Computational Molecular Biology and Bioinformatics

Protein Dynamics

Malay Bhattacharyya

Assistant Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

December, 2021

1. Basics
2. Molecular dynamics approach
3. Secondary structure prediction approach
   - Secondary structure of RNA
   - Secondary structure of protein
4. Hands-on

Protein folding

Protein folding deals with the problem of understanding the tertiary structure (conformation) from the given amino acid sequence.

Broadly speaking, there are three categories of approaches of addressing the problem of protein folding as listed below.

- Molecular dynamics
- Secondary structure prediction
- Protein threading

Due to the existence of conformational isomerism, protein folding is a hard problem to address. Conformational isomerism is a form of stereoisomerism in which the isomers can be inter-converted just by rotations about formally single bonds.

Molecular dynamics

In this, we simulate the actual folding process of a protein, considering mean force fields acting on all atoms in the constituent amino acids, as well as atoms of the solvent (water).

Starting with a random initial conformation, we can calculate motion vectors for the atoms according to different forces. Some of these forces are listed below.

- Covalent bonds
- Hydrophobic effect
- Electrostatic forces
- van der Waals forces
- etc.

Structure prediction

In this, we consider a fixed energy model as follows.

$$E : \Omega \rightarrow \mathbb{R}$$

Here, Ω is the set of all possible conformations.

Protein threading

In this, we use a knowledge-based graph approach. This is because, *de novo* protein structure prediction using physical/chemical energy functions is untractable. Hence, we compute a statistical pseudo-energy functions.

What is protein dynamics?

Proteins structures are not strictly static, but rather populate ensembles of (sometimes similar) conformations. Transitions between these states occur on a variety of length and time scales.

The study of protein dynamics is most directly concerned with the transitions between these states, but can also involve the nature and equilibrium populations of the states themselves.

Understanding protein function requires detailed knowledge about protein dynamics, i.e. the different conformational states the system can adopt.

Molecular dynamics approach

Earlier approaches have proposed to use Boltzmann distribution to define an energy function with terms involving the potentials of amino acids getting paired with each other. This is computed from a reference database of 3D coordinates of protein structures.

Given an energy function E defined on a finite set V of states, the Boltzmann distribution is defined as follows.

$$p(v) = \frac{e^{\frac{-E(v)}{RT}}}{\sum_{w \in V} e^{\frac{-E(w)}{RT}}}$$

Molecular dynamics approach

Note that, the energy function E is not known but the probabilities $p(v)$ for $v \in V$ can be determined from the protein databases. Based on this, we can obtain the energy as follows.

$$E(v) = -RT \ln p(v) - RT \ln \left(\sum_{w \in V} e^{\frac{-E(w)}{RT}} \right)$$

Generally, it is assumed that a protein folds into a unique conformation determined by the global minimum of its free energy.

Without going into the complexity of other effects (e.g., hydrophobic effect, electrostatic force, van der Waals force, etc.), we can perform a fast approximation of such an energy function by considering only the pairwise force contributions.

Molecular dynamics approach

It is reasonable to assume that the average distance between a given pair of amino acids in a reference protein database should correspond to the average energy contribution due to this pair.

Currently, there are many computational software packages that can realize molecular dynamics simulations, such as GROMOS (Scott et al., 1999), GROMACS (Lindahl et al., 2001), NAMD (Phillips et al., 2005), Tinker-HP (L Lagardere et al., 2018).

Secondary structure of RNA

The secondary strcuture of an RNA sequence of length n is an undirected graph $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$, $E \subseteq V \times V$, such that

1. $(i, j) \in E \Leftrightarrow (j, i) \in E$
2. $(\forall 1 \leq i < n)[i, i + 1, \in E]$
3. For $1 \leq i < n$, there exists at most one $j \neq i \pm 1$ for which $(i, j) \in E$
4. If $1 \leq i < k < j \leq n, (i, j) \in E$ and $(k, l) \in E$, then $i \leq l \leq j$

Secondary structure of RNA

There is a one-to-one correspondence between RNA secondary structures and base pairings within the RNA sequence via hydrogen bonds.

Consider the RNA sequence AAGAAACAUCACAU. It might have base pairings among the nucleotides at the positions (2, 14) (between A and U), (3, 7) (between G and C), and (9, 13) (between U and A). But there could be additional possibilities as shown below.

A	A	G	A	A	A	C	A	U	C	A	C	A	U
.	(	(	.	.	.	)	.	)	.	.	.	.	.
.	(	.	.	.	.	)	.	(	.	.	.	)	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.

Secondary structure of protein

Given the amino acid sequence of a protein, we have to predict the tertiary and quaternary structures of the protein.

Lattice models for protein structures

The most basic lattice model used for understanding the secondary structure of proteins is the **HP** model. In this, the 20-letter alphabet of amino acids (and the corresponding variety of forces between them) is reduced to a 2-letter alphabet, namely **H** and **P** with the following interpretations.

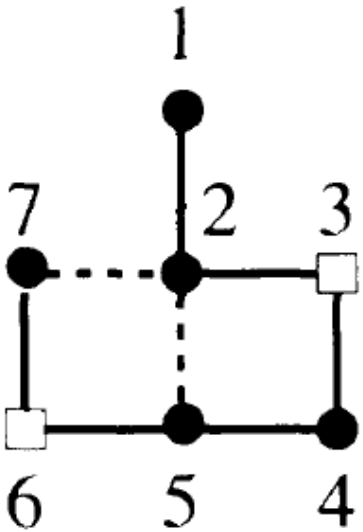
- **H** denotes a *Hydrophobic* amino acid
- **P** denotes a *Polar* (i.e., Hydrophilic) amino acid

Note that, the hydrphobic force is believed to be the predominant force in folding the globular proteins. The energy function is given as follows.

	H	P
H	-1	0
P	0	0

The **HP** model minimizes the hydrophobic force.

Lattice models for protein structures



connected	**vs.**	*topological*
1 — 2		2 — 5
2 — 3		7 — 2
3 — 4		
⋮		

Sample conformation for 1101101. The white beads represent P, the black ones H monomers. The two contacts are indicated via dashed lines.

Hands-on

1. Access the NAMD tool and do the following steps.
   i) Access the amino acid sequence of the S protein (also known as spike protein or surface glycoprotein) of SARS-CoV-2 from NCBI.
   ii) Feed the sequence into the NAMD tool (`https://www.ks.uiuc.edu/Research/namd`) and perform analysis.
   iii) See the latest version of NAMD from here: `https://aip.scitation.org/doi/10.1063/5.0014475`
   iv) Read the following paper to understand its utility.
   Casalino et al., AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. The International Journal of High Performance Computing Applications, p.10943420211006452, 2021.
   Link: `https://journals.sagepub.com/doi/10.1177/10943420211006452`.