LECTURE 11: PROTEIN FOLDING

Protein Folding Is a Highly Cooperative Process

As stated earlier, proteins can be denatured by heat or by chemical denaturants such as urea or guanidium chloride. For many proteins, a comparison of the degree of unfolding as the concentration of denaturant increases has revealed a relatively sharp transition from the folded, or native, form to the unfolded, or denatured, form, suggesting that only these two conformational states are present to any significant extent (Figure 3.56). A similar sharp transition is observed if one starts with unfolded proteins and removes the denaturants, allowing the proteins to fold.

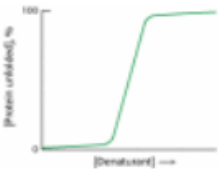

Figure 3.56

Transition from Folded to Unfolded State. Most proteins show a sharp transition from the folded to unfolded form on treatment with increasing concentrations of denaturants.

Protein folding and unfolding is thus largely an *"all or none" process* that results from a *cooperative transition*. For example, suppose that a protein is placed in conditions under which some part of the protein structure is thermodynamically unstable. As this part of the folded structure is disrupted, the interactions between it and the remainder of the protein will be lost. The loss of these interactions, in turn, will destabilize the remainder of the structure. Thus, conditions that lead to the disruption of any part of a protein structure are likely to unravel the protein completely. The structural properties of proteins provide a clear rationale for the cooperative transition.

The consequences of cooperative folding can be illustrated by considering the contents of a protein solution under conditions corresponding to the middle of the transition between the folded and unfolded forms. Under these conditions, the protein is "half folded." Yet the solution will contain no half-folded molecules but, instead, will be a 50/50 mixture of fully folded and fully unfolded molecules (Figure 3.57). Structures that are partly intact

and partly disrupted are not thermodynamically stable and exist only transiently. Cooperative folding ensures that partly folded structures that might interfere with processes within cells do not accumulate.

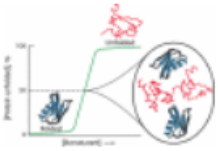

Figure 3.57

Components of a Partially Denatured Protein Solution. In a half-unfolded protein solution, half the molecules are fully folded and half are fully unfolded.

Proteins Fold by Progressive Stabilization of Intermediates Rather Than by Random Search

The cooperative folding of proteins is a thermodynamic property; its occurrence reveals nothing about the kinetics and mechanism of protein folding. How does a protein make the transition from a diverse ensemble of unfolded structures into a unique conformation in the native form? One possibility a priori would be that all possible conformations are tried out to find the energetically most favorable one. How long would such a random search take? Consider a small protein with 100 residues. Cyrus Levinthal calculated that, if each residue can assume three different conformations, the total number of structures would be 3^{100}, which is equal to 5×10^{47}. If it takes 10^{-13} s to convert one structure into another, the total search time would be $5 \times 10^{47} \times 10^{-13}$ s, which is equal to 5×10^{34} s, or 1.6×10^{27} years. Clearly, it would take much too long for even a small protein to fold properly by randomly trying out all possible conformations. The enormous difference between calculated and actual folding times is called *Levinthal's paradox*.

The way out of this dilemma is to recognize the power of *cumulative selection*. Richard Dawkins, in *The Blind Watchmaker,* asked how long it would take a monkey poking randomly at a typewriter to reproduce Hamlet's remark to Polonius, “Methinks it is like a weasel” (Figure 3.58). An astronomically large number of keystrokes, of the order of 10^{40}, would be required. However, suppose that we preserved each correct character and allowed the monkey to retype only the wrong ones. In this case, only a few thousand

keystrokes, on average, would be needed. The crucial difference between these cases is that the first employs a completely random search, whereas, in the second, *partly correct intermediates are retained*.

Figure 3.58

Typing Monkey Analogy. A monkey randomly poking a typewriter could write a line from Shakespeare's *Hamlet*, provided that correct keystrokes were retained. In the two computer simulations shown, the cumulative number of keystrokes is given at the left(more...)

