder cells and various exogenous growth factors in the culture of ESCs presents a problem in that the resulting highly variable culture conditions make the long-term expansion of un-differentiated ESCs challenging. Ideally, chemically defined culture conditions could be developed to maintain ESCs in a pluripotent state indefinitely. Toward this goal, the Ding laboratory at the Scripps Research Institute has identified a small molecule that can preserve the long-term self-renewal of ESCs in the absence of feeder cells and other exogenous growth factors. This novel molecule, called pluripotin, was found to simultaneously inhibit multiple differentiation inducing pathways.



Pluripotin



Neuropathioazol



Reversine

Small molecule modulators of stem-cell fate.

The utility of stem-cells is in their ability to differentiate into all cell types that make up an organism. Differentiation can be achieved in vitro by favoring development toward a particular cell type through the addition of lineage specific growth factors, but this process is typically non-specific and generates low yields of the desired phenotype. Alternatively, inducing differentiation by small molecules is advantageous in that it allows for the development of completely chemically defined conditions for the generation of one specific cell type. A small molecule, neuropathiazol, has been identified which can specifically direct differentiation of multipotent neural stem cells into neurons. Neuropathiazol is so potent that neurons develop even in conditions which normally favor the formation of glial cells, a powerful demonstration of controlling differentiation by chemical means.

Because of the ethical issues surrounding ESC research, the generation of pluripotent cells by reprogramming existing somatic cells into a more "stem-like" state is a promising alternative to the use of standard ESCs. By genetic approaches, this has recently been achieved in the creation of ESCs by somatic cell nuclear transfer and the generation of induced pluripotent stem-cells by viral transduction of specific genes. From a therapeutic perspective, reprogramming by chemical means would be safer than genetic methods because induced stem-cells would be free of potentially dangerous transgenes. Several examples of small molecules which can de-differentiate somatic cells have been identified. In one report, lineage-committed myoblasts were treated with a compound, named reversine, and observed to revert to a more stem-like phenotype. These cells were then shown to be capable of differentiating into osteoblasts and adipocytes under appropriate conditions.

Stem-cell therapies are currently the most promising treatment for many degenerative diseases. Chemical approaches to stem-cell biology support the development of cell-based therapies by enhancing stem-cell growth, maintenance, and differentiation in vitro. Small molecules that have been shown to modulate stem-cell fate are potential therapeutic candidates and provide a natural lean-in to pre-clinical drug development. Small molecule drugs could promote endogenous stem-cells to differentiate, replacing previously damaged tissues and thereby enhancing the body's own regenerative ability. Further investigation of molecules that modulate stem-cell behavior will only unveil new therapeutic targets.

Fluorescence for assessing protein location and function

Fluorophores and techniques to tag proteins

Organisms are composed of cells, which in turn, are composed of macromolecules e.g. proteins, ribosomes, etc. These macromolecules interact with each other, changing their concentration and suffering chemical modifications. The main goal of many biologists is to understand these interactions, using MRI, ESR, electrochemistry, and fluorescence among others. The advantages of fluorescence reside in its high sensitivity, non-invasiveness, safe detection and ability to modulate the fluorescence signal. Fluorescence was mainly observed from small organic dyes attached to antibodies to the protein of interest. Later, fluorophores could directly recognize organelles, nucleic acids, and important ions in living cells. In the past decade, the discovery of green fluorescent protein (GFP), by Roger Y. Tsien, hybrid system and quantum dots have enable assessing protein location and function more precisely. Three main types of fluorophores are used: small organic dyes, green fluorescent proteins, and quantum dots. Small organic dyes usually are less than 1 kD, and have been modified to increase photostability, enhance brightness, and reduce self-quenching. Quantum dots have very sharp wavelength, high molar absorptivity and quantum yield. Both organic dyes and quantum dyes do not have the ability to recognize the protein of interest without the aid of antibodies, hence they must use immunolabeling. Since the size of the fluorophore-targeting complex typically exceeds 200 kD, it might interfere with multiprotein recognition in protein complexes, and other methods should be use in parallel. An advantage includes diversity of

properties and a limitation is the ability of targeting in live cells. Green fluorescent proteins are genetically encoded and can be covalently fused to your protein of interest. A more developed genetic tagging technique is the tetracysteine biarsenical system, which requires modification of the targeted sequence that includes four cysteines, which binds membrane-permeable biarsenical molecules, the green and the red dyes "FlAsH" and "ReAsH", with picomolar affinity. Both fluorescent proteins and biarsenical tetracysteine can be expressed in live cells, but present major limitations in ectopic expression and might cause lose of function. Giepmans shows parallel applications of targeting methods and fluorophores using GFP and tetracysteine with ReAsH for α-tubulin and β-actin, respectively. After fixation, cells were immunolabeled for the Golgi matrix with QD and for the mitochondrial enzyme cytochrome with Cy5.

Protein dynamics

Fluorescent techniques have been used assess a number of protein dynamics including protein tracking, conformational changes, protein-protein interactions, protein synthesis and turnover, and enzyme activity, among others.

Three general approaches for measuring protein net redistribution and diffusion are single-particle tracking, correlation spectroscopy and photomarking methods. In single-particle tracking, the individual molecule must be both bright and sparse enough to be tracked from one video to the other. Correlation spectroscopy analyzes the intensity fluctuations resulting from migration of fluorescent objects into and out of a small volume at the focus of a laser. In photomarking a fluorescent protein can be dequenched in a subcellular area with the use of intense local illumination and the fate of the marked molecule can be imaged directly. Michalet and coworkers used quantum dots for single-particle tracking using biotin-quantum dots in HeLa cells.

One of the best ways to detect conformational changes in proteins is to sandwich said protein between two fluorophores. FRET will respond to internal conformational changes result from reorientation of the fluorophore with respect to the other. Dumbrepatil sandwiched an estrogen receptor between a CFP (cyan fluorescent protein) and a YFP (yellow fluorescent protein) to study conformational changes of the receptor upon binding of a ligand.

FRET can detect dynamic protein-protein interaction in live cells providing the fluorophores get close enough. Galperin *et al.* used three fluorescent proteins to study multiprotein interactions in live cells.

Tetracysteine biarsenical systems can be used to study protein synthesis and turnover, which requires discrimination of old copies from new copies. In principle, a tetracysteine-tagged protein is labeled with FlAsH for a short time, leaving green labeled proteins. The protein synthesis is then carried out in the presence of ReAsH, labeling the new proteins as red.

One can also use fluorescence to see endogenous enzyme activity, typically by using a quenched activity based proteomics (qABP). Covalent binding of a qABP to the active site of the targeted enzyme will provide direct evidence concerning if the enzyme is responsible for the signal upon release of the quencher and regain of fluorescence.

The unique combination of high spatial and temporal resolution, nondestructive compatibility with living cells and organisms, and molecular specificity insure that fluorescence techniques will remain central in the analysis of protein networks and systems biology.

Applications of DNA microarrays in chemical biology

Planar surfaces functionalized with single- or double-stranded nucleic acids have enabled researchers to address a variety of salient biological and biochemical questions in recent years. The general architecture of modern DNA microarrays reflects the historical progression from the sequence-specific probing of whole chromosomes immobilized on glass slides (as early as 1961 with fluorescent *in situ* hybridization) and the low-density porous membrane arrays available since the early 1990s, to the high-density (10^2-10^4 features/mm^2) solid support platforms that exist today. The massively parallel processing capabilities of these picomolar-range contemporary arrays provide for the generation of large data sets and multiplexed analysis. Furthermore, several top-down and bottom-up assembly methodologies provide researchers with the option for "in-house" production of arrays from custom oligonucleotide libraries or the use of commercial genome chips, notably those developed by Affymetrix and Agilent Technologies.

DNA microarrays can be used to conduct several general types of experiments, most of which rely on the hybridization of fluorescently-labeled single-stranded DNA molecules isolated from a biological sample to their single-stranded complement probes presented on an array. One of the earliest conceived applications for DNA microarrays was for single-nucleotide polymorphism (SNP) genotyping. Since SNPs are a "quick and dirty" approach to detect genetic indicators of pathologies and lineages, arrays theoretically provide a facile method for diagnosis; this was confirmed experimentally in the late 1990s in the successful SNP analysis of human tumors . Although there are currently commercially available arrays (e.g. bovine mapping chips) to characterize SNPs, it seems likely that the nascent availability of high-throughput and low-cost pyrosequencing will become the preferred method of recognition, or replace the need for SNP detection altogether with rapid whole-genome sequencing.

A different application of microarray technology that has become the gold standard for RNA analysis in recent years is the widespread utilization of expression microarrays, or "gene chips". Gene chip preparation calls for the quantitative reverse transcription of the total cellular RNA pool into labeled and fragmented single-stranded DNA prior to hybridization-based capture. Up- and down-regulation of genes in response to stressors or disease states are quantitatively compared in cell lines and organisms. Coupled expression microarray and quantitative proteomics experiments have allowed for the in-depth exploration of the oftentimes non-linear relationship between the abundance of a

particular transcribed message and that of its corresponding translated protein. These integrative studies, partially enabled by quantitative DNA microarray technology, have been successfully applied to a variety of biological systems, including yeast, bovine, mouse, bacterial, and human. The expression analysis community has amassed such a significant amount of expression microarray data that they are freely available in public databases.

These types of surfaces can also be used to analyze DNA-protein interactions on a genome-wide scale via chromatin immunoprecipitation, followed by an array-based analysis of the DNA (ChIP-chip). ChIP-chip experiments are enabled by the co-purification of a DNA-binding protein of interest with its corresponding genomic loci when a cross-linked chromatin extract is probed with an antibody to said protein. After purification, amplification and labeling, the DNA is applied to a microarray representing the entire genome; the data are plotted as a histogram that resolves the specific genomic regions associated with that protein. ChIP-chip experiments have provided the scientific community with a wealth of information about the steady-state genomic locations of DNA-binding proteins, such as histones, transcription factors, and polymerase machinery, and have also been successfully applied to studies on the dynamics of transcription factor binding. The data from these experiments may be further manipulated to computationally derive consensus binding sequences for some transcription factors, giving the opportunity for insight into the in vivo behavior of the factor, deeper than simple information about localization.

DNA microarrays are also amenable to the direct analysis of protein-DNA interactions in kinetic binding assays as analyzed by surface plasmon resonance (SPR). This experimental approach also relies on single-stranded DNA immobilized on a high-density array; however, the quantitative readout is based on a change in the optical properties of the DNA-functionalized surface when a protein flowed over the surface binds to the sequence in a particular surface feature. DNA-functionalized arrays analyzed with SPR in this way have yielded kinetic data regarding fundamental molecular biological processes. Recently, SPR analysis of a DNA microarray and components of the DNA replication machinery helped to elucidate the biochemical nuances of the replication fork.

High-density DNA microarrays have emerged as an important component of the chemical biology toolkit. The existing technology allows for the construction of customizable, as well as general, arrays and provides researchers with the opportunity to generate robust data from many different types of biological inputs. Considering the relatively recent shift in the scientific community away from binary perturbation/readout studies and toward “big science” and large data sets, it seems likely that DNA microarrays will continue to enable pertinent biological research for many years to come.

Microfluidics in chemical biology

Due to its physical dimensions, microfluidics both provides a unique platform to utilize chemical biology tools and serves as a chemical biology tool in itself. Defined as the manipulation of fluids through micron sized channels, the field of microfluidics has been

studied extensively over the past twenty years, and much is known about how fluids behave at this scale. As such, this knowledge can, and has been used to manipulate biological samples in ways that cannot be achieved using standard bulk methods.

The main advantages achieved through miniaturization of sample volume with regards to chemical biology applications include the ability to perform high-throughput experiments using a minimum of sample, the means to isolate, amplify and detect rare events from a complex mixture, and the resources to perturb the environment of a cellular sample at the scale of the cell itself(1-3). Through these capabilities researchers have been able to use microfluidics to crystallize proteins(4), perform the polymerase chain reaction(5-6), sequence DNA(5), study protein expression of single cells(7-8), perturb embryonic development in flies(9), culture cells(10) as well as perform many other important biological studies.

The ability to design and manufacture devices to perform microfluidic experiments using well established approaches lends to the utility of studying chemical biology with microfluidics. The most common material used for device manufacturing is polydimethylsioloxane (PDMS)(2). This material is far and away the most popular among researchers due to its compatible properties with biological systems. These characteristics include its relative inertness to most substances, its transparency to ultraviolet and visible light, its malleability and its permeability to gases(2). Additionally, PDMS surfaces can be treated to render them either hydrophilic or hydrophobic, depending on the desired application(2). This versatility allows PDMS to be used in nearly all microfluidic applications. Despite its wide range of uses, there are instances where other materials are preferred. Glass is a common alternative when PDMS is not desirable. Soft lithography is the most common method for making PDMS devices. This technique is relatively cheap and can be used to make nearly any architecture used in microfluidic experiments.

One unique feature that results from miniaturization of the sample vessel is the inevitable increased surface area to volume ratio. This inherent feature of microfluidic experiments can either lend to the advantages of using microfluidics or it can necessitate further refinement of experimental technique. In some instances, it is desirable to be able to direct molecules of interest to the interface between two phases. In this case, the enhanced surface area relative to the total reaction volume lends to the success of the experimental design. In other instances, it is necessary to prevent the migration of molecules to the surface. The most common instance of this is the propensity of protein molecules to adsorb at the interface between either air and water or oil and water. For these applications, it is necessary to modify the surfaces with either a surfactant or some other chemical additive to prevent this undesired effect.

Depending upon the nature of the desired experiment, the manner in which the fluids are manipulated and the number of phases present within the fluid flow can be different. The Reynold's number (Re) determines whether fluid flow is laminar or turbulent. In laminar flow, the exchange of miscible fluids flowing parallel to each other is due to diffusion, and is thus slow. This characteristic has been harnessed to produce stable gradients of

small molecules within fluid streams(11). Rather than using a single liquid phase, it is also possible to use two liquid phases in order to generate droplets. The most common method for generating droplets includes the flow of an aqueous stream perpendicular to an oil stream(12). When these two streams meet at a T-junction, uniform, aqueous droplets are formed that are surrounded by an oil phase. Depending upon the geometry of the microfluidic device as well as the flow rates used, droplets can also be formed using a flow-focusing device.

Microfluidics has a vast potential for single-molecule studies. In order to detect single molecules, it is often necessary to enhance or amplify a signal of interest (13). In bulk methods solutions, an amplified signal from a single molecule will continually be diluted to below the detection limit of nearly every fluorophore or other signal read-out. In small features rendered possible through microfluidics, however, the amplification of a single molecule will be confined within a volume ranging anywhere from nanoliters to picoliters(13). An amplified signal has the potential to grow in intensity above the limit of detection in these small volumes, thus allowing for single-molecule studies (13). The versatility in microfluidic device design and experimental execution combined with the unique size advantages of microfluidics provides nearly endless possibilities for its use as a chemical biology tool. With the advancement of nanofluidic technologies, the combined capabilities of microfluidics and nanofluidics could provide the necessary framework for important biological discoveries using chemical biology tools.

Chapter- 2

Proteomics

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

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

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

Complexity of the problem

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

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

Post-translational modifications

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

Phosphorylation

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

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

Ubiquitination

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

Additional modifications

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

Distinct proteins are made under distinct settings

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

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

Limitations to genomic study

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

Methods of studying proteins

Determining proteins which are post-translationally modified

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

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

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

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

Determining the existence of proteins in complex mixtures

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

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

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

Computational methods in studying protein biomarkers

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

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

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

Establishing protein-protein interactions

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

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

Practical applications of proteomics

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

Biomarkers

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

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

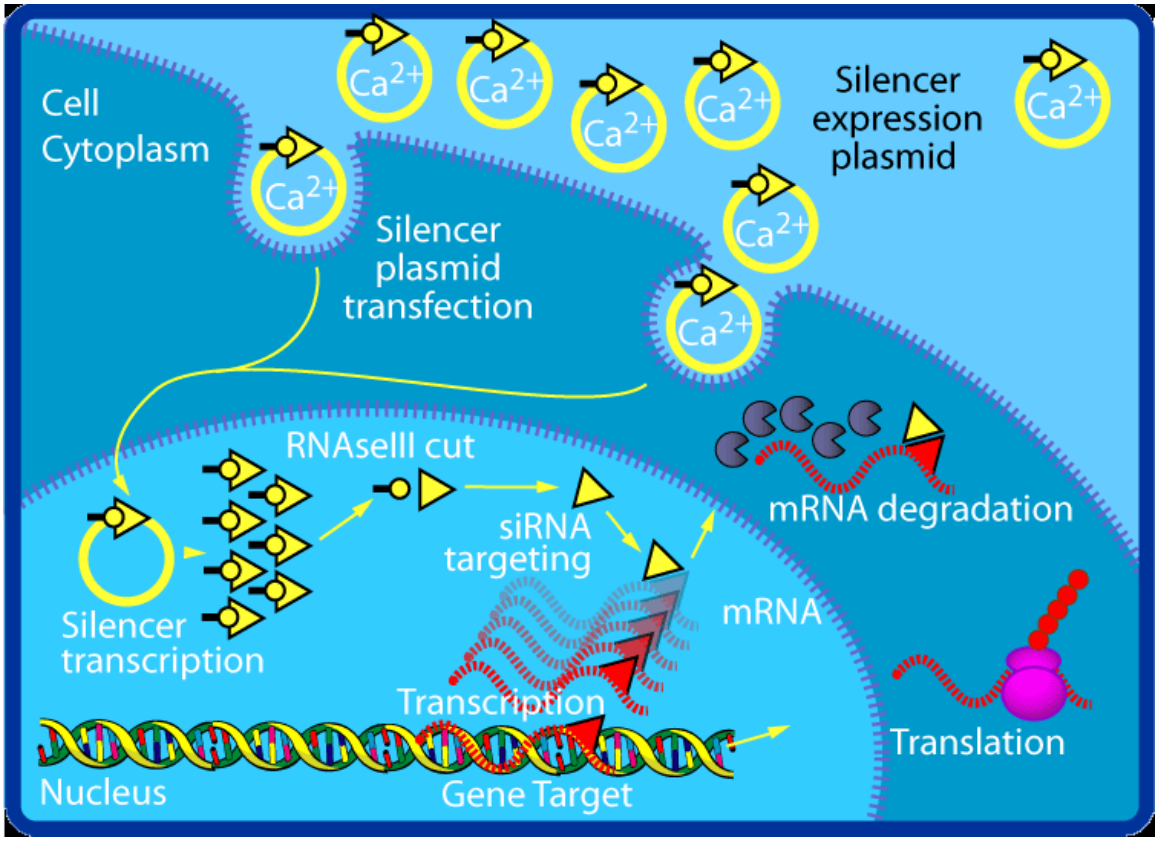
An interesting use of proteomics is using specific protein biomarkers to diagnose disease. A number of techniques allow to test for proteins produced during a particular disease, which helps to diagnose the disease quickly. Techniques include western blot, immunohistochemical staining, enzyme linked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

Current research methodologies

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

Chapter- 3

Small Interfering RNA



Mediating RNA interference in cultured mammalian cells.

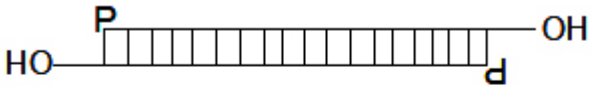
Small interfering RNA (**siRNA**), sometimes known as **short interfering RNA** or **silencing RNA**, is a class of double-stranded RNA molecules, 20-25 nucleotides in length, that play a variety of roles in biology. Most notably, siRNA is involved in the RNA interference (RNAi) pathway, where it interferes with the expression of a specific gene. In addition to their role in the RNAi pathway, siRNAs also act in RNAi-related

pathways, e.g., as an antiviral mechanism or in shaping the chromatin structure of a genome; the complexity of these pathways is only now being elucidated.

siRNAs were first discovered by David Baulcombe's group at the Sainsbury Laboratory in Norwich, England, as part of post-transcriptional gene silencing (PTGS) in plants. The group published their findings in *Science* in a paper titled "A species of small antisense RNA in posttranscriptional gene silencing in plants". Shortly thereafter, in 2001, synthetic siRNAs were shown to be able to induce RNAi in mammalian cells by Thomas Tuschl, and colleagues in a paper published in *Nature*. This discovery led to a surge in interest in harnessing RNAi for biomedical research and drug development.

Structure

siRNAs have a well-defined structure: a short (usually 21-nt) double strand RNA (dsRNA) with 2-nt 3' overhangs on either end:



Schematic representation of a siRNA molecule: a ~19-21basepair RNA core duplex that is followed by a 2 nucleotide 3' overhang on each strand. OH: 3' hydroxyl; P: 5' phosphate.

Each strand has a 5' phosphate group and a 3' hydroxyl (-OH) group. This structure is the result of processing by dicer, an enzyme that converts either long dsRNAs or small hairpin RNAs into siRNAs. siRNAs can also be exogenously (artificially) introduced into cells by various transfection methods to bring about the specific knockdown of a gene of interest. Essentially any gene of which the sequence is known can thus be targeted based on sequence complementarity with an appropriately tailored siRNA. This has made siRNAs an important tool for gene function and drug target validation studies in the post-genomic era.

RNAi induction using siRNAs or their biosynthetic precursors

Dicer protein colored by protein domain.

Transfection of an exogenous siRNA can be problematic because the gene knockdown effect is only transient, particularly in rapidly dividing cells. One way of overcoming this challenge is to modify the siRNA in such a way as to allow it to be expressed by an appropriate vector, e.g., a plasmid. This is done by the introduction of a loop between the two strands, thus producing a single transcript, which can be processed into a functional siRNA. Such transcription cassettes typically use an RNA polymerase III promoter (e.g., U6 or H1), which usually directs the transcription of small nuclear RNAs (snRNAs) (U6 is involved in gene splicing; H1 is the RNase component of human RNase P). It is assumed (although not known for certain) that the resulting siRNA transcript is then processed by Dicer.

RNA activation

It has recently been found that dsRNA can also activate gene expression, a mechanism that has been termed "small RNA-induced gene activation" or RNAa. It has been shown that dsRNAs targeting gene promoters induce potent transcriptional activation of associated genes. RNAa was demonstrated in human cells using synthetic dsRNAs, termed "small activating RNAs" (saRNAs). It is currently not known whether RNAa is conserved in other organisms.

Challenges: avoiding nonspecific effects

Because RNAi intersects with a number of other pathways, it is not surprising that on occasion nonspecific effects are triggered by the experimental introduction of an siRNA. When a mammalian cell encounters a double-stranded RNA such as an siRNA, it may mistake it as a viral by-product and mount an immune response. Furthermore, because structurally related microRNAs modulate gene expression largely via incomplete complementarity base pair interactions with a target mRNA, the introduction of an siRNA may cause unintended off-targeting.

Innate immunity

Introduction of too much siRNA can result in nonspecific events due to activation of innate immune responses. Most evidence to date suggests that this is probably due to activation of the dsRNA sensor PKR, although retinoic acid inducible gene I (RIG-I) may also be involved. The induction of cytokines via toll-like receptor 7 (TLR7) has also been described. One promising method of reducing the nonspecific effects is to convert the siRNA into a microRNA. MicroRNAs occur naturally, and by harnessing this endogenous pathway it should be possible to achieve similar gene knockdown at comparatively low concentrations of resulting siRNAs. This should minimize nonspecific effects.

