

# Research Statement

Orhan Camoglu

Recent advances in molecular biology have resulted in immense amounts of biological data (DNA sequences, protein sequences and structures, gene expression data and protein interaction data). As a result of this growth, scalable techniques to mine and analyze this data have become paramount. These new techniques employ principles in many different areas of computer science especially databases, data mining, machine learning and graph theory. They also impose new challenges. My research focuses on applying database principles and computational techniques on biological data. I enjoy this area since it has a direct practical impact while being grounded in theoretical principles.

## 1 Summary of past research

As a part of my doctoral research, I developed practical applications and databases for mining and analysis of biological data.

- **Protein structure comparison:** Functional and evolutionary relationships between proteins can be discovered via structure comparison. The problem of finding similarities in protein structure databases gains importance as the size of these databases increases. Current techniques sequentially compare the given query protein to all of the proteins in the database to find similarities. Therefore, the cost of similarity queries increases linearly as the volume of the protein databases increase. As the sizes of experimentally determined and theoretically estimated protein structure databases grow, there is a need for scalable searching techniques. Our techniques extract feature vectors on triplets of SSEs (Secondary Structure Elements) of proteins. These feature vectors are then indexed using a multidimensional index structure. Our first technique considers the problem of finding proteins similar to a given query protein in a protein dataset. This technique quickly finds promising proteins using the index structure. These proteins are then aligned to the query protein using a popular pairwise alignment tool such as VAST. We also develop a novel statistical model to estimate the *goodness* of a match using the SSEs. Parts of this project have been published in ISMB 2003, CSB 2003 and JBCB 2004.
- **Protein Classification:** Proteins are classified using their functional and evolutionary relationships. These relationships can be discovered using a wide range of data sources. We proposed a novel technique for automatically generating the SCOP classification of a protein structure with high accuracy. We achieve accurate classification by combining the decisions of multiple methods using the consensus of a committee (or an ensemble) classifier. We later showed that functional and evolutionary relations of proteins can be captured more accurately using a global similarity network. Even more robust representation of the protein space can be realized if multiple sources of information are used. We propose a novel approach for analyzing multi-attribute similarity networks by combining random walks on graphs with Bayesian theory. A multi-attribute network is created by combining sequence and structure based similarity measures. For each attribute of the similarity network, one can compute

a measure of affinity from a given protein to every other protein in the network using random walks. This process makes use of the implicit clustering information of the similarity network, and we show that it is superior to naive, local ranking methods. We then combine the computed affinities using a Bayesian framework. Parts this work have been published in CSB 2004 and JBCB 2005 and a paper is under review for the Bioinformatics journal.

- **Protein interaction networks:** Genome wide protein networks have become reality in recent years due to high throughput methods for detecting protein interactions. Recent studies show that a networked representation of proteins provides a more accurate model of biological systems and processes compared to conventional pairwise analysis. Complementary to the availability of protein networks, various graph analysis techniques have been proposed to mine these networks for pathway discovery, function assignment, and prediction of complex membership. We have built a functional interaction network for *C. elegans* genome by using a wide range of data sources like microarray experiments, gene profiles, images, ortolog information and literature sources. We also developed graph based techniques for analysis of this network. A random walk based technique was developed for the discovery of missing members of a partially known complex. Later, we extended this algorithm to discover new pathways by iteratively performing random walks. We have also developed a flow based two terminal network reliability algorithm for discovery of missing members of partially known complexes. Our developed tools are available at <http://bioserver.cs.ucsb.edu/>.
- **Nearest neighbor queries:** We considered the problem of detecting *data broadness*. A data object is broad if it is one of the k-Nearest Neighbors (k-NN) of many data objects. We introduce a new database primitive called Generalized Nearest Neighbor (GNN) to express data broadness. We show that the standard k-NN query and its flavors ANN and RNN are special cases of GNN. We propose three methods to solve GNN queries. Our methods arrange input datasets into pages with the help of R\*-trees and rank them based on their distance to each other. This work has been published in EDBT 2006.

## 2 Future directions

- **Structural motif discovery and multiple structure alignment:** Protein motifs, specifically active sites and binding sites, play important roles in biochemical reactions. Since active sites of proteins are determined by structure of the participating amino acids rather than their sequence order, structural motifs can be more useful than sequence motifs. Multiple structure alignment of a set of related proteins results in a consensus structure which has the minimum RMSD sum to the protein structures in the set. Functional and evolutionary core of related proteins can be computed via multiple alignment. A popular technique is to compute pairwise alignments, and to construct a multiple alignment from these alignments. This technique, however, will not result in the optimal alignment. I have proposed a new indexing strategy that summarizes the substructures in protein structures and used this index structure to accelerate the database searches. I want to use similar techniques to answer motif queries and perform multiple structure alignment. By application of index based techniques, I aim to reduce the search space for motif discovery. Moreover, I plan to align multiple structures simultaneously by summarizing the substructures in proteins.
- **Representation of protein space:** The protein similarity networks based on sequence and structure information can be used to perform automated classification as well as to obtain a robust representation of protein space. Techniques I have proposed are not limited to only sequence and structure sources, but are general enough to be incorporated with a wide range of information sources. The

only constraint on the source is that it should be able to give a similarity measure between a pair of proteins. I plan to explore construction of better similarity networks using additional data sources such as motif databases, pathway data and gene expression data. As the diversity of the information sources increases in the future, protein networks constructed using integrated approaches will become more accurate.

- **Protein interaction networks:** Analysis of protein interaction networks provides important clues about protein interactions. We have already constructed a genome-wide protein interaction network for *C. elegans*. This network can be improved by using additional information sources and implementing sophisticated machine learning techniques to extract information from different data sources. Upon completion of this network, we can compare it with other interaction networks of other organisms. This can shed new light on the evolution of interactions.

As more protein interaction data is obtained, new and larger protein interaction networks are being built. New scalable techniques are needed for the analysis of these networks. I plan to develop sophisticated techniques that discover functionally related proteins. With their help, proteins can be clustered into functionally related groups. Additionally, new pathways can be discovered using the functional relationships.