

PREDICTING PROTEIN TERTIARY STRUCTURE:

The ROSETTA Approach as an Example

Literature Seminar

Raluca Craciun

Seminar advisor: Dr. Kevin Redding

May 04, 2006

Shelby Hall, Room 151

Introduction

The term “protein” was used for the first time by Berzelius¹, who wrote in a letter: “The name protein that I propose for the organic oxide of fibrin and albumin, I wanted to derive from the Greek word *protas* (of primary importance), because it appears to be the primitive or principal substance of animal nutrition “. Proteins perform a wide variety of functions in an organism: they can be transporters (myoglobin), defensive (antibodies), catalysts (enzymes), signaling (hormones) or structural molecules (keratin, collagen). The function of a protein is dictated by its three dimensional structure, but proteins start out as linear polymers of amino acids when synthesized in cells. Hydrogen bonds between amino acid residues lead to organization of the polypeptide chain into regular, repetitive patterns: elements of secondary structure (α -helices, β -sheets, and turns). Through further interactions (*e.g.* burial hydrophobic groups, formation of salt and disulphide bridges between amino acid side groups), the elements of secondary structure form a unique, 3D shape known as the native conformation of the protein, responsible for its function. This process is called folding.

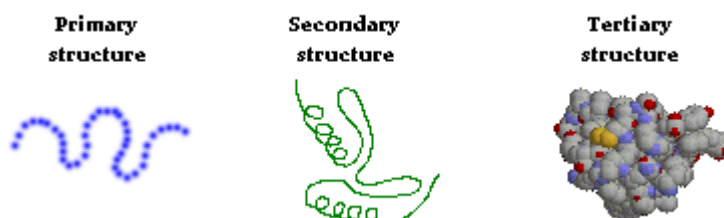


Fig.1 Protein folding pathway

It is important to understand the relationship between sequence, structure and function. This could allow design of synthetic protein sequences to perform specific functions. Drugs², which usually have undesirable side-effects, have been discovered by

trial-and-error testing thousands of compounds for their ability to interfere with the function of the protein. Thus, we need to know the three-dimensional structure of the proteins. A high level of detail and accuracy can be achieved by experimental methods, such as X-ray crystallography and NMR. The prediction of protein structures has been a long-standing problem in computational biology and theoretical chemistry, and the methods that have been developed although not at the same level of accuracy, have improved considerably over the years.

Experimental Methods

All solved protein tertiary structures are deposited into a database. The RCSB (Research Collaboratory for Structural Bioinformatics) Protein Data Bank contains the 3D structures of the proteins determined by various experimental methods. The number of structures deposited has exploded in recent years with the improvement of experimental techniques. As of April 04, 2006 there were 35917 structures, for which information about the authors, method, resolution, and coordinates can be accessed freely at <http://www.rcsb.org/pdb>. The two experimental techniques used to determine the tertiary structures of proteins are X-ray crystallography and NMR spectroscopy.

X-ray crystallography was the technique used to solve for the first time the three dimensional structure of a protein, sperm whale myoglobin³. When X-rays hit a protein crystal they are diffracted causing diffraction patterns that can be converted into electron density maps. From interpreting this maps and considering the chemical sequence, stereochemistry and the nature of any bound ligands a model can be built. The steps involved in crystal structure determination are: protein preparation, crystallization, crystal testing, data collection, model building and refinement. The unique conditions for each

protein needed for growing crystals and their fragile nature make protein crystallization inherently difficult. Errors might arise when the protein structure is positioned in the electron density⁴.

NMR spectroscopy yields structures at a lower resolution compared to X-ray, but offers the advantage of recording data in solution. Thus, temperature, pH, salt concentration can be modified so as to resemble a given physiological fluid. Also, for the case of partially folded proteins, it has been the method of choice since these are difficult to crystallize⁵. The result of the experiment is a set of lower and upper limits, called constraints, for the distance between a pair of atoms. The number of constraints is related to the flexibility of the protein in solution. A sufficient number of constraints lead to a finite number of possible conformations of the protein².