Off-targeting

Off-targeting is another challenge to the use of siRNAs as a gene knockdown tool. Here, genes with incomplete complementarity are inadvertently downregulated by the siRNA (effectively, the siRNA acts as a miRNA), leading to problems in data interpretation and potential toxicity. This, however, can be partly addressed by designing appropriate control experiments, and siRNA design algorithms are currently being developed to produce siRNAs free from off-targeting. Genome-wide expression analysis, e.g., by microarray technology, can then be used to verify this and further refine the algorithms. A 2006 paper from the laboratory of Dr. Khvorova implicates 6- or 7-basepair-long stretches from position 2 onward in the siRNA matching with 3'UTR regions in off-targeted genes.

Possible therapeutic applications and challenges

Given the ability to knock down essentially any gene of interest, RNAi via siRNAs has generated a great deal of interest in both basic and applied biology. There are an increasing number of large-scale RNAi screens that are designed to identify the important genes in various biological pathways. Because disease processes also depend on the activity of multiple genes, it is expected that in some situations turning off the activity of a gene with an siRNA could produce a therapeutic benefit.

However, applying RNAi via siRNAs to living animals, especially humans, poses many challenges. Experimentally, siRNAs show different effectiveness in different cell types in a manner as yet poorly understood: some cells respond well to siRNAs and show a robust knockdown, whereas others show no such knockdown (even despite efficient transfection).

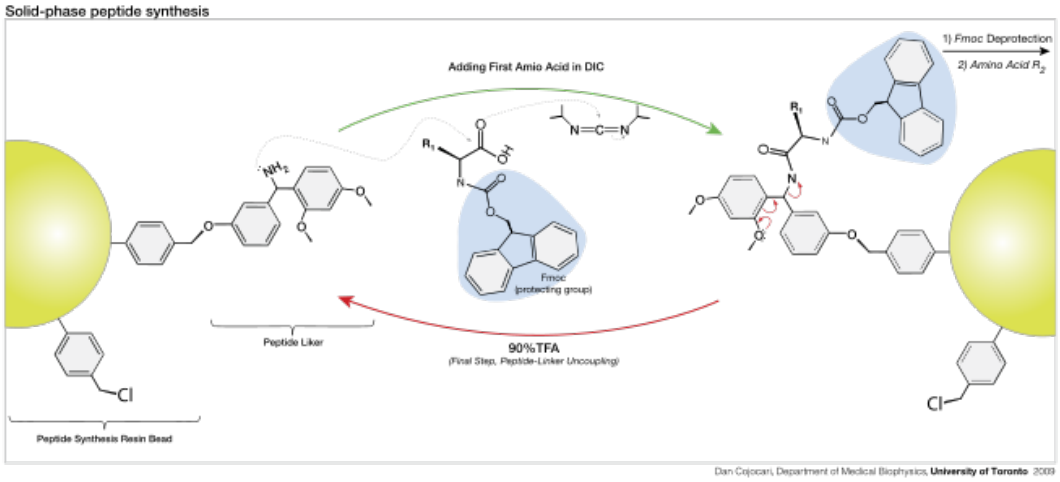
Phase I results of the first two therapeutic RNAi trials (indicated for age-related macular degeneration, aka AMD) reported at the end of 2005 that siRNAs are well tolerated and have suitable pharmacokinetic properties.

Proof of concept trials have indicated that Ebola-targeted siRNAs may be effective as post-exposure prophylaxis in humans, with 100% of non-human primates surviviing a lethal dose of Zaire Ebolavirus, the most lethal strain.

Chapter- 4

Peptide Synthesis

In organic chemistry, **peptide synthesis** is the production of peptides, which are organic compounds in which multiple amino acids are linked via amide bonds which are also known as peptide bonds. The biological process of producing long peptides (proteins) is known as protein biosynthesis.



Solid-Phase Peptide Syntesis on a Rink amide resin using Fmoc-α-amine-protected amino acid

Chemistry

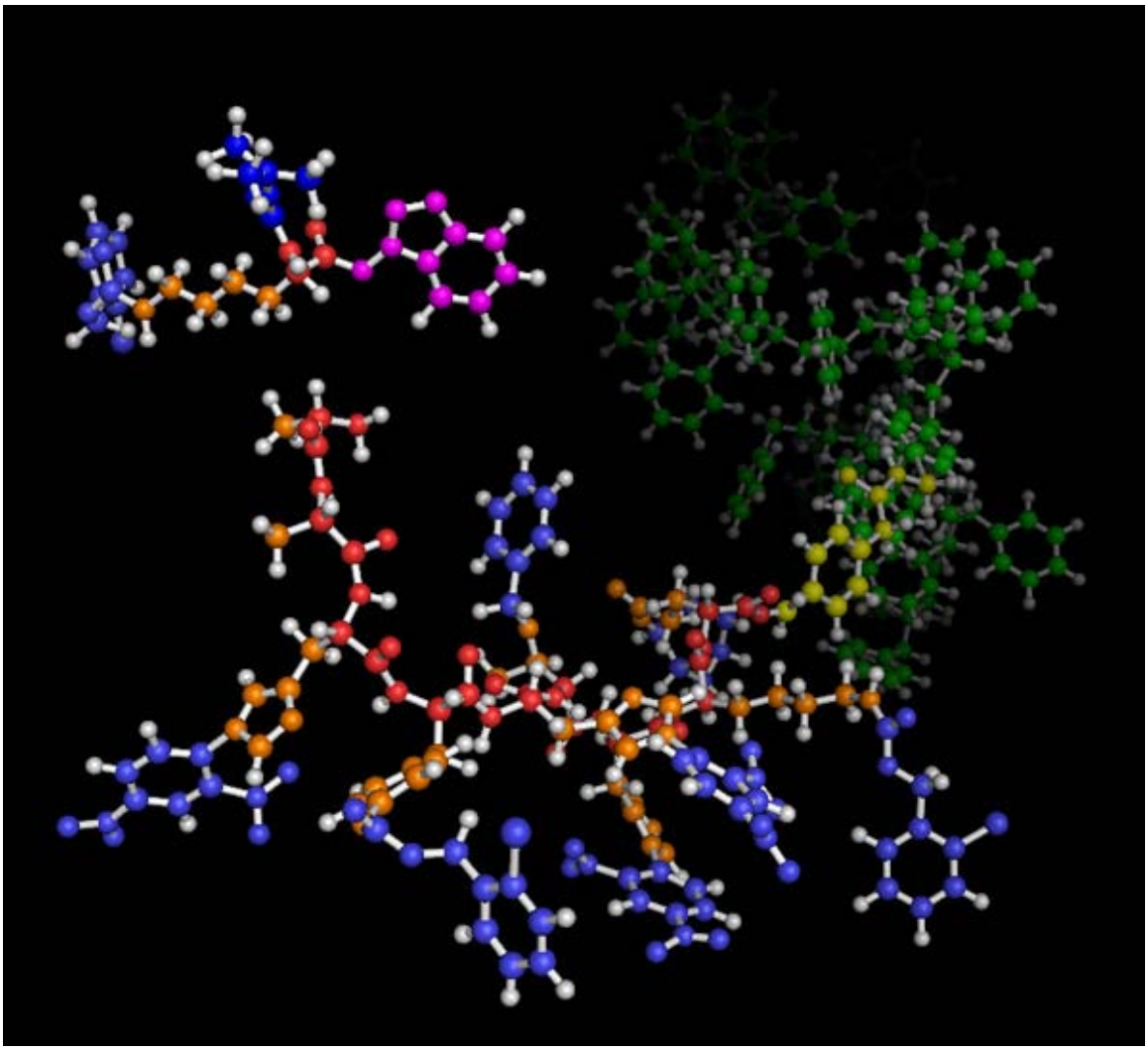
Peptides are synthesized by coupling the carboxyl group or C-terminus of one amino acid to the amino group or N-terminus of another. Due to the possibility of unintended reactions, protecting groups are usually necessary. Chemical peptide synthesis starts at

the C-terminal end of the peptide and ends at the N-terminus. This is the opposite of protein biosynthesis, which starts at the N-terminal end.

Liquid-phase synthesis

Liquid-phase peptide synthesis is a classical approach to peptide synthesis. It has been replaced in most labs by solid-phase synthesis. However, it retains usefulness in large-scale production of peptides for industrial purposes.

Solid-phase synthesis



Coupling step in solid-phase peptide synthesis

Solid-phase peptide synthesis (SPPS), pioneered by Robert Bruce Merrifield, resulted in a paradigm shift within the peptide synthesis community. It is now the accepted method for creating peptides and proteins in the lab in a synthetic manner. SPPS allows the synthesis of natural peptides which are difficult to express in bacteria, the incorporation

of unnatural amino acids, peptide/protein backbone modification, and the synthesis of D-proteins, which consist of D-amino acids.

Small solid beads, insoluble yet porous, are treated with functional units ('linkers') on which peptide chains can be built. The peptide will remain covalently attached to the bead until cleaved from it by a reagent such as anhydrous hydrogen fluoride or trifluoroacetic acid. The peptide is thus 'immobilized' on the solid-phase and can be retained during a filtration process, whereas liquid-phase reagents and by-products of synthesis are flushed away.

The general principle of SPPS is one of repeated cycles of coupling-wash-deprotection-wash. The free N-terminal amine of a solid-phase attached peptide is coupled to a single N-protected amino acid unit. This unit is then deprotected, revealing a new N-terminal amine to which a further amino acid may be attached. The superiority of this technique partially lies in the ability to perform wash cycles after each reaction, removing excess reagent with all of the growing peptide of interest remaining covalently attached to the insoluble resin.

The overwhelmingly important consideration is to generate extremely high yield in each step. For example, if each coupling step were to have 99% yield, a 26-amino acid peptide would be synthesized in 77% final yield (assuming 100% yield in each deprotection); if each step were 95%, it would be synthesized in 25% yield. Thus each amino acid is added in major excess (2~10x) and coupling amino acids together is highly optimized by a series of well-characterized agents.

There are two majorly used forms of SPPS -- **Fmoc** and **Boc**. Unlike ribosome protein synthesis, solid-phase peptide synthesis proceeds in a C-terminal to N-terminal fashion. The N-termini of amino acid monomers is protected by these two groups and added onto a deprotected amino acid chain.

Automated synthesizers are available for both techniques, though many research groups continue to perform SPPS manually.

SPPS is limited by yields, and typically peptides and proteins in the range of 70 amino acids are pushing the limits of synthetic accessibility. Synthetic difficulty also is sequence dependent; typically amyloid peptides and proteins are difficult to make. Longer lengths can be accessed by using native chemical ligation to couple two peptides together with quantitative yields.

Since its introduction over 40 years ago, SPPS has been significantly optimized. First, the resins themselves have been optimized. Furthermore, the ‘linkers’ between the C-terminal amino acid and polystyrene resin have improved attachment and cleavage to the point of mostly quantitative yields. The evolution of side chain protecting groups has limited the frequency of unwanted side reactions. In addition, the evolution of new activating groups on the carboxyl group of the incoming amino acid have improved coupling and decreased epimerization. Finally, the process itself has been optimized. In

Merrifield's initial report, the deprotection of the α-amino group resulted in the formation of a peptide-resin salt, which required neutralization with base prior to coupling. The time between neutralization of the amino group and coupling of the next amino acid allowed for aggregation of peptides, primarily through the formation of secondary structures, and adversely affected coupling. The Kent group showed that concomitant neutralization of the α-amino group and coupling of the next amino acid led to improved coupling. Each of these improvements has helped SPPS become the robust technique that it is today.

Fmoc solid-phase peptide synthesis

The capacity for anhydrous hydrogen fluoride to degrade proteins during the final cleavage conditions led to a new α-amino protecting group based on 9-fluorenylmethyloxycarbonyl (Fmoc). The Fmoc method allows for a milder deprotection scheme. This method utilizes a base, usually piperidine (20-50%) in DMF in order to remove the Fmoc group to expose the α-amino group for reaction with an incoming activated amino acid. Unlike the acid used to deprotect the α-amino group in Boc methods, Fmoc SPPS uses a base, and thus the exposed amine is neutral. Therefore, no neutralization of the peptide-resin is required, but the lack of electrostatic repulsions between the peptides can lead to increased aggregation.

Along with the development of Fmoc SPPS, different resins have also been created to be removed by TFA. Similar to the Boc strategy, two primary resins are used, based on whether a C-terminal carboxylic acid or amide is desired. The Wang resin is the most commonly used resin for peptides with C-terminal carboxylic acids. If a C-terminal amide is desired, the Rink amide resin is used.

Semipermanent protecting groups are t-butyl based, and final cleavage of the protein from the resin and removal of permanent protecting groups is performed with TFA in the presence of scavengers. Water and triisopropylsilane (TIPS) present in a 1:1 ratio are often used as scavengers. Thus, the Fmoc method is orthogonal in two directions: deprotection of the α-amino group and final cleavage from the resin occur by independent mechanisms. The resulting final product is a TFA salt, which is more difficult to solubilize than the fluoride salts generated in Boc SPPS. This method is thus milder than the Boc method because the deprotection/cleavage-from-resin steps occur with different conditions rather than with different reaction rates.

t-Boc solid-phase peptide synthesis

The original method for the synthesis of proteins relied on tert-butoxycarbonyl (Boc) to temporarily protect the α-amino group. In this method, the Boc group is covalently bound to the amino group to suppress its nucleophilicity. The C-terminal amino acid covalently linked to the resin through a linker. Next, the Boc group is removed with acid, such as trifluoroacetic acid (TFA). This forms a positively-charged amino group, which is simultaneously neutralized and coupled to the incoming activated amino acid. Reactions are driven to completion by the use of excess (two- to fourfold) activated amino acid.

After each deprotection and coupling step, a wash with N,N-dimethylformamide (DMF) is performed to remove excess reagents, allowing for high yields (~99%) during each cycle.

Importantly, improvements to the resin have enhanced its ability to withstand the repeated use of TFA during the deprotection step. Furthermore, different resins allow for different functional groups at the C-terminus. The oxymethylphenylacetamidomethyl (PAM) resin results in the conventional C-terminal carboxylic acid. On the other hand, the paramethylbenzhydrylamine (pMBHA) resin yields a C-terminal amide, which is useful in mimicking the interior of a protein.

Permanent side chain protecting groups are typically benzyl or benzyl-based groups. Final removal of the peptide from the linkage occurs simultaneously with side-chain deprotection with anhydrous hydrogen fluoride via hydrolytic cleavage. The final product is a fluoride salt which is relatively easy to solubilize. Importantly, scavengers such as cresol are added to the HF in order to prevent reactive t-butyl cations from generating undesired products. In fact, the use of harsh hydrogen fluoride may degrade some peptides, which was the premise for the development of a milder, base-labile method of SPPS—namely, the Fmoc method.

Some researchers prefer Boc SPPS for complex syntheses . In addition, when synthesizing nonnatural peptide analogs which are base-sensitive (such as depsipeptides), the t-Boc protecting group is necessary. This is because Fmoc SPPS uses base to protect the α-amino group.

Comparison of Boc and Fmoc Solid-Phase Peptide Synthesis

Both the Fmoc and Boc methods offer advantages and disadvantages. The selection of one technique over another is thus made on a case-by-case basis.

	Boc	**Fmoc**
Requires special equipment	Yes	No
Cost of reagents	Lower	Higher
Solubility of peptides	Higher	Lower
Purity of hydrophobic peptides	High	May be lower
Problems with aggregation	Less frequently	More frequently
Synthesis time	~20 min/amino acid	~20-60 min/amino acid
Final deprotection	HF	TFA
Safety	Potentially dangerous	Relatively safe
Orthogonal	No	Yes

Boc SPPS uses special equipment to handle the final cleavage and deprotection step, which requires anhydrous hydrogen fluoride. Because the final cleavage of the peptide with Fmoc SPPS uses TFA, this special equipment is not necessary. The solubility of

peptides generated by Boc SPPS is generally higher than those generated with the Fmoc method, because fluoride salts are higher in solubility than TFA salts. Next, problems with aggregation are generally more of an issue with Fmoc SPPS. This is primarily because the removal of a Boc group with TFA yields a positively-charged α-amino group, whereas the removal of an Fmoc group yields a neutral α-amino group. The steric hindrance of the positively charged α-amino group limits the formation of secondary structure on the resin. Finally, the Fmoc method is considered orthogonal, since α-amino group deprotection is with base, while final cleavage from the resin is with acid. The Boc method utilizes acid for both deprotection and cleavage from the resin. Based on this comparison, one sees that both methods possess advantages and disadvantages. Thus, several factors help to decide which method may be preferable.

Solid supports

The name solid support implies that reactions are carried out on the surface of the support, but this is not the case. Reactions also occur within these particles, and thus the term "solid support" better describes the insolubility of the polymer. The physical properties of the solid support, and the applications to which it can be utilized, vary with the material from which the support is constructed, the amount of cross-linking, as well as the linker and handle being used. Most scientists in the field believe that supports should have the minimum amount of cross-linking to confer stability. This should result in a well-solvated system where solid-phase peptide synthesis can be carried out. Nonetheless, the characteristics of an efficient solid support include:

1. It must be physically stable and permit the rapid filtration of liquids, such as excess reagents
2. It must be inert to all reagents and solvents used during SPPS
3. It must swell extensively in the solvents used to allow for penetration of the reagents
4. It must allow for the attachment of the first amino acid

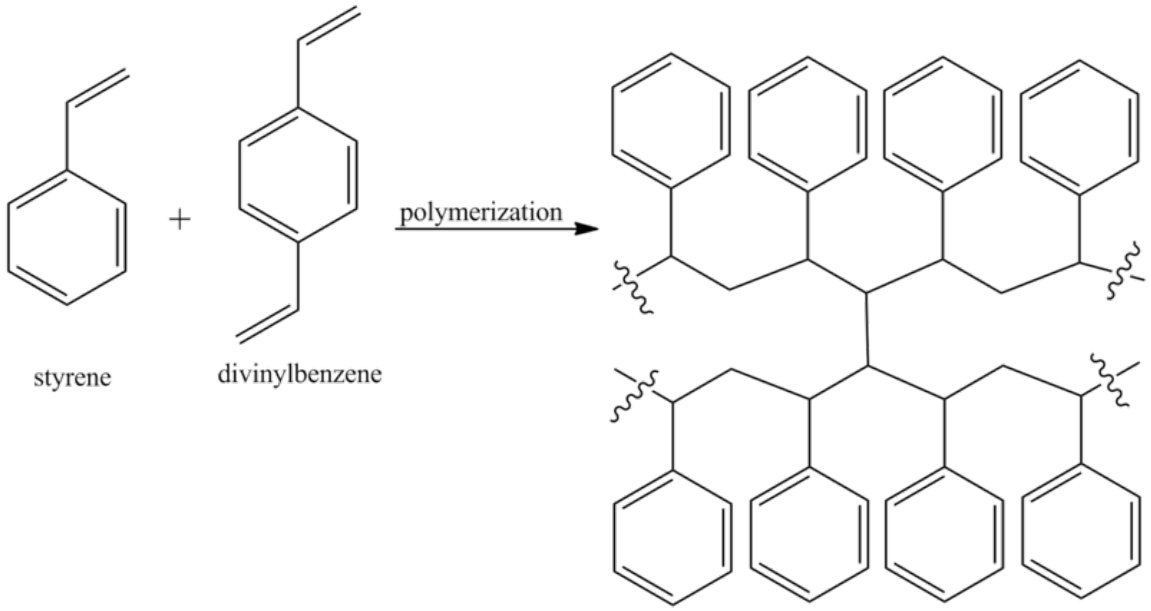
There are four primary types of solid supports:

1. Gel-type supports: These are highly solvated polymers with an equal distribution of functional groups. This type of support is the most common, and includes:
    - Polystyrene: Styrene cross-linked with 1-2% divinylbenzene
    - Polyacrylamide: A hydro**philic** alternative to polystyrene
    - Polyethylene glycol (PEG): PEG-Polystyrene (PEG-PS) is more stable than polystyrene and spaces the site of synthesis from the polymer backbone
    - PEG-based supports: Composed of a PEG-polypropylene glycol network or PEG with polyamide or polystyrene
2. Surface-type supports: Many materials have been developed for surface functionalization, including controlled pore glass, cellulose fibers, and highly cross-linked polystyrene.
3. Composites: Gel-type polymers supported by rigid matrices.

Polystyrene resin

Polystyrene resin is a versatile resin and it is quite useful in multi-well, automated peptide synthesis, due to its minimal swelling in dichloromethane. The initial support used by R. Bruce Merrifield was polysytrene cross-linked with 2% divinylbenzene. This support is sometimes referred to as the 'Merrifield resin.' This produces a hydrophobic bead that is solvated by nonpolar solvent such as dichloromethane. Since then, new resins have been developed with the following advantages:

1. Enhanced swelling or rigidity (a property of mechanical strength)
2. Chemical inertness



polystyrene cross-linked with divinylbenze

Polystyrene cross-linked with divinylbenzene. This is the most common solid support used in SPPS, and was the support pioneered by R. Bruce Merrifield.

Highly cross-linked (50%) polystyrene has been developed that possesses the features of increased mechanical stability, better filtration of reagents and solvents, and rapid reaction kinetics.

Polyamide resin

Polyamide resin is also a useful and versatile resin. It seems to swell much more than polystyrene, in which case it may not be suitable for some automated synthesizers, if the wells are too small.

PEG hybride polystyrene resin

An example of this type of resin is the Tentagel resin. The base resin is polystyrene onto which is attached long chains (Mw ca. 3000 Da) of polyethylene glycol (PEG; also known as polyethylene oxide). Synthesis is caried out on the distal end of the PEG spacer making it suited for long and difficult peptides. In addition it is also attractive for the synthesis of combinatorial peptide libraries and on resin screening experiments. It does not expand much during synthesis making it a preferred resin for robotic peptide synthesis.

PEG based resin

ChemMatrix(R) is a new type of resin which is based on PEG that is crosslinked. ChemMatrix(R) has claimed a high chemical and thermal stability (is compatible with Microwave synthesis) and has shown higher degrees of swellings in acetonitrile, dichloromethane, DMF, N-methylpyrrolidone, TFA and water compared to the polystyrene based resins. ChemMatrix has shown significant improvements to the synthesis of hydrophobic sequences. ChemMatrix is recommended for the synthesis of difficult and long peptides.