The essence of protein folding is the retention of partly correct intermediates. However, the protein-folding problem is much more difficult than the one presented to our simian Shakespeare. First, the criterion of correctness is not a residue-by-residue scrutiny of conformation by an omniscient observer but rather the total free energy of the transient species. Second, proteins are only marginally stable. The free-energy difference between the folded and the unfolded states of a typical 100-residue protein is 10 kcal mol^{-1} (42 kJ mol^{-1}), and thus each residue contributes on average only 0.1 kcal mol^{-1} (0.42 kJ mol^{-1}) of energy to maintain the folded state. This amount is less than that of thermal energy, which is 0.6 kcal mol^{-1} (2.5 kJ mol^{-1}) at room temperature. This meager stabilization energy means that correct intermediates, especially those formed early in folding, can be lost. The analogy is that the monkey would be somewhat free to undo its correct keystrokes. Nonetheless, the interactions that lead to cooperative folding can stabilize intermediates as structure builds up. Thus, local regions, which have significant structural preference, though not necessarily stable on their own, will tend to adopt their favored

structures and, as they form, can interact with one other, leading to increasing stabilization.

MOLECULAR MODELING: A METHOD FOR UNRAVELING PROTEIN STRUCTURE AND FUNCTION

Proteins are fundamental components of all living cells. They exhibit an enormous amount of chemical and structural diversity, enabling them to carry out an extraordinarily diverse range of biological functions. Proteins help us digest our food, fight infections, control body chemistry, and in general, keep our bodies functioning smoothly. Scientists know that the critical feature of a protein is its ability to adopt the right shape for carrying out a particular function. But sometimes a protein twists into the wrong shape or has a missing part, preventing it from doing its job. Many diseases, such as Alzheimer's and "mad cow", are now known to result from proteins that have adopted an incorrect structure.

Identifying a protein's shape, or **structure**, is key to understanding its biological function and its role in health and disease. Illuminating a protein's structure also paves the way for the development of new agents and devices to treat a disease. Yet solving the structure of a protein is no easy feat. It often takes scientists working in the laboratory months, sometimes years, to experimentally determine a single structure. Therefore, scientists have begun to turn toward computers to help predict the structure of a protein based on its sequence. The challenge lies in developing methods for accurately and reliably understanding this intricate relationship.

Levels of Protein Structure

To produce proteins, cellular structures called **ribosomes** join together long chains of subunits. A set of 20 different subunits, called **amino acids**, can be arranged in any order to form a **polypeptide** that can be thousands of amino acids long. These chains can then loop about each other, or **fold**, in a variety of ways, but only one of these ways allows a protein to function properly. The critical feature of a protein is its ability to fold into a conformation that creates structural features, such as surface grooves, ridges, and pockets, which allow it to fulfill its role in a cell. A protein's conformation is usually

Proteins function through their conformation.

described in terms of levels of structure. Traditionally, proteins are looked upon as having four distinct levels of structure, with each level of structure dependent on the one below it. In some proteins, functional diversity may be further amplified by the addition of new chemical groups after synthesis is complete.

The stringing together of the amino acid chain to form a polypeptide is referred to as the **primary structure**. The **secondary structure** is generated by the folding of the primary sequence and refers to the path that the polypeptide backbone of the protein follows in space. Certain types of secondary structures are relatively common. Two well-described secondary structures are the **alpha helix** and the **beta sheet**. In the first case, certain types of bonding between groups located on the same polypeptide chain cause the backbone to twist into a helix, most often in a form known as the alpha helix. Beta sheets are formed when a polypeptide chain bonds with another chain that is running in the opposite direction. Beta sheets may also be formed between two sections of a single polypeptide chain that is arranged such that adjacent regions are in reverse orientation.

The **tertiary structure** describes the organization in three dimensions of all of the atoms in the polypeptide. If a protein consists of only one polypeptide chain, this level then describes the complete structure.

