Optimal Tag SNP Selection for Haplotype Reconstruction

Jin Jun and Ion Măndoiu*

Abstract. In this poster, we propose optimum tag single nucleotide polymorphism (SNP) selection methods based on integer linear programming. Experimental results on simulated data show that haplotype reconstruction based on tag SNPs is nearly as accurate as reconstruction based on all SNPs.

Introduction

The completion of the Human Genome Project and the identification of millions of single nucleotide polymorphisms (SNPs) in the human population has opened the way for large-scale association studies between genetic variation and susceptibility to common diseases. At present, it is prohibitively expensive to directly determine the haplotypes in diploid organisms such as humans. However, there are several robust methods for determining the conflated SNP information in the so called *genotype*. Haplotypes are currently inferred from the genotypes of a large population of individuals via statistical approaches such as the well-known PHASE algorithm [6]. In order to reduce haplotyping costs, a two stage methodology has been recently proposed (see, e.g., [2]). In the first stage, all SNPs of interest are genotyped in a small sample of the population. Common haplotypes are then inferred using statistical methods, and a minimum set of *tag SNPs* is selected. In the second stage, tag SNPs are genotyped in the remaining population. Statistical methods are then used to infer haplotypes over the tag SNPs, and finally the latter are extrapolated to full haplotypes.

Tag SNP Selection

Tag SNP selection has received much attention recently. Most of the proposed methods start by decomposing haplotypes into blocks with reduced haplotype diversity, then select tag SNPs within each block independently. Sebastiani et al. [5] proposed a *Best Enumeration SNP Tags (BEST)* algorithm for finding the minimum set of tag SNPs that uniquely identify a given set of haplotypes. Unfortunately, the worst-case runtime of the BEST algorithm grows exponentially in the number of SNPs and haplotypes as shown in the table below. These runtimes were obtained on a 3.0GHz Pentium4 Dell Optiplex by running BEST on n haplotypes corresponding to the rows of the $n \times n$ identity matrix.

n	10	12	14	16	18	20
BEST time	$<$.01s	2s	29s	14m8s	6h4m	4d18h

Although BEST is typically faster on real haplotype data, its runtime is still not practical for tag SNP selection of large unstructured datasets such as that of Daly et al. [1], which consists of 360 distinct haplotypes with 103 SNPs each. Practical runtime for minimum tag SNP selection can be achieved by using integer programming methods. Let x_i be an integer variable that is set to 1 when SNP i is selected as a tag SNP, and to 0 otherwise. As observed in [4], the problem of selecting the minimum number of SNPs tagging a set $\mathcal{H}$ of haplotypes with n SNPs each can be formulated as the following integer linear program (ILP): minimize $\sum_{i=1}^{n} x_i$ subject to the constraint that $\sum_{i:h_i \neq h'_i} x_i \geq 1$ for every pair of distinct haplotypes $h, h' \in \mathcal{H}$. This ILP can be solved using commercial solvers such as CPLEX (http://www.ilog.com/products/cplex/) or open source optimization packages such as GLPK (http://www.gnu.org/software/glpk/glpk.html). Using CPLEX, it takes less than 7 seconds to select tag SNPs for the Daly et al. [1] dataset that could not be handled by BEST.

Accuracy of Haplotype Reconstruction

To evaluate the accuracy of the haplotypes reconstructed by the two stage methodology described in introduction we used a program developed by R.Hudson [3] to generate populations of 200-400 genotypes over 10-30 SNP sites. For each population, we simulated the first phase of a genotyping study with sample size 50. Then, we inferred haplotypes on the whole population using three methods: (1) by running PHASE on full genotypes, (2) by running PHASE on the tag SNPs selected by the ILP, then extrapolating the inferred haplotypes to full haplotypes, and (3) same as (2) but this time using a random set of tag SNPs of the same size as the set selected by the ILP. The table below gives the number of tag SNPs and the haplotype phasing accuracy of the three methods (averaged over 10 instances of each size). The results show that haplotype reconstruction based on tag SNPs is nearly as accurate as reconstruction based on all SNPs, while inference based on the same number of random SNPs is significantly less accurate.

Pop. Size	#SNPs	#Tag	PHASE/Full	PHASE/ILP	PHASE/Random
200	10	7.20	93.65	94.15	89.00
200	20	11.70	90.50	87.05	70.30
200	30	16.10	93.10	90.30	72.25
400	10	7.00	96.30	96.45	87.67
400	20	13.00	94.03	91.88	77.00
400	30	15.60	94.15	89.92	62.50

References

[1] M.J. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson and E. S. Lander. High resolution haplotype structure in the human genome. *Nature Genetics* 29, pp. 229-232, 2001.

[2] J. Forton, D. Kwiatkowski, K. Rockett, G. Luoni, M. Kimber, and J. Hull, Accuracy of haplotype reconstruction from haplotype-tagging single-nucleotide polymorphisms, *Am. J. Hum. Genet.* 76(3), pp. 438-48, 2005.

[3] R. Hudson. Generating samples under the Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2), pp. 337-338, 2002.

[4] X. Ke and L.R. Cardon, "Effi cient selective screening of haplotype tag SNPs", *Bioinformatics*, 19, pp. 287-288, 2003.

[5] P. Sebastiani, R. Lazarus, S.T. Weiss, L.M. Kunkel, I.S. Kohane, and M.F. Ramoni, Minimal haplotype tagging, *PNAS*, 100(17), pp. 9900–9905, 2003.

[6] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, pp. 978-989, 2001

*Computer Science and Engineering Department, University of Connecticut, 371 Fairfi eld Rd., Unit 2155, Storrs, CT 06269-2155. E-mail: {jinjun,ion}@engr.uconn.edu. Work supported in part by a "Large Grant" from the University of Connecticut's Research Foundation.