Protecting groups

Due to amino acid excesses used to ensure complete coupling during each synthesis step, polymerization of amino acids is common in reactions where each amino acid is not protected. In order to prevent this polymerization, protecting groups are used. This adds additional deprotection phases to the synthesis reaction, creating a repeating design flow as follows:

- Protecting group is removed from trailing amino acids in a deprotection reaction
- Deprotection reagents washed away to provide clean coupling environment
- Protected amino acids dissolved in a solvent such as dimethylformamide (DMF) are combined with coupling reagents are pumped through the synthesis column
- Coupling reagents washed away to provide clean deprotection environment

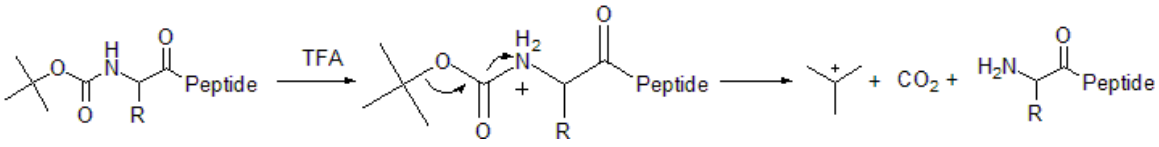
Currently, two protecting groups (t-Boc, Fmoc) are commonly used in solid-phase peptide synthesis. Their lability is caused by the carbamate group which readily releases CO_2 for an irreversible decoupling step.

DMF must be 'peptide grade' i.e. little/no impurities and must also be 'fresh'. This is due to the fact that DMF undergoes photolysis to form carbon monoxide and dimethylamine. Dimethylamine may remove the Fmoc group and, therefore, lead to impurities.

t-Boc protecting group

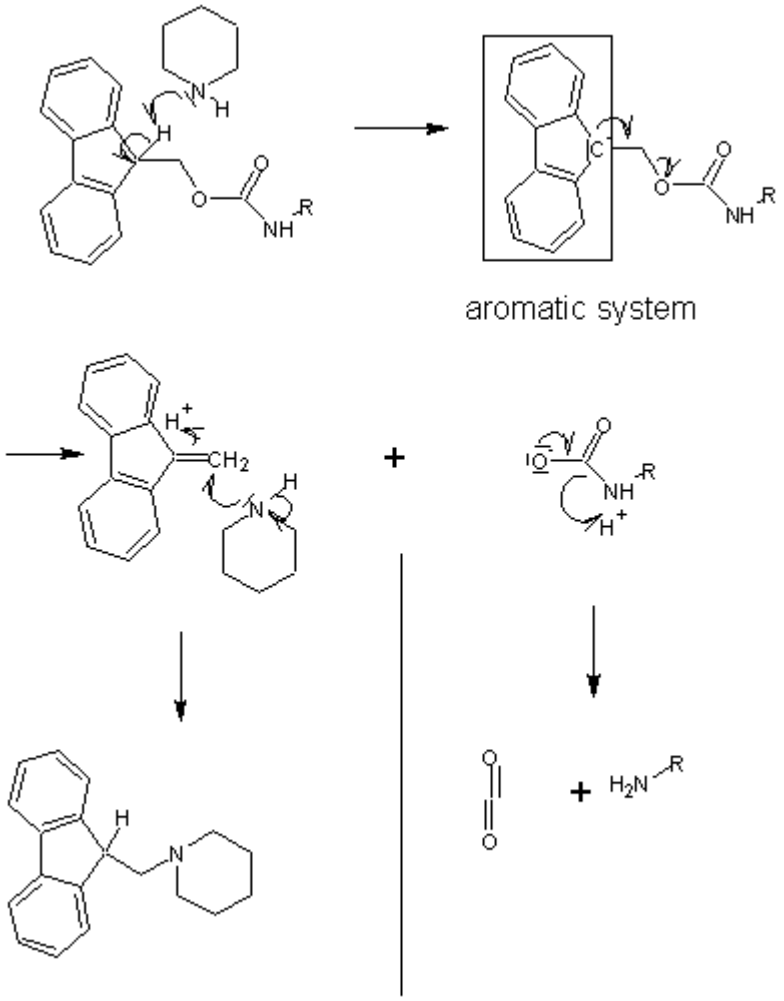
The t-Boc (tert-butyloxycarbonyl or more simply "Boc") group was commonly used for protecting the terminal amine of the peptide, requiring the use of more acid stable groups

for side chain protection in orthogonal strategies. It retains usefulness in reducing aggregation of peptides during synthesis. t-Boc groups can be added to amino acids with t-Boc anhydride and a suitable base.



Boc cleavage

Fmoc protecting group



Fmoc cleavage

Fmoc (9*H*-fluoren-9-ylmethoxycarbonyl) is currently a widely used protective group that is generally removed from the N terminus of a peptide in the iterative synthesis of a peptide from amino acid units. The advantage of Fmoc is that it is cleaved under very mild basic conditions (e.g. piperidine), but stable under acidic conditions, although this has not always held true in certain synthetic sequences. This allows mild acid labile

protecting groups that are stable under basic conditions, such as Boc and benzyl groups, to be used on the side-chains of amino acid residues of the target peptide. This orthogonal protecting group strategy is common in the art of organic synthesis.

FMOC is preferred over BOC due to ease of cleavage; however it is less atom-economical, as the fluorenyl group is much larger than the tert-butyl group. Accordingly, prices for FMOC amino acids were high until the large-scale piloting of one of the first synthesized peptide drugs, enfuvirtide, began in the 1990s, when market demand adjusted the relative prices of the two sets of amino acids.

Because the liberated Fluorenyl group is a chromophore, deprotection by FMOC can be monitored by UV absorbance of the runoff, a strategy which is employed in automated synthesizers.

Benzyloxy-carbonyl (Z) group

The first use of (Z) group as protecting groups was done by Max Bergmann who synthesised oligopeptides.

Another carbamate based group is the benzyloxy-carbonyl (Z) group. It is removed in harsher conditions: HBr/acetic acid or catalytic hydrogenation. Today it is almost exclusively used for side chain protection.

Alloc protecting group

The allyloxycarbonyl (alloc) protecting group is often used to protect a carboxylic acid, hydroxyl, or amino group when an orthogonal deprotection scheme is required. It is sometimes used when conducting on-resin cyclic peptide formation, where the peptide is linked to the resin by a side-chain functional group. The alloc group can be removed using tetrakis(triphenylphosphine)palladium(0) along with a 37:2:1 mixture of methylene chloride, acetic acid, and N-Methylmorpholine (NMM) for 2 hours. The resin must then be carefully washed 0.5% DIPEA in DMF, 3x10 ml of 0.5% sodium diethylthiocarbamate in DMF, and then 5x10 ml of 1:1 DCM:DMF.

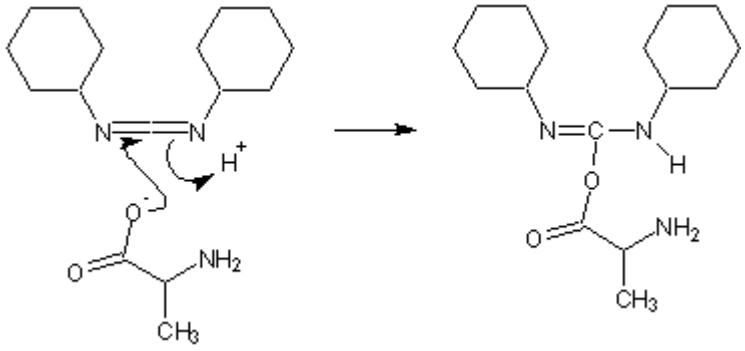
Lithographic protecting groups

For special applications like protein microarrays lithographic protecting groups are used. Those groups can be removed through exposure to light.

Activating groups

For coupling the peptides the carboxyl group is usually activated. This is important for speeding up the reaction. There are two main types of activating groups: carbodiimides and triazolols. However the use of pentafluorophenyl esters (FDPP, PFPOH) and BOP-Cl are useful for cyclising peptides.

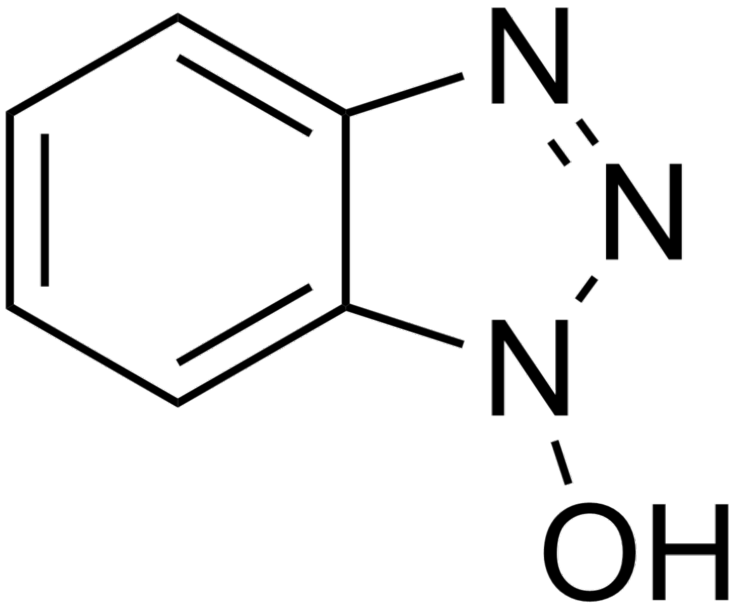
Carbodiimides



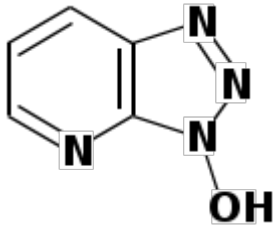
Alanine attaching to DCC

These activating agents were first developed. Most common are dicyclohexylcarbodiimide (DCC) and diisopropylcarbodiimide (DIC). Reaction with a carboxylic acid yields a highly reactive O-acyl-urea. During artificial protein synthesis (such as Fmoc solid-state synthesizers), the C-terminus is often used as the attachment site on which the amino acid monomers are added. To enhance the electrophilicity of carboxylate group, the negatively charged oxygen must first be "activated" into a better leaving group. DCC is used for this purpose. The negatively charged oxygen will act as a nucleophile, attacking the central carbon in DCC. DCC is temporarily attached to the former carboxylate group (which is now an ester group), making nucleophilic attack by an amino group (on the attaching amino acid) to the former C-terminus (carbonyl group) more efficient. The problem with carbodiamides is that they are too reactive and that they can therefore cause racemization of the amino acid.

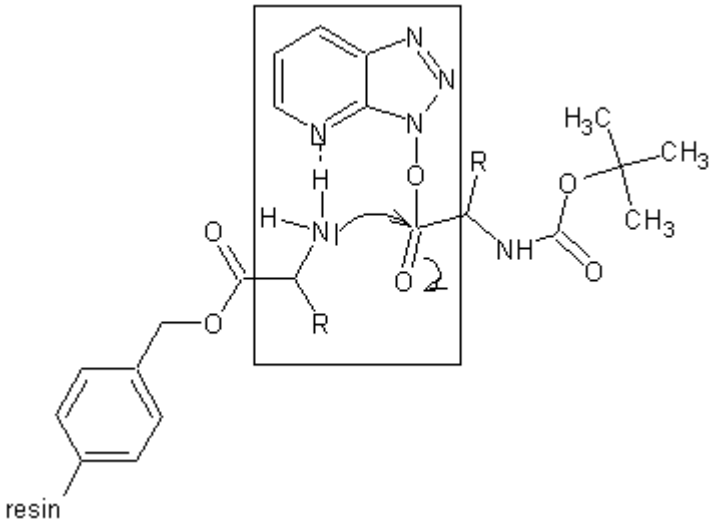
Triazoles



HOBt



HOAt



Neighbouring group effect of HOAt

To solve the problem of racemization, triazoles were introduced. The most important ones are 1-hydroxy-benzotriazole (HOBt) and 1-hydroxy-7-aza-benzotriazole (HOAt). Others have been developed. These substances can react with the O-acylurea to form an active ester which is less reactive and less in danger of racemization. HOAt is especially favourable because of a neighbouring group effect. Recently, HOBt has been removed from many chemical vendor catalogues; although almost always found as a hydrate, HOBt may be explosive when allowed to fully dehydrate and shipment by air or sea is heavily restricted. Alternatives to HOBt and HOAt has been introduced. One of the most promising and inexpensive is ethyl 2-cyano-2-(hydroxyimino)acetate (trade name Oxyma Pure), which is not explosive and has a reactivity of that in between HOBt and HOAt.

Newer developments omit the carbodiimides totally. The active ester is introduced as a uronium or phosphonium salt of a non-nucleophilic anion (tetrafluoroborate or hexafluorophosphate): HBTU, HATU, HCTU, TBTU, PyBOP. Two uronium types of the coupling additive of Oxyma Pure is also available as COMU or TOTU reagent.

Regioselective Disulfide Formation

The formation of multiple native disulfides remains one of the primary challenges of native peptide synthesis by solid-phase methods. Random chain combination typically results in several products with nonnative disulfide bonds. Stepwise formation of disulfide bonds is typically the preferred method, and performed with thiol protecting groups (PGs). Different thiol PGs provide multiple dimensions of orthogonal protection. These orthogonally-protected cysteines are incorporated during the solid-phase synthesis of the peptide. Successive removal of these PGs to allow for selective exposure of free thiol groups, leads to disulfide formation in a stepwise manner. The order of removal of these PGs must be considered so that only one group is removed at a time. Using this method, Kiso *et al.* reported the first total synthesis of insulin by this method in 1993.

The thiol PGs must possess multiple characteristics. First, the PG must be reversible with conditions that do not affect the unprotected side chains. Second, the protecting group must be able to withstand the conditions of solid-phase synthesis. Third, the configuration of the removal of the thiol protecting group must be such that it leaves intact other thiol PGs, if orthogonal protection is desired. That is, the removal of PG A should not affect PG B. Some of the thiol PGs commonly used include the acetamidomethyl (Acm), tert-butyl (But), 3-nitro-2-pyridine sulfenyl (NPYS), 2-pyridine-sulfenyl (Pyr), and triphenylmethyl (Trt) groups. Importantly, the NPYS group can replace the Acm PG to yield an activated thiol.

In the stepwise formation of disulfides to synthesize insulin by Kiso *et al.*, the authors synthesize the A-chain with following protection: CysA6(But); CysA7(Acm); CysA11(But). Thus, CysA20 is unprotected. Synthesis of the B-chain is performed with the following protection: CysB7(Acm) CysB19(Pyr). The first disulfide bond, CysA20-CysB19, was formed by mixing the two chains in 8 M urea, pH 8 (RT) for 50 min. The second disulfide bond, CysA7-CysB7, was formed by treatment with iodine in aqueous acetic acid to remove the Acm groups. The third disulfide, the intramolecular CysA6-CysA11, was formed by the removal of the But groups by methyltrichlorosilane with diphenyl sulfoxide in TFA. Importantly, formation of the first disulfide in 8 M urea, pH 8 does not affect the other PGs, namely Acm and But groups. Likewise, formation of the second disulfide bond with iodine in aqueous acetic acid does not affect the But groups.

Important to the discussion of disulfide bond formation is the order in which disulfides are formed. From a logical standpoint, the order in which the thiol groups are exposed to form disulfides should be of little consequence, since the other cysteines are protected. Practically, however, the order in which disulfides are formed can have a significant effect on yields. This may be because the formation of the CysA20-CysB19 disulfide may place the thiol group of CysB7 in close proximity with both CysA6 and CysA7, leading to multiple disulfide products. This is one manifestation of the reality that solid-phase peptide synthesis is as much art as it is science.

Synthesizing long peptides

Stepwise elongation, in which the amino acids are connected step-by-step in turn, is ideal for small peptides containing between 2 and 100 amino acid residues. Another method is fragment condensation, in which peptide fragments are coupled. Although the former can elongate the peptide chain without racemization, the yield drops if only it is used in the creation of long or highly polar peptides. Fragment condensation is better than stepwise elongation for synthesizing sophisticated long peptides, but its use must be restricted in order to protect against racemization. Fragment condensation is also undesirable since the coupled fragment must be in gross excess, which may be a limitation depending on the length of the fragment.

A new development for producing longer peptide chains is chemical ligation: Unprotected peptide chains react chemoselectively in aqueous solution. A first kinetically controlled product rearranges to form the amide bond. The most common form of native chemical ligation uses a peptide thioester that reacts with a terminal cysteine residue.

Coupling Efficiency Vs. Peptide Length

Peptide

Length	Coupling Efficiency	Coupling Efficiency	Coupling Efficiency	Coupling Efficiency	Coupling Efficiency
0	0.995	0.99	0.98	0.97	0.96
5	0.98	0.95	0.92	0.89	0.85
10	0.96	0.91	0.83	0.76	0.69
15	0.93	0.87	0.75	0.65	0.56
20	0.91	0.83	0.68	0.56	0.46
25	0.89	0.79	0.62	0.48	0.38
30	0.86	0.75	0.56	0.41	0.31
35	0.84	0.71	0.50	0.36	0.25
40	0.82	0.67	0.45	0.30	0.20
45	0.80	0.63	0.41	0.26	0.17
50	0.78	0.60	0.37	0.22	0.14
55	0.76	0.58	0.34	0.19	0.11
60	0.74	0.55	0.30	0.17	0.09
65	0.73	0.53	0.27	0.14	0.07
70	0.71	0.50	0.25	0.12	0.06

Microwave assisted peptide synthesis

Although microwave irradiation has been around since the late 1940s, it was not until 1986 that microwave energy was used in organic chemistry. During the end of the 1980s and 1990s, microwave energy was an obvious source for completing chemical reactions in minutes that would otherwise take several hours to days. Through several technical improvements at the end of the 1990s and beginning of the 2000s, microwave synthesizers have been designed to provide both low and high energy pockets of microwave energy so that the temperature of the reaction mixture could be controlled. The microwave energy used in peptide synthesis is of a single frequency providing maximum penetration depth of the sample which is in contrast to conventional kitchen microwaves.

In peptide synthesis, microwave irradiation has been used to complete long peptide sequences with high degrees of yield and low degrees of racemization. Microwave irradiation during the coupling of amino acids to a growing polypeptide chain is not only catalyzed through the increase in temperature, but also due to the alternating electromagnetic radiation to which the polar backbone of the polypeptide continuously aligns to. Due to this phenomenon, the microwave energy can prevent aggregation and thus increases yields of the final peptide product. There is however no clear evidence that microwave is better than simple heating and some peptide laboratories regard microwave just as a convenient method for rapid heating of the peptidyl resin. Heating to above 50-55 degrees Celsius also prevents aggregation and accelerates the coupling.

Despite the main advantages of microwave irradiation of peptide synthesis, the main disadvantage is the racemization which may occur with the coupling of cysteine and histidine. A typical coupling reaction with these amino acids are performed at lower temperatures than the other 18 natural amino acids. A number of peptides does not survive microwave synthesis or heating in general. One of the more serious side effects is dehydration (loss of water) which for certain peptides can be almost quantitative like pancreatic polypeptide (PP). This side effect is also seen by simple heating without the use of microwave.

Chapter- 5

Metagenomics

Metagenomics is the study of **metagenomes**, genetic material recovered directly from environmental samples. The broad field may also be referred to as **environmental genomics**, **ecogenomics** or **community genomics**. Traditional microbiology and microbial genome sequencing rely upon cultivated clonal cultures. This relatively new field of genetic research enables studies of organisms that are not easily cultured in a laboratory as well as studies of organisms in their natural environment.

Early environmental gene sequencing cloned specific genes (often the 16S rRNA gene) to produce a profile of diversity in a natural sample. Such work revealed that the vast majority of microbial biodiversity had been missed by cultivation-based methods. Recent studies use "shotgun" Sanger sequencing or massively parallel pyrosequencing to get (mostly) unbiased samples of all genes from all members of sampled communities.

History

Origin of the term

The term "metagenomics" was first used by Jo Handelsman, Jon Clardy, Robert M. Goodman, and others, and first appeared in publication in 1998. The term metagenome referenced the idea that a collection of genes sequenced from the environment could be analyzed in a way analogous to the study of a single genome. The exploding interest in environmental genetics, along with the buzzword-like nature of the term, has resulted in the broader use of metagenomics to describe any sequencing of genetic material from environmental (i.e. uncultured) samples, even work that focuses on one organism or gene. Recently, Kevin Chen and Lior Pachter (researchers at the University of California, Berkeley) defined metagenomics as "the application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, bypassing the need for isolation and lab cultivation of individual species."

Environmental gene surveys

Conventional sequencing begins with a culture of identical cells as a source of DNA. However, early metagenomic studies revealed that there are probably large groups of microorganisms in many environments that cannot be cultured and thus cannot be sequenced. These early studies focused on 16S ribosomal RNA sequences which are relatively short, often conserved within a species, and generally different between species. Many 16S rRNA sequences have been found which do not belong to any known cultured species, indicating that there are numerous non-isolated organisms out there.

Early molecular work in the field was conducted by Norman R. Pace and colleagues, who used PCR to explore the diversity of ribosomal RNA sequences. The insights gained from these breakthrough studies led Pace to propose the idea of cloning DNA directly from environmental samples as early as 1985. This led to the first report of isolating and cloning bulk DNA from an environmental sample, published by Pace and colleagues in 1991 while Pace was in the Department of Biology at Indiana University. Considerable efforts ensured that these were not PCR false positives and supported the existence of a complex community of unexplored species. Although this methodology was limited to exploring highly conserved, non-protein coding genes, it did support early microbial morphology-based observations that diversity was far more complex than was known by culturing methods.

Soon after that, Healy reported the metagenomic isolation of functional genes from "zoolibraries" constructed from a complex culture of environmental organisms grown in the laboratory on dried grasses in 1995. After leaving the Pace laboratory, Ed DeLong continued in the field and has published work that has largely laid the groundwork for environmental phylogenies based on signature 16S sequences, beginning with his group's construction of libraries from marine samples.

Longer sequences from environmental samples

Recovery of DNA sequences longer than a few thousand base pairs from environmental samples was very difficult until recent advances in molecular biological techniques, particularly related to constructing libraries in bacterial artificial chromosomes (BACs), provided better vectors for molecular cloning.

Shotgun metagenomics

Advances in bioinformatics, refinements of DNA amplification, and proliferation of computational power have greatly aided the analysis of DNA sequences recovered from environmental samples. These advances have enabled the adaptation of shotgun sequencing to metagenomic samples. The approach, used to sequence many cultured microorganisms as well as the human genome, randomly shears DNA, sequences many short sequences, and reconstructs them into a consensus sequence.

In 2002, Mya Breitbart, Forest Rohwer, and colleagues used environmental shotgun sequencing to show that 200 liters of seawater contains over 5000 different viruses. Subsequent studies showed that there are >1000 viral species in human stool and possibly a million different viruses per kilogram of marine sediment, including many bacteriophages. Essentially all of the viruses in these studies were new species. In 2004, Gene Tyson, Jill Banfield, and colleagues at the University of California, Berkeley and the Joint Genome Institute sequenced DNA extracted from an acid mine drainage system. This effort resulted in the complete, or nearly complete, genomes for a handful of bacteria and archaea that had previously resisted attempts to culture them. It was now possible to study entire genomes without the biases associated with laboratory cultures.