Multimeric proteins, or proteins that consist of more than one polypeptide chain, require a higher level of organization. The **quaternary structure** defines the conformation assumed by a multimeric protein. In this case, the individual polypeptide chains that make up a multimeric protein are often referred to as the **protein subunits**. The four levels of protein structure are hierarchal, that is, each level of the build process is dependent upon the one below it.

How Do Proteins Acquire Their Correct Conformations?

A protein's primary amino acid sequence is crucial in determining its final structure. In some cases, amino acid sequence is the sole determinant, whereas in other cases, additional interactions may be required before a protein can attain its final conformation. For example, some proteins require the presence of a **cofactor**, or a second molecule that

is part of the active protein, before it can attain its final conformation. Multimeric proteins often require one or more subunits to be present for another subunit to adopt the proper higher order structure. In any case, as we stated earlier, the entire process is cooperative, that is, the formation of one region of secondary structure determines the formation of the next region.

Allosteric Proteins

Allosteric proteins can change their shape and function depending on the environmental conditions in which they are found.

Under certain conditions, a protein may have a stable alternate conformation, or shape, that enables it to carry out a different biological function. Proteins that exhibit this characteristic are called **allosteric**. The interaction of an allosteric protein with a specific cofactor, or with another protein, may influence the transition of the protein between shapes. In addition, any change in conformation brought about by an interaction at one site may lead to an alteration in the structure, and thus function, at another site. One should bear in mind, though, that this type of transition affects only the protein's shape, not the primary amino acid sequence. Allosteric proteins play an important role in both metabolic and genetic regulation.

Determining Protein Structure

Traditionally, a protein's structure was determined using one of two techniques: **X-ray crystallography** or **nuclear magnetic resonance (NMR) spectroscopy X-ray Crystallography**

When performing this technique, the molecule under

PROTEIN ENGINEERING

Crystals are a solid form of a substance in which the component molecules are present in an ordered array called a **lattice**. The basic building block of a crystal is called a **unit cell**. Each unit cell contains exactly one unique set of the crystal's components, the smallest possible set that is fully representative of the crystal. Crystals of a complex molecule, like a protein, produce a complex pattern of **X-ray diffraction**, or scattering of X-rays. When the crystal is placed in an X-ray beam, all of the unit cells present the same face to the beam; therefore, many molecules are in the same orientation with respect to the incoming X-rays. The X-ray beam enters the crystal and a number of smaller beams emerge: each one in a different direction, each one with a different intensity. If an X-ray detector, such as a piece of film, is placed on the opposite side of the crystal from the X-ray source, each diffracted ray, called a **reflection**, will produce a spot on the film. However, because only a few reflections can be detected with any one orientation of the crystal, an important component of any X-ray diffraction instrument is a device for accurately setting and changing the orientation of the crystal. The set of diffracted, emerging beams contains information about the underlying crystal structure

study must first be crystallized, and the crystals must be singular and of perfect quality – a time-consuming and difficult task.

If we could use light instead of X-rays, we could set up a system of lenses to recombine the beams emerging from the crystal and thus bring into focus an enlarged image of the unit cell and the molecules therein. But the molecules do not diffract visible light, and X-rays, unlike light, cannot be focused with lenses. However, the scientific laws that lenses obey are well understood, and it is possible to calculate the molecular image with a computer. In effect, the computer mimics the action of a lens.

The major drawback associated with this technique is that crystallization of the proteins is a difficult task. Crystals are formed by slowly precipitating proteins under conditions that maintain their native conformation or structure. These exact conditions can only be discovered by repeated trials that entail varying certain experimental conditions, one at a time. This is a very time consuming and tedious process. In some cases, the task of crystallizing a protein borders on the impossible.

Nuclear Magnetic Resonance (NMR) Spectroscopy

Solution NMR is performed on a solution of macromolecules while the molecules tumble and vibrate with thermal motion.

