A Survey of Computational Methods for Protein Structure Prediction

Michael Untereiner

1 Introduction

In recent years, the cost of whole genome sequencing has been drastically reduced, and as a result, scientists have become inundated with a wealth of genomic data linking mutations in the DNA code to various diseases. While the process which converts DNA to protein in our cells has been understood for several decades, the process by which an expressed protein then takes on its distinct three-dimensional shape remains mysterious, despite significant effort in the literature to uncover the folding mechanisms. As a result, elucidating the structural effects of genetic mutations remains a challenging problem due to a lack of tractable algorithms for modeling the protein folding process. While researchers may be able to correlate changes in protein encoding to disease progression, they are oftentimes unsure why these particular changes spur progression of the disease. Without a comprehensive understanding of the structural effects of mutation throughout the disease pathway, effective molecular treatments are much harder, and in many cases impossible, to design.

Thus, protein structure prediction (PSP) is a foundational problem in bioinformatics. The goal of PSP is to determine, with a high degree of confidence, the three-dimensional configuration of the protein given its linear chain order as stored in the genetic code. The structure of a protein involves both a lowest energy configuration (called the native state), but also a set of transition states and probabilities which dictate the biological function carried out by the protein. Both pieces are invaluable for predicting the effect of particular mutations on the efficacy of the overall protein. In addition, a comprehensive model of disease would take into account interactions between proteins, each of which exhibits individual variations in structure. When considering the workings of an entire cell, the space of protein-protein interactions is vast, and researchers need the ability to selectively examine subsets of proteins which are likely to exhibit abnormal interactions, given particular structural variants. In this way, PSP is a fundamental tool for exploring this search space, in effect providing a mapping between the genetic code and protein function.

2 Background

Proteins are the agents of the cell. They catalyze chemical reactions, transport materials, breakdown waste, and coordinate all of the activities of cellular life. Proteins range from a few hundred amino acids in length to tens of thousands. Remarkably, a set of twenty amino acids can account for the vast complexity and variety demonstrated by proteins. What's more, proteins fold on the millisecond time scale - multiple orders of magnitude faster than the best computational solutions. These two facts speak to the combinatorial ingenuity of the protein molecule, as well as our difficulty in deciphering the secrets which underlie its efficient folding.

A protein is a chain of amino acids. Successive amino acids are connected by peptide bonds, forming a carbon-nitrogen backbone. Amino acids are distinguished by their side chain (also called R group), of which there are twenty distinct types. This side chain determines an amino acid's unique chemical properties and potential interactions with the other amino acids in the chain. Side chains can rotate freely about the central carbon atom, as long as they do not collide with nearby side chains. The angle that an amino acid's side chain occupies in a protein is called its rotamer.

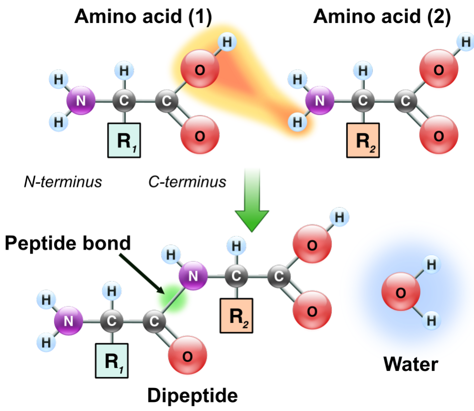

Figure 1. Formation of a peptide bond
www.wikipedia.org/wiki/Amino_acid

Amino acids can be classified as either hydrophobic (repelled from water) or hydrophilic (attracted to water). Once a protein chain is synthesized and released into the aqueous environment of the cell, hydrophobic amino acids cluster together and become buried deep within the interior of the protein, while hydrophilic amino acids rise to the surface of the protein. In this way, the disruption of the water network is minimized and entropy is maximized. This dogma has been accepted by molecular biologists for decades and is called the thermodynamic hypothesis.[1] The protein chain stochastically folds, guided by random interactions with water molecules, until it reaches its native structure, the 3-dimensional configuration which minimizes its free energy.

A pair of amino acids, when brought close together, can form a hydrogen bond, ionic bond, or covalent bond, depending on their side chains. This bond will then restrict the overall folding of the chain by constraining the motion of these two amino acids. The active site of a protein is a section of the chain which selectively binds to a substrate and catalyzes a chemical reaction. Active sites often include many charged and polar side chains, which are particularly useful for transferring energy between different molecules. Bonds between side chains can be broken, and new ones formed, during the activity of the protein. Such kinetic and potential energy transitions between amino acids within the chain are responsible for function.

