Term Project Ideas
CAP 5937 – ST: Bioinformatics

1. SiRNA(RNAi)

siRNA(small interference RNA)s are short (19-23bp) RNA sequences that that 'interfere' with gene expression. These RNA sequences are known to help in 'gene silencing', preventing a gene from being expressed. These sequences, together with ribonucleases, attach to the mRNA of a specific gene, and 'cut' the mRNA, thereby rendering it useless for translation. Gene silencing through siRNA was first observed in *c. elegans* and *drosophila melanogaster*. Efforts are underway to design synthetic siRNAs that suppress the expression of unwanted genes (eg: in cancer cells) in humans and other mammals. A good introduction to siRNA is available from http://www.invivogen.com/siRNA/siRNA_overview.htm.

A term project on this topic will deal with the siRNA technology in depth. The project should provide a overview of any computational problems in designing siRNA sequences. A survey of current solutions to these computational problems should be provided. Improvements to current solutions/solutions to unsolved problems in this area should be investigated.

This topic will be ideal for a team with at least one member that has done undergraduate/graduate coursework in molecular biology.

2. Web interface for CodonOpt

The genetic code is degenerate. Sixty four possible codons code for 21 possible amino acids(including the stop codon). Therefore, each amino acid can be coded anywhere from 1 to 6 different codons. The preference of which codon to use for an amino acid varies from organism to organism. This preference is called 'codon bias'. These preferences may arise from the availability/non-availability of the respective tRNAs. The codon usage tables for different organisms are available.

Codon usage effects gene expression. In laboratory experiments, it is often infeasible to extract a protein from the source organism. To extract any protein from the source organism in large quantities, large colonies of the source organism have to be cultured. There fore, it is a general practice to introduce a gene from the source organism (eg: *plasmodium falciparum*, the malarial parasite) into the dna of e.coli (the target organism), cultivate large colonies of this modified e.coli, and extract the desired protein from these e.coli colonies.

One problem with this procedure is that codon usages might be different in the target organism from the source organism. The introduced gene might be using a rare codon in the target organism. When this happens, the target organism quickly runs out of tRNA molecules for the rare codon, and the translation process will stop. The end result is that the desired protein will not be expressed in the target organism.

Solution to this problem is to design a synthetic gene that is codon-optimized for the target organism. Therefore, keeping the end-product (the amino acid sequence) the same, rare codons in the gene have to be replaced with codons that are preferred in the target organism.

However, synthetic genes created in this fashion may create new problems, especially when used in dna vaccines or in gene therapy. The target organism identifies certain motifs in the synthetic DNA. These motifs modulate the immune response of the target organism. Some motifs might stimulate immune response, while others might suppress it. Depending on the application, it is desirable either to minimize or maximize the immune response to the introduced gene/vaccine.

Therefore, while designing synthetic DNA sequence, it is desirable to:
(1) codon optimize for the target organism
(2) minimize/maximize the immune response to the DNA sequence in the target organism.

We have developed an algorithm for this problem. The algorithm is explained in detail in the paper available from http://vlsi.cs.ucf.edu/bio_info.htm. The implementation of the algorithm is available as a windows executable developed using MFC and C++.

A term project in this topic will:
(1) Understand the algorithm in depth
(2) Provide a web-based implementation for this algorithm. (ideally running on the client machine, as a java applet).

3. Web server for PRUNER

Identifying transcription factor binding sites is another well-known problem in bioinformatics. Many different approaches exist for this problem. One of the approaches is to look at the 5' UTR(untranslated regions) of a set of co-expressed genes, and identify monad patterns in the these untranslated regions.

Monad patterns are patterns of the form (*l*,*d*)-*k*, where *l* is the length of the pattern, *d* is the maximum number of mismatches allowed, and *k* is the minimum number of sequences out of *t* input sequences that have an occurrence of the pattern.