The basic phenomenon of **NMR spectroscopy** was discovered in 1945. In this technique, a sample is immersed in a magnetic field and bombarded with radio waves. These radio waves encourage the nuclei of the molecule to **resonate**, or spin. As the positively charged nucleus spins, the moving charge creates what is called a **magnetic moment**. The **thermal motion** of the molecule—the movement of the molecule associated with the temperature of the material—further creates a **torque**, or twisting force, that makes the magnetic moment "wobble" like a child's top. When the radio waves hit the spinning nuclei, they tilt even more, sometimes flipping over. These resonating nuclei emit a unique signal that is then picked up on a special radio receiver and translated using a decoder. This decoder is called the **Fourier Transform algorithm**, a complex equation that translates the language of the nuclei into something a scientist can understand. By measuring the frequencies at which different nuclei flip, scientists can determine molecular structure, as well as many other interesting properties of the molecule.

In the past 10 years, NMR has proven to be a powerful alternative to X-ray

crystallography for the determination of molecular structure. NMR has the advantage over crystallographic techniques in that experiments are performed in solution as opposed to a crystal lattice. However, the principles that make NMR possible tend to make this technique very time consuming and limit the application to small- and medium-sized molecules.

The Advent of Computational Modeling

Researchers have been working for decades to develop procedures for predicting protein structure that are not so time consuming and that are not hindered by size and solubility constraints. To do this, researchers have turned to computers for help in predicting protein structure from gene sequences, a concept called **homology modeling**. The

complete genomes of various organisms, including humans, have now been decoded and allow researchers to approach this goal in a logical and organized fashion.

Before we go any further, it is important to define some common terminology used in this field.

Some Basic Theory

It is theorized that proteins that share a similar sequence generally share the same basic structure. Therefore, by experimentally determining the structure for one member of a protein family, called a **target**, researchers have a model on which to base the structure of other proteins within that family. Moving a step further, by selecting a target from each superfamily, researchers can study the universe of protein folds in a systematic fashion and outline a set of sequences associated with each folding motif. Many of these sequences may not demonstrate a resemblance to one another, but their identification and assignment to a particular fold is essential for predicting future protein structures using homology modeling.

The scientific basis for these theories is that a strong conservation of protein three-dimensional shape across large evolutionary distances—both within single species, between species, and in spite of sequence variation—has been demonstrated again and again. Although most scientists choose high-priority structures as their targets, this theory provides the option to choose any one of the proteins within a family as the target, rather than trying to achieve experimental results using a protein that is particularly difficult to work with using crystallographic or NMR techniques.

Steps for Maximizing Results

Specific tasks must be carried out to maximize results when determining protein structure using homology modeling.

First, protein sequences must be organized in terms of families, preferably in a searchable database, and a target must be selected. Protein families can be identified and organized

by comparing protein sequences derived from completely sequenced genomes. Targets may be selected for families that do not exhibit apparent sequence homology to proteins with a known three-dimensional structure.

Next, researchers must generate a purified protein for analysis of the chosen target and then experimentally determine the target's structure, either by X-ray crystallography and/or NMR. Target structures determined experimentally may then be further analyzed to evaluate their similarity to other known protein structures and to determine possible evolutionary relationships that are not identifiable from protein sequence alone. The target structure will also serve as a detailed model for determining the structure of other proteins within that family. In favorable cases, just knowing the structure of a particular protein may also provide considerable insight into its possible function. provide considerable insight into its possible function.

PDB: The Protein Data Bank

The **PDB** was the first "bioinformatics" database ever built and is designed to store complex three-dimensional data. The PDB was originally developed and housed at the Brookhaven National Laboratories but is now managed and maintained by the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB is a collection of all publicly available three-dimensional structures of proteins, nucleic acids, carbohydrates, and a variety of other complexes experimentally determined by X-ray crystallography and NMR.

PDB is supported by funds from the National Science Foundation, the Department of Energy, and two units of the National Institutes of Health: the National Institute of General Medical Sciences and the National Library of Medicine.

Protein Modeling at NCBI

The Molecular Modeling Database

NCBI's **Molecular Modeling DataBase (MMDB)**, an integral part of our Entrez information retrieval system, is a compilation of all of the PDB three-dimensional structures of biomolecules. The difference between the two databases is that the MMDB records reorganize and validate the information stored in the database in a way that enables cross-referencing between the chemistry and the three-dimensional structure of macromolecules. By integrating chemical, sequence, and structure information, MMDB is designed to serve as a resource for structure-based homology modeling and protein structure prediction.

