Project guidelines:

1. Literature review
2. Learn and run two existing bioinformatics programs OR
3. Implement one method (can be implemented using a computer language, shell scripting, macros in Excel, Maple etc)
4. Use these methods to analyze some real biological data
5. Write a ~5 page report
6. Prepare a 20 minute presentation.

Project list[§]

Motif finding. Take a group of related proteins and find motifs. Do they match the reported motifs? Why or why not? Might there be other, unreported motifs? What assumptions are made by available motif-finding programs?

Gene finding. Choose a fairly large genome contig from any organism. Predict where the genes are. Take into account codon usage, splice sites, alternative splicing, nested genes, repeats, TATA boxes, Kozak sequences, termination signals, gene products functioning at RNA level. Compare your results and methods with the literature.

Regulatory motifs. Review the literature on algorithms to automatically determine regulatory motifs (short sequence signals) in DNA sequence data. DNA-arrays allow one to find genes that are simultaneously expressed. Those genes are most likely co-regulated, i.e. they share a common sequence signal in their promoter region.

SNP (Single Nucleotide Polymorphism). Review the literature of the methods for detecting SNPs, as well as their application. 'Single nucleotide polymorphisms (SNPs) are common DNA sequence variations among individuals. They promise to significantly advance our ability to understand and treat human disease.' (Excerpt from snp.cshl.org).

Metabolic Pathways. Proteins interact together to perform specific functions. Such networks of interaction are called a molecular pathways. There are two main aspects to this field: how to infer/determine the connections and how to simulate cellular processes. There exist several computational approaches to model molecular pathways.

Genome rearrangements. Genomes are evolving at several scales: from point mutations to large rearrangements. It the late 80s, it became evident that several closely related genomes had genes that were extremely similar (say 99 %id) to one another, but the order of genes along the chromosomes was not preserved. The main algorithms to compare entire genomes include: sorting by reversals (Sankoff), break point graph, Hannenhalli and Pevzner algorithm.

[§] Several ideas from the Turcotte Bioinformatics Course, Ottawa University -- http://www.site.uottawa.ca/~turcotte/teaching/csi-4126/assignments/project/)

Accurate Phylogenetic Reconstruction from Gene-Order Data.

Predicting Gene-Gene (Protein-Protein) interactions. There exist a number of algorithms that allow one to predict whether two genes interact. This includes: text-mining, co-location along the chromosomes, phylogenetic footprinting, etc.

Methods for detecting trans-membrane helices. There is class of transmenbrane proteins whose secondary structure can be reliably predicted. Those proteins are mainly made of helices, such that if the loop connecting the helices i and i + 1 is exposed to the inside of the cell, then the next one will be exposed to the outside of the cell.

Bio-Ethics. Bioinformatics deals with biological and medical data. Accordingly there are numerous related ethical issues: should patenting genes be allowed? How does one handle patient data? How does one deal with genomic data? Imagine that the analysis of a dataset allows one to draw conclusions about a population, a religious group, people who live in a specific region, etc. The consequences can be severe: it could be that this group will be more likely to suffer from certain diseases, such information could be used by insurance companies, employers, etc. to screen candidates. (Project guidelines are flexible for this project).

Reference implementation of affine-gap penalty (NW or SW) alignments. Include visualization of the matrix and traceback steps.

From work in your own lab, if you have any other potential projects which could fit into the project guidelines, you are encouraged to develop a project proposal based on that.