Global Ocean Sampling Expedition

Beginning in 2003, Craig Venter, leader of the privately-funded parallel of the Human Genome Project, has led the Global Ocean Sampling Expedition, circumnavigating the globe and collecting metagenomic samples throughout. All of these samples are sequenced using shotgun sequencing, in hopes that new genomes (and therefore new organisms) would be identified. The pilot project, conducted in the Sargasso Sea, found DNA from nearly 2000 different species, including 148 types of bacteria never before seen. As of 2009, Venter has circumnavigated the globe and thoroughly explored the West Coast of the United States, and is currently in the midst of a two-year expedition to explore the Baltic, Mediterranian and Black Seas.

Pyrosequencing

In 2006 Robert Edwards, Forest Rohwer, and colleagues at San Diego State University published the first sequences of environmental samples generated with so-called next generation sequencing, in this case chip-based pyrosequencing developed by 454 Life Sciences. This technique for sequencing DNA generates shorter fragments than conventional techniques, however this limitation is compensated for by the very large number of sequences generated. In addition, this technique does not require cloning the DNA before sequencing, removing one of the main biases in metagenomics.

MEGAN

In 2007, Daniel Huson and Stephan Schuster developed and published the first stand-alone metagenome analysis tool, MEGAN, which can be used to perform a first analysis of a metagenomic shotgun dataset. This tool was originally developed to analyse the metagenome of a mammoth sample. However in a recent study by Monzoorul et al. 2009, it was shown that adopting the LCA approach (of MEGAN) solely based on bit-score of the alignment leads to a number of false positive assignments especially in the context of metagenomic sequences originating from new organisms. This study proposed a new approach called SOrt-ITEMS which used several alignment parameters to increase the accuracy of assignments.

MG-RAST

In 2007, Folker Meyer and Robert Edwards and a team at Argonne National Laboratory and the University of Chicago released the Metagenomics RAST server (MG-RAST) a community resource for metagenome data set analysis. The SEED based free, public resource has so far (October 2009) been used for the analysis of over 4000 metagenome data sets. As of October 2009 100+ giga-basepairs of DNA have been analyzed via MG-RAST, more than 350 public data sets are freely available for comparison within MG-RAST.

Applications

Metagenomics can improve strategies for monitoring the impact of pollutants on ecosystems and for cleaning up contaminated environments. Increased understanding of how microbial communities cope with pollutants is helping assess the potential of contaminated sites to recover from pollution and increase the chances of bioaugmentation or biostimulation trials to succeed.

Recent progress in mining the rich genetic resource of non-culturable microbes has led to the discovery of new genes, enzymes, and natural products. The impact of metagenomics is witnessed in the development of commodity and fine chemicals, agrochemicals and pharmaceuticals where the benefit of enzyme-catalyzed chiral synthesis is increasingly recognized.

Metagenomic sequencing is being used to characterize the microbial communities from 15-18 body sites from at least 250 individuals. This is part of the Human Microbiome initiative with primary goals to determine if there is a core human microbiome, to understand the changes in the human microbiome that can be correlated with human health, and to develop new technological and bioinformatics tools to support these goals.

It is well known that the vast majority of microbes have not been cultivated. Functional metagenomics strategies are being used to explore the interactions between plants and microbes through cultivation-independent study of the microbial communities.

Microbial diversity

Much of the interest in metagenomics comes from the discovery that the vast majority of microorganisms had previously gone unnoticed. Traditional microbiological methods relied upon laboratory cultures of organisms. Surveys of ribosomal RNA (rRNA) genes taken directly from the environment revealed that cultivation based methods find less than 1% of the bacteria and archaea species in a sample.

Gene surveys

Shotgun sequencing and screens of clone libraries reveal genes present in environmental samples. This provides information both on which organisms are present and what

metabolic processes are possible in the community. This can be helpful in understanding the ecology of a community, particularly if multiple samples are compared to each other.

Environmental genomes

Shotgun metagenomics also is capable of sequencing nearly complete microbial genomes directly from the environment. Because the collection of DNA from an environment is largely uncontrolled, the most abundant organisms in an environmental sample are most highly represented in the resulting sequence data. To achieve the high coverage needed to fully resolve the genomes of underrepresented community members, large samples, often prohibitively so, are needed. On the other hand, the random nature of shotgun sequencing ensures that many of these organisms will be represented by at least some small sequence segments. Due to the limitations of microbial isolation methods, the vast majority of these organisms would go unnoticed using traditional culturing techniques.

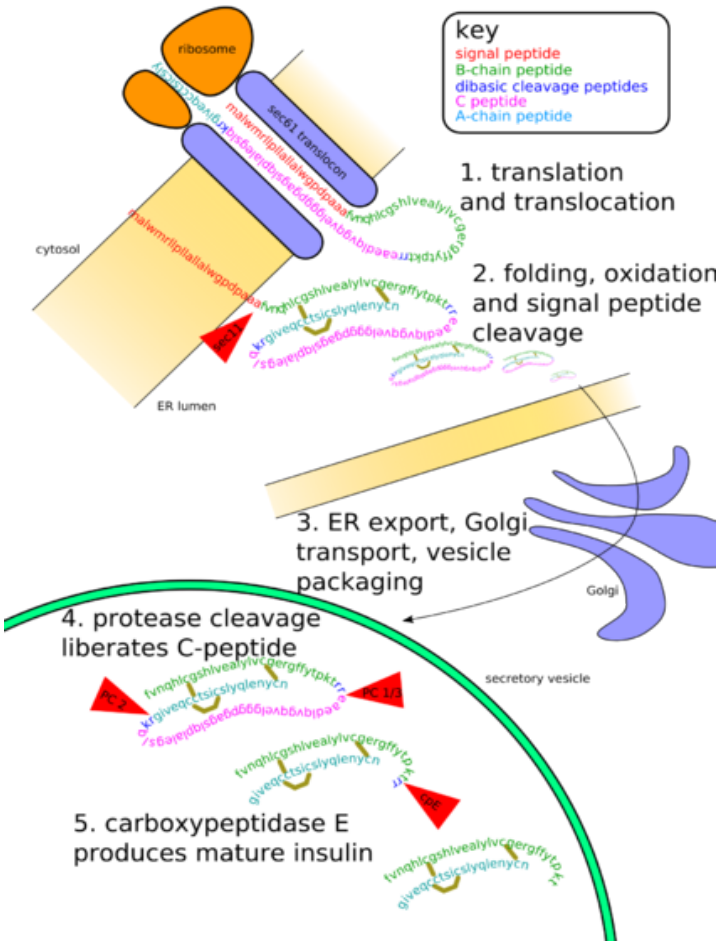
Community metabolism

Many bacterial communities show significant division of labor in metabolism. Waste products of some organisms are metabolites for others. Working together they turn raw resources into fully metabolized waste. Using comparative gene studies and expression experiments with microarrays or proteomics researchers can piece together a metabolic network that goes beyond species boundaries. Such studies require detailed knowledge about which versions of which proteins are coded by which species and even by which strains of which species. Therefore, community genomic information is another fundamental part (as metabolomics or proteomics) to be able to estimate how metabolites are possibly transferred and transformed through a community.

Chapter- 6

Posttranslational Modification

Posttranslational modification (PTM) is the chemical modification of a protein after its translation. It is one of the later steps in protein biosynthesis for many proteins.



The bottom of this diagram shows the modification of primary structure of insulin, as described.

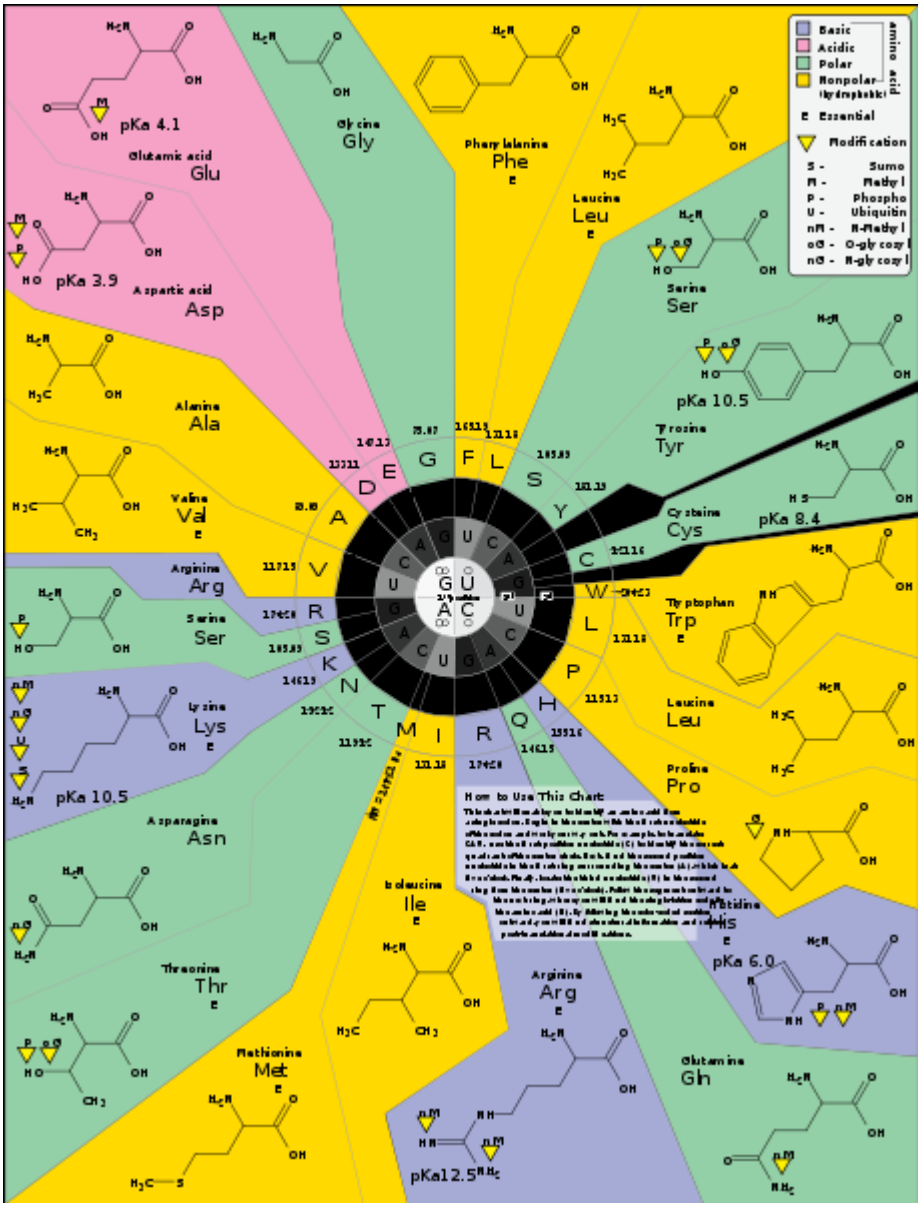
A protein (also called a polypeptide) is a chain of amino acids. During protein synthesis, 20 different amino acids can be incorporated to become a protein. After translation, the posttranslational modification of amino acids extends the range of functions of the protein by attaching to it other biochemical functional groups such as acetate, phosphate, various lipids and carbohydrates, by changing the chemical nature of an amino acid (e.g. citrullination) or by making structural changes, like the formation of disulfide bridges.

Also, enzymes may remove amino acids from the amino end of the protein, or cut the peptide chain in the middle. For instance, the peptide hormone insulin is cut twice after disulfide bonds are formed, and a propeptide is removed from the middle of the chain; the resulting protein consists of two polypeptide chains connected by disulfide bonds. Also, most nascent polypeptides start with the amino acid methionine because the "start" codon on mRNA also codes for this amino acid. This amino acid is usually taken off during post-translational modification.

Other modifications, like phosphorylation, are part of common mechanisms for controlling the behavior of a protein, for instance activating or inactivating an enzyme.

Post-translational modification of proteins is detected by mass spectrometry or Eastern blotting.

PTMs involving addition of functional groups



The genetic code diagram showing the amino acid residues as target of modification.

PTMs involving addition by an enzyme *in vivo*

- acylation, e.g. *O*-acylation (esters), *N*-acylation (amides), *S*-acylation (thioesters)
  - acetylation, the addition of an acetyl group, either at the N-terminus of the protein or at lysine residues. The reverse is called deacetylation.
  - formylation
  - lipoylation, attachment of a lipoate (C_8) functional group
  - myristoylation, attachment of myristate, a C_{14} saturated acid
  - palmitoylation, attachment of palmitate, a C_{16} saturated acid
- alkylation, the addition of an alkyl group, e.g. methyl, ethyl

    - methylation the addition of a methyl group, usually at lysine or arginine residues. The reverse is called demethylation.
    - isoprenylation or prenylation, the addition of an isoprenoid group (e.g. farnesol and geranylgeraniol)
        - farnesylation
        - geranylgeranylation
- amidation at C-terminus
- amino acid addition
    - arginylation, a tRNA-mediation addition
    - polyglutamylation, covalent linkage of glutamic acid residues to tubulin and some other proteins.
    - polyglycylation, covalent linkage of one to more than 40 glycine residues to the tubulin C-terminal tail
- diphthamide formation
- gamma-carboxylation dependent on Vitamin K
- glycosylation, the addition of a glycosyl group to either asparagine, hydroxylysine, serine, or threonine, resulting in a glycoprotein. Distinct from glycation, which is regarded as a nonenzymatic attachment of sugars.
    - polysialylation, addition of polysialic acid, PSA, to NCAM
- glypiation, glycosylphosphatidylinositol (GPI) anchor formation
- heme moiety may be covalently attached
- hydroxylation
- hypusine formation (on conserved lysine of [EIF5A] and aIF5a)
- iodination (e.g. of thyroid hormones)
- nucleotides or derivatives thereof may be covalently attached
    - adenylation
    - ADP-ribosylation
    - flavin attachment
- nitrosylation
- S-glutathionylation
- oxidation
- phosphopantetheinylation, the addition of a 4'-phosphopantetheinyl moiety from coenzyme A, as in fatty acid, polyketide, non-ribosomal peptide and leucine biosynthesis
- phosphorylation, the addition of a phosphate group, usually to serine, tyrosine, threonine or histidine
- pyroglutamate formation
- sulfation, the addition of a sulfate group to a tyrosine.
- selenoylation (co-translational incorporation of selenium in selenoproteins)

PTMs involving non-enzymatic additions *in vivo*

- glycation, the addition of a sugar molecule to a protein without the controlling action of an enzyme.

PTMs involving non-enzymatic additions *in vitro*

- biotinylation, acylation of conserved lysine residues with a biotin appendage
- pegylation

PTMs involving addition of other proteins or peptides

- ISGylation, the covalent linkage to the ISG15 protein (Interferon-Stimulated Gene 15)
- SUMOylation, the covalent linkage to the SUMO protein (Small Ubiquitin-related MOdifier)
- ubiquitination, the covalent linkage to the protein ubiquitin.
- Neddylation, the covalent linkage to Nedd

PTMs involving changing the chemical nature of amino acids

- citrullination, or **deimination**, the conversion of arginine to citrulline
- deamidation, the conversion of glutamine to glutamic acid or asparagine to aspartic acid
- eliminylation, the conversion to an alkene by beta-elimination of phosphothreonine and phosphoserine, or dehydration of threonine and serine, as well as by decarboxylation of cysteine
- carbamylation, the conversion of lysine to homocitrulline

PTMs involving structural changes

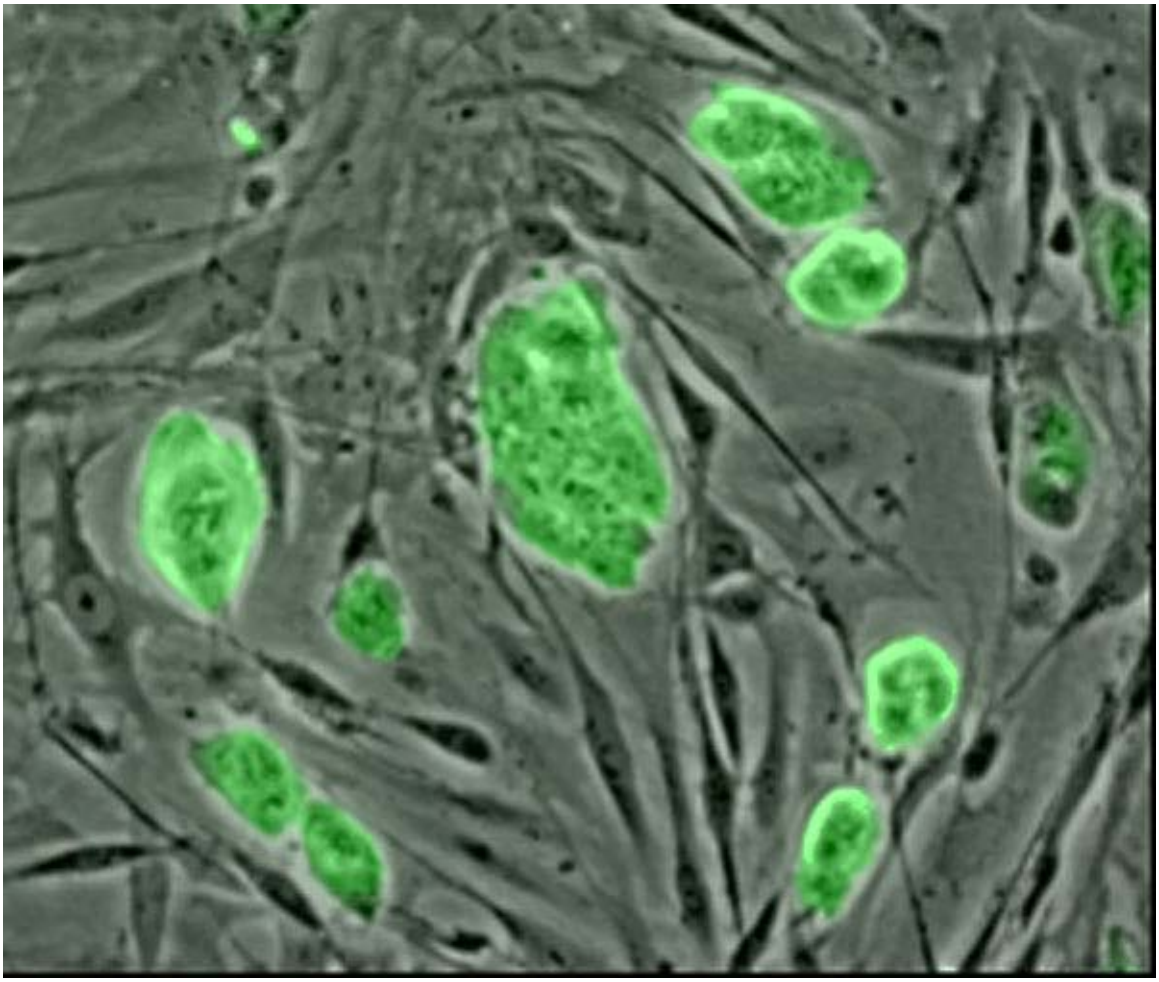
- disulfide bridges, the covalent linkage of two cysteine amino acids
- proteolytic cleavage, cleavage of a protein at a peptide bond
- racemization of proline by prolyl isomerase

Case examples

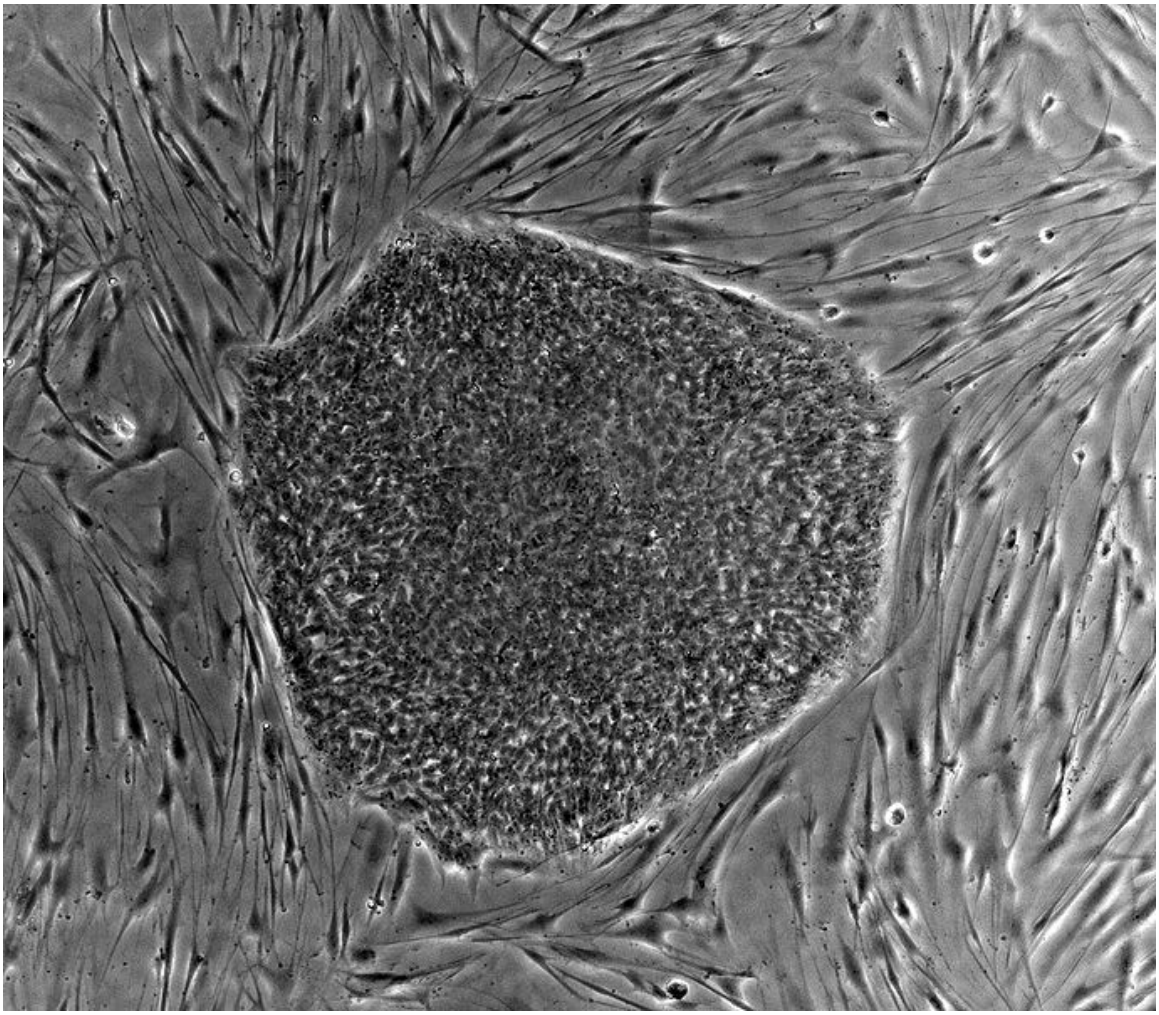
- Cleavage and formation of disulfide bridges during the production of insulin
- PTM of histones as regulation of transcription: RNA polymerase control by chromatin structure
- PTM of RNA polymerase II as regulation of transcription
- Cleavage of polypeptide chains as crucial for lectin specificity

Chapter- 7

Stem Cell



Mouse embryonic stem cells with fluorescent marker



Human embryonic stem cell colony on mouse embryonic fibroblast feeder layer

Stem cells are cells found in all multi cellular organisms. They are characterized by the ability to renew themselves through mitotic cell division and differentiate into a diverse range of specialized cell types. Research in the stem cell field grew out of findings by Ernest A. McCulloch and James E. Till at the University of Toronto in the 1960s.

