

Unsupervised Pattern Discovery in Protein Structures

Tom Milledge and Giri Narasimhan

Bioinformatics Research Group (BioRG)
School of Computing and Information Science

A. Specific Aim

Pattern Discovery refers to the task of identifying relevant and significant commonalities or essential differences in related data. In bioinformatics applications, patterns come in different flavors – in sequences, structures, shapes, images, and in quantitative and temporal data. Since nature tends to reuse these patterns, pattern discovery is an important task in understanding the nature around us. While lot of research has been done on finding patterns in protein sequences, there is not enough progress on finding patterns in protein structures. Protein structures are known to be assemblies of recurrent secondary structures and structural motifs and domains. Our main goal in this proposal is the following:

- **Aim:** *To implement unsupervised pattern discovery tools for protein structure data.*

Prior research has been limited to structure pattern discovery on small sets of protein structures, often in sets where patterns are either known to exist or very likely to exist. Here we propose to perform the search in an unsupervised manner on the entire *Protein Data Bank* (PDB), a publicly available database of all known protein structures. The result of such a search will be a database (or knowledgebase) of structure patterns in proteins. Such databases will help in the important task of functional annotation of proteins. Databases of specialized substructures have recently been used as the basis for protein structure prediction, a holy grail of Bioinformatics.

B. Relevance to High-Performance Computing

The Protein Data Bank (PDB) is a large database containing 33,931 protein structures (as of Nov 2005). Each such structure is a complex molecule with an average of roughly three hundred amino acids consisting of roughly 5000 individual atoms. A structure pattern present in a set of proteins is a set of atoms configured in space in roughly the same way in each of the protein structures, meaning that when these proteins are structurally superimposed then the atoms corresponding to the pattern occupy nearly the same positions in space.

Finding such patterns is a difficult problem because there are an astronomical number of subsets of atoms that would need to be checked. Our **idea** is to look for common triples of atoms and to then use this as a basis for finding larger common patterns. Towards this end, we propose to build a database (technically, a hash table) of triples of atoms found in all protein molecules. Since the triples may be arbitrarily rotated or translated in space within each protein molecule, each such triple will be stored on the basis of the three inter-atomic distances, which is independent of rotations or translations in space. The problem is that even for one protein, one could end up with roughly a trillion triples, unless these are pruned in meaningful ways. Thus clearly, the problem is very compute-intensive as well as memory-intensive. The problem is also inherently parallelizable, lending itself extremely well to computing with a high-performance grid.

C. Research Design and Methods

We plan to implement time- and memory-efficient algorithms for the problem of unsupervised pattern discovery. Our ultimate goal is to build a search engine, which when provided with a new protein structure, can quickly search through the knowledgebase and inform the user of all proteins with which it shares common substructures, thus performing the equivalent of what the well-known search engine BLAST does for protein and DNA sequences. The first step towards this is to show the feasibility of unsupervised pattern discovery techniques on such a large search space. We intend to start with a smaller version of PDB called ASTRAL40, which contains all proteins from the PDB that have at most 40% similarity. Successful completion of this pilot project will enable us to apply for future external funding to build comprehensive databases and algorithms for pattern discovery and, more importantly, to investigate applications of such patterns.

D. Personnel Credentials

The proposed research will be performed by Tom Milledge, who is a PhD candidate working under the supervision of Giri Narasimhan. Tom's PhD dissertation is broadly focused on finding patterns in protein sequences and structures. Prior work by Giri Narasimhan and his Bioinformatics Research Group (BioRG) on pattern discovery has led to two novel algorithms for supervised pattern discovery. The algorithm, GYM, has been successfully applied to the problem of detecting helix-turn-helix (HTH) motifs and homeodomain motifs in protein sequences; the SSPSite algorithm has been used to find sequence-structure patterns from known sequence patterns. The work on SSPSite was done by Tom Milledge. Parts of this project were also completed for single processor machines by two other Masters students, Minchi Hu and Frank Placencia. The research group BioRG is an extremely active and productive group and further information about this group can be found at the following URL: [\[http://biorg.cis.fiu.edu\]](http://biorg.cis.fiu.edu).