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

SCHOOL OF BIO AND CHEMICAL ENGINEERING

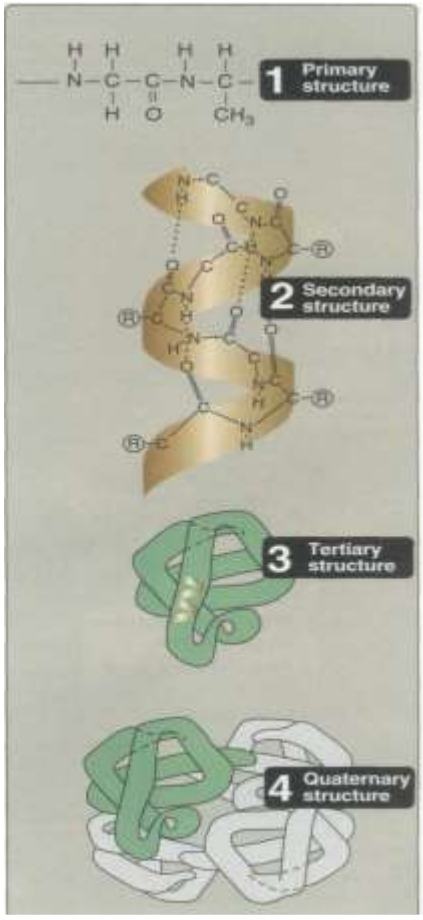
DEPARTMENT OF BIOTECHNOLOGY

UNIT – I –ENZYME AND PROTEIN ENGINEERING– SBTA5202

I. Protein Structure

The twenty amino acids commonly found in proteins are joined together by peptide bonds. The linear sequence of the linked amino acids contains the information necessary to generate a protein molecule with a unique three-dimensional shape. The complexity of protein structure is best analyzed by considering the molecule in terms of four organizational levels, namely, primary, secondary, tertiary, and quaternary.

1.1. PRIMARY STRUCTURE OF PROTEINS



The sequence of amino acids in a protein is called the primary structure of the protein. Understanding the primary structure of proteins is important because many genetic diseases result in proteins with abnormal amino acid sequences, which cause improper folding and loss or impairment of normal function. If the primary structures of the normal and the mutated proteins are known, this information may be used to diagnose or study the disease.

A. Peptide bond

Proteins, amino acids are joined covalently by peptide bonds, which are amide linkages between the α-carboxyl group of one amino acid, and the α-amino group of another. For example, valine and alanine can form the dipeptide valylalanine through the formation of a peptide bond Peptide bonds are not broken by conditions that denature proteins, such as heating or high concentrations of urea. Prolonged exposure to a strong acid or base at elevated temperatures is required to hydrolyze these bonds nonenzymically.

1. Naming the peptide: By convention, the free amino end of the peptide chain (N-Terminal) is written to the left and the free carboxyl end (C-Terminal)to the right. Therefore, all amino sequences are read from the N- to the C-terminal end of the peptide

2. Characteristics of the peptide bond: The peptide bond has a partial double-bond character that is, it is shorter than a single bond, and is rigid and planar. This prevents free rotation around the bond between the carbonyl carbon and the nitrogen of the peptide bond. However, the bonds between the α-carbons and the α-amino or α-carboxyl groups can be freely rotated (although they are limited by the size and character of the R-groups). This allows the polypeptide chain to assume a variety of possible configurations. The peptide bond is generally a trans bond in large part because of steric interference of the R-groups when in the cis position.

3.Polarity of the peptide bond: Like all amide linkages, the -C= 0 and -N H groups of the peptide bond are uncharged, and neither accept nor release protons over the P^H range of 2 to 12. Thus, the charged groups present in polypeptides consist solely of the N-terminal α-amino group, the C-terminal α-carboxyl group, and any ionized groups present in the side chains of the constituent amino acids.

B. Determination of the amino acid composition of a polypeptide

The first step in determining the primary structure of a polypeptide is to identify and quantitate its constituent amino acids. A purified sample of the polypeptide to be analyzed is first hydrolyzed by strong acid at 110^0C for 24 hours. This treatment cleaves the peptide bonds, and releases the individual amino acids, which can be separated by cation-exchange chromatography. In this technique, a mixture of amino acids is applied to a column that contains a resin to which a negatively charged group is tightly attached. The amino acids bind to the column with different affinities, depending on their charges, hydrophobicity, and other characteristics. Each amino acid is sequentially released from the chromatography column by eluting with solutions of increasing ionic strength and pH. The separated amino acids contained in the elute from the column are quantitated by heating them with Ninhydrin-a reagent that forms a purple compound with most.

C. Sequencing of the peptide from its end

Sequencing is a stepwise process of identifying the specific amino acids at each position in the peptide chain, beginning at the N-Terminal end. Phenylisothiocyanate, known as Edman's reagent, is used to label the amino terminal residue under mildly alkaline conditions. The resulting phenylthiohydantoin (PTH) derivative introduces an instability in the N-terminal peptide bond that can be selectively hydrolyzed without cleaving the other peptide bonds. The identity of the amino acid derivative can then be determined. Edman's reagent can be applied repeatedly to the shortened peptide obtained in each previous cycle. This process has been automated and, currently, the repetition of the method can be employed by a machine (sequenator) to determine the sequence of more than 100 amino acid residues, starting at the amino terminal end of a polypeptide.

D. Cleavage of the polypeptide into smaller fragments

Many polypeptides have a primary structure composed of more than100 amino acids. Such molecules cannot be sequenced directly from end to end by a sequenator. However, these large molecules can be cleaved at specific sites, and the resulting fragments sequenced. By using more than one cleaving agent (enzymes and/or chemicals) on separate samples of the purified polypeptide, overlapping fragments can be generated that permit the proper ordering of the sequenced fragments, thus providing a complete amino acid sequence of the large polypeptide.

E. Determination of a protein's primary structure by DNA sequencing

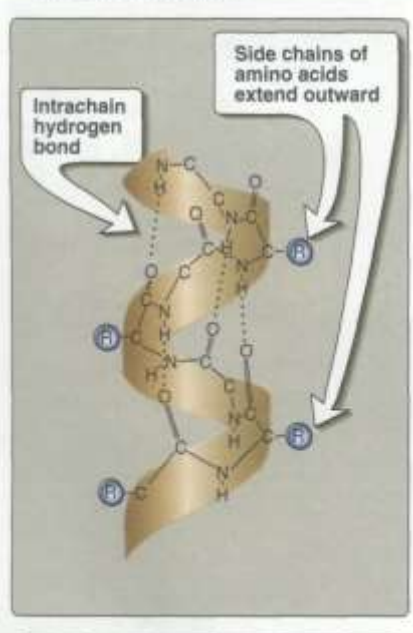
The sequence of nucleotides in a coding region of the DNA specifies the amino acid sequence of a polypeptide. Therefore, if the nucleotide sequence can be determined, it is possible, from knowledge of the genetic code, to translate the sequence of nucleotides into the corresponding amino acid sequence of that polypeptide. This process, although routinely used to obtain the amino acid sequences of proteins, has the limitations of not being able to predict the positions of disulfide bonds in the folded chain, and not identifying any amino acids that are modified after their incorporation into the polypeptide post translational modification,. Therefore, direct protein sequencing is an extremely important tool for determining the true character of the primary sequence of many polypeptides.

1.2. SECONDARY STRUCTURE OF PROTEINS

The polypeptide backbone does not assume a random three-dimensional structure, but instead generally forms regular arrangements of amino acids that are located near to each other in the linear sequence. These arrangements are termed the **secondary structure** of the polypeptide. The α-helix, β-sheet, and β-bend are examples of secondary structures frequently encountered in proteins.

A. α- Helix

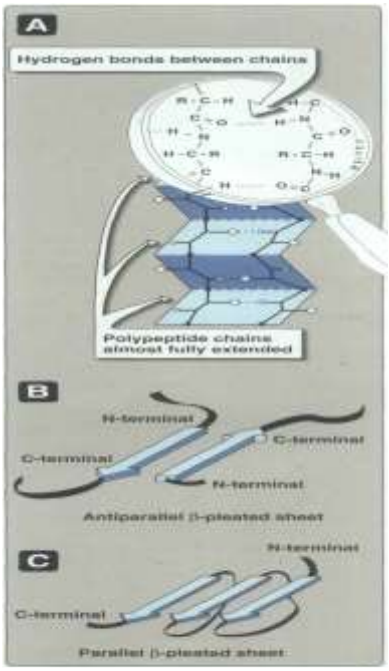
There are several different polypeptide helices found in nature, but the α-helix is the most common. It is a spiral structure, consisting of a tightly packed, coiled polypeptide backbone core, with the side chains of the component amino acids extending outward from the central axis to avoid interfering sterically with each other. A very diverse group of proteins contains α-helices. For example, the keratins are a family of closely related, fibrous proteins whose structure is nearly entirely α-helical. They are a major component of tissues such as hair and skin, and their rigidity is determined by the number of disulfide bonds between the constituent polypeptide chains. In contrast to keratin, myoglobin, whose structure is approximately eighty percent α-helical, is a globular, flexible molecule.



1. Hydrogen bonds: An α-helix is stabilized by extensive hydrogen bonding between the peptide-bond carbonyl oxygens and amide hydrogens that are part of the polypeptide backbone. The hydrogen bonds extend up the spiral from the carbonyl oxygen of one peptide bond to the -NH - group of a peptide linkage four residues ahead in the polypeptide. This ensures that all but the first and last peptide bond components are linked to each other through hydrogen bonds. Hydrogen bonds are individually weak, but they collectively serve to stabilize the helix.

2. Amino acids per turn: Each turn of an α-helix contains 3.6 amino acids. Thus, amino acid residues spaced three or four apart in the primary sequence are spatially close together when folded in the α-helix.

3. Amino acids that disrupt an α-helix: Proline disrupts an α-helix because its imino group is not geometrically compatible with the right-handed spiral of the α-helix. Instead, it inserts a kink in the chain, which interferes with the smooth, helical structure. Large numbers of charged amino acids (for example, glutamate aspartate, histidine, lysine, or arginine) also disrupt the helix by forming ionic bonds, or by electrostatically repelling each other. Finally, amino acids with bulky side chains, such as tryptophan, or amino acids, such as valine or isoleucine, that branch at the β-carbon (the first carbon in the R-group, next to the α-carbon) can interfere with formation of the α-helix if they are present in large numbers.



B. β-sheet

The β-sheet is another form of secondary structure in which all of the peptide bond components are involved in hydrogen bonding. The surfaces of β-sheets appear "pleated," and these structures are, therefore, often called **"β-pleated sheets."** When illustrations are made of protein structure, β-strands are often visualized as broad arrows.

1. Comparison of a β-sheet and an α-helix: Unlike the α-helix, β-sheets are composed of two or more peptide chains (β-strands), or segments of polypeptide chains, which are almost fully extended. Note also that in β-sheets the hydrogen bonds are perpendicular to the polypeptide backbone.

2. **Parallel and antiparallel sheets:** A β-sheet can be formed from two or more separate polypeptide chains or segments of polypeptide chains that are arranged either antiparallel to each other (with the and ends of the β-strands alternating as shown in Figure B, or parallel. When the hydrogen bonds are formed between the

Poly peptide backbones of separate polypeptide chains, they are termed **inter-chain bonds. A** β-sheet can also be formed by a single polypeptide chain folding back on itself (see Figure C). In this case, the hydrogen bonds are **intrachain bonds. In** globular proteins, β-sheets always have a right-handed curl, or twist, when viewed along the polypeptide backbone.

C. β bends (reverse turns)

β - bends reverse the direction of a polypeptide chain, helping it forma compact, globular shape. They are usually found on the surface of protein molecules, and often include charged residues. β - bends are generally composed off our amino acids, one of which may be Proline the imono acid that causes a "kink" in the polypeptide chain. Glycine, the amino acid with the smallest R-group, is also frequently found in β-bends. β -Bends are stabilized by the formation of hydrogen and ionic bonds.

D. Non repetitive secondary structure

Approximately one half of an average globular protein is organized into repetitive structures, such as the α-helix and/or β-sheet. The remainder of the polypeptide chain is described as having a loop or coil conformation. These nonrepetitive secondary structures are not "random," but rather simply have a less regular structure.

E. Supersecondary structures (motifs)

Globular proteins are constructed by combining secondary structural elements (α-helices, β-sheets, nonrepetitive sequences). These form primarily the core region-that is, the interior of the molecule. They are connected by loop regions (for example, β-bends) at the surface of the protein. Super secondary structures are usually produced by packing side chains from adjacent secondary structural elements close to each other. Thus, for example, α-helices and β-sheets that are adjacent in the amino acid sequence are also usually (but not always) adjacent in the final, folded protein. Some of the more common motifs are illustrated in Figure.

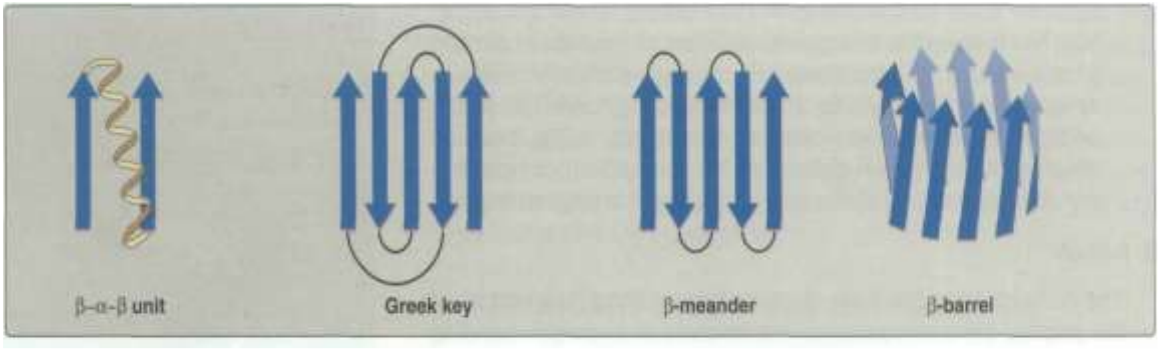


Figure common structural motifs

1.3. TERTIARY STRUCTURE OF GLOBULAR PROTEINS

The primary structure of a polypeptide chain determines its **tertiary structure.** [Note: "Tertiary" refers both to the folding of domains (the basic units of structure and function, see discussion below), and the final arrangement of domains in the polypeptide.] The structure of globular proteins in **aqueous solution** is compact, with a high-density (close packing) of the atoms in the core of the molecule. **Hydrophobic side chains** are buried in the **interior,** whereas **hydrophilic groups** are generally found on the **surface** of the molecule. All hydrophilic groups (including components of the peptide bond) located in the interior of the polypeptide are involved in hydrogen bonds or electrostatic interactions.

A. Domains

Domains are the fundamental functional and three-dimensional structural units of a polypeptide. Polypeptide chains that are greater than 200 amino acids in length generally consist of two or more domains. The core of a domain is built from combinations of **super-secondary structural elements (motifs).** Folding of the peptide chain within a domain usually occurs independently of folding in other domains. Therefore, each domain has the characteristics of as mall, compact globular protein that is structurally independent of the other domains in the polypeptide chain.

B. Interactions stabilizing tertiary structure

The unique three-dimensional structure of each polypeptide is determined by its amino acid sequence. Interactions between the amino acid side chains guide the folding of the polypeptide to form a compact structure. Four types of interactions cooperate in stabilizing the tertiary structures of globular proteins.

1. **Disulfide bonds:** A disulfide bond is a covalent linkage formed from the sulfhydryl group (-SH) of each of **two cysteine residues,** to produce a **cysteine** residue (Figure 2.9). The two cysteine's may be separated from each other by many amino acids in the primary sequence of a polypeptide, or may even be located on two different polypeptide chains; the folding of the polypeptide chain(s)brings the cysteine residues into proximity, and permits covalent bonding of their side chains. A disulfide bond contributes to the stability of the three-dimensional shape of the protein molecule. For example, many disulfide bonds are found in proteins such as immunoglobulin's that are secreted by cells.

2. Hydrophobic interactions: Amino acids with nonpolar side chains tend to be located in the interior of the polypeptide molecule, where they associate with other hydrophobic amino acids. In contrast, amino acids with polar or charged side chains tend to be located on the surface of the molecule in contact with the polar solvent. Proteins located in nonpolar (lipid) environments, such as

a membrane, exhibit the reverse arrangement – that is, hydrophilic amino acid side chains are located in the interior of the polypeptide, whereas hydrophobic amino acids are located on the surface of the molecule in contact with the nonpolar environment In each case, the segregation of R-groups occurs that is energetically most favorable.

3. Hydrogen bonds: Amino acid side chains containing oxygen- or nitrogen-bound hydrogen, such as in the alcohol groups of serine and threonine, can form hydrogen bonds with electron-rich atoms, such as the oxygen of a carboxyl group or carbonyl group of a peptide bond Formation of hydrogen bonds between polar groups on the surface of proteins and the aqueous solvent enhances the solubility of the protein.

4. Ionic interactions: Negatively charged groups, such as the carboxyl group ($-COO^-$ in the side chain of aspartate or glutamate, can interact with positively charged groups, such as the amino group ($-NH_3^+$ in the side chain of lysine.

C. Protein folding

Interactions between the side chains of amino acids determine how a long polypeptide chain folds into the intricate three-dimensional shape of the functional protein. Protein folding, which occurs within the cell in seconds to minutes, employs a shortcut through the maze of all folding possibilities. As a peptide folds, its amino acid side chains are attracted and repulsed according to their chemical properties. For example, positively and negatively charged side chains attract each other. Conversely, similarly charged side chains repel each other. In addition, interactions involving hydrogen bonds, hydrophobic interactions, and disulfide bonds all seek to exert an influence on the folding process. This process of trial and error tests many, but not all, possible configurations, seeking a compromise in which attractions outweigh repulsions. This results in a correctly folded protein with a low energy state

D. Role of chaperones in protein folding

It is generally accepted that the information needed for correct protein folding is contained in the primary structure of the polypeptide. Given that premise, it is difficult to explain why most proteins when denatured do not resume their native conformations under favorable environmental conditions. One answer to this problem is that a protein begins to fold in stages during its synthesis, rather than waiting for synthesis of the entire chain to be totally completed. This limits competing folding configurations made available by longer stretches of nascent peptide. In addition, a specialized group of proteins, named **"chaperones,"** are required for the proper folding of many species of proteins. The chaperones- also known as **"heat shock" Protein-interact** with the polypeptide at various stages during the folding process. Some chaperones are important in keeping the protein unfolded until its synthesis is finished, or act as catalysts by increasing the rates of the final stages in the folding process. Others protect proteins as they fold so that their vulnerable, exposed regions do not become tangled in unproductive encounters.

1.4. QUATERNARY STRUCTURE OF PROTEINS

Many proteins consist of a single polypeptide chain, and are defined as monomeric **proteins.** However, others may consist of two or more polypeptide chains that may be structurally identical or totally unrelated. The arrangement of these polypeptide subunits is called the quaternary structure of the protein. [Note: If there are two subunits, the protein is called dimeric if three subunits trimeric and, if several subunits, multimeric Subunits are held together by non-covalent interactions (for

example, hydrogen bonds, ionic bonds, and hydrophobic interactions). Subunits may either function independently of each other, or may work cooperatively, as in hemoglobin, in which the binding of oxygen to one subunit of the tetramer increases the affinity of the other subunit oxygen.

1.5. Protein Folding

Protein folding is a process in which a polypeptide folds into a specific, stable, functional, three-dimensional structure. It is the process by which a protein structure assumes its functional shape or conformation Proteins are formed from long chains of amino acids; they exist in an array of different structures which often dictate their functions. Proteins follow energetically favorable pathways to form stable, orderly, structures; this is known as the proteins' native structure. Most proteins can only perform their various functions when they are folded. The proteins' folding pathway, or mechanism, is the typical sequence of structural changes the protein undergoes in order to reach its native structure. Protein folding takes place in a highly crowded, complex, molecular environment within the cell, and often requires the assistance of molecular chaperones, in order to avoid aggregation or mis folding. Proteins are comprised of amino acids with various types of side chains, which may be hydrophobic, hydrophilic, or electrically charged. The characteristics of these side chains affect what shape the protein will form because they will interact differently intra molecularly and with the surrounding environment, favoring certain conformations nd structures over others. Scientists believe that the instructions for folding a protein are encoded in the sequence. Researchers and scientists can easily determine the sequence of a protein, but have not cracked the code that governs folding.

1.5.1. Protein Folding theory and experiment

Early scientists who studied proteomics and its structure speculated that proteins had templates that resulted in their native conformations. This theory resulted in a search for how proteins fold to attain their complex structure. It is now well known that under physiological conditions, proteins normally spontaneously fold into their native conformations. As a result, a protein's primary structure is valuable since it determines the three-dimensional structure of a protein. Normally, most biological structures do not have the need for external templates to help with their formation and are thus called self-assembling.

1.5.2. Protein Renaturation

Protein renaturation known since the 1930s. However, it was not until 1957 when Christian Anfinsen performed an experiment on bovine pancreatic RNase A that protein renaturation was quantified. RNase A is a single chain protein consisting of 124 residues. In 8M urea solution of 2-mercaptoethanol, the RNase A is completely unfolded and has its four disulfide bonds cleaved through reduction. Through dialysisof urea and introducing the solution to O2 at pH 8, the enzymatically active protein is physically incapable of being recognized from RNase A. As a result, this experiment demonstrated that the protein spontaneously renatured.

One criteria for the renaturation of RNase A is for its four disulfide bonds to reform. The likelihood of one of the eight Cys residues from RNase A reforming a disulfide bond with its native residue compared to the other seven Cys residues is 1/7. Futhermore, the next one of remaining six Cys residues randomly forming the next disulfide bond is 1/5 and etc. As a result, the probability of RNase A reforming four native disulfide links at random is (1/7 * 1/5 * 1/3 * 1/1 = 1/105). The result of this probability demonstrates that forming the disulfide bonds from RNase A is not a random activity.

When RNase A is reoxidized utilizing 8M urea, allowing the disulfide bonds to reform when the polypeptide chain is a random coil, then RNase A will only be around 1 percent enzymatically active after urea is removed. However, by using 2-mercaptoethanol, the protein can be made fully active once again when disulfide bond interchange reactions occur and the protein is back to its native state. The native state of the RNase A is thermodynamically stable under physiological conditions, especially since a more stable protein that is more stable than that of the native state requires a larger activation barrier, and is kinetically inaccessible.By using the enzyme protein disulfide isomerase (PDI), the time it takes for randomized RNase A is minimized to about 2 minutes. This enzyme helps facilitate the disulfide interchange reactions. In order for PDI to be active, its two active site Cys residues needs to be in the -SH form. Furthermore, PDI helps with random cleavage and the reformation of the disulfide bonds of the protein as it attain thermodynamically favorable conformations.

1.5.3. Posttranslationally Modified Proteins Might Not Renature

Proteins in a "scrambled" state go through PDI to renature, and their native state does not utilize PDI because native proteins are in their stable conformations. However, proteins that are posttranslationally modified need the disulfide bonds to stabilize their rather unstable native form. One example of this is insulin, a polypeptide hormone. This 51 residue polypeptide has two disulfide bonds that is inactivated by PDI. The following link is an image showing insulin with its two disulfide bonds. Through observation of this phenomena, scientists were able to find that insulin is made from proinsulin, an 84-residue single chain. This link provides more information on the structure of proinsulin and its progression on becoming insulin. The disulfide bonds of proinsulin need to be intact before conversion of becoming insulin through proteolytic excision of its C chain which is an internal 33-residue segment. However according to two findings, the C chain is not what dictates the folding of the A and B chains, but instead holds them together to allow formation of the disulfide bonds. For one, with the right renaturing conditions in place, scrambled insulin can become its native form with a 30% yield. This yield can be increased if the A and B chains are cross-linked. Secondly, through analysis of sequences of proinsulin from many species, mutations are permitted at the C chain eight times more than if it were for A and B chains.

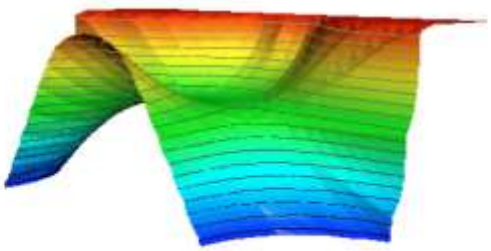
1.5.4. The Protein Folding Process

Considerable evidence suggests that all of the information to describe the three dimensional conformation of a protein is contained within the primary structure. However, for the most part, we cannot fully interpret the information contained within the sequence. To understand why this is true, we need to take a more careful look at proteins and how they fold.

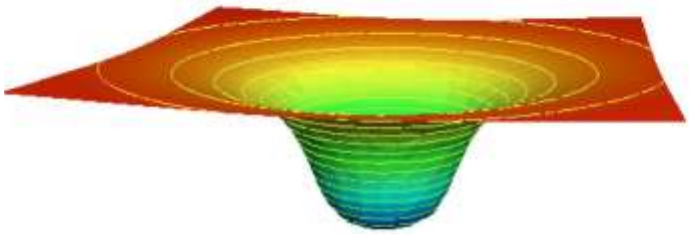
The polypeptide chain for most proteins is quite long. It therefore has ***many*** possible conformations. If you assume that all residues could have 2 possible combinations of ϕ and ψ angles (real peptides can have many more than this), a 100 amino acid peptide could have 2^{100} ($\sim 10^{30}$) possible conformations. If the polypeptide tested a billion conformations/second, it would still take over 10^{13} years to find the correct conformation. (Note that the universe is only $\sim 10^{10}$ years old, and that a 100 residue polypeptide is a relatively small protein.) The observation that proteins cannot fold by random tests of all possible conformations is referred to as the Levinthal paradox.

1.5.5. Folding pathways

In classical transition state theory, the reaction diagram for a spontaneous two state system is considered to have a high-energy starting material, a lower energy product, and an energy barrier between them. While the typical diagram that describes the process (such as the one shown at right) is useful, it is incomplete. The process for the conversion of S to P could actually take many pathways; the pathway shown is merely the minimum energy route from one state to another. The true situation is described by an energy landscape, with the minimum energy route being the equivalent of a pass between two mountains. Thus, although the pathway involves an energy barrier, other pathways require passing through even higher energy states.



A large part of the reason that single pathways (or small numbers of pathways) exist for chemical reactions is that most reactions involve the cleavage and reformation of covalent bonds. The energy barrier for breaking a covalent bond is usually quite high. In protein folding, however, the interactions involved are weak. Because the thermal energy of a protein molecule is comparable to the typical non covalent interaction strength, an unfolded polypeptide is present in a large variety of rapidly changing conformations. This realization led to the Levinthal paradox: because the unfolded protein should be constantly changing its shape due to thermal motions of the different parts of the polypeptide, it seemed unlikely that the protein would be able to find the correct state to begin transiting a fixed folding pathway.



An alternate hypothesis has been proposed, in which *portions* of the protein self-organize, followed by folding into the final structure. Because the different parts of the protein begin the folding process independently, the shape of the partially folded protein can be very variable. In this model, the protein folds by a variety of different paths on an energy landscape. The folding energy landscape has the general shape of a **funnel**. In the folding process, as long as the overall process results in progressively lower energies, there can be a large variety of different pathways to the final folded state.

The folding funnel shown above has a smooth surface. Actual folding funnels may be fairly smooth, or may have irregularities in the surface that can act to trap the

polypeptide chain in misfolded states. Alternatively, the folding funnel may direct the polypeptide into a ***metastable*** state. Metastable states are local minima in the landscape; if the energy barriers that surround the state are high enough, the metastable state may exist for a long time – metastable states are stable for **kinetic** rather than **thermodynamic** reasons.

The difficulty in refolding many proteins *in vitro* suggests that the folded state of at least some complex proteins may be in a metastable state rather than a global energy minimum.

1.5.6. Folding process

The lower energies observed toward the depression in the folding funnel are thought to be largely due to the collapse of an extended polypeptide due to the hydrophobic effect. In addition to the hydrophobic effect, de solvation of the backbone is necessary for protein folding, at least for portions of the backbone that will become buried. One method for desolvation of the backbone is the formation of secondary structure. This is especially true for helical structures, which can form tightly organized regions of hydrogen bonding while excluding water from the backbone structure. A general outline for the process experienced by a folding protein seems to look like this:

A general outline for the process experienced by a folding protein seems to look like

this:

1. Some segments of a polypeptide may rapidly attain a relatively stable, organized structure (largely due to organization of secondary structural Elements).

2. These structures provide nuclei for further folding.

3. During the folding process, the protein is proposed to form a state called a

1.5.7. Molten globule. This state readily rearranges to allow interactions between different parts of the protein.

4. These nucleated, partially folded domains then coalesce into the folded protein. If this general pathway is correct, it seems likely that at least some of the residues within the sequence of most proteins function to guide the protein into the proper folding pathway, and prevent the "trapping" of the polypeptide in unproductive

Partially folded states.

1.5.8. Folding inside cells

Real cells contain **many** proteins at a high overall protein concentration. The protein concentration inside a cell is ~150 mg/ml. folding inside cells differs from most experiments used to study folding *in vitro*:

1. Proteins are synthesized on ribosomes. The entire chain is not available to fold at once, as is the case for an experimentally unfolded protein in a test tube.

2. Within cells, the optimum ionic concentration, pH, and macromolecule Concentration for each protein to fold properly cannot be controlled as tightly as in an experimental system.

3. Major problems could arise if unfolded or partially folded proteins encountered one another. Exposed hydrophobic regions might interact, and form potentially lethal insoluble aggregates within the cell.

One mechanism for limiting problems with folding proteins inside cells volves specialized proteins called **molecular chaperones**, which assist in folding proteins. Molecular chaperones were first observed to be involved in responses to elevated temperature (*i.e*. "heat shock") to stabilize existing proteins and prevent protein aggregation and were called heat-shock proteins (abbreviated as "hsp").

Additional research revealed that heat shock proteins are present in all cells, and that they decrease or prevent non-specific protein aggregation and assist in protein folding.

1.5.9. Thermodynamics of protein folding

In contemplating protein folding, it is necessary to consider different types of amino acid side-chains separately. For each situation, the reaction involved will be assumed to be:

Protein *unfolded* ⟷ Protein *folded*

Note that this formalism means that a negative ΔG implies that the folding process is spontaneous.

First we will look at **polar groups** in an aqueous solvent. For polar groups, the Δ*Hchain* favors the unfolded structure because the backbone and polar groups interact form stronger interactions with water than with themselves. More hydrogen bonds and electrostatic interactions can be formed in unfolded state than in the folded state. This is true because many hydrogen bonding groups can form more than a single hydrogen bond. These groups form multiple hydrogen bonds if exposed to water, but frequently can form only single hydrogen bonds in the folded structure of a protein.

For similar reasons, the Δ*H solvent* favors the folded protein because water interacts more strongly with itself than with the polar groups in the protein. More hydrogen bonds can form in the absence of an extended protein, and therefore the number of the **sum of the ΔH polar** contributions is close to zero, but usually favors the folded structure for the protein slightly. The chain ΔH contributions are positive, while the solvent ΔH contributions are negative. The sum is slightly negative in most cases, and therefore slightly favors folding.

The Δ*S chain* of the polar groups favors the unfolded state, because the chain is much more disordered in the unfolded state. In contrast, the Δ*S solvent* favors the folded

State, because the solvent is more disordered with the protein in the folded state. In most cases, the **sum of the ΔS polar** favors the **unfolded** state slightly. In other words, the ordering of the chain during the folding process outweighs the other entropic factors.

The **ΔG polar** that is obtained from the values of ΔH polar and ΔS polar for the polar groups varies somewhat, but usually tends to favor the unfolded protein. In other words, the folding of proteins comprised of polar residues is usually a nonspontaneous process.

Next, we will consider a chain constructed from **non-polar groups** in aqueous

Solvent. Once again, the Δ*H chain* usually favors the unfolded state slightly. Once again, the reason is that the backbone can interact with water in the unfolded state. However, the effect is smaller for non-polar groups, due to the greater number of favorable van der Waals interactions in the folded state. This is a result of the fact that non-polar atoms form better van der Waals contacts with other non-polar groups than with water; in some cases, these effects mean that the Δ*H chain* for nonpolar residues is slightly negative.

As with the polar groups, the Δ*H solvent* for non-polar groups favors the folded state. In the case of non-polar residues, Δ*H solvent* favors folding more than it does for polar groups, because water interacts much more strongly with itself than it does with non-polar groups,

The **sum of the** Δ**H non-polar** favors folding somewhat. The magnitude of the ΔH nonpolar is not very large, but is larger than the magnitude of the ΔH polar, which also tends to slightly favor folding.

The Δ*S chain* of the non-polar groups favors the less ordered unfolded state. However, the Δ*S solvent* highly favors the folded state, due to the hydrophobic effect. During the burying of the non-polar side chains, the solvent becomes more disordered. The ΔS *solvent* is a major driving force for protein folding.

The Δ**G non-polar** is therefore negative, due largely to the powerful contribution of the Δ*S solvent*.

Adding together the terms for ΔG polar and ΔG non-polar gives a slightly negative overall ΔG for protein folding, and therefore, proteins generally fold spontaneously raising the temperature, however, tends to greatly increase the magnitude of the TΔ*S chain* term, and therefore to result in unfolding of the protein.

The folded state is the sum of many interactions. Some favor folding, and some

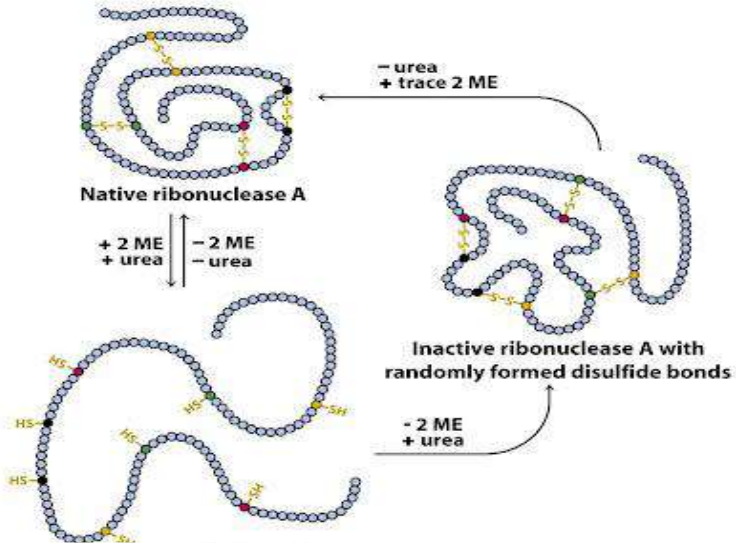
Favor the unfolded state. The qualitative discussion above did not include the magnitudes of the effects. For real proteins, the various ΔH and ΔS values are difficult to measure accurately. However, for many proteins it is possible to estimate the overall ΔG of folding. Measurements of this value have shown that **the overall** Δ**G for protein folding is very small**: only about –10 to –50 k Joules/mol. This corresponds to a few salt bridges or hydrogen bonds.

Studies of protein folding have revealed one other important point: the hydrophobic effect is very important, but it is relatively non-specific. Any hydrophobic group will interact with essentially any other hydrophobic group. While the hydrophobic effect is a major driving force for protein folding, it is the constrains imposed by the more geometrically specific hydrogen bonding and electrostatic interactions in conjunction with the hydrophobic interactions that largely determine the overall folded structure of the protein.

1.5.10. Anfinsen's experiment

1. The Observation

Ribonuclease A (RNaseA) is an extracellular enzyme of 124 residues with four disulfide bonds. In the first phase of the experiment, the S-S bonds were reduced to eight –SH groups (using mercaptoethanol, HS-CH2-CH2-OH); the protein was then denatured with 8 M urea. Under these conditions, the enzyme is inactive and becomes a flexible random polymer. In the second phase, the urea was slowly removed (dialysis); then the –SH groups were oxidized back to S-S bonds. If the protein was able to regain its native structure spontaneously after removal of the urea, we expect that it would also regain its activity. In fact, the activity was >90% of the untreated enzyme. Moreover, sequence analysis showed that nearly all of the correct S-S bonds had been formed.



2. The Control

A reasonable objection can be raised to the above result by suggesting that perhaps RNase A was not completely unfolded in 8 M urea. To address this class of objections, RNase A was first reduced and denatured as above. But in the second phase, the enzyme was first oxidized to form S-S bonds, and then the urea was removed, i.e. the order of steps in the second phase of the experiment was reversed. The resulting activity was only about 1-2% of the untreated enzyme. Sequence analysis showed a random assortment of S-S bonds.

Anfinsen's work showed convincingly that proteins can indeed adopt their native information spontaneously, i.e. sequence determines structure. His demonstration of this fundamental property of proteins opened the problem to a massive amount of experimental and theoretical effort

1.5.11. The Levinthal paradox and kinetics

Levinthal's paradox is a thought experiment, also constituting a self-reference in the theory of protein folding. In 1969, Cyrus Levinthal noted that, because of the very large number of degrees of freedom in an unfolded polypeptide chain, the molecule has an astronomical number of possible conformations. An estimate of 3^{300} or 10^{143} was made in one of his papers. For example, a polypeptide of 100 residues will have 99 peptide bonds, and therefore 198 different phi and psibond angles. If each of these bond angles can be in one of three stable conformations, the protein may misfold into a maximum of 3^{198} different conformations (including any possible folding redundancy). Therefore if a protein were to attain its correctly folded configuration by sequentially sampling all the possible conformations, it would require a time longer than the age of the universe to arrive at its correct native conformation. This is true even if conformations are sampled at rapid (nanosecond or picosecond) rates. The "paradox" is that most small proteins fold spontaneously on a millisecond or even

microsecond time scale. This paradox is central to computational approaches to protein structure prediction.

Levinthal himself was aware that proteins fold spontaneously and on short timescales. He suggested that the paradox can be resolved if "protein folding is sped up and guided by the rapid formation of local interactions which then determine the further folding of the peptide; this suggests local amino acid sequences which form stable interactions and serve as nucleation points in the folding process."[4] Indeed, the protein folding intermediates and the partially folded transition states were experimentally detected, which explains the fast protein folding. This is also described as protein folding directed within funnel-like energy landscapes some computational approaches to protein structure prediction have sought to identify and simulate the mechanism of protein folding. Levinthal also suggested that the native structure might have a higher energy, if the lowest energy was not kinetically accessible. An analogy is a rock tumbling down a hillside that lodges in a gully rather than reaching the base.

1.5.12. Protein Folding Rate

Determining how a protein will fold has been fairly difficult to predict even though the amino acid sequence is known. Instead of analyzing the structure of the protein and analyzing the mechanism of how a protein folds, understanding the kinetics of folding rates has proven to be a much more efficient way of understanding protein folding. The two-state folding kinetics of proteins is mostly studied, which analyzes the folding progress of a protein from its linear chain form, its primary structure, to its folded state, its tertiary structure. This process is dependent on the cooperative nature of the transition state. The kinetics of protein folding can be illustrated through the funnel energy landscape diagram, which is mathematically explained through the Gibbs free energy equation. This energy landscape diagram can follow the tract of the many pathways a protein can take until it reaches its native, or most stable, folded state. As a protein conforms to its most native state, a free energy barrier ends up controlling the kinetics of the protein folding. To illustrate the folding mechanisms, different Go-model simulations are used, which are coarse-grained topology-based models. However, although Go-model simulations provide the folding mechanism of proteins, they lack the ability to predict the folding rates of proteins based on the kinetic or thermodynamic cooperativity demonstrated by two-state proteins. Because of this reason, studies have been done to understand the cooperative nature of the two-state folding of proteins and the factors that affect the folding rates of proteins.

1.5.13. Folding rate trends of Protein

The folding rates of two-state proteins can be understood through two general properties of the folded conformations. One of the trends is that more structurally complex proteins tend to fold at slower rates in comparison to more simple structural proteins. For example, a tertiary structure containing beta sheet proteins and proteins combined with alpha helices and beta sheets tend to fold slower than proteins that are made up of only alpha helices. The second trend is that larger proteins tend to fold a lot more slowly than smaller proteins. The kinetics of alpha helical proteins and structurally complicated proteins such as globular proteins also differ due to long-range tertiary contacts. The transition states of globular proteins are expected to have a higher transitional energy barrier than alpha helical proteins because more entropic energy is required to make a more structurally complicated protein to fold in a more ordered fashion in comparison to a simpler structural protein.

As the chain length of a protein also increases, the free energy barrier exponentially increases as well to reach the transition state of the protein.

In determining the transition state of an in-process folded protein, the native state topology of the protein has to be known in order to predict the structure of the transition state of the protein. Topology refers to the effect of the orientation of objects in space due to deformations of the objects. In the case for proteins, a folded structure might change its orientation in space if the protein is heated up as it would lead to denaturing. To examine this transition state of folded proteins, the formation of the transition state is determined by the free energy barrier that controls the kinetics of the folding reaction. This free energy barrier is the result of the compensation of energy and the loss in entropy due to the new interactions formed in the process of protein folding. The relationship between the kinetics of a folding protein and topology help to explain why the transition state of a protein is dependent upon its native state. This is known as the principle of minimum frustration of energy landscape theory, which can related to the funnel model of folded proteins. The more stable the protein is, the lower the energy it is at, and the energy of the native protein can help give information on how much energy is required for a protein to reach its transition state in the folding process.

1.5.14. COOPERATIVITY OF PROTEINS

The use of Go models helps to give an identification of a protein in its most native state, which is held together by stabilizing interactions between native contacts. These stabilizing interactions are also known as non-additive forces, and these forces play a factor in the kinetics and thermodynamics of protein folding. These non-additive forces can also be thought of as intramolecular interactions that happen spontaneously within the protein such as side-chain ordering and hydrophobic forces. The effect of these non-additive forces have been shown to increase the free energy barrier of the two-state folded protein, and therefore, this makes these Go models more thermodynamically cooperative.

Upon using these Go models, the three-body interactions of the folding rates and what are known as phi values are examined in two-state proteins. The meaning of these phi values gives a relationship between the transition state of a two-state folded protein and its native state. The phi value explains the content of the native structure in its transition state. Therefore, the more native-like the structure of the transition state, the more likely this transition state will conform into its native state in a shorter period of time. In general, phi values improve when the transition state is more like its native state, but the ratio between its transition state and native state is different for each protein that varies in size and its secondary structure.

Many different types of Go models have been developed to better understand the cooperativity of the folding rates of proteins. For example, a Go model has been created in analyzing a small alpha-helical protein also known as a Calpha Go-like model. This model has also been altered by introducing solvent-mediated interactions to the model. The interactions between proteins are instead replaced by a desolvation barrier. Studies have shown that the thermodynamic and kinetic cooperativity of two-state folded proteins increase as the desolvation barrier increases in height. Desolvation is known as the removal of solvent from a material in solution. In general, desolvation has a property whee short-range contact proteins such as those that form alpha-helices have little cooperativity due to desolvation while long-range contacts such as those with a mix of beta sheets and alpha helices are expected to have high cooperativity because long-range contacts require persistence in bringing the proper chains together, and therefore, require a high amount of cooperativity. In conclusion, it is these

topological models with nonadditive forces such as hydrophobic forces of proteins that help to better understand the folding rates of certain proteins.

1.5.15. MOLTEN GLOBULE

The molten globule state is an intermediate conformational state between the native and the fully unfolded states of a globular protein . Many proteins can be observed in this state when partially unfolded at equilibrium, under mild denaturation conditions, or as a transient intermediate kinetic species, being formed rapidly from the unfolded state upon transfer to refolding conditions. The characteristics of the molten globule state are:

1. the presence of a native-like content of secondary structure

2. the absence of a specific tertiary structure produced by the tight packing of amino acid side chains

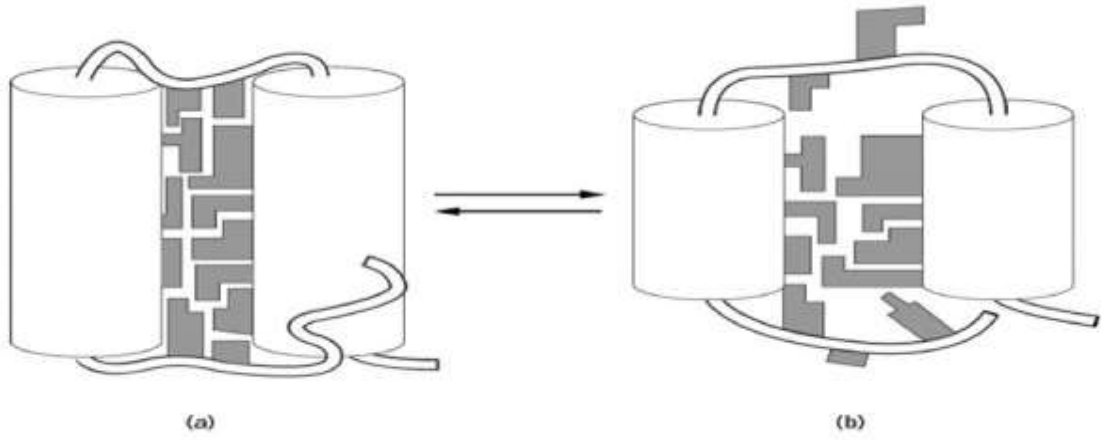
3. compactness in the overall shape of the protein molecule, with a radius 10 to 30% larger than that of the native state

4. the presence of a loosely packed hydrophobic core that increases the hydrophobic surface area accessible to solvent.

Thus, in short, the molten globule is a compact globule with a "molten" side-chain structure that is primarily stabilized by nonspecific hydrophobic interaction

Experimentally, the molten globule state is characterized by having a native-like circular dichroism (CD) peptide spectrum below 250 nm, which arises from the secondary structure, and by an unfolded-like CD spectrum of the aromatic side chains between 250 and 320 nm, due to the absence of specific side-chain packing interactions. Hydrogen atoms of the peptide backbone involved in secondary structure of the molten globule state appear to be protected from hydrogen exchange with the solvent protons, but the protection factor for the molten globule (10 to 1000) is much smaller than that for the native state, which is often greater than 10 6 The nuclear magnetic resonance (NMR) spectrum of the molten globule state is closer to that of the unfolded protein, and there is little, if any, chemical shift dispersion in the spectrum, reflecting the absence of a specific tertiary structure. The individual resonances in the NMR spectrum are, however, broader than those in the unfolded state,

A schematic model of the native (a) and the molten globule (b) states of a protein molecule.



reflecting conformational fluctuations in the molten globule state. It is known that the structure of the molten globule state, as determined by hydrogen exchange and NMR spectra, is heterogeneous in proteins, including a-lactalbumin, cytochrome c, and apo myoglobin. In these proteins, one portion of the structure is more organized and substantially protected in the molten globule state, with other portions of the structure being less organized. Solution X-ray scattering has been used to characterize the molten globule structure. The presence of a clear peak in the Kratky plot and the radius of gyration evaluated from the Guinier plot of the X-ray scattering curve in the molten globule state show that the protein molecule in this state is compact and globular. Limited proteolysis by proteolytic enzymes has also been used for probing the partly folded structures of proteins; the key result is that the molten globule can be sufficiently rigid to prevent extensive proteolysis and it appears to maintain significant native-like structure. A hydrophobic fluorescent dye, such as 8-anilinonaphthalene-1-sulfonate (ANS), which binds to solvent-accessible hydrophobic surfaces of a protein molecule, has also been used for characterizing the molten globule state. The molten globule binds ANS much more strongly than does either the fully folded or fully unfolded form of the protein; the latter can be generated in a concentrated solution of a strong denaturant (6 M guanidinium chloride or 8 M urea).

