From 1D to 3D: Cooperative Determination of a Protein's Structure from its Sequence Using Comparative and De Novo Methods

by
RAY LUO
Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720

Table of Contents

List of Figures

From 1D to 3D: Cooperative Determination of a Protein's Structure from its Sequence Using Comparative and De Novo Methods

by Ray Luo

Abstract

Proteins are crucial in transport, catalysis, regulation, and other functions of the human body. Constructed out of one-dimensional strings of amino acids, proteins are transformed into three-dimensional structures in chemically favorable steps in a process known as protein folding (Figure 1). The proliferation of protein sequence data and the relative poverty of known protein structures have stimulated the need for efficient and flexible computational methods for predicting protein structure from its amino acid sequence. Expensive and time-consuming experimental protein structure determination techniques are being replaced by cheap and fast computational methods, two types of which are dominating current research. One set of methods called comparative alignment compares new sequences to existing sequences with known structures in protein databases. Another set of methods called de novo simulation searches for a stable three-dimensional configuration by calculating thermodynamic quantities for amino acid interactions, thus modeling the folding of a protein. We present a new cooperative approach that utilizes both comparative and de novo methods, reinforcing one method with the results of the other. Sequence comparisons are performed using a dynamic programming alignment algorithm, while simulation of protein folding is done using experimental data and Monte Carlo sampling. Combining sequence alignment with folding simulation, the Cooperative Structure Determination algorithm takes advantage of both protein databases and thermodynamic data.

From 1D to 3D: Cooperative Determination of a Protein's Structure from its Sequence Using Comparative and De Novo Methods

1.0 Introduction

Proteins are responsible for transport, support, regulation, and catalysis in the cell. In order to design proteins that carry out specific functions in the body, we must understand the way its three-dimensional (3D) structure interacts with molecules around it. Driven by large-scale efforts such as the Human Genome Project and the Mouse Genome Sequence Project, DNA and protein sequence data are now accumulating faster than ever. The proliferation of nucleotide sequence information for DNA and amino acid sequence information for proteins, however, has not been matched by an equally rapid rate of 3D structure determination (2:801).

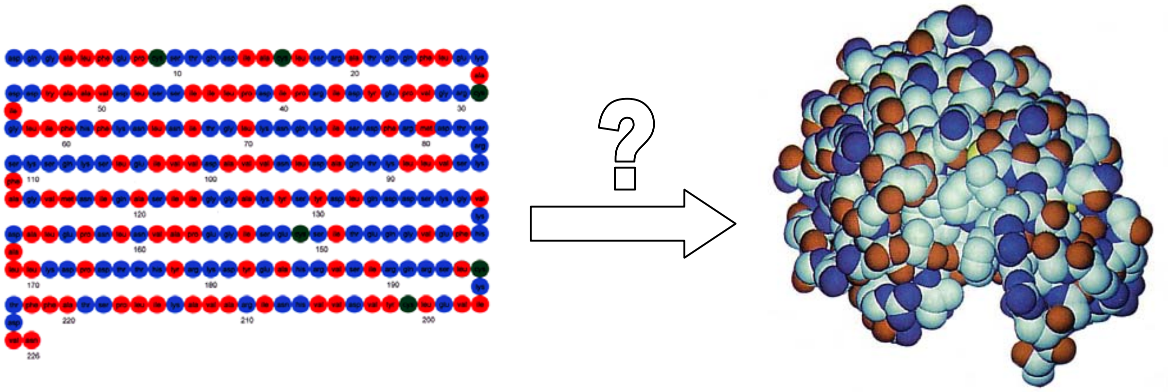


Figure 1. The problem of going from the amino acid sequence to the 3D structure of a protein. Shown here is the sequence and structure for the protein lysozyme, which destroys specific molecules on the surface of invading bacteria by recognizing its 3D configuration. (Source: Campbell, N. A., J. B. Reece, and L. G. Mitchell. *Biology*, 5th ed. CA: Addison Wesley Longman, 1999, pp. 71.)

Due to the unprecedented level of interest in applying protein analysis to drug design and to understanding the chemical basis of diseases involving protein abnormality, fourteen other competing structural genomics centers have been established across North America, all aiming to bridge the gap between protein sequence and protein structure (2:804). Before we can design proteins that have a specific function, we must first understand the mapping between protein sequence and protein structure, for the only way to produce a protein is by manufacturing an amino acid sequence while the only way the protein can function is by interacting with molecules in its 3D form. In order to more quickly achieve our long term goal of designing proteins for use in treating diseases, we must be able to rapidly and accurately identify the structure and function of a protein from its string of amino acid constituents (Figure 1).

Experimental methods for identifying 3D protein structure, including Nuclear Magnetic Resonance (NMR) spectroscopy and X-ray crystallography, are expensive and time consuming (6:316). Computational methods, on the other hand, can be run on a computer, requiring less maintenance and less expensive instruments. Moreover, efficient algorithms have been developed for processing more than one protein sequences at once, making structure determination less time consuming. Computational algorithms require either a database of protein sequence information or a statistical model of the protein folding process. Protein sequence information is widely and freely available from sources such as the Protein Data Bank (PDB) and the Swiss Protein Database. Meanwhile, powerful and flexible statistical models with efficient inference algorithms have been developed for analyzing sequences on a computer (5:2). With these two sources of knowledge, we can use computers to quickly achieve the goal of structure determination instead of running many expensive chemistry experiments. From the newly sequenced amino acid string of a protein, we can determine its structure by comparing the string against proteins in databases and by modeling amino acid interaction.

We begin with a background introduction to the relevant molecular biology of proteins and how they fold, along with a computational formulation of the structure prediction problem. Then we will assess the benefits and costs of current approaches to the problem. Next comes a discussion of our proposed solution combining a sequence alignment algorithm with the simulation of protein folding. After describing methods for evaluating our technique, we summarize our approach and provide directions for future research.

2.0 From Sequence to Protein Structure

Proteins are a type of macromolecule that is responsible for many of the functions of cells, such as transport, regulation, catalysis, support, and movement. Proteins are polymers of amino acids joined together by peptide bonds in a linear sequence. An amino acid consists of a central carbon atom (C) attached to an amino group (NH_2), a carboxyl group (COOH), and one of twenty different side chains. The side chains of amino acids determine its chemical properties. The twenty different side chains can be categorized into electrically charged, polar, hydrophobic, and sulfur-containing varieties. A sequence of amino acids is called a polypeptide. A protein consists of one or more polypeptides. In addition to the side chains of their amino acids, proteins are also distinguished by molecules that attach to one or more polypeptides, known as prosthetic groups. The most prominent prosthetic group is the disulfide bridge, which results from a covalent sulfur-sulfur bond joining two cysteine amino acids.