Some of the best algorithms presented for this problem are SPELLER, MITRA, WINNOWER, etc. We proposed an improved approach called PRUNER. The full paper is available from http://vlsi.cs.ucf.edu/118_VijayaSatya.pdf.

A term project on this topic will focus on:

(1) Understanding the problem thoroughly
(2) Developing a webserver for PRUNER
(3) Implementing some of the other approaches for this problem

(4) Running these algorithms on some biologically relevant data sets to compare the performance of the algorithms

4. Finding Regulatory motifs

Monad patterns are only one type of approximation for regulatory patterns: There are approximations, like profile-based approaches and PWM(Position-Weight Matrix) based approaches. Many motif-detection techniques employ statistical methods, like hidden markov models and bayesian networks.

A term project on this topic will focus on preparing a comprehensive report on all these other approaches/tools/databases available for finding regulatory motifs.

Refer to:

(1) http://www.gene-regulation.com/ TRANSFAC, a database of all known transcription factors and their corresponding binding sites. Links to many other programs

(2) http://labs.systemsbiology.net/bolouri/Mogul/ a webserver with restricted access. Provides links to some motif-detection programs

5. Haplotype Inference

In general, the DNA of all humans is almost similar. However, there are some site sin which a significant percentage of the population(at least 2-5%) have a different base than the rest. These locations are called SNP(Single Nucleotide Polymorphism) sites. It is believed these SNPs are responsible for many of the common diseases. There fore, the first step in understanding the connection between SNPs and diseases is to obtain the SNP information of a large set of individuals. In total. There are bout 10 million SNP locations in the human Genome, roughly one every 300 base pairs.

Humans are diploid - meaning – we have two copies of each chromosome. These two copies are woven together in their natural form. One of these copies is inherited from the mother, and the other is inherited from the father. Doing a complete sequencing of each chromosome is a very costly process. Therefore, only the SNP locations are sequenced. Again, it is a costly procedure to separate both the copies of the chromosome and sequence them separately. Therefore, the two copies are sequenced together. Giving information like this:

SNP	1	2	3	4
Bases	(A,A)	(C,T)	(A,C)	(A,T)

What this is telling us is that both the copies have an ‘A’, in site1, one copy has a ‘C’, and one copy has a ‘T’ in site2, and so on. This does not tell us the exact sequence of bases in each copy. This information is called the ‘genotype’ of the individual. The exact sequence on each copy is called a ‘haplotype’.

Each SNP site is generally bi-allelic. That means that there are only two possibilities at each site. The genotype data is expressed as a vector over the alphabet {0,1,2}. A site is called homozygous if both the copies have the same base at the location. Otherwise, the site is called heterozygous. ‘0’ and ‘1’ indicate that the site is homozygous with the dominant (more common) or the rare base at that location, respectively. A ‘2’ indicates that the site is heterozygous. There fore, the genotype vectors are generally represented as:

SNP	1	2	3	4
Genotype	0	2	1	2

A haplotype is said to *resolve* a genotype if the haplotype and the genotype agree on all non-’2’positions in the genotype. If there are *k* ‘2’s in a genotype, there can be 2^{k-1} distinct pairs of haplotypes that resolve the given genotype.

The following two pairs explain the genotype above:

H	0	0	1	0
H’	0	1	1	1

Or:

H	0	1	1	0
H’	0	0	1	1

Given a set of genotypes, haplotype inference problems deal with deducing a set of haplotypes that resolve the given set of genotypes. There are different versions of the problem, including:

- Finding the minimum number of haplotypes that resolve the given set of genotypes.
- Constructing a ‘phylogenetic tree’, in which edge is labeled by a site, each site appearing at most once in the tree, with the paths to the leaves representing the haplotypes that resolve the given genotypes.

6. Multiple Sequence Alignment (MSA)

Multiple sequence alignment problem is proven to be np-complete. However, multiple sequence alignment is one of the most important problems in computational biology. Many heuristic approaches (for eg: the center star approach) have been proposed for this problem.