For some proteins, the molten globule state is an equilibrium intermediate observed at an intermediate concentration of a strong denaturant (eg, 2 M guanidinium chloride) as part of a denaturant-induced unfolding transition. On the other hand, many globular proteins show a cooperative two-state unfolding transition without the intermediate. Whether or not the molten globule state is observed as a stable intermediate depends upon its stability relative to that of the native and the unfolded states. The unfolding intermediates of carbonic anhydrase and of a-lactalbumin are typical examples of a molten globule state that is stably populated at an intermediate concentration of denaturant. For these proteins, the partially unfolded states at acidic or alkaline pH are identical to the unfolding intermediate in the denaturant-induced transition, so the acidic or alkaline transitions also produce the molten globule state. For some other proteins, such as cytochrome c, apomyoglobin, and b-lactamase, the acidic or alkaline transition is known to produce a more extensively unfolded state; in these cases the addition of salt refolds the protein molecule from the unfolded to the molten globule state. The salt-induced refolding to the molten globule state is caused by counterion binding of the salt to the protein molecule, which eliminates the electrostatic repulsion between the charged groups. Other mildly denaturing processes that lead to the molten globule state include denaturation induced by hydrostatic pressure and by alcohols.

Removal of the bound metal ion in a metal-ion binding protein sometimes results in a molten globule state, as in the case of apo-a-lactalbumin produced by removal of the bound Ca. Covalent modification of a protein can also sometimes result in a molten globule state.

In many globular proteins, the molten globule state is observed at an early stage of the kinetics of refolding from the unfolded state. The early formation of the molten globule might be a way by which these proteins can be folded efficiently without wandering into the huge conformational space available for the proteins. Two experimental techniques, stopped-flow CD and pulsed hydrogen-exchange combined with either two-dimensional NMR or electrospray ionization mass spectrometry, have been used successfully to characterize the transiently formed molten globule-like states during the kinetics of refolding of many globular proteins. The stopped-flow CD studies have shown the rapid formation of the peptide secondary structure occurring within the dead-time of the stopped-flow mixing (~10ms), although how much of the secondary structure is rapidly regained depends on the protein species. The pulsed hydrogen-exchange technique, when combined with two-dimensional NMR, can identify the specific location of stabilized secondary structure segments in a transient

intermediate of a protein. For apomyoglobin and ribonuclease HI, comparison of the kinetic refolding intermediate and the equilibrium molten globule state has shown that the two are identical. Identification of the molten globule intermediate has also been well-established for a-lactalbumin by time-resolved CD and NMR studies. Molten globule-like folding intermediates have also been detected and characterized in many other globular proteins. Nevertheless, this does not necessarily mean that the molten globule state must be an obligatory, universal intermediate of protein folding. Because the formation of the molten globule-like folding intermediate is usually too rapid to be coupled kinetically with the subsequent folding reactions, it is very difficult to determine whether or not the molten globule is an obligatory folding intermediate. Furthermore, several small globular proteins with approximately 60 amino acid residues are known to refold very rapidly to the native state within a few milliseconds without accumulation of the molten globule intermediate.

When considering the role of the molten globule state in protein folding, it is important to address the question as to whether or not the molten globule is a thermodynamic state. Analysis of the cooperativity parameters for denaturant-induced unfolding transitions of some proteins has suggested that the transitions from the molten globule state to the unfolded state and from the native state to the molten globule state are both all-or-none transitions, indicating that the molten globule state is a thermodynamic state. Furthermore, stability studies of mutants of apomyoglobin and cytochrome c have concluded that the molten globule states of these proteins show cooperative unfolding and are stabilized by native-like tertiary interactions, in addition to nonspecific hydrophobic interactions, which is consistent with the proposal that the molten globule state is a distinct thermodynamic state. However, for the best-characterized molten globule of a-lactalbumin, calorimetry, NMR, vibrational Raman spectroscopy, and other techniques have clearly shown that the unfolding of this molten globule is not a cooperative two-state transition. Such diversity in the unfolding behavior of the molten globule state among different proteins may arise from the diversity of the molten globule structure. Because the native tertiary interactions are at least partially lost in the molten globule state, its structure must be more diverse than the native structure, and how cooperatively the molten globule unfolds may depend on how many residual native tertiary interactions are retained in this state. A good example is the molten globule state of the equine lysozyme, which is a calcium-binding protein and homologous to a-lactalbumin. Although the equine lysozyme molten globule apparently resembles that of a-lactalbumin, a rigorous analysis of its spectroscopic and thermodynamic properties has shown that its structure is significantly more highly organized and that its unfolding is a cooperative first-order transition accompanied by a large change in enthalpy. Because of this diversity in the intermediate conformational states of proteins, it is difficult to provide a clear structural definition of the molten globule state. Consequently, this causes some controversy, with the formation of native-like tertiary fold considered to be a characteristic of the molten globule state in certain cases, while in other cases structures with non-native tertiary folds are also molten globules. Furthermore, more than one intermediate conformational state is often observed between the native and fully unfolded states. Nevertheless, the four characteristics itemized at the beginning of this article are those generally accepted as the characteristics of the molten globule state.

It is now well established that not only the native state, but also non-native conformational states, play an important role in a biological cell. The protein states recognized by various molecular chaperones are non-native. The non-native conformation is also required for translocation of a protein across a biological membrane. Various genetic diseases can be caused by the misfolding of translated polypeptides, and this misfolding results from an increased propensity of the mutant proteins to form non-native conformations. Because the molten globule state is regarded as a denatured state under

physiological conditions, it definitely assumes some role in the above phenomena in vivo. It is also true, however, that there is a much greater diversity in the non-native conformations of proteins than in just the conformations characterized as the molten globule state.

1.5.16. BIOPHYSICAL TECHNIQUES FOR THE STUDY OF PROTEIN FOLDING

Rapid Mixing Methods

The mechanism of protein folding can be studied by two different groups of approaches. Equilibrium methods provide information about possible folding intermediate states or deduce rate constants from the molecular fluctuations or dynamic properties of the system. Relaxation methods follow the change of the system evolving toward a new equilibrium after a rapid perturbation of its extrinsic variables, such as temperature, pH, pressure, or solvent composition. The time required for folding varies greatly among proteins, ranging from microseconds to minutes. The smallest protein molecules with no folding intermediates fold on the microsecond timescale, which, on one hand, might make them suitable for in silico-folding simulation studies. On the other hand, such fast reactions make the experimental detection of events during the folding process difficult. Basic techniques in the study of folding kinetics are stopped-flow fluorescence spectroscopy and stopped-flow circular dichroism. These techniques are capable of monitoring the formation of secondary and tertiary structures during the folding reaction with millisecond time resolution. In such experiments, rapid processes occurring within the dead time of the measurement were observed for many proteins. The challenge to resolve this initial burst phase and to reveal the structural changes taking place during the first millisecond stimulated the development of new, rapid kinetic techniques capable of triggering and monitoring the folding process on the sub- millisecond timescale. While the conventional stopped-flow apparatus, in which a small volume of a freshly made mixture containing the reacting components is injected into the measurement cell, is quite economical and offers a wide range of applications, its dead time is usually about 1 millisecond or longer. Continuous-flow methods extend the time resolution to the microsecond time range. In the continuous-flow cell, solutions are mixed under highly turbulent conditions to achieve complete mixing. The kinetics of the reaction is monitored under steady-state flow conditions as a function of the distance downstream from the mixer by using relatively simple and inexpensive detection methods. Using this technique, it has become possible to study the initial collapse and formation of intermediates in the early stage of the folding reaction, during the burst phase.

Real-Time NMR Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy has greatly contributed to our understanding of the protein-folding problem. Hydrogen–deuterium exchange experiments, revealing dynamical events at an atomic level, have illuminated the process of unfolding from the native state and the structure of folding intermediates. The quenched-flow pulse- labeling technique has enabled researchers to study the early stages of protein folding using a conventional NMR instrument. NMR studies have characterized the properties of denaturant-induced equilibrium folding intermediate states such as the molten globule. In equilibrium systems, the rates of conversion between distinct conformational states can be calculated from a line shape analysis of the NMR resonances, and therefore can provide kinetic data on folding. Slow folding reactions such as cis–trans prolyl isomerization can be directly followed by sequential recording of one-dimensional (1D) NMR spectra.This method is particularly useful for discovering intermediates formed at the late stages of the folding process. Using a stopped-

flow device for injection of the protein solution into the NMR tube that already contains the denaturant or the refolding buffer pushes the dead time of mixing below 1 s . One of the first proteins studied by real-time NMR was a-lactalbumin. 1D-NOE (nuclear Overhauser effect) experiments revealed the native-like compactness of the transient molten globule state of a-lactalbumin. These experiments also demonstrated that the transient intermediate closely resembles the well-characterized stable molten globule state formed at low pH. While 1D-NMR spectra have limited resolution, multidimensional NMR can provide high spatial resolution information on the folding process. Because recording multidimensional spectra is time consuming, only slow processes could be followed directly by sequential recording. Balbach and coworkers developed new methods to reconstruct the kinetic history of folding reactions from a single two-dimensional NMR spectrum recorded during the entire time course of the reaction. The basis of these methods is that the line widths and intensities reflect the history of the folding events occurring during spectral accumulation. When applied to a-lactalbumin, the technique demonstrated the cooperative nature of the folding of the main chain.

CHEMICALLY INDUCED NUCLEAR POLARIZATION

Chemically induced nuclear polarization (CIDNP) can be used to probe the solvent accessibility of certain aromatic residues in proteins. The reactive collision of polarizable amino acids such as tryptophan, tyrosine, and histidine with a photoexcited dye such as flavin mononucleotide (FMN) results in an electron transfer (in the case of Trp and Tyr) or proton transfer (His) reaction forming a pair of radicals. Electron-nuclear hyperfine interactions between the two radicals result in a significant enhancement of NMR signals. The ‘‘photosensitizer’’ flavin molecule can be excited by laser as light source. For the photoreaction to take place, the aromatic side chains must be accessible to the photosensitizer, e.g., located on the surface of the protein molecule. The CIDNP spectrum is recorded immediately after the laser flash and corrected by a ‘‘dark’’ spectrum recorded without irradiation. Besides the equilibrium studies of protein surfaces, the technique can be combined with a stopped-flow apparatus and in this way it can be used to study folding intermediates. Using CIDNP pulse-labeling technique, the exposed tryptophan and tyrosine residues in a molten globule state can be identified.

HIGH-PRESSURE NMR SPECTROSCOPY

When high pressure is applied to a protein solution, it shifts the conformational equilibrium of the protein molecules toward lower volume conformers, thereby decreasing the partial molar volume of the protein. The combination of high pressure with hetero nuclear two-dimensional NMR spectroscopy provides atomic resolution information on the structure of the protein molecule at different stages of the folding process. By varying the pressure, one can explore the conformational space from the folded to the unfolded conformer. In recent years, numerous studies using high-pressure NMR spectroscopy have Protein folding been carried out on locally disordered, molten globule unfolded as well as oligomeric or aggregated states of proteins.

PROTEIN FOLDING AND DYNAMICS STUDIED BY MASS SPECTROMETRY

Mass spectrometry of protein molecules has become a rapidly developing field in the last decade. In comparison with NMR spectroscopy, which provides site-specific information averaged in time, mass spectrometry is capable of detecting different conformers coexisting in the protein solution. This method is especially useful for the study of low- populated intermediate states and is free of the

molecular size limitation of NMR spectroscopy. Because of its high sensitivity, a protein concentration in the femtomolar range is sufficient for analysis. Structural and dynamic properties of various conformational states can be studied by hydrogen/deuterium exchange (HDX) combined with mass spectrometry. Recently, Kaltashov and coworkers investigated the conformational ensemble of the molten globule state of ubiquitin. Using protein ion fragmentation in the gas phase, they evaluated the stability of various segments of the protein in the molten globular state. By the method of pulse-labeling HDX-MS, it is possible to study the kinetics of folding and to explore complex folding scenarios with parallel pathways. Co-populated protein conformers can be detected and characterized directly by electrospray ionization mass spectrometry (ESIMS). Protein surface areas in solution may be determined by ESIMS.

Limited proteolysis with ESIMS provides site-specific structural information on different conformational states of the protein molecules including protein aggregates and the amyloid state.

MECHANICAL UNFOLDING OF PROTEINS

In the first studies of the mechanical unfolding of single protein molecules using AFM, the giant sarcomeric protein titin, consisting of a large number of immunoglobulin segments, was used. Because of the heterogeneity of titin domains, it was not possible to assign the individual force peaks to specific domains. Using tandem repeats of a single domain, constructed by protein engineering techniques, it was possible to explain the mechanical characteristics of single domains in terms of their specific structures. Using force-measuring optical tweezers, it is possible to induce mechanical unfolding and refolding of individual molecules. In a recent work, Cecconi and coworkers showed that E. coli ribonuclease H molecule unfolds in a two-state manner and refolds through a transient molten globule-like intermediate. We may expect significant progress in the application of other new techniques such as the study of single-molecule folding kinetics by optical techniques in the near future.

1.5.17. MOLECULAR CHAPERONES

In molecular biology, **molecular chaperones** are proteins that assist the covalent folding or unfolding and the assembly or disassembly of other macromolecular structures. Chaperones are present when the macromolecules perform their normal biological functions and have correctly completed the processes of folding and/or assembly. The chaperones are concerned primarily with protein folding. The first protein to be called a chaperone assists the assembly of nucleosomes from folded histones and DNA and such assembly chaperones, especially in the nucleus, are concerned with the assembly of folded subunits into oligomeric structures.

One major function of chaperones is to prevent both newly synthesised polypeptide chains and assembled subunits from aggregating into nonfunctional structures. It is for this reason that many chaperones, but by no means all, are heat shock proteins because the tendency to aggregate increases as proteins are denatured by stress. In this case, chaperones do not convey any additional stericinformation required for proteins to fold. However, some highly specific 'steric chaperones' do convey unique structural (steric) information onto proteins, which cannot be folded spontaneously. Such proteins violate Anfinsen's dogma.

Various approaches have been applied to study the structure, dynamics and functioning of chaperones. Bulk biochemical measurements have informed us on the protein folding efficiency, and

prevention of aggregation when chaperones are present during protein folding. Recent advances in single-molecule analysis have brought insights into structural heterogeneity of chaperones, folding intermediates and affinity of chaperones for unstructured and structured protein chains.

Properties

- Molecular chaperones interact with unfolded or partially folded protein subunits, e.g. nascent chains emerging from the ribosome, or extended chains being translocated across subcellular membranes.
- They stabilize non-native conformation and facilitate correct folding of protein subunits.
- They do not interact with native proteins, nor do they form part of the final folded structures.
- Some chaperones are non-specific, and interact with a wide variety of polypeptide chains, but others are restricted to specific targets.
- They often couple ATP binding/hydrolysis to the folding process.
- Essential for viability, their expression is often increased by cellular stress.

Main role: They prevent inappropriate association or aggregation of exposed hydrophobic surfaces and direct their substrates into productive folding, transport or degradation pathways.

Location and Function

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

Some chaperone systems work as foldases: they support the folding of proteins in an ATP-dependent manner (for example, the GroEL/GroES or the DnaK/DnaJ/GrpE system). Other chaperones work as holdases: they bind folding intermediates to prevent their aggregation, for example DnaJ or Hsp33.

Macromolecular crowding may be important in chaperone function. The crowded environment of the cytosol can accelerate the folding process, since a compact folded protein will occupy less volume than an unfolded protein chain. However, crowding can reduce the yield of correctly folded protein by increasing protein aggregation. Crowding may also increase the effectiveness of the chaperone proteins such as GroEL, which could counteract this reduction in folding efficiency.

More information on the various types and mechanisms of a subset of chaperones that encapsulate their folding substrates (e.g. GroES) can be found in the chaperonins. Chaperonins are characterized by a stacked double-ring structure and are found in prokaryotes, in the cytosol of eukaryotes, and in mitochondria.

Other types of chaperones are involved in transport across membranes, for example membranes of the mitochondria and endoplasmic reticulum (ER) in eukaryotes. Bacterial translocation—specific

chaperone maintains newly synthesized precursor polypeptide chains in a translocation-competent (generally unfolded) state and guides them to the translocon.

New functions for chaperones continue to be discovered, such as assistance in protein degradation, bacterial adhesin activity, and in responding to diseases linked to protein aggregation (e.g. see prion) and cancer maintenance.

CHEPARONINE

Chaperonins are proteins that provide favourable conditions for the correct folding of other proteins, thus preventing aggregation. Newly made proteins usually must fold from a linear chain of amino acids into a three-dimensional form. Chaperonins belong to a large class of molecules that assist protein folding, called molecular chaperones. The energy to fold proteins is supplied by adenosine triphosphate

GroupI Chaperonins

GroupI Chaperonins are found in bacteria as welas organelles of endosymbiotic origin: chloroplasts and mitochondria. The GroEL/GroES complex in *E. coli* is a Group I chaperonin and the best characterized large (~ 1 MDa) chaperonin complex.

1.GroEL is a double-ring 14mer with a greasy hydrophobic patch at its opening and can accommodate the native folding of substrates 15-60 kDa in size.

2.GroES is a single-ring heptamer that binds to GroEL in the presence of ATP or transition state analogues of ATP hydrolysis, such as ADP-AlF_3. It's like a cover that covers GroEL (box/bottle).

GroEL/GroES may not be able to undo protein aggregates, but kinetically it competes in the pathway of misfolding and aggregation, thereby preventing aggregate formation.

Group II Chaperonins

Group II chaperonins, found in the eukaryotic cytosol and in archaea, are more poorly characterized. TRiC (TCP-1 Ring Complex, also called CCT for chaperonin containing TCP-1), the eukaryotic chaperonin, is composed of two rings of eight different though related subunits, each thought to be represented once per eight-membered ring. TRiC was originally thought to fold only the cytoskeletal proteins actin and tubulin but is now known to fold dozens of substrates.

Mm cpn (Methanococcus maripaludis chaperonin), found in the archaea Methanococcus maripaludis, is composed of sixteen identical subunits (eight per ring). It has been shown to fold the mitochondrial protein rhodanese; however, no natural substrates have yet been identified.

Group II chaperonins are not thought to utilize a GroES-type cofactor to fold their substrates. They instead contain a "built-in" lid that closes in an ATP-dependent manner to encapsulate its substrates, a process that is required for optimal protein folding activity.

Mechanism of action

Chaperonins undergo large conformational changes during a folding reaction as a function of the enzymatic hydrolysis of ATP as well as binding of substrate proteins and cochaperonins, such as GroES. These conformational changes allow the chaperonin to bind an unfolded or misfolded protein, encapsulate that protein within one of the cavities formed by the two rings, and release the protein back into solution. Upon release, the substrate protein will either be folded or will require further rounds of folding, in which case it can again be bound by a chaperonin.

The exact mechanism by which chaperonins facilitate folding of substrate proteins is unknown. According to recent analyses by different experimental techniques, GroEL-bound substrate proteins populate an ensemble of compact and locally expanded states that lack stable tertiary interactions. A number of models of chaperonin action have been proposed, which generally focus on two (not mutually exclusive) roles of chaperonin interior: passive and active. Passive models treat the chaperonin cage as an inert form, exerting influence by reducing the conformational space accessible to a protein substrate or preventing intermolecular interactions e.g. by aggregation prevention. The active chaperonin role is in turn involved with specific chaperonin–substrate interactions that may be coupled to conformational rearrangements of the chaperonin.

Probably the most popular model of the chaperonin active role is the iterative annealing mechanism (IAM), which focus on the effect of iterative, and hydrophobic in nature, binding of the protein substrate to the chaperonin. According to computational simulation studies, the IAM leads to more productive folding by unfolding the substrate from misfolded conformations or by prevention from protein misfolding through changing the folding pathway.

HUMAN CHAPERONE PROTEINS

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

Endoplasmic reticulum

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

- General chaperones: GRP78/BiP, GRP94, GRP170.
- Lectin chaperones: calnexin and calreticulin
- Non-classical molecular chaperones: HSP47 and ERp29
- Folding chaperones:
    - Protein disulfide isomerase (PDI),
    - *Peptidyl prolyl cis-trans-isomerase* (PPI)
    - ERp57

Nomenclature and examples of bacterial and archael chaperons.

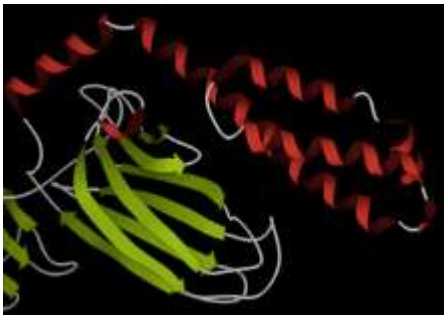
There are many different families of chaperones; each family acts to aid protein folding in a different way. In bacteria like *E. coli*, many of these proteins are highly expressed under conditions of high stress, for example, when the bacterium is placed in high temperatures. For this reason, the term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

Hsp60

Hsp60 (GroEL/GroES complex in *E. coli*) is the best characterized large (~ 1 MDa) chaperone complex. GroEL is a double-ring 14mer with a hydrophobic patch at its opening; it is so large it can accommodate native folding of 54-kDa GFP in its lumen. GroES is a single-ring heptamer that binds to GroEL in the presence of ATP or ADP. GroEL/GroES may not be able to undo previous aggregation, but it does compete in the pathway of misfolding and aggregation.[19] Also acts in mitochondrial matrix as molecular chaperone.

Hsp70

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



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

It has been noted that increased expression of Hsp70 proteins in the cell results in a decreased tendency toward apoptosis.Although a precise mechanistic understanding has yet to be determined, it is known that Hsp70s have a high-affinity bound state to unfolded proteins when bound to ADP, and a low-affinity state when bound to ATP. It is thought that many Hsp70s crowd around an unfolded substrate, stabilizing it and preventing aggregation until the unfolded molecule folds properly, at which time the Hsp70s lose affinity for the molecule and diffuse away. Hsp70 also acts as a mitochondrial and chloroplastic molecular chaperone in eukaryotes.

Hsp90

Hsp90 (HtpG in *E. coli*) may be the least understood chaperone. Its molecular weight is about 90 kDa, and it is necessary for viability in eukaryotes (possibly for prokaryotes as well).Heat shock protein 90 (Hsp90) is a molecular chaperone essential for activating many signaling proteins in the eukaryotic cell.Each Hsp90 has an ATP-binding domain, a middle domain, and a dimerization domain.

Hsp100

Hsp100 (Clp family in *E. coli*) proteins have been studied *in vivo* and *in vitro* for their ability to target and unfold tagged and mis folded proteins. Proteins in the Hsp100/Clp family form large hexameric structures with unfoldase activity in the presence of ATP. These proteins are thought to function as chaperones by processively threading client proteins through a small 20 Å (2 nm) pore, thereby giving each client protein a second chance to fold. Some of these Hsp100 chaperones, like ClpA and ClpX, associate with the double-ringed tetradecameric serine protease ClpP; instead of

catalyzing the refolding of client proteins, these complexes are responsible for the targeted destruction of tagged and misfolded proteins. Hsp104, the Hsp100 of Saccharomyces cerevisiae, is essential for the propagation of many yeast prions. Deletion of the HSP104 gene results in cells that are unable to propagate certain prions.

PROTEOSOME MEDIATED PROTEIN DEGRADATION

The ubiquitin/proteasome system (UPS) is the main eukaryotic cytosolic and nuclear proteolytic pathway serving for selective degradation of cellular proteins. By influencing protein abundance, the proteasome contributes to the dynamic state of cells, which allows a tight control of many biochemical pathways and cellular responses upon changes of the environment.

Ubiquitination and targeting

Proteins are targeted for degradation by the proteasome with covalent modification of a lysine residue that requires the coordinated reactions of three enzymes. In the first step, a ubiquitin-activating enzyme (known as E1) hydrolyzes ATP and adenylylates a ubiquitin molecule. This is then transferred to E1's active-site cysteine residue in concert with the adenylylation of a second ubiquitin.This adenylylated ubiquitin is then transferred to a cysteine of a second enzyme, ubiquitin-conjugating enzyme (E2). In the last step, a member of a highly diverse class of enzymes known as ubiquitin ligases (E3) recognizes the specific protein to be ubiquitinated and catalyzes the transfer of ubiquitin from E2 to this target protein. A target protein must be labeled with at least four ubiquitin monomers (in the form of a polyubiquitin chain) before it is recognized by the proteasome lid.It is therefore the E3 that confers substratespecificity to this system. The number of E1, E2, and E3 proteins expressed depends on the organism and cell type, but there are many different E3 enzymes present in humans, indicating that there is a huge number of targets for the ubiquitin proteasome system.

The mechanism by which a polyubiquitinated protein is targeted to the proteasome is not fully understood. Ubiquitin-receptor proteins have an N-terminal ubiquitin-like (UBL) domain and one or more ubiquitin-associated (UBA) domains. The UBL domains are recognized by the 19S proteasome caps and the UBA domains bind ubiquitin via three-helix bundles. These receptor proteins may escort polyubiquitinated proteins to the proteasome, though the specifics of this interaction and its regulation are unclear.

The ubiquitin protein itself is 76 amino acids long and was named due to its ubiquitous nature, as it has a highly conserved sequence and is found in all known eukaryotic organisms.The genes encoding ubiquitin in eukaryotes are arranged in tandem repeats, possibly due to the heavy transcription demands on these genes to produce enough ubiquitin for the cell. It has been proposed that ubiquitin is the slowest-evolving protein identified to date. Ubiquitin contains seven lysine residues to which another ubiquitin can be ligated, resulting in different types of polyubiquitin chains. Chains in which each additional ubiquitin is linked to lysine 48 of the previous ubiquitin have a role in proteasome targeting, while other types of chains may be involved in other processes.

Unfolding and translocation

After a protein has been ubiquitinated, it is recognized by the 19S regulatory particle in an ATP-dependent binding step. The substrate protein must then enter the interior of the 20S particle to come in contact with the proteolytic active sites. Because the 20S particle's central channel is narrow and gated by the N-terminal tails of the α ring subunits, the substrates must be at least partially unfolded

before they enter the core. The passage of the unfolded substrate into the core is called *translocation* and necessarily occurs after deubiquitination. However, the order in which substrates are deubiquitinated and unfolded is not yet clear. Which of these processes is the rate-limiting step in the overall proteolysis reaction depends on the specific substrate; for some proteins, the unfolding process is rate-limiting, while deubiquitination is the slowest step for other proteins. The extent to which substrates must be unfolded before translocation is not known, but substantial tertiary structure, and in particular nonlocal interactions such as disulfide bonds, are sufficient to inhibit degradation.

The gate formed by the α subunits prevents peptides longer than about four residues from entering the interior of the 20S particle. The ATP molecules bound before the initial recognition step are hydrolyzed before translocation. While energy is needed for substrate unfolding, it is not required for translocation.The assembled 26S proteasome can degrade unfolded proteins in the presence of a non-hydrolyzable ATP analog, but cannot degrade folded proteins, indicating that energy from ATP hydrolysis is used for substrate unfolding. Passage of the unfolded substrate through the opened gate occurs via facilitated diffusion if the 19S cap is in the ATP-bound state.

The mechanism for unfolding of globular proteins is necessarily general, but somewhat dependent on the amino acid sequence. Long sequences of alternating glycine andalanine have been shown to inhibit substrate unfolding, decreasing the efficiency of proteasomal degradation; this results in the release of partially degraded byproducts, possibly due to the decoupling of the ATP hydrolysis and unfolding steps. Such glycine-alanine repeats are also found in nature, for example in silk fibroin; in particular, certain Epstein-Barr virus gene products bearing this sequence can stall the proteasome, helping the virus propagate by preventing antigen presentation on the major histocompatibility complex.

Proteolysis

The mechanism of proteolysis by the β subunits of the 20S core particle is through a threonine-dependent nucleophilic attack. This mechanism may depend on an associated water molecule for deprotonation of the reactive threonine hydroxyl. Degradation occurs within the central chamber formed by the association of the two β rings and normally does not release partially degraded products, instead reducing the substrate to short polypeptides typically 7–9 residues long, though they can range from 4 to 25 residues, depending on the organism and substrate. The biochemical mechanism that determines product length is not fully characterized. Although the three catalytic β subunits have a common mechanism, they have slightly different substrate specificities, which are considered chymotrypsin-like, trypsin-like, and peptidyl-glutamyl peptide-hydrolyzing (PHGH)-like. These variations in specificity are the result of interatomic contacts with local residues near the active sites of each subunit. Each catalytic β subunit also possesses a conserved lysine residue required for proteolysis.

Although the proteasome normally produces very short peptide fragments, in some cases these products are themselves biologically active and functional molecules. Certain transcription factors regulating the expression of specific genes, including one component of the mammalian complex NF-κB, are synthesized as inactive precursors whose ubiquitination and subsequent proteasomal degradation converts them to an active form. Such activity requires the proteasome to cleave the substrate protein internally, rather than processively degrading it from one terminus. It has been suggested that long loops on these proteins' surfaces serve as the proteasomal substrates and enter the central cavity, while the majority of the protein remains outside. Similar effects have been

observed in yeast proteins; this mechanism of selective degradation is known as *regulated ubiquitin/proteasome dependent processing* (RUP).

Ubiquitin-independent degradation

Although most proteasomal substrates must be ubiquitinated before being degraded, there are some exceptions to this general rule, especially when the proteasome plays a normal role in the post-translational processing of the protein. The proteasomal activation of NF-κB by processing p105 into p50 via internal proteolysis is one major example. Some proteins that are hypothesized to be unstable due tointrinsically unstructured regions, are degraded in a ubiquitin-independent manner. The most well-known example of a ubiquitin-independent proteasome substrate is the enzyme ornithine decarboxylase. Ubiquitin-independent mechanisms targeting key cell cycle regulators such as p53 have also been reported, although p53 is also subject to ubiquitin-dependent degradation. Finally, structurally abnormal, misfolded, or highly oxidized proteins are also subject to ubiquitin-independent and 19S-independent degradation under conditions of cellular stress.

1.5.18. Protein folding errors

Proteins can miss function for several reasons. When a protein is miss folded it can lead to denaturation of the protein. Denaturation is the loss of protein structure and function. The miss folding does not always lead to complete lack of function but only partial loss of functionality. The miss functioning of proteins can sometimes lead to diseases in the human body.

Alzheimer's disease

Alzheimer's disease (AD) is a neurological degenerative disease that affects around 5 million Americans, including nearly half of those who are age 85 or older. The predominant risk factors of AD are age, family history, and heredity. Alzheimer’s disease typically results in memory loss, confusion of time and place, misplacing places, and changes in mood and behavior. AD results in dense plaques in the brain that are comprised of fibrillar β-amyloid proteins with a well-orders β-sheet secondary structure. These plaques visually look like voids in the brain figure matter and are directly connected to the deterioration of thought processes. It has been determined that AD is a protein misfolding disease, where the misfolded protein is directly related to the formation of these plaques in the brain.

Mad Cow

Diseases caused by prions, like Mad Cow / Creutzfeldt-Jacob are also, in essence, protein folding disorders. These are caused by a certain protein, named PrP, that will stay in a mis-folded conformation (PrPsc) if encouraged to go into it in the first place. In most people, the PrP protein folds normally, leaving the person healthy. Rarely, a mutation in the PrP gene will allow the protein to be made incorrectly, and it will fold incorrectly, making a PrPsc prion. These prions, when exposed to PrP which is in the process of folding, will encourage that PrP to fold badly too, thus creating another PrPsc. While PrP can be processed and cleaned out of a cell once it has been used, PrPsc is shaped differently enough that it can't be, so it never goes away. PrPsc, much more quickly than with Ab in Alzheimer's, builds up into plaques, handily destroying whatever nervous tissue it's building up in. See the writeups under prion for more on this**.**

Cystic Fibrosis

Besides building up un-processable plaques, protein folding errors can leave behind too little of the effective conformation for it to do its job. This is the case with diseases like Cystic Fibrosis, and many other hereditary diseases. Cystic Fibrosis results from lack of a protein that regulates chloride ion transport through a cell membrane. Findings show that while this protein seems to be forming correctly, there is a problem with one of its associated chaperone proteins. Chaperone proteins help encourage unfolded proteins to fold in the right way by surrounding them and protecting their movement. In Cystic Fibrosis, the chaperone doesn't pull away from the transport protein smoothly, leaving it partially mis-folded and useless. The broken chaperone protein then moves on to do the same thing to another transport protein, and so forth.

1.6. Combinational manipulation of polyketides and non ribosomal peptides

Non-ribosomal peptides are small peptides synthesised mainly by bacteria and fungi. Despite their small size, they are highly diverse in terms of the monomers that can be incorporated. According to the most recent published reports there are 1,164 different non-ribosomal peptides known , which collectively contain over 500 unique monomers, including both proteinogenic and non-proteinogenic L- and D-amino acids, as well as carboxylic acids and amines. Non-ribosomal peptides also exhibit high structural diversity with only 27 % being linear; the remainder having cyclic, branched or other complex primary structures

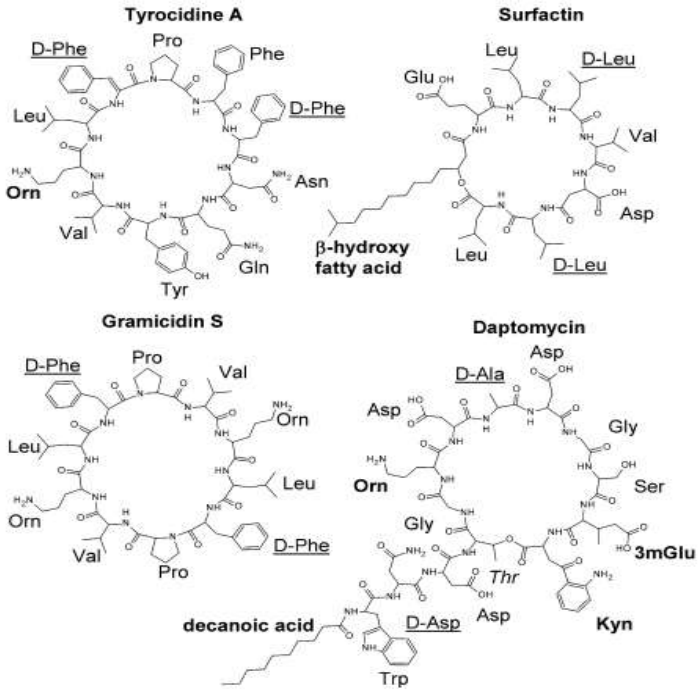


Fig. 1 Structures of some non-ribosomal peptides relevant to biotechnology that are highlighted in this review. Non-proteinogenic amino acids or substituents are labelled in bold, D-isomers are underlined, and a threonine residue in daptomycin that is involved in an atypical ester bond via the side-chain hydroxyl is labelled in italics. Orn, ornithine; 3mGlu, 3-methylglutamate; Kyn, kynurenine; standard three letter abbreviations are used for proteinogenic amino acids

The diversity of non-ribosomal peptides imparts to them many properties of relevance to biotechnology; for example, peptides have been identified with antibiotic, antiviral, anti-cancer, anti-inflammatory, immunosuppressant and surfactant qualities. Importantly for medicine, natural products often need to be modified to improve clinical properties and/or bypass resistance mechanisms. Due to their typically complex structures, most clinical natural product derivatives are

created by means of semisynthesis; a process whereby the natural product is chemically modified post-isolation from biological sources.

1.6.1. The multiple template model of non-ribosomal peptide synthesis

Non-ribosomal peptide synthesis generally follows the multiple template model. According to this model, peptides are synthesised in a modular assembly line-like manner by NRPS enzymes (‘‘the template’’). The modules that comprise an NRPS template may be clustered on a single enzyme or located within multiple distinct enzymes that associate post-translation; and are classified as either initiation, elongation, or termination modules depending on their location in the assembly line (Fig. 2a). Modules act in a concerted but semiautonomous fashion, and are defined by their ability to recognise, activate and incorporate a specific monomer into the final peptide product.

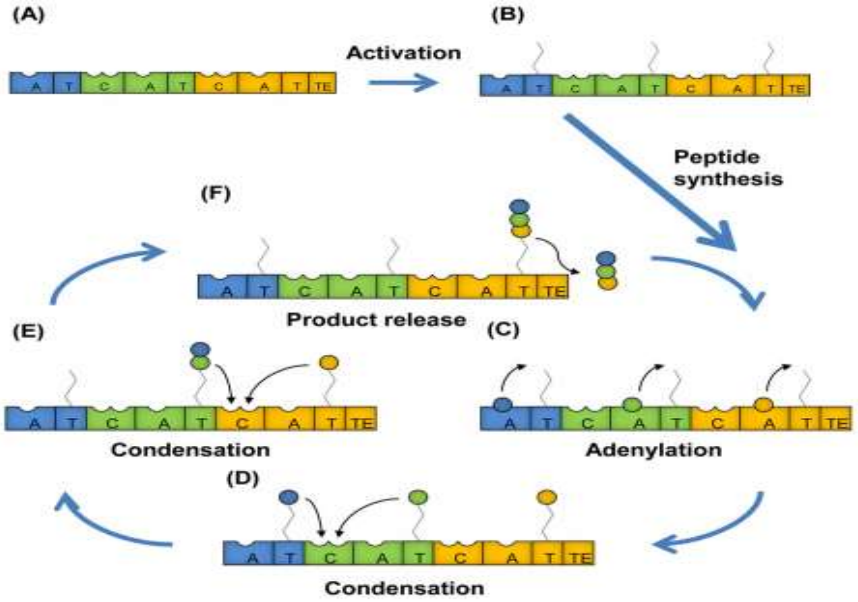


Fig. 2 The multiple template model of non-ribosomal peptide synthesis. **a** A schematic arrangement of domains within a hypothetical three module NRPS that contains an initiation (*blue*), an elongation (*green*) and a termination (*orange*) module. There may be multiple elongation modules present within a single NRPS template. **b** Activation of NRPS modules by post-translational attachment of a 4′-phosphopantetheine (PPT) cofactor to each T domain. **c** Domains within a module act in a semi-autonomous fashion, beginning with the A domain, which activates and tethers a specific monomer substrate to the PPT prosthetic group attached to the T domain immediately downstream. **d** The substrate linked to each T domain is then passed to a C domain, which catalyses peptide bond formation between the donor substrate provided by the upstream module, and the acceptor substrate from the downstream module. C domains are generally located immediately upstream of A domains and C-A-T domain units comprise a basic elongation module. **e** Peptide bond formation hydrolyses the upstream thioester bond, yielding a peptide that is now attached to the downstream T domain, and which goes on to serve as the donor substrate at the C domain of the following module. **f** After addition of the final monomer by the termination module, the peptide product is released by a thioesterase (TE) domain; most commonly by intra-molecular cyclisation to yield a macrocyclic lactone or lactam, or by hydrolysis to yield a linear peptide product. Following product release, the NRPS is returned to step (e) and peptide synthesis can repeat in an iterative fashion

Within each module, an adenylation (A) domain recognises and activates a specific substrate by addition of AMP (Fig. 2c). The activated substrate is then tethered to a flexible 40 - phosphopantetheine (PPT) prosthetic group, which is itself covalently attached to a thiolation (T) domain (also known as a peptidyl carrier protein (PCP) domain) (Fig. 2b). The T domain lies at the heart of the biosynthetic process, with its flexible PPT prosthesis effectively the ‘‘swinging arm’’ of a biomolecular assembly line that transfers peptide intermediates between different domains and modules. Post-attachment of an activated substrate by its A domain partner, a T domain then passes that substrate to a condensation (C) domain, which catalyses peptide bond formation between the donor substrate provided by the T domain immediately upstream, and the acceptor substrate provided by the downstream T domain (Fig. 2d). Following the initial condensation event, the process can

repeat in an iterative fashion, with the previous peptide intermediate now serving as the donor substrate for the C domain of the next module in an NRPS complex (Fig. 2e). Along the way, certain modules may contain additional tailoring domains that modify individual substrates in a directed fashion (e.g., epimerisation (E) domains, for conversion from L- to D-enantiomers). The growing peptide continues to be passed from the T domain of one module to the T domain of the next until the product is released, typically via a hydrolysis or intramolecular cyclisation reaction catalysed by a thioesterase (TE) domain associated with the final module in an NRPS complex (Fig. 2f).

1.6.2. Strategies to create novel peptide products via genetic manipulation of NRPS templates

The modular structure of the NRPS assembly line suggests that it should be possible to rationally alter one or more residues in a non-ribosomal peptide product by substitution or engineering of the module(s) that specify the target residue(s). In nature, the diversity of non-ribosomal peptides is thought to have arisen from point mutation, substitution of domains or modules for alternatives that specify different substrates, and/or the insertion/deletion of modules. The modular structure of the NRPS assembly line suggests that it should be possible to rationally alter one or more residues in a non-ribosomal peptide product by substitution or engineering of the module(s) that specify the target residue(s). In nature, the diversity of non-ribosomal peptides is thought to have arisen from point mutation, substitution of domains or modules for alternatives that specify different substrates, and/or the insertion/deletion of modules.

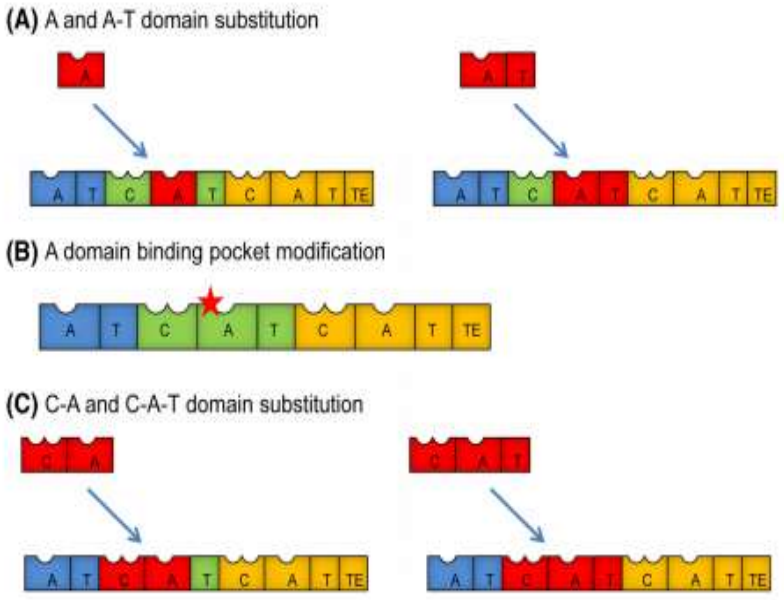


Fig. 3 The three main strategies employed to re-engineer NRPS templates at a genetic level. **a** Substitution of the substrate-specifying A domain, with or without its native T domain partner. **b** Modulation of A domain substrate specificity by site-directed mutagenesis, to avoid major perturbation of tertiary and quaternary structure. **c** Strategies that treat C and A domains as inseparable pairs, to address the phenomenon of strong C domain substrate specificity at the acceptor site

1.6.3. Creation of novel peptide products via A domain substitution

Due to their similarity to A domain substitutions, paired A-T domain substitutions will also be considered in this section (Fig. 3a). The first reported efforts to alter the products of NRPS enzymes by domain substitution targeted a leucine-specifying A-T domain pair, in the termination module of the Bacillus subtilis surfactin NRPS template SrfA-C.

Non-ribosomal peptide synthetases (NRPS) are large modular enzymes that govern the synthesis of numerous biotechnologically relevant products. Their mode of action is frequently compared to an assembly line, in which each module acts in a semi-autonomous but coordinated manner to add a

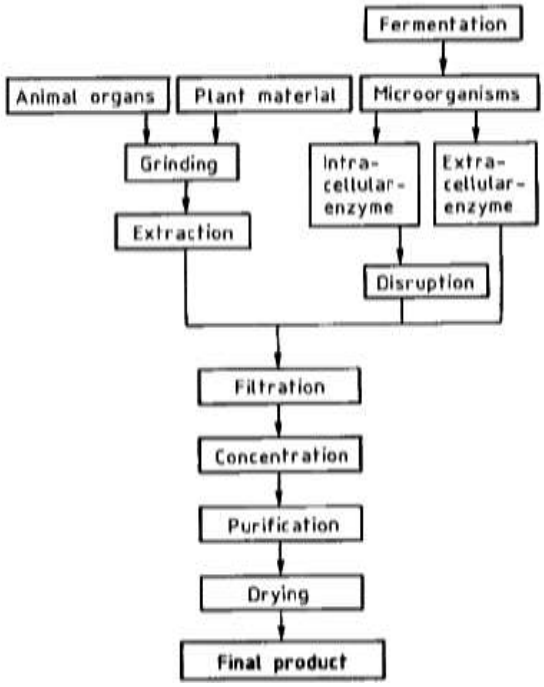
specific monomer to a growing peptide chain, unfettered by ribosomal constraints. The modular nature of these systems offers tantalising prospects for synthetic biology, wherein the assembly line is re-engineered at a genetic level to generate a specific or combinatorial modified product. However, despite some success stories, a ‘‘one size fits all’’ approach to NRPS synthetic biology remains elusive. This review examines both rational and random mutagenesis strategies that have been employed to modify NRPS function, in an attempt to highlight key points that should be considered when seeking to reengineer an NRPS biosynthetic template.

UNIT – II - ENZYME AND PROTEIN ENGINEERING – SBTA5202

2. Production and purification of crude enzyme extracts from plant, animal and microbial sources

The degree of purity of commercial enzymes ranges from raw enzymes to highly- purified forms and depends on the application. Raw materials for the isolation of enzymes are animal organs, plant material, and microorganisms. Enzymes are universally present in living organisms; each cell synthesizes a large number of different enzymes to maintain its metabolic reactions. The choice of procedures for enzyme purification depends on their location. Isolation of intracellular enzymes often involves the separation of complex biological mixtures. On the other hand, extracellular enzymes are generally released into the medium with only a few other components. Enzymes are very complex proteins, and their high degree of specificity as catalysts is manifest only in their native state. The native conformation is attained under specific conditions of pH, temperature, and ionic strength.Hence, only mild and specific methods can be used for enzyme isolation.

Sequence of steps in the isolation of enzymes



2.1 Preparation of Biological Starting Materials

Animal Organs : Animal organs must be transported and stored at low temperature to retain enzymatic activity. The organs should be freed of fat and connective tissue before freezing. Frozen organs can be minced with machines generally used in the meat industry, and the enzymes can be extracted with a buffer solution. Besides mechanical
grinding, enzymatic digestion can also be employed. Fat attached to the organs interferes with subsequent purification steps and can be removed with organic solvents. However, enzymatic activity might be influenced negatively by this procedure.

Plant Material : Plant material can be ground with various crushers or grinders, and the desired enzymes can be extracted with buffer solutions. The cells can also be disrupted by previous treatment with lytic enzymes.

Microorganisms : Microorganisms are a significant source of enzymes. New techniques, summarized under genetic and protein engineering, have much to offer the enzyme industry.A gene

can be transferred into a microorganism to make that organism produce a protein it did not make naturally. Alternatively, modification of the genome of a microorganism can change the properties of proteins so that they may be isolated and purified more easily. Such modifications might, for example, cause the release of intracellular enzymes into the medium; change the net charge and, therefore, the chromatographic properties of proteins; or lead to the formation of fused proteins.

Most enzymes used commercially are extracellular enzymes, and the first step in their isolation is separation of the cells from the solution. For intracellular enzymes, which are being isolated today in increasing amounts, the first step involves grinding to rupture the cells. A number of methods for the disruption of cells are known, corresponding to the different types of cells and the problems involved in isolating intracellular enzymes. However, only a few of these methods are used on an industrial scale.

Mechanical methods	Nonmechanical methods
High pressure (Manton–Gaulin, French-press)	Drying (freeze-drying, organic solvents)
Grinding (ball mill)	Lysis
Ultrasound	physical: freezing, osmotic shock
	chemical: detergents, antibiotics
	enzymatic: enzymes (e.g., lysozyme), antibiotics

2.1.1 Cell Disruption by Mechanical Methods

High-pressure homogenization is the most common method of cell disruption. The cell suspension is pressed through a valve and hits an impact ring (e.g., Manton–Gaulin homogenizer). The cells are ruptured by shearing forces and simultaneous decompression. Depending on the type of machine, its capacity ranges from 50 to 5000 L/h.

