

Sarah Aerni

## Computational Analysis of Gene Regulation and Cancer Genomes

My work in research has been focused on the field of Bioinformatics, particularly in the exploration and creation of tools for analyzing of the human genome and its components. A strong focus has been developing tools which enable biologists to study human disease. In the past I have undertaken three Bioinformatics projects, all of which are ongoing: (i) novel alignment and clustering techniques for DNA sequences from Chronic Lymphocytic Leukemia (CLL) patients; (ii) implementation of DNA motif finding algorithms; and (iii) analysis of chromosomal rearrangements in human tumor cells.

The first project, under the direction of Ben Raphael and Bradley Messmer at University of California at San Diego, involves the creation of tools used to study CLL. The goal of the project is to create tools which permit clustering of DNA sequences from CLL patients. Such clustering can be used for disease prognosis or to point to an antigen triggering the disease, as has been found in other cancers. Existing methods for aligning and clustering DNA sequences are too general and do not make use of specific information present in CLL sequences. I implemented a novel approach to sequence alignment and clustering which uses this information. The programs analyze large numbers of CLL sequences by translating them into alternative alphabets and scoring alignments with customized scoring matrices. The resulting aligned sequences are then clustered using average linkage and maximum linkage hierarchical clustering algorithms. The programs are designed to allow users to change scoring features and to incorporate biological information into the clustering process. The program also includes functionality allowing the creation of random datasets, based on permuting input sequences, permitting the statistical analysis of the significance of clusters based on the frequency of seeing clusters of a certain size randomly. Resulting clusters can then be further analyzed for biological significance. The approach has shown promise and was well received by the CLL community which expressed a strong interest in making the tools publicly available for use in their research.

The second project includes work on motif finding programs which find regulatory elements of genes. This work has been done under the direction of Barbara Wold at Caltech, for incorporation into a package, Cistematic (created by Ali Mortazavi), whose goal is to identify such regulatory elements across multiple genomes. Motif finding programs have been largely ineffective in being able to identify regulatory elements, particularly in the human genome. Therefore, the goal of this project is to develop methods that are more effective across a span of different organisms. The motif finding techniques being implemented by the various tools I am creating include greedy algorithms, Gibb's sampler, and expectation maximization. The primary difference between the tools currently available and these, is the incorporation of a background model which accounts for the dependence of nucleotides based on their sequence position. Unlike many other programs, there is no assumption of independence, and therefore a Markov model has been developed to determine background. While the project is still in progress, Cistematic is already in use by groups of researchers at Caltech and California State University LA, and will soon include the motif finders I have created.

The third project involves studying various tumor genomes to discover "hotspots" for genome rearrangement. Since a tumor genome is a rearranged version of the normal human genome, the locations of rearrangement can be found by sequencing short tags derived from the ends of fragments of DNA from tumors. I wrote programs which used information about the tag sequences to exclude possible mis-mappings and to determine whether rearrangements had taken place. Potential genome rearrangements hotspots were identified by examining the normal human genome at the locations of the tag sequences. The results of this analysis suggest that some rearrangements are located in hotspots and a biological mechanism for some genome rearrangements in tumors. This hypothesis is currently being tested experimentally.