Development and Testing of Algorithmic Solutions for Problems in Computational Genomics and Proteomics

By

Thiruvarangan Ramaraj

Prospectus
For
PhD Comprehensive Exam

College of Engineering
Department of Computer Science
Montana State University- Bozeman
Montana

October 08 2008

Abstract:

Molecular biology has experienced an exponential growth in terms of access to data, volume of results, and conceptual discoveries in recent years. Making full use of the vast amount of molecular biology data poses challenging information science questions on data retrieval, data mining, and data management. Molecular biology has become a computational science. Problems like Multiple Sequence Alignment for structure prediction, reconstructing haplotype data from genotype data, discovering protein classes and characterizing proteins from 2D gel electrophoresis and other proteomic data, building signaling network maps, aligning DNA fragments, guessing evolutionary trees, reconstructing large sequences from smaller ones, and many other issues can be cast in mathematical models. All of these questions can be considered optimization problems to be attacked by means of standard techniques. With the enormous quantity of data available today modern molecular biologists seek the use of computer science approaches in order to solve biological problems. Thus, the field of molecular biology provides computer scientists with numerous challenging practical and theoretical questions.

These issues in computational molecular biology have led to interesting theoretical computer science problems; both in algorithm design and complexity analysis. Numerous uses of graphs and Network Flows represent biological phenomenon at many conceptual levels. A few hundred papers have been published exploiting or extending non trivial classic graph theory for problems in biology. For example the Euler path approach is used in Genome Sequencing problems like Genome Assembly, and applications of tree-design or network-design approximations are used in multiple sequence alignment problems.

We introduce the three problems 1. Antibody Imprinting, 2. Patterns of Protein Expression and 3. Genome Assembly in this proposal, followed by formal description, background, related work, research plan, line of attack and time line of expected progress for each problem.

Introduction to Problem 1: Antibody Imprinting

In the first problem we present a new Multiple Sequence Alignment method to improve the alignment of peptide fragments representing antibody epitopes to discontinuous regions of the one dimensional amino acid sequences of target proteins. A large fraction of protein structures of interest cannot be solved by traditional structural biology techniques such as X-ray crystallography and NMR (Nuclear Magnetic Resonance). Antibodies can either recognize continuous or discontinuous epitopes, but virtually all epitopes that have been analyzed in detail are discontinuous. Discontinuous epitopes can potentially provide extremely useful structural information because with suitable analysis they could reveal distant segments of primary sequence that are in close proximity on the native, folded protein and could reveal changes in protein structure in different functional states when appropriate antibodies are available[2].

In this approach peptide probes selected experimentally from a random peptide library to have a high affinity to antibodies of interest. The computational problem addressed in our previous work on this issue was to align each probe individually to the target protein. These alignments were performed with a program called EPIMAP[1]. Our goal here is to improve the current EPIMAP program in such a way that instead of aligning each probe individually to the target protein, it aligns all the probes simultaneously to the target protein and also considers the substitutability of the aligned probe sequences. We hope that this might lead to better epitope predictions. We want to formalize this as a Multiple Sequence Alignment problem and describe an efficient algorithm that will work well in practice for inputs of interest.

Introduction to Problem 2: Patterns of Protein Expression

In the second problem our aim is to present an efficient algorithm to discover groups of proteins that are up or down regulated as a function of biochemical stimulation that would help characterize responses of cells or tissues as monitored in 2D gel electrophoresis experiments. 2D gel electrophoresis is a widely used technique to separate proteins and modified proteins and can be employed in a multicolor format to detect changes in the levels of proteins and modified proteins in systems of interest. In the simplist multiplex experiment proteins from experimental and control samples are labeled with different colored fluorescent dyes, mixed together and separated on 2D gels. These gels use two sequentially performed separations in acrylamide media. The first dimension is isoelectric focusing, that separates by amino acid composition, and the second dimension separates by molecular weight. This produces a 2D pattern of spots each representing a protein or a modified form of a protein. These 2D patterns are detected by scanning the gels with lasers to detect the fluorescently-labeled proteins.

The different colored images are normalized by their total intensity and the ratios of the different protein forms determine. Protein spots that exhibit statistically significant changes between experimental and control samples are then analyzed by Mass Spectrometry (MS).

The main goal in this project is to look more closely at the experimental data and identify patterns of clusters of proteins that not only change the most in size and intensity but also to consider the patterns of proteins that undergo slight but correlated expression changes. We want to develop a combinatorial algorithm to this problem to identify interesting protein clusters and in many cases try to solve it to optimality.

Introduction to Problem 3: Genome Assembly

Genome assembly refers to the process of taking a large number of short DNA sequences which are generated by shotgun sequencing and put them together to create a representation of the original chromosome sequences from which the DNA originated. Assembling genomes computationally is challenging because genomes often contain large numbers of identical "repeated" sequences (repeats). These repeats can be thousands of nucleotides long and can occur in different locations, especially in large genomes. I studied and reviewed different computational algorithms and mathematical formulation that have been used in the currently available assembly software platforms to gain an appreciation of the breadth and depth of their potential. In this third and final problem I will be working on computational genome assembly and identifying copy number variations. I will be seeking structural variants that would lead to the discovery of gene(s) responsible for the cause of MS in one and not in the other in the genome sequence of Identical Twins. I will be working at the National Center for Genome Resources (NCGR) on this project, were the genome will be sequenced under the supervision of Dr. Joann Mudge, Research Scientist at NCGR, and the computer science work will be supervised by Dr. Brendan Mumey, Associate Professor at Montana State University –Bozeman, Montana.