The rigid cell walls of small bacteria are only partially ruptured at the pressures up to 55 MPa (550 bar) achieved by this method. Higher pressures, however, would result in further heat exposure (2.2 8C per 10 MPa). Hence, the increased enzyme yield resulting from improved cell disruption could be counteracted by partial inactivation

caused by heating and higher shearing forces. Therefore, efficient cooling must be provided.

The ***wet grinding*** of cells in a high-speed bead mill is another effective method of cell disruption [190–193]. Glass balls with a diameter of 0.2–1 mm are used to break the cells. The efficiency of this method depends on the geometry of the stirrer system. A symmetrical arrangement of circular disks gives better results than the normal asymmetrical arrangement. Given optimal parameters such as stirring rate, number and size of glass beads, flow rate, cell concentration, and temperature, a protein release of up to 90 % can be achieved in a single passage.

2.1.2 Cell Disruption by Nonmechanical Methods

Cells may frequently be disrupted by ***chemical, thermal, or enzymatic lysis***. The drying of microorganisms and the preparation of acetone powders are standard procedures in which the structure of the cell wall is altered to permit subsequent extraction of the cell contents.Methods based on enzymes or autolysis have been described in the literature. Ultrasound is generally used in the laboratory. In this procedure, cells are disrupted by shearing forces and cavitation. An optimal temperature must be maintained by cooling the cell suspension because heat is generated in the process.

2.2 Separation of Solid Matter

After cell disruption, the next step is separation of extracellular or intracellular enzymes from cells or cellular fragments, respectively. This operation is rather difficult because of the small size of

bacterial cells and the slight difference between the density of the cells and that of the fermentation medium. Continuous filtration is used in
industry. Large cells, e.g., yeast cells, can be removed by decantation. Today, efficient centrifuges have been developed to separate cells and cellular fragments in a continuous process. Residual plant and organ matter can be separated with simpler centrifuges or filters.

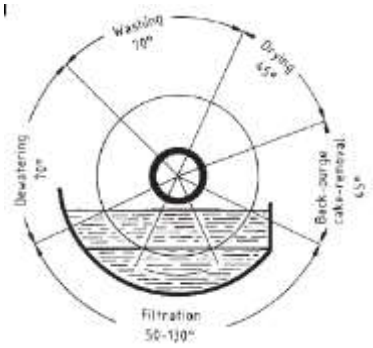
2.2.1 Filtration

The filtration rate is a function of filter area, pressure, viscosity, and resistance offered by the filter cake and medium. For a clean liquid, all these terms are constant which results in a constant flow rate for a constant pressure drop. The cumulative filtrate volume increases linearly with time. During the filtration of suspensions, the increasing thickness of the formed filter cake and the concomitant resistance gradually decrease the flow rate. Additional difficulties may arise because of the compressibility of biological material. In this case, the resistance offered by the filter cake and, hence, the rate of filtration depend on the pressure applied. If the pressure applied exceeds a certain limit, the cake may collapse and total blockage of the filter can result.

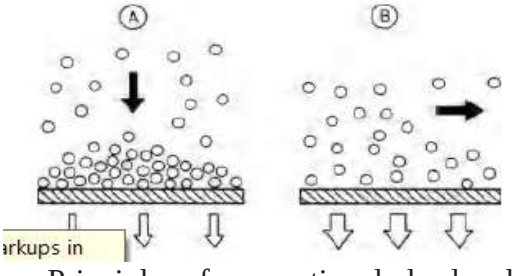
Pressure Filters A filter press (plate filter, chamber filter) is used to filtrate small volumes or to remove precipitates formed during purification. The capacity to retain solid matter is limited, and the method is rather work-intensive. However, these filters are highly suitable for the fine filtration of enzyme solutions.

Vacuum Filters Vacuum filtration is generally the method of choice because biological materials are easily compressible. A rotary vacuum filter is used in the continuous filtration of large volumes. The suspension is usually mixed with a filter aid, e.g., kieselguhr, before being applied to the filter. The filter drum is coated with a thin
layer of filter aid (precoat). The drum is divided into different sections so that the filter cake can also be washed and dried on the filter. The filter cake is subsequently removed by using a series of endless strings or by scraper discharge (knife). The removal of a thin layer of precoat each time exposes a fresh filtering area. This system is useful for preventing an increase in resistance with the accumulation of filter cake during the course of filtration.

Cross-Flow Filtration In conventional methods, the suspension flows perpendicular to the filtering material. In cross-flow filtration, the input stream flows parallel to the filter area, thus preventing the accumulation of filter cake and an increased resistance to filtration. To maintain a sufficiently high filtration rate, thismethodmust consume a relatively large amount of energy, in the form of high flux rates over the membranes. With the membranes now available, permeate rates can be attained. Indeed, in many cases the use of a separator is more economical.



Rotary vacuum filter



Principles of conventional, dead end filtration (A) and cross-flow filtration (B)

The future of this method depends on the development of suitable membranes, but cross-flow filtration can be
conveniently used in recombinant DNA techniques to separate organisms in a closed system.

2.2.2 Centrifugation

The sedimentation rate of a bacterial cell with a diameter of 0.5 mmis less than 1 mm/h. An economical separation can be achieved only by sedimentation in a centrifugal field. The range of applications of centrifuges depends on the particle size and the solids content.

Type of centrifuge	Solids content, %	Particle size, µm
Multichamber separator	0–5	0.5–500
Desludging disk separator	β–10	0.5–500
Nozzle separator	5–25	0.5–500
Decanter	5–40	5–50 000
Sieve centrifuge	5–60	5–10 000
Pusher centrifuge	20–75	100–50 000

Decanters (scroll-type centrifuges) work with low centrifugal forces and are used in the separation of large cells or protein precipitates. Solid matter is discharged continuously by a screw conveyer moving at a differential rotational speed.

Tubular bowl centrifuges are built for very high centrifugal forces and can be used to sediment very small particles. However, these centrifuges cannot be operated in a continuous process. Moreover, solid matter must be removed by hand after the centrifuge has come to a stop. A further disadvantage is the appearance of aerosols.

Separators (disk stack centrifuges) can be used in the continuous removal of solid matter from suspensions. Solids are discharged by a hydraulically operated discharge port (intermittent discharge) or by an arrangement of nozzles (continuous discharge). Bacteria and cellular fragments can be separated by a combination of high centrifugal
forces, up to 15 000 _ gravity, presently attainable, and short sedimentation distances. Disk stack centrifuges that can be sterilized with steam are used for recombinant DNA techniques in a closed system.

2.2.3 Extraction

An elegant method used to isolate intracellular enzymes is liquid–liquid extraction in an aqueous two-phase system This method is based on the incomplete mixing of different polymers, e.g., dextran and poly(ethylene glycol), or a polymer and a salt in an aqueous solution [208]. The first extraction step separates cellular fragments. Subsequent purification can be accomplished by extraction or, if high purity is required, by other methods. The extractability can be improved by using affinity ligands or modified chromatography gels, e.g., phenyl-Sepharose.

2.2.3.1 Flocculation and Flotation Flocculation Separation of bacterial cells or cell debris by filtration or centrifugation can involve considerable difficulties due to their small size and physical properties. The compressible nature of the cells is the primary limiting factor for using filtration as a separation step to remove them. The low permeability of a typical cell cake results in a filtration rate that is often too slow to be practical. In cell removal by centrifugation, the small size and low density difference between the cells or cell debris and the medium results in a low sedimentation rate. Flocculation of cell suspensions has been reported to aid cell separation by both filtration and centrifugation.

Flocculation is the process whereby destabilized particles are induced to come together, make contact, and subsequently form larger aggregates. Flocculating agents are additives capable of increasing the degree of flocculation of a suspension. They can be organic or inorganic, and natural or synthetic. A comprehensive review of various categories of flocculating agents can be found in.

Synthetic organic flocculating agents are by far the most commonly used agents for cell flocculation in industrial processes. They are typically water-soluble, charged polymeric substances with average molecular weight ranging from about 10^3 to greater than $5x10^6$ and are generally referred to as polyelectrolytes. The positively and negatively charged polymers are referred to as cationic and anionic polyelectrolytes, respectively. Polyelectrolytes containing both positive and negative charges are termed polyampholytes. Flocculation of cells by polyelectrolytes is a two-step process. The first step is the neutralization of the surface charge on the suspended cells or cell debris. The second step involves the linkage of these particles to form large aggregates. The various mechanisms and theories of flocculation have been summarized. Flocculant selection for a specific cell separation process is a challenge as many factors can impact flocculation. These factors can have their origin in the broth (cell surface charge and size, ionic strength, pH, cell concentration, and the presence of other charged matter), the polymer (molecular weight, charge and charge density, structure, type), and engineering parameters (mixing and mode and order of addition). The final criteria for flocculant selection should take into consideration all aspects of the flocculation process. These include the cost of the added flocculant, subsequent separation performance, process robustness, and yield. In some cases, flocculation can also provide purification by selectively removing unwanted proteins, nucleic acids, lipids and endotoxin from the cell broth.

Flotation If no stable agglomerates are formed, cells can be separated by flotation. Here, cells are adsorbed onto gas bubbles, rise to the top, and accumulate in a froth. An example is the separation of single cell protein.

2.3 Concentration

The enzyme concentration in starting material is often very low. The volume of material to be processed is generally very large, and substantial amounts of waste material must be removed. Thus, if economic purification is to be achieved, the volume of starting material must be decreased by concentration. Only mild concentration procedures that do not inactivate enzymes can be employed. These include thermal methods, precipitation, and to an increasing extent, membrane filtration.

2.3.1Thermal Methods

Only brief heat treatment can be used for concentration because enzymes are thermolabile. Evaporators with rotating components that achieve a thin liquid film (thin-layer evaporator, centrifugal thin-layer evaporator) or circulation evaporators (long-tube evaporator) can be employed.

1.3.2 Precipitation

Enzymes are very complex protein molecules possessing both ionizable and hydrophobic groups which interact with the solvent. Indeed, proteins can be made to agglomerate and, finally, precipitate by changing their environment. Precipitation is actually a simple procedure for concentrating enzymes.

Precipitation with Salts High salt concentrations act on the water molecules surrounding the protein and change the electrostatic forces responsible for solubility. Ammonium sulfate is commonly used for precipitation; hence, it is an

effective agent for concentrating enzymes. Enzymes can also be fractionated, to a limited extent, by using different concentrations of ammonium sulfate. The corrosion of stainless steel and cement by ammonium sulfate is a disadvantage, which causes additional problems in wastewater treatment. Sodium sulfate is more efficient from this point of view, but it is less soluble and must be used at

temperatures of 35–40 8C. The optimal concentration of salt required for precipitation must be determined experimentally, and generally ranges from 20 to 80 % saturation.

Precipitation with Organic Solvents Organic solvents influence the solubility of enzymes by reducing the dielectric constant of the medium. The solvation effect of water molecules surrounding the enzyme is changed; the interaction of protein molecules is increased; and therefore, agglomeration and precipitation occur. Commonly used solvents are ethanol and acetone. Satisfactory results are obtained only if the concentration of solvent and the temperature are carefully controlled because enzymes can be inactivated easily by organic solvents.

Precipitation with Polymers The polymers generally used are polyethylenimines and poly(ethylene glycols) of different molecular masses. The mechanism of this precipitation is similar to that of organic solvents and results from a change in the solvation effect of the water molecules surrounding the enzyme. Most enzymes precipitate at polymer concentrations ranging from 15 to 20 %.

Precipitation at the Isoelectric Point Proteins are ampholytes and carry both acidic and basic groups. The solubility of proteins is markedly influenced by pH and is minimal at the isoelectric point at which the net charge is zero. Because most proteins have isoelectric points in the acidic range, this process is also called acid precipitation.

2.3.2 Ultrafiltration

A semipermeable membrane permits the separation of solvent molecules from larger enzyme molecules because only the smaller molecules can penetrate the membrane when the osmotic pressure is exceeded. This is the principle of all membrane separation processes, including ultrafiltration. In reverse osmosis, used to separate materials with low molecular mass, solubility and diffusion phenomena influence the process, whereas ultrafiltration and cross-flow filtration are based solely on the sieve effect. In processing enzymes, cross-flow filtration is used to harvest cells, whereas ultrafiltration is employed for concentrating and desalting.

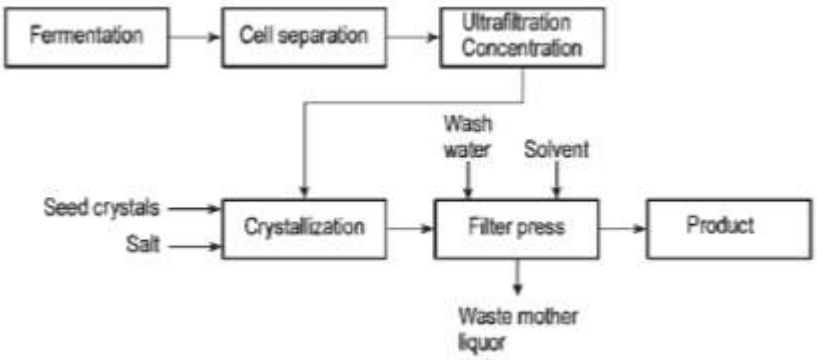
Process	Application	Separation range, M_r
Cross-flow microfiltration	Concentration of bacteria, removal of cell debris	>1 000 000 (or particles)
Ultrafiltration	Concentration of enzymes, dialysis, fractionation	>10 000 (macromolecules)
Reverse osmosis	Concentration of small molecules, desalting	>200

2.4 Purification

For many industrial applications, partially purified enzyme preparations will suffice; however, enzymes for analytical purposes and for medical use must be highly purified. Special procedures employed for enzyme purification are crystallization, electrophoresis, and chromatography.

2.4.1 Crystallization

The rapid growth in the utilization of enzymes in commercial sectors such as agriculture and consumer products requires a cost-effective, industrial-scale purification method. Crystallization, one of the oldest chemical purification technologies, has the potential to fulfill these requirements. Enzyme crystallization is the formation of solid enzyme particles of defined shape and size. An enzyme can be induced to crystallize or form protein-protein interactions by creating solvent conditions that result in enzyme supersaturation. The theory and history of protein crystallization are well documented.Much of the emphasis in enzyme crystallization has focused on obtaining crystals for X-ray diffraction analysis rather than as a purification process.



2.4.2 Electrophoresis

Electrophoresis is used to isolate pure enzymes on a laboratory scale. Depending on the conditions, the following procedures can be used: zone electrophoresis, isotachophoresis, or porosity gradients. The heat generated in electrophoresis and the interference caused by convection are problems associated with a scale-up of this method. An interesting contribution to the industrial application of electrophoresis is a continuous process in which the electrical field is stabilized by rotation.

2.4.3 Chromatography

Chromatography is of fundamental importance to enzyme purification. Molecules are separated according to their physical properties (size, shape, charge, hydrophobic interactions), chemical properties (covalent binding), or biological properties (biospecific affinity). In *gel chromatography* (also called gel filtration), hydrophilic, cross-linked gels with pores of finite size are used in columns to separate biomolecules. Concentrated solutions are necessary for separation because the sample volume that can be applied to a column is limited to ca. 10 % of the column volume. In gel filtration, molecules areseparated according to size and shape. Molecules larger than the largest pores in the gel beads, i.e., above the exclusion limit, cannot enter the gel and are eluted first. Smaller molecules, which enter the gel beads to varying extent depending on their size and shape, are retarded in their passage through the column and eluted in order of decreasing molecular mass. Gel filtration is used commercially for both separation and desalting of enzyme solutions.

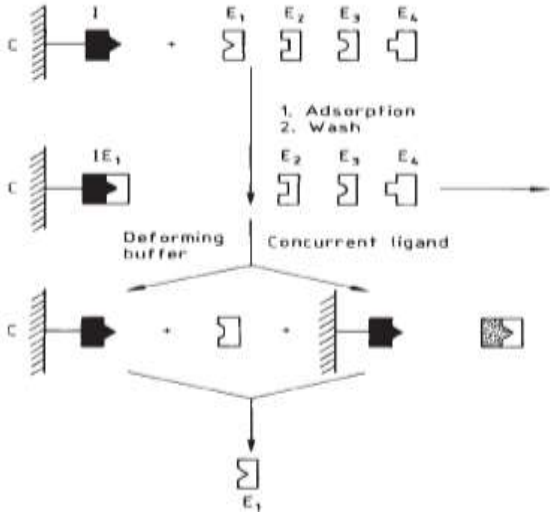
Type of chromatography	Principle	Separation according to
Adsorption	surface binding	surface affinity
Distribution	distribution equilibrium	polarity
Ion exchange	ion binding	charge
Gel filtration	pore diffusion	molecular size, molecular shape
Affinity	specific adsorption	molecular structure
Hydrophobic	hydrophobic chelation	molecular structure
Covalent	covalent binding	polarity
Metal chelate	complex formation	molecular structure

Ion-exchange chromatography is a separation technique based on the charge of protein molecules. Enzyme molecules possess positive and negative charges. The net charge is influenced by pH, and this property is used to separate proteins by chromatography on anion exchangers (positively charged) or cation exchangers (negatively charged). The sample is applied in aqueous solution at low ionic strength, and elution is best carried out with a salt gradient of increasing concentration. Because of the concentrating effect, samples can be applied in dilute form.

For *hydrophobic chromatography*, media derived from the reaction of CNBr-activated Sepharose with aminoalkanes of varying chain length are suitable. This method is based on the interaction of hydrophobic areas of protein molecules with hydrophobicgroups on the matrix. Adsorption occurs at high salt concentrations, and fractionation

of bound substances is achieved by eluting with a negative salt gradient. This method is ideally suited for further purification of enzymes after concentration by precipitation with such salts as ammonium sulfate.

In *affinity chromatography*, the enzyme to be purified is specifically and reversibly adsorbed on an effector attached to an insoluble support matrix. Suitable effectors are substrate analogues, enzyme inhibitors, dyes, metal chelates, or antibodies.The insoluble matrix (C) is contained in a column. The biospecific effector, e.g., an enzyme inhibitor (I), is attached to the matrix. A mixture of different enzymes (E1, E2, E3, E4) is applied to the column. The immobilized effector specifically binds the complementary enzyme. Unbound substances are washed out, and the enzyme of interest (E1) is recovered by changing the experimental conditions, for example by altering pH or ionic strength.



Immunoaffinity chromatography occupies a unique place in purification technology. In this procedure, monoclonal antibodies are used as effectors. Hence, the isolation of aspecific substance from a complex biological mixture in one step is possible. In this procedure, enzymes can be purified by immobilizing antibodies specific for the desired enzyme.Amore generalmethod offers the synthesis of a fusion proteinwith proteinAby ‘‘protein engineering’’. Protein A is a Staphylococcus protein with a high affinity formany immunoglobulins, especially of the IgG class of antibodies. In this way, enzymes that usually do not bind to an antibody can be purified by immunoaffinity chromatography.

Covalent chromatography differs from other types of chromatography in that a covalent bond is formed between the required protein and the stationary phases.

2.5. METHODS FOR ENZYME CHARACTERISATION

Enzyme characterisation simply refers to the determination of the various chemical and physical properties (characteristics) of an enzyme. It involves the use of a series of laboratory procedures. Examples of assays for commonly characterised enzyme properties in biochemical researches include:

1. Determination of the effect of changes in temperature on enzyme’s activity and optimum temperature.
2. Determination of enzyme’s thermal stability.

3. Determination of the effect of changes in pH on enzyme's activity and optimum pH.
4. Determination of pH stability
5. Determination of the effect of changes in substrate concentration on enzyme's activity and kinetic constants e.g Vmax, Km, Kcat, Km/Kcat etc.
6. Determination of substrate specificity
7. Determination of molecular weight (Mw) of enzyme
8. Determination of the effect of metal ions, chelating agent or denaturating agents.
9. Determination of enzyme's isoelectric point (pI)
10. Determination of the effect of duration of incubation
11. Determination of active site fractional saturation (V/Vmax) at a particular substrate concentration.
12. Determination of Enzyme's turnover number (Vmax/ E T)
13. Determination of activation energy (Ea)
14. Determination of salt tolerance

2.6. DEVELOPMENT OF ENZYMATIC ASSAYS

Enzyme assays are laboratory methods for measuring enzymatic activity. They are vital for the study of enzyme kinetics and enzyme inhibition.

2.6.1. Enzyme Units

The quantity or concentration of an enzyme can be expressed in molar amounts, as with any other chemical, or in terms of activity in enzyme units.

2.6.2. Enzyme activity

Enzyme activity = moles of substrate converted per unit time = rate × reaction volume.

Enzyme activity is a measure of the quantity of active enzyme present and is thus dependent on conditions, *which should be specified*.

The SI unit is the katal, 1 katal = 1 mol s^{-1}, but this is an excessively large unit.

A more practical and commonly used value is enzyme unit (U) = 1 μmol min^{-1}.

Enzyme activity as given in katal generally refers to that of the assumed natural target substrate of the enzyme. Enzyme activity can also be given as that of certain standardized substrates, such as gelatin, then measured in *gelatin digesting units* (GDU), or milk proteins, then measured in *milk clotting units* (MCU). The units GDU and MCU are based on how fast one gram of the enzyme will digest gelatin or milk proteins, respectively. 1 GDU equals approximately 1.5 MCU.

An increased amount of substrate will increase the rate of reaction with enzymes, however once past a certain point, the rate of reaction will level out because the amount of active sites available has stayed constant.

2.6.3. Specific activity

The specific activity of an enzyme is another common unit.

This is the activity of an enzyme per milligram of total protein (expressed in μmol min^{-1} mg^{-1}).

Specific activity gives a measurement of enzyme purity in the mixture. It is the micro moles of product formed by an enzyme in a given amount of time (minutes) under given conditions per milligram of total proteins. Specific activity is equal to the rate of reaction multiplied by the volume of reaction divided by the mass of total protein. The SI unit is katal/kg, but a more practical unit is μmol/mgmin.

Specific activity is a measure of *enzyme processivity* (the capability of enzyme to be processed), at a specific (usually saturating) substrate concentration, and is usually constant for a pure enzyme.

An active site titration process can be done for the elimination of errors arising from differences in cultivation batches and/or misfolded enzyme and similar issues. This is a measure of the amount of active enzyme, calculated by e.g. titrating the amount of active sites present by employing an irreversible inhibitor. The specific activity should then be expressed as μmol min^{-1} mg^{-1} active enzyme. If the molecular weight of the enzyme is known, the turnover number, or μmol product per second per μmol of active enzyme, can be calculated from the specific activity. The turnover number can be visualized as the number of times each enzyme molecule carries out its catalytic cycle per second.

2.6.4. Purification fold: This factor or parameter provides information on the degree of purity of an enzyme or protein after being subjected to a series of purification processes

$$\text{Purification fold} = \frac{\textit{Specific enzyme activity after purification}}{\textit{Specific enzyme activity of the crude homogenate}}$$

Yield or Recovery: This parameter indicated the percentage of the total enzyme obtained after a purification step. It is calculated by taking the ratio (expresses as percent) of total activity of enzyme in purified fraction to the total activity of enzyme in the crude or un purified homogenate

$$\%\text{Yield or Recovery} = \frac{\textit{Total enzyme activity in purified fraction}}{\textit{Total enzyme activity in crude homogenate}} \; x\; 100\%$$

Amount of protein (mg)= Protein concentration (mg/ml) x volume of sample (ml)

2.6.5. Other related terminology

The **rate of a reaction** is the concentration of substrate disappearing (or product produced) per unit time (mol L^{-1} s^{-1}).

The **% purity** is 100% × (specific activity of enzyme sample / specific activity of pure enzyme). The impure sample has lower specific activity because some of the mass is not actually enzyme. If the specific activity of 100% pure enzyme is known, then an impure sample will have a lower specific activity, allowing purity to be calculated and then getting a clear result.

Step	Vol (ml)	Total protein (mg)	Total activity (units)	Specific activity (units/mp)	Purification (fold)	Yield (%)
Homogenate	900	43600	48000	1.1	(1)	(100)
pH 4.2 sínatant	650	4760	28000	5.9	5	58
$(NH_4)_2SO_4$ppt	140	1008	18667	18.5	17	39
S-Sepharose	57	7.1	7410	1044	949	15
SephadexG-75	35	2.45	3266	1333	1211	7

Table : A typical enzyme purification table (Adopted from Isolation of cathepsin by R.N.Pike)

2.7. Types of assay

All enzyme assays measure either the consumption of substrate or production of product over time. A large number of different methods of measuring the concentrations of substrates and products exist and many enzymes can be assayed in several different ways. Biochemists usually study enzyme-catalysed reactions using four types of experiments:

- **Initial rate experiments**. When an enzyme is mixed with a large excess of the substrate, the enzyme-substrate intermediate builds up in a fast initial transient. Then the reaction achieves a steady-state kinetics in which enzyme substrate intermediates remains approximately constant over time and the reaction rate changes relatively slowly. Rates are measured for a short period after the attainment of the quasi-steady state, typically by monitoring the accumulation of product with time. Because the measurements are carried out for a very short period and because of the large excess of substrate, the approximation that the amount of free substrate is approximately equal to the amount of the initial substrate can be made.[5][6] The initial rate experiment is the simplest to perform and analyze, being relatively free from complications such as back-reaction and enzyme degradation. It is therefore by far the most commonly used type of experiment in enzyme kinetics.

- Progress curve experiments. In these experiments, the kinetic parameters are determined from expressions for the species concentrations as a function of time. The concentration of the substrate or product is recorded in time after the initial fast transient and for a sufficiently long period to allow the reaction to approach equilibrium. Progress curve experiments were widely used in the early period of enzyme kinetics, but are less common now.

- Transient kinetics experiments. In these experiments, reaction behaviour is tracked during the initial fast transient as the intermediate reaches the steady-state kinetics period. These experiments are more difficult to perform than either of the above two classes because they require specialist techniques (such as flash photolysis of caged compounds) or rapid mixing (such as stopped-flow, quenched flow or continuous flow).

- Relaxation experiments. In these experiments, an equilibrium mixture of enzyme, substrate and product is perturbed, for instance by a temperature, pressure or pH jump, and the return to equilibrium is monitored. The analysis of these experiments requires consideration of the fully reversible reaction. Moreover, relaxation experiments are relatively insensitive to

mechanistic details and are thus not typically used for mechanism identification, although they can be under appropriate conditions.

Enzyme assays can be split into two groups according to their sampling method: continuous assays, where the assay gives a continuous reading of activity, and discontinuous assays, where samples are taken, the reaction stopped and then the concentration of substrates/products determined.

2.7.1. Continuous assays

Continuous assays are most convenient, with one assay giving the rate of reaction with no further work necessary. There are many different types of continuous assays.

Spectrophotometric

In spectrophotometric assays, you follow the course of the reaction by measuring a change in how much light the assay solution absorbs. If this light is in the visible region you can actually see a change in the color of the assay, and these are called colorimetric assays. The MTT assay, a redox assay using a tetrazolium dye as substrate is an example of a colorimetric assay.

UV light is often used, since the common coenzymes NADH and NADPH absorb UV light in their reduced forms, but do not in their oxidized forms. An oxidoreductase using NADH as a substrate could therefore be assayed by following the decrease in UV absorbance at a wavelength of 340 nm as it consumes the coenzyme**.**

Direct versus coupled assays

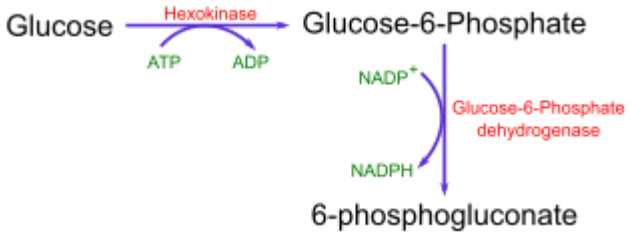


Fig Coupled assay for hexokinase using glucose-6-phosphate dehydrogenase.

Even when the enzyme reaction does not result in a change in the absorbance of light, it can still be possible to use a spectrophotometric assay for the enzyme by using a coupled assay. Here, the product of one reaction is used as the substrate of another, easily detectable reaction. For example, figure 1 shows the coupled assay for the enzyme hexokinase, which can be assayed by coupling its production of glucose-6-phosphate to NADPH production, using glucose-6-phosphate dehydrogenase.

Fluorometric

Fluorescence is when a molecule emits light of one wavelength after absorbing light of a different wavelength. Fluorometric assays use a difference in the fluorescence of substrate from product to measure the enzyme reaction. These assays are in general much more sensitive than spectrophotometric assays, but can suffer from interference caused by impurities and the instability of many fluorescent compounds when exposed to light.

An example of these assays is again the use of the nucleotide coenzymes NADH and NADPH. Here, the reduced forms are fluorescent and the oxidised forms non-fluorescent. Oxidation reactions can therefore be followed by a decrease in fluorescence and reduction reactions by an increase. Synthetic substrates that release a fluorescent dye in an enzyme-catalyzed reaction are also available, such as 4-methylumbelliferyl-β-D-galactoside for assaying β-galactosidase or 4-methylumbelliferyl-butyrate for assaying *Candida rugosa* lipase.

Calorimetric

Chemiluminescence of luminol

Calorimetry is the measurement of the heat released or absorbed by chemical reactions. These assays are very general, since many reactions involve some change in heat and with use of a microcalorimeter, not much enzyme or substrate is required. These assays can be used to measure reactions that are impossible to assay in any other way.

Chemiluminescent

Chemiluminescence is the emission of light by a chemical reaction. Some enzyme reactions produce light and this can be measured to detect product formation. These types of assay can be extremely sensitive, since the light produced can be captured by photographic film over days or weeks, but can be hard to quantify, because not all the light released by a reaction will be detected.

The detection of horseradish peroxidase by enzymatic chemiluminescence (ECL) is a common method of detecting antibodies in western blotting. Another example is the enzyme luciferase, this is found in fireflies and naturally produces light from its substrate luciferin.

Light scattering

Static light scattering measures the product of weight-averaged molar mass and concentration of macromolecules in solution. Given a fixed total concentration of one or more species over the measurement time, the scattering signal is a direct measure of the weight-averaged molar mass of the solution, which will vary as complexes form or dissociate. Hence the measurement quantifies the stoichiometry of the complexes as well as kinetics. Light scattering assays of protein kinetics is a very general technique that does not require an enzyme.

Microscale thermophoresis

Microscale thermophoresis (MST) measures the size, charge and hydration entropy of molecules/substrates at equilibrium. The thermophoretic movement of a fluorescently labeled substrate changes significantly as it is modified by an enzyme. This enzymatic activity can be measured with high time resolution in real time. The material consumption of the all optical MST method is very low, only 5 µl sample volume and 10nM enzyme concentration are needed to measure the enzymatic rate constants for activity and inhibition. MST allows analysts to measure the modification of two different substrates at once (multiplexing) if both substrates are labeled with different fluorophores. Thus substrate competition experiments can be performed.

2.7.2. Discontinuous assays

Discontinuous assays are when samples are taken from an enzyme reaction at intervals and the amount of product production or substrate consumption is measured in these samples**.**

Radiometric

Radiometric assays measure the incorporation of radioactivity into substrates or its release from substrates. The radioactive isotopes most frequently used in these assays are ^{14}C, ^{32}P, ^{35}S and ^{125}I. Since radioactive isotopes can allow the specific labelling of a single atom of a substrate, these assays are both extremely sensitive and specific. They are frequently used in biochemistry and are often the

only way of measuring a specific reaction in crude extracts (the complex mixtures of enzymes produced when you lyse cells).

Radioactivity is usually measured in these procedures using a scintillation counter.

Chromatographic

Chromatographic assays measure product formation by separating the reaction mixture into its components by chromatography. This is usually done by high-performance liquid chromatography (HPLC), but can also use the simpler technique of thin layer chromatography. Although this approach can need a lot of material, its sensitivity can be increased by labelling the substrates/products with a radioactive or fluorescent tag. Assay sensitivity has also been increased by switching protocols to improved chromatographic instruments (e.g. ultra-high pressure liquid chromatography) that operate at pump pressure a few-fold higher than HPLC instruments (see High-performance liquid chromatography#Pump pressure).

2.7.3. Factors to control in assays

Several factors effect the assay outcome and a recent review summarizes the various parameters that needs to be monitored to keep an assay up and running.

- **Salt Concentration:** Most enzymes cannot tolerate extremely high salt concentrations. The ions interfere with the weak ionic bonds of proteins. Typical enzymes are active in salt concentrations of 1-500 mM. As usual there are exceptions such as the halophilic algae and bacteria.
- **Effects of Temperature:** All enzymes work within a range of temperature specific to the organism. Increases in temperature generally lead to increases in reaction rates. There is a limit to the increase because higher temperatures lead to a sharp decrease in reaction rates. This is due to the denaturating (alteration) of protein structure resulting from the breakdown of the weak ionic and hydrogen bonding that stabilize the three-dimensional structure of the enzyme active site.[16] The "optimum" temperature for human enzymes is usually between 35 and 40 °C. The average temperature for humans is 37 °C. Human enzymes start to denature quickly at temperatures above 40 °C. Enzymes from thermophilic archaea found in the hot springs are stable up to 100 °C. However, the idea of an "optimum" rate of an enzyme reaction is misleading, as the rate observed at any temperature is the product of two rates, the reaction rate and the denaturation rate. If you were to use an assay measuring activity for one second, it would give high activity at high temperatures, however if you were to use an assay measuring product formation over an hour, it would give you low activity at these temperatures.
- **Effects of pH:** Most enzymes are sensitive to pH and have specific ranges of activity. All have an optimum pH. The pH can stop enzyme activity by denaturating (altering) the three-dimensional shape of the enzyme by breaking ionic, and hydrogen bonds. Most enzymes function between a pH of 6 and 8; however pepsin in the stomach works best at a pH of 2 and trypsin at a pH of 8.
- **Substrate Saturation:** Increasing the substrate concentration increases the rate of reaction (enzyme activity). However, enzyme saturation limits reaction rates. An enzyme is saturated when the active sites of all the molecules are occupied most of the time. At the saturation point, the reaction will not speed up, no matter how much additional substrate is added. The graph of the reaction rate will plateau.

- Level of crowding, large amounts of macromolecules in a solution will alter the rates and equilibrium constants of enzyme reactions, through an effect called macromolecular crowding.

2.8. TECHNIQUES FOR IMMOBILIZATION OF ENZYMES AND OVERVIEW OF APPLICATIONS OF IMMOBILIZED ENZYME SYSTEM

Enzymes are biocatalyst that carries out all the essential biochemical reactions inside the body of an organism. Their unique feature is that they remain unaltered after the reaction is completed. Therefore, they can be used again and again. But the limitation of soluble enzymes is their isolation from the product and the substrate. Most of the Enzymes in the living organism are attached to the cell membrane or entrapped within the cells. This observation led to the concept that pure isolated enzymes may actually perform better when they are immobilized on a solid support. The term immobilized enzyme is used to denote "enzymes physically confined or localized in a defined region of space with retention of their catalytic activities and which can be used repeatedly and continuously". Immobilization is beneficial because it facilitates work up product isolation. Some of the potential advantages and disadvantages of immobilization are highlighted below.

Soluble Enzyme + Substrate---------- Product (single time usage of enzyme)

Immobilized Enzyme + Substrate---------Product (Repeated usage of enzyme)

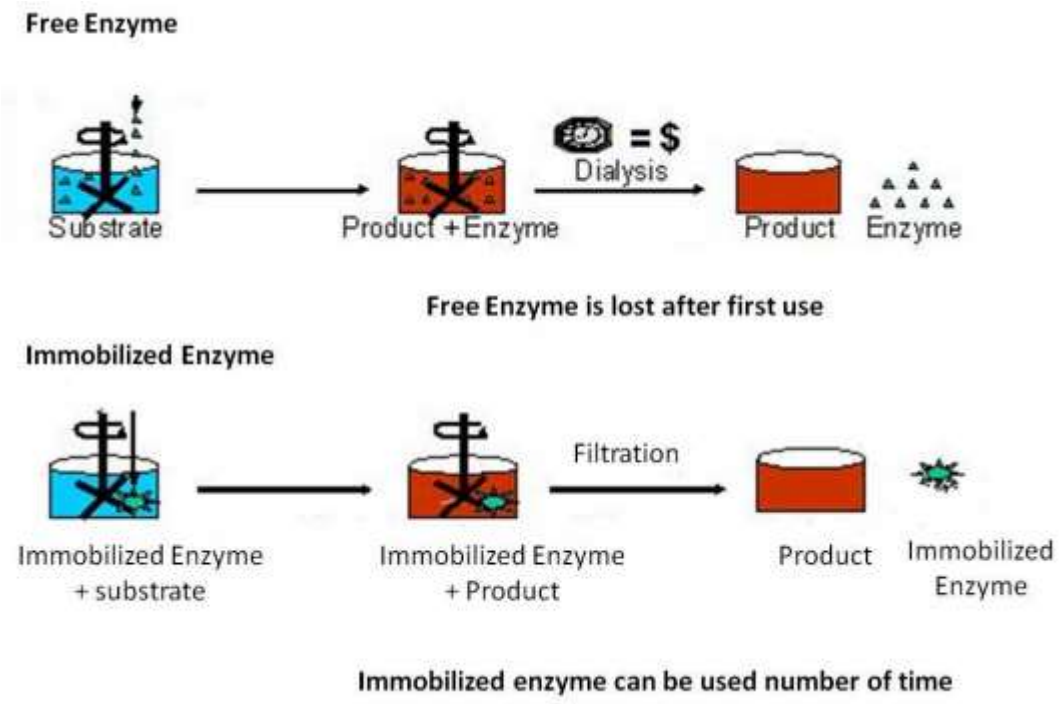


Fig Enzyme immobilisation

Advantages:

1. Easy recovery of the product

2. Product is free from enzyme, so there is no cost of purification of enzyme.

3. Enzyme can be used repeatedly

4. The enzyme generally get stabilized after adsorption

Disadvantages:

1. Loss of catalytic properties for some enzymes

2. Some enzymes become unstable

3. Additional cost of immobilization

4. Differential limitations.

Intensive study in the area of enzyme immobilization started in mid 1950 and has since continued. The first industrial use of immobilized enzymes was reported in 1967 by Chibata and co-workers who immobilized Aspergillus oryzae amino acylase for the resolution of synthetic , DL amino acids into the corresponding optically active enantiomers. At present, the use of immobilized enzyme is well established in various industries.

Number of important points should be kept in mind while immobilizing an enzyme

1. The biological activity of the enzyme should be retained

2. The enzyme should be more stable as compared its soluble counterpart.

3. The cost of immobilization should not be too high

4. It should be used repeatedly.

2.8.1. Enzyme immobilization Techniques

The various methods of enzyme immobilization are broadly classified as

1. Reversible immobilization

a. Adsorption

2. Irreversible immobilization

a. Covalent coupling

b. Entrapment and microencapsulation

c. Crosslinking

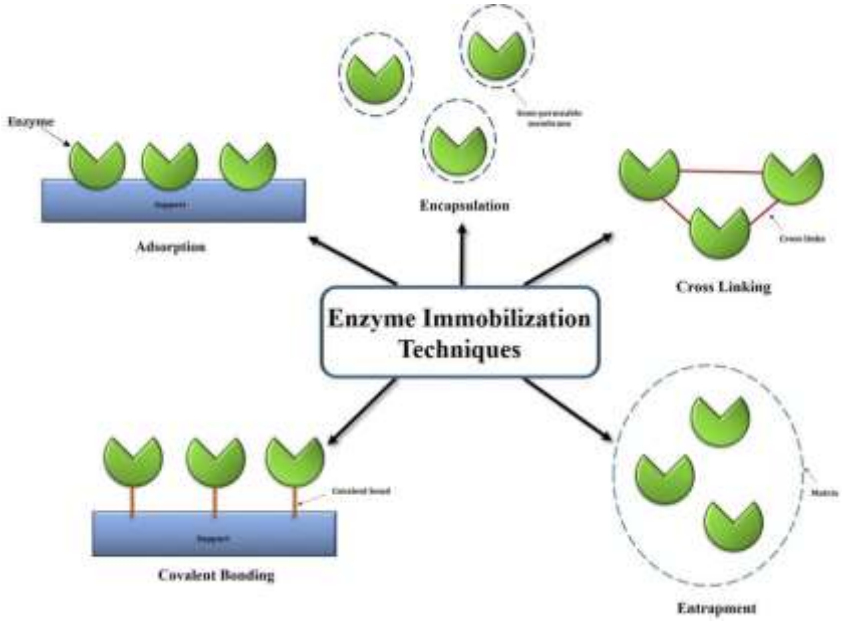


Figure : Diagrammatic representation of the various methods of immobilization

The advantage of reversible immobilization over irreversible systems can be summarized as follows:

1. No chemical modification of the enzyme is required

2. If the enzyme gets activated during use, it can be replaced in reversible immobilization.

3. The immobilization of enzyme by adsorption of bioaffinity can be accomplished rapidly

2.8.1.1. Reversible immobilization

Adsorption

Immobilization by adsorption is the easiest and fastest method. The adsorption is dependent on the experimental variables such as pH, nature of solvent, ionic strength, quantity of enzyme and adsorbent, the time and temperature. A close control of these variables is required owing to the relatively weak binding forces between protein and adsorbent (hydrogen bonds, van der Waals forces, hydrophobic interactions, etc.). Enzymes can be immobilized by simply mixing the enzymes with the matrix, under appropriate conditions of pH and ionic strength. Adsorption process is based on vander Waal forces, ionic and hydrogen bonding as well as hydrophobic interactions, which are very weak forces, but in large number, impart sufficient binding strength. Adsorbed enzymes can be protected from agglomeration, proteolysis and interaction with hydrophobic interfaces. In order to prevent chemical modification and damage to enzyme, the existing surface properties of enzymes and support are need to be considered. The adsorption through physical method generally involves multipoint protein adsorption between a single protein molecule and a number of binding sites on the immobilization surface.

The main disadvantage of this method is that the enzyme is easily desorbed by factors like pH, temperature fluctuations, changes in substrate and ionic concentrations. Few advantages of dsorption methods are

- Easy to carry out
- No reagent are required
- Minimum activation step involved
- Comparatively cheap method
- Less disruptive to protein than chemical methods

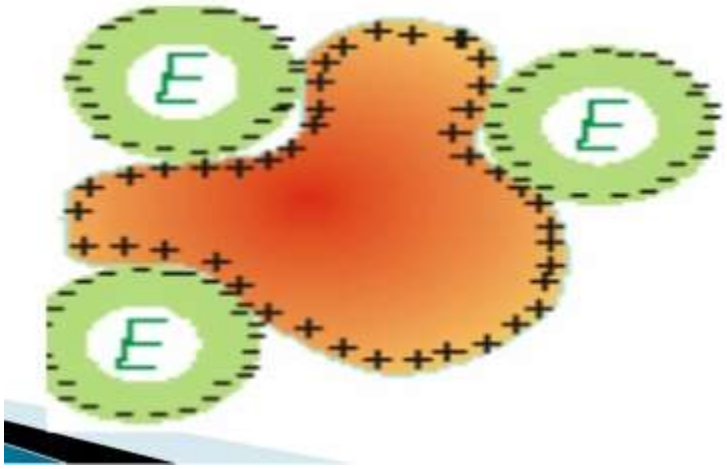


Fig Adsorption

2.8.2. Irreversible immobilization

Covalent coupling

Covalent coupling is the most frequently used approach of immobilization in which covalent bonds are formed between surface amino acids of the enzyme and the matrix. Hydrophilic amino acids which are likely to be present on the protein surface, can be exploited for this purpose. ε-amino group of lysine residue, cysteine (via SH), tyrosine, histidine, aspartic and glutamic acids, tryptophan and arginine mostly takes part in bond formation. A number of chemical reagents and protocols are available for covalently linking an enzyme to the matrix.

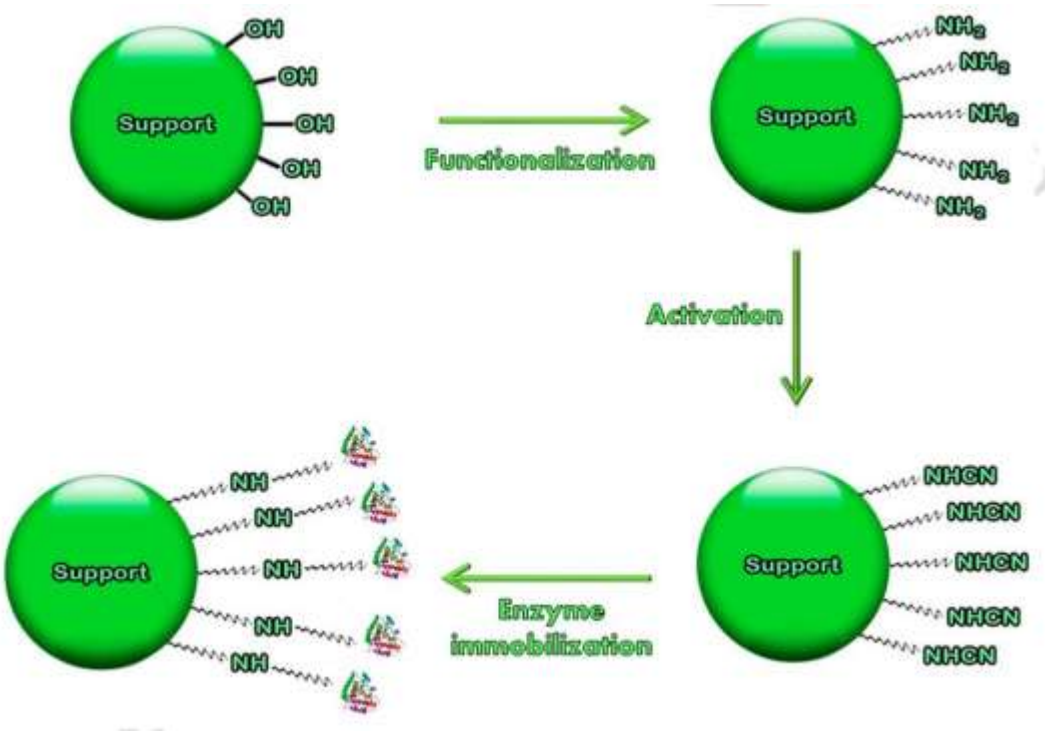


Figure : Covalent binding

Most commonly used methods of covalent bonding are:

1. Diazoation: It is based on the diazo linkage between protein and aryldiazonium electrophilic groups of the matrix.

2. Formation of Peptide bond: bond formation between amino/carboxyl groups of support and amino or carboxyl group of the enzyme

3. Poly functional reagents: use of bi-functional or multifunctional reagent (glutaraldehyde) which forms bonding between the amino group of the support and amino group of the enzyme

4. Amidination reaction: The matrix containing amido ester functional groups can be used for immobilization of protein.

5. Thiol–disulphide interchanged reaction: This methods is used for protein bonding via thiol groups of both carrier and protein.

6. Akylation and arylation: This methods is based on alkylation of amino, phenolic and thiol groups of protein with reactive matrix containing halides, vinyle, sulphonile etc.

