Projects

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.

Contact: vbafna@cs.ucsd.edu

Project goals

1. Describe the perfect phylogeny algorithm, and summarize its complexity in the presence of recurring mutations. Come up with a heuristic for handling recurrent mutations. Additionally, summarize and critique the approach in [12].

2. Download the data referenced at the site above, and produce a SNP matrix. For aligning the DNA, feel free to contact Fjola Bjornsdottir (Fjola Bjornsdottir fbjornsd@cs.ucsd.edu).

3. Implement the heuristic on the SNP matrix, and compare your phylogenies with existing phylogenies.
2 Fast test for multiple locus mapping

Recall that the goal of genetic mapping is to identify genetic markers that are correlated with incidence of a disease. The reasoning is that under a common evolutionary history, different alleles are correlated. However, the correlation is broken over long distances due to recombination. Therefore a test for correlated loci helps isolate the genetic cause of a phenotype.

For many complex diseases, multiple loci (genes) participate to confer susceptibility to the disease. A recent study of Evans et al.[5] points to the importance of measuring gene-gene interactions. However, this is a significant computational challenge To identify even \( d \approx 5 \) genes that interactively associate with the disease, the naive approach would require a test of all subsets of 5 SNPs. On a genome wide scan with \( 1M \) SNPs over a \( 10^4 \) individuals, this would require an astounding \( 10^4 \cdot 10^{6.5} = 10^{34} \) computations, as well as multiple-testing problems. Your goal is the following:

**Problem 1:** Given a case-control sample of genotypes, devise an efficient algorithm to filter all pairs of SNPs that might interactively associate with the phenotype.

For ease of exposition, assume that we have a binary \((n \times m)\) matrix \( M \) of haplotypes with \( n \) haplotypes (rows), and \( m \) columns (SNPs). Also assume that \( m \simeq 10^6 \) is large. Consider two SNPs \( s_1, s_2 \) with frequencies of allele ‘1’ being \( p_1 \) and \( p_2 \), respectively. Let \( q_1 = 1 - p_1 \), and \( q_2 = 1 - p_2 \). Define the labeled-hamming-distance as

\[
\mathcal{H}[s_1, s_2] = \min\{p_1p_2 + q_1q_2, p_1q_2 + p_2q_1\}
\]

Note that \( \mathcal{H} \) is closely related to the standard hamming distance which is given by \((p_1q_2 + q_1p_2)n\). \( \mathcal{H} \) is the minimum hamming distance (averaged over \( n \)) when the ones and zeroes of the SNPs can be relabeled. The critical observation is that if \( s_1, s_2 \) are correlated, then \( \mathcal{H} \) is ‘low’, otherwise ‘high’. Our first step will be to formalize and quantify the \( \mathcal{H} \) gap for pairs of correlated, and uncorrelated SNPs.

Next, we will consider the combinatorial problem of identifying pairs of SNPs with low labeled-hamming distance. The goal here is not to detect the interacting (correlated) pairs, but to filter out a small subset of pairs that include the truly interacting pairs, and to do this efficiently \( o(nm^2) \) time. If the subset of pairs filtered is small enough, we can exhaustively test all pairs for epistatic interactions. Formally, identify all pairs of SNPs whose hamming distance is less than \( d_1 \). This is possible if a typical pair of SNPs has expected hamming distance \( d_2n \) where \( d_2 >> d_1 \). We propose the following randomized protocol:

**PairedSNPs(\( \delta, k \))**

Repeat \( l \) times
1. Select \( k \) rows (haplotypes) at random as \( i_1, \ldots, i_k \).
2. Hash each SNP \( j \) in the location \( h_j = ⟨M[i_1,j], M[i_2,j], \ldots, M[i_k,j]⟩ \).
3. Also, hash SNP \( j \) to bitwise complement \( h_j^\prime \).
4. For each hash bin, filter all pairs of SNPs that fall in the bin at least \((1 - \delta)le^{-kd_1}\) times.

Consider a pair of SNPs \( s_1, s_2 \). Let \( C \) be the count of the number of times the pair appears together. If \( \mathcal{H}[s_1, s_2] = d_1 \), the probability that the pair passes the filter in a single iteration is \( p = e^{-kd_1} \). In \( l \) iterations the expected number of times a pair appears together is \( \mu_1 = le^{-kd_1} \). On the other hand, a pair of SNPs with high hamming distance \( d_2 >> d_1 \) has a lower expected count of \( \mu_2 = le^{-kd_2} \). Using Chernoff’s inequality, the two counts can be separated with high probability.