The two broad types of mammalian stem cells are: embryonic stem cells that are isolated from the inner cell mass of blastocysts, and adult stem cells that are found in adult tissues. In a developing embryo, stem cells can differentiate into all of the specialized embryonic tissues. In adult organisms, stem cells and progenitor cells act as a repair system for the body, replenishing specialized cells, but also maintain the normal turnover of regenerative organs, such as blood, skin, or intestinal tissues.

Stem cells can now be grown and transformed into specialized cells with characteristics consistent with cells of various tissues such as muscles or nerves through cell culture. Highly plastic adult stem cells from a variety of sources, including umbilical cord blood and bone marrow, are routinely used in medical therapies. Embryonic cell lines and

autologous embryonic stem cells generated through therapeutic cloning have also been proposed as promising candidates for future therapies.

Properties

The classical definition of a stem cell requires that it possess two properties:

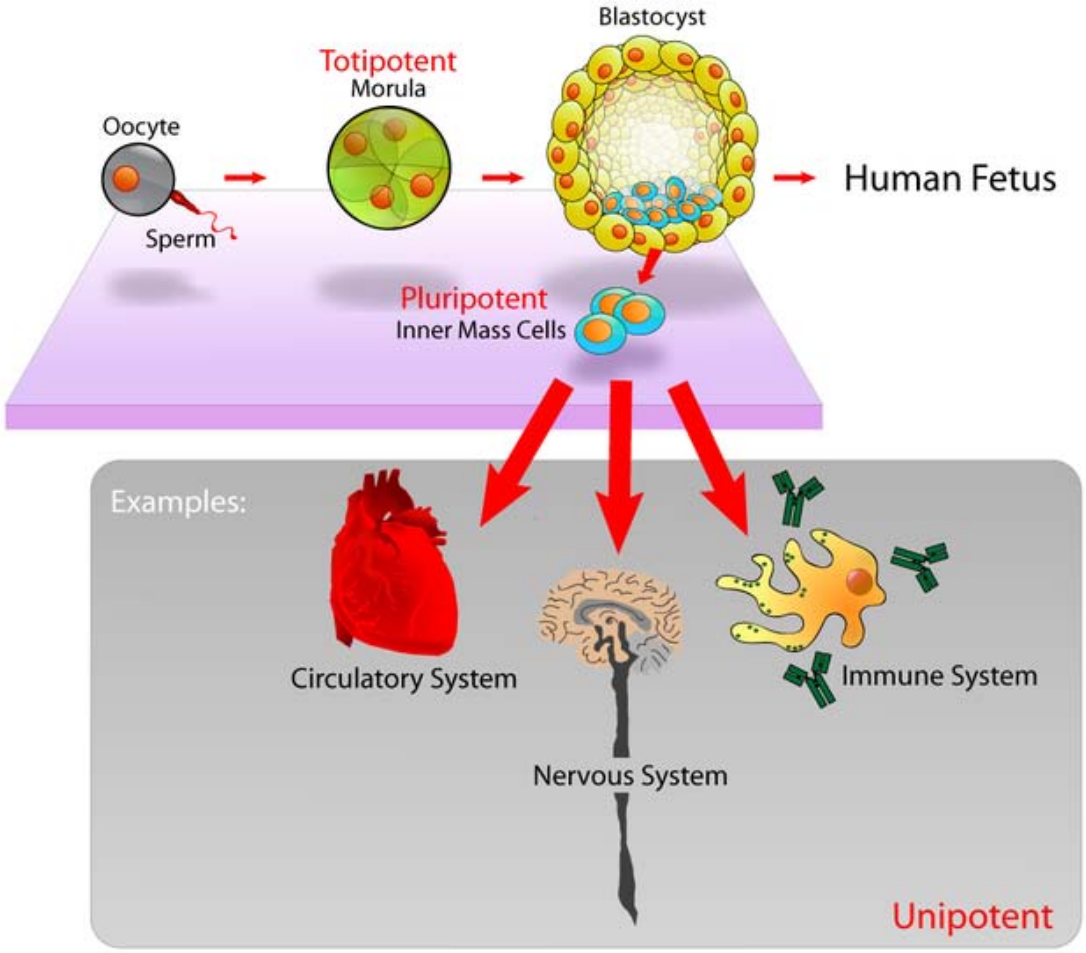
- *Self-renewal* - the ability to go through numerous cycles of cell division while maintaining the undifferentiated state.
- *Potency* - the capacity to differentiate into specialized cell types. In the strictest sense, this requires stem cells to be either totipotent or pluripotent - to be able to give rise to any mature cell type, although multipotent or unipotent progenitor cells are sometimes referred to as stem cells.

Self-renewal

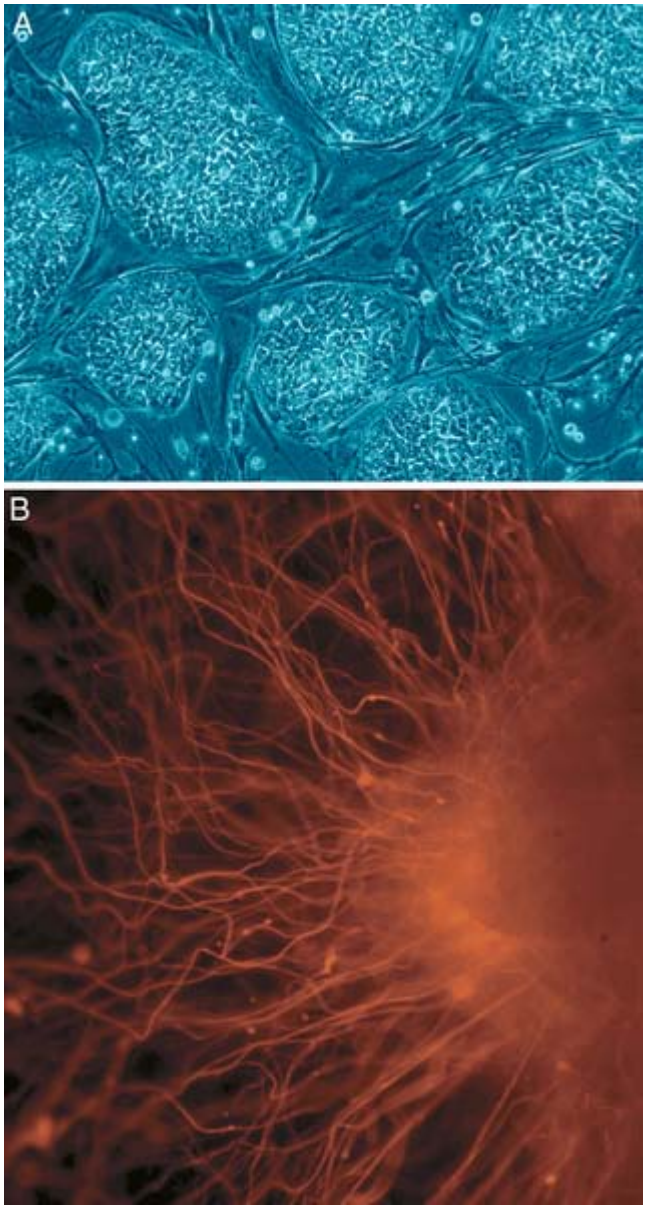
Two mechanisms exist to ensure that the stem cell population is maintained:

1. Obligatory asymmetric replication - a stem cell divides into one father cell that is identical to the original stem cell, and another daughter cell that is differentiated
2. Stochastic differentiation - when one stem cell develops into two differentiated daughter cells, another stem cell undergoes mitosis and produces two stem cells identical to the original.

Potency definitions



Pluripotent, embryonic stem cells originate as inner mass cells within a blastocyst. The stem cells can become any tissue in the body, excluding a placenta. Only the morula's cells are totipotent, able to become all tissues and a placenta.



Human embryonic stem cells
A: Cell colonies that are not yet differentiated.
B: Nerve cell

Potency specifies the differentiation potential (the potential to differentiate into different cell types) of the stem cell.

- Totipotent (a.k.a omnipotent) stem cells can differentiate into embryonic and extraembryonic cell types. Such cells can construct a complete, viable, organism. These cells are produced from the fusion of an egg and sperm cell. Cells produced by the first few divisions of the fertilized egg are also totipotent.
- Pluripotent stem cells are the descendants of totipotent cells and can differentiate into nearly all cells, i.e. cells derived from any of the three germ layers.

- Multipotent stem cells can differentiate into a number of cells, but only those of a closely related family of cells.
- Oligopotent stem cells can differentiate into only a few cells, such as lymphoid or myeloid stem cells.
- Unipotent cells can produce only one cell type, their own, but have the property of self-renewal which distinguishes them from non-stem cells (e.g. muscle stem cells).

Identification

The practical definition of a stem cell is the functional definition - a cell that has the potential to regenerate tissue over a lifetime. For example, the gold standard test for a bone marrow or hematopoietic stem cell (HSC) is the ability to transplant one cell and save an individual without HSCs. In this case, a stem cell must be able to produce new blood cells and immune cells over a long term, demonstrating potency. It should also be possible to isolate stem cells from the transplanted individual, which can themselves be transplanted into another individual without HSCs, demonstrating that the stem cell was able to self-renew.

Properties of stem cells can be illustrated *in vitro*, using methods such as clonogenic assays, where single cells are characterized by their ability to differentiate and self-renew. As well, stem cells can be isolated based on a distinctive set of cell surface markers. However, *in vitro* culture conditions can alter the behavior of cells, making it unclear whether the cells will behave in a similar manner *in vivo*. Considerable debate exists whether some proposed adult cell populations are truly stem cells.

Embryonic

Embryonic stem cell lines (ES cell lines) are cultures of cells derived from the epiblast tissue of the inner cell mass (ICM) of a blastocyst or earlier morula stage embryos. A blastocyst is an early stage embryo—approximately four to five days old in humans and consisting of 50–150 cells. ES cells are pluripotent and give rise during development to all derivatives of the three primary germ layers: ectoderm, endoderm and mesoderm. In other words, they can develop into each of the more than 200 cell types of the adult body when given sufficient and necessary stimulation for a specific cell type. They do not contribute to the extra-embryonic membranes or the placenta.

Nearly all research to date has taken place using mouse embryonic stem cells (mES) or human embryonic stem cells (hES). Both have the essential stem cell characteristics, yet they require very different environments in order to maintain an undifferentiated state. Mouse ES cells are grown on a layer of gelatin and require the presence of Leukemia Inhibitory Factor (LIF). Human ES cells are grown on a feeder layer of mouse embryonic fibroblasts (MEFs) and require the presence of basic Fibroblast Growth Factor (bFGF or FGF-2). Without optimal culture conditions or genetic manipulation, embryonic stem cells will rapidly differentiate.

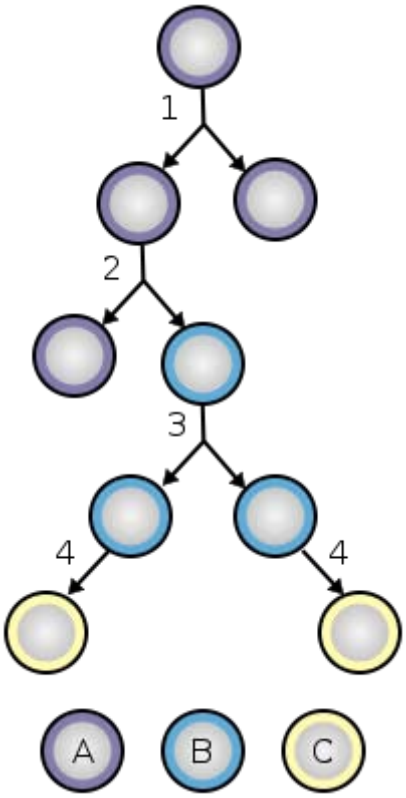
A human embryonic stem cell is also defined by the presence of several transcription factors and cell surface proteins. The transcription factors Oct-4, Nanog, and Sox2 form the core regulatory network that ensures the suppression of genes that lead to differentiation and the maintenance of pluripotency. The cell surface antigens most commonly used to identify hES cells are the glycolipids SSEA3 and SSEA4 and the keratan sulfate antigens Tra-1-60 and Tra-1-81. The molecular definition of a stem cell includes many more proteins and continues to be a topic of research.

After nearly ten years of research, there are no approved treatments using embryonic stem cells. The first human trial was approved by the US Food & Drug Administration in January 2009. However, as of August 2010, the first human trial had not yet been initiated. The first human medical trial for embryonic stem cells started in Atlanta on October 13, 2010 for spinal injury victims. ES cells, being pluripotent cells, require specific signals for correct differentiation - if injected directly into another body, ES cells will differentiate into many different types of cells, causing a teratoma. Differentiating ES cells into usable cells while avoiding transplant rejection are just a few of the hurdles that embryonic stem cell researchers still face. Many nations currently have moratoria on either ES cell research or the production of new ES cell lines. Because of their combined abilities of unlimited expansion and pluripotency, embryonic stem cells remain a theoretically potential source for regenerative medicine and tissue replacement after injury or disease.

Fetal

Fetal stem cells are primitive cell types found in the organs of fetuses.

Adult



Stem cell division and differentiation. A - stem cell; B - progenitor cell; C - differentiated cell; 1 - symmetric stem cell division; 2 - asymmetric stem cell division; 3 - progenitor division; 4 - terminal differentiation

Also known as somatic (from Greek Σωματικός, "of the body") stem cells and germline (giving rise to gametes) stem cells, they can be found in children, as well as adults.

Pluripotent adult stem cells are rare and generally small in number but can be found in a number of tissues including umbilical cord blood. A great deal of adult stem cell research has focused on clarifying their capacity to divide or self-renew indefinitely and their differentiation potential. In mice, pluripotent stem cells are directly generated from adult fibroblast cultures. Unfortunately, many mice don't live long with stem cell organs.

Most adult stem cells are lineage-restricted (multipotent) and are generally referred to by their tissue origin (mesenchymal stem cell, adipose-derived stem cell, endothelial stem cell, dental pulp stem cell, etc.).

Adult stem cell treatments have been successfully used for many years to treat leukemia and related bone/blood cancers through bone marrow transplants. Adult stem cells are also used in veterinary medicine to treat tendon and ligament injuries in horses.

The use of adult stem cells in research and therapy is not as controversial as embryonic stem cells, because the production of adult stem cells does not require the destruction of an embryo. Additionally, because in some instances adult stem cells can be obtained from the intended recipient, (an autograft) the risk of rejection is essentially non-existent in these situations. Consequently, more US government funding is being provided for adult stem cell research.

An extremely rich source for adult mesenchymal stem cells is the developing tooth bud of the mandibular third molar. While considered multipotent they may prove to be pluripotent. The stem cells eventually form enamel (ectodrm), dentin,periodontal ligament, blood vessels, dental pulp, nervous tissues, including a minimum of 29 different unique end organs. Because of extreme ease in collection at 8–10 years of age before calcification and minimal to no morbidity will probably constitute a major source for personal banking, research and multiple therapies. These stem cells have been shown capable of producing hepatocytes.

Amniotic

Multipotent stem cells are also found in amniotic fluid. These stem cells are very active, expand extensively without feeders and are not tumorigenic. Amniotic stem cells are multipotent and can differentiate in cells of adipogenic, osteogenic, myogenic, endothelial, hepatic and also neuronal lines. All over the world, universities and research institutes are studying amniotic fluid to discover all the qualities of amniotic stem cells, and scientists such as Anthony Atala and Giuseppe Simoni have discovered important results.

From an ethical point of view, stem cells from amniotic fluid can solve a lot of problems, because it's possible to catch amniotic stem cells without destroying embryos. For example, the Vatican newspaper "Osservatore Romano" called amniotic stem cell "the future of medicine".

It's possible to collect amniotic stem cells for donors or for autologuous use: the first US amniotic stem cells bank opened in 2009 in Medford, MA, by Biocell Center Corporation and collaborates with various hospitals and universities all over the world.

Induced pluripotent

These are not adult stem cells, but rather reprogrammed cells (e.g. epithelial cells) given pluripotent capabilities. Using genetic reprogramming with protein transcription factors, pluripotent stem cells equivalent to embryonic stem cells have been derived from human adult skin tissue. Shinya Yamanaka and his colleagues at Kyoto University used the transcription factors Oct3/4, Sox2, c-Myc, and Klf4 in their experiments on cells from human faces. Junying Yu, James Thomson, and their colleagues at the University of Wisconsin–Madison used a different set of factors, Oct4, Sox2, Nanog and Lin28, and carried out their experiments using cells from human foreskin.

As a result of the success of these experiments, Ian Wilmut, who helped create the first cloned animal Dolly the Sheep, has announced that he will abandon nuclear transfer as an avenue of research.

Frozen blood samples can be used as a source of induced pluripotent stem cells, opening a new avenue for obtaining the valued cells.

Lineage

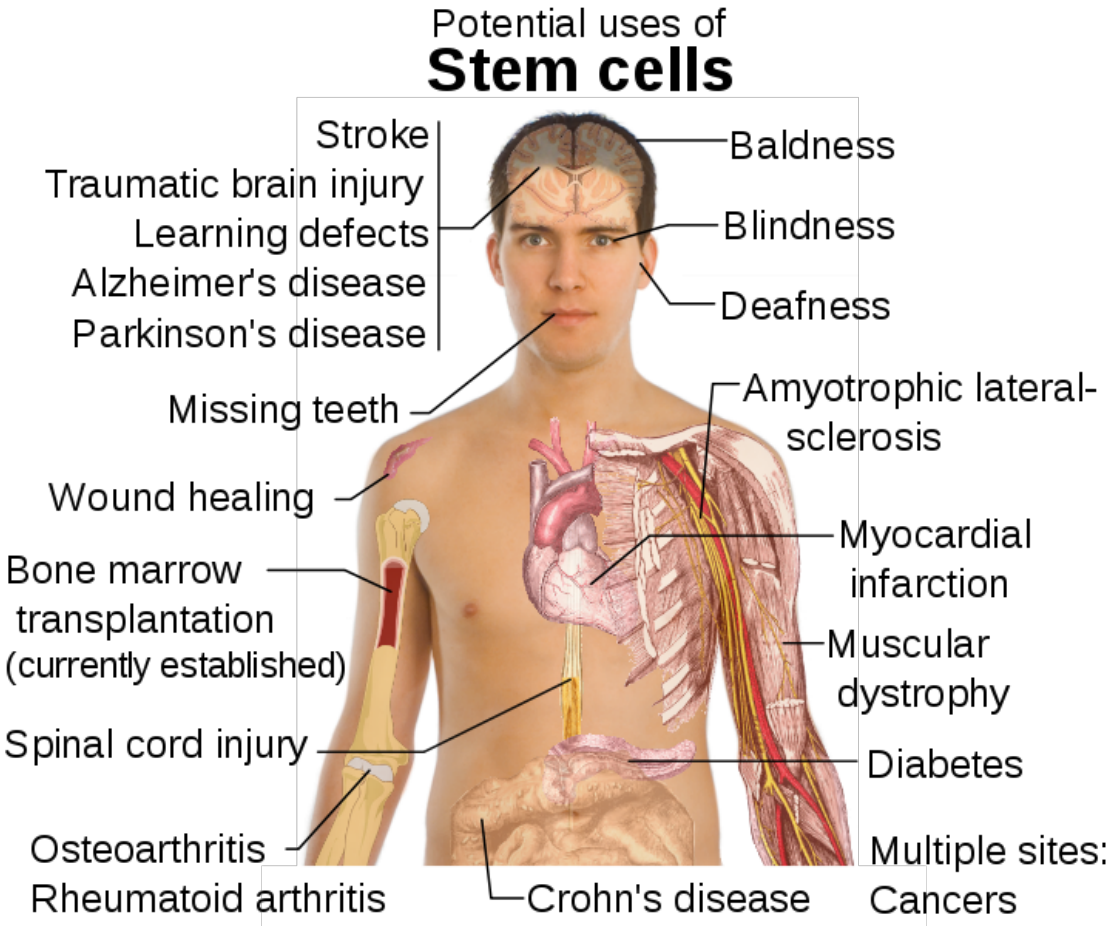
To ensure self-renewal, stem cells undergo two types of cell division. Symmetric division gives rise to two identical daughter cells both endowed with stem cell properties. Asymmetric division, on the other hand, produces only one stem cell and a progenitor cell with limited self-renewal potential. Progenitors can go through several rounds of cell division before terminally differentiating into a mature cell. It is possible that the molecular distinction between symmetric and asymmetric divisions lies in differential segregation of cell membrane proteins (such as receptors) between the daughter cells.

An alternative theory is that stem cells remain undifferentiated due to environmental cues in their particular niche. Stem cells differentiate when they leave that niche or no longer receive those signals. Studies in *Drosophila* germarium have identified the signals dpp and adherens junctions that prevent germarium stem cells from differentiating.

The signals that lead to reprogramming of cells to an embryonic-like state are also being investigated. These signal pathways include several transcription factors including the oncogene c-Myc. Initial studies indicate that transformation of mice cells with a combination of these anti-differentiation signals can reverse differentiation and may allow adult cells to become pluripotent. However, the need to transform these cells with an oncogene may prevent the use of this approach in therapy.

Challenging the terminal nature of cellular differentiation and the integrity of lineage commitment, it was recently determined that the somatic expression of combined transcription factors can directly induce other defined somatic cell fates; researchers identified three neural-lineage-specific transcription factors that could directly convert mouse fibroblasts (skin cells) into fully functional neurons. This "induced neurons" (iN) cell research inspires the researchers to induce other cell types implies that *all* cells are totipotent: with the proper tools, all cells may form all kinds of tissue.

Treatments



Diseases and conditions where stem cell treatment is promising or emerging. Bone marrow transplantation is, as of 2009, the only established use of stem cells.

Medical researchers believe that stem cell therapy has the potential to dramatically change the treatment of human disease. A number of adult stem cell therapies already exist, particularly bone marrow transplants that are used to treat leukemia. In the future, medical researchers anticipate being able to use technologies derived from stem cell research to treat a wider variety of diseases including cancer, Parkinson's disease, spinal cord injuries, Amyotrophic lateral sclerosis, multiple sclerosis, and muscle damage, amongst a number of other impairments and conditions. However, there still exists a great deal of social and scientific uncertainty surrounding stem cell research, which could possibly be overcome through public debate and future research, and further education of the public.

One concern of treatment is the possible risk that transplanted stem cells could form tumors and have the possibility of becoming cancerous if cell division continues uncontrollably.

Stem cells, however, are already studied extensively. While some scientists are hesitant to associate the therapeutic potential of stem cells as the first goal of the research, they find the investigation of stem cells as a goal worthy in itself.