Advantage of Covalent bonding method:

- Strong linkage of enzyme to the support
- No leakage or desorption problem
- Comparatively simple method

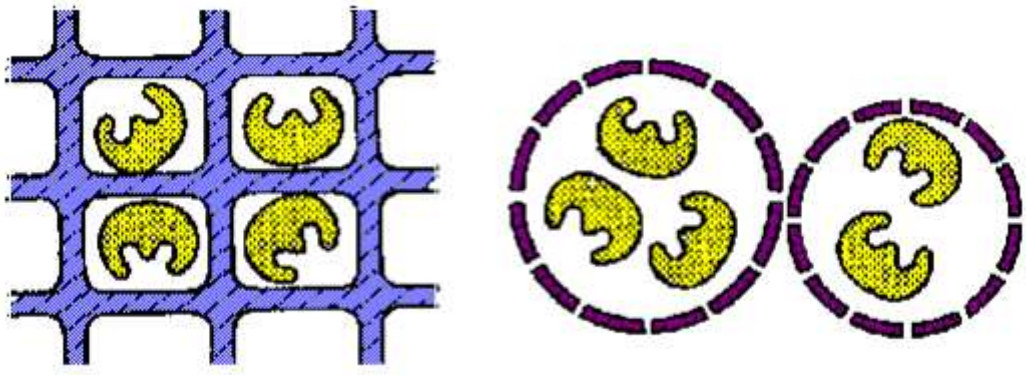
- A variety of support with different functional group available
- Wide applicability

Disadvantage of Covalent bonding method:

- Chemical modification of enzyme leading to functional conformational loss
- Enzyme inactivation by change in the conformational when undergoes reactions at active site
- This can be overcome through immobilization in the presence of enzyme substrate or a competitive inhibitor

Entrapment

Entrapment and encapsulation is based on the occlusion of an enzyme within a constraining structure, but tight enough to release an enzyme while allowing penetration of a substrate. However, due to diffusion limitations, such methods are often unsuitable for the immobilization of enzymes hydrolyzing macromolecular substrates.



entrapped in a matrix

entrapped in droplets

Methods of Entrapment

- Inclusion in the gels: enzymes trapped in gels
- Inclusion in fibers: enzymes supported on fiber formate
- Inclusion in microcapsules: enzymes entrapped in microcapsules formed by monomer mixtures such as polyamine, calcium alginate

Advantage of Entrapment method:

- Fast
- Cheap (low cost matrix available)
- Mild conditions are required
- Less chance of conformational change in the enzyme

Disadvantage of Entrapment method:

- Leakage of enzyme

- Pore diffusion limitation
- Chance of microbial contamination

Cross linking

This method involves attachment of biocatalysts to each other by bi- or multifunctional reagents or ligands. In this way, very high molecular weight typically insoluble aggregates are formed. Cross-linking is a relatively simple process. It is not a preferred method of immobilization as it does not use any support matrix. So they are usually gelatinous and not particularly firm. Since it involves a bond of the covalent kind, biocatalyst immobilized in this way frequently undergoes changes in conformation with a resultant loss of activity. Still it finds good use in combination with other support dependent immobilization technologies, namely to minimize leakage of enzymes already immobilized by adsorption.

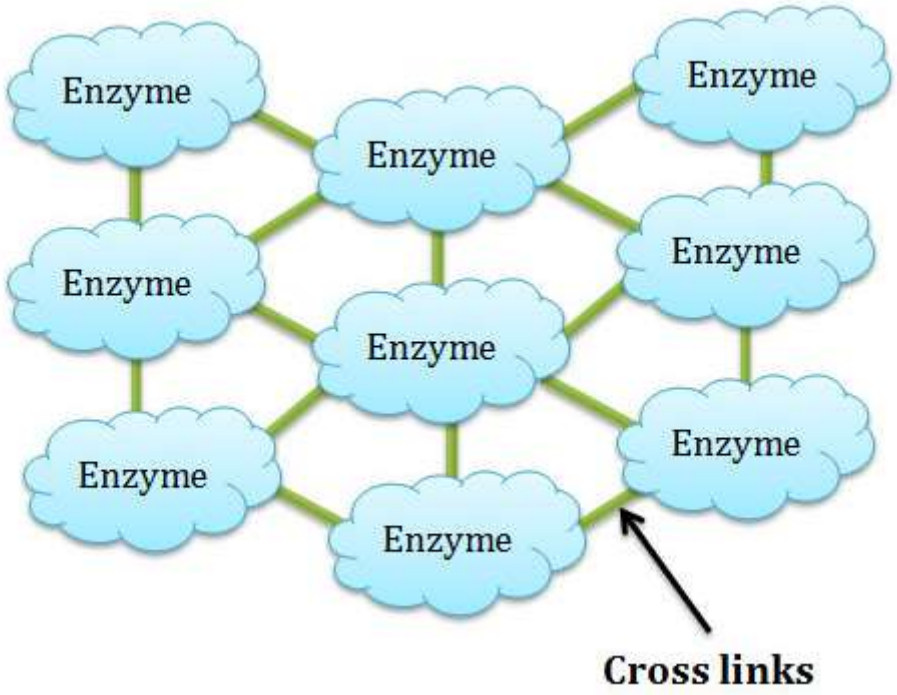


Figure : Cross-linking

Advantage of Cross linking method:

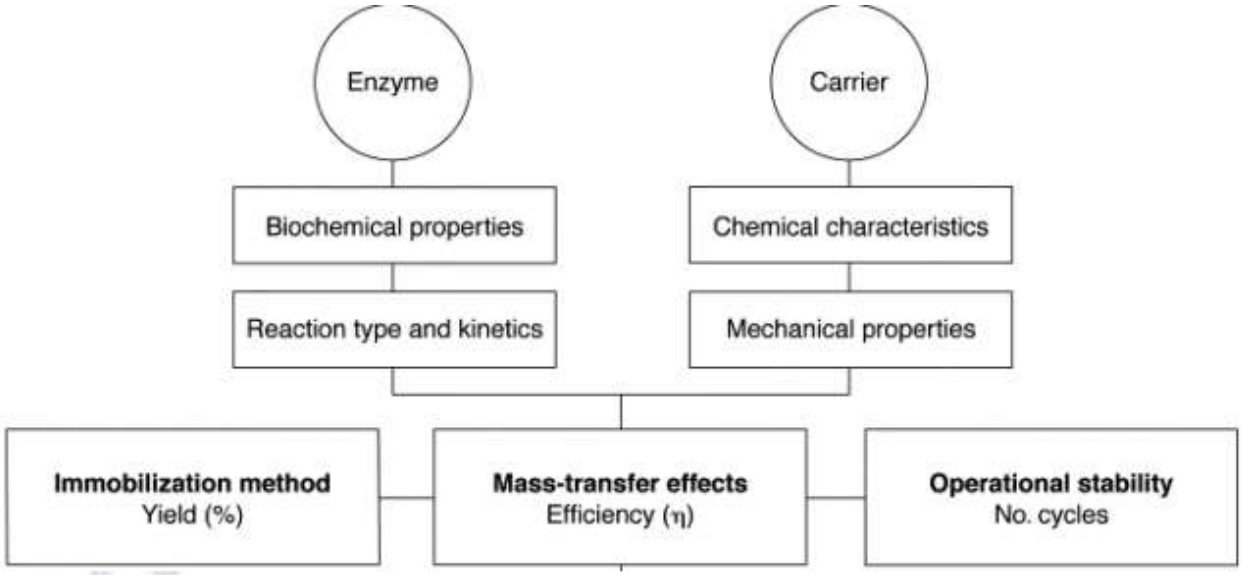
- It is used mostly as a means of stabalizing adsorbed enzyme and also for preventing leakage

Disadvantage of Cross linking method:

- Cross linking may cause significant change in the active site of enzyme which may lead to loss of activity

2.8.2. Biochemical Properties of the Immobilized Enzymes

The biochemical properties of the immobilized enzyme are different when compared to the free enzyme, this is due to the change in environment of the immobilized enzyme



- the physical and chemical properties of the support matrix and interactions of the matrix with substrates or products changes the kinetics of the immobilized enzyme.
- Generally there is a decrease in the rate of enzyme catalyzed reaction because the matrix restric the diffusion of the substrate towards the enzyme.
- The Km of the enzyme also changes after immobilization because of diffusion limitations. If the matrix is positively charged and substrate is also positively charges due to electrostatic repulsion the substrate will not come in the viccinity of the enzyme hence Km is altered.
- Sometimes the 3D structure of the enzyme is also changed which also result in altering the kinetic properties of enzyme.
- The performance of the immobilized enzyme can be improved further by studying the structural changes of the immobilized enzyme

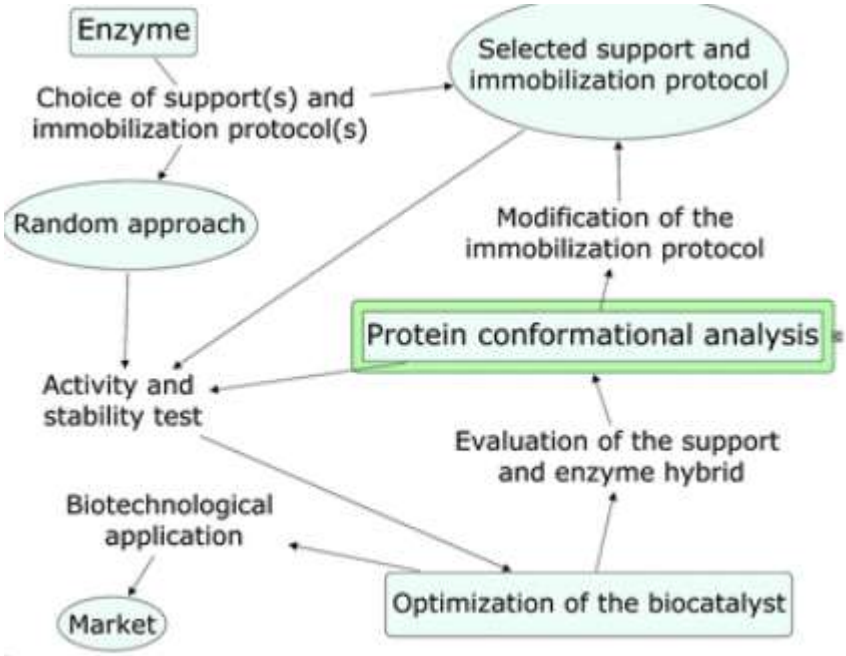


Figure Structural studies improves the performance of the immobilized enzyme

2.8.3. Applications of enzyme immobilization

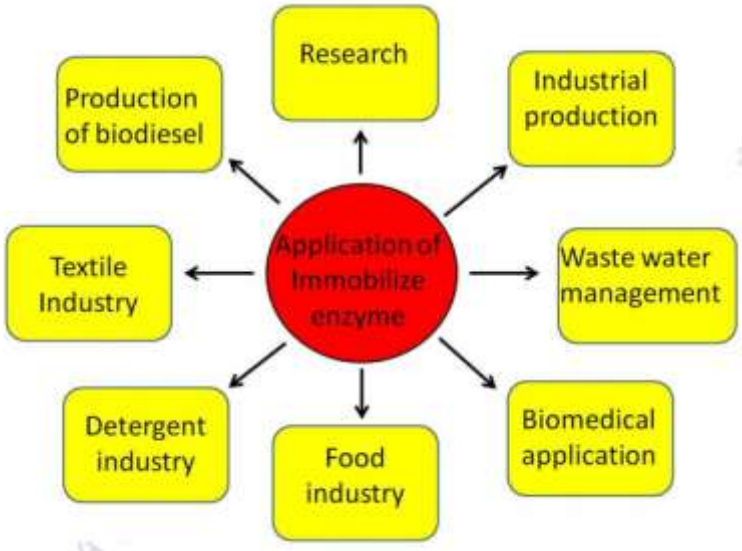


Fig **Applications of enzyme immobilization**

2.8.3.1. Biomedical Application

Biosensor: Biosensor are electronic monitoring devices that make use of an enzyme's specificity and the technique of enzyme immobilization

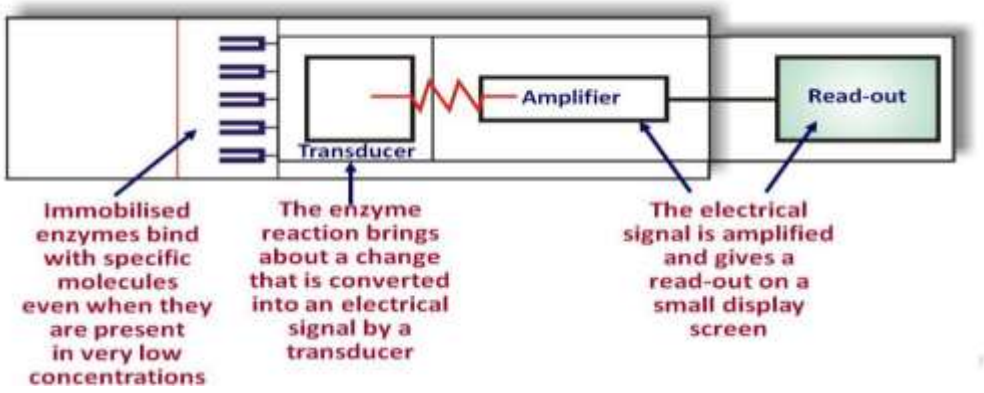


Fig Biosensor

A biosensor has been developed for detecting glucose in the blood of diabetics

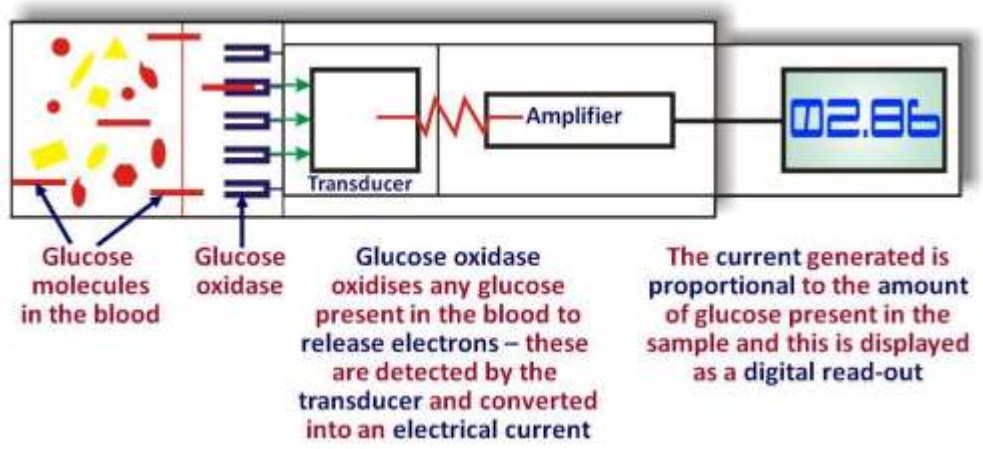


Figure 17: Detecting glucose in blood

Table 1: Immobilized enzymes used as biosensors

Enzymes	Inhibitor	Immobilization matrix	Samples
Biosensors for pesticides determination			
Acetylcholinesterase	Paraoxon	Multiwalled carbon nanotubes	Real Water sample
Acetylcholinesterase	Paraoxon	Entrapment in PVA-SbQ	Spiked
Variants	Carbofuran	Polymer	River water sample
Catalase	Azide	Gelatine with GA	Fruit Juice
Biosensors for the heavy metals determination			

Urease	Hg2+, Cu Cd	Entrapment in sol-gel matrix	Tap and River water
Glucose oxidase	Hg2+	Cross-linking with GA and BSA	Spiked water
Biosensors for the determination of other chemical components			
Butyrylcholinesterase	A-chaconine, α-solanine	Cross-linking with BSA	GA Agriculture
Acetylcholinesterase	Anatoxine-a	Entrapment in PVA-SbQ	Fresh water

2.8.3.2. Industrial applications of immobilized enzymes

2.8.3.2.1.

Table 2 shows some of the immobilized enzymes used for the synthesis of various antibiotics.

Enzyme	Immobilization support	Antibiotic produced
Penicillin acylase from E. coli	Polyacrylamide gel	Cephalexin
Penicillin G acylase	Nylon hydrolon membrane	Cephalexin
Penicillin G acylase from E. coli	Silica gel	6-APA
Penicillin G acylase	Eupergit	6-APA
Acetyl xylan estrase	CKEAs using glutaraldehyde	Desacetyl β-lactam
Penicillin acylase	Poly-N-isopropylacrylamide	Cephalexin

2.8.3.2.2. Immobilized enzymes in food industry

Immobilized enzymes are used in the processing of food samples and its analysis.

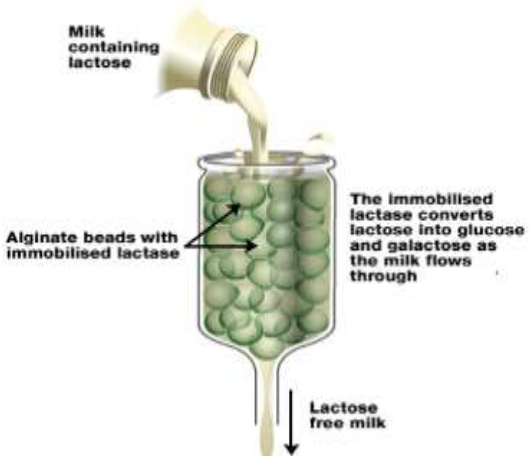


Figure 18: Removal of lactose from milk by Immobilized enzyme

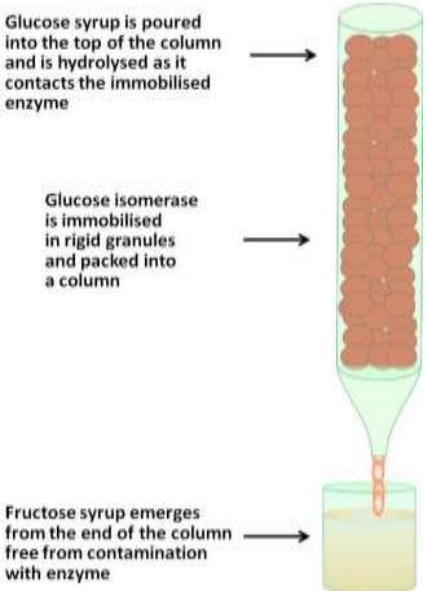


Figure 19: Conversion of glucose syrup into fructose corn syrups

Table 3 shows the processing of various food substrates using respective immobilized enzymes.

Table 3: Immobilized enzymes used in food industry

Enzyme	Immobilization support	Food substrate
β-galactosidase and amyloglucosidas	Bone powder	Lactose, Whey, Whey permeates, skimmed milk
Pectinase	Amino exchange resin	Pectin
Laccase	Silica gel	Ine, fruit juice and beer processing
Trypsin	cellulose	B-lactoglobulin
Tyrosinase	Polyacrylic acid carbon nanotubes	Phenolic in red wine
Pectinase	Amino exchange resin	Pectin solution

2.8.3.2.3. Biodiesel production Biodiesel has gained importance for its ability to replace fossil fuels which are likely to run out within a century.

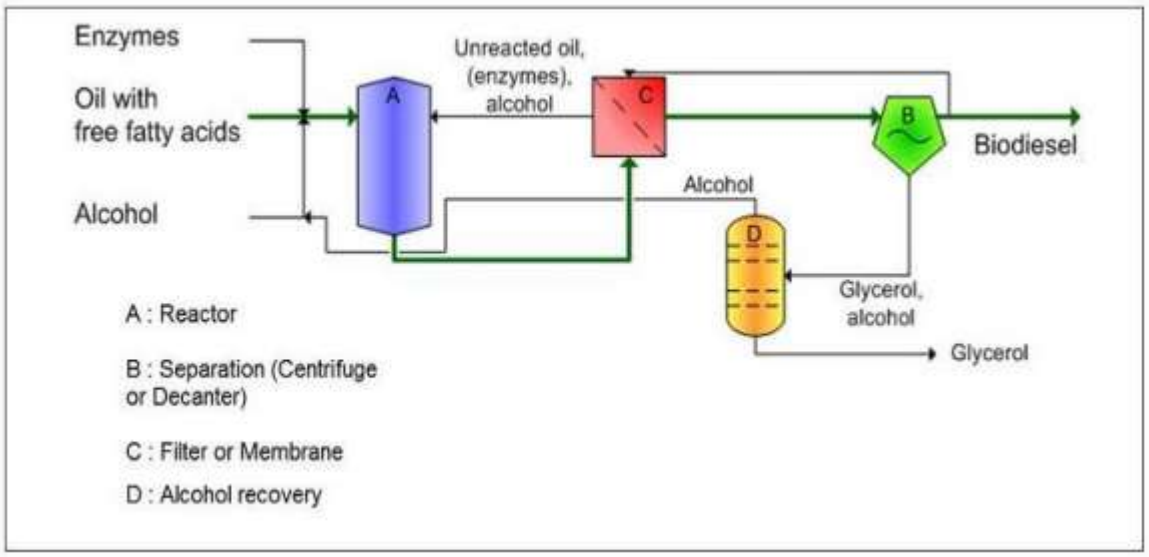


Figure 20: Enzymatic Biodiesel Production

Table 4: Immobilized lipases used biodiesel production

Source of lipase	Immobilization support	substrate
T. lanuginous	Polyurethane foam	Canola oil and methanol
C. antarcrica	Ceramic beads	Waste cooking oil
P. fluoescens	Porous kaolonite	Safflower oil
P. expansum	Silica gel (resin D4020)	Waste oil
T. lanuginous	Microporous polymeric matrix	Sunflower, soyabean
C. nigosa	chitosan	Rapeseed oil
Rhozopus oryzae	Biomass support particles	Jatropha oil

2.8.3.2.4. Immobilized enzymes for bioremediation

Bioremediation is a technique that involves the use of enzyme and biological organism to remove pollutants from a contaminated site.

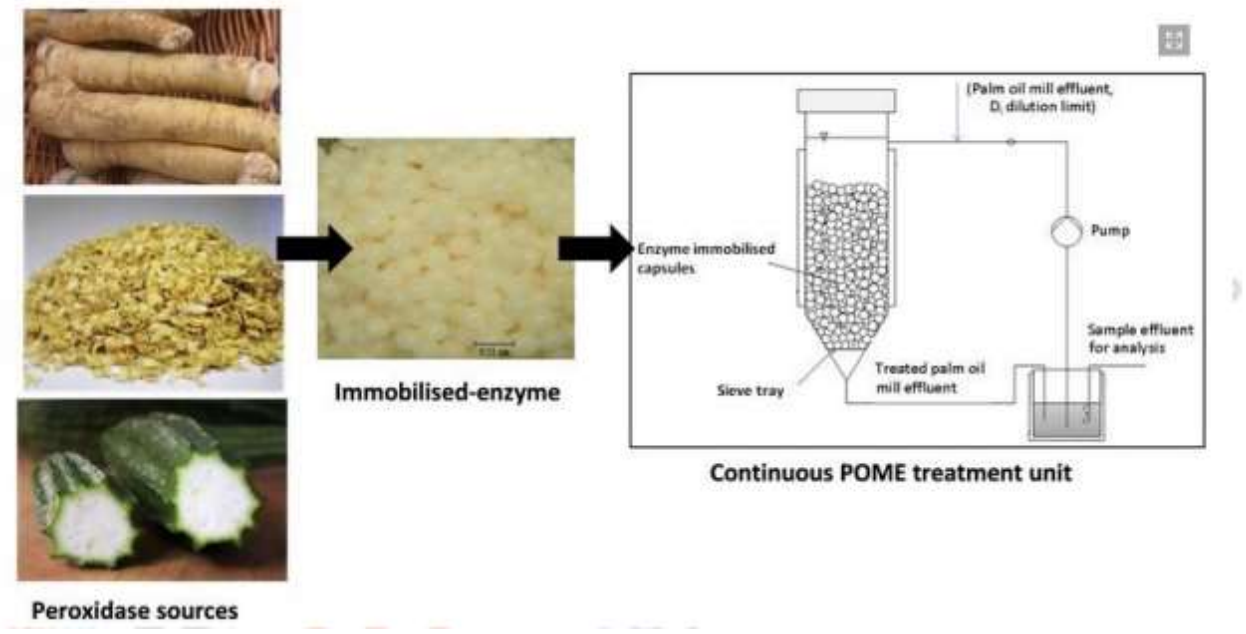


Figure 21: Peroxidase immobilized on support and used for continuous palm oil mill effluent (POME) treatment

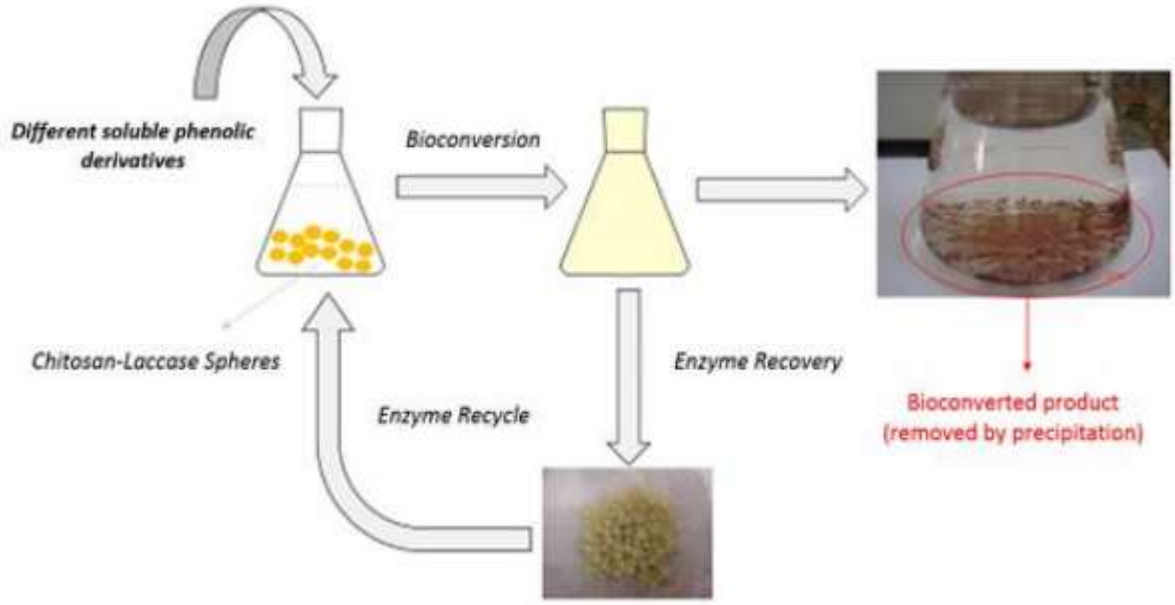


Figure 22: Removal of phenolic derivative by immobilized enzyme.

Table 5 lists some recent research about the use of enzymes in different immobilized forms for dye and phenolic compounds removal.

Table 5: Immobilized Enzymes in bioremediation

Enzyme	Immobilization support	substrate
Lipase	Polypropylene membrane	Dimethylphthalate
Laccase	Silica	Reactive dye
Polyphenol oxidase	Chitosen coated polysulphone membrane	Industrial phenolic effluent
Polyphenol oxidase	Celite-545	Textile and non textile dyes
peroxidase	Con A-sephadex	Textile dye
laccase Epoxy	activated carriers	Synthethic reactive dye
Fungal laccase	Porous glass beads	Anthraquinone and indigoid dyes

2.9. ABZYMES AND THEIR APPLICATIONS

An abzyme (from antibody and enzyme), also called *catmab* (from *catalytic monoclonal antibody*), is a monoclonal antibody with catalytic activity. Molecules which are modified to gain new catalytic activity are called synzymes. Abzymes are usually artificial constructs, but are also found in normal humans (anti-vasoactive intestinal peptide autoantibodies) and in patients with autoimmune diseases such as systemic lupus erythematosus, where they can bind to and hydrolyze DNA. Abzymes are potential tools in biotechnology, e.g., to perform specific actions on DNA.

Enzymes function by lowering the activation energy of the transition state, thereby catalyzing the formation of an otherwise less-favorable molecular intermediate between reactants and products. If an antibody is developed to a stable molecule that's similar to an unstable intermediate of another (potentially unrelated) reaction, the developed antibody will enzymatically bind to and stabilize the intermediate state, thus catalyzing the reaction. A new and unique type of enzyme is produced.

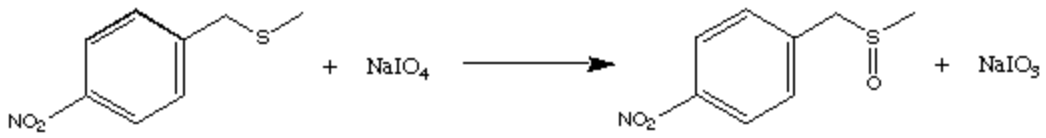
HIV treatment

In a June 2008 issue of the journal Autoimmunity Reviews, researchers S Planque, Sudhir Paul, Ph.D, and Yasuhiro Nishiyama, Ph.D of the University Of Texas Medical School at Houston announced that they have engineered an abzyme that degrades the superantigenic region of the gp120 CD4 binding site. This is the one part of the HIV virus outer coating that does not change, because it is the attachment point to T lymphocytes, the key cell in cell-mediated immunity. Once infected by HIV, patients produce antibodies to the more changeable parts of the viral coat. The antibodies are ineffective because of the virus' ability to change their coats rapidly. Because this protein gp120 is necessary for the HIV virus to attach, it does not change across different strains and is a point of vulnerability across the entire range of the HIV variant population.

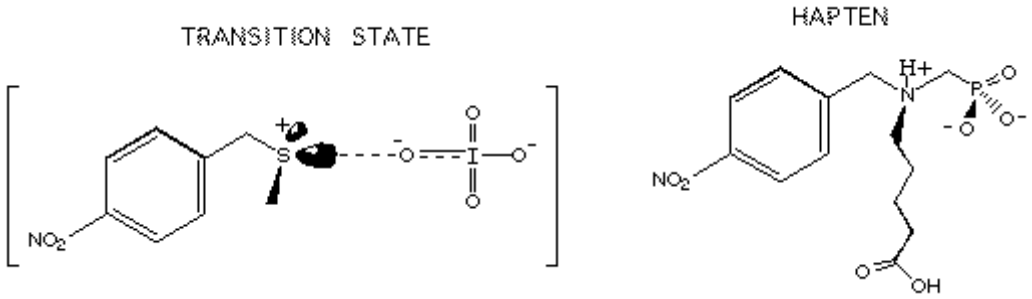
The abzyme does more than bind to the site, it actually destroys the site, rendering the HIV virus inert, and then can attach to other viruses. A single abzyme can destroy thousands of HIV viruses. Human clinical trials will be the next step in producing treatment and perhaps even preventative vaccines and microbicide

Abzymes: Catalytic Antibodies

By exploiting the highly specific antigen binding properties of antibodies, experimental strategies have been devised to produce antibodies that catalyze chemical reactions. These catalytic antibodies, or *abzymes*, are selected from monoclonal antibodies generated by immunizing mice with haptens that mimic the transition states of enzyme-catalyzed reactions. For example, the 28B4 abzyme catalyzes periodate oxidation of *p*-nitrotoluene-methyl sulfide to sulfoxide, as shown below, where electrons from the sulfur atom are transferred to the more electronegative oxygen atom.



The rate of this reaction is promoted by enzyme catalysts that stabilize the transition state of this reaction, thereby decreasing the activation energy and allowing for more rapid conversion of substrate to product. In this case, the transition state is thought to involve a transient positive charge on the sulfur atom and a double-negative charge on the periodate ion as shown below on the left.



In order to generate abzymes complementary in structure to this transition state, mice were immunized with an aminophosphonic acid hapten, as shown above at the right. Obviously, its structure mirrors the structure and electrostatic properties of the sulfoxide transition state. Of the hapten-binding monoclonal antibodies produced with this hapten, many were found to catalyze sulfide oxidation but with a wide range of binding affinities and catalytic efficiencies. In particular, abzyme 28B4 binds hapten with high affinity (K_d = 52 nM) and exhibits a correspondingly high degree of catalytic efficiency (k_3/K_M = 190,000 $M^{-1}s^{-1}$).

Elucidation of the molecular structure of abzyme 28B4 bound to the hapten reveals much about the nature of its catalytic action. Highly specific structural and electrostatic interactions create a remarkable degree of structural complementarity between the antigen-binding site and the sulfoxide

transition state analog as illustrated in the following series of three-dimensional views of the antibody-hapten complex.

2.10 ENZYME ELECTRODES

The enzyme electrode is a miniature chemical transducer which functions by combining an electrochemical procedure with immobilized enzyme activity. This particular model uses glucose oxidase immobilized on a gel to measure the concentration of glucose in biological solutions and in the tissues *in vitro*.

Applications of "Wired"Enzyme Electrodes

The electrochemical detection of physiologically important molecules using enzyme-based biosensors has been an area of intense activity for a number of years, the most successful application being determination of glucose. One successful approach has involved in ding redox-active centers (mediators) and enzymes in a polymeric matrix immobilized on an electrode surface. A series of such enzyme-based systems has been developed by Heller, and are generally referred to as "wired"enzyme electrodes

"Wired" Enzyme Electrodes for Determination of Glucose

The concept involved in amperometric enzyme biosensors is conversion of a chemical signal (in this case, the enzyme reaction) to an analytical signal(a current) using the working electrode as the transducer. A schematic diagram for such a sensor is shown in . The enzyme is immobilized on the surface of an electrode, and this immobilized layer is covered by a membrane. The function of the membrane is to provide stability, and it can also be used to prevent potential interferants from reacting with the enzyme. The electrode assembly is placed in the solution containing the analyte, which can readily diffuse through the membrane, and into the immobilized enzyme layer.

"Wired" Enzyme Electrodes for Determination of Hydrogen Peroxide

Biofuel Cells Based on "Wired" Enzyme Electrodes

A more recent application of "wired" enzyme technology is its use for the cathode of a biofuel cell (15). In this example, the cathode reaction was the four-electron reduction of oxygen to water, and the anode was the oxidation of glucose, again using a "wired" glucose oxidase electrode. In previous fuel cells, reduction of water has been achieved using either noble metal cathodes at pH 0 or

activated carbon cathodes at pH 14; high temperature was required in both cases. In contrast, in this example, reduction of oxygen to water at a current density of 5 mA cm was achieved at 37.5°C in pH 5 citrate buffer using an enzyme electrode based on laccase.

The scheme for mediated laccase reduction of water is shown in . The electrode material was carbon cloth (i.e. a large surface area electrode), to which the osmium-containing redox polymer was covalently attached. The laccase enzyme was electrostatically bound to the osmium centers. One major advantage of this approach is that the redox potential of the covalently-bound osmium complex can be altered by varying the ligands. In this case, bidentate dimethyl-bipyridine and tridentate terpyridine were used, giving the osmium complex a redox potential of +0.78 V (vs. NHE). This value is close to the redox potential of laccase under these conditions (+0.82 V); that is, the redox potential of the osmium complex is adjusted to minimize the overpotential required for laccase reduction. Another advantage of this approach is that the electrode reactions are so selective that the reactions of glucose at the cathode and oxygen at the anode are insignificant, which eliminates the need for a membrane to separate the two electrodes into two compartments

2.11. Enzyme multiplied immunoassay technique

Enzyme multiplied immunoassay technique, or EMIT, is a common method for screening urine and blood for drugs, both legal or illicit. First introduced by Syva Company in 1973, it is the first homogeneous immunoassay to be widely used commercially.

A mix and read protocol has been developed that is exceptionally simple and rapid. The most widely used applications for EMIT are for therapeutic drug monitoring (serum) and as a primary screen for abused drugs and their metabolites (urine). The US patents covering the major aspects of the method, 3,817,837 and 3,875,011, have expired. While still sold by Siemens Healthcare under its original trade name, EMIT, assay kits with different names that employ the same technology are supplied by other companies. The test is not particularly accurate, especially with regard to test results for cannabis. When the Food and Drug Administration approved EMIT, it did so with the strict provision that positive test results should be confirmed by an alternative testing method

UNIT – III - ENZYME AND PROTEIN ENGINEERING – SBTA5202

3. PROTEIN ENGINEERING

Protein engineering is the process of developing useful or valuable proteins. It is a young discipline, with much research taking place into the understanding of protein oldingand recognition for protein design principles. There are two general strategies for protein engineering, 'rational' protein design and directed evolution. These techniques are not mutually exclusive; researchers will often apply both. In the future, more detailed knowledge of protein structure and function, as well as advancements in high-throughput technology, may greatly expand the capabilities of protein engineering. Eventually, even unnatural amino acids may be incorporated, thanks to a new method that allows the inclusion of novel amino acids in the genetic code

3.1. Rational design

In rational protein design, the scientist uses detailed knowledge of the structure and function of the protein to make desired changes. In general, this has the advantage of being inexpensive and technically easy, since site-directed mutagenesis techniques are well-developed. However, its major drawback is that detailed structural knowledge of a protein is often unavailable, and, even when it is available, it can be extremely difficult to predict the effects of various mutations. Computational protein design algorithms seek to identify novel amino acid sequences that are low in energy when folded to the pre-specified target structure. While the sequence-conformation space that needs to be searched is large, the most challenging requirement for computational protein design is a fast, yet accurate, energy function that can distinguish optimal sequences from similar suboptimal ones.

3.2. Directed evolution

In directed evolution, random mutagenesis is applied to a protein, and a selection regime is used to pick out variants that have the desired qualities. Further rounds of mutation and selection are then applied. This method mimics natural evolution and, in general, produces superior results to rational design. An additional technique known as DNA shuffling mixes and matches pieces of successful variants in order to produce better results. This process mimics the recombination that occurs naturally during sexual reproduction. The advantage of directed evolution is that it requires no prior structural knowledge of a protein, nor is it necessary to be able to predict what effect a given mutation will have. Indeed, the results of directed evolution experiments are often surprising in that desired changes are often caused by mutations that were not expected to have that effect. The drawback is that they require high-throughput, which is not feasible for all proteins. Large amounts of recombinant DNA must be mutated and the products screened for desired qualities. The sheer number of variants often requires expensive robotic equipment to automate the process. Furthermore, not all desired activities can be easily screened for.

Examples of engineered proteins

Using computational methods, a protein with a novel fold has been designed, known as Top7,as well as sensors for unnatural molecules. The engineering of fusion proteinshas

yielded rilonacept, a pharmaceutical that has secured FDA approval for the treatment of cryopyrin-associated periodic syndrome.

Another computational method, IPRO, successfully engineered the switching of cofactor specificity of Candida boidinii xylose reductase.[3] Iterative Protein Redesign and Optimization (IPRO) redesigns proteins to increase or give specificity to native or novel substrates and cofactors. This is done by repeatedly randomly perturbing the structure of the proteins around specified design positions, identifying the lowest energy combination of rotamers, and determining whether the new design has a lower binding energy than previous ones.

Computation-aided design has also been used to engineer complex properties of a highly ordered nano-protein assembly. A protein cage, E. coli bacterioferritin (EcBfr), which naturally shows structural instability and an incomplete self-assembly behavior by populating two oligomerization states, is the model protein in this study. Through computational analysis and comparison to its homologs, it has been found that this protein has a smaller-than-average dimeric interface on its two-fold symmetry axis due mainly to the existence of an interfacial water pocket centered around two water-bridged asparagine residues. To investigate the possibility of engineering EcBfr for modified structural stability, a semi-empirical computational method is used to virtually explore the energy differences of the 480 possible mutants at the dimeric interface relative to the wild type EcBfr. This computational study also converges on the water-bridged asparagines. Replacing these two asparagines with hydrophobic amino acids results in proteins that fold into alpha-helical monomers and assemble into cages as evidenced by circular dichroism and transmission electron microscopy. Both thermal and chemical denaturation confirm that, all redesigned proteins, in agreement with the calculations, possess increased stability. One of the three mutations shifts the population in favor of the higher order oligo merization state in solution as shown by both size exclusion chromatography and native gel electrophoresis.

Enzyme engineering

Enzyme engineering is the application of modifying an enzyme's structure (and, thus, its function) or modifying the catalytic activity of isolated enzymes to produce new metabolites, to allow new (catalyzed) pathways for reactions to occur, or to convert from some certain compounds into others (biotransformation). These products will be useful as chemicals, pharmaceuticals, fuel, food, or agricultural additives. An *enzyme reactor* consists of a vessel containing a reactional medium that is used to perform a desired conversion by enzymatic means. Enzymes used in this process are free in the solution.

3.3. PROTEIN SPLICING

Protein splicing is an intramolecular reaction of a particular protein in which an internal protein segment (called an intein) is removed from a precursor protein with a ligation of C-terminal and N-terminal external proteins (called exteins) on both sides. The splicing junction of the precursor protein is mainly a cysteine or a serine, which are amino acids containing a nucleophilic side chain. The protein splicing reactions which are known now do not require

exogenous cofactors or energy sources such as adenosine triphosphate (ATP) or guanosine triphosphate (GTP). Normally, **splicing** is associated only with pre-mRNA splicing.

Protein splicing was unanticipated and discovered by two groups (Anraku and Stevens) in 1990. They both discovered a *Saccharomyces cerevisiae* VMA1 in a precursor of a vacuolar H^+-ATPase enzyme. The amino acid sequence of the N- and C-termini corresponded to 70% DNA sequence of that of a vacuolar H^+-ATPase from other organisms, while the amino acid sequence of the central position corresponded to 30% of the total DNA sequence of the yeast HO nuclease.

GENERAL ORGANIZATION OF AN INTEIN

Inteins usually vary in size from 134 to 650 amino acid residues, although inteins of 1308 and 1650 residues are also know. Inteins are conventionally divided into two large groups, classical inteins and mini-inteins (Fig. 1). A classical intein consists of two domains, Hint, which catalyses protein splicing, and a central endonuclease domain. In mini-inteins, the central endonuclease domain is replaced by a linker sequence, which lacks catalytic activity. Analysis of most inteins showed that an average intein harbors ten conserved amino acid sequence motifs: A, N2, B, N4, C, D, E, H, F, and G (Fig. 1a). Mini-inteins lack central motifs C, D, E, and H

Motif A is a short N-terminal sequence of 13 residues, of which two (the first and the last ones) are highly conserved, suggesting their immense importance for splicing initiation and completion. Position 1 at the N end of an intein is almost always occupied by Cys and, in extremely rare cases, by Ala, Gln, or Ser. Position 13 is occupied by Gly or, in rare cases, Ala, Lys, Thr, Arg, Tyr, or Asn.

Motif N2 consists of 7 residues, of which Asp5 or Glu5 is highly conserved and is most often preceded by Gly.

Motif B consists of 14 residues. Position 10 is occupied by His in all known inteins. Position 7 is most often occupied by Thr [9]. These two conserved amino acid residues are involved in splicing initiation.

Motif N4 consists of 16 residues, including highly conserved Asp or Glu in position 11. As in motif N2, this residue is usually preceded by Gly10. However, motif N4 is lacking in some inteins (SceVMA, CtrVMA, CeuClpP, and some others) [11]. Motifs A, N2, B, and N4 form the N-terminal splicing domain, whose function is to facilitate disruption of the peptide bond at the N end of the intein [11]. On average, the N-terminal domain is 150–200 residues in size [11]. Most of the above conserved amino acid residues are absolutely essential for N-terminal cleavage. Amino acid substitutions in the corresponding positions often fully abolish the initiation of cleavage and splicing of the precursor protein. Motifs C and E form a basis of the DOD endonuclease domain [13]. Like known DOD endonucleases, these motifs have sequences of nine and ten residues, which form the centers recognizing double-stranded DNA and are separated by a linker of 90–130 residues. The active center involves conserved Gly residues, which are in positions 3 and 9 of motif C and 4 and 10 of motif E. In addition, the motifs each harbor catalytically active Asn and Lys

Motif D (eight residues) is in the linker between motifs C and E. Substitution of its Lys2 completely abolishes endonuclease activity of the DOD domain in some cases [9]. This indicates that, together with motifs C and E, motif D is involved in the formation of the active endonuclease domain.

Motif H consists of 19 amino acid residues, of which Leu13–Leu14 are rather conserved. These residues are probably involved in the intein–DNA interaction. Motifs C, D, E, and H form the DOD endonuclease domain, which occurs in many, thought not all, inteins. It should be noted that, according to the available experimental data, the DOD domain is unnecessary for protein splicing [8, 9, 11, 12, 14] but ensures intein homing (see below).

Motifs F and G form the C-terminal splicing domain, which is 25–40 residues in size [11]. Motif F consists of 16 residues, half of which are highly conserved (table). Motif G is a short C-terminal sequence of eight residues, of which seven belong to the intein and one is the N-terminal residue of C-extein. Motifs F and G are separated by a small linker, usually consisting of two to five residues. The last amino acid is Asn in most cases (or, extremely rarely, Gln or Asp); the last but one is His. The two last amino acid residues play an important role in hydrolyzing the peptide bond at the C end of the intein [9, 11], while the N-terminal residue of C-extein (hereafter referred to as residue +1) is critical for extein ligation. Position +1 is occupied by Ser, Thr, or Cys in the majority of known C-exteins.

MAIN FEATURES OF INTEINS

Although structurally heterogeneous, all inteins have some features in common. Four main features of the intein sequence are now recognized.

(1) An intein-coding gene has a sequence absent from its homologs of other organisms.

(2) A mature protein differs in size from the product deduced from its coding sequence by more than 100 residues.

(3) A protein has specific motifs A, B, F, and G (the presence or absence of the DOD endonuclease domain is not considered to distinguish an intein).

(4) A protein has four conserved amino acid residues: Ser, Thr, or Cys at the N end of a putative intein; His–Asn or His–Gln at the C end of the intein; and Ser, Thr, or Cys at the N end of the C-extein.

APPLICATIONS.

- Rapid purification of target proteins
- Temperature sensitive control of protein activity by conditionally splicing inteins.

Solid-phase peptide synthesis

Solid-phase peptide synthesis (SPPS), pioneered by Robert Bruce Merrifield, caused a paradigm shift within the peptide synthesis community, and it is now the standard method for synthesizing peptides and proteins in the lab. SPPS allows for 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.

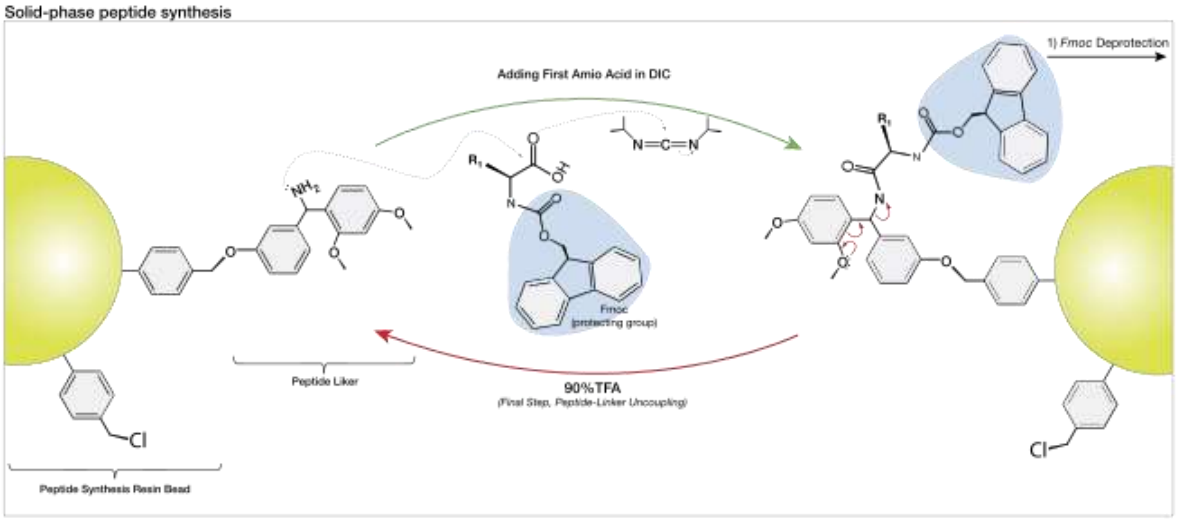
Small porous beads are treated with functional units ('linkers') on which peptide chains can be built. The peptide will remain covalently attached to the bead until cleaved from it by a reagent such

as anhydrous hydrogen fluoride or trifluoroacetic acid. The peptide is thus 'immobilized' on the solid-phase and can be retained during a filtration process while liquid-phase reagents and by-products of synthesis are flushed away.

The general principle of SPPS is one of repeated cycles of deprotection-wash-coupling-wash. The free N-terminal amine of a solid-phase attached peptide is coupled (see below) 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.

