Protein Structure Prediction and Comparison

Xiang Fu

October 22, 2024

1. Protein Structure Prediction and Comparison

Proteins come in all different shapes, and these shapes are crucial for their biological functions. The shape or structure of a protein is determined by its amino acid sequence (primary structure) and the way it folds into specific three-dimensional shapes, which can be predicted using various bioinformatics tools and algorithms.

—

The shape of a protein influecnes its function

- A protein's function is highly dependent on its three-dimensional structure. The specific shape allows the protein to interact with other molecules, fit into binding sites, and catalyze chemical reactions. If the shape changes (due to mutations or environmental factors), the function of the protein may be altered or lost.
- For example, enzymes have active sites that are precisely shaped to bind specific substrates, and any changes to their structure can impact their ability to catalyze reactions.

—

A protein typically folds into the same shape every time

- Despite the complexity of folding, a protein typically folds into the same three-dimensional structure each time, given identical environmental conditions. This is due to the physical and chemical properties of the amino acids in the protein and the interactions between them (such as hydrogen bonds, hydrophobic interactions, and disulfide bridges).
- Misfolding of proteins can lead to diseases such as Alzheimer's, Parkinson's, and cystic fibrosis, where the incorrectly folded proteins aggregate or malfunction.

—

Central Dogma: DNA to mRNA (transcription), mRNA to to Protein Sequence (translation), to Protein

Question. How do we get the structure of a protein from its sequence? How do we sequence a protein?

Both are unsolved as of now.

1.1. Protein Structure Prediction Problem

- Input: An amino acid string corresponding to a protein.
- Output: The 3-D shape of the protein.

Nature has devised a "maghic algorithm" solving this biological problem.

Question 1: What is the 3-dimensional protein corresponding to a string of amino acids?

1. *ab initio* Protein Structure Prediction
2. Homology modeling

Question 2: How can we compare two similar proteins on the level of structure?

1. RMSD

2. Contact Maps

1.2. What we can do now?

Align Potein Sequences

- Same as DNA sequence alignment except using amino acids instead of nucleotides
- Sometimes helpful, but other times cna be very misleading

For exmaple, hemoglobin subunit alpha sequences can be very different, but the structures are very similar.

Solve for individual protein structures using Cryogenic electron microscopy "Cryo-EM"

The lectron microscope needed can cost $5< or more and cost a fortune to run.

Just for humans, there are between 600,000 and 6 million protein isoforms.

Barring a amjor innovation, we will never be able to experimentally determine the structure of all proteins.

1.3. Some Necessary Bio Background

A protein's **primary structure** refers to the amino acid sequence of the protein, which is the linear sequence of amino acids linked together by peptide bonds. This sequence is determined by the genetic code in the DNA and is unique to each protein. The primary structure dictates the protein's higher levels of structure (secondary, tertiary, and quaternary) and ultimately its function.

The correct order of amino acids in the primary structure is critical because even a single change can alter the protein's function, sometimes drastically. For example, a change in a single amino acid in the protein hemoglobin leads to the disease sickle cell anemia.

A **secondary structure** is a repeating substructure that forms as a substructure of the oberall folded protein.

A protein's **tertiary structure** describes its final 3D shape after the polypeptide chain has folded and is chemically stable. This is what we most commonly refer to as the "structure" of a protein.

Some proteins have a **quaternary structure**, which describes the protein's interaction with other copies of itself to form a single functional unit, or a **multimer**.

Hemoglobin is a multimer consisting of two alpha subunits and two beta subunits.

—

Proteins are typically comprised of independetly folding **domains** that are each responsible for a specific interaction or function.

Proteins domains can be thought of a moleculare buildings blocks. A protein can have many domains, and many domaisn are reused across proteins.

Issue. Proteins are flexibale and can therefore form a huge number of shapes.

- A good analogy for polypeptide flexibility is the "Rubik's Twist" puzzle.

A polypeptide with n amino acids has $n-1$ peptide bonds. If each bond has k stable conformations, then the polypeptide has k^2n-2 potential structures.

—

Proteins seek the lowest energy conformation

- We can view protein folding as finding the tertiary structure that is the most stable given a polypeptide's primary structure (i.e. has loweest potential energy.)

1.4. *ab initio* Protein Structure Prediction

Biochemists have produced scoring functions called **force fields** that compute the potential energy of a candidate protein structure.

Proteins seek the lowest energy conformation.

This is an optimization problem, and the search space is the combination of

—

A Local Search Algorithm for Protein Structure Prediction

1. Start with an arbitrary protein conformation.
2. Make slight changes to the structure in a varietty of ways to produce "neighbors".
3. Consider the neighbor with optimal score. Is its score better than the current structure?
   - If "yes", update the current structure to this neighbor and iterate at step 2.
   - If "no", return the current structure.

Idea. Provide some "jiggle" to allow candidate solutions to "bounce" out of the local optima.

1.4.1. Quantifying "Jiggle"

When considering a "neighbor", S' of a candidate protein structure S:
- If $\text{energy}(S') < \text{energy}(S)$, update $S = S'$
- If $\text{energy}(S') > \text{energy}(S)$, then update $S = S'$ with probability proportional to $\Delta \text{ energy} = \text{energy}(S) - \text{energy}(S')$.

Classic function: $\exp(\Delta \frac{\text{energy}}{T})$, where T is a "temperature" constant or function. This is called **simulated annealing** because of the analogy ofreducing the temperature of a metal slowly.