Our genetic material is contained in the molecule DNA, consisting of a sequence of nucleotide bases, and situated in the nucleus of a cell (3:78). When the cell needs to manufacture

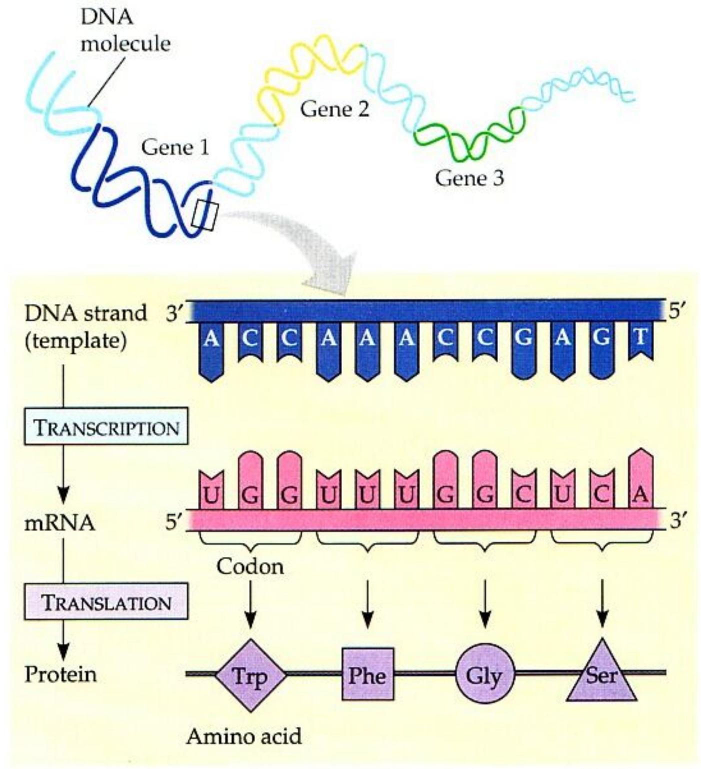


Figure 2. Production of the amino acids of a protein from the genes specifying their sequence. DNA is transcribed onto RNA, which is translated into amino acids. (Source: Campbell, N. A., J. B. Reece, and L. G. Mitchell. *Biology*, 5th ed. CA: Addison Wesley Longman, 1999, pp. 298.)

a certain protein, the nucleotide bases of a DNA strand is copied onto an RNA molecule in a process known as transcription (Figure 2). Transported outside the nucleus, the messenger RNA molecule is translated by an organelle called ribosome. Ribosome constructs the amino acid sequence, or the primary structure, of a protein under the direction of RNA. Each sequence of three nucleotide bases codes for one amino acid. The completed linear amino acid sequence cannot yet function as a protein, because it cannot interact with the molecules in its environment. To become a functional protein, the polypeptide must utilize all the information encoded in its amino acid sequence to fold into its 3D structure. The conformational changes that take place as a protein folds are guided by a chaperone, a special type of protein that protects its surrogate protein from unwanted chemical reactions with neighboring molecules as it folds.

The extent of protein folding can be divided into the formation of secondary and tertiary structures. As Figure 3 shows, secondary structure refers to one of the several repeated motifs

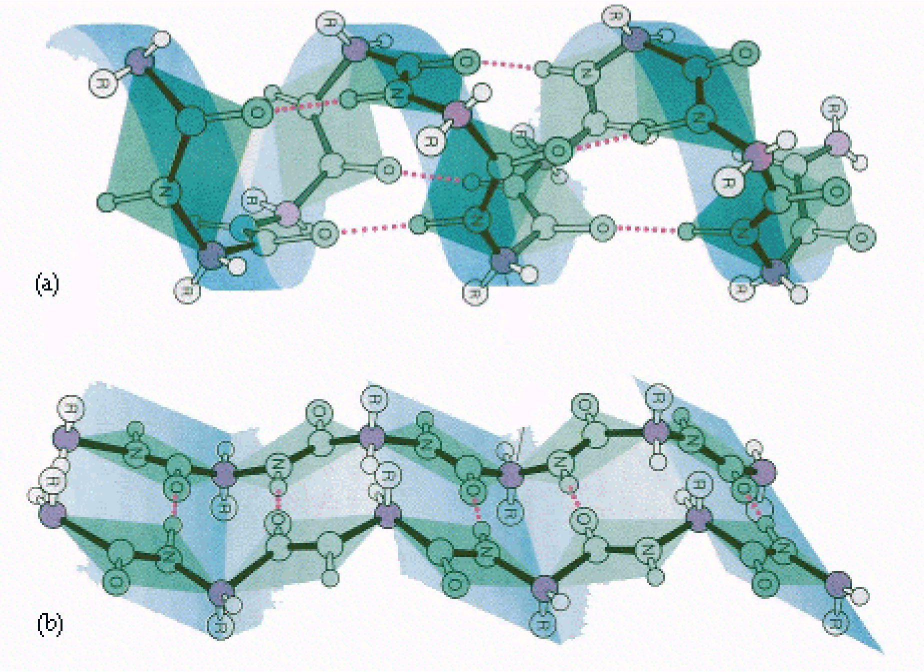


Figure 3. Secondary structure of proteins is the result of hydrogen bonding. (a) Alpha helix is formed when side chains protrude outward from the polypeptide. (b) Beta sheet is formed when two or more polypeptides are aligned side-by-side. (Source: “Secondary Structure.” http://sosnick.uchicago.edu/precpsecstru.html.)

polypeptides can fold into. Formed by hydrogen bonding between near-by amino acids, secondary structure captures the local patterns of interaction at an area in the surface of the protein. Alpha helix and beta sheet are two of the most ubiquitous structural motifs (Figure 3). The bending and folding of the primary or secondary structure of a protein constitutes its tertiary structure. The tertiary structure captures the global patterns of interaction between a protein and the molecules on which it acts. Prosthetic groups like the iron heme group and the disulfide bridge are parts of the tertiary structure that help stabilize the protein. Finally, the combination of two or more polypeptide chains in a multi-subunit protein is known as its quaternary structure. The quaternary structure represents complex interactions that take place in a protein linking together multiple polypeptides. Figure 4 shows the different levels of protein 3D structure (4:920). We are interested in providing a description of the tertiary and quaternary structure of a