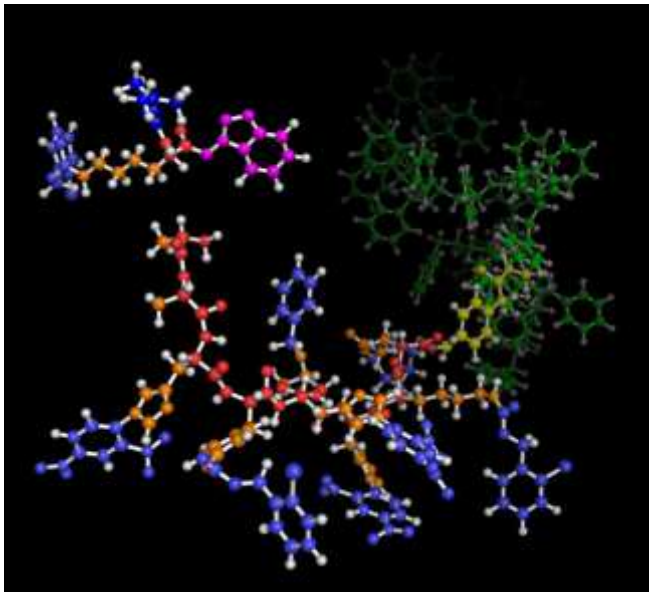
The overwhelmingly important consideration is to generate extremely high yield in each step. For example, if each coupling step were to have 99% yield, a 26-amino acid peptide would be synthesized in 77% final yield (assuming 100% yield in each deprotection); if each step were 95%, it would be synthesized 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.[*citation needed*]

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 either of 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.[2] 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.



Novel Proteins

Pets such as dogs and cats are often fed diets consisting of chicken, beef, lamb and fish. If an intolerance develops, it may be hard to discern exactly what ingredient causes the problem. For these pets, offering a diet consisting of novel proteins such as bison, duck, rabbit, fish they haven't eaten before, venison, kangaroo or egg, either alone or with a single carbohydrate source such as potato or rice, can help resolve the problem.

Use of novel protein

Increasing the production and use of novel proteins is one of our research lines to help building new protein value chains. Durable success and impact of these value chains depend entirely on the ability to use the technical and nutrition functionality of proteins in a final product, and – even more important – on the acceptance of novel proteins by consumers and regulatory bodies.A typical example of a new value chain is the FoodWaste2Feed project. The acceptance of novel proteins by consumers is studied in several programs, focussing on various protein sources and on different target groups.Most novel proteins to be used as food or food ingredient will have to be approved prior to market introduction under the Novel Food Regulation (Regulation (EC) No 258/97). Wageningen UR developed a Guideline for producers of novel proteins on how to fill the NFR application dossiers (Guideline LEI 14-075). There is no pre-market authorisation procedure for novel feed ingredients derived from non-animal sources, a notification will suffice (Regulation (EC) No 767/2009). For

proteins derived from animals, like insect proteins, the situation is rather complex and depends among others on the destination of the feed (food producing animal or not).

3.4. Site-directed mutagenesis

Site-directed mutagenesis is a procedure used to induce a specific mutation in a cell. It may be used for a host of reasons, including the generation of restriction sites, investigating the role of a gene or regulatory element by knock-out, understanding the role a particular amino acid has in a protein, or the creation of new and 'better' proteins with, for example, greater thermal stability or more efficient catalytic ability.

The method quite simply involves template DNA - the DNA to be mutated, usually bacterial DNA - and an oligonucleotide carrying the reverse complement of the desired mutation which can anneal to the template and be used as a primer for DNA synthesis. For instance, if a TGG codon is present in the bacterial DNA, and the desired mutation is AAT, then the oligonucleotide primer should read ATT (all read from 5' to 3'). Once the mutagenic primer is annealed to its template, the complete structure is called a *heteroduplex*, owing to the differences between the strands. The heteroduplex is used to transform a cell - most often *E. coli* - and it is left there overnight.

In theory, both strands of the heteroduplex should be replicated at equal frequency to give a 50/50 mixture of mutant to template DNA in the cell. In practice, mutant recovery in this way is poor for two reasons: firstly, because of the cell's intrinsic mismatch repair system, and secondly, because the template DNA is methylated, and methylated DNA is preferentially replicated by the host cell machinery. Consequently, higher-efficiency mutagenesis approaches have been developed to raise the percentage mutant recovery from around 0.1% to as high as 50%. These approaches work on the principle that once the template DNA has been used to copy the mutant strand it is of no further use, and can only hinder mutant recovery.

Basic mechanism

The basic procedure requires the synthesis of a short DNA primer. This synthetic primer contains the desired mutation and is complementary to the template DNA around the mutation site so it can hybridize with the DNA in the gene of interest. The mutation may be a single base change (a point mutation), multiple base changes, deletion, or insertion. The single-strand primer is then extended using a DNA polymerase, which copies the rest of the gene. The gene thus copied contains the mutated site, and is then introduced into a host cell as a vector and cloned. Finally, mutants are selected by DNA sequencing to check that they contain the desired mutation.

The original method using single-primer extension was inefficient due to a low yield of mutants. This resulting mixture contains both the original unmutated template as well as the mutant strand, producing a mixed population of mutant and non-mutant progenies. Furthermore the template used is methylated while the mutant strand is unmethylated, and the mutants may be counter-selected due to presence of mismatch repair system that favors the methylated template DNA, resulting in

fewer mutants. Many approaches have since been developed to improve the efficiency of mutagenesis.

3.5. Quik change

Quikchange is one high-efficiency mutagenesis approach that has been developed. The plasmid template is denatured and mutagenic primers are annealed to each strand. The primers are de-phosphorylated so that although they can be extended, there cannot be ligation between the end of the synthesised strand and the start of the primer. Once DNA synthesis is complete, the template DNA and mutagenic DNA are denatured in each of the two PCR-like products. The mutant strands cannot be reused for DNA synthesis because when the primers anneal to them, they have no template material to copy. Instead the parental strands are reused: one new mutagenic primer is added to each, DNA is synthesised, the products are denatured and then the parental strands used again. Unlike conventional PCR, which makes products exponentially, this is a linear amplification: two new mutated strands are made in each 'cycle'. At the end of this amplification period, the parental templates are recognised by a methylation-specific restriction enzyme called Dpn I. Because the mutated DNA is unmethylated, it goes unrecognised by this enzyme and remains intact. The consequence is that parental DNA cannot be used to transform *E. coli*, while mutant DNAcan.Although currently the mutant products are all linearised, their lengthy (approx 40 nt) complementary primers can anneal to result in a double-stranded circular plasmid containing a homoduplex of only mutant DNA. The *E. coli* repair system will complete ligation where the dephosphorylated primers fail to do this, to form complete plasmids readyforhostcellreplication.

A Quik change protocol might look something like this:

1. Design 2 long (25-40nt) complementary primers containing the mutations
- complementarity is important for circularisation of the mutant plasmid later on
- the melt temperature of the primers should be around 78C
- there is no need for either primer to have a 5' phosphate as there is no ligation step

2. Mix the template plasmid, primers, dNTPs and a thermostable polymerase and run for 16-25 thermal cycles

3. Digest (methylated) DNA with Dpn I

4. Transform *E. coli* with the remaining DNA and leave overnight

5. Pick four colonies and isolate plasmid DNA

6. Sequence the plasmid to ensure that the mutation has been correctly inserted

3.6. Uracil-containing DNA method

This approach is based on the simple notion that (deoxy)uracil is not a usual component of DNA. It involves the following protocol:

1. Grow the template DNA to contain a high proportion of deoxyuracil (dU) by growing it in an *E. coli* mutant:
- dut- (which lacks dUTPase; an enzyme which normally prevents the incorporation of uracil into DNA)
- ung- (which lacks uracil glycosylase; an enzyme which normally removes uracil from DNA)

2. Anneal the mutagenic primers, as usual, and begin DNA synthesis

3. The next step can either be performed in vivo or in vitro:

In vivo: transform the heteroduplex DNA into a wild-type *E. coli* which retains its uracil glycosylase function (ung+). The template DNA, which is rich in dU, will then be repaired using the newly-incorporated mutant strand as a template. The product is a homoduplex DNA containing only mutant strands.

In vitro: extract the heteroduplex DNA from *E. coli* and treat with uracil glycosylase to remove the parental DNA. Then synthesise a new strand with DNA polymerase and dNTPs, using the mutant strand as a template. Both of these are performed in the test tube. The product, again, is a homoduplex DNA containing only mutant strands.

3.7. Cassette mutagenesis

Cassette mutagenesis is a technique employed to introduce multiple mutations to the same region of DNA. A cassette (block of DNA) is designed to contain all of the desired mutations and then given ligatable ends to facilitate its insertion into the wild-type DNA. Quikchange, described above, can be used to generate suitable restriction sites for its insertion, and the cassette should have both 5' phosphorylation and 4-base 'sticky' overhangs at each end in order to encourage its insertion into the host molecule.

3.8. PCR mutagenesis

PCR mutagenesis is similar in principle to Quikchange. The target plasmid is heated to denature, mutagenic primers (forward and reverse) are added to each strand, and roughly 8 cycles are performed to amplify the mutant plasmid (fewer cycles are ideal to minimise the risk of error). The methylated (parental) DNA is then treated with Dpn I and *E. coli* is transformed with the mutant homoduplexes and left to grow overnight. Again, the plasmid DNA is isolated from selected colonies and sequenced to ensure that the desired mutation has been incorporated.

3.9. Sticky feet PCR

Sticky feet PCR is used to generate *insertional* mutations in the wild-type DNA. The mutagenic primer contains a series of bases (the desired insertion) which is not present in the template DNA. Because it cannot form complementary pairs with the template DNA upon annealing, the desired insertion 'loops out'. When the primer is extended, to generate heteroduplex DNA, the parental strand is digested, as usual, using Dpn I. This leaves a single-stranded mutant strand, containing the insertion, which then itself acts as a template for new DNA synthesis; the nascent DNA strand will contain the complement of the insertion. The product is homoduplex DNA containing both strands with the insertional mutation.

The size of insertion that can be generated by sticky feet is limited, however, by the size of oligonucleotide primer that can be accurately synthesised (certainly no more than 80 nucleotides in length). Deletions are performed in a similar manner, except the mutagenic primer *lacks* the bases which need to be deleted (it contains only the flanking sequences). Upon annealing, this causes the deletion bases in the *template* DNA to 'loop out' because they have nothing to anneal to. Unlike with insertions, there is no size limitation to deletions because the oligonucleotide only need be big enough to correspond to the flanking regions of the desired site of deletion.

3.10. Random mutagenesis

Early approaches to mutagenesis rely on methods which are entirely random in the mutations produced. Cells or organisms may be exposed to mutagens such as UV radiation or mutagenic chemicals, and mutants with desired characteristics are then selected. Hermann Muller discovered that x-rays can cause genetic mutations in fruit flies (published in 1927),and went on to use the *Drosophila* mutants created for his studies on genetics. For *Escherichia coli*, mutants may be selected first by exposure to UV radiation, then plated onto agar medium. The colonies formed are then replica-plated, one in rich medium, another in minimal medium, and mutants that have specific nutritional requirements can then be identified by their inability to grow in minimal medium. Similar procedures may be repeated with other types of cells and with different media for selection.

A number of methods for generating random mutations in specific proteins were later developed to screen for mutants with interesting or improved properties. These methods may involve the use of doped nucleotides in oligonucleotides synthesis, or conducting a PCR reaction in conditions that enhance misincorporation of nucleotides (error-prone PCR), for example by reducing the fidelity of replication or using nucleotide analogues. PCR products which contain mutation are then cloned into an expression vector and the mutant proteins produced can then be characterised.

In animal studies, alkylating agents such as *N*-ethyl-*N*-nitrosourea (ENU) have been used to generate mutant mice. Ethyl methanesulfonate (EMS) is also often used to generate animal and plant mutants.Random mutagenesis is an incredibly powerful tool for altering the properties of enzymes. Imagine, for example, you were studying a G-protein coupled receptor (GPCR) and wanted to create

a temperature-sensitive version of the receptor or one that was activated by a different ligand than the wild-type.

1. Error-prone PCR. This approach uses a "sloppy" version of PCR, in which the polymerase has a fairly high error rate (up to 2%), to amplify the wild-type sequence. The PCR can be made error-prone in various ways including increasing the MgCl2 in the reaction, adding MnCl2 or using unequal concentrations of each nucleotide. Here is a good review of error prone PCR techiques and theory. After amplification, the library of mutant coding sequences must be cloned into a suitable plasmid. The drawback of this approach is that size of the library is limited by the efficiency of the cloning step. Although point mutations are the most common types of mutation in error prone PCR, deletions and frameshift mutations are also possible. There are a number of commercial error-prone PCR kits available, including those from Stratagene and Clontech

2. Rolling circle error-prone PCR is a variant of error-prone PCR in which wild-type sequence is first cloned into a plasmid, then the whole plasmid is amplified under error-prone conditions. This eliminates the ligation step that limits library size in conventional error-prone PCR but of course the amplification of the whole plasmid is less efficient than amplifying the coding sequence alone. More details can be found here.

3. Mutator strains. In this approach the wild-type sequence is cloned into a plasmid and transformed into a mutator strain, such as Stratagene's XL1-Red. XL1-red is an *E.coli* strain whosedeficiency in three of the primary DNA repair pathways (*mutS, mutD* and *mutT*) causes it to make errors during replicate of it's DNA, including the cloned plasmid. As a result each copy of the plasmid replicated in this strain has the potential to be different from the wild-type. One advantage of mutator strains is that a wide variety of mutations can be incorporated including substitutions, deletions and frame-shifts. The drawback with this method is that the strain becomes progressively sick as it accumulates more and more mutations in it's own genome so several steps of growth, plasmid isolation, transformation and re-growth are normally required to obtain a meaningful library.

4. Temporary mutator strains. Temporary mutator strains can be built by over-expressing a mutator allele such as *mutD5* (a dominant negative version of *mutD*) which limits the cell's ability to repair DNA lesions. By expressing *mutD5* from an inducible promoter it is possible to allow the cells to cycle between mutagenic (*mutD5* expression on) and normal (*mutD5* expression off) periods of growth. The periods of normal growth allow the cells to recover from the mutagenesis, which allows these strains to grow for longer than conventional mutator strains. If a plasmid with a temperature-sensitive origin of replication is used, the mutagenic plasmid can easily be removed restore normal DNA repair, allowing the mutants to be grown up for analysis/screening. An example of the construction and use of such a strain can be found here. As far as I am aware there are no commercially available temporary mutator strains.

5. Insertion mutagenesis. Finnzymes have a kit that uses a transposon-based system to randomly insert a 15-base pair sequence throughout a sequence of interest, be it an isolated insert or plasmid. This inserts 5 codons into the sequence, allowing any gene with an insertion to be expressed (i.e. no frame-shifts or stop codons are cause). Since the insertion is random, each copy of the sequence will have different insertions, thus creating a library.

6. Ethyl methanesulfonate (EMS) is a chemical mutagen. EMS aklylates guanidine residues, causing them to be incorrectly copied during DNA replication. Since EMS directly chemically modifies DNA, EMS mutagenesis can be carried out either in vivo (i.e. whole-cell mutagenesis) or in vitro.

7. Nitrous acid is another chemical mutagen. It acts by de-aminating adenine and cytosine residues causing transversion point mutations (A/T to G/C and vice versa). **Note***: I have only mentioned two chemical mutagens but there are many others. Hirokazu Inoue has written an excellent article describing some of them and their use in mutagenesis*

8. DNA Shuffling is a very powerful method in which members of a library (i.e. copies of same gene each with different types of mutation) are randomly shuffled. This is done by randomly digesting the library with DNAseI then randomly re-joining the fragments using self-priming PCR. Shuffling can be applied to libraries produced by any of the above method and allows the effects of different combinations of mutations to be tested.

3.11. RECOMBINANT PROTEINS EXPRESSING METHODS

Choosing an appropriate method for expressing a recombinant protein is a critical factor in obtaining the desired yields and quality of a recombinant protein in a timely fashion. Selecting a wrong expression host can result in the protein being misfolded or poorly expressed, lacking the necessary posttranslational modifications or containing inappropriate modifications. Factors to consider when selecting an expression system include the mass of the protein and number of disulfide bonds, type of posttranslational modifications desired on the expressed protein, and the destination of the expressed protein. The intended application of the purified recombinant protein is also critical in the decision-making process and the applications can be categorized into four broad areas: structural studies, in vitro activity assays, antigens for antibody generation, and in vivo studies. The purpose of this chapter is to help guide the investigator in the decision-making process for choosing an appropriate expression system. However, even with the described guidelines there are many circumstances when it is not obvious a priori which expression system is the best choice, and the use of multiple expressions systems must be attempted before an optimal system is identified. Numerous expression systems are currently being used in academic and industrial settings. Some of these systems are too new and insufficiently tested to comment on their utility. In addition, some established systems for expressing recombinant proteins, such as transgenic animals, are too technically challenging, time consuming and prohibitively expensive to be a viable option for the average laboratory. For the purpose of this chapter, only Escherichia coli, Pichia pastoris, baculovirus/insect cell, and mammalian expression systems will be considered. These four systems have straightforward protocols, are readily accessible either from colleagues or from research product companies (e.g., Invitrogen, EMD-Novagen, Stratagene, and Promega), and are relatively inexpensive for small-scale production. The characteristics and available options of these expression systems will be briefly reviewed with the focus on the differences between the systems. Strategies will then be presented to help guide the investigator in making the best choice for an expression system.

Escherichia coli

The bacteria E. coli was the first host used to express recombinant proteins and is still considered to be the workhorse in the field. Using the E. coli system offers a rapid and simple method for expressing recombinant proteins due to its short doubling time. Consequently, the assessment of recombinant gene expression in E. coli can take less than a week. The growth media for E. coli are inexpensive and there are relatively straightforward methods to scale-up bioproduction. In E. coli, recombinant proteins are normally either directed to the cytoplasm or to the periplasm and, to a lesser extent, secreted. Proteins directed to the cytoplasm are the most efficiently expressed, giving yields of up to 30% of the biomass. However, the high expression of recombinant protein can often lead to the accumulation of aggregated, insoluble protein that forms inclusions bodies. Inclusion bodies have been observed not only with eukaryotic proteins but also to a lesser extent with overexpressed proteins from prokaryotes including E. coli. The rate of translation and folding in E. coli is almost 10-fold higher than that observed in eukaryotic cells, and this presumably contributes to the inclusion body formation of eukaryotic proteins. Inclusion bodies can be a significant hindrance in obtaining soluble, active protein in some situations. However, in some cases, inclusion bodies are advantageous because they are resistant to proteolysis, easy to concentrate by centrifugation, minimally contaminated with other proteins, and, with some effort, able to be refolded to form active, soluble proteins.

E. coli: Temperature and molecular chaperones

Several methods have been described for maximizing the formation of soluble, properly folded proteins in the cytoplasm and minimizing inclusion body formation. The most straightforward method involves lowering the temperature to 15–30 C during the expression period. Presumably, the reduced temperature slows the rate of transcription, translation, and refolding, thereby allowing for proper folding. In addition, lower temperature has been shown to decrease heat shock protease activity. Some investigators have co expressed molecular chaperones in the cytoplasm along with the recombinant protein for promoting protein solubility. The utility of this approach appears to be quite protein-specific, and therefore needs to be tested individually for each recombinant protein of interest.

E. coli: Fusion partners

Alternatively, a method that promotes solubility with many proteins is to fuse the recombinant protein at either the N-terminus or C-terminus to a soluble fusion tag. Fusion partners that have been shown to increase solubility of recombinant proteins include glutathione-S-transferase (GST), thioredoxin, maltose-binding protein (MBP), small ubiquitin-modifier (SUMO), and N-utilization substrate (NusA). Both GST and MBP have the added advantage of also being an affinity purification tag. Unfortunately, no single tag appears to work for all recombinant proteins, and multiple fusion partners may need to be evaluated for promoting soluble expression. Fusion tags can be removed from the recombinant protein by several strategies, and a widespread approach involves adding a protease site between the fusion partner and the recombinant protein that can be cleaved with the specific protease. This approach must be carefully tested since removal of the fusion tag can, in some cases, render the recombinant protein insoluble.

E. coli: Disulfide bond formation

E. coli is normally inefficient in promoting the correct formation of disulfide bonds when recombinant proteins are expressed in the cytoplasm; normally disulfide bond formation occurs only in the periplasm where it is catalyzed by the Dsb system. Consequently, if disulfide bond formation is needed, the recombinant protein can be directed to the periplasm via a cleavable signal peptide (e.g., pelB). However, a major disadvantage of periplasmic expression is the significant reduction in production yields. Through engineering of the E. coli genome, a more suitable environment for disulfide bond generation in the cytoplasm can be induced by disrupting the thioredoxin reductase (trxB) and glutathione reductase (gor) genes in the Dsb system which in turn enables thioredoxin and glutaredoxin to promote cytoplasmic reduction of cysteines. These engineered strains are commercially available through EMD-Novagen (Origami). If additional disulfide bond formation is still needed, the recombinant protein can be fused to thioredoxin, and the fusion protein expressed in a trxB/gor E. coli strain.

E. coli: Posttranslational modifications

Finally, it is important to recognize that E. coli has a limited capacity for posttranslational modifications compared to eukaryotic organisms. For example, E. coli does not support enzyme-mediated N-linked glycosylation, O-linked glycosylation, amidation, hydroxylation, myristoylation, palmitation, or sulfation.

PICHIA PASTORIS

Yeast is another traditional, powerful tool for expressing recombinant proteins and has been used successfully to express a multitude of proteins. Yeast has many of the advantageous features of E. coli such as a short doubling time and a readily manipulated genome, but also has the additional benefits of a eukaryote that includes improved folding and most posttranslational modifications. The first yeast routinely used for recombinant protein expression was Saccharomyces cerevisiae. However, in the last 15 years, P. pastoris has become the yeast of choice because it typically permits higher levels of recombinant protein expression than does S. cerevisiae. P. pastoris is a methyltropic yeast, and can use methanol as its only carbon source. The growth of P. pastoris in methanol-containing medium results in the dramatic transcriptional induction of the genes for alcohol oxidase (AOX) and dihydroxyacetone synthase. After induction, these proteins comprise up to 30% of the P. pastoris biomass. Investigators have exploited this methanol-dependent gene induction by incorporating the strong, yet tightly regulated, promoter of the alcohol oxidase I (AOX1) gene into the majority of vectors for expressing recombinant proteins. The P. pastoris expression vectors integrate in the genome whereas by contrast, S. cerevisiae vectors use the more unstable method of replicating episomally. The length of time to assess recombinant gene expression with the P. pastoris method is approximately 3–4 weeks which includes the transformation of yeast, screening the transformants for integration, and an expression timecourse. An appealing feature of P. pastoris is the extremely high cell densities achievable under appropriate culture conditions. Using inexpensive medium, the P. pastoris culture can reach 120 g/l of dry cell weight density. An important caveat is that the induction medium requires a low percentage of methanol. In large-scale cultures, the amount of methanol becomes a fire hazard requiring a new level of safety conditions.

P. pastoris has been used to obtain both intracellular and secreted recombinant proteins. Like other eukaryotes, it efficiently generates disulfide bonds and has successfully been used to express proteins containing many disulfide bonds. To facilitate secretion, the recombinant protein must be engineered to carry a signal sequence. The most commonly used signal sequence is the pre-pro sequence from S. cerevisiae a-mating factor. Because P. pastoris secretes few endogenous proteins, purification of the recombinant protein from the medium is a relatively simple task. If proteolysis of the recombinant protein is a concern, expression can be completed using the pep4 protease-deficient strain of P. pastoris selecting an Expression System 135. This strain has reduced vacuole peptidase A activity which is responsible for activation of carboxypeptidase Y and protease B1.

Yeast has the posttranslational capacity to add glycans at both specific asparagine residues (N-linked) and serine/threonine residues (O-linked). These glycan structures are substantially different from the modifications added by insect and mammalian cells. In P. pastoris the N-linked glycan is a high mannose type and usually contains 8–17 mannoses, which is quite different from S. cerevisiae structures that consist of approximately 50–150 mannose residues. Similar to insect and mammalian cells, the consensus sequence for N-linked glycans in yeast is Asn-Xaa-Ser/Thr. Two groups have completed extensive engineering to create P. pastoris strains that produce complex N-linked glycan structures comparable to those produced by mammalian cells. Wever, only the strains developed by Roland Contreras' group are available to investigators and must be licensed through Research Corporation Technologies. The O-linked structures in P. pastoris have not been studied comprehensively but are known to be formed by the addition of one to four mannose residues to serines/threonines. Several reports have indicated that expression of certain proteins in P. pastoris resulted in the addition O-linked glycans not observed when the protein was expressed endogenously in mammalian cells.

Baculovirus/Insect Cells

Baculovirus-mediated expression in insect cells offers another useful tool for generating recombinant proteins. Baculovirus is a lytic, large (130 kb), double-stranded DNA virus, and the Autographa californica virus is the most commonly used baculovirus isolate for recombinant expression. Baculovirus is routinely amplified in insect cell lines derived from the fall armyworm Spodoptera frugiperda (Sf 9, Sf 21), and recombinant protein expression is completed either in the aforementioned lines or in a line derived from the cabbage looper Trichoplusia ni (High-Five). Originally, creating recombinant baculoviruses involved cotransfecting the gene of interest flanked by baculovirus sequence with baculovirus DNA into insect cells, and screening for rare homologous recombination events. Recombinants were identified by screening plaques with a modified morphology, and often additional rounds of plaque screening were required to ensure that the recombinant viral preparation was not contaminated with wild-type virus. This lengthy and laborious process for generating recombinant viruses has been largely replaced by using site-specific transposition (Bac-to-Bac or BaculoDirect, Invitrogen) or an improved homologous recombination method with an engineered 136 William H. Brondyk baculovirus containing a lethal mutation in orf1629 (flashBAC from Oxford Expression Technologies or BacMagic from EMD-Novagen). Both of these approaches overcome the requirement to isolate plaques because the efficiency of recombination is 100%. Following one or two rounds of amplifying the recombinant baculovirus, the

investigator can quantify the baculovirus concentration stock either by the plaque assay or by using the newer, more rapid real-time PCR or antibody-based assays. The improvements in creating and quantifying recombinant baculoviruses have dramatically reduced the time for evaluating baculovirus expression to approximately 3 weeks, including a time-course study for optimizing expression.

The most common promoters used with baculovirus expression are the polH and p10 promoters, both of which induce a high level of expression in the very late phase of the baculovirus infection. During this phase, cells undergo cell death with the concomitant release of proteases, which can result in degradation of the expressed recombinant protein. To reduce proteolysis of the recombinant protein, promoters active in earlier phases of the lytic cycle such as the basic promoter have been used. Alternatively proteolytic activity can be minimized by using constructs deleted in the chiA and v-cath genes, which encode chitinase and a cathepsin protease, respectively. Baculovirus-mediated expression is routinely used to generate both cytoplasmic and secreted recombinant proteins. Efficient secretion generally requires the presence of a signal peptide. Both insect and mammalian signal sequences can promote entry into the insect cell secretory pathway. Insect cells were originally grown in serum-containing medium which complicated purification of the secreted proteins. Recent advances in media development permit the replacement of serum with protein hydrolysates derived from either animal tissues or plants, thereby greatly simplifying protein purification. However, the high cost of this specialized media can limit its use for large-scale bioproduction. Insect cells efficiently generate disulfide bonds in recombinant proteins. They also produce the majority of the posttranslational modifications found in mammalian cells. However, the N-linked glycan structure formed in most insect cells is the predominantly fucosylated paucimannose structures (Man3GlcNAc2-N-Asn). This finding has prompted the recent generation of insect cell lines that produce glycoproteins with the complex N-linked glycans normally found in mammalian cells. A transgenic Sf-9 insect line expressing several glycosyltransferases is commercially available (Mimic cell line, Invitrogen) and produces N-linked glycans containing a biantennary, sialylated structure. There are only a few reports describing the O-linked glycans structures generated by insect cells.

MAMMALIAN CELLS

Mammalian expression methods have conventionally been considered to be the least efficient vehicle for expressing recombinant proteins. However, recent advances have significantly improved the expression levels from mammalian cell lines. For example, stably transfected Chinese hamster ovary (CHO) cells have been reported to express recombinant antibodies up to a level of a few grams per liter. While many cell lines and expression strategies have been tested, this chapter will focus on transient transfection in human embryonic kidney (HEK293) cells and stable transfection with CHO cells.

The HEK293 cell line was derived from human embryonic kidney cells transformed with adenovirus. HEK293 cells can be transiently transfected with a high efficiency (>80%) using certain cationic lipids, calcium phosphate, or polyethyleneimine as transfection reagents. For large-scale transient transfections (>100 ml), calcium phosphate or polyethyleneimine reagents are more cost-effective options when compared to cationic lipids. Transient transfections have been performed at even the bioreactor level but for most laboratories this scale is technically challenging. The transient

transfection method is relatively easy, and the evaluation for a given recombinant protein can be made in less than 2 weeks.

CHO cells are commonly used for mammalian expression when large quantities of recombinant protein are needed. For example, most therapeutic antibodies currently on the market are manufactured using this method. The standard method for stable CHO expression involves transfecting dihydrofolate reductase (DHFR)-deficient CHO cells with a DHFR selection cassette along with an expression cassette containing the gene of interest. Dihydrofolate reductase converts dihydrofolate into tetrahydrofolate which is required for the de novo synthesis of purines, certain amino acids, and thymidylic acid. Methotrexate, which binds and inhibits DHFR, is used as a selection agent and only those cells that have integrated the DHFR selection cassette will survive. Sequentially increasing the concentration of methotrexate will result in amplification of the DHFR gene along with the linked gene of interest. Following at least one round of selection with the drug methotrexate, the stably transfected pools are subcloned using limiting dilution cloning into multiwell plates. Typically only a small percentage of the screened subclones will be expressing the recombinant gene at a high level since in the majority of the clones, the expression cassette has integrated into the heterochromatin region which is transcriptionally inactive. Unfortunately, the entire selection and screening process takes at least 2–3 months, making this the major drawback of the CHO 138 William H. Brondyk method. However, recent high-throughput methods based on flow cytometry or automation have increased the ease in rapidly screening and selecting high expressing clones. Another development has been to use specific cis-acting DNA elements flanking the recombinant gene cassette that confer active transcription to integration sites. Unfortunately, the majority of these DNA elements is owned by companies and must be licensed for use in the laboratory, and, even with the aforementioned advances in CHO expression, the timelines for generating a high expressing CHO clone have not changed considerably.

Mammalian expression systems are used primarily to generate secreted rather than intracellular recombinant proteins. Serum-free media have been developed for both the CHO and HEK293 cell lines, which simplifies the purification of secreted recombinant proteins. However, the cost of the media is quite high, making large-scale bioproduction rather costly. Mammalian cells contain the most superior folding and disulfide bond formation when compared to other expression hosts. The N-linked and O-linked glycan structures formed by mammalian cells are extremely varied and are not only dependent on the protein but also on the mammalian cell type used as the expression host. Furthermore, the cell culture conditions such as nutrient content, pH, temperature, oxygen levels and ammonia concentration can significantly affect the glycosylation profile. N-linked glycosylation can result in oligomannose, hybrid, and complex structures, and the structures all contain the Man3GlcNac2 core. The oligomannose glycans can have two to six additional mannoses and the mannoses can be phosphorylated or sulfated. The most common complex structures have two to four Gal b1,4-GlcNac2 attached to the mannoses which result in bi-, tri-, and tetra-antennary branches. The branches can terminate with sialic acid, and fucose can also be attached to the structures. Hybrid structures contain features of both the oligomannose and complex structures. O-Glycosylation structures can be classified into eight types based on their core structures: O-GalNAc-type glycosylation, O-GlcNAc-type glycosylation, O-fucosylation, O-mannosylation, O-glucosylation, phosphoglycosylation, O-glycosaminoglycan-type glycosylation, and collagen-type glycosylation.

3.12. Proteins structure: Crystallography

X-ray crystallography is essentially a form of very high resolution microscopy. It enables us to visualize protein structures at the atomic level and enhances our understanding of protein function. Specifically we can study how proteins interact with other molecules, how they undergo conformational changes, and how they perform catalysis in the case of enzymes. Armed with this information we can design novel drugs that target a particular protein, or rationally engineer an enzyme for a specific industrial process.

In all forms of microscopy, the amount of detail or the resolution is limited by the wavelength of the electro-magnetic radiation used. With light microscopy, where the shortest wavelength is about 300 nm, one can see individual cells and sub-cellular organelles. With electron microscopy, where the wavelength may be below 10 nm, one can see detailed cellular architecture and the shapes of large protein molecules. In order to see proteins in atomic detail, we need to work with electro-magnetic radiation with a wavelength of around 0.1 nm or 1 Å, in other words we need to use X-rays.

In light microscopy, the subject is irradiated with light and causes the incident radiation to be diffracted in all directions. The diffracted beams are then collected, focused and magnified by the lenses in the microscope to give an enlarged image of the object. The situation with electron microscopy is similar only in this case the diffracted beams are focused using magnets. Unfortunately it is not possible to physically focus an X-ray diffraction pattern, so it has to be done mathematically and this is where the computers come in. The diffraction pattern is recorded using some sort of detector which used to be X-ray sensitive film, but nowadays is usually an image plate or a charge-coupled device (CCD).

The diffraction from a single molecule would be too weak to be measurable. So we use an ordered three-dimensional array of molecules, in other words a crystal, to magnify the signal. Even a small protein crystal might contain a billion molecules. If the internal order of the crystal is poor, then the X-rays will not be diffracted to high angles or high resolution and the data will not yield a detailed structure. If the crystal is well ordered, then diffraction will be measurable at high angles or high resolution and a detailed structure should result. The X-rays are diffracted by the electrons in the structure and consequently the result of an X-ray experiment is a 3-dimensional map showing the distribution of electrons in the structure.

A crystal behaves like a three-dimensional diffraction grating, which gives rise to both constructive and destructive interference effects in the diffraction pattern, such that it appears on the detector as a series of discrete spots which are known as reflections. Each reflection contains information on all atoms in the structure and conversely each atom contributes to the intensity of each reflection. As with all forms of electro-magnetic radiation, X-rays have wave properties, in other words they have both an amplitude and a phase. In order to recombine a diffraction pattern, both of these parameters are required for each reflection. Unfortunately, only the amplitudes can be recorded experimentally all phase information is lost. This is known as "the phase problem". When crystallographers say they have solved a structure, it means that they have solved "the phase

problem". In other words they have obtained phase information sufficient to enable an interpretable electron density map to be calculated.

Crystal structure determination

Firstly we need to obtain a pure sample of our target protein. We can do this by either isolating it from its source, or by cloning its gene into a high expression system. The sample then needs be assessed for suitability according to the following criteria:

1. Is it pure and homogeneous? we can test this by various electrophoretic methods and mass spectrometry .
2. Is the protein soluble and folded? if protein estimations suggest that a lot of protein is being lost, then it may be due to precipitation. The degree of ordered secondary structure can be tested with circular dichroism if this is very low then the protein may be misfolded. This may occur if the protein is being produced faster than it can fold and may result in the formation of insoluble inclusion bodies. Attenuating the induction can alleviate this problem e.g. using a lower temperature.
3. Is the sample monodisperse? in other words is the sample free from aggregation? This can be monitored using a dynamic light scattering (DLS) device.
4. Is the protein still active? check with activity assays
5. Is the sample stable? Occasionally good protein crystals will form overnight at room temperature, but usually it may take several days to one or two weeks before suitable crystals can grow. Therefore, ideally the sample needs to remain stable over that period

If the sample fails one or more of the above criteria, it may be worthwhile returning to the expression and purification protocols and trying something different, such as the addition of ligands known to interact with the protein, or adding extra purification steps. In extreme cases it may be worthwhile switching to a different expression system altogether or working with a mutated or truncated construct. It may be possible to refold protein successfully using chaotropic reagents such as urea. Aggregated or polydisperse samples may be made monodisperse by simply changing pH or adding some salt. However, without DLS, this is very difficult to assess.

Crystallization

Before beginning trials the sample needs to be concentrated and transferred to dilute buffer containing little or no salt if the protein is happy under these conditions. This can easily be achieved using centrifugal concentrators. In order to screen a reasonable number of conditions we need at least 200 m l of protein at 10 mg/ml. If this is not the case then you may need to scale up the expression and purification to make it so.

If a similar protein has already been crystallized then it is definitely worth trying the conditions used to grow crystals of this protein. In any case if you have enough material one would normally subject it to one or more sparse matrix screens. To date the total number of different conditions in our repertoire of screens comes to about 400.

We normally use these tissue culture trays to set up crystallizations with up to 24 different conditions per tray. The method used is hanging drop vapour diffusion it has the advantage of being the least expensive on protein. The set up is as follows:

The well is prepared first and usually contains 1ml of a buffered precipitant solution such as polyethylene glycol or ammonium sulfate or even a mixture of PEG and salt. Sometimes additives are also included such as detergents or metal ions which may enhance the crystallization. Then 1 m l of the concentrated protein sample is pipetted onto a siliconized coverslip, followed by 1 m l of the well solution. The coverslip is then inverted over the well and sealed using a bead of vacuum grease. This is then left undisturbed for at least 24 hours to equilibrate. At the start of the experiment, the precipitant concentration in the drop is half that of the well. Equilibration then takes place via the vapour phase. Given the relatively large volume of the well, its concentration effectively remains the same. The drop however loses water vapour to the well until the precipitant concentration equals that of the well. Hopefully, if the conditions have been favourable, at some point during this process the protein has become supersaturated and been driven out of solution in the form of crystals. All too often however these trials result in precipitate or the formation of salt crystals, or nothing happens at all and the drops remain clear. I would estimate that the success rate at this stage is less than 0.1%.

If no promising leads are found then there are several possible courses of action. We can add various things to the sample which may affect crystallization. We can work at a different temperature, temperature can have a profound affect on protein solubility. Temperatures of 4° C and 18° C are typically used. If we have already been round this cycle more than once, it may be time to go back to the purification and expression and try something different, such as working with a fragment of our target protein.

If however we are lucky enough to get one or more "hits" in the screens, then we do follow-up experiments which will be variations on a theme where the theme is the successful set of conditions. Essentially we need to refine all variables and possibly introduce some new ones in order to achieve our goal, which is large, single crystals (see below). Things to try at this stage include varying the concentrations of all components in the crystallization, slight pH changes, using additives, switching to similar buffers or precipitants, or even using different crystallization methods (e.g. dialysis). Occasionally good crystals will form overnight, but more typically they will take from several days to several weeks to grow.

3.13. X-ray diffraction

As mentioned above, X-rays are electromagnetic waves of the same nature as visible light or radio waves, the only difference being the very short wavelength of around 1 Å (Ångström, which is 10-10 meter). For comparison, the wavelength of visible light is approximately between 400 and 700 nm (one nm is 10 Å). X-rays may be generated using various laboratory sources or at synchrotrons, where very high intensity and highly focused X-rays can be generated. To obtain X-ray data from a crystal, it needs to be placed in a monochromatic (single wavelength) X-ray beam.

Subsequently, it is repeatedly exposed to the X-ray beam, while changing its orientation (usually rotating). Each exposure provides an image, similar to that shown above. Each spot on the image is a diffracted X-ray beam, which emerged from the crystal and was registered by the X-ray detector. Thousands of diffraction spots need to be collected to solve a protein structure. Depending on the type of the crystal (cell dimensions and symmetry), different strategies for data collection are followed and a different amount of data is collected. Usually the crystal is rotated in the X-ray beam one degree a time, and exposed to X-rays for a short period (seconds to minutes, depending on the intensity of the X-ray source). The intensities of these spots are subsequently used to calculate the electron density of the molecules within the crystal. The electron density, in turn, will tell us where the atoms are located, information which can be used to build a model of the molecule or molecules in the crystal.

Using crystallographic terminology, this process is called **X-ray data collection**. When the X-rays hit the crystal, a phenomenon called **X-ray diffraction** takes place. Diffraction is a common physical phenomenon and occurs when a wave (of any nature) encounters an obstacle, which can be any material object. This results in bending of the wave around that object, also called scattering of waves. Another way for diffraction to occur is when a wave encounters a small opening, a small hole or a slit. This causes spreading of the wave in all directions. In practice, in both cases, the obstacle and the hole/slit start to act as a new wave source, sending around waves with slightly different direction of propagation, as compared to the original wave. The "new" scattered waves interact with each other, resulting in another physical phenomena called interference, which translated to normal language simply means addition of waves.

X-ray diffraction is caused by the interaction of electromagnetic waves with the matter inside the crystals, and particularly with the electrons. These waves get scattered by the electrons, or each electron becomes a small X-ray source of its own. Scattered waves from all the electrons within each atom are added to each other, giving diffracted waves from each atom, etc. When the scattered waves are added, they may either get stronger or cancel each other. Those which get stronger are registered by the X-ray detector, as in the figure above. Interestingly, we do not necessarily need X-rays to observe interference, we can, for example go to a lake nearby, through two stones into the water and then observe how the waves from the two stones either reinforce each other or become weaker.

3.14. Nuclear Magnetic Resonance Spectroscopy

NMR spectroscopy allows structure determination in solution under conditions that approximate the physiological environment of a protein. It is based on the observation of physical phenomena exhibited when nuclei absorb energy from a radio frequency source at certain characteristic frequencies in the presence of strong external magnetic fields. The position of the nuclei in the molecule effects the electronic environment of the nucleus and thus affects the absorption frequency. The frequency differences observed in the resultant spectrum can be used to determine the molecular structure of the sample. NMR has low sensitivity and the data obtained is noisy. It is used for smaller proteins.

NMR spectroscopy is one of only two techniques that can provide detailed structural information about macromolecules at atomic resolution. This detailed view of molecular structure results from a laborious examination of a number of conformationally sensitive parameters and the application of distance geometry programs to provide high-resolution structures of peptides and proteins to about 30,000 MW. However, unlike the organic chemist, who has long characterized small molecules by applying empirical "rules" associating the chemical shift with structure and conformation, the use of chemical shift as a tool to understand biological conformation has not been widely employed. The major difficulty has been a poor understanding of the link between chemical shifts and structural parameters. While the theoretical difficulties remain largely unsolved, there now exists a large body of NMR chemical shift data for peptides and small proteins that can be used to develop empirical relationships. In an early statistical study, Szilagyi and Jardetzky identified a significant correlation between aH chemical shifts and helical and b-sheet structures. In the absence of other effects, helical conformations produce up-field shifts while b-structures shift the a proton downfield. A smoothed plot of 1 Ha chemical shifts as a function of sequence can be used to readily identify.

Regions of secondary structure. A closely related method assigns a chemical shift index (CSI) to each residue in a protein, by comparison with a table of chemical shifts corresponding to random structure. Regions where the CSI is clustered with negative values are assigned as a-helical; those with positive values are assigned to b-structure. In contrast to optical methods for determination of protein structure, NMR provides information on the location of secondary structural elements within the protein sequence. Oldfield has even suggested that suf- ficiently accurate data may be used to predict the three-dimensional structure of the protein from chemical-shift data alone. In order to apply chemicalshift information to predictions of secondary structure, the chemical shifts must be assigned to particular residues in the protein. This is a tedious task that requires the measurement and analysis of 2-D and often 3-D spectra. In addition, the limit of about 30K for a protein that produces sufficiently narrow lines seriously hampers the general application of the method. However, the chemical-shift index serves as a useful check on further model refinement in high-resolution NMR studies of small proteins.

3.15. Ultraviolet-Visible (UV-Vis) absorption spectroscopy, Circular Dichroism (CD), Fourier Transform Infrared (FTIR) and fluorescence spectroscopy. These techniques can be utilised to study the structure of proteins as structural changes can have a major impact on their activity, stability and toxicity, and consequently can compromise the efficacy and shelf life of products.

UV-Vis can be used as a sensitive measure of subtle changes in protein structure and also to determine the protein concentration and purity of a protein solution.

CD in the far UV region (180–260nm) provides information regarding different forms of regular secondary structure found in proteins whereas the near UV region (240-360nm) can provide a detailed fingerprint of the tertiary structure. It can provide information about interaction between ligands or cofactors for e.g. DNA-protein interaction. CD can also be very useful in the comparison of batches of pharmaceuticals and we can provide some additional help with the analysis using objective pattern recognition techniques. We also provide advice and consultancy on obtaining a good quality CD spectral measurement

.

FTIR facilitates the structural analysis of proteins in different chemical environments, which makes it a valuable tool for the biotechnology and pharmaceutical industry. This technique can be utilised to analyse the structure of protein therapeutics at higher concentrations, than CD. The facilities available in house - ATR accessories specially designed for protein solutions and powders - enable the analysis of proteins in formulation buffer as well as in powder form.

Fluorescence spectroscopy can also provide tertiary structural information. Changes in the local environment of tryptophan residues can be followed by changes in the emission spectra.

Proteins are highly diversified class of biomolecules Differences in their chemical properties, such as charge shape, size and solubility, enable them to perform many biological functions. These functions include – enzyme catalysts, metabolic regulation, binding and transport of small molecules, gene regulation, immunological defense and cell structure.

The cellular activities and functions involve one or more proteins. Their central place in the cell is reflected in the fact that genetic information is ultimately expressed as proteins,

The basic building blocks of proteins are amino acids. There are about 20 amino acids found in proteins, all of which share certain structural features. These features are:

Carboxyl (acid) (-COOH) group

An amino (basic) (-NH) group

They differ from each other with respect to their side chains. Amino acids of proteins are linked together by peptide bonds between their carboxyl and –amino group to form linear polymers. Proteins have 3 or 4 levels of structural organization. and complexity. The primary structure of a protein is the sequence of amino acids in its polypeptide chain or chains. Secondary structure is formed and stabilized by the interaction of amino acids that are fairly close to one another on the polypeptide chain. The polypeptide with its primary and secondary structure can be coiled or organized along three axes to form a more complex, three dimensional shape. Thus, level of organization is the tertiary structure

A number of colorimetric and photometric methods are used for the determination of proteins. Photocolorimetric methods are based on the so called “colour” reactions for functional group of protein molecules. Among these are reactions for peptide groups and folin’s test for amino acid aromatic radicals (tyrosin and trypthophan). The biuret test is more specific since peptide bond occurs only in proteins and peptides. It is widely used in clinico-biochemical examination. The Lowry’s method, based on folin’s reaction is highly sensitive but of low specificity, since free aromatic amino acids and numerous materials containing a phenolic group produce a similar colouration. Photonephelometric methods for protein 2 concentration determination are based on the estimation of the degree of turbidity (or clouding) of a protein suspension in solution. These methods have not gained wide acceptance in practice.

Spectrophotometric methods are sub-divided into direct and indirect methods. The latter method represents a sensitive and accurate variant of the photocolorimetrically techniques. After the induction of the colour reaction of a protein, the coloured solution is measured spectrophotometrically and the protein. Concentration is estimated by the percentage of monochromatic light energy absorbed by the colour solution.

The direct method is based on the measure of light absorption by protein solution in the ultra violet spectra region at 200-220nm (characteristic absorption due to aromatic amino acid radicals, chiefly tryphotophan and tyrosine). These methods are easy to handle and require no preliminary colouration of the solution to be induced by a chromogenic agent. The 200-220nm spectrophotomery is more specific than that at 230nm. Since in the latter case, the additional absorption due to various low molecular aromatic compound, which are found in biological materials that interferes with the measurement accuracy.

The local dye "*Uri isi*" which was purchased at Nsukka market is used locally for dying grey hair. It is believed to undergo some reactions with certain chemical components of the hair in the presence of hydrogen peroxide. When applied on the hair in the presence of hydrogen peroxide, the grey colour of the hair is changed to dark colour. Preliminary screening showed that the dye reacts with proteins to produce a change in colour. The present study attempts to design a new colorimetric method for estimation of proteins based on the colour reaction between local dye "*uri isi*" and proteins.

