1 Projects

1.1 Population history

National Geographic and IBM are teaming up to do the human genography project, which traces the history of a human individual from the first (African) ancestor.

(see https://www5.nationalgeographic.com/genographic/).

To do this, people usually study mtDNA (or the Y chromosomes). The reason is simple: unlike in the autosomal chromosomes, there is no recombination, implying that each DNA sequence has a unique parent. If it were also true that there were no recurrent mutations, known efficient algorithms for character based phylogenies could reconstruct the tree uniquely. Recurrent mutations could make the problem harder, but not insurmountable. However, as more individuals are genotyped, the complexity of these algorithms will grow. Complete mtDNA of over 1500 humans is available. (EX:http://www.bch.umontreal.ca/ogmp/projects/other/mt_list.html).

Problem: Given a collection of mtDNA, use the variations to build a character based phylogeny efficiently. The algorithm should scale to 100,000 individuals, and should allow for the placement of specific mutations on individual lineages.

Skills: Algorithm development skills, understanding of Phylogeny.

Reading:
Gusfield’s book has a good introduction to perfect phylogeny and other tree reconstruction methods. The following paper discusses the use of heuristics to address recurrent mutations, and incomplete data.

Contact: Fjola Bjornsdottir (fbjornsd@ucsd.edu)

1.2 Identifying ncRNA in genomic regions

Large scale efforts are underway to sequence large parts of the genomes of many species (the ENCODE project). The goal of this project is to identify functions of regions that might be unusually conserved across mammalian species. (See http://www.genome.gov/10005107). The goal of this project is to identify regions that might encode non-coding RNA.
A good place to start is to look for conserved secondary structure in untranslated part of non-coding genes. The tools RNAz and QRNA look for conserved secondary structure in aligned regions. However, it is possible that the multiple alignments for ncRNA are not correct (as primary sequence conservation is not as high, and there are more indels). Other tools like RNAscf look for conserved secondary structure in unaligned sequences.

In this project, you can start by developing a tool to identify the conserved consensus structure in multiple alignments of orthologus (sequences with a common ancestor, and having similar functions) sequences. We will provide you with ∼200 alignments from known non-coding RNA, and some controls.

**PROBLEM:** Identify the conserved consensus RNA structure in an alignment of genomic sequences. Modify your algorithm to include shifted stacks. Use your tools to differentiate true RNA from false.

**REQUIREMENT:** Understanding of RNA structure and existing algorithms (covariance models).

**ASSIGNED PAPERS:**


### 1.3 Querying biological networks

Protein protein interactions are key to many biological processes. To take an example, the extracellular signal might be in the form of a small peptide, or an organic compound. The signal could bind to a receptor protein on the surface of the cell. To transmit the signal to the nucleus, so as to switch other genes on or off, a chain of events take place. Each protein in the chain interacts with the next protein on the chain, and activates/de-activates it. This chain of proteins is often called a *pathway*. Proteins may also interact to form complexes which function together.

A protein interaction graph is a natural representation of these interactions. Each node \( u \) corresponds to a distinct protein, and we define an edge \((u, v) \in E\) if and only if proteins \( u \) and \( v \) interact. Clearly, a graph of all interactions gives us great insight about the cellular processes. Traditionally, each interaction was mapped using careful experimentation, but high throughput techniques are being developed that allow us to measure all these interactions in a few experiments. Protein interaction networks are now available for many species, including yeast, drosophila, human, and so on.

An interesting question that arises is the following: how similar are these networks across different species? The question is relevant because we often study specific processes, or drug-responses in other *model* organisms, such as yeast, rat, dog and so on. Imagine that you have a
pathway (a small subgraph of \(k\) nodes) in a species, and you want to identify the corresponding sub-graph in a large protein interaction network.

