1 Subjects

- What is protein tertiary structure?
- Computing a folding in the 2D HP model

2 Notes

2.1 Protein

Protein is made up of amino acids, which then folds into a three-dimensional structure. Experiments show that they can unfold and refold, which leads to the believe that the three-dimensional structure can be found computationally from the information contained in the amino acid sequence.

The thermodynamical hypothesis states that the native structure of a protein is the one for which the free energy is at a global minimum.

If we were able to construct the tertiary structure from the primary structure, we would be able to answer scientific questions like:

- What does the structure of protein x look like?
- Can we predict the binding of of molecule x to y?
- Does molecule z has the potential to become a good drug candidate?
- Can we find a promising subset of drug candidates from a huge database containing millions of compounds?

A bit of background on genes, DNA etc. A cell has 46 chromosomes, DNA molecules, which store genetic information. DNA's consist of nucleotides (ACGT). Before a gene comes into use, its coding DNA is transcribed to RNA which in turn is translated into a protein.

An amino acid consists of a central carbon atom C_α which is bonded to an amino group and a carboxyl group and a side-chain. The backbone in proteins is the sequence of amino groups, C_α and carboxyl groups. A gene reveals the blueprint of a protein, the structure is what we will try to find.

Primary the sequence of amino acids

Secondary is either a α-helix or a β-strand. They are secondary structures that form while the protein is folded. This is usually presented as a sequence of characters corresponding to the individual amino-acids of e.g. α, β.

Tertiary the three-dimensional form the protein takes after folding.

However, protein folding remains one fo the great unresolved problems of molecular biology. So why do we even want to predict the protein structures?

Because X-Ray crystallography is expensive and time-consuming, and some important proteins (membrane proteins) are difficult or impossible to crystallize. Furthermore, doing simulations may provide us with new knowledge about the physics, and help us understand mis-folding which causes diseases.

2.2 HP Lattice models

In the HP-lattice models we want to predict protein structures by minimizing free energy, under the assumption that formation of a hydrophobic core is a primary force in protein structure formation.

The HP-part is that we take the amino-acids and map them to either hydrophilic (water loving) or hydrophobic (water hating) amino acids. This isn't always so black and white, since some amino acids are not hydrophilic or vice versa in all contexts.

The lattice part is that we embed it into a lattice (i.e. a grid) and avoid any overlaps. The aim is then to maximize the number of non-bonded H-H contacts, and we just increment (or decrement for minimization) the score by one for each pair of neighbouring non-diagonal, non-bonded lattice points which are both H's.

The result is that we the hydrophobic residues tend to be on the inside, forming a hydrophobic core, and the hydrophilic residues on the outside.

- Residues are represented by a single atom, what about the side-chains? Bonds are not mimicing 'reality'
- Electrostatic interactions (repulsion/attraction) are not considered
- Energies are very short range
- Two residues must be an odd distance (of at least three) apart to be in contact
- Does not reveal the structure of any particular protein
- For short sequences, all conformations can be found
- It's easy to understand for non-biologists and it is easy to implement

The number of possible valid folds for a sequence of length L on a two-dimensional square lattice approaches 2.638^n, so the number of solutions is exponential in the length of the sequence being folded. Finding an optimal solution is NP-complete!

Various algorithms has been found:

- U-Fold: $\frac{1}{4}$ of the optimal
- S-Fold: Still $\frac{1}{4}$ of the optimal (though often better)
- C-Fold: Between $\frac{1}{4}$ and $\frac{1}{3}$ of the optimal

- Newman found a linear time $\frac{1}{3}$ approximatino
- There are numerous heuristics, like genetic, ant-systems etc.

U-Fold:

- Mark every even H with e and every odd H with o.
- Match even's from the left with odd's from the right (match 1). Then odd's from the left with even's from the right (match 2).
- Pick the biggest of match 1 and match 2
- Fold based on the match, such that every other grid point is occupied by a matched h.

A h can form at most 2 bonds with h's of opposite parity:

$$OPT(S) \leq 2\min(|Even(S)|, |Odd(S)|)$$

We then make a HP fold such that:

$$\begin{aligned} HP - Score(F) \geq & \text{ "Size of matching"} \\ & \geq \frac{1}{2}\min(|Even(S)|, |Odd(S)|) \\ & \geq \frac{1}{2}(\frac{1}{2}OPT(S)) \\ & \geq \frac{1}{4}OPT(S) \end{aligned}$$