QUESTIONS TO PRACTICE:

1. **Discuss the Steps of Protein engineering**
2. **Define Inteins and its applications**
3. **Describe the Protein splicing mechanism**
4. **Enumerate on the methods involved in production of novel proteins**
5. **Give a detailed explanation of methods for expressing recombinant proteins**
6. **Give an account on the characterization of protein structure**

UNIT – IV - ENZYME AND PROTEIN ENGINEERING – SBTA5202

4.1. Incorporation of Noncanonical Amino Acids into Engineered Proteins

There are two generic (and complementary) strategies for metabolic incorporation of noncanonical amino acids into proteins – the so-called residue-specific and site-specific methods. The residue-specific approach involves replacement of all (or a fraction) of one of the natural amino acid residues. This method has its origins in the work of Cohen and coworkers, who showed in the 1950s that near-quantitative replacement of methionine by selenomethionine could be accomplished in bacterial cells. This observation has had revolutionary consequences for protein science and engineering, in that it provides the basis of the multiwavelength anomalous diffraction method for crystallographic structure determination.

The site-specific approach allows replacement of a single amino acid residue by a noncanonical analog. In this approach, a heterologous transfer RNA(tRNA)/aminoacyl-tRNA

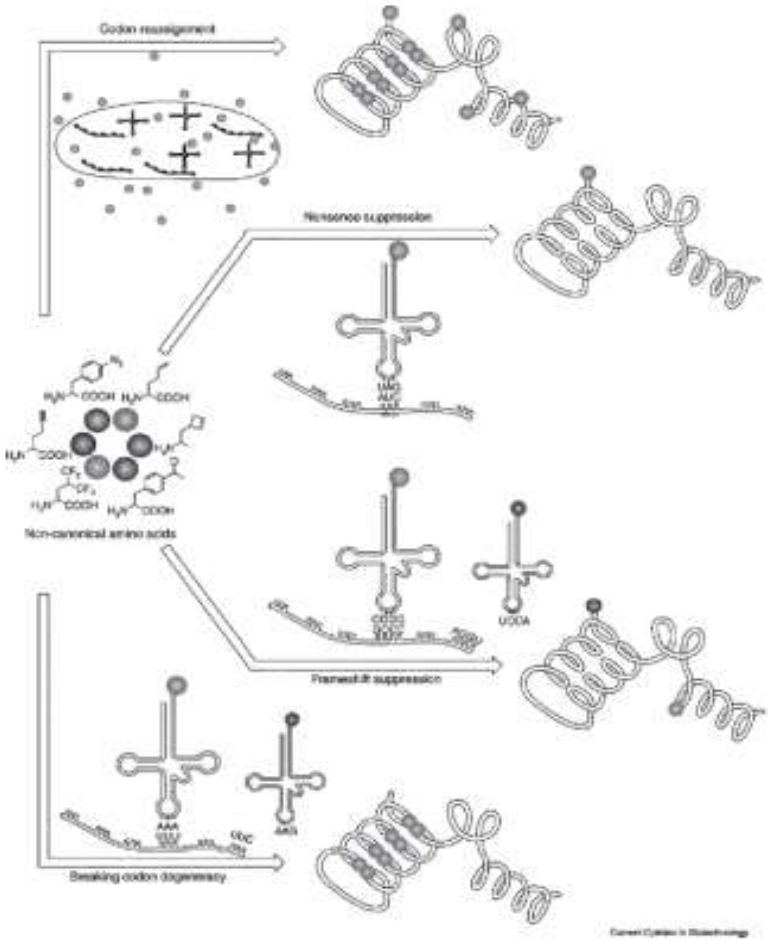


Fig. 1 Methods for incorporation of noncanonical amino acids. Residue-specific incorporation by sense codon reassignment enables replacement of all, or a fraction, of the corresponding canonical residues. Nonsense suppression, frameshift suppression, and breaking codon degeneracy can all be used to place noncanonical amino acids at specific sites. (Reprinted from Link et al. 2003; with permission from Elsevier)

synthetase pair is used to deliver the analog in response to a nonsense or four-base codon. In 1996, Drabkin and coworkers used an Escherichia coli tRNA/glutaminyl-tRNA synthetase pair for amber codon suppression in mammalian cells, and showed that the suppressor tRNA was not charged by any of the mammalian aminoacyl-tRNA synthetases. Shortly thereafter, Furter (1998) introduced a yeast tRNA/phenylalanyl-tRNA synthetase (PheRS) Noncanonical Amino Acids in Protein Science

and Engineering 129 pair into E. coli for site-specific incorporation of the noncanonical amino acid p-fluorophenylalanine. Since then, amber codon suppression has become the most common method for site-specific incorporation of noncanonical amino acids in vivo. Schultz and coworkers have been especially successful in producing orthogonal suppressor tRNA/aminoacyl-tRNA synthetase pairs for incorporation of chemically, structurally, and spectroscopically diverse amino acid analogs. Site-specific incorporation has also been accomplished in Xenopus oocytes using microinjected messenger RNAs and chemically misacylated amber suppressor tRNAs.

Sisido have pioneered the use of four-base codons (frameshift suppression) for site-specific introduction of noncanonical amino acids into proteins, and have employed this strategy to label streptavidin with fluorophores for fluorescence resonance energy transfer (FRET) experiments. Much of the work reported to date with four-base codons involves in vitro translation, but design of appropriate orthogonal tRNA/aminoacyl-tRNA synthetase pairs enables use of the method in bacterial cells. Anderson and coworkers have reported orthogonal tRNA/leucyl-tRNA synthetase (LeuRS) pairs for four-base, amber, and opal suppression. Anderson have reported use of a four-base codon with an amber codon for incorporation of two noncanonical amino acids into a recombinant protein using two orthogonal sets. An analogous five-base codon strategy has also been described.

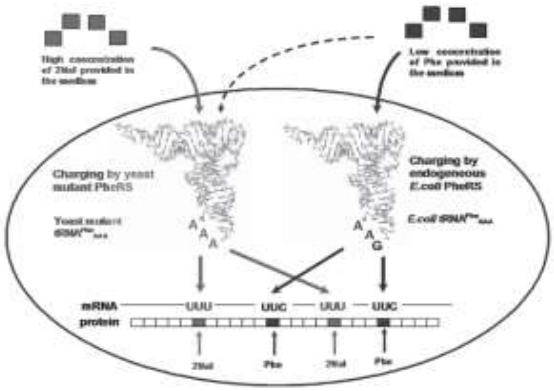


Fig. 2 Breaking the degeneracy of phenylalanine codons in *Escherichia coli*. The endogenous *E. coli* phenylalanyl-tRNA synthetase (*PheRS*) charges Phe to $tRNA^{Phe}_{GAA}$. The plasmid-borne yeast PheRS charges 2-naphthylalanine (*2Nal*) to yeast $tRNA^{Phe}_{AAA}$. UUC codons are decoded predominantly as Phe, while UUU codons are decoded predominantly as 2-naphthylalanine. *mRNA* messenger RNA, *tRNA* transfer RNA. (Reprinted with permission from Kwon et al. 2003. Copyright 2003 American Chemical Society)

Reassignment of sense codons can also be used for site-specific incorporation of noncanonical amino acids, although the fidelity of the method is lower than that of nonsense or frameshift suppression (Fig. 2). Because the 20 canonical amino acids are encoded by 61 sense codons, the genetic code is highly degenerate. For example, phenylalanine is coded by two codons, UUC and UUU. In E. coli, both codons are read by a single tRNA, which decodes UUC via Watson–Crick base-pairing and UUU through a "wobble" interaction. Reassignment of the UUU codon was achieved by introducing into an E. coli expression host a mutant yeast PheRS capable of charging 2-naphthylalanine, and a mutant yeast tRNAPhe equipped with an AAA anticodon. Expression of dehydrofolate reductase led to preferential incorporation of phenylalanine at UUC codons and of 2-naphthylalanine at UUU codons. The generality and quantitative specificity of this method have not yet been established.

4.1.1. Translational Fidelity

Aminoacyl-tRNA Synthetases Translational fidelity is controlled in large measure by the aminoacyl-tRNA synthetases, which match the 20 canonical amino acids with their cognate tRNAs. The remarkable capacity of the synthetases to discriminate among the natural amino acids might lead one to expect noncanonical substrates to be excluded by the translational apparatus (for more details see the chapter by Mascarenhas et al., this volume). In fact, many noncanonical amino acids are activated by the wild-type synthetases at rates that support efficient protein synthesis in bacterial cells. For analogs that are activated more slowly, addition of plasmid-encoded copies of the cognate

synthetase can restore the rate of protein synthesis to levels characteristic of overexpressed recombinant proteins, and synthetase engineering has enabled further expansion of the set of useful amino acids. Szostak and coworkers have described a screen for identifying noncanonical amino acid substrates that are susceptible to enzymatic aminoacylation. Using the screen, they identified 59 previously unknown amino acid substrates.

4.2. Choice of protein scaffold for protein engineering

When engineering a new functionality in a protein, many aspects must be taken into account. First, it is necessary to know as much as possible about the starting structure in order to assess its potential. It is important to consider whether the protein will work in a particular selection or screening system, whether it will tolerate the changes introduced, and whether its production is simple and scalable enough for future applications. These are just a few examples of the assets to be considered in a structural framework.

The features we are seeking in a structural framework fall into two categories: experimental demands, and application demands. The fi rst is associated with the method used in the engineering experiment, while the latter depends on the intended use of the product.

A number of questions concerning the experimental procedure:

- Does an adequate assay, selection or screening system exist, or should the framework provide a means for testing the newly established functionality? If for example you are seeking a protein where signal change can be measured as a function of binding, certain scaffolds such as periplasmic binding proteins will facilitate this more easily than others
- Does the applied methodology have requirements for the framework? For example, small single - chain proteins are preferable for the application of phage and ribosome display and for the construction of fusion proteins. Likewise, cysteine - free scaffolds are useful when unique cysteines should be introduced to which effector compounds can be coupled. Further, a robust scaffold with high thermodynamic stability is preferable because it can compensate for any destabilizing effects of newly introduced functional residues, an effect often observed in rational design approaches.
- The expression of functional molecules is also important in selection and screening systems; for example, poorly or insolubly expressed proteins will not be able to complement a missing functionality and thus unstable variants are often eliminated in the process. A number of questions, concerning applicability:
- Under what conditions should the fi nal product be active, should it be especially stable or degradable, and does it need to be localized specifi cally?
- Is large - scale production feasible, what are the protein yields, and is there an easy purifi cation?
- High thermodynamic stability, reversible folding, and high expression levels are what you will be looking for. The absence of disulfi de bonds or free cysteines is also advantageous because it allows the expression of functional molecules in the reducing environment of the bacterial cytoplasm, which usually produces higher yields than periplasmic or eukaryotic expression or refolding in vitro.

Proteins can be optimized to improve chemical robustness, thermodynamic stability or recombinant expression yields before using them as a framework in an engineering experiment. However, if considered well, the choice of a framework may also relieve the need to engineer many of these properties, so that attention can be focused on the property in question.

Apart from practical considerations, the choice of the structural framework can also be important for the new functionality that is introduced. Using a partial binding pocket and adjusting it to fi t a new ligand may be easier to achieve than introducing a new one from scratch. Obviously, studying the structure of the framework is essential in rational design approaches, but it can also be advantageous in directed evolution experiments. A detailed knowledge of the protein structure can reveal important parts that are better left untouched and help focus on the variable regions that can be subjected to randomization. To a certain degree, sequence alignments will also provide this type of information. Highly conserved residues are often important for folding or stability of the protein, while variable regions are free to evolve.

4.3. Applications of Molecular Modelling and Structure predictions to Protein engineering

Structure Predictions:

Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence — that is, the prediction of its folding and its secondary, tertiary, and quaternary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Every two years, the performance of current methods is assessed in the CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction). A continuous evaluation of protein structure prediction web servers is performed by the community project CAMEO3D.

Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the local secondary structures of proteins based only on knowledge of their amino acid sequence only. For proteins, a prediction consists of assigning regions of the amino acid sequence as likely alpha helices, beta strands (often noted as "extended" conformations), or turns. The success of a prediction is determined by comparing it to the results of the DSSP algorithm (or similar e.g. STRIDE) applied to the crystal structure of the protein. Specialized algorithms have been developed for the detection of specific well-defined patterns such as transmembrane helices and coiled coils in proteins.[1]

The best modern methods of secondary structure prediction in proteins reach about 80% accuracy;[3] this high accuracy allows the use of the predictions as feature improving fold recognition and ab initio protein structure prediction, classification of structural motifs, and refinement of sequence alignments. The accuracy of current protein secondary structure prediction methods is assessed in weekly benchmarks such as LiveBench and EVA.

Background

Early methods of secondary structure prediction, introduced in the 1960s and early 1970s,[4][5][6][7][8] focused on identifying likely alpha helices and were based mainly on helix-coil transition models.[9] Significantly more accurate predictions that included beta sheets were introduced in the 1970s and relied on statistical assessments based on probability parameters derived from known solved structures. These methods, applied to a single sequence, are typically at most about 60-65% accurate, and often underpredict beta sheets.[1] The evolutionary conservation of secondary structures can be exploited by simultaneously assessing many homologous sequences in a multiple sequence alignment, by calculating the net secondary structure propensity of an aligned column of amino acids. In concert with larger databases of known protein structures and modern machine learning methods such as neural nets and support vector machines, these methods can achieve up 80% overall accuracy in globular proteins.[10] The theoretical upper limit of accuracy

is around 90%,[10] partly due to idiosyncrasies in DSSP assignment near the ends of secondary structures, where local conformations vary under native conditions but may be forced to assume a single conformation in crystals due to packing constraints. Limitations are also imposed by secondary structure prediction's inability to account fortertiary structure; for example, a sequence predicted as a likely helix may still be able to adopt a beta-strand conformation if it is located within a beta-sheet region of the protein and its side chains pack well with their neighbors. Dramatic conformational changes related to the protein's function or environment can also alter local secondary structure.

Historical perspective

To date, over 20 different secondary structure prediction methods have been developed. One of the first algorithms was Chou-Fasman method, which relies predominantly on probability parameters determined from relative frequencies of each amino acid's appearance in each type of secondary structure.[11] The original Chou-Fasman parameters, determined from the small sample of structures solved in the mid-1970s, produce poor results compared to modern methods, though the parameterization has been updated since it was first published. The Chou-Fasman method is roughly 50-60% accurate in predicting secondary structures.

The next notable program was the GOR method, named for the three scientists who developed it — *G*arnier, *O*sguthorpe, and *R*obson, is an information theory-based method. It uses the more powerful probabilistic technique of Bayesian inference.[12] The GOR method takes into account not only the probability of each amino acid having a particular secondary structure, but also the conditional probability of the amino acid assuming each structure given the contributions of its neighbors (it does not assume that the neighbors have that same structure). The approach is both more sensitive and more accurate than that of Chou and Fasman because amino acid structural propensities are only strong for a small number of amino acids such as proline and glycine. Weak contributions from each of many neighbors can add up to strong effects overall. The original GOR method was roughly 65% accurate and is dramatically more successful in predicting alpha helices than beta sheets, which it frequently mispredicted as loops or disorganized regions.[1]

Another big step forward, was using machine learning methods. First artificial neural networks methods were used. As a training sets they use solved structures to identify common sequence motifs associated with particular arrangements of secondary structures. These methods are over 70% accurate in their predictions, although beta strands are still often underpredicted due to the lack of three-dimensional structural information that would allow assessment of hydrogen bonding patterns that can promote formation of the extended conformation required for the presence of a complete beta sheet.[1] PSIPRED and JPRED are some of the most known programs based on neural networks for protein secondary structure prediction. Next, support vector machines have proven particularly useful for predicting the locations of turns, which are difficult to identify with statistical methods

Extensions of machine learning techniques attempt to predict more fine-grained local properties of proteins, such as backbone dihedral angles in unassigned regions. Both SVMs[15] and neural networks[16] have been applied to this problem.[13] More recently, real-value torsion angles can be accurately predicted by SPINE-X and successfully employed for ab initio structure prediction.[17]

Other improvements

It is reported that in addition to the protein sequence, secondary structure formation depends on other factors. For example, it is reported that secondary structure tendencies depend also on local environment,[18] solvent accessibility of residues,[19] protein structural class,[20] and even the organism from which the proteins are obtained.[21] Based on such observations, some studies have

shown that secondary structure prediction can be improved by addition of information about protein structural class,[22] residue accessible surface area[23][24] and also contact number information.[25]

Tertiary structure

The practical role of protein structure prediction is now more important than ever. Massive amounts of protein sequence data are produced by modern large-scale DNAsequencing efforts such as the Human Genome Project. Despite community-wide efforts in structural genomics, the output of experimentally determined protein structures—typically by time-consuming and relatively expensive X-ray crystallography or NMR spectroscopy—is lagging far behind the output of protein sequences.

The protein structure prediction remains an extremely difficult and unresolved undertaking. The two main problems are calculation of protein free energy and finding the global minimum of this energy. A protein structure prediction method must explore the space of possible protein structures which is astronomically large. These problems can be partially bypassed in "comparative" or homology modeling and fold recognition methods, in which the search space is pruned by the assumption that the protein in question adopts a structure that is close to the experimentally determined structure of another homologous protein. On the other hand, the *de novo* or ab initio protein structure predictionmethods must explicitly resolve these problems. The progress and challenges in protein structure prediction has been reviewed in Zhang 2008.[26]

4.4. Ab initio protein modelling

Energy- and fragment-based methods

Ab initio- or *de novo*- protein modelling methods seek to build three-dimensional protein models "from scratch", i.e., based on physical principles rather than (directly) on previously solved structures. There are many possible procedures that either attempt to mimic protein folding or apply some stochastic method to search possible solutions (i.e.,global optimization of a suitable energy function). These procedures tend to require vast computational resources, and have thus only been carried out for tiny proteins. To predict protein structure *de novo* for larger proteins will require better algorithms and larger computational resources like those afforded by either powerful supercomputers (such as Blue Gene or MDGRAPE-3) or distributed computing (such as Folding@home, the Human Proteome Folding Project and Rosetta@Home). Although these computational barriers are vast, the potential benefits of structural genomics (by predicted or experimental methods) make *ab initio* structure prediction an active research field.[26]

As of 2009, a 50-residue protein could be simulated atom-by-atom on a supercomputer for 1 millisecond.[27] As of 2012, comparable stable-state sampling could be done on a standard desktop with a new graphics card and more sophisticated algorithms.[28]

Evolutionary covariation to predict 3D contacts

As sequencing became more commonplace in the 1990s several groups used protein sequence alignments to predict correlated mutations and it was hoped that these coevolved residues could be used to predict tertiary structure (using the analogy to distance constraints from experimental procedures such as NMR). The assumption is when single residue mutations are slightly deleterious,

compensatory mutations may occur to restabilize residue-residue interactions. This early work used what are known as *local* methods to calculate correlated mutations from protein sequences, but suffered from indirect false correlations which result from treating each pair of residues as independent of all other pairs.[29][30][31]

In 2011, a different, and this time *global* statistical approach, demonstrated that predicted coevolved residues were sufficient to predict the 3D fold of a protein, providing there are enough sequences available (>1,000 homologous sequences are needed).[32] The method, EVfold, uses no homology modeling, threading or 3D structure fragments and can be run on a standard personal computer even for proteins with hundreds of residues. The accuracy of the contacts predicted using this and related approaches has now been demonstrated on many known structures and contact maps,[33][34][35] including the prediction of experimentally unsolved transmembrane proteins.[36]

4.5. Comparative protein modeling

Comparative protein modelling uses previously solved structures as starting points, or templates. This is effective because it appears that although the number of actual proteins is vast, there is a limited set of tertiary structural motifs to which most proteins belong. It has been suggested that there are only around 2,000 distinct protein folds in nature, though there are many millions of different proteins.

These methods may also be split into two groups:

Homology modeling

is based on the reasonable assumption that two homologous proteins will share very similar structures. Because a protein's fold is more evolutionarily conserved than its amino acid sequence, a target sequence can be modeled with reasonable accuracy on a very distantly related template, provided that the relationship between target and template can be discerned through sequence alignment. It has been suggested that the primary bottleneck in comparative modelling arises from difficulties in alignment rather than from errors in structure prediction given a known-good alignment.[37] Unsurprisingly, homology modelling is most accurate when the target and template have similar sequences.

Protein threading

[38] scans the amino acid sequence of an unknown structure against a database of solved structures. In each case, a scoring function is used to assess the compatibility of the sequence to the structure, thus yielding possible three-dimensional models. This type of method is also known as **3D-1D fold recognition** due to its compatibility analysis between three-dimensional structures and linear protein sequences. This method has also given rise to methods performing an **inverse folding search** by evaluating the compatibility of a given structure with a large database of sequences, thus predicting which sequences have the potential to produce a given fold.

Side-chain geometry prediction

Accurate packing of the amino acid side chains represents a separate problem in protein structure prediction. Methods that specifically address the problem of predicting side-chain geometry include dead-end elimination and the self-consistent mean field methods. The side chain conformations with low energy are usually determined on the rigid polypeptide backbone and using a set of discrete side chain conformations known as "rotamers." The methods attempt to identify the set of rotamers that minimize the model's overall energy.

These methods use rotamer libraries, which are collections of favorable conformations for each residue type in proteins. Rotamer libraries may contain information about the conformation, its frequency, and the standard deviations about mean dihedral angles, which can be used in sampling.[39] Rotamer libraries are derived from structural bioinformatics or other statistical analysis of side-chain conformations in known experimental structures of proteins, such as by clustering the observed conformations for tetrahedral carbons near the staggered (60°, 180°, -60°) values.

Rotamer libraries can be backbone-independent, secondary-structure-dependent, or backbone-dependent. Backbone-independent rotamer libraries make no reference to backbone conformation, and are calculated from all available side chains of a certain type (for instance, the first example of a rotamer library, done by Ponder and Richards at Yale in 1987).[40] Secondary-structure-dependent libraries present different dihedral angles and/or rotamer frequencies for α-helix, β-sheet, or coil secondary structures.[41]Backbone-dependent rotamer libraries present conformations and/or frequencies dependent on the local backbone conformation as defined by the backbone dihedral angles $\phi$and ψ, regardless of secondary structure.[42]

The modern versions of these libraries as used in most software are presented as multidimensional distributions of probability or frequency, where the peaks correspond to the dihedral-angle conformations considered as individual rotamers in the lists. Some versions are based on very carefully curated data and are used primarily for structure validation,[43] while others emphasize relative frequencies in much larger data sets and are the form used primarily for structure prediction, such as the Dunbrack rotamer libraries.[44]

Side-chain packing methods are most useful for analyzing the protein's hydrophobic core, where side chains are more closely packed; they have more difficulty addressing the looser constraints and higher flexibility of surface residues, which often occupy multiple rotamer conformations rather than just one.[45][46]

Prediction of structural classes

Statistical methods have been developed for predicting structural classes of proteins based on their amino acid composition,[47] pseudo amino acid composition[48][49][50][51] and functional domain composition.[52]

Quaternary structure

In the case of complexes of two or more proteins, where the structures of the proteins are known or can be predicted with high accuracy, protein–protein docking methods can be used to predict the structure of the complex. Information of the effect of mutations at specific sites on the affinity of the complex helps to understand the complex structure and to guide docking methods.

4.6. Molecular modelling:

Molecular modelling encompasses all theoretical methods and computational techniques used to model or mimic the behaviour of molecules. The techniques are used in the fields of computational chemistry, drug design, computational biology and materials science for studying molecular systems ranging from small chemical systems to large biological molecules and material assemblies. The simplest calculations can be performed by hand, but inevitably computers are

required to perform molecular modelling of any reasonably sized system. The common feature of molecular modelling techniques is the atomistic level description of the molecular systems. This may include treating atoms as the smallest individual unit (the Molecular mechanics approach), or explicitly modeling electrons of each atom (the quantum chemistry approach).

Molecular mechanics

Molecular mechanics is one aspect of molecular modelling, as it refers to the use of classical mechanics/Newtonian mechanics to describe the physical basis behind the models. Molecular models typically describe atoms (nucleus and electrons collectively) as point charges with an associated mass. The interactions between neighbouring atoms are described by spring-like interactions (representing chemical bonds) and van der Waals forces. The Lennard-Jones potential is commonly used to describe van der Waals forces. The electrostatic interactions are computed based on Coulomb's law. Atoms are assigned coordinates in Cartesian space or in internal coordinates, and can also be assigned velocities in dynamical simulations. The atomic velocities are related to the temperature of the system, a macroscopic quantity. The collective mathematical expression is known as a potential function and is related to the system internal energy (U), a thermodynamic quantity equal to the sum of potential and kinetic energies. Methods which minimize the potential energy are known as energy minimization techniques (e.g., steepest descent and conjugate gradient), while methods that model the behaviour of the system with propagation of time are known as molecular dynamics.

$$E = E_{\text{bonds}} + E_{\text{angle}} + E_{\text{dihedral}} + E_{\text{non-bonded}}$$
$$E_{\text{non-bonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

This function, referred to as a potential function, computes the molecular potential energy as a sum of energy terms that describe the deviation of bond lengths, bond angles and torsion angles away from equilibrium values, plus terms for non-bonded pairs of atoms describing van der Waals and electrostatic interactions. The set of parameters consisting of equilibrium bond lengths, bond angles, partial charge values, force constants and van der Waals parameters are collectively known as a force field. Different implementations of molecular mechanics use different mathematical expressions and different parameters for the potential function. The common force fields in use today have been developed by using high level quantum calculations and/or fitting to experimental data. The technique known as energy minimization is used to find positions of zero gradient for all atoms, in other words, a local energy minimum. Lower energy states are more stable and are commonly investigated because of their role in chemical and biological processes. A molecular dynamics simulation, on the other hand, computes the behaviour of a system as a function of time. It involves solving Newton's laws of motion, principally the second law, $\mathbf{F} = m\mathbf{a}$. Integration of Newton's laws of motion, using different integration algorithms, leads to atomic trajectories in space and time. The force on an atom is defined as the negative gradient of the potential energy function. The energy minimization technique is useful for obtaining a static picture for comparing between states of similar systems, while molecular dynamics provides information about the dynamic processes with the intrinsic inclusion of temperature effects.

Variables

Molecules can be modelled either in vacuum or in the presence of a solvent such as water. Simulations of systems in vacuum are referred to as *gas-phase* simulations, while those that include the presence of solvent molecules are referred to as *explicit solvent* simulations. In another type of

simulation, the effect of solvent is estimated using an empirical mathematical expression; these are known as *implicit solvation* simulations.

Applications

Molecular modelling methods are now routinely used to investigate the structure, dynamics, surface properties and thermodynamics of inorganic, biological and polymeric systems. The types of biological activity that have been investigated using molecular modelling include protein folding, enzyme catalysis, protein stability, conformational changes associated with biomolecular function, and molecular recognition of proteins, DNA, and membrane complexes.

Elementary introduction to Molecular Mechanics and Dynamics

Background

The "mechanical" molecular model was developed out of a need to describe molecular structures and properties in as practical a manner as possible. The range of applicability of molecular mechanics includes:

- Molecules containing thousands of atoms.
- Organics, oligonucleotides, peptides, and saccharides (metallo-organics and inorganics in some cases).
- Vacuum, implicit, or explicit solvent environments.
- Ground state only.
- Thermodynamic and kinetic (via molecular dynamics) properties.

The great computational speed of molecular mechanics allows for its use in procedures such as molecular dynamics, conformational energy searching, and docking. All the procedures require large numbers of energy evaluations.

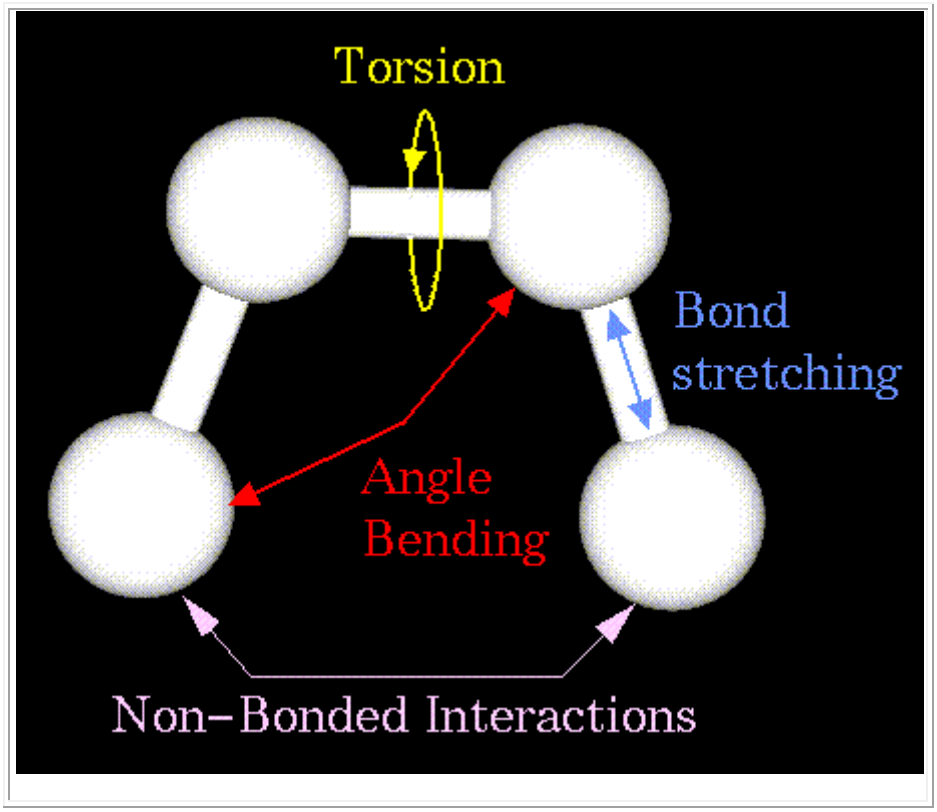
Molecular mechanics methods are based on the following principles:

- Nuclei and electrons are lumped into atom-like particles.
- Atom-like particles are spherical (radii obtained from measurements or theory) and have a net charge (obtained from theory).
- Interactions are based on springs and classical potentials.
- Interactions must be preassigned to specific sets of atoms.
- Interactions determine the **spatial distribution** of atom-like particles and their **energies**.

Note how these principles differ from those of quantum mechanics.

The Anatomy of a Molecular Mechanics Force-Field

The mechanical molecular model considers atoms as spheres and bonds as springs. The mathematics of spring deformation can be used to describe the ability of bonds to stretch, bend, and twist:



Non-bonded atoms (greater than two bonds apart) interact through van der Waals attraction, steric repulsion, and electrostatic attraction/repulsion. These properties are easiest to describe mathematically when atoms are considered as spheres of characteristic radii.

The object of molecular mechanics is to predict the energy associated with a given conformation of a molecule. However, molecular mechanics energies have no meaning as absolute quantities. Only differences in energy between two or more conformations have meaning. A simple molecular mechanics energy equation is given by:

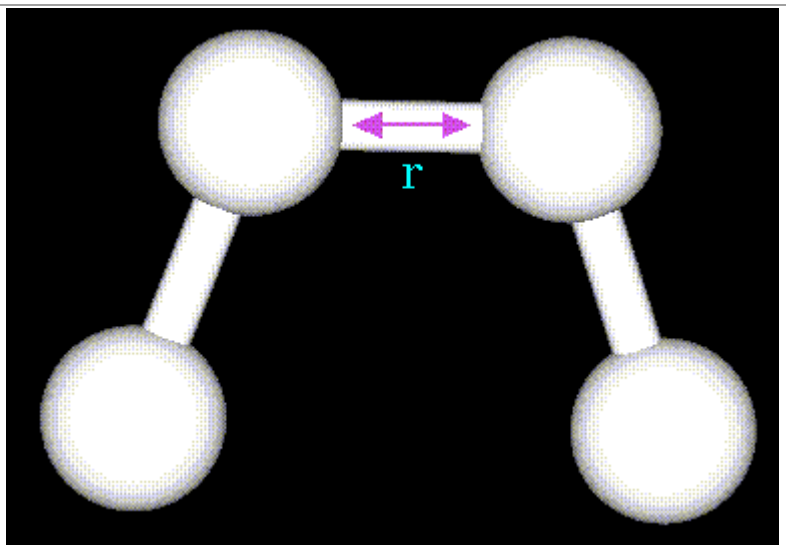
> **Energy = Stretching Energy +Bending Energy + Torsion Energy + Non-Bonded Interaction Energy**

These equations together with the data (parameters) required to describe the behavior of different kinds of atoms and bonds, is called a force-field. Many different kinds of force-fields have been developed over the years. Some include additional energy terms that describe other kinds of deformations. Some force-fields account for coupling between bending and stretching in adjacent bonds in order to improve the accuracy of the mechanical model.

The mathematical form of the energy terms varies from force-field to force-field. The more common forms will be described.

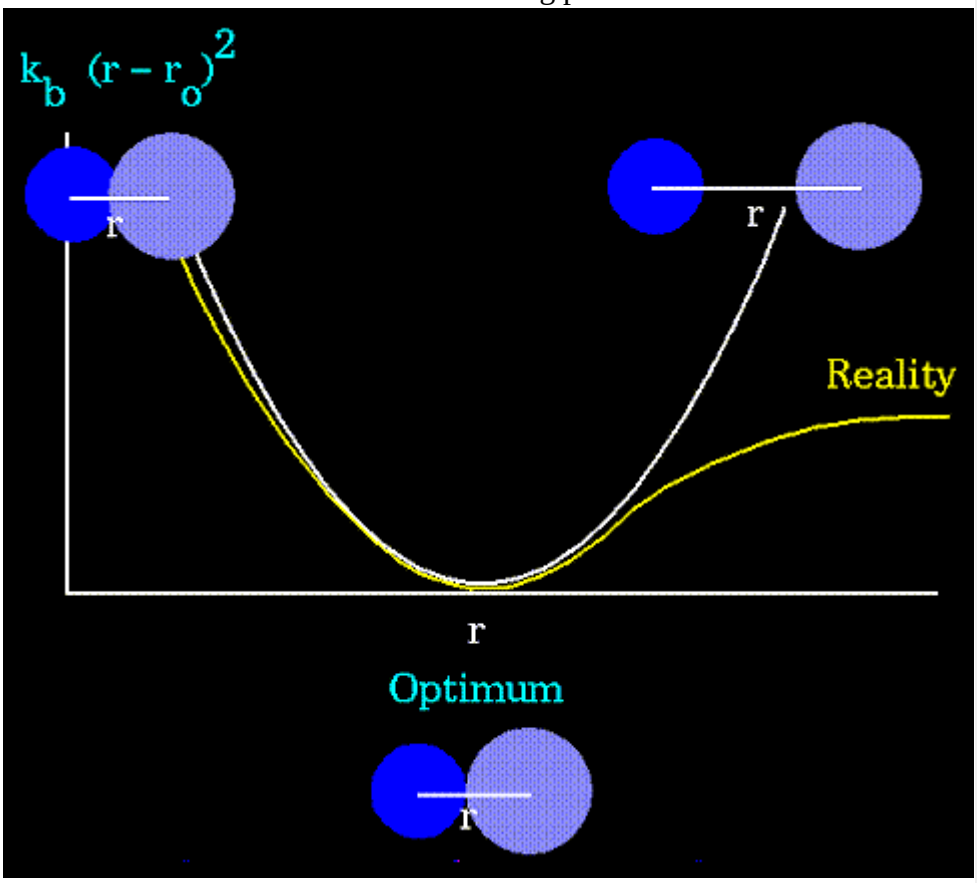
- **Stretching Energy**

$$E = \sum_{bonds} k_b \, (r - r_o)^2$$



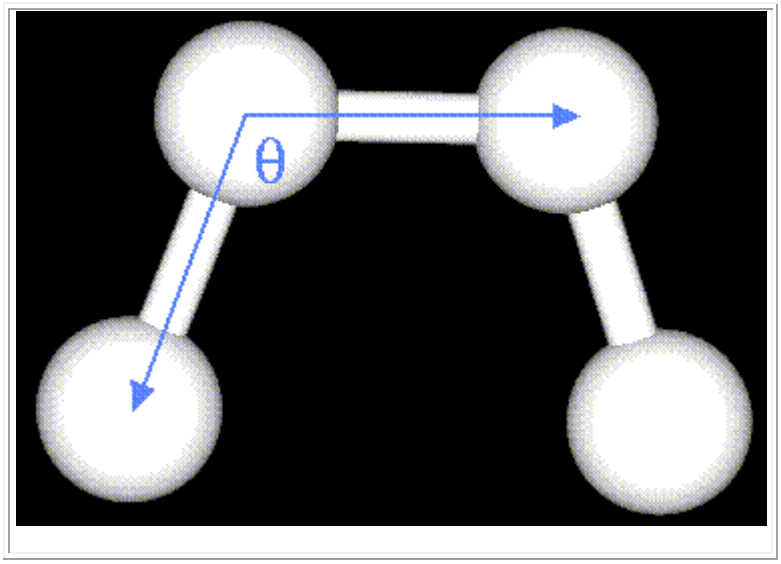
The stretching energy equation is based on Hooke's law. The "k*b*" parameter controls the stiffness of the bond spring, while "r*o*" defines its equilibrium length. Unique "k*b*" and "r*o*" parameters are assigned to each pair of bonded atoms based on their types (e.g. C-C, C-H, O-C, etc.). This equation estimates the energy associated with vibration about the equilibrium bond length. This is the equation of a parabola, as can be seen

in the following plot:

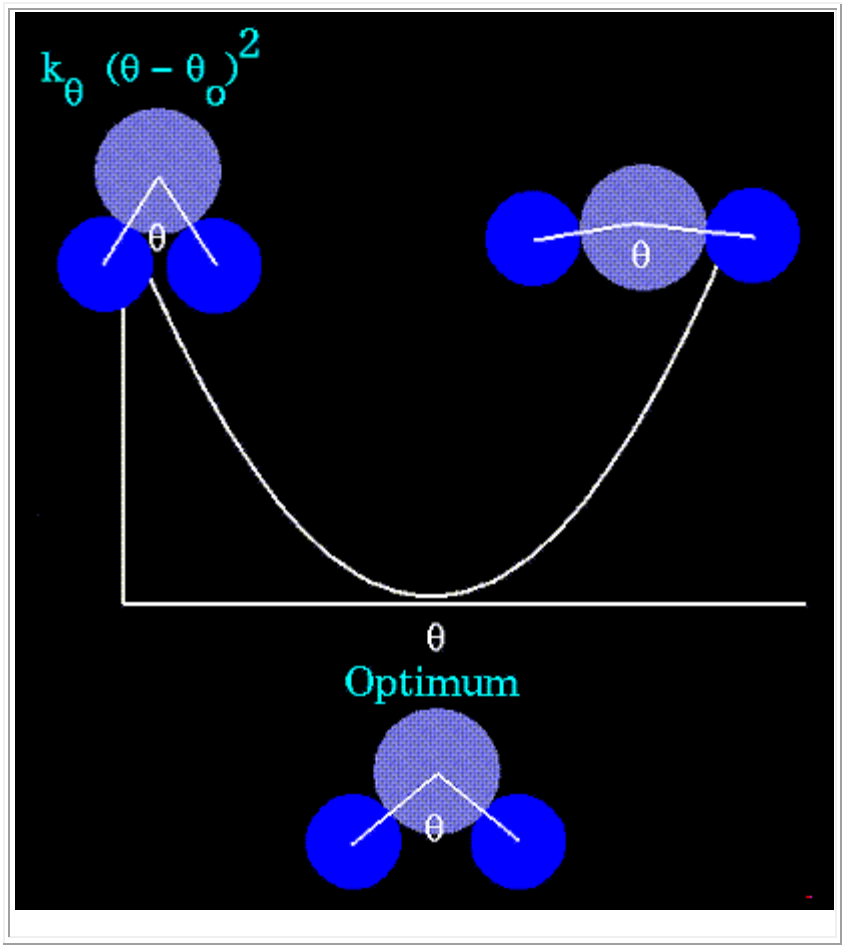


- Notice that the model tends to break down as a bond is stretched toward the point of dissociation.
- **Bending Energy**

$$E = \sum_{\text{angles}} k_\theta (\theta - \theta_o)^2$$

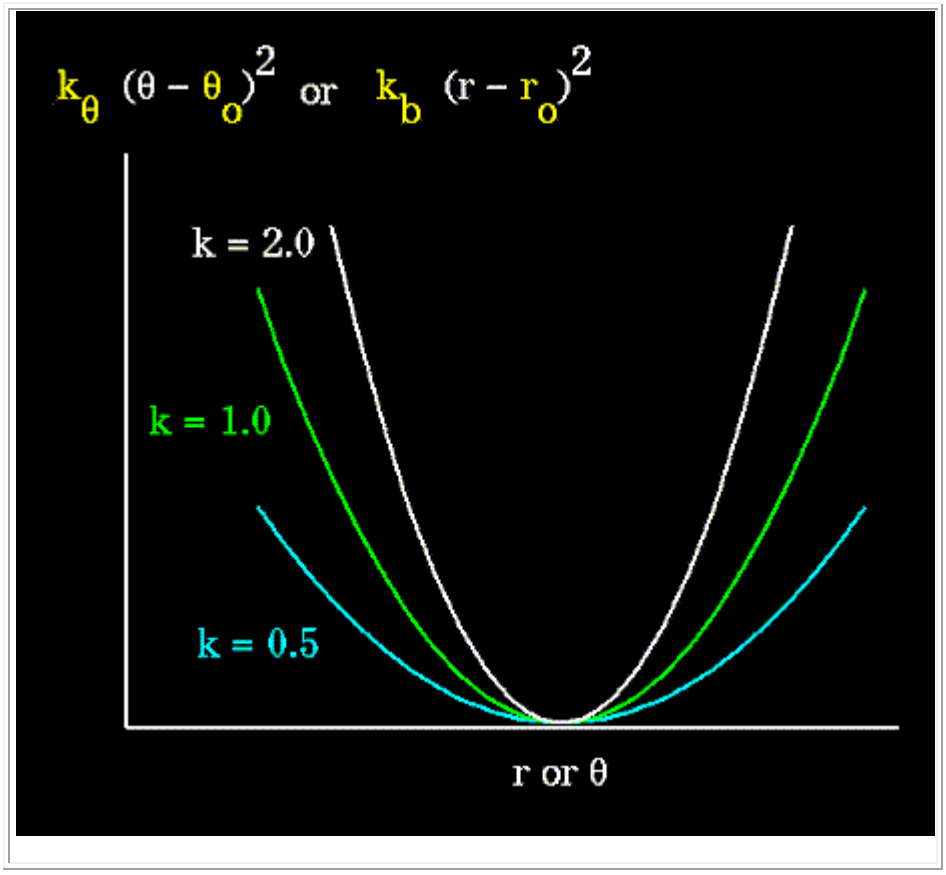


- The bending energy equation is also based on Hooke's law. The "k*theta*" parameter controls the stiffness of the angle spring, while "theta*o*" defines its equilibrium angle. This equation estimates the energy associated with vibration about the equilibrium bond angle:



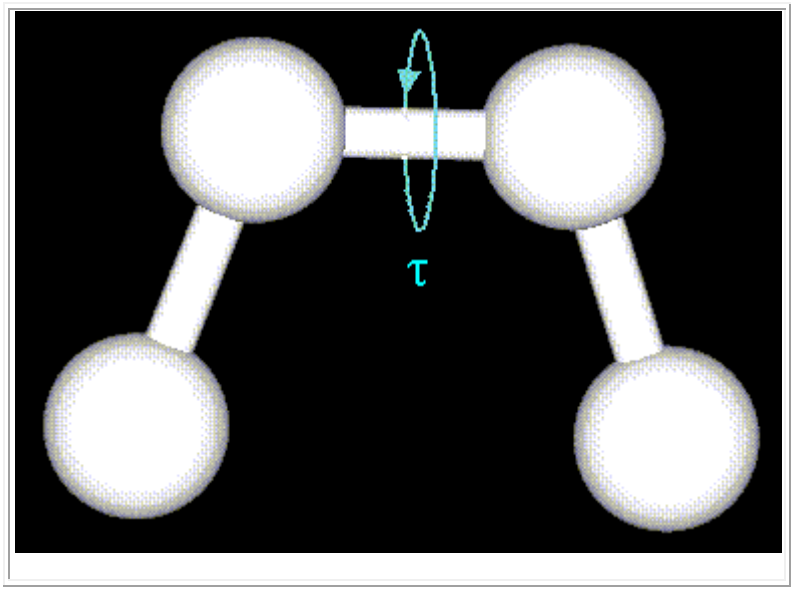
- Unique parameters for angle bending are assigned to each bonded triplet of atoms based on their types (e.g. C-C-C, C-O-C, C-C-H, etc.). The effect of the "k*b*" and "k*theta*" parameters is to broaden or steepen the slope of the parabola. The larger the value of "k", the more

energy is required to deform an angle (or bond) from its equilibrium value. Shallow potentials are achieved for "k" values between 0.0 and 1.0. The Hookeian potential is shown in the following plot for three values of "k":

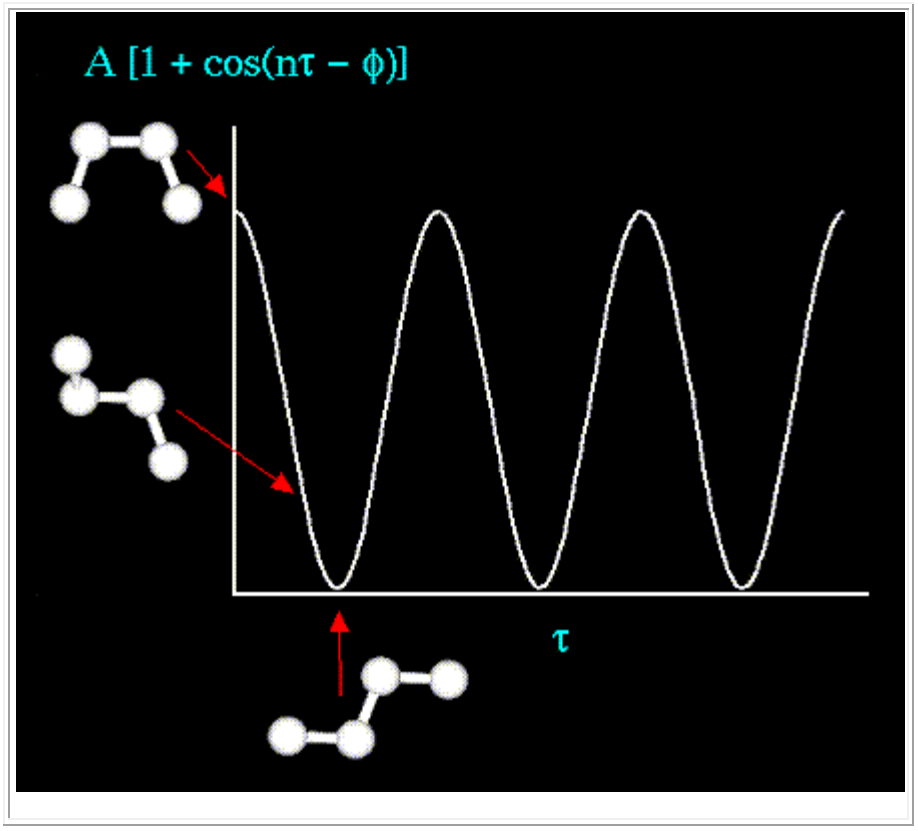


- **Torsion Energy**

$$E = \sum_{\text{torsions}} A\,[1 + \cos(n\tau - \phi)]$$

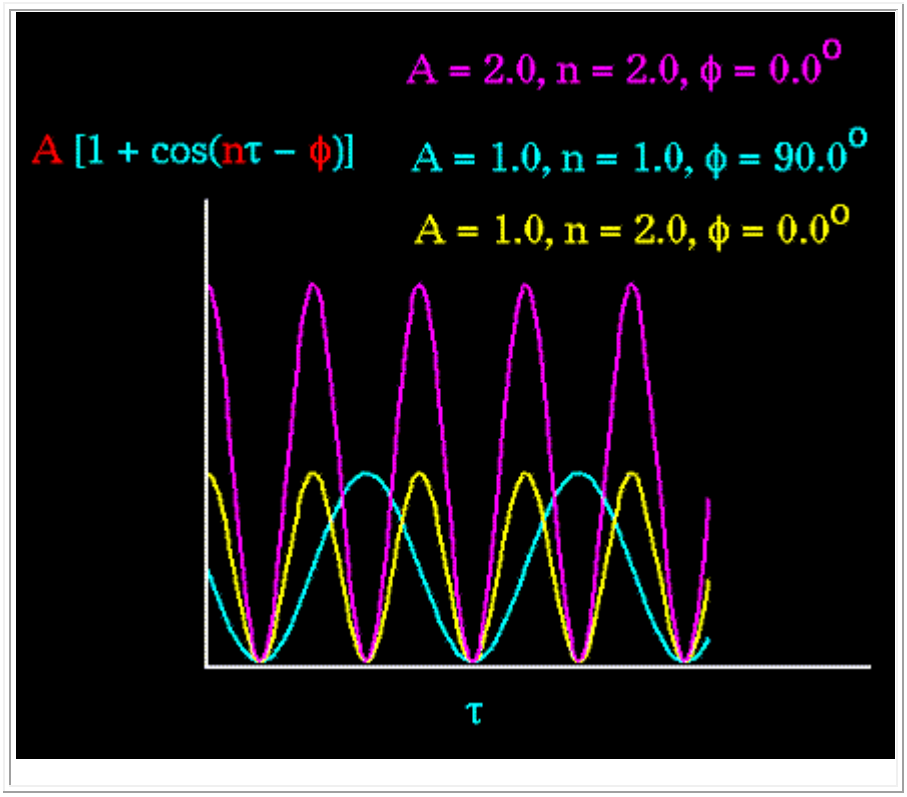


- The torsion energy is modeled by a simple periodic function, as can be seen in the following plot:



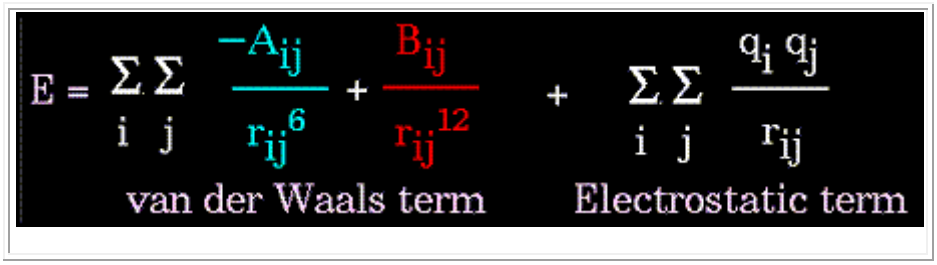
- The torsion energy in molecular mechanics is primarily used to correct the remaining energy terms rather than to represent a physical process. The torsional energy represents the amount of energy that must be added to or subtracted from the Stretching Energy + Bending Energy + Non-Bonded Interaction Energy terms to make the total energy agree with experiment or rigorous quantum mechanical calculation for a model dihedral angle (ethane, for example might be used a a model for any H-C-C-H bond).

