

# Computer Science 549 – Computational Biology

Prof. Steven Skiena

Fall 2011

Semester Project

Proposal due: Thursday, October 27, 2011

Progress report due: Thursday, November 17, 2011

Project due: Thursday, December 8, 2011 (in class)

September 22, 2011

The project will involve concentrated work in one research project related to computational biology. Below I list several possible projects. You may also choose your own topic if you can convince me that what you want to do is interesting.

My hope is that projects will involve interaction between life science and computational students. Thus there is a rule that every group *must* have at least one Computer Science student to prevent the non-CS students from clumping up together.

While I anticipate that much of the work will be done as the deadline approaches, it is important to get started early enough to discover insurmountable roadblocks in data acquisition or problem definition before it is too late. The project proposal and progress reports have been instituted to ensure people get serious well in advance of the final deadline.

Each group is responsible to turn in 3-5 page project proposals/literature search and progress reports as of the dates above. I will award roughly 50% of the grade for each project on the strength of the preliminary reports. This is to encourage starting early, and to make sure that you and I both know what you are to do before it is too late to avoid trouble. Each group will have to turn in a final written report/WWW site and have a quick meeting with during finals week – during which I will ask if all members participated equally.

My hope is that one or more of these projects will lead to published work. This has been the case most times I have taught such a course. The projects I believe are best have been starred (\*). Those marked PhD are reserved for groups of PhD students *only*. I particularly recommend them to students thinking about doing research under me.

1. *PacBio Assembly* (\* PhD only) - new sequencing instruments from Pacific Biosciences can sequence much longer fragments of DNA than any other sequencing technology over 2000bp compared to 100-500bp), but at a much higher error rate (typically 15% error). The long read length makes the instruments very attractive for de novo assembly of complex genomes, but the high error rate prevents traditional approaches from being used. Design and implement an algorithm for assembling these reads de novo, such as using spaced seeds for finding overlaps or an error correction pipeline that can be applied directly to the reads. Requires strong algorithm and programming skills.

There is 30x to 100x coverage of long reads of a few strains of *E. coli* publicly available from PacBio: <http://www.pacbiodevnet.com/Share/Datasets/E-coli-Outbreak> Assembling *E.coli* from just the long reads should be challenging but not impossible.

2. Reconstructing Books from NGrams (\*)- Google has made available collections of all runs of five consecutive words in scanned books. The question is to what extent large hunks of the book text itself can be reconstructed from this. It is a sequence assembly problem, but on a big alphabet (words) with

all ngrams (ie. perfect coverage). Start by trying to assemble one book (from Project Guttenberg) before thinking about the much larger problem of a mix of all books from a given year.

3. Designing Autocorrelated/uncorrelated Genes (\* PhD only) – One recent observation is that codon usage within a gene seems to be autocorrelated: codons are reused locally in the hopes of reusing a tRNA already in the neighborhood. Can we design genes which use codons to maximize and minimize autocorrelation for a given set of codons? See Cannarozzi cited in Plotkin and Kudla’s 2011 nature review: “Synonymous but not the Same”.
4. Genome Compression Against Reference (\*) – How much data is necessary to encode one’s genome? Can you compress it, using a database genome as a reference?
5. *Fear of Diseases and Faith in Drugs* (\*) – Use TextMap Access data to chart changes in frequency and sentiment across archival news, patent, and Pubmed depositories for the most common diseases and drugs, to answer questions like:
  - Do research and public (newspaper) sources exhibit similar trends in volume and sentiment on various diseases?
  - Which diseases do people fear more (less) than before? Does this coincide with actual frequencies of death?
  - What trends are evident in the frequency and sentiment of drugs?
  - How does the preferred drug change its association with a given disease per year?

*Need:* Groups which are a mix of CS and non CS-types

6. *Patent Analysis* (\*) – Patents are very important in the biotechnology and pharmaceutical industries. Can we identify which patents are most valuable by doing NLP and citation network analysis on the complete set of patents?

This project will involve using the existing TextMap Access patent depository and improving text-mining tools to retrieve and analyze patents:

- Spider and parse patent documents.
- Analyze/cluster them by similarity, ownership, citations, etc. to determine the relative significance of them.
- Estimate the value of the patent portfolio of each company.

There is a literature on such analysis to start from – find it!

7. *Making DNA patterns* (\*) – DNA origami is a fascinating technology. Read the literature (e.g. Rothmund PWK (2006) “Folding DNA to Create Nanoscale Shapes and Patterns”, Nature 440: 297-302), understand the design problems and see what you can make/simulate with existing tools. Then see if you can go farther...
8. *Inverse Protein Folding* – For a given shape, find an amino acid string with a fold approximating the shape. Seek algorithms that improve over current approximation methods on 2D/3D lattices, at least heuristically?