The "Hotter" the temperature, the more "jiggle".

Because the search space is so large, and we need to run an algorithm with a lot of initial structures, *ab initio* algorithm still are **extremely slow finish**.

1.5. Homology Modeling

Using the **known protein strcture** of a **homologous protein** as a template, we can in theory improve both the accuracy and speed of protein structure prediction. This idea serves as the foundation of **homology modeling** for protein structure prediciton (a.k.a **comparative modeling**).

If we do not know which template to use before we begin, how could we find a suitable template?

Once natural thing to do would be to search for similar *sequences* for our novel protein in a database using some form of sequence alignment.

Once we have a template, how might we use what we have learned to perform homology modeling?

One approach is to include an extra "similarity term" in our energy function, The more similar a structure is to the template, the more this similarity term decreases the function we are minimizing

$$f(S) = \text{energy}(S) - \text{similarity } (S, \text{template}) \tag{1}$$

We assume that very conserved regions in two genes correspond to essentially identical structures in the proteins.

We then use **fragment libraries**, or known protein substructures, to fill in the non-conserved regions and produce a final structure. This approach to homology modeling is called **fragment assembly**.

The answer to Question 2 will help us using the homology modeling.

1.6. Comparing Protein Structures

Question 2: How can we compare two similar proteins on the level of structure?

Goal. Develop a "distance function $d(S, T)$ that quantifies shapes S and T are.

To define $d(S, T)$, first translate/flip/rotate S so that the resulting shape is as similar to T as possible. Then, determine how different the shapes are.

First, we will first translate S and T to have their **centroids** (a.k.a. **center of mass**) at 0. The centroid is the point (x, y) where x is the average of x-coordinates and y is the average of y-coordinates of the shape.

Next, we rotate and flip S to resemble T as closely as possible.

We can use the **root mean square deviation (RMSD)** between the two shapes:

$$\text{RMSD}(s, t) = \sqrt{\frac{1}{n} \cdot (d(s_1, t_1))^2 + d(s_2, t_2)^2 + ... + d(s_n, t_n)^2)} \tag{2}$$

the square root of the average squared distance between corresponding points in the vectors.

1.7. RMSD on Proteins

RMSD calculates distances between corresponding points. In pratice, researchers take the main carbon atom from each amino acid to vectorize a structure.

How do we get corresponding points? We can use protein sequence alignment.

Any gap columsn will not contribute to RMSD.

1.8. Kabsch Algorithm

Uses Singular Value Decomposition to find flip/rotation of one shape to minimize RMSD.

1. Calculate corss covariance matrix $H = S^T T$
2. Apply SVD to H

$$H = U\Sigma V^T \tag{3}$$

3. Calculate d which specifies if there is reflection (just need the sign of d)

$$d = \det(UV^T) = \det(U)\det(V) \tag{4}$$

4. Calculate R

$$R = U \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{pmatrix} \cdot V^T \tag{5}$$

It turns out $R * T$ minimizes the RMSD between S and T.

—

Small protein changes can have ahuge impact on RMSD

Here are two protein structure that are identical except for changing a single bond angle. The Kabsch algorithm will completely misalign these.

However, notice that intrapotein distances $d(s_i, s_j)$ are similar to the distances $d(t_i, t_j)$.

We can compare structures locally using *intraprotein distance*.

1.9. Contact Maps

For some threshold t, given a structure S, color cell (i, j) black if $d(s_i, s_j) < t$ and white otherwise.

How might we use a contact map to look for local regions of similarity in protein structures?

Q per residue offers a single value for how much two proteins differ locally.

Q per residue (Qres): defined as follows:

$$Q_{\text{res}}^{(i)} = \frac{1}{N-k} \sum_{j \neq i-1,i,i+1}^{\text{residues}} \exp\left[-\frac{\left[d(s_i, s_j) - d(t_i, t_j)\right]^2}{2\sigma_{i,j}^2} \right] \tag{6}$$

- N is the number of amino acids in each protein;
- k is equal to 2 when i is at either the start or the end of the protein, and k is equal to 3 otherwise;
- The variance term, $\sigma_{i,j}^2$ is equal to $|i-j|^{0.15}$, so that nearby amino acids have more influance.

—

What happens to the interior term of the sum if $d(s_i, s_j)$ is comparable to $d(t_i, t_j)$?

It heads toward $\exp(0) = 1$.

What heppens to the interor term of the sum is vert dufferent to $d(t_i, t_j)$?

It heads toward $\exp(\infty) = 0$.

1.10. CASP and AlphaFold

Critical Assessment of Protein Structure Prediction (CASP): contest run every two years since 1994 that tests structure prediction algorithms against each other on known (hidden) protein structures.

People used methods we described for most of this time (simulated annealing, homology modeling), and typically tested their algorithm using RMSD and Qres.

—

A Solution Out of Left Field - AlphaFold

This is where things stood until 2018... **AlphaFold**!

AlphaFold obtained a mdain RMSD of 1.6, but to be trustworthy for a sentitive application like designing drug targets, it would need an RMSD about 90% lower.

AlphaFold does well but is "trained" using a database of known structures, which makes it more likely to correctly predict known structures. But proteins with structures dissimilar to any known structure possess some of the most scientific interest.

We may never again see such an *improvement* to the state of the art in a problem that has puzzled biologies for fifty years.