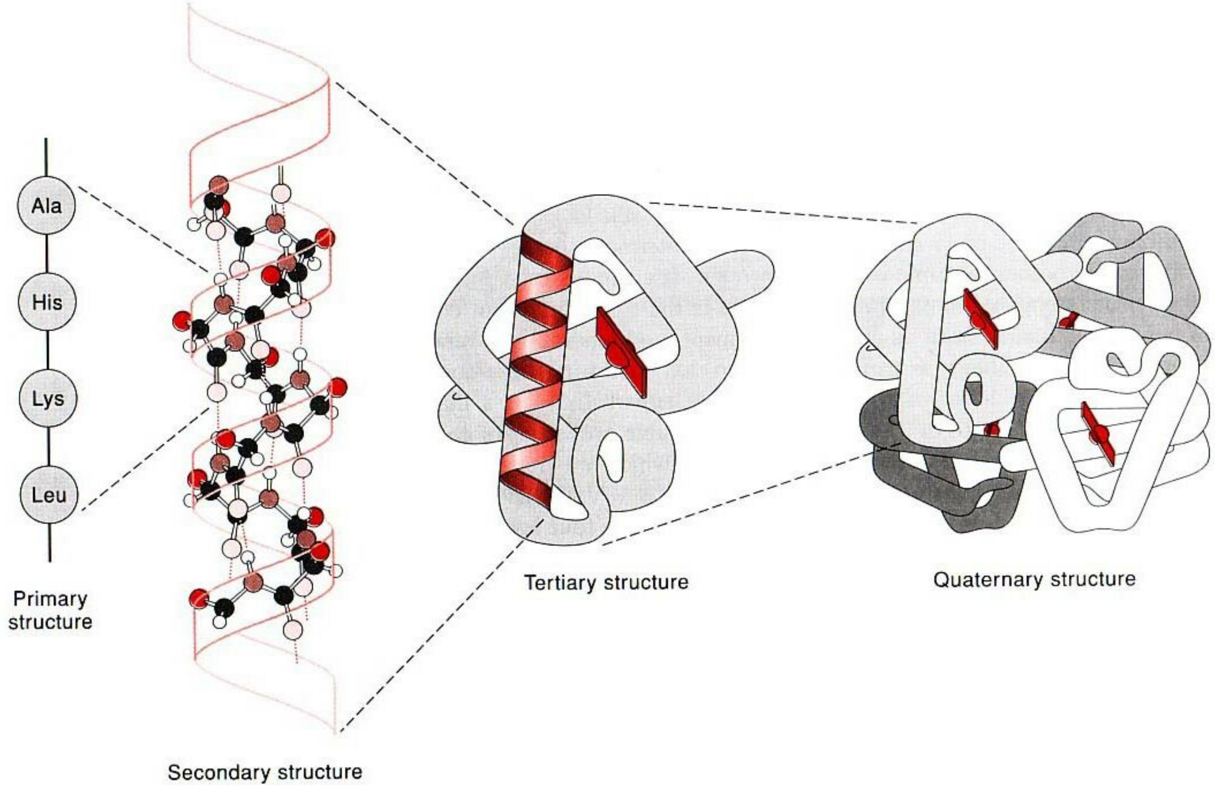


Figure 4. Different levels of protein structure. Primary structure is the amino acid sequence. Secondary structure refers to local structural motifs (alpha helix here). Tertiary structure is the bending and folding of secondary structure. Quaternary structure links multiple subunits. (Source: Change, R. *Physical Chemistry for the Chemical and Biological Sciences*, 3rd ed. CA: University Science Books, 2000, pp. 920.)

novel protein given only the primary structure.

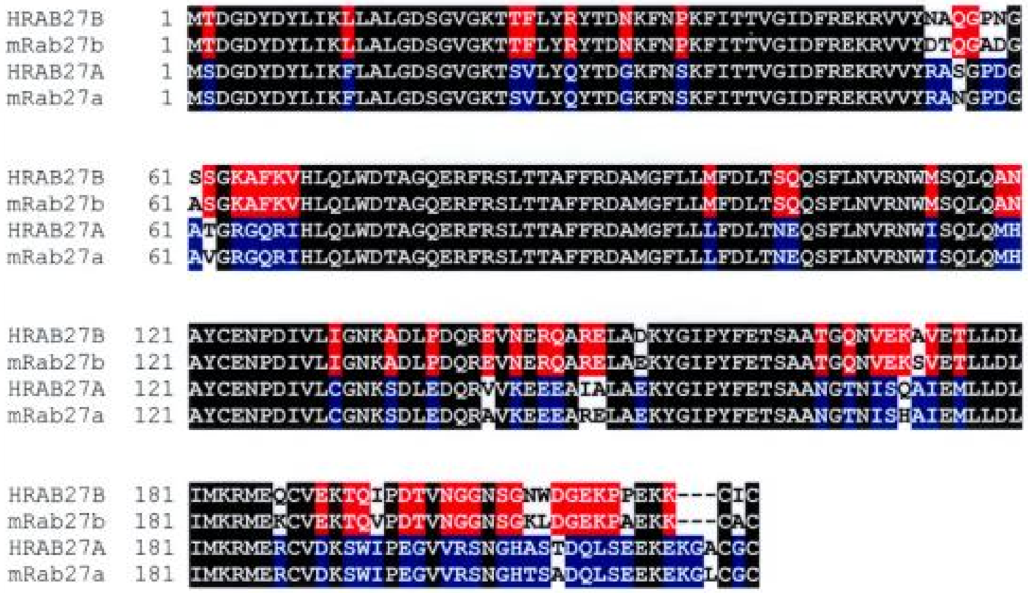
Designing proteins with specific patterns of interaction with its neighbors requires the knowledge of its 3D structure. Since the production of proteins starts with their amino acid sequence, we need to predict the tertiary or quaternary structure of a protein based on its primary structure. Protein structure determination can be accomplished using computation alone, without expensive and exhaustive experimental analysis. We need a cheap, efficient, and flexible computational technique for predicting protein structure. Using information gathered from protein databases and research on the chemistry of protein folding, we can proceed either by comparing our sequence against a set of proteins with known structures or hypothesize a physical model for the folding of our sequence into a 3D configuration.

3.0 Current Approaches

The two main computational approaches for solving the protein structure problem are the comparative methods and the de novo methods (1:93). The comparative methods attempt to match a polypeptide with one out of a set of proteins for which 3D configurations are known. The de novo methods search the set of possible configurations of a given polypeptide for the thermodynamically most stable form.