Contrarily, supporters of embryonic stem cell research argue that such research should be pursued because the resultant treatments could have significant medical potential. It is also noted that excess embryos created for in vitro fertilization could be donated with consent and used for the research.

The recent development of iPS cells has been called a bypass of the legal controversy. Laws limiting the destruction of human embryos have been credited for being the reason for development of iPS cells, but they are less efficient and reliable than natural stem cells. Various methods are being developed to bypass this problem by removing mutation.

Research patents

The patents covering a lot of work on human embryonic stem cells are owned by the Wisconsin Alumni Research Foundation (WARF). WARF does not charge academics to study human stem cells but does charge commercial users. WARF sold Geron Corp. exclusive rights to work on human stem cells but later sued Geron Corp. to recover some of the previously sold rights. The two sides agreed that Geron Corp. would keep the rights to only three cell types. In 2001, WARF came under public pressure to widen access to human stem-cell technology.

These patents are now in doubt as a request for reviewing the US Patent and Trademark Office has been filed by non-profit patent-watchdogs The Foundation for Taxpayer & Consumer Rights, and the Public Patent Foundation as well as molecular biologist Jeanne Loring of the Burnham Institute. According to them, two of the patents granted to WARF are invalid because they cover a technique published in 1993 for which a patent had already been granted to an Australian researcher. Another part of the challenge states that these techniques, developed by James A. Thomson, are rendered obvious by a 1990 paper and two textbooks.

The outcome of this legal challenge is particularly relevant to the Geron Corp. as it can only license patents that are upheld.

Chapter- 8

Synthetic Biology

Synthetic biology is a new area of biological research that combines science and engineering. Synthetic biology encompasses a variety of different approaches, methodologies and disciplines, and many different definitions exist. What they all have in common, however, is that they see synthetic biology as the design and construction of new biological functions and systems not found in nature.

A light programmable biofilm made by the UT Austin / UCSF team during the 2004 Synthetic Biology competition, displaying "Hello World"

History of the term

The term "synthetic biology" has a history spanning the twentieth century.. The first use was in Stéphane Leduc's publication of La Biologie Synthétique in 1912. In 1974, the Polish geneticist Waclaw Szybalski used the term "synthetic biology", writing:

Let me now comment on the question "what next". Up to now we are working on the descriptive phase of molecular biology. ... But the real challenge will start when we enter the synthetic biology phase of research in our field. We will then devise new control elements and add these new modules to the existing genomes or build up wholly new genomes. This would be a field with the unlimited expansion potential and hardly any limitations to building "new better control circuits" and ..... finally other "synthetic" organisms, like a "new better mouse". ... I am not concerned that we will run out of exciting and novel ideas, ... in the synthetic biology, in general.

When in 1978 the Nobel Prize in Physiology or Medicine was awarded to Arber, Nathans and Smith for the discovery of restriction enzymes, Waclaw Szybalski wrote in an editorial comment in the journal *Gene*:

The work on restriction nucleases not only permits us easily to construct recombinant DNA molecules and to analyze individual genes, but also has led us into the new era of synthetic biology where not only existing genes are described and analyzed but also new gene arrangements can be constructed and evaluated.

Biology

Biologists are interested in learning more about how natural living systems work. One simple, direct way to test our current understanding of a natural living system is to build an instance (or version) of the system in accordance with our current understanding of the system. Michael Elowitz's early work on the Repressilator is one good example of such work. Elowitz is generally considered to be the father of synthetic biology. Elowitz had a model for how gene expression should work inside living cells. To test his model, he built a piece of DNA in accordance with his model, placed the DNA inside living cells, and watched what happened. Slight differences between observation and expectation highlight new science that may be well worth doing. Work of this sort often makes good use of mathematics to predict and study the dynamics of the biological system before experimentally constructing it. A wide variety of mathematical descriptions have been used with varying accuracy, including graph theory, Boolean networks, ordinary differential equations, stochastic differential equations, and Master equations (in order of increasing accuracy). Good examples include the work of Adam Arkin, Jim Collins and Alexander van Oudenaarden.

Chemistry

Biological systems are physical systems that are made up of chemicals. Around 100 years ago, the science of chemistry went through a transition from studying natural chemicals

to trying to design and build new chemicals. This transition led to the field of synthetic chemistry. In the same tradition, some aspects of synthetic biology can be viewed as an extension and application of synthetic chemistry to biology, and include work ranging from the creation of useful new biochemicals to studying the origins of life. Eric Kool's group at Stanford, the Foundation for Applied Molecular Evolution, Carlos Bustamante's group at Berkeley, Jack Szostak's group at Harvard, and David McMillen's group at University of Toronto are good examples of this tradition. Much of the improved economics and versatility of synthetic biology is driven by ongoing improvements in gene synthesis.

Engineering

Engineers view biology as a *technology.* Synthetic Biology includes the broad redefinition and expansion of biotechnology, with the ultimate goals of being able to design and build engineered biological systems that process information, manipulate chemicals, fabricate materials and structures, produce energy, provide food, and maintain and enhance human health and our environment . A good example of these technologies include the work of Chris Voigt, who redesigned the Type III secretion system used by *Salmonella typhimurium* to secrete spider silk proteins, a strong elastic biomaterial, instead of its own natural infectious proteins. One aspect of Synthetic Biology which distinguishes it from conventional genetic engineering is a heavy emphasis on developing foundational technologies that make the engineering of biology easier and more reliable. Good examples of engineering in synthetic biology include the pioneering work of Tim Gardner and Jim Collins on an engineered genetic toggle switch, a riboregulator, the Registry of Standard Biological Parts, and the International Genetically Engineered Machine competition (iGEM).

Studies in synthetic biology can be subdivided into broad classifications according to the approach they take to the problem at hand: photocell design, biomolecular engineering, genome engineering, and biomolecular-design. The photocell approach includes projects to make self-replicating systems from entirely synthetic components. Biomolecular engineering includes approaches which aim to create a toolkit of functional units that can be introduced to present new orthogonal functions in living cells. Genome engineering includes approaches to construct synthetic chromosomes for whole or minimal organisms. Biomolecular-design approach refers to the general idea of the de novo design and combination of biomolecular components. The task of each of these approaches is similar: To create a more synthetic entry at a higher level of complexity by manipulating a part of the proceeding level.

Re-writing

Re-writers are Synthetic Biologists who are interested in testing the idea that since natural biological systems are so complicated, we would be better off re-building the natural systems that we care about, from the ground up, in order to provide engineered surrogates that are easier to understand and interact with. Re-writers draw inspiration from refactoring, a process sometimes used to improve computer software. Drew Endy

and his group have done some preliminary work on re-writing (e.g., Refactoring Bacteriophage T7). Oligonucleotides harvested from a photolithographic or inkjet manufactured DNA chip combined with DNA mismatch error-correction allows inexpensive large-scale changes of codons in genetic systems to improve gene expression or incorporate novel amino-acids. As in the T7 example above, this favors a synthesis-from-scratch approach.

Challenges

Opposition to Synthetic Biology

Opposition by civil society groups to Synthetic Biology has been led by the ETC Group who have called for a global moratorium on developments in the field and for no synthetic organisms to be released from the lab. In 2006 38 civil society organizations authored an open letter opposing voluntary regulation of the field and in 2008 ETC Group released the first critical report on the societal impacts of synthetic biology which they dubbed "Extreme Genetic Engineering".. Other groups opposing Synthetic Biology developments include Friends of the Earth, Alliance for Humane Biotechnology, International Center for Technology Assessment and Centro Ecologico (Portuguese).

Safety and Security

In addition to numerous scientific and technical challenges, synthetic biology raises questions for ethics, biosecurity, biosafety, involvement of stakeholders and intellectual property. To date, key stakeholders (especially in the US) have focused primarily on the biosecurity issues, especially the so-called dual-use challenge. For example, while the study of synthetic biology may lead to more efficient ways to produce medical treatments (e.g. against malaria), it may also lead to synthesis or redesign of harmful pathogens (e.g., smallpox) by malicious actors . Proposals for licensing and monitoring the various phases of gene and genome synthesis began to appear in 2004. A 2007 study compared several policy options for governing the security risks associated with synthetic biology. Other initiatives, such as OpenWetWare, diybio, biopunk, biohack, and possibly others, have attempted to integrate self-regulation in their proliferation of open source synthetic biology projects. However the distributed and diffuse nature of open-source biotechnology may make it more difficult to track, regulate, or mitigate potential biosafety and biosecurity concerns.

An initiative for self-regulation has been proposed by the International Association Synthetic Biology that suggests some specific measures to be implemented by the synthetic biology industry, especially DNA synthesis companies. Some scientists, however, argue for a more radical and forward looking approaches to improve safety and security issues. They suggest to use not only physical containment as safety measures, but also trophic and semantic containment. Trophic containment includes for example the design of new and more robust forms of auxotrophy, while semantic containment means the design and construction of completely novel orthogonal life-forms.

Social and Ethical

Online discussion of “societal issues” took place at the SYNBIOSAFE forum on issues regarding ethics, safety, security, IPR, governance, and public perception (summary paper). On July 9–10, 2009, the National Academies' Committee of Science, Technology & Law convened a symposium on "Opportunities and Challenges in the Emerging Field of Synthetic Biology" (transcripts, audio, and presentations available).

Some efforts have been made to engage social issues "upstream" focus on the integral and mutually formative relations among scientific and other human practices. These approaches attempt to invent ongoing and regular forms of collaboration among synthetic biologists, ethicists, political analysts, funders, human scientists and civil society activists. These collaborations have consisted either of intensive, short term meetings, aimed at producing guidelines or regulations, or standing committees whose purpose is limited to protocol review or rule enforcement. Such work has proven valuable in identifying the ways in which synthetic biology intensifies already-known challenges in rDNA technologies. However, these forms are not suited to identifying new challenges as they emerge, and critics worry about uncritical complicity.

An example of efforts to develop ongoing collaboration is the "Human Practices" component of the Synthetic Biology Engineering Research Center in the US and the SYNBIOSAFE project in Europe, coordinated by IDC, that investigated the biosafety, biosecurity and ethical aspects of synthetic biology. A report from the Woodrow Wilson Center and the Hastings Center, a prestigious bioethics research institute, found that ethical concerns in synthetic biology have received scant attention.

In January 2009, the Alfred P. Sloan Foundation funded the Woodrow Wilson Center, the Hastings Center, and the J. Craig Venter Institute to examine the public perception, ethics, and policy implications of synthetic biology. Public perception and communication of synthetic biology is the main focus of COSY: Communicating Synthetic Biology, that showed that in the general public synthetic biology is not seen as too different from 'traditional' genetic engineering . To better communicate synthetic biology and its societal ramifications to a broader public, COSY and SYNBIOSAFE published a 38 min. documentary film in October 2009 .

Key enabling technologies

There are several key enabling technologies that are critical to the growth of synthetic biology. The key concepts include standardization of biological parts and hierarchical abstraction to permit using those parts in increasingly complex synthetic systems.. Achieving this is greatly aided by basic technologies of reading and writing of DNA (sequencing and fabrication), which are improving in price/performance exponentially (Kurzweil 2001). Measurements under a variety of conditions are needed for accurate modeling and computer-aided-design (CAD).

DNA sequencing

DNA sequencing is determining the order of the nucleotide bases in a molecule of DNA. Synthetic biologists make use of DNA sequencing in their work in several ways. First, large-scale genome sequencing efforts continue to provide a wealth of information on naturally occurring organisms. This information provides a rich substrate from which synthetic biologists can construct parts and devices. Second, synthetic biologists use sequencing to verify that they fabricated their engineered system as intended. Third, fast, cheap and reliable sequencing can also facilitate rapid detection and identification of synthetic systems and organisms.

Fabrication

A critical limitation in synthetic biology today is the time and effort expended during fabrication of engineered genetic sequences. To speed up the cycle of design, fabrication, testing and redesign, synthetic biology requires more rapid and reliable *de novo* DNA synthesis and assembly of fragments of DNA, in a process commonly referred to as gene synthesis.

In 2000, researchers at Washington University, mentioned synthesis of the 9.6 kbp Hepatitis C virus genome from chemically synthesized 60 to 80-mers. In 2002 researchers at SUNY Stony Brook succeeded in synthesizing the 7741 base poliovirus genome from its published sequence, producing the second synthetic genome. This took about two years of painstaking work. In 2003 the 5386 bp genome of the bacteriophage Phi X 174 was assembled in about two weeks. In 2006, the same team, at the J. Craig Venter Institute, has constructed and patented a synthetic genome of a novel minimal bacterium, *Mycoplasma laboratorium* and is working on getting it functioning in a living cell.

In 2007 it was reported that several companies were offering the synthesis of genetic sequences up to 2000 bp long, for a price of about $1 per base pair and a turnaround time of less than two weeks. By September 2009, the price had dropped to less than $0.50 per base pair with some improvement in turn around time. Not only is the price judged lower than the cost of conventional cDNA cloning, the economics make it practical for researchers to design and purchase multiple variants of the same sequence to identify genes or proteins with optimized performance.

In 2010, Venter's group announced they had been able to assemble a complete genome of millions of base pairs, insert it into a cell, and cause that cell to start replicating.

Modeling

Models inform the design of engineered biological systems by allowing synthetic biologists to better predict system behavior prior to fabrication. Synthetic biology will benefit from better models of how biological molecules bind substrates and catalyze reactions, how DNA encodes the information needed to specify the cell and how multi-

component integrated systems behave. Recently, multiscale models of gene regulatory networks have been developed that focus on synthetic biology applications. Simulations have been used that model all biomolecular interactions in transcription, translation, regulation, and induction of gene regulatory networks, guiding the design of synthetic systems.

Measurement

Precise and accurate quantitative measurements of biological systems are crucial to improving understanding of biology. Such measurements often help to elucidate how biological systems work and provide the basis for model construction and validation. Differences between predicted and measured system behavior can identify gaps in understanding and explain why synthetic systems don't always behave as intended. Technologies which allow many parallel and time-dependent measurements will be especially useful in synthetic biology. Microscopy and flow cytometry are examples of useful measurement technologies.

Chapter- 9

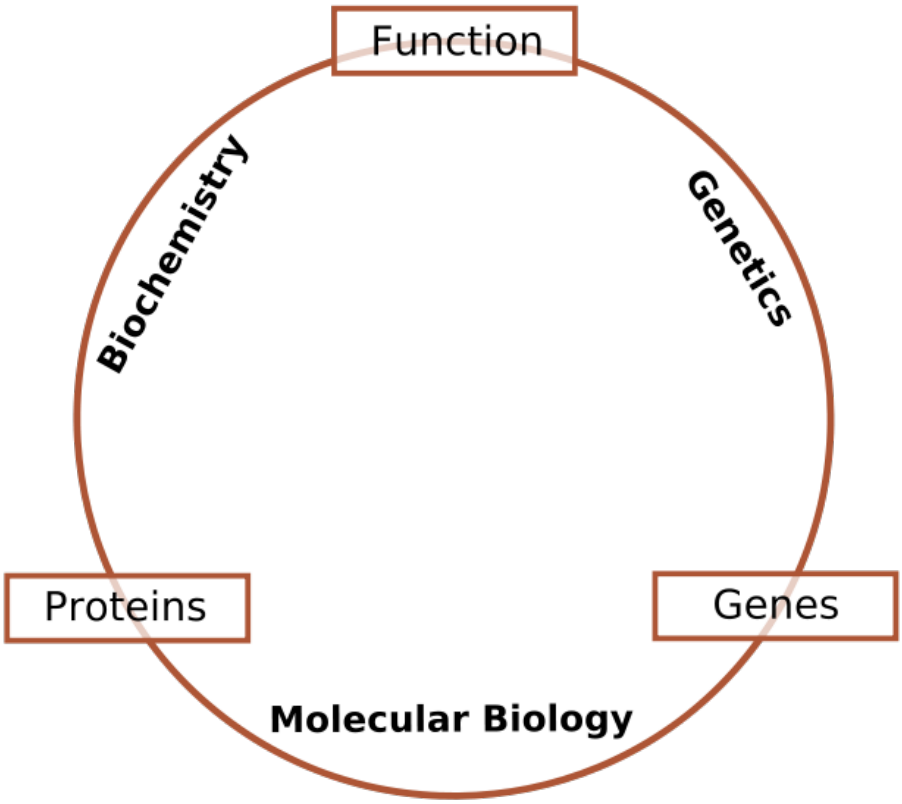
Molecular Biology

Molecular biology is the branch of biology that deals with the molecular basis of biological activity. This field overlaps with other areas of biology and chemistry, particularly genetics and biochemistry. Molecular biology chiefly concerns itself with understanding and the interactions between the various systems of a cell, including the interactions between the different types of DNA, RNA and protein biosynthesis as well as learning how these interactions are regulated.

Writing in *Nature* in 1961, William Astbury described molecular biology as

not so much a technique as an approach, an approach from the viewpoint of the so-called basic sciences with the leading idea of searching below the large-scale manifestations of classical biology for the corresponding molecular plan. It is concerned particularly with the *forms* of biological molecules and [...] is predominantly three-dimensional and structural—which does not mean, however, that it is merely a refinement of morphology. It must at the same time inquire into genesis and function.

Relationship to other biological sciences



Schematic relationship between biochemistry, genetics, and molecular biology

Researchers in molecular biology use specific techniques native to molecular biology, but increasingly combine these with techniques and ideas from genetics and biochemistry. There is not a defined line between these disciplines. The figure above is a schematic that depicts one possible view of the relationship between the fields:

- *Biochemistry* is the study of the chemical substances and vital processes occurring in living organisms. Biochemists focus heavily on the role, function, and structure of biomolecules. The study of the chemistry behind biological processes and the synthesis of biologically active molecules are examples of biochemistry.
- *Genetics* is the study of the effect of genetic differences on organisms. Often this can be inferred by the absence of a normal component (e.g. one gene). The study of "mutants" – organisms which lack one or more functional components with respect to the so-called "wild type" or normal phenotype. Genetic interactions (epistasis) can often confound simple interpretations of such "knock-out" studies.
- *Molecular biology* is the study of molecular underpinnings of the processes of replication, transcription, translation, and cell function. The central dogma of molecular biology where genetic material is transcribed into RNA and then translated into protein, despite being an oversimplified picture of molecular

biology, still provides a good starting point for understanding the field. This picture, however, is undergoing revision in light of emerging novel roles for RNA.

Much of the work in molecular biology is quantitative, and recently much work has been done at the interface of molecular biology and computer science in bioinformatics and computational biology. As of the early 2000s, the study of gene structure and function, molecular genetics, has been among the most prominent sub-field of molecular biology.

Increasingly many other loops of biology focus on molecules, either directly studying their interactions in their own right such as in cell biology and developmental biology, or indirectly, where the techniques of molecular biology are used to infer historical attributes of populations or species, as in fields in evolutionary biology such as population genetics and phylogenetics. There is also a long tradition of studying biomolecules "from the ground up" in biophysics.

Techniques of molecular biology

Since the late 1950s and early 1960s, molecular biologists have learned to characterize, isolate, and manipulate the molecular components of cells and organisms. These components include DNA, the repository of genetic information; RNA, a close relative of DNA whose functions range from serving as a temporary working copy of DNA to actual structural and enzymatic functions as well as a functional and structural part of the translational apparatus; and proteins, the major structural and enzymatic type of molecule in cells.

Expression cloning

One of the most basic techniques of molecular biology to study protein function is expression cloning. In this technique, DNA coding for a protein of interest is cloned (using PCR and/or restriction enzymes) into a plasmid (known as an expression vector). This plasmid may have special promoter elements to drive production of the protein of interest, and may also have antibiotic resistance markers to help follow the plasmid.

This plasmid can be inserted into either bacterial or animal cells. Introducing DNA into bacterial cells can be done by transformation (via uptake of naked DNA), conjugation (via cell-cell contact) or by transduction (via viral vector). Introducing DNA into eukaryotic cells, such as animal cells, by physical or chemical means is called transfection. Several different transfection techniques are available, such as calcium phosphate transfection, electroporation, microinjection and liposome transfection. DNA can also be introduced into eukaryotic cells using viruses or bacteria as carriers, the latter is sometimes called bactofection and in particular uses Agrobacterium tumefaciens. The plasmid may be integrated into the genome, resulting in a stable transfection, or may remain independent of the genome, called transient transfection.

In either case, DNA coding for a protein of interest is now inside a cell, and the protein can now be expressed. A variety of systems, such as inducible promoters and specific cell-signaling factors, are available to help express the protein of interest at high levels. Large quantities of a protein can then be extracted from the bacterial or eukaryotic cell. The protein can be tested for enzymatic activity under a variety of situations, the protein may be crystallized so its tertiary structure can be studied, or, in the pharmaceutical industry, the activity of new drugs against the protein can be studied.

Polymerase chain reaction (PCR)

The polymerase chain reaction is an extremely versatile technique for copying DNA. In brief, PCR allows a single DNA sequence to be copied (millions of times), or altered in predetermined ways. For example, PCR can be used to introduce restriction enzyme sites, or to mutate (change) particular bases of DNA, the latter is a method referred to as "Quick change". PCR can also be used to determine whether a particular DNA fragment is found in a cDNA library. PCR has many variations, like reverse transcription PCR (RT-PCR) for amplification of RNA, and, more recently, real-time PCR (QPCR) which allow for quantitative measurement of DNA or RNA molecules.

Gel electrophoresis

Gel electrophoresis is one of the principal tools of molecular biology. The basic principle is that DNA, RNA, and proteins can all be separated by means of an electric field. In agarose gel electrophoresis, DNA and RNA can be separated on the basis of size by running the DNA through an agarose gel. Proteins can be separated on the basis of size by using an SDS-PAGE gel, or on the basis of size and their electric charge by using what is known as a 2D gel electrophoresis.

Macromolecule blotting and probing

The terms *northern*, *western* and *eastern* blotting are derived from what initially was a molecular biology joke that played on the term *Southern blotting*, after the technique described by Edwin Southern for the hybridisation of blotted DNA. Patricia Thomas, developer of the RNA blot which then became known as the *northern blot* actually didn't use the term. Further combinations of these techniques produced such terms as *southwesterns* (protein-DNA hybridizations), *northwesterns* (to detect protein-RNA interactions) and *farwesterns* (protein-protein interactions), all of which are presently found in the literature.

Southern blotting

Named after its inventor, biologist Edwin Southern, the Southern blot is a method for probing for the presence of a specific DNA sequence within a DNA sample. DNA samples before or after restriction enzyme digestion are separated by gel electrophoresis and then transferred to a membrane by blotting via capillary action. The membrane is then exposed to a labeled DNA probe that has a complement base sequence to the

sequence on the DNA of interest. Most original protocols used radioactive labels, however non-radioactive alternatives are now available. Southern blotting is less commonly used in laboratory science due to the capacity of other techniques, such as PCR, to detect specific DNA sequences from DNA samples. These blots are still used for some applications, however, such as measuring transgene copy number in transgenic mice, or in the engineering of gene knockout embryonic stem cell lines.

