# CSCI2951-N: Advanced Algorithms in Computational Biology Homework 3

Due: 11:59PM Wednesday November 12, 2014

**Please handin your submission by emailing it to sorin@cs.brown.edu with subject "csci2951-N HW3 handin"**

The Homework contains problems worth 120 points; i.e. there are 20 extra points.

## Problem 1. The Search for the mysterious "1 Fisher" unit for measuring LD (50 points)

This is a very difficult open problem in genetics. We discussed in class some of the methodological difficulties with the measures of LD used in the literature. The two desiderata, or axioms, formulated in the literature as good properties LD measures to have are: (1) Independence of Allele Frequencies, and (2) Overcoming the "curse of the pairwise" we need multimarkers measure. Namely all LD measures in use are pairwise, and essentially the extension of any of them to 3 or more SNPs or markers makes them behave difficult analytically; when pairwise the measure has magically good properties; once one try to generalize it to more than two sites, the measures gets complex formulas that are not easy to use or interpret.

In the class chapter on Algorithms for Tagging SNPs selection we discussed two algorithms, the *LD-Select algorithm* included in the Tagger software package, and the *Informativeness algorithm*. Both algorithms take as input a haplotype matrix with $n$ rows, haplotypes corresponding to $n$ individuals, and $m$ columns, namely SNPs. The algorithms select respectively a set of SNPs called "tagging SNPs" that "capture" the power of of the entire set of $m$ SNPs. The mathematical formal objective function of LD-select is the "Dominating Set problem"and the formal objective function of Informativeness is the "Set Cover problem." These are both famous NP-comlete problems when about on general graphs. Moreover, I presented in class a reduction from one problem to the other. The reduction shows that if one can find an exact polynomial time algorithm for one of the problems, then one can get an exact polynomial time algorithm for the other. The "Dominating Set" article in wikipedia contains good references and the standard proofs of these mathematical reductions results. Informativeness is an "information theoretic" measure: the one bit of information is "distinguishing one person from another when they have different alleles at a SNP". Informativeness also satisfies axiom (2) above, as it can be extended easily and uniquely (in mathematical terms, a "conservative extension") to multi SNPs or multi markers.

Design and implement an algorithm for Dominating Set and one for Set Cover, so you can have your own programs for LD-select based on a general threshold $r^2 = \theta$ and and your program for

Informativeness. Try to convert LD measured, with $r^2$, as in the graph constructed for LD-select SNPs selection, to Informativeness. Study the set of tagging SNPs picked by the LD-select, namely how "informative" they are in the sense of Informativeness vis a vis the information in the entire set of SNPs. A set of SNPs of the $n$ x $m$ input matrix contains information about the entire set of SNPs (all the D-edges). If a set of SNPs contains only 30% of the total number of D-edges, then this is the *information threshold* $\alpha$ for that set of SNPs.

Is there a correlation between $r^2$ in LD-select and the information threshold $\alpha$ of Informativeness? Use your algorithms for Dominating Set and Set Cover and use also, if you can, the reduction between the two algorithms, maybe as a hybrid algorithm, or iterative algorithm, to attempt to calibrate the conversion of LD measured by $r^2$ into information threshold $\alpha$. It is hard to compare the LD in two regions of same genome or two different genomes. Think about ways to calibrate biases due to allele frequencies specificificity to those genomic regions. Can you come up with your definition of "1 Fisher" unit for LD that is more robust? Use HapMap data to obtain LD information from different regions of the same genome as benchmark, and try to compare them to each other via your convertion to Informativeness's information threshold $\alpha$..

## Problem 2. Adopt a Disease/GWAS and present it to our class (50 points)

There are materials posted for the classes devoted to Computational Workflows of GWAS. The article "Genetic Mapping of Human Disease" presents an overview and it appendix has the list of GWASes cited and overviewed and it is a list for you to consider. Pick one that may be of interest to you. Use the Tutorial on Statistical Methods for GWAS to follow up and present the deails of their computational workflow. There are two articles giving advise on how to analyse a GWAS presenting also caveats and critical analysis. Finally, the article "Gilt by association" presents the findings of a GWAS using police/detective terminology to present the weight of the evidence in the case. Use a similar presentation motif for your presentation. Be creative in finding the corresponding metaphores that well describe the genetic and genomic evidence of association. The big mistery is what exactly the leading SNP (or few leading SNPs) tells us about the disease: such a SNP or few such leading SNPs are found at the end of a very large amount of work by thousands of people, tens of hospitals, operation spending huge amounts of money to get there.

## Problem 3. Herbert Simon's Hora and Tempus Puzzle: A parable on Evolution (20 points)

Herbert Simon (1962): "Let me introduce the topic of evolution with a parable. There once were two watchmakers, named Hora and Tempus, who manufactured very fine watches. Both of them were highly regarded, and the phones in their workshops rang frequently  new customers were constantly calling them. However, Hora prospered, while Tempus became poorer and poorer and finally lost his shop. What was the reason?

"The watches the men made consisted of about 1.000 parts each. Tempus had so constructed his that if he had one partly assembled and had to put it down to answer the phone say it immediately fell to pieces and had to be reassembled from the elements. The better the customers liked his watches, the more they phoned him, the more difficult it became for him to find enough uninterrupted time to finish a watch.

"The watches that Hora made were no less complex than those of Tempus. But he had designed them so that he could put together subassemblies of about ten elements each. Ten of these subassemblies, again, could be put together into a larger subassembly; and a system of ten of the latter subassemblies constituted the whole watch. Hence, when Hora had to put down a partly assembled watch in order to answer the phone, he lost only a small part of his work, and he assembled his watches in only a fraction of the manhours it took Tempus."

Can you find a quantitative argument to witness the difference in speed between the two watchmakers?