3.1 Comparative Techniques

Databases like the PDB archive over 20,000 extensively studied proteins along with their sequences, 3D structures, and functions. Comparative techniques take advantage of this vast existing source of information by aligning new sequences against sequences found in databases,



```
HRAB27B     1 MTDGDYDYLIKLLALGDSGVGKTTFLYRYTDNKFNPKFITTVGIDFREKRVVYNAQGPNG
mRab27b     1 MTDGDYDYLIKLLALGDSGVGKTTFLYRYTDNKFNPKFITTVGIDFREKRVVYDTQGADG
HRAB27A     1 MSDGDYDYLIKFLALGDSGVGKTSVLYQYTDGKFNSKFITTVGIDFREKRVVYRASGPDG
mRab27a     1 MSDGDYDYLIKFLALGDSGVGKTSVLYQYTDGKFNSKFITTVGIDFREKRVVYRANGPDG

HRAB27B    61 SSGKAFKVHLQLWDTAGQERFRSLTTAFFRDAMGFLLMFDLTSQQSFLNVRNWMSQLQAN
mRab27b    61 ASGKAFKVHLQLWDTAGQERFRSLTTAFFRDAMGFLLMFDLTSQQSFLNVRNWMSQLQAN
HRAB27A    61 ATGRGQRIHLQLWDTAGQERFRSLTTAFFRDAMGFLLLFDLTNEQSFLNVRNWISQLQMH
mRab27a    61 AVGRGQRIHLQLWDTAGQERFRSLTTAFFRDAMGFLLLFDLTNEQSFLNVRNWISQLQMH

HRAB27B   121 AYCENPDIVLIGNKADLPDQREVNERQARELADKYGIPYFETSAATGQNVEKAVETLLDL
mRab27b   121 AYCENPDIVLIGNKADLPDQREVNERQARELAEKYGIPYFETSAATGQNVEKSVETLLDL
HRAB27A   121 AYCENPDIVLCGNKSDLEDQRVVKEEEAIALAEKYGIPYFETSAANGTNISQAIEMLLDL
mRab27a   121 AYCENPDIVLCGNKSDLEDQRAVKEEEARELAEKYGIPYFETSAANGTNISHAIEMLLDL

HRAB27B   181 IMKRMEQCVEKTQIPDTVNGGNSGNWDGEKPPEKK---CIC
mRab27b   181 IMKRMEKCVEKTQVPDTVNGGNSGKLDGEKPAEKK---CAC
HRAB27A   181 IMKRMERCVDKSWIPEGVVRSNGHASTDQLSEEKEKGACGC
mRab27a   181 IMKRMERCVDKSWIPEGVVRSNGHTSADQLSEEKEKGLCGC
```

Figure 5. The comparative methods align sequences against a set of proteins with known structures in a database. Shown here are four sets of sequences aligned against each other. Amino acids whose abbreviations are in black are aligned correctly; those in blue and red are matched against similar amino acids (in terms of side chain category); those in white are matched against a vastly different amino acid. (Source: "Protein alignment." http://www.biomedcentral.com/content/figures/1471-2156-2-2-2.jpg.)

looking for the best overall match (Figure 5). Searching for a good alignment among 20,000 candidates is computationally intensive. Although there are heuristics for intelligent search of large databases, none of the existing algorithms are guaranteed to find the best match. Moreover, heuristic algorithms tend to assume that amino acids in a polypeptide interact independently and that polypeptides have similar lengths and few alignment gaps. (5:33) Comparative methods cannot handle sequences that do not resemble those found in databases. Nor can comparative methods distinguish between two proteins like alpha globin and glutathione S-transferase that have similar alignment sequences but entirely different functions (5:12). Finally, uninformed use of comparative methods can be computationally expensive. Without a knowledge of which proteins in a database of over 20,000 are likely to yield a comparative match, we may need to compare the new sequence with an exorbitant number of existing proteins.

3.2 De Novo Techniques

Recent progress in physical chemistry allows us to predict the mechanisms of simple protein folding using kinetic and thermodynamic analysis (7:70, 9:362). Instead of relying on database matches, de novo techniques search for a stable 3D configuration for a sequence of amino acids interacting among themselves. The reaction of one amino acid with another in the same sequence yields a looped structure that may be more stable than the linear sequence. Of all the possible interactions among the amino acids, those that yield the most favorable structures are most likely to occur in practice. Relying on experimental tabulation of the energetics of amino acid dimerization reactions, de novo techniques search over all the possible ways in which a protein can fold for a folded structure that is thermodynamically most favorable (Figure 6). De

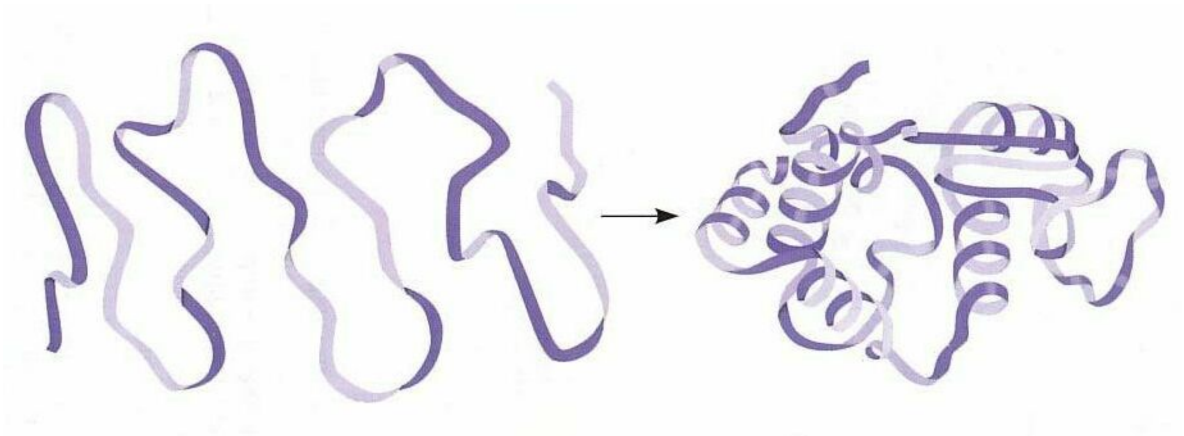


Figure 6. The de novo methods simulate the process of protein folding. Searching over all the different ways in which a protein can fold, de novo methods look for 3D structures associated with particular ways of folding that are thermodynamically most favorable. Shown at the right is a stable (i.e. favorable) structure for the polypeptide at the left. Note the folding nuclei forming loops at the right. (Source: Campbell, N. A., J. B. Reece, and L. G. Mitchell. *Biology*, 5th ed. CA: Addison Wesley Longman, 1999, pp. 76.)

novo methods assume local interactions among a group of amino acids, forming what is known as a folding nucleus. Searching for the 3D structure of a complex protein with lowest energy can be computationally prohibitive. Although flexible statistical algorithms like Markov chain Monte Carlo (MCMC) can lower the running time in searching a high-dimensional space, bounds on the probability of finding the best configuration are difficult to achieve. Like comparative methods, de novo techniques treat two proteins with different functions the same way if they have similar amino acid sequences.

4.0 A Cooperative Approach

Instead of relying on comparative or de novo methods alone, we would like to leverage both our catalog of existing proteins and our understanding of protein folding in determining 3D structure. Protein folding simulation allows us to narrow down a set of viable candidates for sequence alignment; sequence alignment allows us to isolate secondary structures for protein folding simulation. We will flesh out the details starting with comparative alignment.