Protein structure is hierarchical and classified at four levels. Primary structure refers to the linear ordering of amino acids in the chain. Secondary structure denotes local motifs produced by hydrogen bonding between side chains. The two most common of these motifs are alpha helices and beta sheets. Tertiary structure includes ionic and covalent bonding between amino acids which are far apart in the primary structure, but brought close together during folding. Finally, quaternary structure occurs when multiple chains assemble into a protein complex.

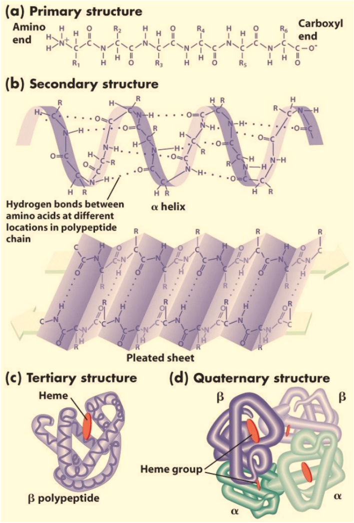

Figure 2. Hierarchy of protein structure. www.wikipedia.org/wiki/Amino_acid

3 Problem Definition

There are many formalizations of the protein folding problem, each of which motivate different algorithms for structure prediction. We will examine them one by one. All of them require a scoring function, which stands in for the energy function in that it ranks configurations by how close they are to the native state. Scoring functions examine the pairwise contacts which occur in a given configuration in relation to a scoring table, and then add up the scores for each contact. The simplest scoring function classifies each amino acid as hydrophobic (H) or hydrophilic (P), and assigns a score of 1 for each H-H contact and 0 otherwise.[2] A more complex scoring function could use a 20x20 table which specifies unique scores for every possible pairwise contact between amino acid side chains.

3.1 On-Lattice Models

On-lattice models discretize the configuration space by placing the protein on a lattice.[2] Two beads are used to represent each amino acid: one for the backbone carbon-nitrogen, one for the side chain. These models are approximations of the configuration because in reality, bond lengths may fluctuate slightly and side chain rotamers are continuous. Despite these inaccuracies, on-lattice models are useful because they bound the search space and can give close enough approximations that realistic configurations can be derived with minimal further computational work.

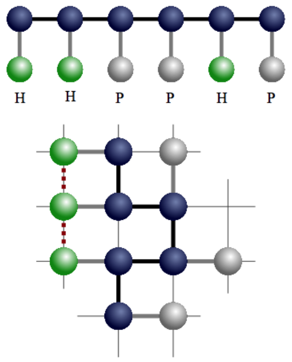

Figure 3. On-lattice model in 2D for protein of length 6. www.wikipedia.org/wiki/Dihedral_angle

In the graphic above, the backbone beads are colored blue and side chains have been classified as either hydrophilic or hydrophobic. Using the simple H-H scoring function described earlier, we can search through the possible configurations for a protein of length 6 (when placed on a 2D lattice) and find an optimal configuration with the maximum possible score of 2. More sophisticated on-lattice models use face-centered cubic models and larger scoring tables.

3.2 Off-Lattice Models

Off-lattice models parameterize the backbone dihedral angles (denoted ϕ and ψ) between successive amino acids in the chain. Contacts between side chains can be determined by calculating distances and accepting those which fall below a given threshold for bond formation. Off-lattice models are more realistic because they don't discretize the search space, allowing for real-valued dihedral angles and thus more complex backbone conformations.[7] An additional challenge of using off-lattice models is that we must implement collision detection between atoms in the protein to prevent illegal rotations.

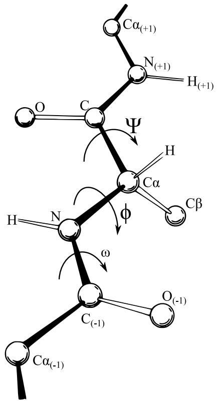

Figure 4. Backbone of a protein molecule. www.wikipedia.org/wiki/Dihedral_angle

3.3 Molecular Dynamics

Molecular dynamics simulations use Newton's equations of motion to calculate the trajectories of all the atoms within the protein through time. These simulations calculate the net force on every particle in the system and then approximate the integral over a small time step to update all of the positions and velocities within the system. Molecular dynamics is a appealing alternative to other models because it is the most physically realistic, because quantum mechanical effects can be ignored. However, this approach is the most computationally intensive and while attempts at coarse-graining proteins by clumping atoms together into single units of force have been met with modest success,[6] improving runtime in this way defeats the purpose of using Newtonian mechanics in the first place, namely the increased accuracy of predicted structures.