\[
Pr(C > (1 + \delta)\mu_2) \leq e^{-\frac{3}{2}\mu_2}
\]

\[
Pr(C < (1 - \delta)\mu_1) \leq e^{-\frac{3}{2}\mu_1}
\]
To take an example, let $d_1 = 0.3$, and $d_2 = 0.5$. Then, choosing $l = 1000$, and $k = 5$, $\mu_1 = 116$, and $\mu_2 = 31$. Choosing the cut-off as 62, two SNPs with low hamming distance fail to make the cut-off with probability $\leq 5 \cdot 10^{-7}$, whereas two SNPs with hamming distance 0.5 make the cut-off only with probability $3 \cdot 10^{-5}$. Only the filtered pairs ($10^{-5}$ fraction of all pairs) need to tested, and any association test can applied on the filtered pairs. This significantly reduces the computational burden without disrupting established testing methodology.

**Problem:** Given a case-control sample of genotypes, devise an efficient algorithm to filter all pairs of SNPs that might interactively associate with the phenotype.

**Skills:** Algorithm development and coding skills, understanding of LD.

**Reading:**
Read and summarize Evans paper[5]. Note that a different approach is presented in Brinza and Zelikovsky[4], and might also serve as a starting point for a possible project. Please contact the instructor for the dataset in Rana et al.[11].

**Contact:** vbafna@cs.ucsd.edu

**Project goals**
1. Read and summarize the papers referenced[5, 11].
2. Formalize the ideas presented above.
3. Design and implement an algorithm which takes a case-control data-set as input, and 'quickly' identifies SNP pairs that interactively associate.
3 Detecting balancing selection

Allelic variants that are causal for a disease, or linked to causal genes are expected to be under negative selection. However, in isolated cases, balancing (particularly co-dominant) selection might play a role. A balancing selection is useful in a number of contexts: for genes involved in host response to pathogens, heterozygosity helps maintain diversity. Second, differing alleles in regulatory regions influence downstream gene expressions. A case for balancing selection is also made for psychiatric disorders like bipolar disorder (BD), where “the grand creativity, sexual attractiveness, and disinhibition in the throes of mania give a clear route to short-term reproductive success...”[16].

As balancing selection is likely to impact tests of association, it is important to determine the regions under balancing selection prior to association testing.

**Problem 2:** Devise an algorithm to detect regions under balancing selection using genotype data.

The problem has been studied before, but mostly in a more general context, identifying regions that depart from neutral selection. Some of the classical work here are statistics based on the work of Tajima[17], Fu and Li[6, 8], and Hudson et al.[7].

\[
E(S) = \sum_{i=1}^{n} \frac{1}{\theta} = (\gamma + \log n)\theta
\]

Thus, \(\theta\) can be estimated either by computed the average number of pairwise differences \(k\), or independently from the number of segregating sites \(S\) as \(\theta_{S} = \frac{S}{\gamma + \log n}\). Tajima’s D-statistic is described as \(D = k - \theta_{S}\). Under neutral evolution, \(D \approx 0\). Under positive selection, the heterozygosity \(k\) decreases (Fig. 1(b)), implying \(D < 0\). Under balancing selection (Fig. 1(c)), \(k\) increases, implying \(D > 0\). While powerful and widely used, Tajima’s D, and other statistics often fail to detect departures from neutrality. As Hudson points out, a sudden increase in frequency of a specific haplotype is a signature of selection[7]. However, for small \(n\), such an increase would decrease \(k\) and \(\theta_{S}\) by the same amount. Therefore, other tests need to be devised to counter this.

Sabeti et al. propose a long range haplotype (LRH) test for locating patterns of higher than expected homozygosity. The idea is as follows: Recent positive selection for a mutation would cause a core haplotype carrying the proposed protective mutation to rapidly rise in frequency. Define \(EHH\) as the probability that two randomly chosen chromosomes carrying the core haplotype are identical. Clearly, at the mutation of interest, the \(EHH\) value is 1. Under neutral

![Diagram](image-url)
models, the EHH value decays quickly, disappearing around (25kb)(~ 0.02cM) from the starting locus. Sabeti et al. find candidate genes in which the decay is much slower (250kb), giving a strong signal for recent positive selection, that is missed by using Tajima’s D-statistic. The method has been extended slightly to use comparisons across populations\cite{19}, and for genome-wide tests\cite{15}. The test works for recent positive selection where recombination does not destroy the homozygosity. Its effectiveness in testing for older selective events, and for balancing selection is unknown, and false-negatives are still very likely. Indeed, LRH can also be reduced by recent mutational events even in the absence of recombination. This effect is observed in Fig. 1(d), where the additional mutations on lineages 6,7,8 can lower the EHH value.