Many times biologists face the problem of finding a ‘consensus’ sequence, given a set of sequences that are expected t have the same function. Similarly, given a gene/protein sequence in a set of organisms, it is often necessary to estimate how the common ancestor would have looked like.

A term project on this topic will explain the problem in detail, discuss some heuristics to solve this problem. Additionally, the project deals in depth with any one practical application of this problem in biology. It will also be necessary to provide review of the different tools available to solve the selected biology problem.

7. Constrained Multiples Sequence Alignment (CMSA)

Constrained multiple sequence alignment is a more restricted version of the multiple sequence alignment problem. The problem is to align a set of sequences, given a constraint sequence S_c. For each position i in S_c, there has to a be column j in the alignment such that all the rows(that is, all the strings) have the character $S_c[i]$, in that column. In other words, we are looking for an alignment of all the input sequences, along with the constraint sequence, such that each character of the constraint sequence aligns against the same character in all the input sequences.

Obviously, unlike the MSA problem, all instances of CMSA do not have a solution. For a solution to exist, the constraint should be a a subsequence of all the input sequences.

CMSA is more applicable to protein sequences. Some times, in a protein family, some positions are crucial: certain residues have to be present in these positions for the protein to belong to this family. The other positions are more flexible: they can tolerate changes up to some extent. CMSA can be used to see how well a give set of sequences fit into a given protein family.

A term project on this problem will study the problem in detail, and discuss the different approaches available for solving this problem.

8. Protein Folding problem

The amino acid sequence produced due during translation does not stay in the linear form. The amino acids interact with each other to form two-dimensinal shapes, called α-helices and β-sheets. These α-helices and β-sheets interact further, producing complex three-dimensional shapes. Further, blocks of 3-D shapes come together to form a protein. These different levels in the structure are called secondary, tertiary and quaternary structures of the given protein. Protein folding problem deals with predicting the secondary, tertiary or quaternary structure of a given amino acid sequence.

A term project on this problem will study the problem in detail, and discuss the existing solutions/heuristics for this problem.

9. Protein Docking Problem

An important problem is to understand how proteins interact with each other. Protein-protein interactions helps in understanding the root causes of some diseases, and the effective ness of the drugs developed for the disease.

The protein problem is essentially a computational geometry problem given two or more three dimensional structures what is the most stable arrangement of these structures? Each point on the surface of each structure (each molecule/atom) has a certain affinity/aversion to each point on the surface of the other structure.

A term project on this problem will study the problem in detail, and discuss the existing solutions/heuristics for this problem.

10. Evolution, Phylogenetics, Comparative Genomics

The project will deal with an interesting computational problem in any of the above topics. Some of the actual problems in this area include:

- Calculating the evolutionary distance between two organisms based on variations in some highly conserved genes
- Constructing phylogenetic trees

11. Microarray Data analysis

A term project in this area deal with computational problems in the acquiring, processing, and interpreting microarray data.

For an introduction to microarrays, refer to the tutorial on micro arrays, posted on the class web page.

12. Oligo Design for Microarrays

Designing unique oligonucleotide sequences for a given set of genes. This oligonucleotide sequences should match closely with the target gene, but should not match with any of the other genes in the given set.

13. Microarray Image compression

Micro array images are often very huge. Scientists do not want to lose any information, even in the background. Therefore, using lossy image-compression techniques is not preferred.

A microarray image has a specific structure – it generally consists bright spots on dark background. The spots form a regular grid. A lossless compression technique which takes advantage of the structure of microarray images will be able to achieve good compression ratios. A few techniques have already been presented, but none of them achieve impressive compression ratios.

14. Data mining problems

There are scores of other problems in bioinformatics that can be broadly classified as data mining problems. Some of topics include:

- Automatic keyword extraction for gene clustering
- Automated term disambiguation of biological terms

  - Multiple terms in biology have the same acronyms. When the acronym used, it is some times difficult to figure out what the acronym is actually referring to.
- Extracting SiRNA sequences from published papers