CSE280 Class Projects (Suggested)

Vineet Bafna

January 15, 2009

Projects

1 Haplotype assembly from single molecule amplification

Dr. Kun Zhang (bioengineering) has devised a novel technique for amplifying large genomic regions, sampled randomly from a genome, and the goal is to use the technology for haplotype phasing (separating maternal and paternal alleles).

The protocol

The protocol consists of the following steps:

a) Shearing. The DNA is sheared/size selected so that each piece is around 200Kb.

b) Dilution. The DNA is diluted so that each genomic locus is sampled once. This critical step ensures that we have the maternal or paternal, but not both, chromosomal fragments, at each locus. In reality, you can assume a poisson distribution on the count with $\lambda = 1$.

c) Amplification: The WGA technique is used to amplify the genomic region. While most of the DNA is sampled, the distribution is random, and not uniform.

d) Genotyping: The sample is arrayed, or sequenced to reveal genotypes. Locations that are sampled once will show up as homozygous alleles.

1.1 Computational Project A:

Problem

Given a set of genotypes from a few experiments as above, reconstruct the complete haplotype sequences. We can formulate the problem as follows: The (hidden) haplotype data for an individual is a binary string $s$, where each bit represents an allele. The second haplotype is simply the bitwise complementary string $\bar{s}$. Each experiment gives us a string with 4 characters 0, 1, 2, ‘-’, with the following interpretation. 0, 1 are simply the allelic forms, implying that we have sampled from one of the two haplotypes. 2 implies that we have sampled both haplotypes $0/1$ at that location, and ‘-’ implies that we have not sampled from that location. The goal is to reconstruct $s$ given a collection of experiments.

Skills: Prior knowledge of algorithms, programming.

Reading: Read Zhang’s paper on WGA for the protocol. A possible approach to haplotype reconstruction is in Bansal et al. Genome Research 2008, but this problem is easier. Discuss with the instructor.

Data-sets should be forthcoming in a month, but use simulated data in the interim. This project is of research interest to me, and to Prof. Zhang. For extra work, also do project B.

1.2 Computational Project B:

Give a design for this experiment. By design, you must provide the following: Dilution: The amount the original sample must be diluted. b) The number of experiments, so that with high probability, each genomic locus is sampled exactly once in some experiment. c) A trade-off between the amount of sequencing, and genotype success, given a distribution of WGA distribution. d) The length of the haplotypes obtained. Your design should be based on extensive simulations, and/or theoretical analysis.

Skills: Mild Programming/Scripting, Basic statistics.
2 Structural variation analysis using transcripts

Paired-end reads can be used to detect structural variations in the genome. A genomic clone of fixed length \( L \) (Ex: \( L = 2000 \)bp) is chosen from a sample, and the sequence of its ends is determined. Normally, one would expect that the two end sequences map to the reference genomic sequence at locations that are \( L \) apart. If however, they map aberrantly, that is indicative of a structural variation in the sampled genome. We have devised techniques to resolve the location of a genomic breakpoint using a collection of overlapping reads.

Additionally, sequencing ends of transcripts can also detect aberrations in genes due to structural variation.

The goal of this project is to identify the structural variations that disrupt human genes based on simultaneous analysis of paired-end sequencing data from genomes and transcripts. The input is a collection of mapped locations of paired-end reads (genomic and transcripts). We will start with transcripts first.

**Problem:** Given a collection of aberrantly mapping ends

1. Identify all fused transcripts, or transcripts that combine the prefix and suffix of two transcripts. Specifically, your algorithm should output a small number of fused transcripts that explain the aberrant mappings.
2. Classify the mapping into one of 2 cases: a) the putative protein product is a fusion of the two proteins, or b): Only the regulatory part of one gene is used. Give confidence values for your assessments.
3. Revise your estimates by incorporating genomic data.
4. **Difficult:** Estimate the amount of sequencing needed to detect such fusion events.

**Skills:** Algorithms, Statistics, Genomics (eukaryotic gene structure)

**Research:** Bashir et al., 2008, Eichler’s papers.
3 Eukaryotic proteogenomics pipeline

The goal of proteogenomic project is to use data from the proteome (in the form of MS2 spectra) to gain knowledge about the genome. For example, identified peptides can be used to reveal or refine gene annotations [7]. As mass spectrometry becomes more widespread, peptide identifications become more reliable, and compute power increases, proteogenomics may soon become an integral part of gene annotation efforts.

In this project, you will develop a general proteogenomic pipeline that can be used for any eukaryotic organism. The pipeline should consist of several steps to accept unidentified spectra and the genome of the organism of interest and produce coordinates of likely coding sequences on the genome similar to Castellana et al. Arabidopsis thaliana is a good test organism, and chromosome sequences can be downloaded from the Arabidopsis Information Resource (TAIR) at https://www.arabidopsis.org.