Problem 1

Algorithm Improvement and Validation Using Known 3D Structures for Mapping Discontinuous Antibody Epitopes to Reveal Structural Features of Proteins

Introduction:

Proteins are large organic compounds composed of linear polypeptide chains. To fully understand the biological role of a protein one requires knowledge of its structure, function, and mechanism. There are about 1,000,000 different protein forms in human cells and each protein form has its own folded functional structure. Whenever the three dimensional structure of a protein could be determined, the information has provided important insights into mechanisms of action and may be extremely useful in drug design. With the increased amount of proteins becoming available traditional methods like X-ray crystallography or Nuclear Magnetic Resonance (NMR) are often not feasible for protein structure determination. Therefore, computationally predicting the structure of proteins is an increasingly popular approach are becoming increasingly popular.

Background:

This project is a collaboration work with Dr. Algirdas Jesaitis from the Microbiology Department and Dr. Edward Dratz from the Chemistry and Biochemistry department at Montana State University. One of the main aims of biological research is to understand how cells work and define the rules by which they interact and function. A main concept is “form follows function”, which means that if we know the detailed shape of a molecule on a molecular scale we are then much more likely to understand the function of that molecule. Antibodies that bind to the surfaces of a protein of interest can carry information on the structure of the protein and the present project strives to extract that information.

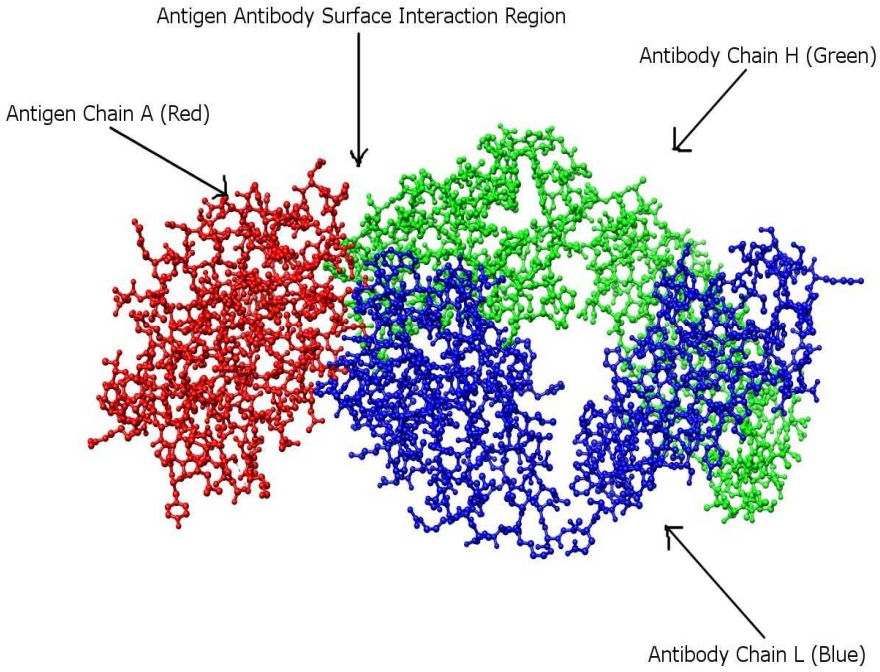
The general structure of all antibodies is very similar, but a small region at the tips of the protein is extremely variable. Differences in the variable regions allow millions of antibodies with slightly different binding site structures to exist. This region is known as the hyper variable region. Each of these variants can bind to a different target, known as an antigen. This huge diversity of antibodies allows the immune system to recognize an equally wide diversity of antigens. The unique part of the antigen recognized by an antibody is called an epitope. The alignments of the antibody epitopes to the discontinuous regions of the one dimensional amino acid sequence of a target protein indicates how segments of the protein sequence must be folded together and provide long range constraints for providing information on the 3-D protein structure. Antibodies can recognize either continuous or discontinuous epitopes, but all of the known antibody-protein antigen structures have discontinuous epitopes (Rubenstein, et al, 2008). Discontinuous epitopes can provide useful structural information in the antibody imprinting process, by revealing segments of primary sequence that are in close proximity on the native, folded protein.

Related Work:

The Antibody Imprinting Method:

Antibody imprinting works by using antibodies to select peptide sequences from large random peptide libraries [2]. The sequences selected by the antibody are called “probes” and after several rounds of selection and enrichment have relatively high affinity for the antibody (that binds the protein of interest). Our group has developed an algorithm called EPIMAP that computationally evaluates which protein residues form the epitope surface, based on the probe sequences selected and the one-dimensional antigen protein sequence.

Step 1: Find an antibody that binds the protein under conditions of interest.



1lk3.pdb(Source:www.pdb.org)
Antigen Antibody Reaction (Using Chimera Program)

Step 2: Probe the antibody with a random peptide library of short peptide sequences (displayed on bacteriophage or alternatively on ribosomes, yeast or other library host).

Probe sequences are used to witness protein structure. The bacteriophages that bear the strongly binding peptide probes are are diluted to form individual clones on bacterial lawns. A number of clones are picked and the DNA that codes for the selected peptide is amplified by PCR and the DNA is sequenced from the selected phage DNA clones. We then carry out our analysis using a number of these selected peptide sequences that serve as a witness to discontinuous epitopes and thus the folded structure of the protein

EPIMAP Program:

A new algorithm was developed to computationally predict the epitope where the antibody binds to the target protein, based on aligning individual mimetic probe sequences. The probe amino acid sequence, s, is aligned to one or more regions of the target protein amino acid sequence, t. The probe amino acid sequence(s) is usually about 8 -20 amino acids long and the target amino acid sequence (t) is several hundred in length. This method is approached as a more complex combinatorial alignment problem than is commonly used to assess relatedness of protein or DNA sequences. This method allows sequence reversals of the probe amino acid sequence to align to the underlying protein (target) sequence and complex gap patterns are allowed in matching probe and the target sequences. These gaps can be large numerous when the epitope is discontinuous.