References: Kleinberg’s RECOMB 99 paper. The inverse protein folding problem on 2D and 3D lattices. Piotr Bermana, Bhaskar DasGupta, Dhruv Mubayic, Robert Sloan, Gyrgy Turn, Yi Zhang Inverse protein folding in 3D hexagonal prism lattice under HP model. Alireza Hadj Khodabakhshi, Jan Manuch, Arash Rafiey and Arvind Gupta *Need:* two to three CS/AMS types

9. *Designing Secondary Structures in Coding Regions (\*)* – It is remarkable that sophisticated regulatory structures can be built within highly-constrained coding regions of genes. We propose a study of the extent to which such structures can be constructed. Can such stem loops be built essentially anywhere within coding regions, or is the potential restricted to rare amino acid sequences? Reimplement the Cohen-Skienna RNA design algorithm and provide the answer. *Need:* two or three people with algorithmic chops.
10. *Centrifugal Separation Strategies* – Centrifuges separate materials in solution based on the difference in density between them. Develop an abstract model of centrifugal separation of materials, and devise algorithms which provably give you the largest amount of pure material possible. (from Dmitri Gnatenko) *Need:* One algorithmically inclined CS-type.
11. *Exploiting DNA Ancestry Data* – Personal DNA ancestry test kits are now commercially available (e.g. <http://www.dnaancestryproject.com/>) enabling the tracking of family histories from such sequence data/databases. Get such data and something interesting with it.
12. *Data Mining/Analysis Projects (\*)* – Experiment with a data mining/analysis technology on an interesting problem / data set. This project is good **IFF** you have a good application and idea of where you will get data from.
  - (a) *Building HMMs* – Build a general-purpose Hidden-Markov Model implementation, supporting one or more learning algorithms, general topologies, probability computations, etc. Report the results of experiments with your program on some classification problem, perhaps one we discussed in class or in natural language processing (actor type recognition, sentence boundary identification).
  - (b) *Using SVMs* – Use support vector machines (SVMs) to solve some classification problem, perhaps one we discussed in class or in natural language processing and report the results.
  - (c) *Using Bayesian Networks* – Use Bayesian networks to model data from an entity, citation, or other relevant network.
  - (d) *Phylogenies of Texts/Disciplines* – Use our text analysis system to generate distances between entities or document sets, and then phylogenetic tree algorithms to reconstruct history of (say) geopolitical splits, language/cultural groups, or partitioning of academic fields..
13. *Racehorses as a Model Organism?* – Extensive genealogical and performance data is available for thoroughbred race horses. Further, these pedigrees are tremendously inbred, which makes them a very interesting model organism for genetic studies. To what extent can you get and parse the genealogical and performance data, and make predictions of which breeding pairs will be most successful? *Need:* one or more CS-types who can write scripts to download and parse web data. – MS
14. *Theoretical Algorithm Problems (PhD only)* – Each of these requires at least one algorithmically strong CS-type, and a keen sense of when to give up and move to something else.
  - (a) *Shortest Subset Subsequence* – As discussed in algorithm reading group, what is the string on an alphabet of size  $n$  such that every one of the  $2^n$  subsets is represented by a substring of minimum size, ie. equal to the cardinality.
  - (b) *Path Consistent Walk in a String* – Can you find a low-period string defining a path between two vertices in a character-labeled path, i.e. a string? The problem is hard for trees and dags.  
A related problem: Given two strings  $S_1$  and  $S_2$ , can you find a string  $S$  such that both  $S_1$  and  $S_2$  can be realized by a back and forth walk on  $S$ ?
  - (c) *Finding the Most Frequent Subsequence* – Given a string  $S$ , which string  $S'$  occurs most often as a scattered subsequence of  $S$ ? Note that the number of candidates and possible occurrences are exponential. Can you prove this is NP-complete, and maybe give an approximation algorithm? Can you use dynamic programming to count the number of occurrences of a candidate  $S'$ ?

15. *Software Survey Papers* – Write a 10-15 page in-depth survey paper on the state of software available for one of the following topics of interest, or pick your own topic:

- Haplotyping
- Synthetic biology
- RNA interference.
- Genetic linkage analysis.
- Computational challenges in Epigenomics
- Small RNAs, microRNAs and other noncoding RNAs

In all cases, discuss the different programs available for the problem, and how they compare by performance, speed, input requirements, and algorithms. Your paper should be based on some actual experience with the programs.