MMDB records have value-added information compared to the original PDB entries, including explicit chemical graph information, uniformly derived secondary structure definitions, structure domain information, literature citation matching, and molecule-based assignment of taxonomy to each biologically derived protein or nucleic acid chain.

NCBI has also developed a three-dimensional structure viewer, called **Cn3D**, for easy interactive visualization of molecular structures from Entrez. Cn3D serves as a visualization tool for sequences and sequence alignments. What sets Cn3D apart from other software is its ability to correlate structure and sequence information. For example, using Cn3D, a scientist can quickly locate the residues in a crystal structure that correspond to known disease mutations or conserved active site residues from a family of **sequence homologs**, or sequences that share a common ancestor. Cn3D displays structure-structure alignments along with the corresponding structure-based sequence alignments to emphasize those regions within a group of related proteins that are most conserved in structure and sequence. Cn3D also features custom labeling options, high-quality graphics, and a variety of file exports that together make Cn3D a powerful tool for literature annotation.

PDBeast: Taxonomy in MMDB

Taxonomy is the scientific discipline that seeks to catalog and reconstruct the evolutionary history of life on earth. NCBI's Structure Group, in collaboration with NCBI's taxonomists, has undertaken **taxonomy annotation** for the structure data stored

in MMDB. A semi-automated approach has been implemented in which a human expert checks, corrects, and validates automatic taxonomic **PDBeast: Taxonomy in MMDB**

Taxonomy is the scientific discipline that seeks to catalog and reconstruct the evolutionary history of life on earth. NCBI's Structure Group, in collaboration with NCBI's taxonomists, has undertaken **taxonomy annotation** for the structure data stored in MMDB. A semi-automated approach has been implemented in which a human expert checks, corrects, and validates automatic taxonomic

assignments.

The **PDBeast** software tool was developed by NCBI for this purpose. It pulls text-descriptions of "Source Organisms" from either the original PDB-Entries or user-specified information and looks for matches in the NCBI Taxonomy database to record taxonomy assignments.

COGs: Phylogenetic Classification of Proteins

The database of **Clusters of Orthologous Groups** of proteins (COGs) represents an attempt at the **phylogenetic classification** of proteins— a scheme that indicates the evolutionary relationships between organisms—from complete genomes. Each COG includes proteins that are thought to be **orthologous**. Orthologs are genes in different species derived from a common ancestor and carried on through evolution. COGs may be used to detect similarities and differences between species for identifying protein families and predicting new protein functions and to point to potential drug targets in disease-causing species. The database is accompanied by the **COGnitor** program, which assigns new proteins, typically from newly sequenced genomes, to pre-existing COGs. A Web page containing additional structural and functional information is now associated with each COG. These hyperlinked information pages include: systematic classification of the COG members under the different classification systems; indications of which COG members (if any) have been characterized genetically and biochemically; information on the domain architecture of the proteins constituting the COG and the three-dimensional structure of the domains if known or predictable; a succinct summary of the common

structural and functional features of the COG members as well as peculiarities of individual members; and key references.

Detecting New Sequence Similarities: BLAST against MMDB

Comparison, whether of structural features or protein sequences, lies at the heart of biology. The introduction of **BLAST**, or The Basic Local Alignment Search Tool, in 1990 made it easier to rapidly scan huge databases for overt **homologies**, or sequence similarities, and to statistically evaluate the resulting matches. BLAST works by comparing a user's unknown sequence against the database of all known sequences to determine likely matches. Sequence similarities found by BLAST have been critical in several gene discoveries. Hundreds of major sequencing centers and research institutions around the country use this software to transmit a query sequence from their local computer to a BLAST server at the NCBI via the Internet. In a matter of seconds, the BLAST server compares the user's sequence with up to a million known sequences and determines the closest matches.