To evaluate the quality of the alignments this method uses a two part scoring system that is composed of a substitution score and an epitope gap cost. The substitution score is calculated with a substitution matrix and the epitope gap cost is calculated by examining the number of amino acid residues skipped along the target amino acid sequence between successive aligned probe positions. The computational problem is thus to find an alignment that maximizes the score. Different gap cost models have been used to evaluate the quality of the alignments. EPIMAP program implements this algorithm and produces a list of the top-scoring alignment(s) of the probe to protein. Typically 50-100 probe sequences will be known experimentally and are individually aligned using EPIMAP.

Input: The linear amino acid sequences of the peptide probe and the target protein.

Output: A mapping of probe positions to target positions (usually including insertions and deletions).

Sample Alignment Result:

```
                                  . G IGN DP.IRW      .         .         .         .   probe GIGNDPIRW bp 0 score 6.55 22x
                                    G IGN DP.IRW      .         .         .         .   probe GIGNDPIRW bp 0 score 6.55 2x
                                                                .  WRI P DN  GMG    .   probe GMGNDPIRW bp 0 score 6.68
                                                                .  WR LP DN  GIG    .   probe GIGNDPLRW bp 0 score 6.75
                                                                       .GNGVGGRS .      probe GNGVGGRSV bp 0 score 7.07
                                                                   WRI P DNGNG      .   probe GNGNDPIRW bp 0 score 7.13

...AL8EMIQFYLEEVMPQAENQDPDIKAHVN8LGENLKTLRLRLRRCHRFLPCENGGG8GGK8KAVEQVKNAFNKLQEKGIYKAM8EF...  <- Target Protein
      ^ ^^^ ^^^ ^^^ ^^^^^^^^                                    ^  ^^^ ^^^^^^ ^ ^  ^        <- known epitope positions
...5678901234567890123456789012345678901234567890123456789012345678901234567890123456789...  <- target numbers
...     6         7         8         9         0         1         2         3         4...
```

To solve the alignment problem a branch-and-bound algorithm was used[1]

Research Plan:

Improvement of the Algorithm:

In the above described work prior efforts developed the EPIMAP algorithm that computationally evaluated the epitope based on aligning individual mimetic probe sequences derived from an experimental process called antibody imprinting for the protein of interest. The goal of my work is to improve the current approach to examine and evaluate all possible ways to map all the peptide epitope sequences (probe sequence set) simultaneously onto the sequence of the target protein (protein target sequence) in question, which would greatly increase the size of the search space but could well lead to better epitope predictions. . This problem can be approached as a Multiple Sequence Alignment problem.

The Multiple Sequence Alignment Problem:

As mentioned previously the principal prior goal of protein sequence alignment is to discover biological similarities among proteins. Multiple sequence alignment can be a useful technique for studying and analyzing sequence-structure relationships. So Multiple sequence alignments remains an important area of research as biological inferences can be made from the conservation or variation within the aligned positions, especially with reference to the structure of at least one of the aligned sequences.

In protein sequence alignment, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region is among lineages. The absence of substitutions, or the presence of only very conservative substitutions in a particular region of the sequence, suggests that this region has structural or functional importance. In the previous EPIMAP approach only two sequences are aligned at a time, one probe sequence at a time against the target sequence. Multiple sequence alignment incorporates more than two sequences at a time. In our approach the goal would be to align all the probes sequences against the target sequence at once. This method should allow improved identification of the conserved sequence across the target. This conserved sequence can be used in conjunction with the structural information of the target protein.

Multiple sequence alignment problems are computationally difficult to produce and most formulations of the problem are NP complete combinatorial problems. To solve this probe-target alignment problem a branch-and-bound algorithm will be used to find optimal and suboptimal alignments and also to prove that the corresponding decision problem is NP complete.

Secondary Structure Scoring Function:

As mentioned earlier proteins are composed of amino acids that are joined together in peptide chains. The size and shape of the different amino acid side chain cause space constraints that limit the range of rotation of the bonds adjoining the peptide bonds. The two of the most easily achievable conformations correspond to the most common forms of protein secondary structure, namely the α-helix and β-sheet. The α-helix is formed when the amino acid backbone curls around at 3.6 amino acids per turn. β-sheets are formed when two adjacent strands of peptide lie in a plane and form hydrogen bonds between their respective backbones. The chains in are formally 2.0 amino acids per turn with 180 degree turns between residues.

The current EPIMAP version does not take into account any gap cost evaluation for the secondary structure. An important objective is to improve our scoring function by including secondary structure considerations.

Proposed Lines of Attack:

1. To develop means to examine and evaluate and implement Multiple Sequence Alignment methods in which all the epitope-mimetic peptide (probe sequences) can be mapped onto the sequence of the target protein in question simultaneously, to better recognize discontinuous epitopes that provide proximity constraints on the 3-D structure of the protein.
2. Study the Computational Complexity and devise efficient algorithms to find optimal and suboptimal solutions
3. Generate synthetic data sets which can be used for machine-learning and cross-validation approaches.
4. Test the systems devised on cases where the epitopes are known from antibody-antigen x-ray structures and experimental probe sequences have been obtained.