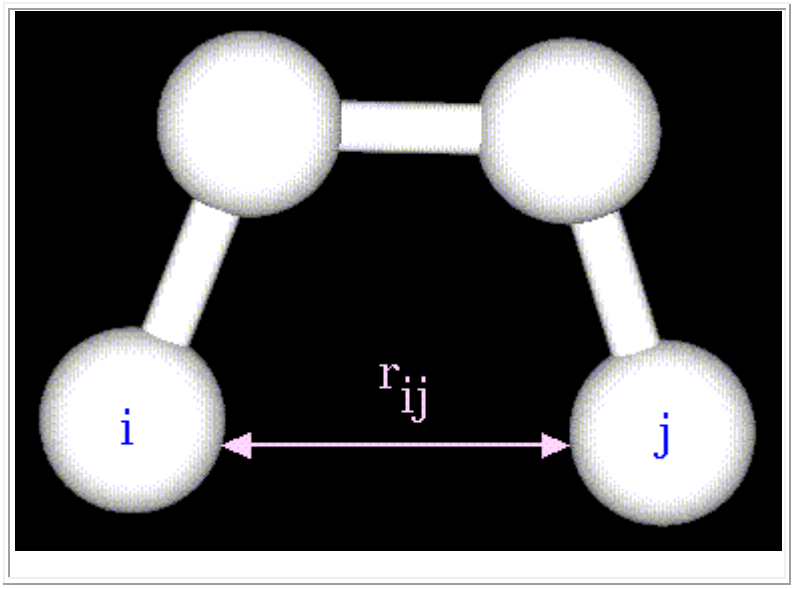
- The "A" parameter controls the amplitude of the curve, the n parameter controls its periodicity, and "phi" shifts the entire curve along the rotation angle axis (tau). The parameters are determined from curve fitting. Unique parameters for torsional rotation are assigned to each bonded quartet of atoms based on their types (e.g. C-C-C-C, C-O-C-N, H-C-C-H, etc.). Torsion potentials with three combinations of "A", "n", and "phi" are shown in the following plot:



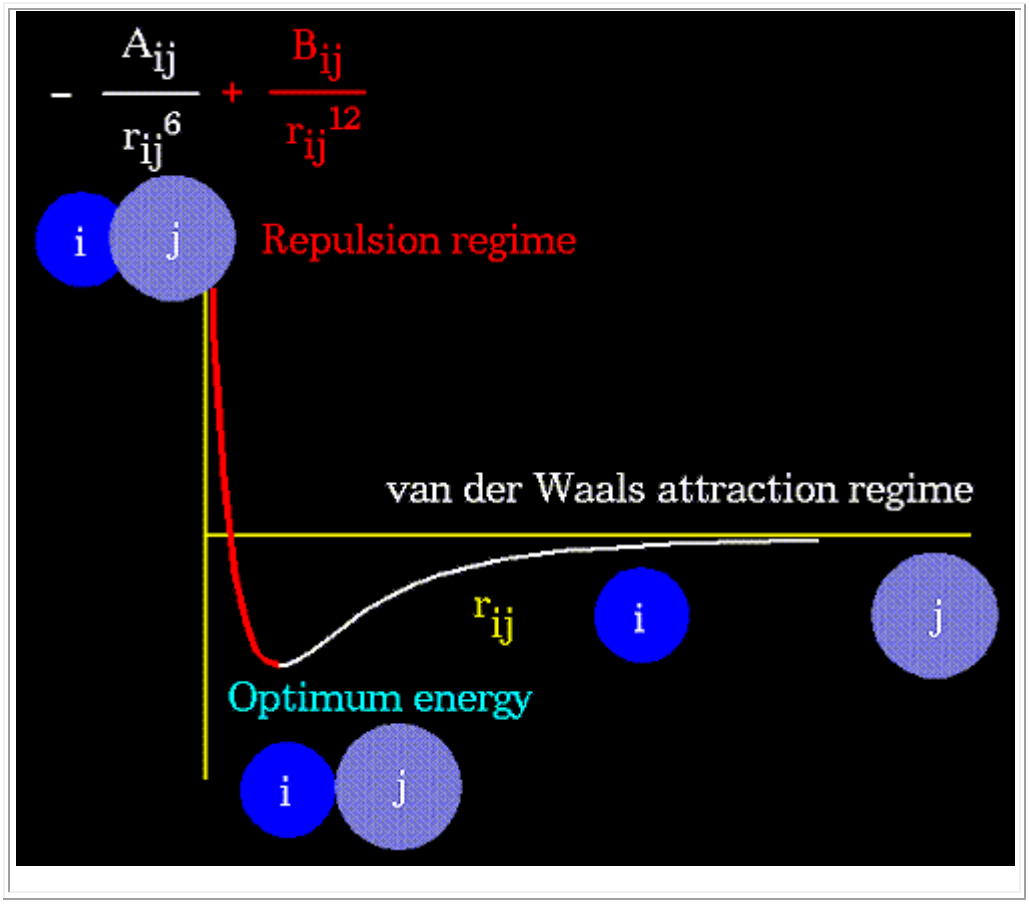
- Notice that "n" reflects the type symmetry in the dihedral angle. A CH3-CH3 bond, for example, ought to repeat its energy every 120 degrees. The *cis* conformation of a dihedral angle is assumed to be the zero torsional angle by convention. The parameter phi can be used to synchronize the torsional potential to the initial rotameric state of the molecule whose energy is being computed.
- **Non-Bonded Energy**

  The non-bonded energy represents the pair-wise sum of the energies of all possible interacting non-bonded atoms i and j:

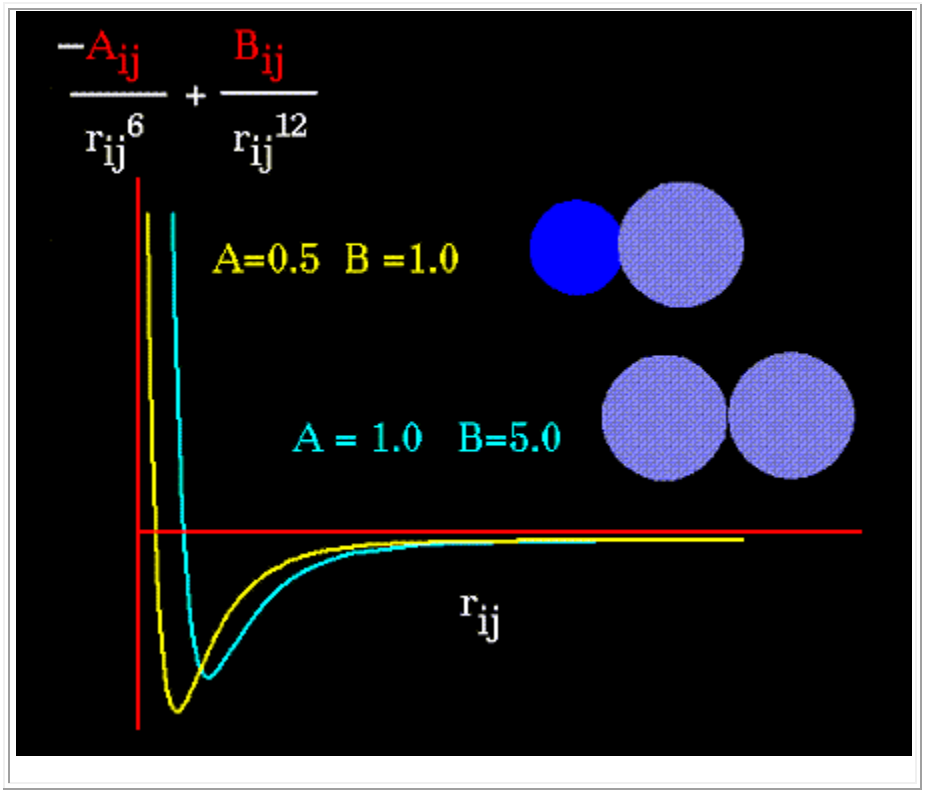




The non-bonded energy accounts for repulsion, van der Waals attraction, and electrostatic interactions. van der Waals attraction occurs at short range, and rapidly dies off as the interacting atoms move apart by a few Angstroms. Repulsion occurs when the distance between interacting atoms becomes even slightly less than the sum of their contact radii. Repulsion is modeled by an equation that is designed to rapidly blow up at close distances (1/r^12 dependency). The energy term that describes attraction/repulsion provides for a smooth transition between these two regimes. These effects are often modeled using a 6-12 equation, as shown in the following plot:



The "A" and "B" parameters control the depth and position (interatomic distance) of the potential energy well for a given pair of non-bonded interacting atoms (e.g. C:C, O:C, O:H, etc.). In effect, "A" determines the degree of "stickiness" of the van der Waals attraction and "B" determines the degree of "hardness" of the atoms (e.g marshmallow-like, billiard ball-like, etc.).



The "A" parameter can be obtained from atomic polarizability measurements, or it can be calculated quantum mechanically. The "B" parameter is typically derived from crystallographic data so as to reproduce observed average contact distances between different kinds of atoms in crystals of various molecules.

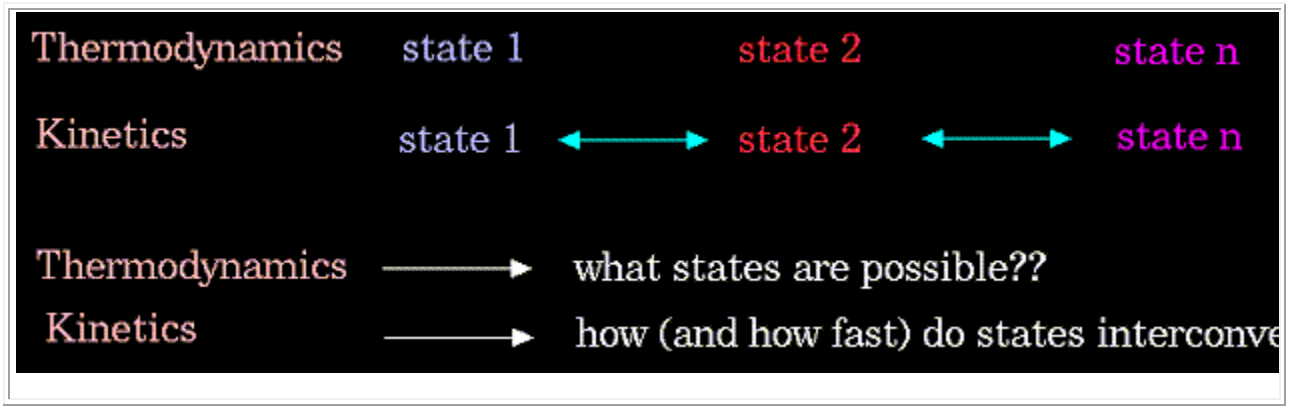
The electrostatic contribution is modeled using a Coulombic potential. The electrostatic energy is a function of the charge on the non-bonded atoms, their interatomic distance, and a molecular dielectric expression that accounts for the attenuation of electrostatic interaction by the environment (e.g. solvent or the molecule itself). Often, the molecular dielectric is set to a constant value between 1.0 and 5.0. A linearly varying distance-dependent dielectric (i.e. 1/r) is sometimes used to account for the increase in environmental bulk as the separation distance between interacting atoms increases.

Partial atomic charges can be calculated for small molecules using an *ab initio* or semiempirical quantum technique (usually MOPAC or AMPAC). Some programs assign charges using rules or templates, especially for macromolecules. In some force-fields, the torsional potential is calibrated to a particular charge calculation method (rarely made known to the user). Use of a different method can invalidate the force-field consistency.

Molecular Dynamics

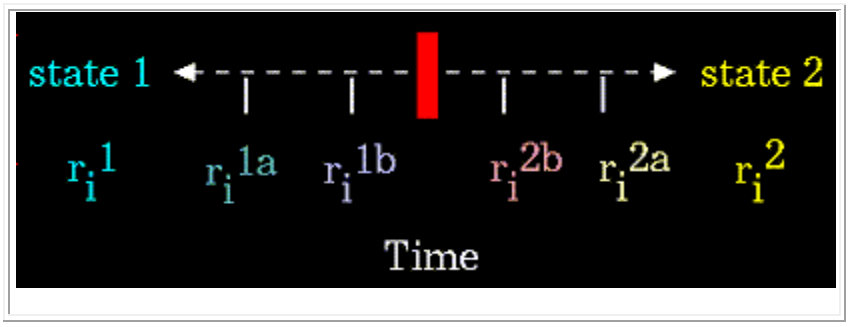
In the broadest sense, molecular dynamics is concerned with molecular motion. Motion is inherent to all chemical processes. Simple vibrations, like bond stretching and angle bending, give rise to IR spectra. Chemical reactions, hormone-receptor binding, and other complex processes are associated with many kinds of intra- and intermolecular motions.

The **driving force** for chemical processes is described by **thermodynamics**. The **mechanism** by which chemical processes occur is described by **kinetics**. Thermodynamics dictates the energetic relationships between different chemical states, whereas the sequence or rate of events that occur as molecules transform between their various possible states is described by kinetics:



Conformational transitions and local vibrations are the usual subjects of molecular dynamics studies. Molecular dynamics alters the intramolecular degrees of freedom in a step-wise fashion, analogous to energy minimization. The individual steps in energy minimization are merely directed at establishing a down-hill direction to a minimum. The steps in molecular dynamics, on the other hand, meaningfully represent the changes in atomic position, r*i*, over time (i.e. velocity).

For the "i" atoms of the system:



Newton's equation is used in the molecular dynamics formalism to simulate atomic motion:

$$\text{force} = \text{mass} \times \text{acceleration} \quad (F_i = m_i a_i)$$

The rate and direction of motion (velocity) are governed by the forces that the atoms of the system exert on each other as described by Newton's equation. In practice, the atoms are assigned initial velocities that conform to the total kinetic energy of the system, which in turn, is dictated by the desired simulation temperature. This is carried out by slowly "heating" the system (initially at absolute zero) and then allowing the energy to equilibrate among the constituent atoms. The basic ingredients of molecular dynamics are the calculation of the force on each atom, and from that information, the position of each atom throughout a specified period of time (typically on the order of picoseconds = 10^-12 seconds).

The force on an atom can be calculated from the change in energy between its current position and its position a small distance away. This can be recognized as the derivative of the energy with respect to the change in the atom's position:

$$-\frac{dE}{dr_i} = F_i$$

Energies can be calculated using either molecular mechanics or quantum mechanics methods. Molecular mechanics energies are limited to applications that do not involve drastic changes in electronic structure such as bond making/breaking. Quantum mechanical energies can be used to study dynamic processes involving chemical changes. The latter technique is extremely novel, and of limited availability (Gaussian03 is an example of such a program).

Knowledge of the atomic forces and masses can then be used to solve for the positions of each atom along a series of extremely small time steps (on the order of femtoseconds = 10^-15 seconds). The resulting series of snapshots of structural changes over time is called a trajectory. The use of this method to compute trajectories can be more easily seen when Newton's equation is expressed in the following form:

$$-\frac{dE}{dr_i} = m_i \frac{d^2 r_i}{dt^2}$$

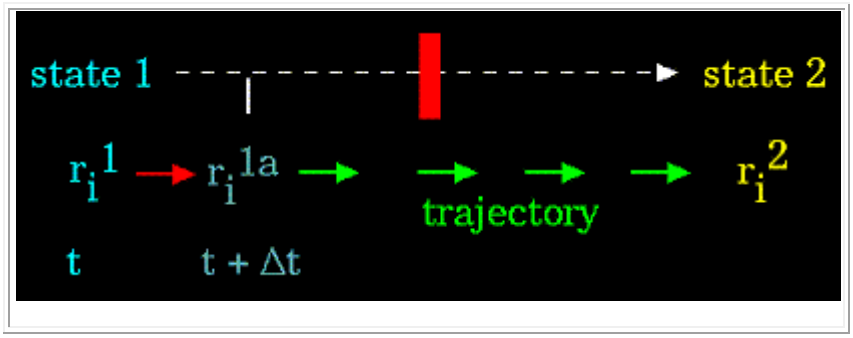
In practice, trajectories are not directly obtained from Newton's equation due to lack of an analytical solution. First, the atomic accelerations are computed from the forces and masses. The velocities are next calculated from the accelerations based on the following relationship:

$$a_i = \frac{dv_i}{dt}$$

Lastly, the positions are calculated from the velocities:

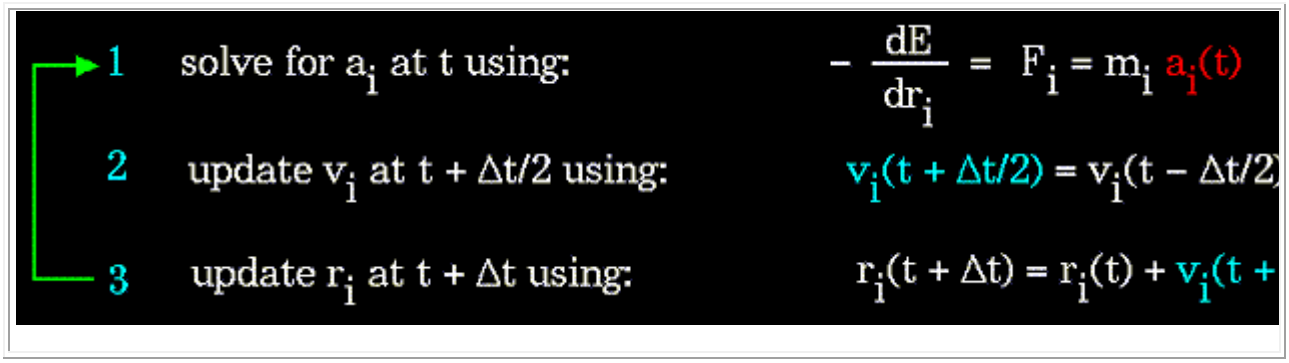
$$v_i = \frac{dr_i}{dt}$$

A trajectory between two states can be subdivided into a series of sub-states separated by a small time step, "delta t" (e.g. 1 femtosecond):



The initial atomic positions at time "t" are used to predict the atomic positions at time "t + delta t". The positions at "t + delta t" are used to predict the positions at "t + 2*delta t", and so on.

The "leapfrog" method is a common numerical approach to calculating trajectories based on Newton's equation. The steps can be summarized as follows:



The method derives its name from the fact that the velocity and position information successively alternate at 1/2 time step intervals.

Molecular dynamics has no defined point of termination other than the amount of time that can be practically covered. Unfortunately, the current picosecond order of magnitude limit is often not long enough to follow many kinds of state to state transformations, such as large conformational transitions in proteins.

Molecular dynamics calculations can be performed using both HyperChem and Gaussian programs.

Quantum mechanics:

Definition of Computational Chemistry

- Computational Chemistry: Use mathematical approximations and computer programs to obtain results relative to chemical problems.
- Computational *Quantum* Chemistry: Focuses specifically on equations and approximations derived from the postulates of quantum mechanics. Solve the Schrödinger equation for molecular systems.
- *Ab Initio* Quantum Chemistry: Uses methods that do not include any empirical parameters or experimental data.

What's it Good For?

- Computational chemistry is a rapidly growing field in chemistry.
  - Computers are getting faster.
  - Algorithims and programs are maturing.
- Some of the almost limitless properties that can be calculated with computational chemistry are:
  - Equilibrium and transition-state structures
  - dipole and quadrapole moments and polarizabilities
  - Vibrational frequencies, IR and Raman Spectra
  - NMR spectra
  - Electronic excitations and UV spectra
  - Reaction rates and cross sections
  - thermochemical data

Motivation

- Schrödinger Equation can only be solved exactly for simple systems.
  - Rigid Rotor, Harmonic Oscillator, Particle in a Box, Hydrogen Atom
- For more complex systems (i.e. many electron atoms/molecules) we need to make some simplifying assumptions/approximations and solve it numerically.
- However, it is still possible to get very accurate results (and also get very crummy results).
  - In general, the "cost" of the calculation increases with the accuracy of the calculation and the size of the system.

Getting into the theory...

- Three parts to solving the Schrödinger equation for molecules:
  - Born-Oppenheimer Approximation
    - Leads to the idea of a potential energy surface
  - The expansion of the many-electron wave function in terms of Slater determinants.
    - Often called the "Method"
  - Representation of Slater determinants by molecular orbitals, which are linear combinations of atomic-like-orbital functions.
    - The basis set

The Born-Oppenheimer Approximation

- Now we can solve the electronic part of the Schrödinger equation separately.

$$\hat{H}_{el}\psi_{el}(r;R) = E_{el}\psi_{el}(r;R)$$

$$\hat{H}_{el} = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_\alpha \sum_i \frac{Z_\alpha e'^2}{r_{i\alpha}} + \sum_j \sum_{i>j} \frac{e'^2}{r_{ij}}$$

- BO approximation leads to the idea of a potential energy surface.

$$U(R) = E_{el} + V_{NN}$$

$$V_{NN} = \sum_\alpha \sum_{\alpha>\beta} \frac{Z_\alpha Z_\beta e'^2}{r_{\alpha\beta}}$$

Nuclear Schrödinger Equation

Once we have the Potential Energy Surface (PES) we can solve the nuclear Schrödinger equation.

- Solution of the nuclear SE allow us to determine a large variety of molecular properties.

  An example are vibrational energy levels.

4.7. Energy Minimization

The potential energy calculated by summing the energies of various interactions is a numerical value for a single conformation. This number can be used to evaluate a particular conformation, but it may not be a useful measure of a conformation because it can be dominated by a few bad interactions. For instance, a large molecule with an excellent conformation for nearly all atoms can have a large overall energy because of a single bad interaction, for instance two atoms too near each other in space and having a huge van der Waals repulsion energy. It is often preferable to carry out energy minimization on a conformation to find the best nearby conformation. Energy minimization is usually performed by gradient optimization: atoms are moved so as to reduce the net forces on them. The minimized structure has small forces on each atom and therefore serves as an excellent starting point for molecular dynamics simulations.

Energy minimization is usually performed in Cartesian coordinates, by optimizing along pathways in -dimensional space, where is the number of particles. This pathway can be the gradient, , where

In other words, each Cartesian component, , of the gradient equals the derivative of the potential energy with respect to that component. Only those interactions involving particle contribute to the gradients of the Cartesian coordinates of ( ). The components of constitute a path, **P**, in -dimensional space. Finding the minimum along this pathway typically involves an interpolation of two points in -space to find a new point where . Usually, however, at the new point, so a new path is chosen and minimization proceeds. It is possible to set at each new point, but it is more efficient to choose the new pathway to be orthogonal to all previous paths. This method of ``conjugate gradients'' is perhaps the most popular method of energy minimization. Details of this method can be found in Reference [16].

It is also possible to minimize the energy of a conformation by optimizing the dihedral angle degrees of freedom, rather than the Cartesian coordinates. The minimization occurs in -dimensional space, where is the number of dihedral angles. Torques, or derivatives of the forcefield with respect to dihedral angles, take the place of the gradient. We have found that ``torque minimization,'' when followed by Cartesian minimization, produces an overall lower-energy conformation than Cartesian minimization alone. Neither method, however, can guarantee that the lowest possible conformation (the global minimum) will be reached. The process of moving along pathways in conformational space usually ends at a ``local minimum'' - a well in the potential energy surface, where the energy is lower than for all other nearby conformations, but not necessarily lower than other local minima.

4.8. De Novo Protein Design

In rational protein design proteins can be redesigned from the sequence and structure of a known protein, or completely from scratch in *de novo* protein design. In protein redesign, most of the residues in the sequence are maintained as their wild-type amino-acid while a few are allowed to mutate. In *de novo* design the entire sequence is designed anew, based on no previous sequence.

Both *de novo* designs and protein redesigns can establish rules on the sequence space: the specific amino acids that are allowed at each mutable residue position. For example, the composition of the surface of the RSC3 probe to select HIV-broadly neutralizing antibodies was restricted based on evolutionary data and charge balancing. In fact, many of the earliest attempts on protein design were heavily based on empirical "rules" on the sequence space. Moreover, the design of fibrous proteins, usually follows strict rules on the sequence space. Collagen-based designed proteins, for example, are often composed of Gly-Pro-X repeating patterns. With the advent of computational techniques, however, the design of proteins with no human intervention in sequence selection has become possible.

4.9. PROTEIN DATABASES

Protein databases are more specialized than primary sequence databases.

They contain information derived from the primary sequence databases.

Some contain protein translations of the nucleic acid sequences.

Some contain sets of patterns and motifs derived from sequence homologs.

GenBank - the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.

PIR Protein Information Resource -a comprehensive, non-redundant, expertly annotated, fully classified and extensively cross-referenced protein sequence database.

SWISS-PROT & TrEMBL - SWISS-PROT is a curated protein sequence database. is a computer-annotated supplement of SWISS-PROT that contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT.

TIGR - a collection of curated databases containing DNA and protein sequence, gene expression, cellular **role, protein family, and taxonomic data for microbes, plants and humans.**

MOTIF, PATTERN & PROFILE DATABASES

ALIGN - a compendium of sequence alignments: it is a companion resource to *PRINTS*.

BLOCKS - multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.

DOMO - a database of homologous protein domain families.

HOMSTRAD - a curated database of structure-based alignments for homologous protein families.

InterPro- Integrated Resource of Protein Domains and Functional Sites - InterPro is an integrated documentation resource for protein families, domains and sites, developed initially as a means of rationalising the complementary efforts of the PROSITE, PRINTS, Pfam and ProDom database projects. Each combined InterPro entry includes functional descriptions and

literature references, and links are made back to the relevant member database(s), allowing users to see at a glance whether a particular family or domain has associated patterns, profiles, fingerprints, etc. Merged and individual entries (i.e., those that have no counterpart in the companion resources) are assigned unique accession numbers. Each InterPro entry lists all the matches against SWISS-PROT and TrEMBL (more than 1,000,000 hits in total).InterPro aims to reduce duplication of effort in the labour-intensive, rate-limiting process of annotation, and will facilitate communication between the disparate resources. By uniting these databases, we capitalise on their individual strengths, producing a single entity that is far greater than the sum of its parts.

PFam - a database of multiple alignments of protein domains or conserved protein regions. The alignments represent some evolutionary conserved structure which has implications for the protein's function. Profile hidden Markov models (profile HMMs) built from the Pfam alignments can be very useful for automatically recognizing that a new protein belongs to an existing protein family, even if the homology is weak.

PRINTS ñ Protein Fingerprint Database - a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.

PRINTS-S ñ relational cousin of the PRINTS Database

ProDom - an automatic compilation of homologous domains.ProDom families were generated automatically using PSI-BLAST with a profile built from the seed alignments of Pfam-A 4.3 families.

ProSite - is a database of protein families and domains

consisting of biologically significant sites, patterns and profiles.

Protein Profiles - online cross-references to the Oxford University Press Protein Profiles project.

ProtoMap - site offers an exhaustive classification of all the proteins in the SWISSPROT and TrEMBL databases, into groups of related proteins.The resulting classification splits the protein space into well defined groups of proteins, most of them are closely correlated with natural biological families and superfamiliesfor comprehensive evaluation results). The hierarchical organization may help to detect finer subfamilies that make up known families of proteins as well as interesting relations between protein families.

SBASE - protein domain library sequences that contains 237.937 annotated structural, functional, ligand-binding and topogenic segments of proteins, cross-referenced to all major sequence databases and sequence pattern collections.

SYSTERS - SYSTERS cluster set contains sequences from SWISS-PROT , TrEMBL, PIR, Wormpep, and MIPS Yeast protein translations which are sorted into disjoint clusters. fragmental sequences build single sequence clusters, while the remaining sequences are contained in clusters of non-redundant sequences per cluster.

PROTEIN STRUCTURE DATABASES

CATH Protein Structure Classification ñ a hierarchical domain classification of protein structures in the Brookhaven protein databank.

FSSP Fold Classification based on Structure-Structure Alignment of Proteins - based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB).

Library of Protein Family Cores - structural alignments of protein families and computed average core structures for each family.Useful for building models, threading, and exploratory analysis.

ModBase a database of three-dimensional protein models calculated by comparative modeling**.**

PRESAGE - a database of proteins, each of which has a collection of annotations reflecting current experimental status, structural assignments models, and suggestions.

RCSB Protein Data Bank - single international repository for the processing and distribution of 3-D macromolecular structure data primarily determined experimentally.

Protein Loop Classification - Conformational clusters and consensus sequences for protein loops derived by computational analysis of their structures.

SCOP ñ Structural Classification of Proteins - a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known.

Sloop Database ñ Sloop Database of Super Secondary Fragments - a classification of protein loops.

3 Dee ñ Database of Protein Domain Definitions - contains structural domain definitions for all protein chains in the Protein Databank (PDB)that have 20 or more residues and are not theoretical models.

GENOMES

DEAMBULUM ñ contains the GENOMES: Viruses, Archaea,Bacteria, Fungi, Plants, Animals, and Man.

FlyBase - a comprehensive database for information on the genetics and molecular biology of *Drosophila*. It includes data from the Drosophila Genome Projects and data curated from the literature.

GeneCards - database of human genes, their products and their involvement in diseases.

GeneCensus Genome Comparisons

GenDis ñ Human Genetic Disease Database

Genome Database - Regions of the human genome, including genes, clones, amplimers (PCR markers), breakpoints, cytogenetic markers, fragile sites, ESTs, syndromic regions, contigs and repeats. Maps of the human genome, including cytogenetic maps, linkage maps, radiation hybrid maps, content contig maps, and integrated maps. These maps can be displayed graphically via the Web.Variations within the human genome including mutations and polymorphisms, plus allele frequency data.

KEGG: Kyoto Encyclopedia of Genes and Genomes - information pathways that consist of interacting molecules or genes and to provide links from the gene catalogs produced by genome sequencing projects.

PROTEOME ñ The BioKnowledge Library of Public Human PSD, *Caenorhabditis elegans* (WormPD), *Saccharomyces cerevisiae* (YPD) and *S. pombe* (PombePD).

Saccharomyces Genome Database - a scientific database of the molecular biology and genetics of the yeast *Saccharomyces cerevisiae*.

WhiteHead Institute for Genomic Research ñ information on the Neurospora crassa Genome Database, Human SNP Database, Human Physical Mapping Project, Mouse Genetic and Physical Mapping Project,Rat Genetic Mapping Project, Mouse RH Mapping Project, Genome Center ftp Archive (Data)

WORMBASE - a repository of mapping, sequencing and phenotypic information about the *C. elegans* nematode

TRANSCRIPTIONAL REGULATION DATABASES & ALGORITHMS

COMPEL - Database on composite regulatory elements affecting gene transcription in eukaryotes.

EDP ñ Eukaryotic Promoter Database - an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally.

RegulonDB ñ A database on transcriptional regulation in *Escherichia coli*.

TRANSFAC ñ The Transcription Factor Database

TRDD ñ Transcription Regulatory Region Database

FastM and ModelInspector A program for the generation of models for regulatory regions in DNA sequences.

FunSiteP - Recognition and classification of eukaryotic promoters.

PatSearch Search for potential transcription factor binding sites.

Promoter Inspector - Prediction of promoter regions in mammalian genomic sequences.

Mat Inspector - Search for potential transcription factor binding sites.

RSATools Regulatory Sequence Analysis Tools

S_Compsearch for NFATp/AP-1 Comp. Elements

TRADAT- TRAnscription Databases and Analysis Tools

2Zip - Computational Approaches to Identify Leucine Zippers

OTHER

BIND - full descriptions of interactions, molecular complexes and pathways.

BioMagRes Bank ñ NMR-derived protein structures.

Cytomer ñ A relational database of physiological systems, organs and cell types.

ENZYME ñ Enzyme Nomenclature Database

Enzyme Structures Database - contains the known enzyme structures that have been deposited in the Brookhaven Protein Data Bank (the PDB).

Gene Ontology Consortium ñ attempts to produce a dynamic controlled vocabulary that can be applied to all eukaryotes.

Human Transcript Database a curated source for information related to RNA molecules that have been sequenced.

LIGAND- Database for enzymes, compounds, and reactions.

Metabolic Pathways of Biochemistry - graphically represents all major metabolic pathways, primarily those important to human biochemistry.

NDB ñ Nucleic Acid Database Project - assembles and distributes structural information about nucleic acids.

PMD ñ Protein Mutant Database - covers natural as well as artificial mutants, including random and site-directed ones, for all proteins except members of the globin and immunoglobulin families.

REBase ñ Restriction Enzyme Database ñ contains detailed information about restriction enzymes, methylases, the microorganisms from which they have been isolated, recognition sequences, cleavage sites, methylation specificity, the commercial availability of the enzymes, and references.

Radar ñ Rapid Automatic Detection and Alignment of Repeats
in protein sequences.

rRNA Database ñ all about ribosomal RNA.

S/MARtDB - information about scaffold/matrix attached regions.

TargetDB -database of peptides targeting proteins to cellular locations.

Transpath ñ Signal Transduction Browser - an information system on gene-regulatory pathways. Focuses on pathways involved in the regulation of transcription factors in different species, mainly human, mouse and rat.Elements of the relevant signal transduction pathways like hormones, receptors, enzymes and transcription factors are stored together with information about their interaction and references in an object-oriented database.

TOOLS

CLUSTALW ñ Multiple sequence alignment tool

ProteinProspector - Proteomics tools for mining sequence databases in conjunction with Mass Spectrometry experiments.

ReBASE Information Tool -ReBASE query tool.

SeqHound ñ database sequence fetch program.

SignalP - predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes.

SIMILARITY, HOMOLOGY SEARCH

These algorithms are designed for the comparison of a protein sequence against sequence databases to detect similar or homologous proteins.Conserved regions usually have similar amino acid sequence and/or structural similarities.Perform at least three separate searches using different algorithms.If default settings do not detect any similar proteins, try varying the PAM matrix values.Lower matrix values are best for identifying short regions of sequence with very high similarity. Higher PAM matrices are able to detect longer, weaker matches.Simultaneously, adjust the gap penalty value around the default value.

BLAST- The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships.The BLAST search algorithm is designed to find close matches rapidly. It is faster than the S-W algorithm.

BLITZperforms a sensitive and extremely fast comparison of a protein sequence against the SWISS-PROT protein sequence database using the Smith-Waterman algorithm.The Smith-Waterman algorithm is able to detect short matching regions such as binding sites in the middle of long sequences.

Bic-sw - Smith & Waterman algorithm implementation for protein database searches

BMC Search Launcher

FASTA ñ detects patches of regional similarity rather than the best alignment between the query sequence and the database sequences. Very fast, but complete sensitivity is sacrificed.

GeneMatcher - The Smith-Waterman (S-W) search algorithm used by the FDF server is about 5% more sensitive towards divergent matches than the BLAST algorithm. This significantly increases the chances of finding distant homologs of your query sequence in the databases.FDF software incorporates a frameshift-tolerant search algorithm. This feature is particularly useful when searching for potential coding sequences in low-quality DNA sequences, such as those found in EST databases.

MPsearch - MPSRCH is a biological sequence comparison tool that implements the true Smith and Waterman algorithm. This algorithm exhaustively compares every letter in a query sequence with every letter in the database.

Paralign and SWMMX - searches a number of sequence databases for sequences similar to your amino acid query sequence using two very sensitive algorithms. You can choose between the well-known Smith-Waterman optimal local alignment algorithm or a new algorithm called ParAlign, which is much faster but still almost as sensitive.

Pfam ñ HMM Search - Unlike standard pairwise alignment methods (e.g. BLAST, FASTA), Pfam HMMs deal sensibly with multidomain proteins.

SAS ñ Sequences Annotated by Structure - will perform a FASTA search of the given sequence against the proteins of known structure in the PDB and return a multiple alignment of all hits, each annotated by structural features.

Scanps 2.3 - Fast implementation of the true Smith & Waterman algorithm for protein database searches.

MOTIF, PATTERN & PROFILE SEARCH

There are a limited number of families into which most proteins are grouped.Proteins within a given family generally have a shared function.Conserved regions are usually important for function or for maintaining a specific 3D structure. Conserved regions usually have similar amino acid sequence and/or structural similarities.Domains are distinct functional regions of a protein, often linked together by a flexible region.Motifs are recurring substructures found in many proteins.Proteins of 500 or more amino acids most likely contain discrete functional domains.Regions of low complexity often separate domains.Long stretches of repeated residues, particularly proline, glutamine, serine, or threonine, often indicate linker sequences.Approximately 2000-3000, out of a predicted 10,000-20,000, different protein families have been characterized.Roughly, half of the proteins encoded in a new genome can be placed in a known family based on their amino acid sequence.

CDD A Conserved Domain Database and Search Service

eMatrix ñ fast and accurate sequence analysis using minimal-risk scoring matrices.

eMotif Scan ñ sequence database search using eMatrix regular expressions.

eMotif Search ñ protein classification search.

InterProScan ñ queries a protein sequence against InterPro.

Kangaroo - Kangaroo is a pattern search program. Given a sequence pattern the program will find all the records that contain that pattern.

MEME ñ Multiple EM for Motif Elicitation - MEME is a tool for discovering motifs in a group of related DNA or protein sequences.Takes as input a group of DNA or protein sequences (the *training set*) and outputs as many motifs as requested. MEME uses statistical modeling techniques to automatically choose the best width, number of occurrences, and description for each motif.

MOTIF - findssequence motifs in a query sequence, also provides functional and genomic information of the found motifs using DBGET and LinkDB as the hyperlinked annotations. Results presented graphically, and, where available, 3D structures of the found motifs can be examined by RasMol program when the hits are found in PROSITE database.Also, given a profile generated from the multiple sequence alignment, or, retrieved from a motif library such as PROSITE or Pfam, you can align a protein sequence with the profile.

Network Protein Sequence Analysis -this multi-algorithm server offers two pattern and signature searches: PATTINPROT: scan a protein sequence or a protein database for one or several pattern(s) andPROSCAN: scan a sequence for sites/signatures against PROSITE database.

PFam HMM Search - Analyzes a protein query sequence to find Pfam domain matches.

PPSearch - Protein motifs searches

PredictProtein - this multi-algorithm server searches the PROSITE Database to detect **functional motifs** and PRODOM to detect **protein domains.**

ProDom BLAST ñ BLAST homology search against all domain sequences in ProDom.

ProfileScan Server- compares a protein or nucleic acid sequence against a profilelibrary (PROSITE or Pfam).

ProtoMap ñ classifies a new protein sequence.

Pscan - uses information derived from the PRINTS database to detect functional fingerprints in protein.

P-val FingerPRINTScan - find the closest matching PRINTS fingerprint/s to a query sequence.

ScanProsite - Scans a protein sequence for the occurrence of patterns stored in the PROSITE database.

SMART ñ Simple Modular Architecture Research Tool

SPRINT ñ Search the PRINTS-S Database.

3motif ñ searches by eMOTIF, PDB Structure or BLOCKS accession number.

SECONDARY SEARCH

Folding and coiling due to H-bond formation determines secondary structure.H-bonds form between carboxyl and amino groups of nonadjacent amino acids.A single polypeptide can have both helical and sheet regions.Non-helix and sheet regions can form bends, loops or turns.

BTPRED ñ The Beta-Turn Prediction Server ñ temporarily down

CPHModels - predicts protein structure using comparative (homology) modelling.

COILS - compares a sequence to a database of known parallel two-stranded coiled-coils and derives a similarity score. By comparing this score to the distribution of scores in globular and coiled-coil proteins, the program then calculates the probability that the sequence will adopt a coiled-coil conformation.

Garnier Peptide Structure Tool - is an implementation of the original Garnier Osguthorpe Robson algorithm (GOR I) for predicting protein secondary structure. Secondary structure prediction is notoriously difficult to do accurately. The GOR I alogorithm is one of the first semi-successful methods.

HTH - gives a practical estimation of the probability that the sequence is a helix-turn-helix motif.

Jpred^2 - takes either a protein sequence or a mulitple alignment of protein sequences, and predicts secondary structure. It works by combining a number of modern, high quality prediction methods to form a consensus.

META PredictProtein ñ this multi-algorithm server utilizes eight different algorithms for predicting secondary structure.

MultiCoil - program predicts the location of coiled-coil regions in amino acid sequences and classifies the predictions as dimeric or trimeric. The method is based on the PairCoil algorithm.

PairCoil - predicts the location of coiled-coil regions in amino acid sequences by use of Pairwise Residue Correlations.

PredictProtein ñ this multi-algorithm server utilizes two algorithms to predict secondary structure.

PREDATOR - an accurate algorithm for secondary structure prediction based on recognition of potentially hydrogen-bonded residues in the amino acid sequence.

PSA Protein Structure Prediction Server - determines the probable placement of secondary structural elements along a query sequence.

PSIPRED

Structure Prediction Server ñ this multi-algorithm server uses the PHD algorithm to predict secondary structure.

SOSUI

SSpro - Protein secondary structure prediction based on Bidirectional Recurrent Neural Networks (BRNNs).

Tandem Repeats Finder - a program to locate and display tandem repeats (two or more adjacent, approximate copies of a pattern of nucleotides) in DNA sequences.

Tmpred ñ Prediction of Transmembrane Regions and Orientation - makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occuring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring.

TMHMM - predicts transmembrane helices and the predicted location of the intervening loop regions.

TERTIARY STRUCTURE

Tertiary structure results from folding of thesecondary structural elements.Tertiary structure is stabilized by bonds formed between amino acid R groups (H-bonds, ionic interactions, covalent bonds, hydrophobic interactions).

Dali - compares the coordinates of a query protein structure andcompares them against those in the Protein Data Bank. The output consists of a multiple alignment of structural neighbours.

SWEET - a program for constructing 3D models of saccharides
from their sequences using standard nomenclature.

3D-pssm - A Fast, Web-based Method for Protein Fold Recognition using 1D and 3D Sequence Profiles coupled with Secondary Structure and Solvation Potential Information.

TraDES - a New Way to Customize and Explore Protein Conformational Space.

PROTEIN CHEMISTRY

Compute pI/MW

CUTTER: A tool to generate and analyze proteolytic fragments.

FindMod Tool - predicts potential protein post-translational modifications (PTM) and find potential single amino acid substitutions in peptides.

GlycoMod Tool - predicts the possible oligosaccharide structures that occur on proteins from their experimentally determined masses.

PEPSTATS: Protein Statistics - outputs a report of simple protein sequence information including: molecular weight, number of residues, average residue weight, charge, isoelectric point, for each type of amino acid: number, molar percent, DayhoffStat, for each physico-chemical class of amino acid: number, molar percent.

Phospepsort4

PredAcc - Protein side chains relative solvent accessibility prediction.

ProtParam Tool - allows the computation of various physical and chemical parameters for a given protein stored in SWISS-PROT or TrEMBL or for a user entered sequence.

YinOYang 1.2 Prediction Server - produces neural network predictions for O-þ-GlcNAc attachment sites in eukaryotic protein sequences.

PROTEIN SEQUENCE FOR ANALYSIS

Analyze the following sequence.

- **For each type of search use three different search**
- **Similarity/homology searches**
- **Motif,**
- **Pattern & Profile searches**
- **Secondary Structure Prediction**
- **Tertiary Structure Prediction**

SRYPGQVSFGGIGGLNDQIRELREVIELPLKNPELFLRVGIKPPKGVLLYGPPGTGKTL LARAVASSLETNFLKVVSSAIVDKYIGESARLIREMFGYAKGTRALHHLHGRDRCHR WQAFQRGYICRQRNPAYTYGAPQPARRFRLSRQDQDHHGDEPPRYPRPCFAACRPSR SQD

QUESTIONS TO PRACTICE:

1. **What are non-canonical amino acids? Discuss its applications in protein engineering.**
2. **Discuss the Aminoacyl t-RNA synthetases structure.**
3. **Engineering of t RNA and Aminoacyl t-RNA synthetase for the site specific in corporation of unnatural amino acids into proteins in vivo.**
4. **Write a brief note on protein scaffolds and its choice for protein engineering**
5. **Brief on protein structure prediction methods**
6. **What is molecular modeling? What its applications in protein engineering**
7. **Discuss in brief about energy minimization**
8. **Give a detailed account on protein databases**

UNIT – V - ENZYME AND PROTEIN ENGINEERING – SBTA5202

5. APPLICATIONS OF DIRECTED EVOLUTION TOOLS

5.1 Applications in enzyme engineering

Enzyme biocatalysis is increasingly viewed as a competitive and cost-effective alternative for the manufacturing of fine chemicals, pharmaceuticals, and agrochemical intermediates. Enzymes have major appeal for catalysis because of their high turnover number and refined level of selectivity, particularly in the synthesis of single-enantiomer compounds. Until recently, most of the successful industrial applications of enzymes have been limited to hydrolytic enzymes such as lipases, esterases, acylases, and hydantoinases. This situation is changing with the emergence of enzymes that perform a wide range of transformations, including asymmetric reduction, oxidation, and carbon–carbon bond formation.

Historically, microbial culture has been the most important route for enzyme discovery, even though only a small fraction of all microbes can be sampled by this method. This classical strategy has rapidly been replaced by high-throughput methods based on genomic sequence discovery. However, even these strategies are limited by the natural ability of enzymes to perform only a well-defined set of transformations. Directed evolution has been used with great success in recent years for the diversification of gene sequences and optimization of enzyme phenotypes. By surveying the available gene sequence space, specific traits are created through screening of libraries consisting of $10^4 - 10^{10}$ individuals. In all cases, optimal assay development is critical to the success in optimizing the fitness landscape of these enzymes.