All survey papers must appropriately cite all sources, and be written in *your own words*. Any student caught plagiarizing from Web or published sources will receive a Q grade and be subject to academic dishonesty proceedings. Don't be stupid – I know how to use Google, too. *Need:* each survey paper should be done individually by either a CS-type or a biologist

## Life Science Projects

Several projects have been contributed by faculty from life science departments affiliated with Stony Brook. Many are looking for student to do larger projects with, so this is an opportunity to make a good impression. Be polite and respectful of their time. I encourage you to discuss the project with them – however, contact me first by email with your intention so I can make sure that they are not flooded with responses.

1. High-Performance Assembly (\*) – (from Michael Schatz mschatz@cshl.edu)

- Genome Indexing on the GPU - A critical step in genomics is indexing the genome for rapid searches. Today the most popular indexes are suffix arrays and the closely related Burrows Wheeler Transform used in the leading algorithms Bowtie, BWA, Vmatch, BLASR, and many others. Computing these indexes on large genomes requires many hours of computation. The goal of the project will be to design and implement a parallel indexing algorithm that can exploit the parallelism of a GPU. Requires strong algorithm and programming skills.
- Assembly Forensics - In every large genome assembly, the assembler will make mistakes called mis-assemblies such as leaving out or reordering different segments of the genome. These commonly occur because of repeats in the genome introduce false overlaps between distant segments of the genome. The goal of this project will be to enhance and extend the assembly forensics pipeline which can identify potential mis-assemblies using various statistical tests of the data (<http://genomebiology.com/content/9/3/R55>). This includes updating the pipeline for current sequencing methods and more significantly enhancing the forensic markers into a probabilistic framework. Students with GUI development experience may also wish to enhance the Assembly Visualization program Hawkeye: <http://genomebiology.com/2007/8/3/R34>. Requires strong statistical and/or programming skills.
- Parallel Assembly and Analysis - A major challenge in genomics is executing sequence analysis in the presence of huge volumes of data. A promising solution uses the Hadoop open-source version of MapReduce. <http://www.nature.com/nbt/journal/v28/n7/full/nbt0710-691.html>. The goal of this project is to enhance the parallel assembler Contrail (<http://contrail-bio.sf.net>) with new features, or to implement a sequence analysis algorithm in Hadoop such as for identifying differentially expressed genes or single nucleotide polymorphisms. Requires strong parallel graph algorithms design and/or programming skills.

2. Population Dynamics and Ecology – A new faculty member in Ecology and Evolution has a research focus on Antarctic penguins and their population dynamics. Two projects are:

- The Euler-Lotka equation is a transcendental equation for population growth rate as a function of life history traits such as mortality, litter size, and lifespan. I need someone to figure out a way to efficiently map solutions to the Euler - Lotka equation in the multidimensional space of the input parameters. A general solution would be ideal but I'm specifically interested in doing this when mortality over lifespan is taken as a beta distribution. This would involve both writing efficient code for finding a solution to the Euler equation, and then finding an efficient way of searching parameter space (e.g., adaptively using smaller step sizes where the solution is changing more rapidly).
- The population dynamics of penguin populations are complicated by skipped breeding, whereby breeding-age adults skip breeding due to poor pre-breeding season environmental conditions. I'd like someone to help extend multivariate state space models for studying penguin population dynamics to incorporate skipped breeding, and to use these extended models to study how sensitive parameter estimates are to missing data or short time series.

(from Heather Lynch, hlynch@life.bio.sunysb.edu, <http://lynchlab.wordpress.com/>)

3. Bioinformatics and Viruses (\*) – (from Laurie Krug, krug@notes.cc.sunysb.edu)

- Codon deoptimization of essential viral genes (PhD only) – To define the role for viral genes in replication and pathogenesis in the host we need to examine the defects upon infection with a knock-out virus, a recombinant virus that has a stop mutation in the open-reading frame. However, these viruses are often replication defective and a virus stock for infecting cells and mice is very difficult. Complementation strategies that supply the wild-type sequence in trans lead to recombination and repair of the mutant virus, re-generating a wild-type virus. We seek to design a codon-swapped expression construct that will generate the protein yet lack DNA sequence homology that would enable recombination. Our first codon-altered ORF was not expressed well in culture. Therefore, we need to re-design this construct and others that balance codon-swapping with codon and codon-pair optimization. Predictions of recombination and translation efficiencies might also be considered.
- Explore transcription factor elements in the promoters of different herpesvirus gene classes – Herpesviruses infect for life using a strategy of latency. The role of stimuli that break this latent stage and lead to reactivation and spread is an intense area of investigation. We are interested in how these stimuli impact cellular signaling that ultimately drives lytic viral gene expression. We have identified a set of genes that are expressed early during a reactivation event prior to peak expression of the viral transactivator RTA. We would like to apply bioinformatics analysis to identify patterns of transcription factor recognition elements in these murine gammaherpesvirus promoters and their human gammaherpesvirus counterparts.
- Determining the impact of herpesvirus infection on cellular gene expression – We are interested in identifying groups of genes that are upregulated, downregulated, or not changed during murine gammaherpesvirus infection. First, we would like to define the stable, house-keeping genes to use as a group of genes for normalization, as a point of comparison across multiple arrays. We have performed seven microarray experiments so far, with 4-8 samples per experiment, and 3,000-30,000 cellular genes per sample. There is a tremendous amount of data to mine. We expect that different methods of normalization and cluster analysis will be explored to generate a robust data set.