Exploring Synthetic Data sets:

Introduction and Motivation:

The problem addressed in this section is to mine the PDB for known protein-antibody complexes in order to develop synthetic test cases to tune the performance of EPIMAP. There are over 225 examples of known protein-antibody structures in the PDB [e.g. Rubinstein, et al, 2008]. We have done a significant amount of work in this research to date and have developed an automated system for creating synthetic test cases for EPIMAP from these structures. These test cases permit the efficient training of scoring parameters within EPIMAP with a goal to improve its overall epitope-prediction accuracy.

We follow a standard machine-learning cross-validation approach in which some of the data is held out as a test set and the remaining data is used to train with. In order to create the training data, we have developed a random walk based approach on the antibody based on certain rules for generating probe sets. While we are still collecting test set data, our initial results appear to have been useful in refining the performance of EPIMAP.
We expect that our synthetic data sets will be very useful in studying and understanding the problem much better and further improve the performance of EPIMAP. This initial work was presented in part as a poster at the New Mexico Bioinformatics Symposium, Proteomics including NanoBioIT, March 8th – 9th 2007, Santa Fe, New Mexico.

Methodology:

In our approach we identify 10 PDB (Protein Data Bank) antibody-antigen complexes. Out of the 10 complexes 5 are used as training data sets and 5 are used as test data sets. We identify the antibody region that lie close to the antigen surface [how] and create a set random probes sequences each of length 9 by doing a random walk on the antibody residues following certain rules. Once we have a set of synthetic probe sequences we mutate them using two amino acid substitution probability matrices. The matrices used are the BLOSUM40 substitution matrix and the Tangel frequency of interaction matrix and retain an unmutated set of probes (equivalent to mutating with an Identity Matrix). After the data sets were generated we run them on EPIMAP, using a range of gap scoring parameters. Table 1 (title it) shows some of the selected results obtained.

PDB Code for complex	Sub. Matrix	Prob. Matrix	Maximum Gap Cost	Gap Extension Cost	Alignment Score (Before Filtering)	Alignment Score (After Filtering)
1BVK	tangel	Tangel	2	0.3	1.72	1.37
1BVK	Pearson	Identity	1	0.3	2.92	2.58
1C08	tangel	Tangel	2	0.3	1.60	0.41
1KB5	tangel	Tangel	1	0.3	0.99	1.05

Table 1: EPIMAP Score for the synthetic data set generated.

EPIMAP Parameter Space Explored:

Maximum Gap Cost: 1.0, 1.5, 2.0
Gap Extension Cost: 0.3, 0.5, 0.8
Maximum Top Scores to keep: 3
Maximum Top Solutions to keep: 10
Substitution Matrices Used: Pearson Matrix, tangel Matrix and Argos Matrix.

For each of the 5 training cases we have 81 different combinations of EPIMAP Parameters that were run through EPIMAP. The results from this program are then fed into another program that our group developed called the EPIFILTER program. This program selects the best mutually compatible alignments (one for each probe sequence in the set). After we get the results we pick out the parameters that yielded the highest alignment score. Using these selected parameters we run the 5 test cases that were generated at the beginning with the EPIMAP and EPIFILTER program.

Alignment Score Mechanism:

The alignment scoring mechanism that we have usee so far is simple. The average frequency for the epitope residues within the proximity to the antibody and divide by the average frequency of all the residues

If, α =Average frequencies for the Epitope Residues
If, β = Average frequencies for the entire protein

Then the Alignment Score (S) is defined as $S = \alpha / \beta$

At the present time not all of our experiments have been run (several hours of computational time is required for each EPIMAP run). We found some interesting results for a selected set of cases. This is a work in progress and we have not found one single combination of EPIMAP parameters that works well in all cases, at this point in the process. We expect that synthetic data sets will be invaluable [how can you be so confident that the random walk sets will be invaluable—maybe the selection by the antibody is more guided by preferred interactions—just need to have thought this out] for improving and refining EPIMAP and the antibody imprinting approach to computational epitope prediction.

Problem 2

Analysis of 2D Electrophoresis Gel Data using Differential Two Color Fluorescent Dye Detection

Project Overview:

This is a local collaboration project with Dr. Edward Dratz's laboratory from the Chemistry and Biochemistry Department at Montana State University. The general goal of this research project is to discover and implement an efficient combinatorial algorithm for our research work in proteomics. This project is about analyzing 2D gels which are produced by high resolution gel electrophoresis techniques. The currently used techniques used focus on the that undergo the largest changes in amounts resulting from biological stimulation or disease states compared to controls. The main goal of this project is to investigate 2D electrophoresis gel patterns and identify the protein clusters that undergo correlated modifications and to seek to identify the global pattern of change. We want to develop and implement an algorithm to identify clusters of proteins and use the information in the pattern change to characterize the proteins in the samples.

Introduction:

Biomedical researchers are very much interested in understanding what proteins are involved in biological responses and understanding the networks of responses. Proteomics is an experimental approach that analyzes all of the proteins in samples and identifies the proteins that change in response to biological stimuli or disease state. Computational proteomics concentrates on identification of proteins that change with stimuli, and characterization of protein modifications that change qualitatively and quantitatively. A proteomic experiment involves the analysis of samples that contains complex mixtures of proteins that typically occur at a wide range of different concentrations. The process of identifying proteins from these biological samples is complex and the low abundance proteins are mostly not observed with current methods. New, more powerful detection methods are under development (at MSU and elsewhere) and prospects are good that changes in the lowest level proteins will be detectable in a year or so. The most dominant proteomic methods used are 2D electrophoretic gels and shotgun liquid chromatography. After 2DE proteins are digested with proteolytic enzymes and identified by mass spectrometry. In the shotgun method the complex mixtures of proteins are digested into hyper complex mixtures of smaller peptide components of the proteins before LC-MS/MS.