4.1 Sequence Alignment

Comparative analysis begins with the alignment of a new polypeptide X with a polypeptide Y found in the database. We assume that X evolved from Y in a minimum number of amino acid changes (10:2). Given X and Y, we wish to find an alignment of X against Y that minimizes the number of radical amino acid changes in going from Y to X. Comparing the degree of alignment of X against different Ys found in the database, we hope to find a protein whose 3D structure is similar to X. We represent the two polypeptides using one letter amino acid abbreviations. Bars indicate gaps in the amino acids of one sequence that have no correlate in the other sequence. Plus signs indicate amino acids in the two sequences that have similar side chain categories: electrically charged, polar, hydrophobic, or sulfur-containing (Section 2.0). Thus, bars indicate deletions while plus signs indicate chemical similarity. One possible alignment is shown below.

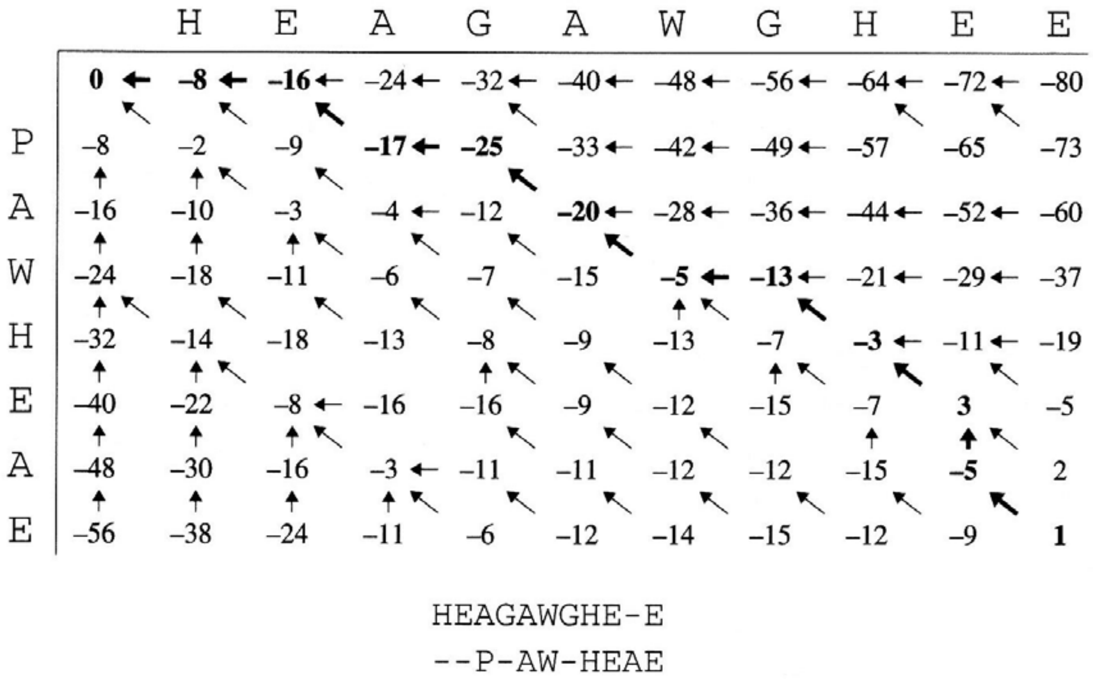
```
Y:  GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
M:  GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
X:  GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE (5:12)
```

Here Y is alpha globin, X is glutathione S-transferase, and M is the matching sequence between them. In order to evaluate an alignment we need: (1) a 20-by-20 substitution table $S(X_i, Y_j)$ assigning a score to each possible amino acid pair that indicates their degree of similarity (scores greater than zero indicate similarity, scores less than zero indicate dissimilarity), and (2) a function D(n) that assigns a negative score to gaps based on a penalty for opening a gap and a penalty for lengthening the gap, where n is the gap length (n is 2 for X above) (10:2). Intuitively, the substitution table tells us whether any pair of amino acids have similar side chains. If two amino acids have the same side chain, then they are the same amino acid, and must have a high substitution score between them. If two amino acids do not have the same side chain, but their side chain belongs to the same category (e.g. hydrophobic), then their score should be smaller, but not necessarily negative. If two amino acids have side chains that belong to different categories (e.g. hydrophobic and polar), then we assign them a negative score. The actual values of the scores are determined from chemical experiments by assuming that all pairs of amino acids in the protein interact independently. Once we have a method for evaluating different alignments, we apply what is known as a dynamic programming algorithm to find the alignment with the highest score.

Let X be the sequence PAWHEAE, and let Y be the sequence HEAGAWGHEE. Note that X and Y have different lengths, so we expect gaps to appear in their alignment. Label X with an index i from 1 to 7 and Y with an index j from 1 to 10, so that X_1 is p, X_4 is H, Y_{10} is E, and Y_0 is just before the beginning of sequence Y. Laying out the sequences in two dimensions as in Figure 7, we construct a table C(i, j) starting from the top left (X_1, Y_1) that gives the score of the best alignment between the subsequences X_1 to X_i and Y_1 to Y_j using the substitution table S(X_i, Y_j) and the negative gap function D(n). (5:20, 10:4)

$$C(i,j) = \max \begin{Bmatrix} C(i-1, j-1) + S(X_i, Y_j) \\ C(i-1, j) + D(1) \\ C(i, j-1) + D(1) \end{Bmatrix} \tag{1}$$

The recurrence for C(i, j) states that the best alignment up to (X_i, Y_j) results from one of the following: (1) a substitution of one amino acid Y_j of Y for X_i of X, (2) a deletion of X_i from the



		H	E	A	G	A	W	G	H	E	E
	0	**–8**	**–16**	–24	–32	–40	–48	–56	–64	–72	–80
P	–8	–2	–9	**–17**	**–25**	–33	–42	–49	–57	–65	–73
A	–16	–10	–3	–4	–12	**–20**	–28	–36	–44	–52	–60
W	–24	–18	–11	–6	–7	–15	**–5**	**–13**	–21	–29	–37
H	–32	–14	–18	–13	–8	–9	–13	–7	**–3**	–11	–19
E	–40	–22	–8	–16	–16	–9	–12	–15	–7	**3**	–5
A	–48	–30	–16	–3	–11	–11	–12	–12	–15	**–5**	2
E	–56	–38	–24	–11	–6	–12	–14	–15	–12	–9	**1**

```
HEAGAWGHE-E
--P-AW-HEAE
```

