**Improving the Prediction of Secondary Structures in Proteins by Finding Coupling Patterns of Hydrogen Bond Pairs**

**Introduction**

Protein structures are vital for the understanding of biological processes. They are also essential for the advancement of many other fields such as material science, medicine and drug design. Yet solving protein structures experimentally is laborious, expensive and time consuming. Protein sequencing, on the other hand, is much faster and cheaper. These differences, in price and difficulty, create an ever-growing gap between the number of protein sequences and the number of solved protein structures. Protein structure prediction (PSP) methods are partially bridging this gap (Schwede, 2013). My work aims to further the development of these PSP methods.

The main approach to PSP is homology modeling. This method is the most widely used because it is fast and easy to implement. However, it is only useful for modeling proteins that have homologues with known structures. On the other hand, the *ab initio* approach can be applied to any protein regardless of the presence of solved homologues, but it is slower more complex, and prone to produce inaccurate models. *Ab initio* methods predict protein structures by performing conformational searches. These searches are guided by energy functions, and involve generating, optimizing and sorting through many conformations for a single protein. This study is mostly relevant to *ab initio*.

The different approaches to *ab initio* PSP use different types of EFs. The direct approach uses physics based energy functions to simulate the actual process of protein folding. This approach yields accurate results, but requires much computational power and takes a long time (Beberg *et al*., 2009). Thus, it is impractical for bridging the gap between the known sequences of proteins and solved structures. The alternative approach uses knowledge based EFs, which are learned by statistical analysis of solved protein structures. They are used to search the space of possible conformations. These knowledge based EFs are less computationally demanding, therefore, most PSP tools use them. My study is particularly applicable to knowledge based EFs.

The modeling of secondary structure elements (SSEs) is one of the processes that make knowledge based PSP more computationally efficient. SSEs are the building blocks of a protein’s tertiary structure, and they mostly consist of alpha helices and beta sheets. Alpha helices only involve short distance hydrogen bonds (SDHB), and are thus very similar in structure. Beta sheets involve many long distance hydrogen bonds (LDHB), and this results in a large variety of types and structures. These SDHB and LDHB consist of backbone hydrogen bonds (BHB), which are hydrogen bonds, formed between carbonyl oxygen atoms and amine hydrogen atoms in the backbone peptide bonds. Treating proteins as blocks of SSEs to be arranged greatly reduces the complexity and thus the computational power required for PSP (Grainger *et al*., 2010).

Modeling SSEs is a useful step in knowledge based PSP but it is also a major challenge. The first step in modeling SSEs is secondary structure prediction, where each residue in the protein sequence is assigned a secondary structure. Nowadays, this step is pretty straightforward and there are software tools available for secondary structure prediction (Jones, 1999). The second step in modeling SSEs is predicting how SSEs assemble. This is simple in the case of alpha helices because they only involve SDHB. The many LDHB present in beta sheets allow the formation of many wrong, beta-like structures in folding simulations. The great variety of beta sheets and wrong structures make predicting beta sheets the primary challenge in modeling SSEs.

The work of Ami Levi, from our lab (Levi-Moonshine *et al.*, 2009), tackles this challenge, by introducing and using the concept of hydrogen bond pairs (HBPs), which are pairs of proximal BHBs in beta sheets. Ami characterized each type of HBP by the relative sequence positions of the amino acids involved in the HBP. Each type of HBP has a grade that represents its frequency in native structures. Frequent, rare and nonexistent HBPs receive high, low and negative grades respectively. Ami used this list of HBPs to define an energy term for PSP and implemented it in the MESHI software package.

In my work I build on and extend the concept of HBP by looking at patterns of HBPs. I first introduce a new concept called HBP coupling. Two HBPs are coupled if they share a hydrogen bond, and each HBP may be coupled with more than one HBP. A pattern of HBPs is a graph where each HBP is linked to every other HBP it is coupled to. These patterns are actually subgraphs of a larger one that holds all the HBP coupling information on proteins from a large and diverse dataset. One of my aims in this study is to build this graph of patterns.

**Methods**

* Meshi and my program for creating a weighted graph of coupled H bond pairs and other files needed for analysis
* Meshi-srv1
* Cytoscape
* Cytostruct plugin and my modifications
* PyMOL and my script for finding proteins containing a subgraph, sorting them by how much of the graph they contain and displaying a protein and coloring the relevant residues and H bonds

**Results**

* The graph

**Discussion**

**References**

Beberg, A.L. *et al.* (2009) Folding@home: Lessons From Eight Years of Volunteer Distributed Computing. *IEEE*, 1-8.

Grainger, B. *et al*. (2010) Re-Evaluating the ‘‘Rules’’ of Protein Topology. *Journal of Computational Biology*, **17**, 1371-1384.

Jones, J.T. (1999) Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *Journal of Molecular Biology*, **292**, 195-202.

Kalisman, N. *et al.* (2005) MESHI: a new library of Java classes for molecular modeling. *Bioinformatics*, **21**, 3931–3932.

Levi-Moonshine, A. *et al*, (2009) Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. *Bioinformatics*, **25**, 2639-2645.

Schwede, T. (2013) Protein Modeling: What Happened to the ‘‘Protein Structure Gap’’? *Structure*, **21**, 1531-1540.