4 Protein Folding as Graph Search

Protein structure prediction is modeled well as graph search because we end up with an approximation of the native state, but also a set of configurations which are reachable from the native state, which are likely to be related to protein function.

4.1 Markov Chain Monte Carlo

With both on-lattice and off-lattice models, we can construct a Markov chain to model the folding pathway.[7] We start with a configuration of the protein to be folded. Next, we randomly pick one amino acid in the chain, and rotate its angles (on the lattice or not), update all the other amino acid positions according, check for collisions, and calculate a score for the new configuration. We repeat this process until for a given number of steps, or until our score drops below a threshold. Alternatively, we can initialize n searches in parallel and continually restart the searches, refining our start configurations to be the n best-scored configurations found thus far.

Before exploring paths in our Markov chain, we will typically generate independent random configurations and pick ones with the best score, in order to improve our starting state and reduce the number of steps in our chain needed to obtain a good score. A further improvement is to only select amino acids at each step which have not formed bonds in the current configuration, and also to select amino acids with smaller side chains (and thus greater rotational freedom) with higher probability.[1] Another strategy entails weighting amino acid choices by their potential to increase the score (using a conditional expectation from the score table).

4.2 Simulated Annealing

Graph search can be further refined by using a technique called simulated annealing. We introduce a temperature factor T which represents random noise, in this case coming from the water network. These disturbances have a certain probability proportional to T of breaking existing bonds between amino acids in our current configuration. Over the course of our path traversal, T decreases until it reaches zero, at which point we are back to our normal Markov chain monte carlo algorithm. Simulated annealing is a useful tool because it resembles how proteins actually fold in vivo,[1] and allows us to throw out bonds which may have formed early in the pathway yet trap our random walk in a locally optimal configuration.

4.3 Genetic Algorithms

Genetic algorithms identify subsequences of the protein chain which have the highest score per amino acid. These subsequences are local motifs which maximize the score possible for a given number of amino acids, and are viewed as solutions to a smaller, local folding problem. The whole chain is then partitioned into such regions with high scores and low scores, and the algorithm recombines the high score structures to match amino acid sequences with low fitness, hopefully creating new structures with higher scores. Ideally, the subsequences of high score slowly aggregate until the entire chain is covered, but this is not guaranteed.[3]

One advantage of this approach, and a motivation for modifying the start configurations of graph search with output from a genetic algorithm, is that we can readily calculate global optimal structures for proteins of small length, and use genetic approaches to approximate optimal structures for moderate size proteins,[4] which has the potential to greatly improve the performance of graph search. Unfortunately, genetic algorithms do not mimic the physical folding of proteins to any significant degree, and so for particular protein chains, they may give erroneous answers which actually impair graph search.

5 Machine Learning and Protein Design

Machine learning techniques leverage dependencies between the primary and tertiary structure of known proteins to learn model parameters which accurately predict protein structure. These models are particularly powerful because they do not attempt to solve the general protein folding problem, but rather rely on training data which comes specifically from human biology.

Bayesian learners have been shown to be particularly well-suited for this type of protein structure prediction. We can initialize their prior knowledge to reflect well-established constraints on individual amino acids in the protein chain. Then, we can update these priors with the observations of experimentally verified protein structures. The concept of a ramachandran plot (shown on the right) will be useful to illustrate this process.