Figure 7. C(i, j) table for the algorithm, aligning X: `--P-AW-HEAE` with Y: `HEAGAWGHE-E`. X is on the vertical axis; Y is on the horizontal axis. Starting from C(0, 0), calculate the maximum over the values C(i-1, j-1)+S(X_i, Y_j), C(i-1, j)+D, and C(i, j-1)+D. Continue from top to bottom and left to right until the entire table has been marked. Initialize C(0, 0) to be 0 and C(i, 0) and C(0, j) to be negative. Read off the answer in C(m, n), where m is the length of X and n is the length of Y. (Source: Durbin, R., et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. UK: Cambridge University Press, 1999, pp. 21.)

subsequence of X so that X_i is replaced by – in Y, or (3) a deletion of Y_j from the subsequence of Y so that Y_j is replaced by – in X. Note that we have used a constant D(1) for the gap function that assigns a constant score to any amino acid deletion. A more sophisticated algorithm would use a function D(n) that varies nonlinearly with gap length and with amino acid composition. We set C(0, 0) at 0 and C(i, 0) and C(0, j) to be some negative values based on the gap function. We then apply the recurrence for C(i, j) for each entry in the table row by row from top to bottom, and within each row, cell by cell from left to right. C(7, 10) contains the score of the best alignment. To actually obtain the best alignment, keep a pointer for each cell (i,

j) to one of the cells (i-1, j-1), (i-1, j), or (i, j-1) from which we calculated C(i, j). Once we have filled in C(7, 10), we can follow the pointer backward to C(0, 0) and obtain the alignment with the highest score. From Figure 2, the aligned X is `--P-AW-HEAE`, and the aligned Y is `HEAGAWGHE-E`. We have just described a variation of the Needleman-Wunsch algorithm. (5:19) Since we have performed a constant number of calculations per entry of the cell, the running time complexity of the algorithm is on the order of mn, where m and n are the lengths of X and Y, respectively.

The two main problems with the Needleman-Wunsch algorithm are: (1) choosing an appropriate set of sequences Y among the over 20,000 in the PDB to compare X against (i.e. a sampling problem), and (2) proposing a substitution table derived from empirical data. We will see that both problems can be solved using the de novo method of protein folding simulation.

4.2 Protein Folding Simulation

Based on the primary structure of a protein, we can estimate the probability of the appearance of certain secondary structures like alpha helix and beta sheet. Based on the secondary structure of a protein, we can calculate the probability of different 3D configurations, the tertiary structure of the protein. Using experimental data at both the primary and secondary structure levels, we can build a complex model of the 3D fold of a protein using the thermodynamic concept of Gibbs free energy (4:165, 12:325).

When one amino acid side chain reacts with another, or when one amino acid reacts with a molecule of water, the enthalpy of the reaction ΔH is the measurable heat released in a constant pressure environment. The entropy of the reaction ΔS is a measure of the change in the degree of disorder as the reactants interact to form the products. Gibbs free energy ΔG is defined as:

$$\Delta G = \Delta H - T\Delta S \tag{2}$$

Gibbs free energy is a measure of the spontaneity of the reaction. If ΔG is less than zero (heat released and/or increase in disorder), the reaction is thermodynamically favorable. If ΔG is greater than zero (heat absorbed and/or decrease in disorder), the reaction is unfavorable. ΔG is equal to zero for reactions at equilibrium, when the forward reaction of reactants forming products and the backward reaction of products forming reactants balance each other out. The more negative ΔG is, the more favorable the reaction.

Each step in protein folding can be treated as: (1) an amino acid reacting with another amino acid (also known as a dimerization reaction), or (2) transferring an amino acid from a solution in water to a solution in a hydrophobic solvent. In the latter case, we note that the exterior of a protein faces water, while the interior of a protein lacks electrical charge (nonpolar) and is thus “afraid” of water. Chemists have made experimental tabulations of ΔGs for the reactions of different amino acids with each other, and the reactions of amino acids with hydrophobic solvents. Using these ΔGs as primitives, we can calculate the overall ΔGs of a reaction involving multiple amino acids if we knew the mechanism of protein folding. Since we do not how a new polypeptide would fold, we must hypothesize a mechanism and calculate the ΔG associated with those particular steps of reactions leading from an amino acid sequence to a 3D protein configuration. Since the most favorable reaction results in the most stable product, and hence the most negative ΔG, we would be searching through the space of possible 3D structures, each with one or more proposed mechanism for the way the protein folds.

Figure 8 shows a simplified version of the energetics of protein folding. Although the overall ΔG for the transformation of an unfolded protein into a folded protein is negative, the reaction begins with a change in conformation that increases the free energy. The rate of the reaction k depends on both the temperature T and the size of this initial activation energy needed to reach the transition state E_a.