Large Scale Protein Analysis:

The most widely used method for defining a proteome and global changes in the protein has been 2D electrophoresis. Gel electrophoresis is a very common technique used in the separation of protein molecules using an electric current applied to a gel matrix. 2D gel electrophoresis (2-DE_first separates proteins based on their isoelectric point (pI), in the first dimension and separates by molecular weight in the second dimension. Plain 2D electrophoresis gel is an older technique and recently it has been improved using

multicolor, multiplex detection. In the new approach different protein samples are covalently labeled with different fluorescent dyes, the different colored labeled proteins are then mixed together and separated on the same gel. After separation the patterns of proteins and are detected by laser scanning. Covalent labeling of protein samples with different-colored fluorescent dyes before 2-D electrophoresis enables much more accurate analysis of differences in protein abundance between samples. The end product of this experiment is a pattern of spots and smears that displays each protein and modified protein in a complex proteome. The preferred technique used to identify each spot of interest is the Mass Spectrometry (MS) for determining the identity of the proteins or modified proteins. The analysis of such image series intends to identify those proteins or modified that change in abundance, reflecting biochemical and biomedical conditions in an organism.

Research Plan:

Source of Experimental Data:

For this project we will be using several data sets created by Dr. Dratz's laboratory from the Chemsitry and Biochemistry Department at Montana State University and their collaborators at other centers. One data set consists of intensity values and x,y coordinates representing protein in human plasma comparing type 2 diabetic with healthy control humans. Another data set studies the responses of human white blood cells to innate immune stimulators and a third data set studies the responses of developing rat or mouse brains to essential fatty acids.

Initial Work:

The data obtained looks pretty imposing. It lists spots vertically (~680) with their spot intensities under "blue", "green" and "red" columns referring to the colors of the dye used to image the spots. The blue dye is ignored in the initial data sets as it had a high background and an improved blue dye has been designed and prepared for future experiments. The green and red are labeled either diabetic or control (healthy). All the pools were compared for both green and red dyes and found that the red dye was less "well-behaved" under the experimental conditions used in the initial experiments, in that it seems to introduce more noise. Future experiments will use modified experimental conditions that will provide better red dye data. So in this initial data set we can try comparing diabetic with healthy just using the "green" data.

Spot	t- Test	p-Value
fraction 1, spot X 1624 Y 800, green cd_align	0.999811136	8.20304E-05
fraction 1, spot X 830 Y 1046, green cd_align	0.997290263	0.001178421
fraction 1, spot X 1927 Y 1040, red cd_align	0.996856532	0.001367341
fraction 1, spot X 306 Y 1326, green cd_align	0.990557699	0.004120222
fraction 1, spot X 2001 Y 226, green cd_align	0.990031398	0.004351032
fraction 1, spot X 324 Y 515, red cd_align	0.985124241	0.006508994
fraction 1, spot X 974 Y 439, green cd_align	0.985080862	0.006528118
fraction 1, spot X 1451 Y 796, red cd_align	0.980105802	0.00872704
fraction 1, spot X 853 Y 1546, red cd_align	0.979769676	0.008876006
fraction 1, spot X 1356 Y 966, red cd_align	0.979296971	0.009085589
fraction 1, spot X 728 Y 969, red cd_align	0.977438787	0.009910431
fraction 1, spot X 1934 Y 803, red cd_align	0.970174836	0.013149994
fraction 1, spot X 240 Y 376, red cd_align	0.96766203	0.0142763
fraction 1, spot X 1476 Y 1366, red cd_align	0.967233007	0.014468892
fraction 1, spot X 851 Y 1376, green cd_align	0.965438252	0.015275498

Table: Top 15 t-Test spots

By looking into the data the interesting columns were the normalized value column for the control, green in this case and normalized value column for diabetic sample, red in this case. We calculated the ratios of diabetic to control for each sample. It was interesting to see that the ratios were somewhat different when the dyes are interchanged. The ratios were different when the dyes were interchanged apparently because different dyes quench the proteins to a somewhat different extend. Also a t-test was computed on all the spots and sorted by significance. In this data we found that the green dye is definitely outperforming the red, but there were still some significant red spots.

Microarray data is used in the systematic study of gene expression, their functions, and their interactions. Analogously, proteomics is the study of proteins, protein complexes, their localization, their interactions, and posttranslational modifications. Technologies for DNA microarray data and proteomics data investigations are increasingly leading to new insights into mechanisms of living systems.

Proteomics data evokes all proteins in any given cell and also the set of all protein isoforms and modifications. Proteomics complements functional genomics approaches, including microarray-based expression profiles, systematic phenotypic profiles at the cell and organism level, systematic genetics and small-molecule-based arrays. Proteomics would not be practical without the previous achievements of genomics, which provides the gene sequences and the blueprint of possible gene products that are the focal point of proteomic studies.

The next step would be is to try some simple clustering of the spots using k-means and also some hierarchical techniques.

At first we wanted to try some existing software like GeneSpring and/or MeV and Principal Component Analysis

Algorithm Implementation Approach:

Clustering data has been widely recognized as a powerful technique in Computer Science and Statistics and is being studied extensively in recent years. Significant amount of research has been done in discovering interesting classes of genes in microarray data based on their expression similarities across multiple samples or conditions. Proteomics is often considered next step in the study of biological systems after genomics. In our initial approach we treated the proteomics data same as the microarray data. The input is in the form of a m*n matrix. Each row represents the intensity of each fraction for different samples and dyes. There have been many efforts carried out recently on analyzing proteomics data. The main goal in Clustering to classify objects into different groups, or more precisely, partitioning of a data set into subsets, so that the data in each subset share some common trait.