The overview of the pipeline should look roughly similar to Figure 1.

Problem: Design a proteogenomic pipeline so that users can easily solve the following problem: Given a collection of mass spectra and a genome, identify likely coding regions.

Skills: Coding skills, understanding of tandem mass spectrometry and peptide identification.

Reading:
Read and summarize Castellana paper [7]. The specifics of the splice graph database are presented in [7]. Information on AUGUSTUS can be found in [7].

Contact: Natalie Castellana (ncastell@cs.ucsd.edu)

Project goals

1. Read and summarize the papers referenced [7, 7, 7].

2. Design the flow of the pipeline, indicating clearly the input and output of each component. Implement or build on current scripts to execute the components. Keep in mind that the number of chromosomes may vary and the size of the genome may grow to several orders of magnitude larger than Arabidopsis.

Figure 1: An overview of the proteogenomic pipeline similar to the process used in Castellana et al. Dark rounded rectangles represent pipeline inputs and outputs. Light rectangles represent subroutines in the pipeline. There are two different flows represented: if the organism has a known proteome and if the organism does not. The dotted lines indicate extra subroutines and data that must be collected and reported in the case of a known proteome. The subroutine 'Determine Genes' is highlighted and explained further in Figure 2.
Figure 2: A zoom of the DetermineGenes component of the pipeline. In some flows, the genes will be determined by the gene prediction tool AUGUSTUS\(^2\). When AUGUSTUS is used, an optional input in addition to the peptides is sequence homology and Expressed Sequence Tag (EST) alignments. However, when AUGUSTUS is not available, or not suitable, simply clustering peptides based on genomic location is sufficient.

4 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 \(^9\).
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.
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…”[13]. 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 1: 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[14], Fu and Li[4, 6], and Hudson et al.[5].

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 $\theta_S = S \frac{\gamma + \log n}{\gamma + \log n}$. Tajima’s D-statistic is described as $D = k - \theta_S$.

Under neutral evolution, $D \simeq 0$. Under positive selection, the heterozygosity $k$ decreases (Fig. 3(b)), implying $D < 0$. Under balancing selection (Fig. 3(c)), $k$ increases, implying $D > 0$. While powerful and widely used,
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 2:** 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{10, 11, 5}

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.

\[
\begin{align*}
\text{max} \sum_k C_k \\
\text{s.t.} \\
\text{for all individuals } i, j, \text{ position } k \\
P_i + P_j + C_k &\leq 2 + q_{ijk} \\
N_i + N_j + C_k &\leq 2 + q_{ijk} \\
P_i + N_j + C_k &\leq 3 - q_{ijk} \\
\text{for all individuals } i, \\
N_i + P_i &\leq 1 \\
\sum_i N_i + \sum_j P_j &\geq n_1 \\
N_i, P_i, C_k &\in \{0, 1\}
\end{align*}
\]

Figure 4: 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, $\sum_i P_i \simeq \sum_i N_i$ indicates a balancing selection.
6 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 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. 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 \[^7\] and simulated annealing and ILP solutions are presented in Bashir et al. 2007 \[^3\]. ILP extensions, on small data set sizes have been proposed by Dasgupta et al. \[^7\].
7 Mapping short reads

Goal: Build a tool for matching (aligning) a collection of methylated DNA short sequences to a reference DNA sequence.

Skills: Good programming skills (Java/C). Scripting language might slow you down here.

Motivation: di-sulfite sequencing is a technique for identifying methylated regions of the genome. The reaction transforms all ‘C’s to a ‘T’, except for ‘C’s that are methylated. Many local researchers are sequencing di-sulfite treated samples. The goal is to map these sequences fast to the reference genome to identify regions of the genome that are methylated.

Work plan:

1. Take the input data file of sequences, and map them efficiently to the genome. You can use the publicly available program, Maq (http://maq.sourceforge.net), but later you want to create your own. The output of the program should provide, for each input sequence, its genomic coordinates.
2. Write a script to randomly change C to a T, each with a fixed input probability $p$ (C2).
3. Write a script to introduce random sequencing errors with some probability $p_1$. The sequencing errors include indels and small substitutions (C2).
4. Devise a strategy to map di-sulfite reacted sequences back to the genome. Test the speed and accuracy of your mapping (C2).
5. Devise and implement an algorithm to map the sequences in the presence of sequencing errors. Test the speed and accuracy of your error-tolerant mapping against Maq (C3).

8 MS2 with charged losses

Talk to instructor

9 MS2 identification of higher charge peptides

Talk to instructor

References