Theoretical Methods

The self-assembly of protein molecules with a huge number of degrees of freedom into a unique three-dimensional shape that performs a specific function is an example of biological self-organization⁶. The number of gene sequences in public and private databases increase at a much faster rate than the number of solved protein structures⁶. Experimental methods are time consuming and expensive. It is believed that approximately one quarter of known protein sequences have a known function and another quarter could be related to proteins of known structure through sequence homology⁷. So, determination of the structures and function of the encoded proteins could be immensely aided by computational methods of structure prediction. Methods for predicting protein structure are classified² according to the relationship between the target protein and proteins with known structure in databases. Thus, *homology modeling* methods are applied when a clear

evolutionary relationship between the target and a protein of known structure can be detected from the sequence. *Fold recognition* methods can be used when the structure of the target protein turns out to be related to that of a protein with known structure, although the relationship is difficult or impossible to detect from the sequences. Although of lower quality than homology modeling methods, fold recognition methods can sometimes detect remote evolutionary relationship. When neither of these apply, methods used for predicting the structure are classified as techniques for new folds or ab initio methods.

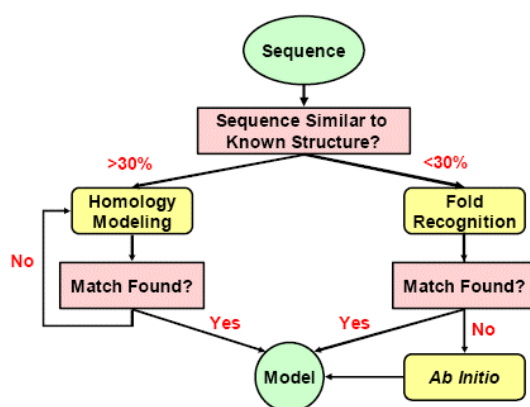


Fig. 2 Theoretical methods

The pioneering work of Anfinsen⁸ showed that if a protein is gently denatured and then returned to normal physiological conditions of temperature, pH, salt concentration, etc., it spontaneously regains its function, and therefore its structure. This led to the conclusion that there is sufficient information contained in the protein sequence alone to allow correct folding from any of a large number of unfolded states. However, this raised the question whether the polypeptide can sample the enormous number of possible conformations available to it⁹. If at least two possible conformations are assumed for each residue, then the chain would need approximately 10^{10} years to sample all conformations

and find the most stable structure. Nevertheless proteins can fold in the timescale of seconds or less. It is not completely understood how the amino acid sequence determines the folding of the protein. This puzzle is called the Protein Folding Problem.

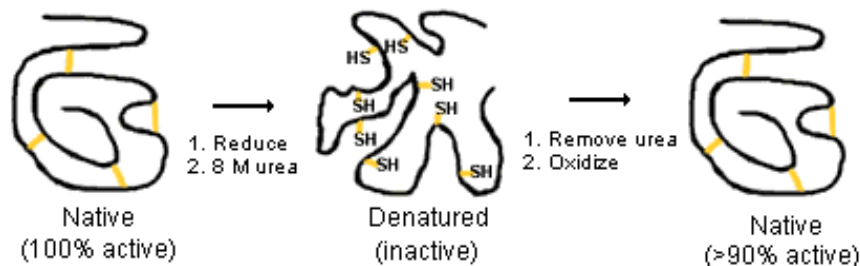


Fig. 3 Schematic diagram of Anfinsen’s experiment showing chemical denaturation and subsequent renaturation of bovine pancreatic ribonuclease.

Methods that try to predict the 3D structure of a protein given only its amino acid sequence are called *ab initio* methods. The name suggests, in analogy with quantum mechanics, that only the laws of physics are used¹⁰. However most successful methods use a blend of physics, fold recognition, and statistical probability. Since these methods are using knowledge from many chemical studies of model compounds and are not making predictions on the basis of first principles of quantum mechanics, the traditional term “*ab initio*” has been replaced with “new” or “novel” fold¹¹. It is generally assumed that a protein sequence folds to a native conformation that is at or near the global free-energy minimum¹². Thus, finding the native conformation requires the development of an accurate potential and of an efficient protocol for searching the vast conformational space.