Data clustering is a common technique for statistical data analysis and is used in many fields. Clustering algorithms are applied to these data to discover expression patterns that may be the key in understanding many genetic pathways. Many different approaches have been previously developed for clustering proteins. There are abundant software packages available to identify subsets of interesting proteins. The main objective of this research project is to identify interesting clusters of proteins and look into the global pattern of change. The currently used methods only take into consideration the proteins that change the most and neglect the protein patterns that change in more subtle patterns and thus neglect global pattern of change. We suspect that looking into the patterns of change could well provide valuable information in characterizing the networks of proteins interacting in the samples.

We hope our algorithmic approach would be capable of discovering the expression patterns for the proteins that have less striking changes in the 2D gel data. We hope for moderately sized problems we can often guarantee that the solution found is optimal. The input to this algorithm would consist of data with each fraction (spot) against the normalized peak values (spot intensity) for the green dye and the red dye which are used to image the spots. The patterns would be observed in a time series after stimulus, in a dose-response patter, or in the present case—a pattern of protein changes that occurs in comparing samples from numerous patients and controls or several groups of patients and controls. The data is typically represented in large expression matrices.

Problem Complexity:

The complexity of the clustering problem will depend on the exact problem formulation and the various functions that will be used in the evaluation of the cluster. We expect that the related optimization problem will be NP – complete.

Proposed Lines of Attack:

1. Since we are looking at the 2D gel electrophoresis data we have normalized peak values for each sample in this case diabetic versus control. First green dye is used for control and red dye is used for diabetic, then they interchange the dyes by using red for control and green for diabetic. As a first line of attack we did some simple analysis by calculating the ratio of diabetic to control in each sample. Some preliminary results are included in the expected results section.

2. Adapt microarray clustering techniques like k-means clustering and some other hierarchical techniques to the experimental data available to try some simple clustering of the spots.

3. Describe a clustering algorithm to apply to these data to identify interesting classes of proteins and look at their expression patterns to seek understanding of functional protein networks.

Expected Results:

We want to develop an efficient algorithm for analyzing and clustering proteins to identify the pattern changes and characterize the proteins in the samples and hopefully find new insights into proteins relevant to diabetes. We would also like to use other available clustering approaches for the analysis of the protein expression data and compare it with the algorithm implemented in this work.

Problem 3

Multiple Sclerosis (MS) – Identical Twins Genome Sequencing Project

Introduction:

Contrary to our previous beliefs, identical twins are not genetically identical (references!!). The presumption has always been that identical twins are identical down to the details of their DNA sequences. But recent genetic studies have shown the DNA of identical twins are not always identical.. Sometimes there are subtle differences in the DNA due to the Copy Number Variations (CNVs)(references). Researchers suggest that these subtle differences may lead to a better understanding of genetic diseases when studies are conducted on discordant monozygotic twins (A pair of twins where one has a disorder and the other does not). In those discordant monozygotic twins, one might have diseases like Multiple Sclerosis (MS), Parkinson's and the other does not. Studies indicate that CNV might play a critical role and this can be effectively studies in identical twins.

Copy Number Variation:

Copy number variation of DNA sequences refers to the differences in the number of copies of a particular region in the genome. Deletions, Insertions, and complex multi-site variants are collectively termed copy number variations or copy number polymorphisms (CNPs). Recent multiple studies have discovered copy number variation of DNA segments ranging from kilo bases to mega bases in size in all humans and mammals and other organisms examined. Differences in the DNA sequence of our genomes contribute to each individual's uniqueness. CNVs are important because they can lead to expression of different amounts of proteins and modified proteins and these submicroscopic variations can influence many traits, including susceptibility to disease.

The CNVs influence gene expression, phenotypic variation and adaptation by disrupting genes and altering gene dosage and can cause disease, as in micro deletion or micro duplication disorders or confer risk to complex disease traits such as HIV-1 infection [reference] and glomerulonephritis [30]. CNVs often represent an appreciable minority of causative alleles at genes at which other types of mutation are strongly associated with specific diseases: CHARGE syndrome21 and Parkinson's and Alzheimer's disease [30]. Furthermore, CNVs can influence gene expression indirectly through position effects, predispose to deleterious genetic changes, or provide substrates for chromosomal change in evolution [30].

CNVs that are benign variants will not directly cause disease, but there are several instances where CNVs that affect critical developmental genes do cause disease. Genes that are involved in immune system and brain development and activity tend to be enriched in CNVs. Capturing all the variation in the genome will help identify those genes that play important roles in human health, hunting for genes underlying common diseases.

Multiple Sclerosis (MS):

Multiple Sclerosis, also known as disseminated sclerosis or encephalomyelitis disseminata is a complex, inflammatory, autoimmune, demyelinating disease of the central nervous system, affecting nearly one million people worldwide. It strikes most often the early adult years. Its symptoms include numbness, impaired vision, and loss of balance, weakness, bladder dysfunction, and psychological changes.

Genetics and Multiple Sclerosis:

Research scientists and physicians who have heavily studied the question of what makes humans susceptible to MS are convinced that the genes that we inherit from our ancestors are an important factor. Many research groups have searched for the genes that make humans susceptible to developing MS. Understanding how genes contribute to determining who gets MS will provide major clues to the cause and may point to ways of preventing and treating MS. Even though in the classical sense, MS is not considered as an inherited disorder, there is strong evidence to support the fact that genetic factors play an important and significant role in MS.