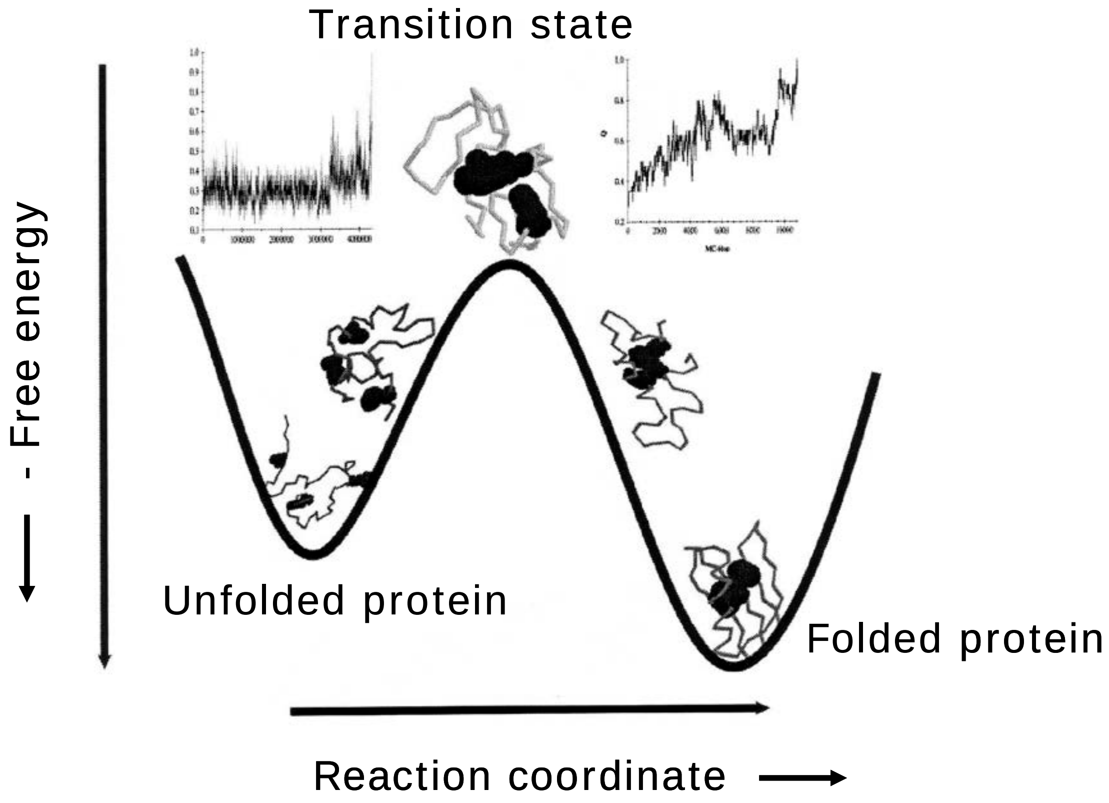


Figure 8. Change in free energy as the protein folds into its native state. The reaction coordinate refers to the time scale of the reaction. The free energy is more negative going to the bottom. The graph at the upper left shows the conformational state of the protein before reaching the transition state. Once past the transition state (upper right), the protein folds rapidly. (Source: Mirny, L. and E. Shakhnovich. "Protein Folding Theory: From Lattice to All-Atom Models." *Annual Review of Biophysics and Biomolecular Structure*, vol. 30, pp. 397.)

$$k = Ae^{-E_a/RT} \tag{3}$$

A is a constant frequency factor and R is the ideal gas constant. (4:470) Equation 3 is an empirical law discovered by Arrhenius. It states that as temperature increases, the reaction rate increases exponentially. Similarly as the activation energy increases, the reaction rate decreases exponentially. Once the activation energy has been overcome, the protein is at a transition state, when it is ready to assume one of the many possible folded structure that will lower the overall free energy. Figure 8 is an idealization. For a real protein, many peaks and valleys in the free energy diagram are observed. Thus we must search for the folded structure that yields the lowest free energy. Regarding ΔGs for each 3D configuration as an energy function on the different possible protein folds, the problem becomes the search for a global minimum of an arbitrarily complex energy function. Due to the presence of activation energies, however, the search for a global minimum is not as simple as going down hill on the free energy function. Rather, we must try going uphill once in a while and see where that leads us. Due to the multi-dimensionality and complexity of protein folding, the global minimization problem can be very difficult to solve efficiently.

One of the most popular set of techniques for unconstrained global optimization is Markov chain Monte Carlo (MCMC), a number of efficient probabilistic sampling algorithms based on the simulation of a Markov chain (6:327, 9:388). Without going into too much technical detail here, let us just note that MCMC algorithms are flexible and efficient, as long as a proper stationary distribution that approximates the energy function of interest is chosen. However, it is often difficult to provide bounds on the probability of finding the global minimum, because sampling on the Markov chain begins before it reaches the stationary distribution. Moreover, the time necessary to run an MCMC algorithm increases quickly as the complexity of the energy function increases. With realistic proteins of over 500 amino acids, we must run MCMC algorithms a long time before we can be reasonably sure of finding a global minimum. Thus we need a way to cluster the amino acid sequence data.

One way to improve the performance of the search for a Gibbs free energy minimum is to use ΔGs for the formation of secondary structures as primitives, instead of ΔGs for amino acid dimerization reactions. If we can break a polypeptide of length N into $\frac{N}{2}$ blocks of roughly equal length with an average block length of 2, then we have reduced the number of possible conformations from N^N to $\left(\frac{N}{2}\right)^{\frac{N}{2}} = \sqrt{\frac{N^N}{2^N}}$, an at least 2^N decrease in complexity. We will see that comparative sequence alignment provides us with a suitable partition of the polypeptide into natural manageable blocks.

4.3 Cooperation

An overview of the Cooperative Structure Determination algorithm (CSD) is presented in Figure 9. We apply the sequence alignment algorithm to the new polypeptide against a random initial set of proteins from the PDB. The gaps (`-`) in the best alignment found for the new polypeptide are used to break the sequence into hypothetical protein blocks intended to model secondary structure. If overlapping blocks are found, we treat each block as a separate motif. Then apply the protein folding simulation algorithm using ΔGs for each motif calculated from amino acid dimerizations. Match general 3D structure properties found by folding simulation to properties of candidates annotated on proteins from the PDB. Eliminate unpromising candidates from further consideration. Change the substitution table used for sequence alignment to reflect the 3D structure found. For example, if no glycine-tryptophan amino acid dimerization is ever found, then we may decrease the values of S(glycine, tryptophan) and S(tryptophan, glycine).

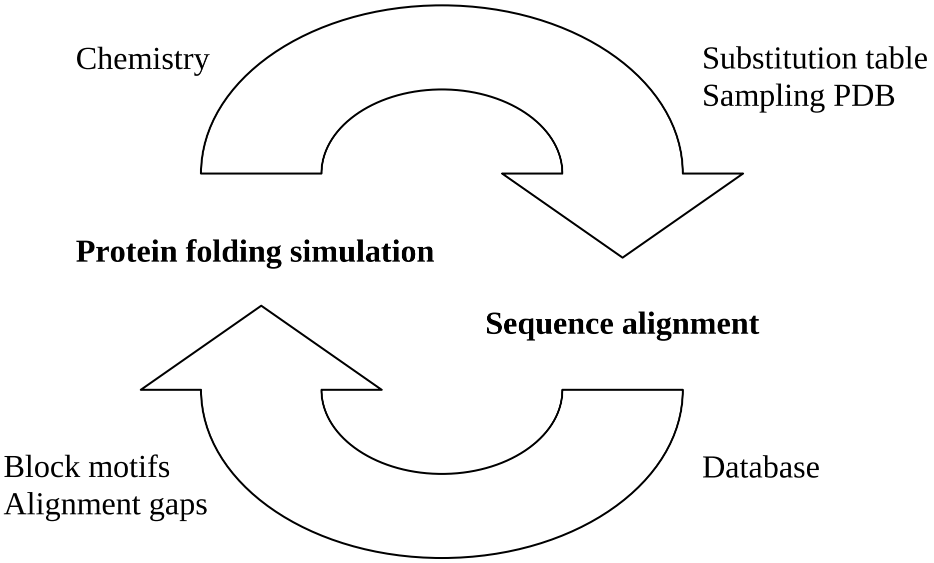


Figure 9. The Cooperative Structure Determination algorithm (CSD).
Match an unknown polypeptide against proteins in a database, then use the gaps in the alignment to improve the performance of protein folding simulation. Use the predicted structure to refine the substitution table and eliminate impossible candidates.