Not all significant homologies are readily and easily detected, however. Some of the most interesting are subtle similarities that do not always rise to statistical significance during a standard BLAST search. Therefore, NCBI has extended the statistical methodology used in the original BLAST to address the problem of detecting weak, yet significant, sequence similarities. **PSI-BLAST**, or **Position-Specific Iterated BLAST**, searches sequence databases with a profile constructed using BLAST alignments, from which it then constructs what is called a position-specific score matrix. For protein analysis, the new **Pattern Hit Initiated BLAST**, or **PHI-BLAST**, serves to complement the profile-based searching that was previously introduced with PSI-BLAST. PHI-BLAST further incorporates hypotheses as to the biological function of a query sequence and restricts the analysis to a set of protein sequences that is already known to contain a specific pattern or motif.

Structure Similarity Searching Using VAST

As just noted, a sequence-sequence similarity program provides an alignment of two sets of sequences. A structure-structure similarity program provides a three-dimensional structure superposition. Structure similarity search services are based on the premise that some measure can be computed between two structures to assess their similarities, much the same way a BLAST alignment is predicted.

VAST, or the **V**ector **A**lignment **S**earch **T**ool, is a computer algorithm developed at NCBI for use in identifying similar three-dimensional protein structures. VAST is capable of detecting structural similarities between proteins stored in MMDB, even when no sequence similarity is detected.

VAST Search is NCBI's structure-structure similarity search service that compares three-dimensional coordinates of newly determined protein structures to those in the MMDB or PDB databases. VAST Search creates a list of structure neighbors, or related structures, that a user can then browse interactively. VAST Search will retrieve almost all structures with an identical three-dimensional fold, although it may occasionally miss a few structures or report chance similarities.

The Conserved Domain Database

The Conserved Domain Database (CDD) is a collection of sequence alignments and profiles representing protein domains conserved in molecular evolution. It includes domains from SMART and Pfam, two popular Web-based tools for studying sequence domains, as well as domains contributed by NCBI researchers. **CD Search**, another NCBI search service, can be used to identify conserved domains in a protein query sequence. CD-Search uses RPS-BLAST to compare a query sequence against specific matrices that have been prepared from conserved domain alignments present in CDD. Alignments are also mapped to known three-dimensional structures and can be displayed using Cn3D (see above).

Conserved Domain Architecture Retrieval Tool

PROTEIN ENGINEERING

NCBI's **Conserved Domain Architecture Retrieval Tool (CDART)** displays the functional domains that make up a protein and lists other proteins with similar domain architectures. CDART determines the domain architecture of a query protein sequence by comparing it to the CDD, a database of conserved domain alignments, using **RPS-BLAST**

The Conserved Domain Architecture Retrieval Tool then compares the protein's domain architecture to that of other proteins in NCBI's non-redundant sequence database. Related sequences are identified as those proteins that share one or more similar domains. CDART displays these sequences using a graphical summary that depicts the types and locations of domains identified within each sequence. Links to the individual sequences as well as to further information on their domain architectures are also provided. Because protein domains may be considered elementary units of molecular function and proteins related by domain architecture may play similar roles in cellular processes, CDART serves as a useful tool in comparative sequence analysis.

The Conserved Domain Architecture Retrieval Tool then compares the protein's domain architecture to that of other proteins in NCBI's non-redundant sequence database. Related sequences are identified as those proteins that share one or more similar domains. CDART displays these sequences using a graphical summary that depicts the types and locations of domains identified within each sequence. Links to the individual sequences as well as to further information on their domain architectures are also provided. Because protein domains may be considered elementary units of molecular function and proteins related by domain architecture may play similar roles in cellular processes, CDART serves as a useful tool in comparative sequence analysis.

- **Folding motifs** are independent folding units, or particular structures, that recur in many molecules.
- **Domains** are the building blocks of a protein and are considered elementary units of molecular function.

- **Families** are groups of proteins that demonstrate sequence homology or have similar sequences.

Superfamilies consist of proteins that have similar folding motifs but do not exhibit sequence similarity.