Next Generation Sequencing Technology:

For the past several years, the main sequencing DNA methodology has been Sanger sequencing. This sequencing method has major limitations and remains prohibitively costly and time consuming for many genome projects. Recently there have been many sequencing technologies developed that have the potential to overcome these limitations, but at the same time the data produced by these technologies pose new challenges in assembling the sequence reads.

The new sequencing technologies, called next-generation sequencers, have the ability to process millions of sequence reads in parallel. The next-generation sequencers are much less expensive to use and are much faster. Some of the commercially available sequencers are:

1. Roche (454) GS FLX sequencer
2. 2. Illumina genome analyzer
3. 3. Applied Biosystems SOLiD sequencer.

All these sequencers produce shorter read lengths, 35 – 250 Base Pairs depending on the platform, than capillary sequencers--which produce 650 – 800 base pair read lengths. The shorter reads are harder to assemble and can impact the utility of the data for various applications such as de novo assembly and genome resequencing.

Genome Sequencing of Identical Twins:

From the numerous research studies conducted it is clearly evident that identical twins, genetically identical from conception, are not identical for various traits and diseases. This project involves the complete genome sequence of twins where one has MS and the other the other does not. I will be working at NCGR (National Center for Genome Resources) under the direct supervision of Dr. Joann Mudge, Research Scientist at NCGR and Dr. Brendan Mumey, Associate Professor, Department of Computer Science here at Montana State University – Bozeman.

National Center for Genome Resources (NCGR) in Santa Fe, New Mexico, is a nonprofit research institute dedicated to improving human health and nutrition, focusing on collaborative research at the intersection of bioscience, computing, and mathematics. Research scientists and software engineers at NCGR study the influence of genetic variability of both host and pathogen on infectious disease progression and develop scientific software solutions to support and enable those studies.

The sequencing technology used at NCGR's Genome Sequencing Center is Illumina Genome Analyzer. Illumina Genome Analyzer System is based on the Solexa sequencing technology, providing a high-speed, massively parallel genetic analysis system for genetic analysis and functional genomics. Some of the highlight features of this technology are, it has scalable ultra-high throughput and it requires sample input as low as 100 ng - 1µg, enabling a host of applications where sample is limited. Also it is simple, fast, and automated.

NCGR has undertaken the research project of sequencing the genomes of identical twins, one having MS and other does not. I will be a part of this research project working primarily on assembly, assembly validation, assembly post processing, identification of Copy Number Variants (CNVs), Single Nucleotide polymorphisms (SNPs) Structural Variants and difference discovery.

Genome Assembly:

An important and a critical step in the sequencing of genomes is to assemble all the short reads or fragments that are extracted from the sample to form a set of contiguous sequences (contigs), that represents the DNA in the sample. Assembling genome computationally is challenging because genomes contain large numbers of identical sequences (repeats). These repeats can be thousands of nucleotides long and can occur in different locations, especially in large genomes.

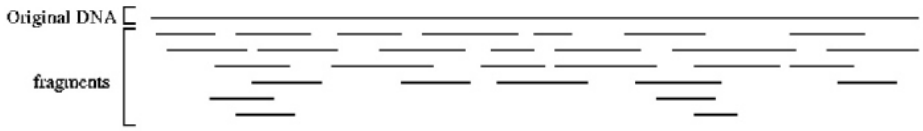


Figure 1. Whole Genome Shotgun Sequencing

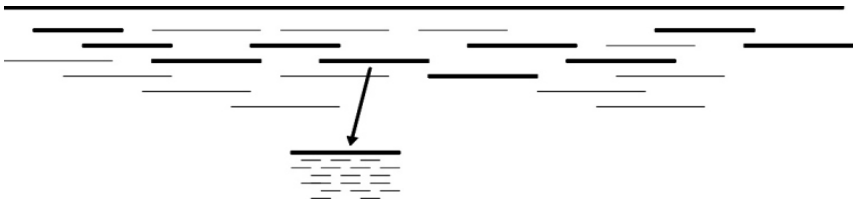


Figure 2. BAC by BAC (Hierarchical) Sequencing

Figure 1 represents shotgun sequencing where the DNA sequence of an organism is sheared into a large number of small fragments. Figure 2 represents the BAC by BAC Sequencing approach, where the long lines represent individual BAC's. A minimum tiling path (thick lines) is chosen, such that each base in the genome is covered by at least one BAC and the overlap between the BAC's is minimized. Each BAC in the tiling path is sequenced through the shotgun method.

Comparison: Sanger Reads Vs Solexa Short Reads:

The following table shows how much the amount of sequence read increase from Sanger technology to the Solexa short read technology for different organisms

Organism	Genome size	Sanger Reads with 8X Coverage (1000 nt)	Solexa Short Reads with 100X Coverage (36 nt)
Virus, Phage Φ-X174	5,400	43	15,000
Bacterium *Escherichia coli* (million reads)	4,000,000	32,000	11,111,111
Nematode, *Caenorhabditis elegans* (million reads)	98,000,000	784,000	272,222,222
Plant, *Arabidopsis thaliana* (million reads)	157,000,000	1,256,000	436,111,111
Mammal, *Homo sapiens* (billion reads)	3,200,000,000	25,600,000	8,888,888,889

Research Plan:

Step1: The first step is to sequence the genome of the Identical Twins. We are expecting to have the data available by the end of November. Once we have the sequence data, the next step is to assemble the short read sequences and look into the contig assembly.

Step 2: After the assembly process, the next logical step would be to validate the assembly by aligning the contigs to the reference genome. This can be done by existing methods like GMAP (Genomic Mapping and Alignment Program) or BALST (Basic Local Alignment Search Tool) from NCBI. One way is to make sure all the contigs from the assembly had at least one hit with the reference genome. This is one way to look in the quality of the assembly, but we are also researching different ideas for validation and also possibly some assembly post processing.