Improving catalytic activity/stability

One of the most popular applications of directed evolution is to improve enzyme activity or stability under well-defined process conditions. By screening for initial activity and residual activity at an elevated temperature, both the thermostability and activity of mesophilic subtilisin E and p-nitrobenzyl esterase were significantly increased. Similarly, a directed evolution approach was successfully used to enhance the specific activity of a thermophilic 3-isopropylmalate dehydrogenase at lower temperatures, demonstrating the flexibility of this method in tailoring desirable enzymatic traits. In addition to thermal properties, enzymes with enhanced activity have also been created. In one example, directed evolution was used to improve the hydrolysis rate of organophosphorus hydrolase for several Directed evolution tools in bioproduct and bioprocess development poorly degraded pesticides (25 to 700 fold), suggesting that this approach may be useful in generating other variants that could rapidly decontaminate structurally similar chemical warfare agents. Directed

evolution approaches have also been used to enhance catalytic activities in non-natural environments such as organic solvents, for organic-phase syntheses. Moore and Arnold created several p-nitrobenzyl esterase variants that were up to 60-fold more active in 30% dimethylformamide. Another recent work using error-prone PCR was described to achieve a five-fold improvement in the amylase activity at pH 10, an alkaline pH required for the paper industry and as a detergent additive.

Expanding specificity

Another application of directed evolution is to fine-tune the specificity of enzymes. Many successful examples have been demonstrated that are useful for the production of important industrial products. The E. coli D-2-keto-3-deoxy-6-phosphogluconate (KDPG) aldolase,which catalyzes the highly specific reversible aldol reaction on D-configurated KDPG substrates, was subjected to DNA shuffling and screening, and one variant was isolated capable of accepting both D- and L-glyceraldehyde as substrates in a non-phosphorylated form. More recently, the P450 BM-3 monoxygenase, normally specific for medium chain fatty acids, has been evolved to accept small hydrocarbon substrates and convert them at very high rates.

Perhaps the most dramatic success in this area is the use of directed evolution to create novel specificity and activity. Sun et al. used combinatorial mutagenesis to change the substrate specificity of galactose oxidase to use glucose as a substrate. One variant (with only three point mutations) exhibited activity against D-glucose and oxidized other primary and secondary alcohols. Family shuffling of two homologous biphenyl dioxygenases created several variants with enhanced substrate specificity towards ortho-substituted polychlorinated

biphenyls and other aromatic compounds such as benzene, suggesting the feasibility to expand the biodegradability of other highly recalcitrant pollutants.

In addition to substrate specificity, product specificity can also be altered by directed evolution. Wild-type toluene 4-monooxygenase (T4MO) of Pseudomonas stutzeri OX1 oxidizes toluene to p-cresol (96%) and oxidizes benzene sequentially to phenol, catechol, and

1,2,3-trihydroxybenzene. To synthesize novel dihydroxy and trihydroxy derivatives of benzene and toluene, DNA shuffling of the alpha-hydroxylase fragment of T4MO (TouA) and saturation mutagenesis of the TouA active site residues were used to generate random mutants. Several variants were isolated to form 4-methylresorcinol, 3-methylcatechol, and methylhydroquinone from o-cresol, whereas wild-type T4MO formed only 3-methylcatechol.

These variants also formed catechol, resorcinol, and hydroquinone from phenol, whereas wild-type T4MO formed only catechol. These reactions show the potential synthesis of important intermediates for pharmaceuticals.

Changing stereo- and enantio-selectivity

Often the production of enantiomerically pure compounds is of extreme importance, particularly in the pharmaceutical industry. In this respect, directed evolution has been useful in creating enzymes with desirable enantioselectivity. May et al. were the first to demonstrate Rubin-Pitel et al. the feasibility to invert the enantioselectivity of D-hydantoinase to generate an enzyme that has enhanced selectivity towards L-5-(2-methylthioethyl)hydantoin. Similarly, inversion of enantioselectivity of a lipase was achieved towards (R)-selectivity with E =30 (comparing to E = 1.1 for the wild type enzyme). Perhaps the best industrial success was demonstrated with the synthesis of cis-(1S, 2R)-indandial, a key precursor of an inhibitor of HIV protease, by toluene dioxygenase . In three rounds of screening, several variants with up to three-fold decrease in production of the undesirable 1-indenol (only 20% from 60%) were obtained. In addition to enantioselectivity, the steroselectivity can be easily altered by directed evolution. Williams et al. demonstrated that stereospecificity of tagatose-1,6-bisphosphate aldolase can be altered by 100-fold via three rounds of DNA shuffling and screening. The resulting mutant catalyzes the formation of carbon-carbon bonds with unnatural diastereoselectivity, where the >99:<1 preference for the formation of tagatose 1,6-bisphosphate was switched to a 4:1 preference for the diastereoisomer, fructose 1,6-bisphosphate.

5.2. Applications in pathway engineering

Metabolic pathway engineering is a rapidly growing area with great potential to impact industrial biocatalysis. As enzymes are the central components in metabolic pathways, the strategy for the generation of sequence diversity and the screening/selection methods can be readily applied for pathway engineering. Directed evolution can be used to optimize an existing pathway, but the ability of this evolutionary approach to create new pathways that are capable of synthesizing novel compounds may be the most promising aspect for the future.

Carotenoids are important antioxidants and food additives that have been attracting commercial attention in recent years. Unfortunately, the synthesis of useful quantities from conventional chemical routes or from natural microorganisms is often costly and limited. The colorful nature of carotenoids makes them easy to detect via high-throughput screening. As a result, gene clusters for carotenoid synthesis have been introduced into E. coli and by performing directed evolution on two phytoene desaturases and two lycopene cyclases, several novel carotenoids were produced. More recently, the C30 carotene synthase CrtM from Staphylococcus aureus was subjected to one round of mutagenesis and screening, and variants

capable of synthesizing C-40 carotenoids were identified. This plasticity of CrtM with respect to its substrate and product range highlights the potential in creating further new carotenoid backbone structures. As a result, previously unknown C-45 and C-50 carotenoid backbones were obtained from the appropriate isoprenyldiphosphate precursors. Similar strategies have been applied successfully to evolve pathways for porphyrin synthesis. Polyketides belong to a second class of important bioactive compounds and efforts have been directed towards the generation of novel structures for uses as antibiotics or anti-cancer agents. The modular nature of the polyketide synthases (PKS) renders polyketide synthesis inherently amenable to directed evolution strategy, particularly in the engineering of novel polyketide structures. Typically a given PKS can generate only one product. However, Shen et al. reported that a minimal PKS from Streptomyces coelicolor is capable of generating more than 30 different structures, suggesting the flexibility in engineering a large Directed evolution tools in bioproduct and bioprocess development number of useful structures by a single PKS. By systematically deleting domains of the erythromycin PKS or exchanging domains with other PKS modules, several variants were obtained that are capable of generating more than 50 different polyketide. These examples imply the feasibility of creating entirely novel products via directed evolution of metabolic pathways.

Directed evolution can also be used as a powerful tool in optimizing an entire metabolic pathway. Functional evolution of an arsenic resistance operon has been accomplished by three rounds of shuffling and selection, resulting in cells that grew in 0.5 M arsenate, a 40- fold increase in resistance. Ten mutations were located in arsB, encoding the arsenite membrane pump, resulting in a 4-fold to 6-fold increase in arsenite resistance. While arsC, the arsenate reductase gene, contained no mutations, its expression level was increased, and
the rate of arsenate reduction was increased 12-fold.
Directed evolution has also been shown to enable the construction of artificial networks of transcriptional control elements in living cells. By applying directed evolution to genes comprising a simple genetic circuit, a nonfunctional circuit containing improperly matched components can evolve rapidly into a functional one. Such an approach is likely to result in a
library of genetic devices with a range of behaviors that can be used to construct more complex genetic circuits.

5.3 Protein Engineering of Enzymes Involved in Bioplastic Metabolism

The petroleum industry has optimized profi ts by producing value-added coproducts, such as plastics and chemicals, in addition to primary liquid fuels. A similar coproduct strategy applied to biorefineries processing cellulosic biomass to liquid fuels and/or energy would transform a technology that is marginally economic, depending on oil prices, to a sustainable business with enhanced revenue streams from multiple coproducts. The challenge is fi nding a biobased coproduct that is compatible with a biorefi nery scenario and where markets warrant its production on a similar scale as liquid fuels and/or energy. **Polyhydroxyalkanoate (PHA)** bioplastics represent a coproduct that would be entirely compatible with either production of liquid fuels by hydrolyzing the residual biomass after PHA extraction or by alternative thermochemical processes. PHA bioplastics possess properties making them suitable replacements for many of the applications currently served by petroleum-based plastics, thus providing tremendous market potential.

The value-added PHA bioplastic needs to meet a number of key criteria:

- Material properties suitable for very large markets
- Biologically achievable in plant crops
- Realistic production scenarios in a biorefinery

PHAs are a natural component of numerous organisms in multiple ecosystems and exist as both high and low molecular weight molecules. High molecular weight PHAs – greater than 60 000 Daltons– are linear polyesters that accumulate in a wide range of gram-positive and gramnegative bacteria , as well as some Archaea, as an intracellular granular storage material when the microbes are faced with an unfavorable growth environment, such as a limitation in an essential nutrient. The monomer unit composition of these polymers is largely dictated by the available carbon source as well as the native biochemical pathways present in the organism. When favorable growth conditions return, the polymer is degraded and provides carbon for microbial growth. Low molecular weight PHAs – less than 15 000 Daltons and sometimes called cPHB– are primarily composed of the monomer 3-hydroxybutyrate, a natural ketone body found in human blood. Low molecular weight PHAs have been found in a wide range of prokaryotes and eukaryotes, including humans, and are believed to be a constituent of every living cell. High molecular weight PHAs have attracted considerable interest from industry since chemically they are polyesters and, from a physical properties perspective, they are thermoplastics which can be melt-processed into various fi nal forms. For these reasons we refer to these materials as PHA bioplastics. PHA bioplastics are present throughout nature and are truly biocompatible molecules. Th ey are biodegradable in all biologically active environments and can be produced industrially from renewable resources.

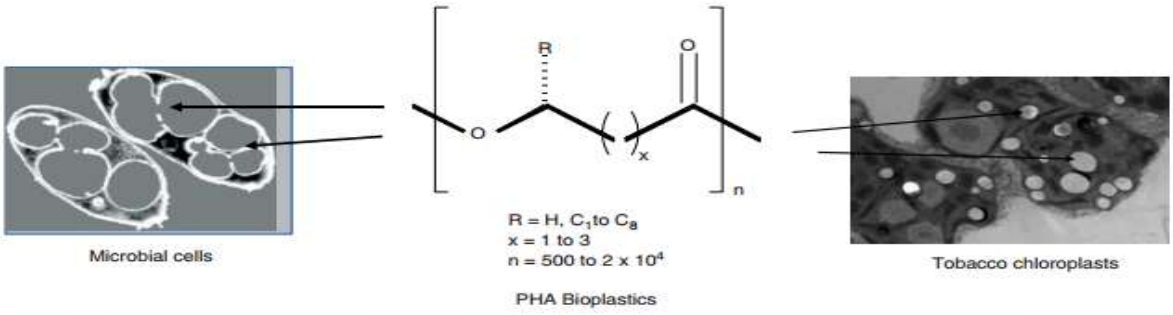


Figure 1. Chemical and physical structure of PHA Bioplastics. Over 150 different monmeric units have been observed in PHA Bioplastics.[51] The chemical structures of the most commonly observed monomeric units are summarized in the diagram. PHAs accumulate as distinct intracellular granules in both native and engineered producers. Electron micrographs of PHA granules in microbial cells (left panel) and tobacco plastids (right panel: Bohmert *et al.*, unpublished results) are shown.

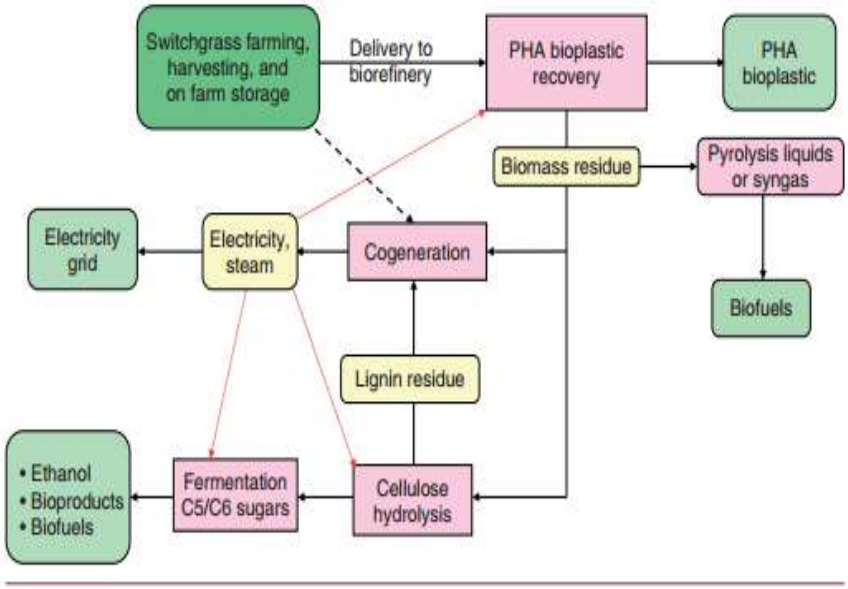
Micro organisms capable of producing PHA bioplastics have been found in a range of environments,including soil, marine, sewage, and even groundwater aquifers. PHAs are already present in fi elds used for large-scale agriculture since many of the bacteria present in the rhizosphere, or the soil that surrounds the roots of plants, produce PHA bioplastics as a storage material. Most of the rhizobia that form symbiotic relationships with legumes contributing to nitrogen fixation are also capable of producing PHA bioplastics. These bacteria live within root nodules of many of the legumes that are grown in large-scale agriculture. Biocompatibility of PHA bioplastics in animals has been demonstrated in feeding trials with sheep, pigs, and broiler chicks, where repeated ingestion of PHAs was not found to adversely affect animal health. In general, milled granules of PHAs added to feed were digestible by animals if they were pre-treated with a substance, such as sodium hydroxide, to lower the molecular weight of the polymer. PHAs pre-treated in this fashion were found to provide added nutritional value to the feed. Other areas in which PHA bioplastics have shown to be safe include medical applications where the material is implantedor as a solid carbon substrate or support for aquaculture de-nitrifi cation processes. Studies with the monomer 3-hydroxybutyrate suggest that it may have value in nutraceutical and therapeutic applications.

In native producers, PHA bioplastics are produced intracellularly as storage granules by a number of enzyme pathways, encoded by distinct sets of genes, which convert cellular intermediates such as acetyl-CoA or intermediates of fatty acid pathways found in all cell types to polymer. PHA bioplastics are currently being commercialized based on microbial fermentation technologies using a two-stage production process. The first stage involves fermentation of a microbe which accumulates PHA intracellularly using a renewable feedstock such as corn wet mill dextrose, cane sugar or even vegetable oil. Once the fermentation process is complete, the bacterial cells are harvested and the PHA polymer is extracted from the cells using either a solvent extraction process or an aqueous process in which the non-PHA component of the microbial cell is digested either chemically or enzymatically and then removed from the PHA polymer.

Realizable production scenarios in a biorefinery

Potential biorefi nery scenarios for PHA bioplastic production from switchgrass are illustrated in the Fig below. The switchgrass would be grown commercially by farmers, harvested using existing equipment, and either stored as large bales at the farm for up to six months or pelletized for longerterm storage. The switchgrass would then be transported as needed to centralized biorefi neries and processed throughout the year. In the simplest scenario, the PHA is recovered using a solvent extraction process and residual dry biomass is used for bioenergy production in a cogeneration (also called combined heat and power) facility. Th e most likely channel to market for the first generation

of PHA bioplastic from switchgrass would be through conversion to chemical intermediates like esters, or blending with fermentation produced PHA bioplastics to expand their properties as described above.



PHA bioplastics are a value-added coproduct that possess a market size compatible with large-scale production of biofuels and/or energy from plants. They have material properties suitable for accessing the markets currently served by petroleum based plastics and their production has been demonstrated in several leading candidate bioenergy crops. Production of these materials in biomass crops has the potential to significantly improve the economics of biomass biorefineries producing liquid fuels and/or energy. We have discussed reasonable production scenarios that will take time to implement but present very attractive business opportunities with exciting revenue streams, providing an economic framework for a truly sustainable business.

5.4. Bioengineering of Sequence - Repetitive Polypeptides

Protein - based materials, which correspond to polymers of tandemly repeated oligopeptide sequence motifs, have been the focus of signifi cant research interest over the past two decades . The intellectual driving force for this process has come from two distinct directions: fi rst, from interest in the fundamental polymer science of architecturally uniform macromolecules; and second from interest in the structural biology of native, protein - based materials. From the viewpoint of fundamental polymer science, protein - based materials represent an approach to understand the effect of polymer architectural parameters (composition, sequence and molar mass) on macromolecular properties. Ribosomal protein synthesis ensures a uniformity of polymer microstructure that is impossible to achieve using the conventional synthetic methods employed for organic polymerization reactions. Thus, non - natural polypeptide sequences can been synthesized with near - absolute control of architectural parameters, and these biologically synthesized poly(α - amino acids) can be considered as model uniform polymers. These synthetic protein - based materials may provide insight into the fundamental aspects of polymer physical chemistry both in solution and in the solid - state, potentially enabling the creation of material constructs that display novel behavior . Furthermore, the observed control of polypeptide primary structure also implies the ability to defi ne a higher - order structure through the progression of protein structural hierarchy. Secondary and super - secondary elements,

and the interactions between them, can be specifi ed through the sequence identity although, as with more conventional targets of protein design, the currently limited ability of theoretical approaches to reliably defi ne the relationship between amino acid sequence and higher - order molecular and supramolecular structure is a signifi cant constraint upon the design of novel polypeptide architectures. Nevertheless, genetic engineering methods have been employed to create artifi cial polypeptides of defi ned sequence that self - assemble into structurally defi ned supramolecular aggregates, including lamellar crystallites surface - stabilizing coatings , smectic liquid crystalline mesophases , thermoresponsive nanoparticles and nanostructured hydrogels , on the basis of structural features programmed into the polypeptide sequences at the molecular level. These de novo - designed biomaterials provide an indication of the potential for biosynthesis to provide novel materials through the near - absolute control of macromolecular architecture.

The second factor that has motivated the investigation of protein - based materials lies in the desire to understand the chemical, biological and mechanical properties that underlie the native biological function of fi brous proteins . Natural evolutionary processes have afforded an array of structurally diverse protein - based materials that are produced within organisms as a natural consequence of their life cycle. These native protein - based materials usually display low complexity sequences that consist of tandem repeats of a fundamental oligopeptide motif that displays limited plasticity in amino acid sequence, and thus they bear a nominal similarity to the repeat sequences of conventional organic polymers. The unique structural and functional properties of these native materials presumably arise as a consequence of their sequence specifi city, which strongly infl uences the mode of self - assembly of the polymer chain into the supramolecular architectures that underlie their materials properties. Most notably, the materials properties of these native proteins often surpass the performance of synthetic materials within the relatively narrow compass of environmental conditions that defi ne these biological systems. Structural variants of these native proteins have been envisioned for technological applications as high - performance materials and, indeed, have provided the intellectual driving force for the development of conventional polymer science during the last century. Dragline silk fi bers from the spider Nephila clavipes display a unique combination of high tensile and compressive strength that presumably originates in the segmented structure of the fi broin proteins that comprise the dragline fiber.

The biosynthesis of artifi cial protein - based polymers derived from sequence - repetitive polypeptides has developed in conjunction with the fundamental advances in recombinant DNA cloning and protein expression techniques over the past 25 years. Although this technology was not developed for the synthesis of protein - based materials, these techniques were soon applied to the synthesis of sequence - repetitive polypeptides based on the canonical repeats observed for native fi brous proteins such as elastin, collagen, keratin and silk. This approach met with mixed results in that, although signifi cant knowledge was obtained with respect to cloning and expression of repetitive polypeptides, considerable challenges remained to be addressed, including the development of better methods to stabilize highly repetitive DNA sequences, to optimize recombinant protein yield, to promote appropriate post - translational modifi cation, and to process the protein into a form that approximates that of the native state of the protein from which the sequence was originally derived.

Block Copolymers as Targets for Materials Design

Synthetic copolymers consisting of well - defi ned blocks of compositionally dissimilar monomers spontaneously self - assemble in the solid state into ordered domains of similar blocks. These hybrid

materials have been extensively studied and often have unique, technologically signifi cant properties in comparison to blends of the respective homopolymers. For example, copolymers comprising distinct blocks of different mechanical and chemical properties have been employed as polymer surfactants, pressure - sensitive adhesives, blend compatibilizers, thermoplastic elastomers, mineralization templates and lithographic resists. In contrast, block co - polypeptides have not garnered as much attention. However, the recent development of biosynthetic and chemosynthetic methods for the preparation of well - defi ned block copolymers of peptide sequences promises the potential for rapid advancements. These materials could be potentially interesting based on the diverse structures and functions observed for naturally occurring protein materials, in which the repetitive sequence pattern induces a regular secondary structure within the individual domains of the block co - polypeptide that has an important effect upon the supramolecular organization of the material. Segregation of the blocks into compositionally, structurally and spatially distinct domains occurs in analogy with synthetic block copolymers, affording ordered structures on the nanometer to micrometer size range. The sequence control and structural uniformity of these natural block copolymers is presumably responsible for their unique materials properties. The genetic engineering of synthetic polypeptides enables the preparation of block copolymers composed of complex sequences in which the individual blocks may have different mechanical, chemical or biological properties. The utility of these protein materials depends on the ability to functionally emulate or enhance the materials properties of conventional polymer systems, while retaining the benefi ts of greater control over the sequence and microstructure that protein engineering affords for the construction of materials. This precise control of macromolecular architecture provides an opportunity for tailoring technologically signifi cant materials properties for directed applications, for example, in biomedicine.

Amphiphilic Block Copolymers

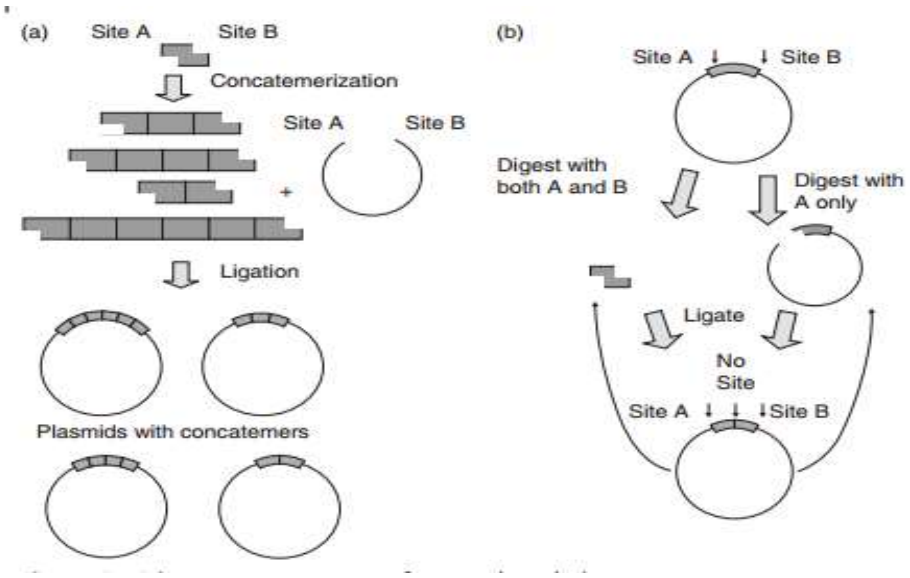
One notable characteristic of technological signifi cance for many conventional block copolymers is the property of amphiphilicity , that is, a difference in hydrophilic versus lipophilic (hydrophobic) character between the respective blocks. Amphiphilic block copolymers represent a special class of materials that are composed of compositionally defi ned blocks that have signifi cantly different interaction affi nities for aqueous solutions. These hybrid materials have attracted scientifi c interest due to their complex phase behavior in selective solvents, which parallels and complements that of small - molecule surfactant amphiphiles such as phospholipids. Amphiphilic diblock (AB) and triblock (ABA or BAB) copolymers undergo selective segregation of the hydrophobic domain in aqueous solvents to form micellar structures in which the corona of the micelle is derived from the hydrophilic block (A) and the core of the micelle from the hydrophobic block (B). The identity and sequence of the individual block units within the polymer dictates the nature of the supramolecular assembly. In contrast to small - molecule surfactants, the phase behavior of amphiphilic block copolymers can be modifi ed conveniently through manipulation of the macromolecular architecture – that is, the length, composition and sequence of the individual blocks. In addition, the hydrophilic/lipophilic balance can be adjusted systematically by variation of the relative lengths of the hydrophilic and hydrophobic blocks. These materials display several key features that may confer advantages over conventional surfactant amphiphiles in controlled delivery and release applications, namely, very low critical micelle concentrations, slow unimer – micelle exchange rates, high aggregate stabilities, and a controllable range of aggregate sizes and morphologies.

Elastin - Mimetic Block Copolymers

Elastin is a native protein - based material that is the primary structural component underlying the elastomeric mechanical response of compliant tissues in vertebrates and, therefore, has potential signifi cance for human health as a medical biomaterial for the preparation of tissue - engineered analogues of native elastin - containing human systems. Moreover, elastin - mimetic polypeptides display a well - defi ned correlation between repeat sequence and macromolecular properties (vide infra), which enables the creation of a wide variety of synthetic elastin analogues with tailorable biophysical properties. The elastomeric domains of elastin comprise structurally similar oligopeptide motifs that are tandemly repeated in the native protein sequence. The local secondary structure and macromolecular thermodynamic and viscoelastic properties of the elastomeric domains can be emulated by synthetic polypeptides that are composed of a concatenated sequence of native oligopeptide motifs; the most common of which is the pentapeptide (Val - Pro - Gly - Val - Gly). Polypeptides based on these pentameric repeat sequences undergo reversible, temperature - dependent, hydrophobic assembly from aqueous solution in analogy to the phase behavior of native tropoelastin, the soluble precursor of crosslinked elastin. This process results in a spontaneous phase separation of the polypeptide above a critical solution temperature, Tt , which is near ambient temperature in vitro. This inverse temperature transition coincides with a conformational rearrangement of the local secondary structure within the pentapeptide motifs.

Strategies for the Construction of Synthetic Genes Encoding Sequence - Repetitive Polypeptides

The synthesis of the protein - based materials based upon complex sequence repeats is best accomplished using the techniques of recombinant DNA (rDNA) technology and bacterial protein expression. The advantage of these methods lies in the ability to directly produce, with high fi delity, synthetic polypeptides of exact amino acid sequence and high molecular weight, as opposed to chemically synthesized oligopeptides, which are essentially limited to low degrees of polymerization (< 60 residues). With regards to the discussion herein, the term ‘ protein - based material ’ implies a sequence - repetitive polypeptide, or a multidomain protein consisting of one or more sequence - repetitive polypeptides, that is encoded within a synthetic DNA expression cassette. As materials properties generally scale to some degree with chain length, the biosynthesis of these protein polymers usually requires the construction and expression of large, synthetic genes containing multiple direct repeats of a ‘ monomeric ’ DNA sequence of approximately 50 to 150 base pairs in length. As automated DNA synthesis technology is currently limited to the production of oligodeoxynucleotides of lengths corresponding to about a hundred bases, sequences encoding medium to high - molecular - weight polypeptides cannot be obtained by direct synthesis of the entire gene. In addition, such repetitive DNA sequences may be unstable with respect to homologous recombination, and this may result in the structural instability of plasmid clones in vivo. Therefore, synthetic procedures for the cloning and expression of the repetitive genes may require special experimental considerations beyond conventional DNA manipulations (vide infra). Two main approaches have been described that are complementary in experimental methodology: (i**) DNA cassette concatemerization** and (ii) **recursive directional ligation.** Both strategies involve the chemical synthesis of the corresponding DNA sequence encoding the desired peptide repeat motif, enzymatically induced concatemerization, ligation of the concatemer into a plasmid vector, propagation in a bacterial host and, fi nally, expression of the repetitive polypeptides. However, the two strategies differ signifi cantly in the method that is employed for generation of the concatemers and subsequent manipulation of the cloning vectors.



(i)DNA Cassette Concatemerization

This experimental protocol involves the construction of double - stranded oligonucleotide segments (DNA ‘ monomers ’) containing nonpalindromic, cohesive ends. Generation of the cohesive - ended DNA monomers is generally accomplished through the use of restriction endonucleases capable of recognizing and cleaving nonpalindromic sequences. The sizes of the oligopeptide repeats are usually chosen such that they could be conveniently encoded within single DNA cassettes of approximately 50 to 150 base pairs in length prior to concatemerization. Self - ligation of the DNA monomers proceeds in a head - to - tail fashion to generate a library of concatemers which differ in length by increments of the monomer. Preparative agarose gel electrophoresis is used to fractionate the concatemers according to the degree of concatemerization. Concatemers within the desired size range are extracted from the gel and used directly in subsequent cloning steps. A critical consideration for the successful application of this procedure is the effi cacy of cloning and screening a population of concatemers to identify a construct of appropriate size. Unless concatemeric DNA cassettes corresponding to individual bands are excised from the gel, it is diffi cult to isolate and clone concatemers of determinate size using this approach. Usually, the sizes of individual concatemers are identifi ed through screening a population of clones in parallel using either colony screening polymerase chain reaction(PCR) or restriction digestion of isolated plasmid - based constructs. Often, this process may require the screening of a large number of clones to identify a cassette of the desired size. DNA cassettes corresponding to very high degrees of concatemerization have been isolated using this procedure, although it is typically challenging to isolate a clone corresponding to a specifi c size. Although laborious, these protocols have been widely employed for the synthesis of artifi cial genes encoding sequence - repetitive polypeptides based on natural sequences, as well as artifi cial proteins having no natural parallel. However, diffi culties have been reported in obtaining long concatemers as cloned inserts using this approach . Modifi ed concatemerization strategies have been described in which DNA adaptors have been appended to the termini to facilitate cloning into conventional plasmid - based vectors [4, 49] , although these approaches do not necessarily address the problems associated with low yields of long concatemers.

Recursive Directional Ligation

In contrast to the DNA cassette concatemerization approach, recursive directional ligation permits the isolation of concatemers of determinate size through a controlled oligomerization process that is

facilitated by the DNA manipulation experiments. Although several variations of the basic protocol have been described [46] , the general procedure involves iterative directional insertions in a plasmid - based vector in which smaller concatemers are joined together recursively to form larger ones. The size of the DNA product that results from the cloning procedure corresponds to the sum of the initial DNA reactants. Thus, two DNA monomers can be joined together in a plasmid to form a dimeric construct. Two equivalents of the resulting dimeric construct can be joined to form a tetrameric construct, and so on. Repetitive application of this process, in which the products from a prior step are employed as the reactants in a successive step, can afford large concatemeric cassettes of determinate size. This procedure relies on the judicious choice of restriction sites at the termini of the DNA cassettes to facilitate the directional cloning process. As for DNA cassette concatemerization, restriction endonucleases that recognize and cleave nonpalindromic sites are very useful for the generation of cohesive - ended DNA fragments that are competent for selective ligation. Recursive directional ligation has the advantage that synthetic genes of determinate size and sequence can be obtained, although the process can be labor intensive for the assembly of large genes. For example, the assembly of a concatemer that encodes 32 (2 5) repeats of the basic sequence motif requires at least fi ve iterations of the directional cloning process in which the size of the concatemeric construct is doubled at the end of each step. Nevertheless, a signifi cant number of sequence - repetitive polypeptides have been produced from synthetic genes assembled in this fashion. Recursive directional ligation is no doubt the technique of choice for the creation of synthetic genes of defi ned size and sequence. In theory, neither gene assembly strategy places any restriction on the size of the cloned DNA concatemers, although practical considerations (e.g. the effi ciency of transformation of plasmid - based constructs and genetic instability of repetitive DNA sequences) may limit the effective size of cloned DNA inserts. It has been found that the ease of isolation of long DNA concatemers depends heavily on the identity of the DNA sequence. For many of the elastin - derived constructs, very large synthetic genes ($\geq$ 8000 base pairs) can be obtained that encode sequence - repetitive polypeptides, whereas for other polypeptide sequences the isolation of long DNA concatemers becomes very diffi cult due to genetic instability leading to a recombinative loss of the majority of the coding sequence.

5.5. Application of protein folding to design new drug

Proteins are polymeric chains of amino acids that organisms and cells rely on for signaling, pathogen clearing, mobility, catalysis, recognition, shape, ordering, and stability. The precise ordering of the amino acids in a protein sequence determines how the protein folds into a 3D structure, and thus its biological function. As our knowledge of the connection between sequence, structure, and function has advanced, interest has grown in designing proteins on a sequence level to produce novel folds and function. Brute-force experimental approaches to resolving protein structures and designing protein sequences for new functions remain time consuming and expensive, and add little to our understanding of the physical principles required for both problems. Protein structure prediction aims to determine accurately the full 3D structure of a protein given only its amino acid sequence. Structure prediction is challenging if only low homology templates exist. De novo protein design is the inverse problem given a rigid or flexible backbone structure, one aims to determine a sequence that will fold into that structure. Different sequences can fold into the same structure, so there is degeneracy in the protein design space. The existence and accuracy of protein structures as templates for protein design can have a significant impact on potential success. For this reason, the ability to produce viable

protein templates through protein structure prediction is important for protein design, and for advancement in biotechnology and drug discovery. Figure 1 schematically shows the roadmap and key challenges in protein structure prediction and de novo protein design. The past few years have shown impressive applications of computational structure prediction and design to biotechnology, spanning peptide or antibody therapeutics, novel biocatalysts, and self-assembling nanomaterials.

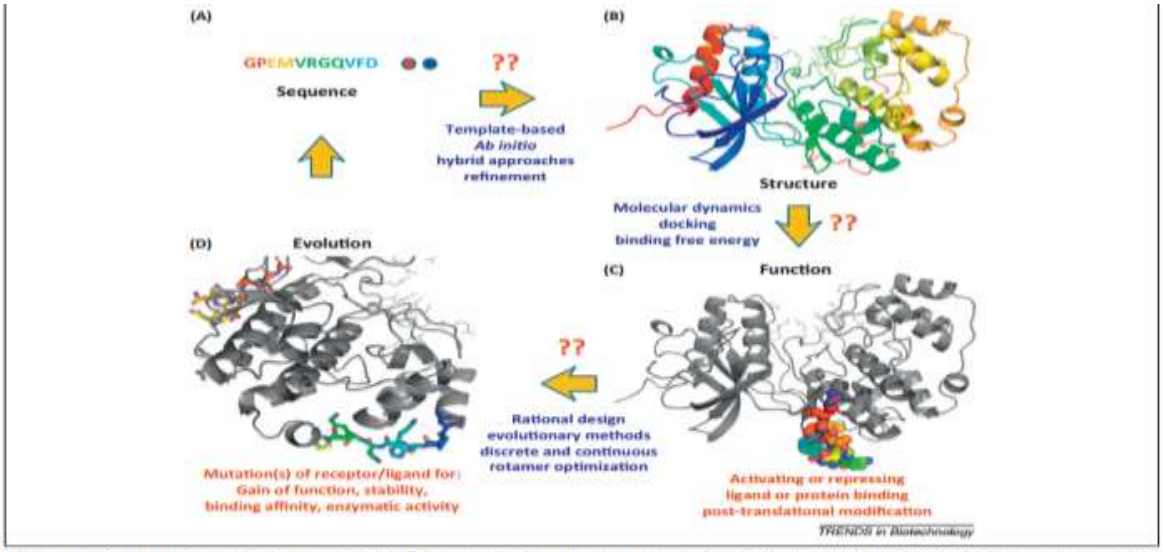


Figure 1. Roadmap of key challenges in understanding how to predict protein sequence to structure to function and design. Structure prediction begins with a primary amino acid sequence **(A)** and aims to predict the full 3D structure **(B)** of that sequence. **(C)** Other proteins, peptides, small molecules, or cofactors may form critical interactions with the protein structure critical to its function. Docking with or without binding free energy calculations may be required to find the most probable conformation for a ligand bound to a receptor protein. Understanding how structure leads to function remains a challenge. The protein structure may be subsequently post-translationally modified, and as most methods have focused in predicting the structures of canonical-amino-acid-containing proteins, the literature is lacking in the ability accurately represent post-translationally modified protein structures. The solution or accurate prediction of the 3D structure of a protein allows it to be used in a design context. **(D)** Biotechnological applications of protein design shown in the literature include designing/redesigning the receptor protein via site-specific mutations to change its binding affinity toward a ligand, change its fold, increase its stability, and create new or alternative enzymatic activity. The ligand of a peptide can be amenable to similar design strategies to design new sequences to bind more strongly to the receptor and compete with its native binding partner (antagonism) or to bind to and activate through a series of specific interactions with the receptor a particular downstream function (agonism). Upon design of the receptor or ligand peptide with new sequences, the cycle begins again as even a few mutations can cause structural conformation and topology changes. The structure shown in the figure is the mitogen-activated protein (MAP) kinase extracellular signal-regulated kinase (ERK)2. The ligand bound is the kinase interaction motif of MAP kinase phosphatase (MKP)3.

State-of-the-art advances and challenges in protein structure prediction and refinement

The consistent determination of structure from sequence is one of the greatest unsolved problems in nature and has recently passed the 50-year milestoneAccurately predicting the 3D structure of a protein involves a series of steps performed on a sequence of amino acids: secondary structure prediction (identifying whether local segments are helical, beta-strand, or loop), structural alignment to candidate template structures, conformational sampling, and selection (Figure 2A and Box 1). A predicted structure may then undergo refinement, in an attempt to improve the accuracy of that structure. Historically, most refinement methods degrade rather than improve the accuracy of the predicted structure, making protein structure refinement a substantial unsolved problem in its own right.

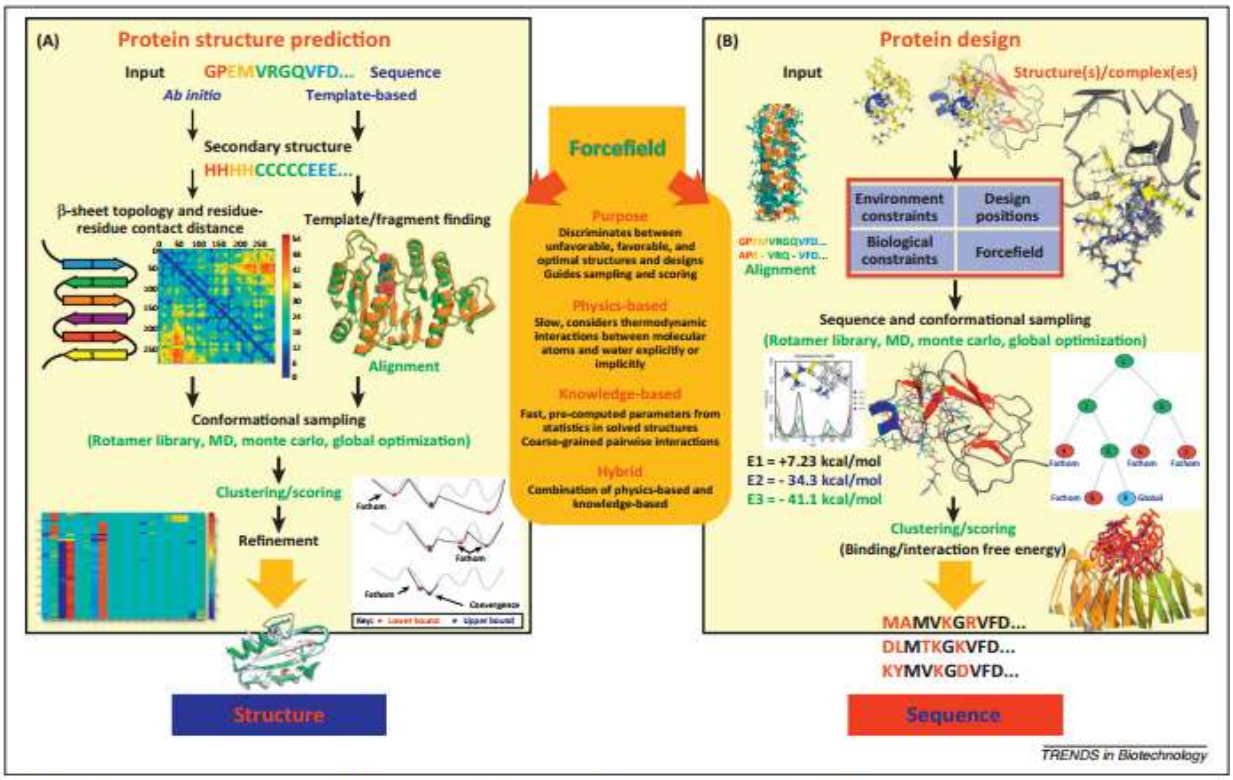
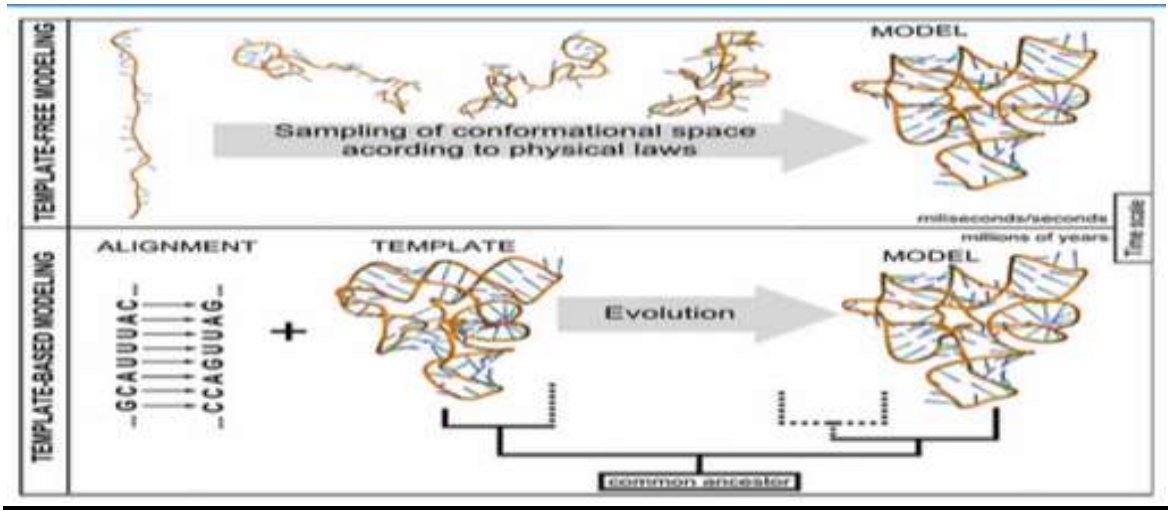


Figure 2. Detailed view of connections and differences between **(A)** protein structure prediction and **(B)** protein design. Dynaneomics image used with permission.

The Critical Assessment of Techniques for Protein Structure Prediction (CASP) occurs biennially and recently completed its tenth experiment. For the CASP competition, prediction targets are categorized into two groups depending on the availability of structural templates: (i) template-based modeling, in cases for which templates are available; and (ii) free modeling, in cases for which templates are not available.



Successful protein designs with biotechnological applications

Design of proteins and peptides for therapeutic applications:

Over 200 peptides, proteins, or antibody therapeutics have been marketed as of 2010. Computational approaches have recently been applied to design new proteins and peptides for therapeutic applications. Elucidation of the sequences, structures, and interaction patterns of several disease-

related proteins have allowed for the application of computational approaches for peptide therapeutic design [48]. Craik et al. predict that by 2020 we will see more prevalence of peptides as drugs, while outlining the challenges to meeting that outcome. Here, we review timely applications by target.

Cancer. Generally, therapeutic proteins/peptides can: (i) interfere with signal transduction cascades; (ii) arrest the cell cycle through modulation of cyclin-dependent kinase activity; or (iii) directly induce apoptosis by modulation of the proteins controlling apoptosis Cysteine-rich intestinal protein 1 (CRIP1) is an early biomarker for breast cancer. Hao et al. used phage display to identify peptide sequences that bound to CRIP1. Subsequently, they computationally redesigned the scaffold sequence to optimize the binding free energy to increase its affinity for CRIP1, finding experimentally that it improved the IC50 27.5x over the phage-displayed sequence.

HIV. A computational method using side-chain grafting and to transplant a continuous structural epitope, 4E10, into scaffold proteins for conformational stabilization and immune presentation was developed. . The method produces epitope-containing designs that bind stronger to monoclonal antibody (mAb) 4E10 than 4E10 alone, and inhibits neutralization by HIV+ sera. Floudas and coworkers designed HIV-1 entry inhibitors starting from the structure of the C14linkmid peptide in complex with the hydrophobic core of gp41. C14linkmid is a crosslinked peptide derived from the C-terminal heptad repeat gp41. A global optimization-based sequence selection was performed with a distance-dependent force- field originally developed for protein folding to select candidate sequences from the vast combinatorial space. These sequences were reranked using fold-specificity calculations, which sample conformations near the template structure with substitutions dictated by the newly designed sequences. It aims to determine how favorably a new sequence folds into the fold of the design template. A subset of top-ranked sequences identified in the fold-specificity stage was evaluated using approximate bindingaffinity calculations, which approximate the binding equilibrium constant. The best design had an IC50 between 29 and 253 mM for different HIV-1 donors and mutants. This de novo design approach was made into an interactive web interface, Protein WISDOM.

Alzheimer's disease. Eisenberg performed computationally guided design to predict and experimentally validate peptide inhibitors of fibril formation by the t protein associated with Alzheimer's disease, as well as an amyloid promoting the sexual transmission of HIV. The designs bind to the end of the steric zipper and inhibit elongation. Focusing on the t protein inhibitor methodology, for a rotameric, fixed-backbone sequence optimization, they inverted the chirality of the design target to enable use of the Rosetta suite of tools. They designed L-amino acid sequences that favorably interact with a fixed-atom D version of the scaffold. Subsequently, the scaffold was reverted to its native L- form, and D-amino-acid-containing peptides were used as inhibitors experimentally. The designed D-peptides were then verified for shape complementarity, noting that D-Leu2 of the peptide was designed to clash with the target VQIVYK on the opposite sheet, and upon alanine substitution, inhibitory activity ceased. Introducing a tight-binding interface and clashes destroying the ability of a cascade of amyloid-forming sequences to propagate is effective for inhibition. Pande and coworkers, guided by observations made in simulations of Ab42, designed a noncanonical and D-amino-acid-containing peptide that organizes Ab42 into stable oligomers.

Self-assembling proteins/peptides

Controlling ordered (i.e., crystals) or disordered (i.e., hydrogels) self-assembly of proteins is a critical test of our understanding of both structure and interactions, having applications in biologically inspired materials. Lanci et al. computationally designed a protein crystal starting from an idealized

homotrimeric parallel coiled-coil template and redesigned the interfaces. They utilized strictly physics-based energy functions to discriminate favorable interfaces. Stranges et al. took the solvent-exposed b strands of two monomeric proteins and redesigned them to form an intermolecular b sheet symmetric homodimer with near atomic-level accuracy. . This design demonstrated the creation of unique stabilizing interactions at an interface. King et al. designed symmetric self-assembling complexes to atomic level accuracy. They performed symmetric docking of subunits followed by redesign at the interfaces to design cage-like nanomaterials with tetrahedral or octahedral point group symmetry. The designed structures were confirmed experimentally by crystallography and electron microscopy to high agreement. The control over such selfassembling can be used to design advanced functional materials and molecular machines.