Northern blotting

The northern blot is used to study the expression patterns of a specific type of RNA molecule as relative comparison among a set of different samples of RNA. It is essentially a combination of denaturing RNA gel electrophoresis, and a blot. In this process RNA is separated based on size and is then transferred to a membrane that is then probed with a labeled complement of a sequence of interest. The results may be visualized through a variety of ways depending on the label used; however, most result in the revelation of bands representing the sizes of the RNA detected in sample. The intensity of these bands is related to the amount of the target RNA in the samples analyzed. The procedure is commonly used to study when and how much gene expression is occurring by measuring how much of that RNA is present in different samples. It is one of the most basic tools for determining at what time, and under what conditions, certain genes are expressed in living tissues.

Western blotting

Antibodies to most proteins can be created by injecting small amounts of the protein into an animal such as a mouse, rabbit, sheep, or donkey (polyclonal antibodies)or produced in cell culture (monoclonal antibodies). These antibodies can be used for a variety of analytical and preparative techniques.

In western blotting, proteins are first separated by size, in a thin gel sandwiched between two glass plates in a technique known as SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis). The proteins in the gel are then transferred to a PVDF, nitrocellulose, nylon or other support membrane. This membrane can then be probed with solutions of antibodies. Antibodies that specifically bind to the protein of interest can then be visualized by a variety of techniques, including colored products, chemiluminescence, or autoradiography. Often, the antibodies are labeled with enzymes. When a chemiluminescent substrate is exposed to the enzyme it allows detection. Using western blotting techniques allows not only detection but also quantitative analysis.

Analogous methods to western blotting can be used to directly stain specific proteins in live cells or tissue sections. However, these *immunostaining* methods, such as FISH, are used more often in cell biology research.

Eastern blotting

Eastern blotting technique is to detect post-translational modification of proteins. Proteins blotted on to the PVDF or nitrocellulose membrane are probed for modifications using specific substrates.

Arrays

A DNA array is a collection of spots attached to a solid support such as a microscope slide where each spot contains one or more single-stranded DNA oligonucleotide fragment. Arrays make it possible to put down a large quantity of very small (100 micrometre diameter) spots on a single slide. Each spot has a DNA fragment molecule that is complementary to a single DNA sequence (similar to Southern blotting). A variation of this technique allows the gene expression of an organism at a particular stage in development to be qualified (expression profiling). In this technique the RNA in a tissue is isolated and converted to labeled cDNA. This cDNA is then hybridized to the fragments on the array and visualization of the hybridization can be done. Since multiple arrays can be made with exactly the same position of fragments they are particularly useful for comparing the gene expression of two different tissues, such as a healthy and cancerous tissue. Also, one can measure what genes are expressed and how that expression changes with time or with other factors. For instance, the common baker's yeast, *Saccharomyces cerevisiae*, contains about 7000 genes; with a microarray, one can measure qualitatively how each gene is expressed, and how that expression changes, for example, with a change in temperature. There are many different ways to fabricate microarrays; the most common are silicon chips, microscope slides with spots of ~ 100 micrometre diameter, custom arrays, and arrays with larger spots on porous membranes (macroarrays). There can be anywhere from 100 spots to more than 10,000 on a given array.

Arrays can also be made with molecules other than DNA. For example, an antibody array can be used to determine what proteins or bacteria are present in a blood sample.

Allele Specific Oligonucleotide

Allele specific oligonucleotide (ASO) is a technique that allows detection of single base mutations without the need for PCR or gel electrophoresis. Short (20-25 nucleotides in length), labeled probes are exposed to the non-fragmented target DNA. Hybridization occurs with high specificity due to the short length of the probes and even a single base change will hinder hybridization. The target DNA is then washed and the labeled probes that didn't hybridize are removed. The target DNA is then analyzed for the presence of the probe via radioactivity or fluorescence. In this experiment, as in most molecular biology techniques, a control must be used to ensure successful experimentation. The Illumina Methylation Assay is an example of a method that takes advantage of the ASO technique to measure one base pair differences in sequence.

Antiquated technologies

In molecular biology, procedures and technologies are continually being developed and older technologies abandoned. For example, before the advent of DNA gel electrophoresis (agarose or polyacrylamide), the size of DNA molecules was typically determined by rate sedimentation in sucrose gradients, a slow and labor-intensive technique requiring expensive instrumentation; prior to sucrose gradients, viscometry was used.

Aside from their historical interest, it is often worth knowing about older technology, as it is occasionally useful to solve another new problem for which the newer technique is inappropriate.

History

While molecular biology was established in the 1930s, the term was first coined by Warren Weaver in 1938. Warren was the director of Natural Sciences for the Rockefeller Foundation at the time and believed that biology was about to undergo a period of significant change given recent advances in fields such as X-ray crystallography. He therefore channeled significant amounts of (Rockefeller Institute) money into biological fields.

Clinical significance

Clinical research and medical therapies arising from molecular biology are covered under gene therapy

Chapter- 10

Biochemistry

Biochemistry, sometimes called **biological chemistry**, is the study of chemical processes in living organisms. Biochemistry governs all living organisms and living processes. By controlling information flow through biochemical signalling and the flow of chemical energy through metabolism, biochemical processes give rise to the incredible complexity of life. Much of biochemistry deals with the structures and functions of cellular components such as proteins, carbohydrates, lipids, nucleic acids and other biomolecules although increasingly processes rather than individual molecules are the main focus. Over the last 40 years biochemistry has become so successful at explaining living processes that now almost all areas of the life sciences from botany to medicine are engaged in biochemical research. Today the main focus of pure biochemistry is in understanding how biological molecules give rise to the processes that occur within living cells which in turn relates greatly to the study and understanding of whole organisms.

Among the vast number of different biomolecules, many are complex and large molecules (called *biopolymers*), which are composed of similar repeating subunits (called *monomers*). Each class of polymeric biomolecule has a different set of subunit types. For example, a protein is a polymer whose subunits are selected from a set of 20 or more amino acids. Biochemistry studies the chemical properties of important biological molecules, like proteins, and in particular the chemistry of enzyme-catalyzed reactions.

The biochemistry of cell metabolism and the endocrine system has been extensively described. Other areas of biochemistry include the genetic code (DNA, RNA), protein synthesis, cell membrane transport, and signal transduction.

History

Originally, it was generally believed that life was not subject to the laws of science the way non-life was. It was thought that only living beings could produce the molecules of life (from other, previously existing biomolecules). Then, in 1828, Friedrich Wöhler

published a paper on the synthesis of urea, proving that organic compounds can be created artificially.

The dawn of biochemistry may have been the discovery of the first enzyme, diastase (today called amylase), in 1833 by Anselme Payen. Eduard Buchner contributed the first demonstration of a complex biochemical process outside of a cell in 1896: alcoholic fermentation in cell extracts of yeast. Although the term "biochemistry" seems to have been first used in 1882, it is generally accepted that the formal coinage of biochemistry occurred in 1903 by Carl Neuberg, a German chemist. Previously, this area would have been referred to as physiological chemistry. Since then, biochemistry has advanced, especially since the mid-20th century, with the development of new techniques such as chromatography, X-ray diffraction, dual polarisation interferometry, NMR spectroscopy, radioisotopic labeling, electron microscopy and molecular dynamics simulations. These techniques allowed for the discovery and detailed analysis of many molecules and metabolic pathways of the cell, such as glycolysis and the Krebs cycle (citric acid cycle).

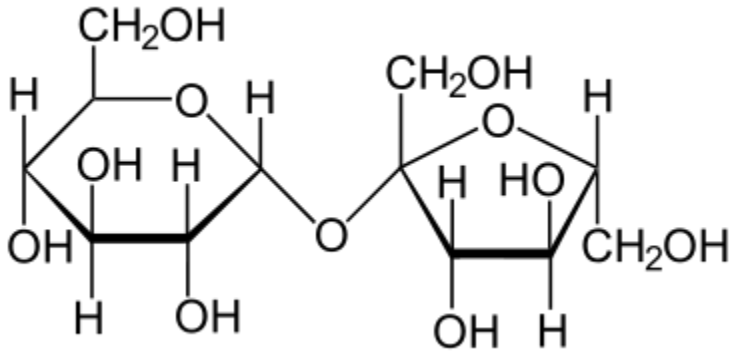
Another significant historic event in biochemistry is the discovery of the gene and its role in the transfer of information in the cell. This part of biochemistry is often called molecular biology. In the 1950s, James D. Watson, Francis Crick, Rosalind Franklin, and Maurice Wilkins were instrumental in solving DNA structure and suggesting its relationship with genetic transfer of information. In 1958, George Beadle and Edward Tatum received the Nobel Prize for work in fungi showing that one gene produces one enzyme. In 1988, Colin Pitchfork was the first person convicted of murder with DNA evidence, which led to growth of forensic science. More recently, Andrew Z. Fire and Craig C. Mello received the 2006 Nobel Prize for discovering the role of RNA interference (RNAi), in the silencing of gene expression.

Today, there are three main types of biochemistry. Plant biochemistry involves the study of the biochemistry of autotrophic organisms such as photosynthesis and other plant specific biochemical processes. General biochemistry encompasses both plant and animal biochemistry. Human/medical/medicinal biochemistry focuses on the biochemistry of humans and medical illnesses.

Biomolecules

The four main classes of molecules in biochemistry are carbohydrates, lipids, proteins, and nucleic acids. Many biological molecules are polymers: in this terminology, ***monomers*** are relatively small micromolecules that are linked together to create large macromolecules, which are known as ***polymers***. When monomers are linked together to synthesize a biological polymer, they undergo a process called dehydration synthesis.

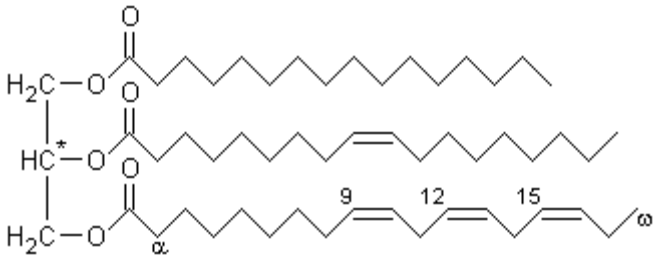
Carbohydrates



A molecule of sucrose (glucose + fructose), a disaccharide.

Carbohydrates are made from monomers called *monosaccharides*. Some of these monosaccharides include glucose ($C_6H_{12}O_6$), fructose ($C_6H_{12}O_6$), and deoxyribose ($C_5H_{10}O_4$). When two monosaccharides undergo dehydration synthesis, water is produced, as two hydrogen atoms and one oxygen atom are lost from the two monosaccharides' hydroxyl group.

Lipids

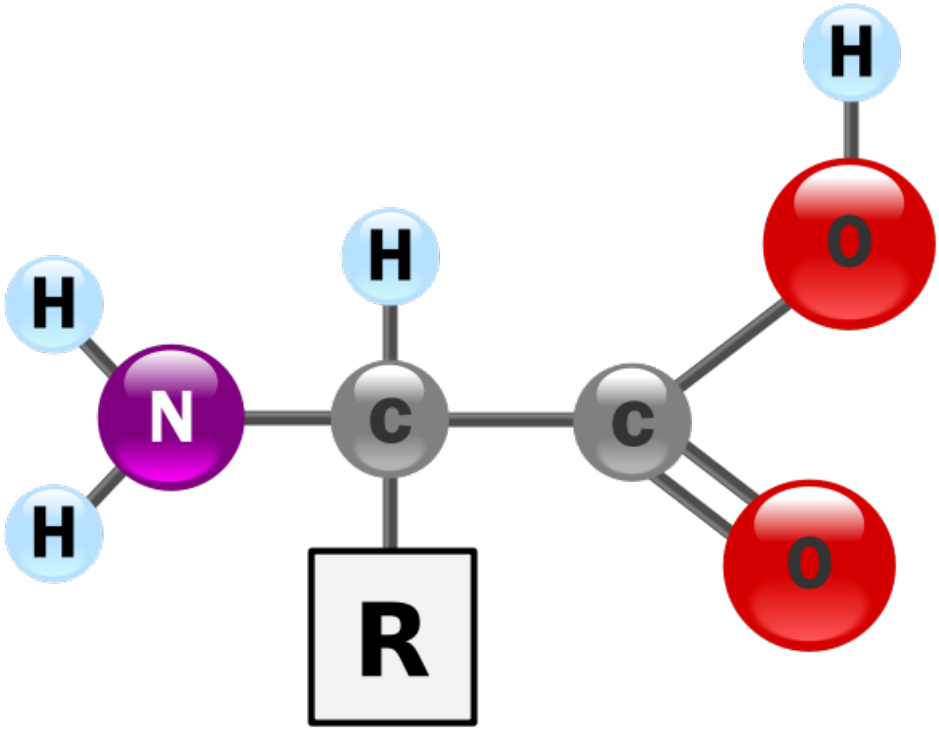


A triglyceride with a glycerol molecule on the left and three fatty acids coming off it.

Lipids are usually made from one molecule of glycerol combined with other molecules. In triglycerides, the main group of bulk lipids, there is one molecule of glycerol and three fatty acids. Fatty acids are considered the monomer in that case, and may be saturated (no double bonds in the carbon chain) or unsaturated (one or more double bonds in the carbon chain).

Lipids, especially phospholipids, are also used in various pharmaceutical products, either as co-solubilisers (e.g. in parenteral infusions) or else as drug carrier components (e.g. in a liposome or transfersome).

Proteins

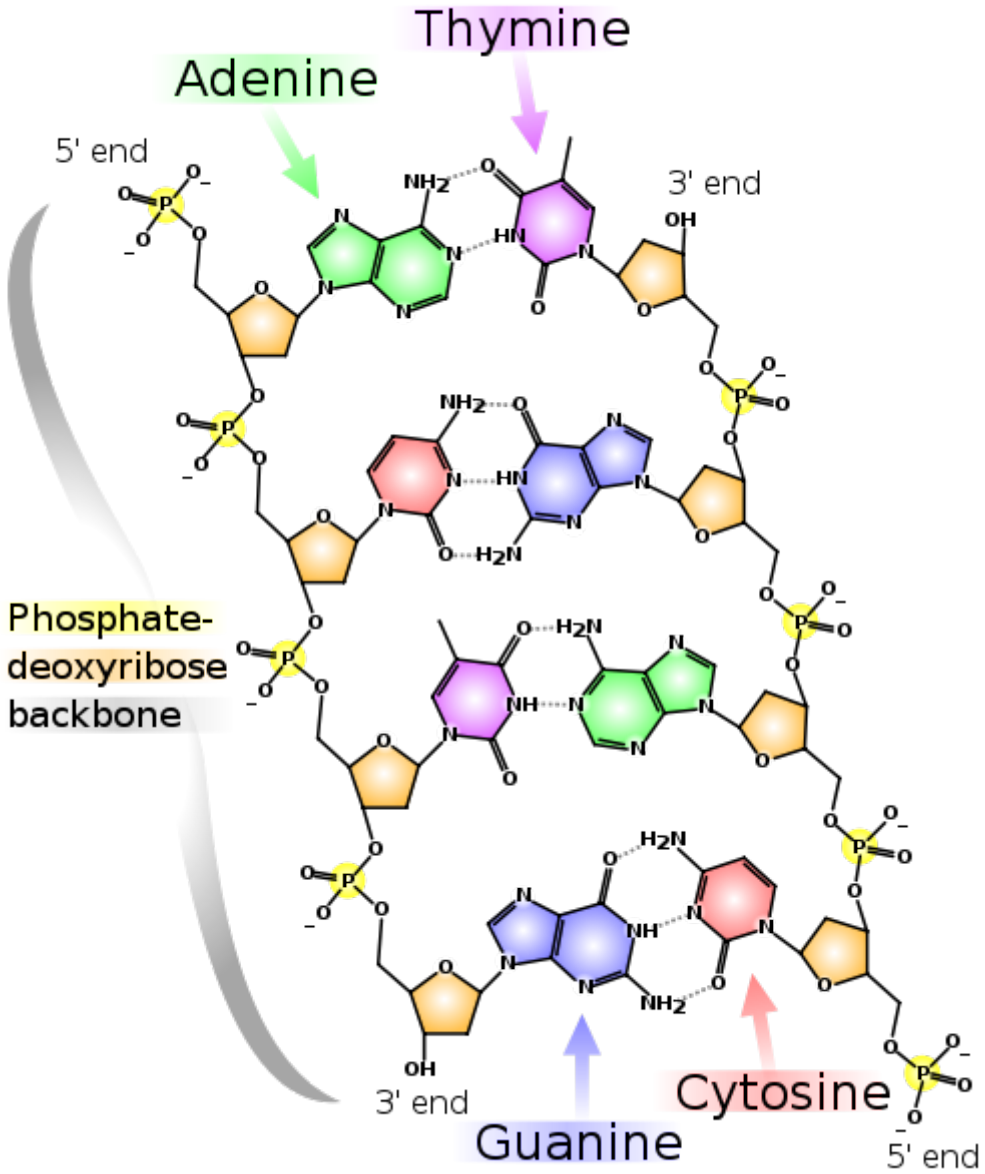


The general structure of an α-amino acid, with the amino group on the left and the carboxyl group on the right.

Proteins are very large molecules – macro-biopolymers – made from monomers called *amino acids*. There are 20 standard amino acids, each containing a carboxyl group, an amino group, and a side chain (known as an "R" group). The "R" group is what makes each amino acid different, and the properties of the side chains greatly influence the overall three-dimensional conformation of a protein. When amino acids combine, they form a special bond called a peptide bond through dehydration synthesis, and become a ***polypeptide***, or protein.

To determine if two proteins are related or in other words to decide whether they are homologous or not, scientists use sequence-comparison methods. Methods like Sequence Alignments and Structural Alignments are powerful tools that help scientist identify homologies between related molecules. The relevance of finding homologies among proteins goes beyond forming an evolutionary pattern of protein families. By finding how similar two protein sequences are, we acquire knowledge about their structure and therefore their function.

Nucleic acids



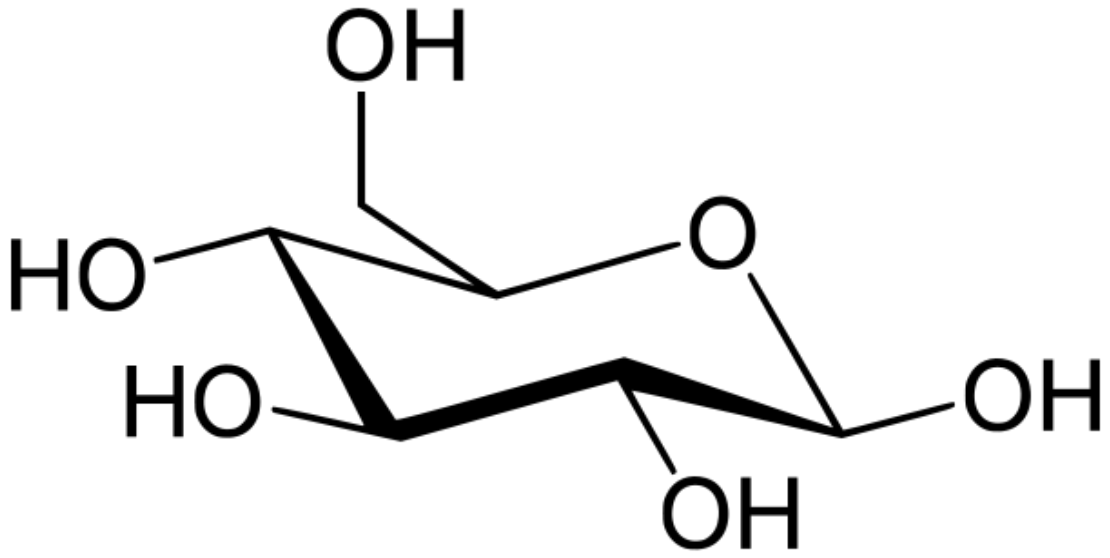
The structure of deoxyribonucleic acid (DNA), the picture shows the monomers being put together.

Nucleic acids are the molecules that make up DNA, an extremely important substance which all cellular organisms use to store their genetic information. The most common nucleic acids are deoxyribonucleic acid and ribonucleic acid. Their monomers are called nucleotides. The most common nucleotides are Adenine, Cytosine, Guanine, Thymine, and Uracil. Adenine binds with thymine and uracil; Thymine only binds with Adenine; and Cytosine and Guanine can only bind with each other.

Carbohydrates

The function of carbohydrates includes energy storage and providing structure. Sugars are carbohydrates, but not all carbohydrates are sugars. There are more carbohydrates on Earth than any other known type of biomolecule; they are used to store energy and genetic information, as well as play important roles in cell to cell interactions and communications.

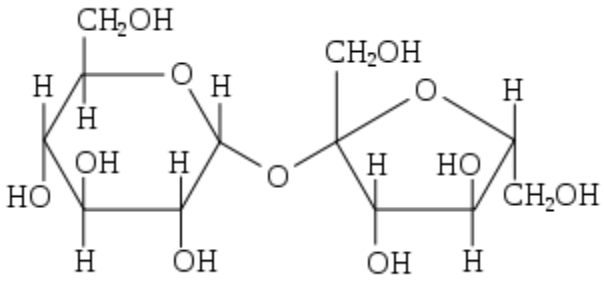
Monosaccharides



Glucose

The simplest type of carbohydrate is a monosaccharide, which among other properties contains carbon, hydrogen, and oxygen, mostly in a ratio of 1:2:1 (generalized formula $C_nH_{2n}O_n$, where n is at least 3). Glucose, one of the most important carbohydrates, is an example of a monosaccharide. So is fructose, the sugar commonly associated with the sweet taste of fruits. Some carbohydrates (especially after condensation to oligo- and polysaccharides) contain less carbon relative to H and O, which still are present in 2:1 (H:O) ratio. Monosaccharides can be grouped into aldoses (having an aldehyde group at the end of the chain, e. g. glucose) and ketoses (having a keto group in their chain; e. g. fructose). Both aldoses and ketoses occur in an equilibrium (starting with chain lengths of C4) cyclic forms. These are generated by bond formation between one of the hydroxyl groups of the sugar chain with the carbon of the aldehyde or keto group to form a hemiacetal bond. This leads to saturated five-membered (in furanoses) or six-membered (in pyranoses) heterocyclic rings containing one O as heteroatom.

Disaccharides

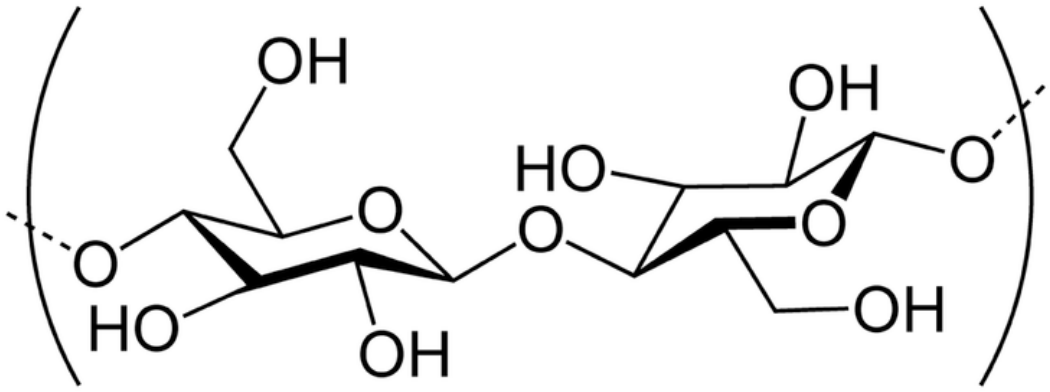


Sucrose: ordinary table sugar and probably the most familiar carbohydrate.