Step 3: Identify CNVs, SNPs, and Structural Variants

Expected Results:

The above described steps hopefully would lead to the identification and difference discovery of the genes that are responsible for the cause of MS in one twin and not in the other.

Final Conclusions and Approximate Time Line of Expected Progress:

We have introduced three interesting research problems in computational genomics and proteomics and presented with some background, related work and proposed solutions and improvements to each of the problems.

Problem 1: Significant amount of work has been done and the next logical step would be is to implement the multiple sequence alignment approach and improve the alignment.

Problem 2: We started with doing some initial analysis on the experimental data by calculating ratios of diabetic to control and also computed t-test on all the spots. The next step is in designing and implementing an efficient algorithm for clustering proteins. Once the methods are developed I would like to see it applied to one or more additional data sets.

Problem 3: The Identical Twin genome sequence data should be available by the end of November 2008 and I will be at NCGR working on the computational analysis of the genome sequence. I will be spending the entire spring semester at NCGR working on this research project.

In fall 2008 the goal would be to implement the clustering algorithm for problem 2 to discover interesting classes of proteins and look at their expression patterns. In spring 2009 as mentioned above I will be working on the MS-Identical Twin project. By summer 2009/fall2009 we plan on having all the three problems completed.

References:

1. Brendan Mumey, Brian W. Bailey, Bonnie Kirkpatrick, Algirdas J. Jesaitis, Thomas Angel, Edward A. Dratz: A New Method for Mapping Discontinuous Antibody Epitopes to Reveal Structural Features of Proteins. Journal of Computational Biology 10(3/4): 555-567 (2003)

2. Padlan, E. 1996. X-ray crystallography of antibodies. *Adv. Protein Chem*. 49, 57-133

3. Brendan Mumey, Nathaniel Ohler, Thomas Angel, Algirdas Jesaitis, Edward Dratz. Filtering Epitope Alignments to Improve Protein Surface Prediction. ISPA Workshops 2006: 648-657

4. Baker, D., and Sali, A.2001.Protein structure prediction and structural genomics, Science 294(5540), 93-96

5. A. Fiser, R.K.G. Do, A. Sali Protein Sci. 9, 1753 (2000).

6. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition Nature 1992;358:86–89

7. Salamov, A. A. & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest- neighbor algorithms and multiple sequence alignments. J. Mol. Biol.247, 11-15.

8. Brendan Mumey, Louise Showe, Micheal Showe. Discovering Classes in Microarray Data using Island Counts, Journal of Combinatorial Optimization (2007) 13:207-216

9. Jacques Colinge, Keiryn L. Bennett. Introduction to Computational Proteomics, PLOS Computational Biology, July 2007, Volume 3, Issue 7.

10. A. Malcom Campbell, Laurie J.Heyer (2002) Discovering Genomics, Proteomics and Bioinformatics.

11. Madeira, S.C., Arlindo L. Oliveira. 2004. Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)

12. Ben-Dor A, Friedman N, Yakhini Z (2001). Class Discovery in Gene Expression Data. In RECOMB ’01:Proceedings of the fifth annual international conference on Research in Computational Molecular Biology

13. Bickel D(2003). Robust Cluster Analysis of Microarray Gene Expression Data with The number of clusters determined biologically. Bioinformatics 19(7):818-824

14. Mardis ER, The impact of next-generation sequencing technology on genetics. Trends Genet. 2008 Mar;24(3):133-41.

15. Daniel Zerbino, Ewan Birney, Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs, Genome Research published March 18th 2008

16. Sajjad Hossain, Navid Azimi, and Steven Skiena, Crystallizing short-read assemblies around lone Sanger reads ???

17. Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer, SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo sequencing, Genome Research 2007 17:1697-1706

18. John D Kecceciogh u, Eugene W Meyers, Combinatorial Algorithms for DNA sequence assembly

19. Mihai Pop, Steven L. Salzberg, Martin Shumway, Genome Sequence Assembly: Algorithms and Issues, IEEE July 2002

20. A. Kalyanaraman, S.J Emrich, P.S Schnable, S. Aluru, Assembling genomes on large-scale parallel computers, Journal of Parallel and Distributed Computing 67 (2007) 1240-1255

21. Weichum Huang, and Gabor T Marth, EagleView: a genome assembly viewer for next generation sequencing technologies, Genome Research Published online Jun 11th 2008

22. Pavel A. Pevzner, and Haixu Tang, Fragment Assembly with double-barreled data, Bioinformatics, April 2, 2001

23. Mark J. Chaisson, and Pavel A. Pevzner, Short read fragment assembly of bacterial genomes, Genome Research 2008 18: 324-330

24. Pavel A. Pevzner, Haixu Tang and Maichael S. Waterman, An Wulerian path approach to DNA fragment assembly, 9748-9753, PNAS, August 14th 2001, vol.98 no.17

25. Andreas Sundquist, Mostafa Ronaghi, Haixu Tang, Pavel Pevzner, and Serafim Batzoglou, Whole-Genome Sequencing and Assembly with High Throughput, Short-Read Technologies, PLoSone May 2007, Issue 5, e484

26. Dan Gusfield, Algorithms on Strings and Sequences, Computer Science and Computational Biology ???

27. Basic Local Alignment Search Tool, Altschul, et al 1990.

28. http://www.cbcb.umd.edu/research/assembly_primer.shtml

29. http://www.ncgr.org

30. http://www.wikipedia.org

31. Global variation in copy number in the human genome, Redon, et al, Nature, Vol 444, 23 November 2006