As the theoretical methods were developed, they were tested for the ability to predict a structure that was already known¹³. CASP (Critical Assessment of Structure Prediction), a competition that started in 1994, is intended to identify the progress made in the area of protein structure prediction and the future work that needs to be done.

Experimentalists who are about to determine the structure of a protein are asked to make available the sequence of a protein. Predictors produce and deposit models for these proteins before the structures are published. Three assessors compare the model with the native structures as soon as they are available. The results are made available to the whole scientific community, via World Wide Web and by publication of a special issue of the journal *Proteins: Structure, Function, and bioinformatics*. The experiment is repeated every two years, with the last edition (CASP6) being held in 2004¹⁴.

One of the most successful “ab initio” approaches to the prediction of protein structure has proven to be the ROSETTA method developed by the group of David Baker from the University of Washington, Seattle. ROSETTA is a fragment-based method that deals with vast conformational space by pre-selecting structural fragments from a library of solved protein structures. It is a method used when the target protein is considered to be a novel fold (i.e. if the other methods could not detect any evolutionary relationship to proteins with known structures). Of course, as the number of solved structures deposited in databases increases, the probability of finding related proteins also increases, but the number of occasions where neither homology modeling nor fold recognition can be applied is still sizeable².

The ROSETTA method is based on the experimental observation that the preferences of local sequences to adopt certain secondary structures bias, but do not uniquely define the local structure of a protein¹⁵. The underlying principle is that the distribution of conformations sampled for a local segment of the chain is reasonably approximated by the distribution of structures adopted by that sequence, and by closely

related sequences, in known protein structures. The general procedure is to collect structures adopted by short sequence segments in known three-dimensional structures and then use their combination to produce a large number of possible 3D models for the target. A scoring function that discriminates between the many assemblies compatible with the distribution of local structures, is used to choose the final model on the basis of energy considerations. The steps involved in ROSETTA are:

- Split the sequence into fragments
- Search database of known structure for regions with similar sequence and select fragments
- Assemble fragments

The method starts with the fully extended chain, which is split into 9-residue fragments. This length was chosen, because it was found that the correlation between the local sequence and local structure is stronger for fragments 9 residues long than it is for other lengths (of <15 residues). For each 9-residue segment in the protein being folded, a number of 25 nearest-sequence neighbors in the structure database are chosen. These are classified as nearest based on a distance measure that compares the amino acid frequency distributions at each position in the two segments.

$$DISTANCE = \sum_i^9 \sum_{aa}^{20} |S(aa,i) - (X(aa,i))|$$

where $S(aa,i)$ and $X(aa,i)$ are the frequency of amino acid aa at position i in 9-residue segments of either the sequence being folded (S) or of one of the proteins in database (X). All of these nearest-neighbor sequences are taken from the database of proteins of known structures, composed of X-ray structures with resolution of 2.5 Å or better and <50% sequence identity. The assembly of these generated fragments into protein-like structures

is done by using a simulated annealing Monte Carlo search with Metropolis criterion¹⁶. The search starts arbitrarily and a 9-residue fragment in the target is replaced by a randomly chosen fragment from those 25 nearest neighbors. The torsion angles in the target chain are also replaced with torsion angles from the selected fragment. For the resulting conformation, the energy is evaluated and moves that decrease the energy are retained. If the new energy is higher, the move is accepted with probability $\exp(\Delta E/T)$, where T is the artificial temperature of the "annealing" process. T is gradually reduced from 2500 to 10 linearly over 10,000 attempted moves. The final conformation is called a decoy and the process is repeated many times until a sufficient number (on the order of thousands) of decoys are generated. The scoring function is derived based on Bayesian statistics.