Run the sequence alignment algorithm again against the revised candidate set to generate another set of block motifs for the new polypeptide. Run the protein folding simulation algorithm again based on the revised block motifs. Continue either until the block motifs from successive runs are very similar, or until the substitution table changes little from one iteration to the next. If only one candidate from the PDB is found at any point during the algorithm, check that it has a similar alignment as the new polypeptide and terminate. In going from the alignment to the simulation, we are tackling the problem of statistical clustering: grouping the amino acids of a sequence into manageable chunks of secondary structure. In going from the simulation to the alignment, we are taking on the problem of statistical sampling: choosing a set of proteins from the database to compare against. Thus we overcome the difficulties of comparative methods (sampling) by applying de novo techniques. Similarly we overcome the difficulties of de novo methods (clutering) by applying comparative techniques. The CSD algorithm allows us to use both database protein information and chemical reaction data to predict the 3D protein structure of an amino acid sequence. Note, however, that no functional predictions can be made.

4.4 Evaluation

We can test our algorithm by feeding it a protein whose 3D structure is known. We should be able to find its structure in the PDB and propose a set of similar proteins. After running the CSD algorithm on several proteins with known structures, we can calculate the percentage correct and compare it to related methods. We can also compare the running times of our algorithm against other methods with comparable accuracy rates. CSD should out perform algorithms based on sequence alignment or folding simulation alone. Finally, we can enter a sequence matching competition such as the CASP2 (8:4).

5.0 Conclusion

We have introduced the computational problem of 3D protein structure determination from a 1D amino acid sequence. Comparative and de novo methods of protein analysis were presented along with their costs and benefits. We then presented details of sequence alignment and protein folding simulation methods, which take advantage of protein database information and thermodynamic considerations, respectively. Finally we united the two approaches into a cyclical cooperative technique that utilizes information generated from sequence alignment to bolster folding simulation, and vice versa (Figure 10).

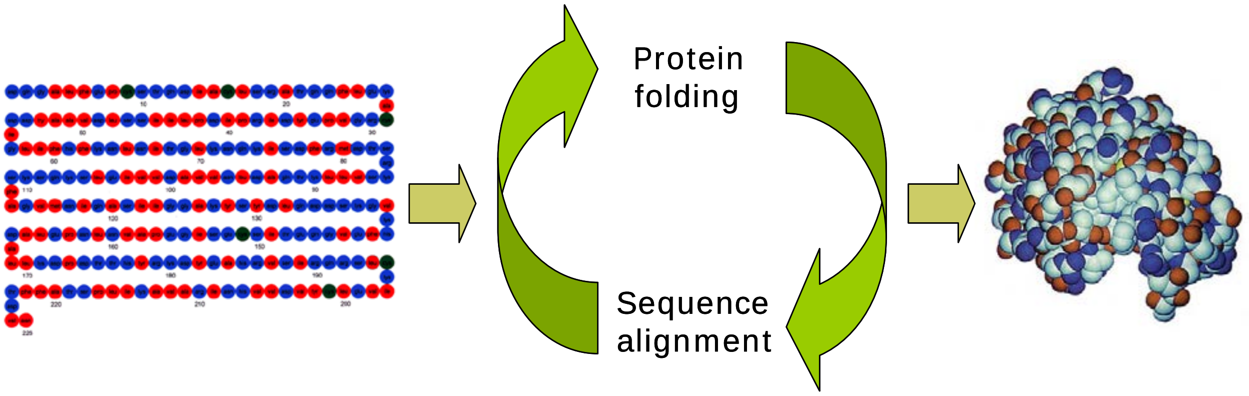


Figure 10. Protein 3D structure determination using an iterative cooperative algorithm (CSD) that cycles between sequence alignment and protein folding simulation. (Protein picture source: Campbell, N. A., J. B. Reece, and L. G. Mitchell. *Biology*, 5th ed. CA: Addison Wesley Longman, 1999, pp. 71.)

Recently, a different attempt to combine sequence alignment and energy minimization has been made by Standley et al. (11:133) Unlike our method, they choose to apply either a comparative method or a de novo method based on the predicted secondary structure. The energy minimization route is similar to our own: 3D structures are screened against a protein database. The sequence alignment route, however, does not utilize the clustering strategy we employ to improve the performance of folding simulation. Moreover our algorithm is iterative, so that each protein sequence is analyzed numerous times. The abundance of recent research in combining comparative and de novo methods testifies to the importance of integrative methods.

Although the 3D structure of a protein can always be determined with sufficient patience, the function of a protein can often elude us. Recall the sequences of alpha globin and glutathione S-transferase presented in 4.1 (3:12). Alpha globin is a human protein involved in the transport of oxygen in blood. Glutathione S-transferase is a rice protein that catalyzes a reaction with herbicides, metabolizing exogenous compounds. Although the alignment score of alpha globin and glutathione S-transferase is quite high, they do not resemble each other functionally. The CSD algorithm can find similar protein structures, but it cannot tell us whether two proteins are functionally similar, for it is not designed to solve the protein function problem. Thus the next step in protein analysis is the determination of protein function.

References

1. Baker, D., and A. Sali. "Protein Structure Prediction and Structural Genomics." *Science*, vol. 294, 2001, pp. 93-96.

2. Brenner, S. E. "A Tour of Structural Genomics." *Nature Reviews*, vol. 2, 2001, pp. 801-809.

3. Campbell, N. A., J. B. Reece, and L. G. Mitchell. *Biology*, 5th ed. CA: Addison Wesley Longman, 1999.

4. Chang, R. *Physical Chemistry for the Chemical and Biological Sciences*, 3rd ed. CA: University Science Books, 2000.

5. Durbin, R., et al. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. UK: Cambridge University Press, 1999.

6. Friesner, R. A., and J. R. Gunn. "Computational Studies of Protein Folding." *Annual Review of Biophysics and Biomolecular Structure*, vol. 25, 1996, pp. 315-342.

7. Grantcharova, V., et al. "Mechanisms of Protein Folding." *Current Opinion in Structural Biology*, vol. 11, 2001, pp. 70-82.

8. Karplus, K., et al. "Predicting Protein Structure Using Hidden Markov Models." *Proteins: Structure, Function, and Genetics*, sup. 1, 1997, pp. 134-139.

9. Mirny, L., and E. Shakhnovich. "Protein Folding Theory: From Lattice to All-Atom Models." *Annual Review of Biophysics and Biomolecular Structure*, vol. 30, pp. 361-396.

10. Myers, E. "Sequence Comparison Algorithms in Molecular Biology." Department of Computer Science, University of Arizona, Tucson, AZ 85721, TR91-29, 1991.

11. Standley, D. M., et al. "Protein Structure Prediction Using a Combination of Sequence-Based Alignment, Constrained Energy Minimization, and Structural Alignment." *Proteins: Structure, Function, and Genetics*, sup. 5, 2001, pp. 133-139.

12. White, S. H., and W. C. Wimley. "Membrane Protein Folding and Stability: Physical Principles." *Annual Review of Biophysics and Biomolecular Structure*, vol. 28, 1999, pp. 319-365.