In this respect, the EHH statistic can be considered as a restricted version of a test proposed by Hudson. For the Drosophila Sod locus, Hudson observed that 19 of 63 locations were identical in 31 of 41 individuals. Using simulations, Hudson showed that this phenomenon represented a departure from neutrality, and was indicative of selective forces. The approach did not ask for the identical locations to be consecutive, and is therefore difficult to test directly. Here, we will use this idea to test for selection using combinatorial optimization.

Problem 3: Given a collection of $n$ individuals with $m$ segregating sites, determine if there is a subset of $n_1 \leq n$ individuals and $m_1 \leq m$ sites, such that the $n_1$ individuals are identical at the $m_1$ sites?

Problem:

Skills: Algorithm development and coding skills, understanding of LD, Integer linear programming, combinatorial optimization.

Contact: vbafna@cs.ucsd.edu

Project goals

1. Read and summarize Sabeti and Hudson’s papers\cite{13, 14, 7}

2. Formalize the ideas presented above, or describe your own algorithm.

3. Design and implement an algorithm for detecting adaptive selection, as well as balancing selection. Test your algorithm on the data-sets provided.

Figure 2: An Integer Linear Program for detecting common segregation sites, as a signal against neutral selection. A large value of $\sum_k C_k$ indicates selection pressure. Further, $\sum_i P_i \approx \sum_i N_i$ indicates a balancing selection.

$$\max \sum_k C_k$$

s.t.

For all individuals $i, j$, position $k$

$P_i + P_j + C_k \leq 2 + q_{ijk}$

$N_i + N_j + C_k \leq 2 + q_{ijk}$

$P_i + N_j + C_k \leq 3 - q_{ijk}$

For all individuals $i$,

$N_i + P_i \leq 1$

$\sum_i N_i + \sum_j P_j \geq n_1$

$N_i, P_i, C_k \in \{0, 1\}$
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 (Primer Approximation Multiplex PCR), 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 the size of the resulting PCR product need not be less than 2-3kb. 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 occurring 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. A breakpoint coordinate, $(x, y)$ flanked on the left by primer $l_x$ and on the right by primer $r_y$, is considered to uncovered if $(x - l_x) + (r_y - y) > d$. Where $d$ is the maximum possible amplification distance (2-3kb).

The goal is to find a subset with a fixed number of primers such that the number of uncovered breakpoints is minimized. It is important that solutions be able to handle primer input sets of typical sizes (at least thousands of primers).

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

**reading:** The original method is presented by Liu and Carson \cite{9} and simulated annealing and ILP solutions are presented in Bashir et al. 2007 \cite{3}. ILP extensions, on small data set sizes have been proposed by Dasgupta et al. \cite{266}.
5 Identification of MS/MS spectra from mixtures of peptides

In the last 40 years, tandem mass spectrometry (MS/MS) has gone from an exotic experimental technique for measuring the mass of small biomolecules to become the major approach to protein identification. Indeed, the widespread availability of these instruments has motivated much more ambitious goals such as identifying post-translationally modified (PTM) proteins\[18, 2\], quantifying differential protein and PTM expression levels across different samples (e.g. healthy vs. diseased states) and even to determine full protein sequences completely from scratch (de-novo assembly\[1\]).

However, the mainstream computational techniques used to analyze mass spectrometry data are still very similar to the initial database search approach over 10 years ago - first digest the proteins into peptides (smaller protein subsequences) and then compare the fragment masses in each peptide’s MS/MS spectrum to the theoretical masses generated from a database of known protein sequences. In essence, a mass spectrum is simply an n-dimensional vector with a non-zero value in the m\textsuperscript{th} dimension if mass m is detected by the mass spectrometer. Since most amino acids have different masses, any peptide can easily be converted to a set of theoretical masses for all possible prefix or suffix fragments. A more extensive introduction to the data will be given in the first project meeting (details are also available in \[10\]).

Although current algorithms are generally successful in identifying spectra generated from a single peptide, none is currently able to identify mixture spectra - spectra generated from two or more peptides. Mixture spectra occur in a number of different situations and the ability to identify these could lead to significant improvements in the number of identified proteins and, in certain experimental setups, even allow for more accurate peptide quantification.

Contact: Nuno Bandeira (bandeira@cs.ucsd.edu)

Project goals:

1. Implement an algorithm to identify mixture spectra from pairs of peptides by combining previously identified spectra from isolated peptides.

2. Test the above implementation by simulating mixture spectra using an existing database of spectra from isolated peptides.

3. Propose a scoring procedure to separate correct from false identifications.

Required skills: proficiency in C++ and algorithms
Desirable skills: familiarity with tandem mass spectrometry

References