4. Polymorphism Analysis on HIV– My laboratory studies HIV replication and one project relates to identifying determinants of disease progression in the viral genome. We recently found sequence polymorphisms in a region of the viral genome that bears such determinants in virus isolated from some

HIV-infected individuals. These individuals were tested and found to possess genes (HLA alleles) that normally would induce their immune system to make antibodies against this region. However, the patients do not so we are wondering whether the polymorphisms permit escape from immune targeting. There is software that can be used to predict the probability that this is what is occurring but understanding it and employing it is too challenging for us life science types. Several computational systems have been used to predict peptide immunogenicity: Some are based on physicochemical properties of the peptide (POPI); some are motif-based (RANKPEP); some involve machine learning methods (e.g., artificial neural network, hidden Markov model). These several approaches would enable the student to do some evaluation of the different strategies. (from Dr. Susan Wantanabe mitsue@optonline.net)

5. FlexE Server (\*) – Proteins are one of the most important biological units in our living beings. We have implemented a method in python that allows us to distinguish good protein models from bad ones based on the proteins intrinsic flexibility nature. We would be interested in porting this scripts into a server that the scientific community could use. Once setup we would send a paper of medium impact for publication. We would like the server to be done using the Django web framework. It would be great if a student would like to take the time to program a nice robust server, the project has a clear line of work so it would be perfect for a short term involvement in a lab. (from Alberto Perez alberto.perez17@gmail.com)
6. Long Non-Coding RNAs – Long non-coding (lnc) RNAs play pivotal roles in regulating developmental chromatin states and nuclear function, including the direct modulation of pathways important for cellular differentiation. We have performed next-generation RNA sequencing (Illumina, paired-end) of total and nuclear poly(A)+ RNA fractions obtained from mouse embryonic stem cells (ESC) and neural progenitor cells (NPC). The aim of this project is to identify and characterize novel lncRNAs that are enriched in the nucleus and differentially expressed between these two cell states.  
The project involves: (1) computation of expression values (FPKM) for annotated genes and de novo assembled transcripts across the four samples; (2) classification of putative lncRNAs based on their position relative to protein-coding genes (antisense, intronic, intergenic, ) (3) correlation of lncRNA loci with published datasets of lncRNAs, transcription factor binding, histone modifications, DNA methylation and lamina association. (from David Spector spector@cshl.edu and Jan Bergmann jbergmann@cshl.edu)
7. Computational analysis of genome-wide DNA methylation alterations in a large cancer genome dataset (\*) – The cancer genome is my field of study and there is a new datatype that has emerged over the last year as one of the most interesting and largely unexplored areas of cancer genome alterations. We have our own dataset of 100 tumor and normal samples for one type of cancer; there are lots of other datasets publically available now, over 1000 in total, but no meta-analysis has been made. (DNA methylation alterations as detected by Illumina infinium platform) The project would be to write programs to merge the datasets and perform a meta-analysis, which I would help with, and create a publicly available database (useful to the cancer research community) for viewing results. I have people in my group familiar with database and web page design, and familiarity with Perl, R, Matlab, C++, Python . (from Scott Powers powers@cshl.edu)
8. Bacterial Sequence Analysis (\*) – The aim is to implement, develop or improve software tools for analyzing and comparing whole-genome sequences of relatively closely related bacteria, to facilitate identification of characteristic features, conserved and variable genes or regions, and sites and types of mobile elements, repeated sequences, inversions, duplications, insertions, deletions and replacements. These tools will be applied initially to help generate a reference basic genome of commensal E. coli, based on currently available sequences, and to analyze evolutionary mechanisms. The reference genome and software tools are meant to facilitate analysis of newly determined genome sequences of both commensal and pathogenic E. coli strains, and should also be applicable for similar analyses of any other groupings of relatively closely related bacteria. (from Sergi Maslov maslov@bnl.gov and Bill Studier studier@bnl.gov)