$$P(struct | seq) = \frac{P(seq | struct)P(struct)}{P(seq)} \propto P(seq | struct)P(struct)$$

$P(struct|seq)$ is the probability of observing the specific structure given the sequence of amino acids, $P(seq|struct)$ is the probability of observing the specific sequence given the structure, $P(seq)$ is the probability of observing the sequence (always set to one), and $P(struct)$ is the prior probability of observing the structure. $P(seq|struct)$ is affected by hydrophobic burial and residue pair interactions and is given by the following expression¹⁵:

$$P(seq | struct) \cong \prod_{i < j} \frac{P(r_{ij} | aa_i aa_j)}{P(r_{ij})}, \text{ where } r_{ij} \text{ is the distance between residues } i \text{ and } j \text{ and}$$

the score is the product over all pairs of residues. $P(struct)$ is affected by the 3-D structure only and depends on helix-strand packing, strand-strand packing, sheet configuration and van der Waals interactions. $P(struct)$ is given by the expression:

$P(\text{struct}) \sim \exp(-\text{radius of gyration}^2)$, where the radius of gyration is a property characterizing the size of a particle of any shape.

The resulting conformations that have properties inconsistent with known protein structures are eliminated. All of the remaining decoys are clustered, with the assumption that there will be more conformations near the native structure than any other minimum. The "distance" between two decoys is given by C_{α} -RMSD. The decoy with 100 closest neighbors is located and the distance to the 100th closest neighbor (or 0.3 nm, whichever is greater) is used as a threshold for the cluster¹⁷. The decoy that has the largest number of neighbors within the threshold distance is identified as the top cluster center. All the other members in the cluster are removed and the process is repeated for another cluster. The top 5 cluster centers are the top ranked models and are submitted as predictions for a certain target at CASP competition.

The first edition of CASP in which ROSETTA was used to predict protein structures was CASP3 (1998)¹⁸. For a number of targets the predictions were not correct. This was the case for proteins containing β -sheets or having a large number of residues (>150). Successful predictions were made for smaller proteins, some of them being the best of all other models submitted. For example, the model submitted for the transcriptional activator, MarA (target 79), predicted a 99-residue segment (out of 129 residues) with an C_{α} -RMSD of 6.4 Å. RMSD is the root mean square deviation of alpha-carbon coordinates after optimal superposition of the predicted and experimentally determined structure. The model was able to approximate the function. The prediction of two α helix-loop- α helix domains, a motif seen in many DNA-binding proteins, suggested

that it binds DNA (Fig. 4). Also, the model had a sequence match with 3 other known proteins, all of which bind DNA.

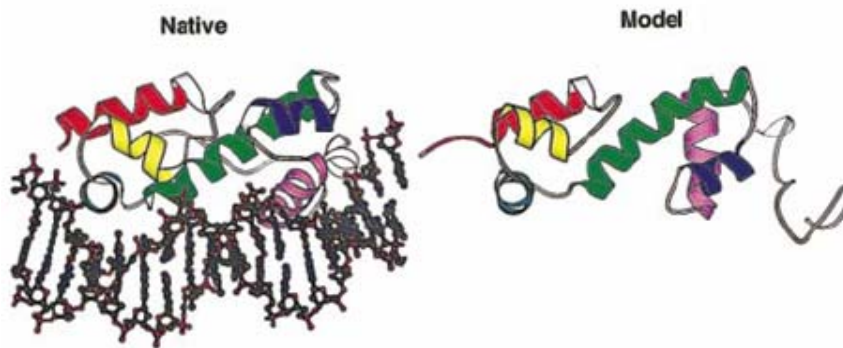


Fig. 4 Left: crystal structure of MarA, target 79 bound to double-stranded DNA.