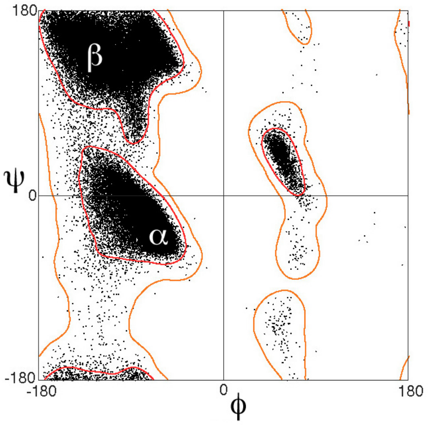

Figure 5. Example ramachandran plot.
www.wikipedia.org/wiki/Ramachandran_plot

A ramachandran plot shows the possible rotamers of a single amino acid's side chain - those which don't result in atomic collisions - in terms of the dihedral angles on either side (ϕ and ψ). The two largest clusters represent two main secondary structure positions: alpha helix and beta sheet. Note that this plot represents an average over all twenty amino acids. Plots can be made for each of the individual twenty amino acids as well. Our bayesian learner starts with these twenty plots, treating them as conditional distributions over each of the amino acids in our protein chain. Next, during training, our learning algorithm updates these distributions with the observed dihedral angles in human proteins determined by experiment, until most of the probability mass has shifted to a small area.[1] The conditional distributions can be further subdivided on adjacent amino acids with only modest increases in memory resources, until a powerful predictive model has been learned.

The limitation of this learning paradigm is that the experimental data itself is not entirely reliable. First of all, there are limits on the resolution of experimentally determined structures. Also, experimental methods crystallize proteins in order to validate their structure with x-ray diffraction, resulting in modified structures which do not perfectly match structure in vivo. Neural networks also show some promise in structural prediction, but they have the additional limitation that oftentimes we have no idea why they work.[7]

6 Discussion

There is no one single algorithm for protein structure prediction. Depending on the size of the protein chain, the heterogeneity of the amino acids present, the resemblance to known structures, and many other factors, one algorithm will likely perform much better than the others. Molecular dynamics and related particle swarm algorithms have the highest fidelity to protein folding in vivo. Graph search approximates folding pathways but is not physically realistic. Genetic algorithms and machine learning techniques are detached from the physical pathways of protein folding, but can still be useful. For many applications, we do not explicitly care how a protein folds, as long as we reliably approximate the native state. In order to understand the functional role of a protein, however, we often require a more detailed procedure for enumerating the state transitions which are readily available from the lowest energy configuration.

Going forward, more protein structures will be experimentally determined, and these new data points will enhance the attractiveness of machine learning solutions. It will always be important, however, to strive towards a unified, generalizable computational solution for the protein folding problem that combines all the approaches enumerated above. The space of potentially useful protein sequences is inexhaustibly vast, and machine learning alone will not be enough to truly revolutionize the way that we design drugs and treat disease. As computational power continues to increase, more work needs to be done to develop flexible hierarchical models for PSP in pursuit of the structural basis for protein function.

References

1. Andrej Sbreveali, Eugene Shakhnovich, and Martin Karplus. How does a protein fold? Nature, 369(6477):248-251, May 1994.

2. Bonnie Berger and Tom Leighton. Protein folding in the hydrophobic-hydrophilic (hp) is np-complete. In Proceedings of the second annual international conference on Computational molecular biology, New York, NY, USA, 1998. ACM.

3. Ron Unger and John Moult. Genetic algorithm for 3d protein folding simulations. In Proceedings of the 5th International Conference on Genetic Algorithms, pages 581-588, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

4. Trent Higgs, Bela Stantic, Tamjidul Hoque, and Abdul Sattar. Genetic algorithm feature-based resampling for protein structure prediction. In IEEE Congress on Evolutionary Computation [1], pages 1-8.

5. Alena Shmygelska and Holger Hoos. An ant colony optimisation algorithm for the 2d and 3d hydrophobic polar protein folding problem. BMC Bioinformatics, 6(1):30, 2005.

6. L. G. P. Hernndez, K. R. Vzquez and R. G. Jurez, ”Estimation of 3D Protein Structure by means of parallel Particle Swarm Optimization,” IEEE Congress on Evolutionary Computation, Barcelona, 2010, pp. 1-8. doi: 10.1109/CEC.2010.5586549

7. J. M. Bahi, N. Ct and C. Guyeux, ”Chaos of protein folding,” The 2011 International Joint Conference on Neural Networks, San Jose, CA, 2011, pp. 1948-1954. doi: 10.1109/IJCNN.2011.6033463

8. Zhou, neng-fa; Bar-Noy, Amotz. A Probabilistic Constraint-based Approach to Protein Structure Predication.

9. Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, et al. (2010) Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized. PLoS ONE 5(11): e13714

10. Schmidler S.C., Liu J.S., Brutlag D.L. (2002) Bayesian Protein Structure Prediction. In: Gatsonis C. et al. (eds) Case Studies in Bayesian Statistics. Lecture Notes in Statistics, vol 162. Springer, New York, NY