**Abstract problem:** Given a small query subgraph \(G_Q = (V_Q, E_Q)\) of \(k\) nodes, a larger graph \(G = (V, E)\) of \(n\) nodes, and a similarity relation \(R \subseteq V_Q \times V\) between nodes in \(Q\) and \(G\), map each \(q \in V_Q\) to a unique \(h(q) \in V\), such that \((q, h(q)) \in R\), and for all \((q, q') \in E_Q\), \((h(q), h(q')) \in E\).

**Skills:** Algorithmic, programming skills.

**Reading:** The following describe straight algorithmic approaches that is suitable for restricted sub-graphs, paths, and trees. The question is to extend these to more general sub-graphs.


### 1.4 Primer design for cancer genomics

Many cancers are marked by massive structural changes to the genome. Identification of these events is important for diagnostic purposes. Here, we assume that the change is in the form of a genomic deletion. The approximate location of the deleted region is known, but the boundaries of the deletion vary from patient to patient.

Polymerase chain reaction (PCR) is an established technique for amplifying a small (1-3Kb) region of DNA. It needs a pair of unique primers (20 bp long sequence) on either end to start the reaction. The amplified region can be investigated to see which primer pair was amplified. In the specific protocol we are considering, PAMP, a collection of primers on either end of the deletion are selected. Denote these left and right primers by their positions \(l_1, l_2, \ldots, (r_1, r_2, \ldots)\), where the positions are all increasing

\[ l_1 < l_2 < \ldots < r_1 < r_2 < r_3 \ldots \]

If the primer pair \(l_i, r_j\) was amplified, we know that the left end-point of the deletion is between \(l_i\) and \(l_{i+1}\), and the right end-point is between \(r_{j-1}\) and \(r_j\).

Unfortunately, this protocol is complicated by the fact that some of these primers interact with each other, and cannot be used together. Also, the primers should be as closely spaced together as possible but do not need to be less than 1kb apart. Densely packed primers make it more expensive. Loosely packed primers make detection of breakpoints harder.

**Problem:** the input to the problem is a set of increasing positions corresponding to uniquely occuring 20-mers, on the left and right of the deletion. Denote these as

\[ l_1 < l_2 < \ldots < r_1 < r_2 < r_3 \ldots \]

In addition, we have a number of forbidden pairs \((l_i, l_j)\), or \((l_i, r_j)\) of positions that cannot occur together. The solution is a subset

\[ l_{i_1} < l_{i_2} < \ldots r_{j_1} < r_{j_2} \ldots \]
such that no forbidden pair is present. The coverage error between $l_{ik}, l_{ik+1}$ is given by

$$e_k = \max\{0, l_{ik+1} - l_{ik} - 1000\}$$

The goal is to find a subset with a fixed number of primers such that $\sum_k e_k$ is minimized.

**skills:** You need some knowledge of combinatorial optimization. Try and solve this using branch and bound, simulated annealing, or ILP.

**reading:**

### 1.5 Efficient construction and search of a splice exon graph

Define a *splice exon graph* as follows: each node is associated with a protein sequence (corresponding to an exon). Two nodes are connected by an edge if the corresponding exons can splice together.

A *mass spectrum* is a collection of spectral *peaks*. Each peak is defined by a mass, and an intensity. Ignoring intensity, the spectrum of a peptide (a small protein sequence) is simply a collection of masses of fragments of that peptide which act as a fingerprint for the peptide. In interpreting the spectrum, we associate many short peptide tags with it.

**Problem:** The goal of this project is to build a tool to search the splice exon-graph with a collection of short peptides. Each peptide might either lie entirely within an exon, or might be a substring of the concatenation of spliced exon pairs, thereby corresponding to a path in the splice-exon graph. Describe the complexity of your algorithm in terms of the total length of all the peptide queries, $k$, the sum of the lengths of all the exon sequences, $n$, and the number of edges, $m$.

**Skills:** Algorithmic, programming skills.


Tanner S, Shen Z, Ng J, Florea L, Guigo R, Briggs SP, Bafna V. Improving gene annotation using peptide mass spectrometry. Genome Res. 2007 Jan 2;

### 1.6 de novo Repeat analysis

### 1.7 Analysis of paired peptides

### 1.8