Right: Best submitted model using ROSETTA

These results encouraged the authors of ROSETTA method to continue improving it, especially the search strategy for proteins larger than 100 residues. Over the next editions of CASP, ROSETTA made noticeable improvements, and its predictions were consistently among the best. At CASP6 (2004), it was able to predict the folded structure of a 70-residue α - β protein from *Thermus thermophilus* with a C_{α} -RMSD of 1.6 Å, making it the most accurate de novo structure prediction in the history of CASP (Fig. 5)¹⁹.

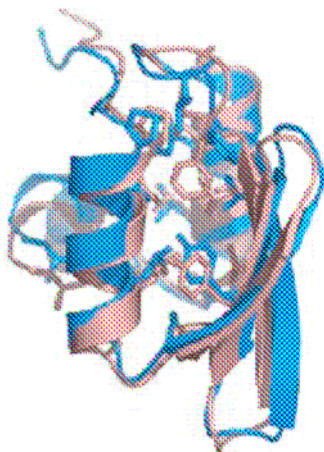


Fig. 5 Blind structure prediction for CASP6 target T0281, a protein of unknown function from *Thermus thermophilus*. Model (blue) and crystal structure (pink; PDB code 1WHZ) showing core side chains.

Conclusions

Experimental techniques offer the best picture of a protein's tertiary structure, but can be time consuming. It is believed that the PDB database will continue to double every three years, but the number of polypeptide sequences coming from genome projects will continue to double every few months. Thus the answer lies in methods for predicting the three-dimensional structure. The CASP competition has monitored the progress made in this area. Homology modeling and fold recognition methods have improved over the years as the PDB increased, since they rely on sequence or structure similarity with proteins of known structures. Ab initio methods hold the key for proteins that show new folds. ROSETTA, a mainly ab initio structure prediction algorithm has proven successful over the years.

References:

1. Vickery, H.B., *Yale J.Biol.Med* **1950**, 22, 387-393
2. Tramontano, A. *Protein Structure Prediction*; Wiley-VCH, Weinheim, 2006
3. Kendrew, J.C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D. C., Shore, V. C. *Nature* **1960**, 185, 422–427.
4. Borgstahl, G.E., Williams, D.R., Getzoff, E.D. *Biochemistry* **1995**, 34, 6278-6287
5. Wuthrich, K. *Les Prix Nobel, The Nobel Prize*; Tore Frängsmyr : Stockholm, 2003
6. Baker, D. *Nature* **2000**, 405, 39-42
7. Kelley, L.A., MacCallum, R.M., Sternberg, M.J. *J. Mol. Biol.* **2000**, 299, 499-520
8. Anfinsen, C. *Science*, **1973**, 181, 223-230

9. Levinthal, C. *Mossbauer Spectroscopy in Biological Systems*; J.C.M. Tsibiris :Univ of Illinois Press 1969
10. Kretsinger, R.H., Ison, R.E., Hovmöller, S. *Methods in Enzymology* **2004**, 383, 1-27
11. Osguthorpe, D. J *Current Opinion in Structural Biology* **2000**, 10, 146-152
12. Bonneau, R., Baker, D. *Annu. Rev. Biomol. Struct.* **2001**, 30, 173-189
13. Mount, D. W. *Bioinformatics .Sequence and Genome analysis*; Cold Spring Harbor, New York, 2001
14. <http://predictioncenter.org/>
15. Simons, K. T., Koopeberg, C., Huang, C., Baker, D. *J.Mol.Biol* **1997**, 268, 209-225
16. Rohl, C., Strauss, C.E.M., Misura, K.M.S., Baker, D. *Methods in Enzymology* **2004**, 383, 67-93
17. Bonneau, R., Strauss, C.E.M., Rohl, C., Chivian, D., Bradley, P., Malmström, L., Robertson, T., Baker, D. *J.Mol.Biol* **2002**, 322, 65-78
18. Simons, K. T., Bonneau, R., Ruczinski, I., Baker, D. *Proteins: Structure, Function and Genetics* **1999**, 37, 171-176
19. Bradley, P., Misura, K.M.S., Baker, D. *Science* **2005**, 309, 1868-1871