Two monosaccharides can be joined together using dehydration synthesis, in which a hydrogen atom is removed from the end of one molecule and a hydroxyl group (—OH) is removed from the other; the remaining residues are then attached at the sites from which the atoms were removed. The H—OH or H_2O is then released as a molecule of water, hence the term *dehydration*. The new molecule, consisting of two monosaccharides, is called a *disaccharide* and is conjoined together by a glycosidic or ether bond. The reverse reaction can also occur, using a molecule of water to split up a disaccharide and break the glycosidic bond; this is termed *hydrolysis*. The most well-known disaccharide is sucrose, ordinary sugar (in scientific contexts, called *table sugar* or *cane sugar* to differentiate it from other sugars). Sucrose consists of a glucose molecule and a fructose molecule joined together. Another important disaccharide is lactose, consisting of a glucose molecule and a galactose molecule. As most humans age, the production of lactase, the enzyme that hydrolyzes lactose back into glucose and galactose, typically decreases. This results in lactase deficiency, also called *lactose intolerance*.

Sugar polymers are characterised by having reducing or non-reducing ends. A reducing end of a carbohydrate is a carbon atom which can be in equilibrium with the open-chain aldehyde or keto form. If the joining of monomers takes place at such a carbon atom, the free hydroxy group of the pyranose or furanose form is exchanged with an OH-side chain of another sugar, yielding a full acetal. This prevents opening of the chain to the aldehyde or keto form and renders the modified residue non-reducing. Lactose contains a reducing end at its glucose moiety, whereas the galactose moiety form a full acetal with the C4-OH group of glucose. Saccharose does not have a reducing end because of full acetal formation between the aldehyde carbon of glucose (C1) and the keto carbon of fructose (C2).

Oligosaccharides and polysaccharides



Cellulose as polymer of β-D-glucose

When a few (around three to six) monosaccharides are joined together, it is called an *oligosaccharide* (*oligo-* meaning "few"). These molecules tend to be used as markers and signals, as well as having some other uses. Many monosaccharides joined together make a polysaccharide. They can be joined together in one long linear chain, or they may be branched. Two of the most common polysaccharides are cellulose and glycogen, both consisting of repeating glucose monomers.

- *Cellulose* is made by plants and is an important structural component of their cell walls. Humans can neither manufacture nor digest it.
- *Glycogen*, on the other hand, is an animal carbohydrate; humans and other animals use it as a form of energy storage.

Use of carbohydrates as an energy source

Glucose is the major energy source in most life forms. For instance, polysaccharides are broken down into their monomers (glycogen phosphorylase removes glucose residues from glycogen). Disaccharides like lactose or sucrose are cleaved into their two component monosaccharides.

Glycolysis (anaerobic)

Glucose is mainly metabolized by a very important ten-step pathway called glycolysis, the net result of which is to break down one molecule of glucose into two molecules of pyruvate; this also produces a net two molecules of ATP, the energy currency of cells, along with two reducing equivalents in the form of converting NAD^+ to NADH. This does not require oxygen; if no oxygen is available (or the cell cannot use oxygen), the NAD is restored by converting the pyruvate to lactate (lactic acid) (e. g. in humans) or to

ethanol plus carbon dioxide (e. g. in yeast). Other monosaccharides like galactose and fructose can be converted into intermediates of the glycolytic pathway.

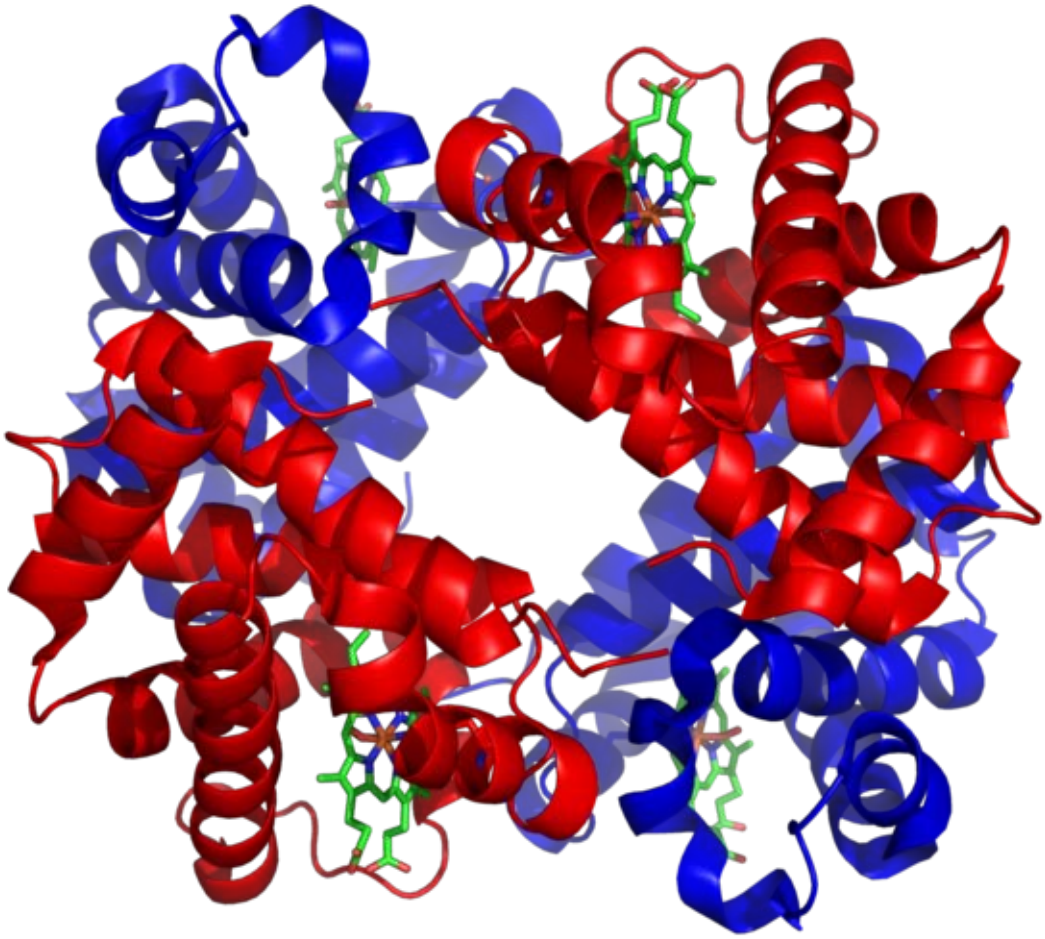
Aerobic

In aerobic cells with sufficient oxygen, like most human cells, the pyruvate is further metabolized. It is irreversibly converted to acetyl-CoA, giving off one carbon atom as the waste product carbon dioxide, generating another reducing equivalent as NADH. The two molecules acetyl-CoA (from one molecule of glucose) then enter the citric acid cycle, producing two more molecules of ATP, six more NADH molecules and two reduced (ubi)quinones (via $FADH_2$ as enzyme-bound cofactor), and releasing the remaining carbon atoms as carbon dioxide. The produced NADH and quinol molecules then feed into the enzyme complexes of the respiratory chain, an electron transport system transferring the electrons ultimately to oxygen and conserving the released energy in the form of a proton gradient over a membrane (inner mitochondrial membrane in eukaryotes). Thereby, oxygen is reduced to water and the original electron acceptors NAD^+ and quinone are regenerated. This is why humans breathe in oxygen and breathe out carbon dioxide. The energy released from transferring the electrons from high-energy states in NADH and quinol is conserved first as proton gradient and converted to ATP via ATP synthase. This generates an additional *28* molecules of ATP (24 from the 8 NADH + 4 from the 2 quinols), totaling to 32 molecules of ATP conserved per degraded glucose (two from glycolysis + two from the citrate cycle). It is clear that using oxygen to completely oxidize glucose provides an organism with far more energy than any oxygen-independent metabolic feature, and this is thought to be the reason why complex life appeared only after Earth's atmosphere accumulated large amounts of oxygen.

Gluconeogenesis

In vertebrates, vigorously contracting skeletal muscles (during weightlifting or sprinting, for example) do not receive enough oxygen to meet the energy demand, and so they shift to anaerobic metabolism, converting glucose to lactate. The liver regenerates the glucose, using a process called gluconeogenesis. This process is not quite the opposite of glycolysis, and actually requires three times the amount of energy gained from glycolysis (six molecules of ATP are used, compared to the two gained in glycolysis). Analogous to the above reactions, the glucose produced can then undergo glycolysis in tissues that need energy, be stored as glycogen (or starch in plants), or be converted to other monosaccharides or joined into di- or oligosaccharides. The combined pathways of glycolysis during exercise, lactate's crossing via the bloodstream to the liver, subsequent gluconeogenesis and release of glucose into the bloodstream is called the Cori cycle.

Proteins

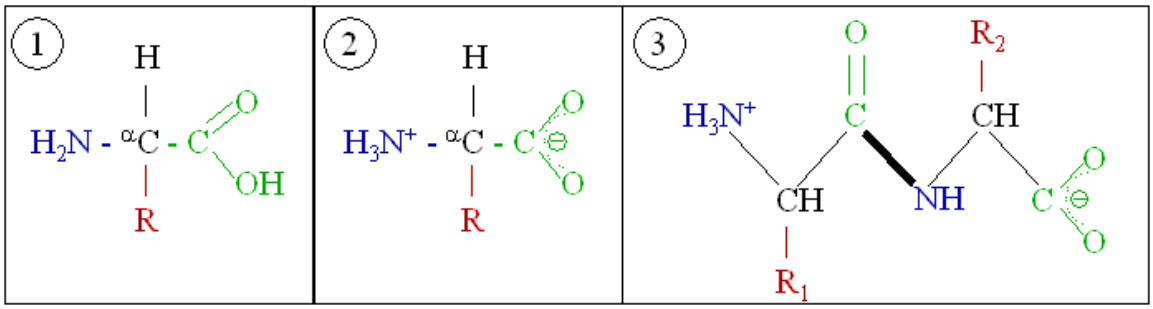


A schematic of hemoglobin. The red and blue ribbons represent the protein globin; the green structures are the heme groups.

Like carbohydrates, some proteins perform largely structural roles. For instance, movements of the proteins actin and myosin ultimately are responsible for the contraction of skeletal muscle. One property many proteins have is that they specifically bind to a certain molecule or class of molecules—they may be *extremely* selective in what they bind. Antibodies are an example of proteins that attach to one specific type of molecule. In fact, the enzyme-linked immunosorbent assay (ELISA), which uses antibodies, is currently one of the most sensitive tests modern medicine uses to detect various biomolecules. Probably the most important proteins, however, are the enzymes. These molecules recognize specific reactant molecules called *substrates*; they then catalyze the reaction between them. By lowering the activation energy, the enzyme speeds up that

reaction by a rate of 10^{11} or more: a reaction that would normally take over 3,000 years to complete spontaneously might take less than a second with an enzyme. The enzyme itself is not used up in the process, and is free to catalyze the same reaction with a new set of substrates. Using various modifiers, the activity of the enzyme can be regulated, enabling control of the biochemistry of the cell as a whole.

In essence, proteins are chains of amino acids. An amino acid consists of a carbon atom bound to four groups. One is an amino group, —NH_2, and one is a carboxylic acid group, —COOH (although these exist as —NH_3^+ and —COO^- under physiologic conditions). The third is a simple hydrogen atom. The fourth is commonly denoted "—R" and is different for each amino acid. There are twenty standard amino acids. Some of these have functions by themselves or in a modified form; for instance, glutamate functions as an important neurotransmitter.



Generic amino acids (1) in neutral form, (2) as they exist physiologically, and (3) joined together as a dipeptide.

Amino acids can be joined together via a peptide bond. In this dehydration synthesis, a water molecule is removed and the peptide bond connects the nitrogen of one amino acid's amino group to the carbon of the other's carboxylic acid group. The resulting molecule is called a *dipeptide*, and short stretches of amino acids (usually, fewer than around thirty) are called *peptides* or polypeptides. Longer stretches merit the title *proteins*. As an example, the important blood serum protein albumin contains 585 amino acid residues.

The structure of proteins is traditionally described in a hierarchy of four levels. The primary structure of a protein simply consists of its linear sequence of amino acids; for instance, "alanine-glycine-tryptophan-serine-glutamate-asparagine-glycine-lysine-…". Secondary structure is concerned with local morphology (morphology being the study of structure). Some combinations of amino acids will tend to curl up in a coil called an α-helix or into a sheet called a β-sheet; some α-helixes can be seen in the hemoglobin schematic above. Tertiary structure is the entire three-dimensional shape of the protein. This shape is determined by the sequence of amino acids. In fact, a single change can change the entire structure. The alpha chain of hemoglobin contains 146 amino acid residues; substitution of the glutamate residue at position 6 with a valine residue changes the behavior of hemoglobin so much that it results in sickle-cell disease. Finally quaternary structure is concerned with the structure of a protein with multiple peptide

subunits, like hemoglobin with its four subunits. Not all proteins have more than one subunit.

Ingested proteins are usually broken up into single amino acids or dipeptides in the small intestine, and then absorbed. They can then be joined together to make new proteins. Intermediate products of glycolysis, the citric acid cycle, and the pentose phosphate pathway can be used to make all twenty amino acids, and most bacteria and plants possess all the necessary enzymes to synthesize them. Humans and other mammals, however, can only synthesize half of them. They cannot synthesize isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine. These are the essential amino acids, since it is essential to ingest them. Mammals do possess the enzymes to synthesize alanine, asparagine, aspartate, cysteine, glutamate, glutamine, glycine, proline, serine, and tyrosine, the nonessential amino acids. While they can synthesize arginine and histidine, they cannot produce it in sufficient amounts for young, growing animals, and so these are often considered essential amino acids.

If the amino group is removed from an amino acid, it leaves behind a carbon skeleton called an α-keto acid. Enzymes called transaminases can easily transfer the amino group from one amino acid (making it an α-keto acid) to another α-keto acid (making it an amino acid). This is important in the biosynthesis of amino acids, as for many of the pathways, intermediates from other biochemical pathways are converted to the α-keto acid skeleton, and then an amino group is added, often via transamination. The amino acids may then be linked together to make a protein.

A similar process is used to break down proteins. It is first hydrolyzed into its component amino acids. Free ammonia (NH_3), existing as the ammonium ion (NH_4^+) in blood, is toxic to life forms. A suitable method for excreting it must therefore exist. Different strategies have evolved in different animals, depending on the animals' needs. Unicellular organisms, of course, simply release the ammonia into the environment. Similarly, bony fish can release the ammonia into the water where it is quickly diluted. In general, mammals convert the ammonia into urea, via the urea cycle.

Lipids

The term lipid comprises a diverse range of molecules and to some extent is a catchall for relatively water-insoluble or nonpolar compounds of biological origin, including waxes, fatty acids, fatty-acid derived phospholipids, sphingolipids, glycolipids and terpenoids (e.g. retinoids and steroids). Some lipids are linear aliphatic molecules, while others have ring structures. Some are aromatic, while others are not. Some are flexible, while others are rigid.

Most lipids have some polar character in addition to being largely nonpolar. Generally, the bulk of their structure is nonpolar or hydrophobic ("water-fearing"), meaning that it does not interact well with polar solvents like water. Another part of their structure is polar or hydrophilic ("water-loving") and will tend to associate with polar solvents like water. This makes them amphiphilic molecules (having both hydrophobic and

hydrophilic portions). In the case of cholesterol, the polar group is a mere -OH (hydroxyl or alcohol). In the case of phospholipids, the polar groups are considerably larger and more polar, as described below.

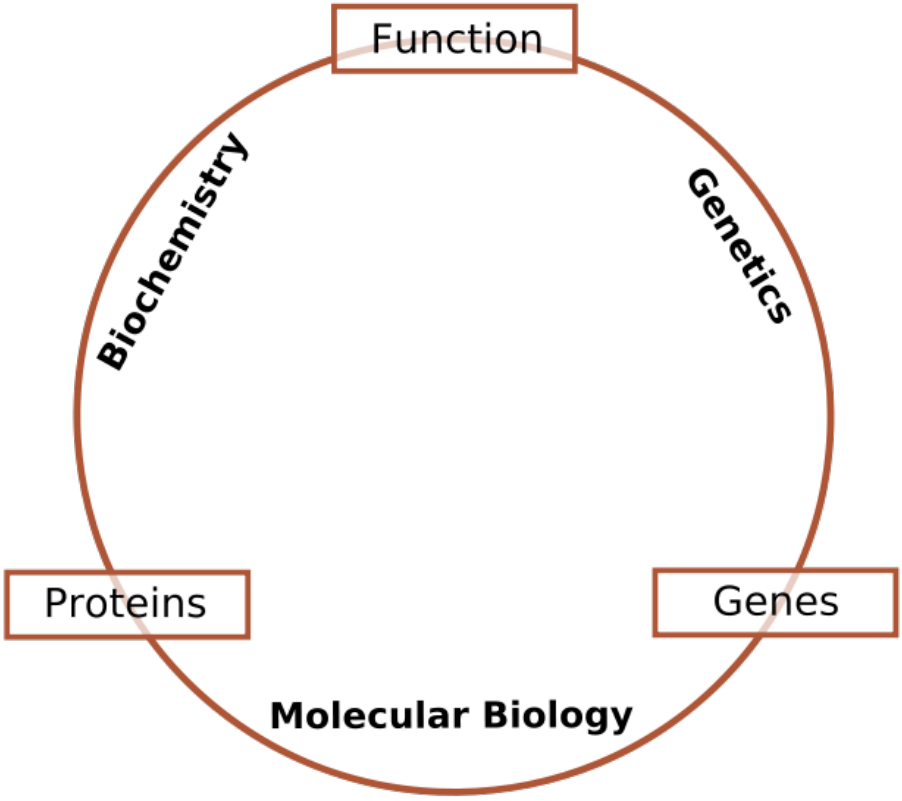
Lipids are an integral part of our daily diet. Most oils and milk products that we use for cooking and eating like butter, cheese, ghee etc., are composed of fats. Vegetable oils are rich in various polyunsaturated fatty acids (PUFA). Lipid-containing foods undergo digestion within the body and are broken into fatty acids and glycerol, which are the final degradation products of fats and lipids.

Nucleic acids

A nucleic acid is a complex, high-molecular-weight biochemical macromolecule composed of nucleotide chains that convey genetic information. The most common nucleic acids are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Nucleic acids are found in all living cells and viruses. Aside from the genetic material of the cell, nucleic acids often play a role as second messengers, as well as forming the base molecule for adenosine triphosphate, the primary energy-carrier molecule found in all living organisms.

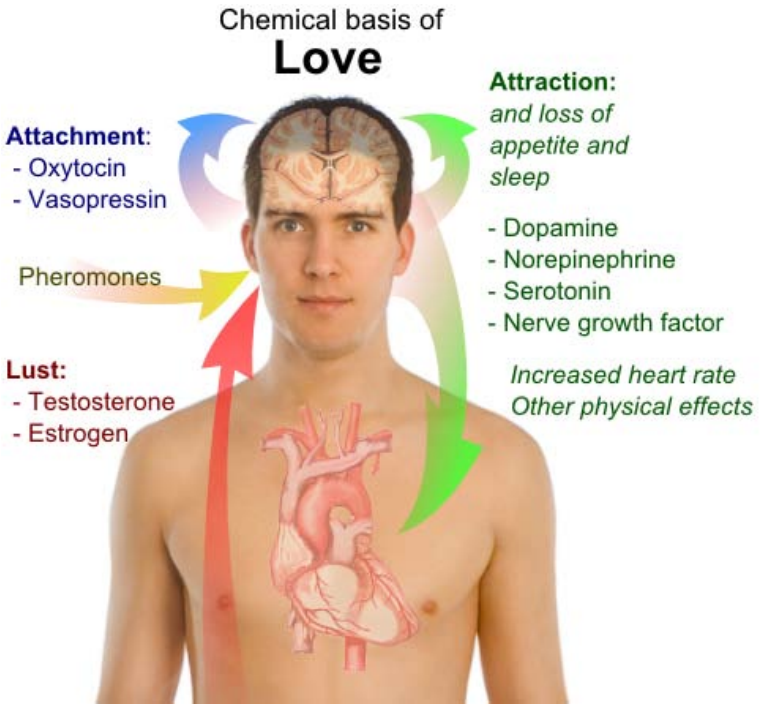
Nucleic acid, so called because of its prevalence in cellular nuclei, is the generic name of the family of biopolymers. The monomers are called nucleotides, and each consists of three components: a nitrogenous heterocyclic base (either a purine or a pyrimidine), a pentose sugar, and a phosphate group. Different nucleic acid types differ in the specific sugar found in their chain (e.g. DNA or deoxyribonucleic acid contains 2-deoxyriboses). Also, the nitrogenous bases possible in the two nucleic acids are different: adenine, cytosine, and guanine occur in both RNA and DNA, while thymine occurs only in DNA and uracil occurs in RNA.

Relationship to other "molecular-scale" biological sciences



Schematic relationship between biochemistry, genetics and molecular biology

Researchers in biochemistry use specific techniques native to biochemistry, but increasingly combine these with techniques and ideas from genetics, molecular biology and biophysics. There has never been a hard-line between these disciplines in terms of content and technique. Today the terms *molecular biology* and *biochemistry* are nearly interchangeable. The following figure is a schematic that depicts one possible view of the relationship between the fields:



Simplistic overview of the chemical basis of love, one of many applications that may be described in terms of biochemistry.

- *Biochemistry* is the study of the chemical substances and vital processes occurring in living organisms. Biochemists focus heavily on the role, function, and structure of biomolecules. The study of the chemistry behind biological processes and the synthesis of biologically active molecules are examples of biochemistry.
- *Genetics* is the study of the effect of genetic differences on organisms. Often this can be inferred by the absence of a normal component (e.g. one gene). The study of "mutants" – organisms which lack one or more functional components with respect to the so-called "wild type" or normal phenotype. Genetic interactions (epistasis) can often confound simple interpretations of such "knock-out" studies.
- *Molecular biology* is the study of molecular underpinnings of the process of replication, transcription and translation of the genetic material. The central dogma of molecular biology where genetic material is transcribed into RNA and then translated into protein, despite being an oversimplified picture of molecular biology, still provides a good starting point for understanding the field. This picture, however, is undergoing revision in light of emerging novel roles for RNA.
- *Chemical Biology* seeks to develop new tools based on small molecules that allow minimal perturbation of biological systems while providing detailed information about their function. Further, chemical biology employs biological systems to create non-natural hybrids between biomolecules and synthetic devices (for example emptied viral capsids that can deliver gene therapy or drug molecules).