A Dense Graph Model for Haplotype Inference

Sharlee Climer[1], Alan R. Templeton[2], and Weixiong Zhang[3]

Phasing genotype data to identify the composite haplotype pairs is a widely-studied problem due to its value for genome-wide association studies, population genetics research, and other significant endeavors. The accuracy of the phasing is crucial as identification of haplotypes is frequently the first step of expensive and vitally important studies. We present a combinatorial approach to this problem, which we call Splitting Heirs, that is based on a dense graph model. We have tested Splitting Heirs with several popular existing phasing methods including PHASE, HAP, and Pure Parsimony, on seven sets of real biological haplotype data. Our method yields the highest accuracy obtainable by these methods in all cases. Furthermore, Splitting Heirs is robust and had higher accuracy than any of the other approaches for the two data sets with high recombination rates. The success of Splitting Heirs validates the assumptions made by the dense graph model and highlights the benefits of finding globally optimal solutions.

Many of the methods previously used for haplotype inference have favored reduction of the cardinality of unique haplotypes. Pure Parsimony [2] is an extreme case in which a set of haplotypes are found such that the number of unique haplotypes is the least possible. Splitting Heirs favors reduced cardinality, but simultaneously considers other favorable properties and does not always yield a strictly parsimonious solution.

Some of the previous algorithms (e.g. PHASE [6]) have favored additional properties, such as pair-wise similarities between haplotypes. That is, if a potential haplotype is similar to one in the current solution, it is favored. In contrast, Splitting Heirs favors *cluster-wide* similarities by favoring solutions in which many haplotypes are similar to a number of other haplotypes. The dense graph model can be used to quantify the quality of a solution with regard to reduced cardinality and cluster-wide similarities.

Let h equal the number of unique haplotypes in a solution. Consider a graph with h nodes, in which each node represents a haplotype in the solution. The weight on an edge in the graph is set equal to the distance between the two haplotypes that are endpoints of the edge. Distances between haplotypes can be defined in various ways. A simple distance measure is just the number of sites in which they differ. If pair-wise similarities were the only concern, a graph to consider would contain only edges that connect each haplotype with its nearest neighbor. When relying completely on simple pair-wise distances, it is possible to have $h/2$ disjoint subgraphs with arbitrarily large distances between them. In real populations, we would expect to find *clusters* of haplotypes that are similar to each other, so it is desirable to enforce similarities beyond a single nearest neighbor.

In a *dense* graph model, the density of the graph is required to be greater than or equal to a given value, α. The density of a graph can be defined as e/h, where e is the number of edges in the graph. By considering these additional edges, similarities beyond single nearest neighbors are taken into consideration. We evaluate the quality of the dense graph solution using:

$$C_D = \sum_{i=1}^{e} w_i d_i + \sum_{i=1}^{h} u_i \tag{1}$$

where d_i is the distance of edge i and w_i and u_i are weights. In our experiments, we used a constant u_i value and $w_i = 1$ for all i.

The dense graph with the minimum cost C_D is considered optimal. We have cast this model as an Integer Linear Program (IP). The constraints of our IP require that the selected haplotypes resolve all of the genotypes. These constraints are similar to the constraints for the Pure Parsimony IP formulation. The key differences between our IP and the Pure Parsimony IP is that our objective function is Equation (1), and we add the following constraint to ensure the density of the graph: $e/h \geq \alpha$. Like Pure Parsimony, this problem may require exponential time to compute in the worst case. However, we were able to obtain globally optimal solutions using ILOG's Cplex 8.11, which is a generic IP solver.

On some occasions, the optimal dense graph may have more than one pair of haplotypes that can resolve a given genotype. When this is the case, Splitting Heirs assumes that common haplotypes are very common, and assigns the pair that contains the haplotype with the highest frequency in the set. Alternate pairs, along with their frequencies, are also provided for the user.

The dense graph model is biologically intuitive as it utilizes three widely accepted principles: the number of unique haplotypes within a given population is relatively small, many haplotypes are similar to others, and common haplotypes are very common. PHASE incorporates the first two of these principles in its priors. However, PHASE favors pairs of

[1]Department of Computer Science and Engineering, Washington University in St. Louis, MO, USA. E-mail: `sharlee@climer.us`
[2]Department of Biology, Washington University in St. Louis, MO, USA. E-mail: `temple_a@biology.wustl.edu`
[3]Department of Computer Science and Engineering, Washington University in St. Louis, MO, USA. E-mail: `zhang@cse.wustl.edu`

	Data	n	Pars	# of HAP	Gen. EM	Wrong PHASE	Split	# Het. Sites	Pars	# of HAP	Sites EM	Wrong PHASE	Split
(a)	A	80	5	11	10	4	4	154	5	13	13	4	4
	B	39	4	4	4	4	4	44	4	4	4	4	4
	C	39	1	0	0	2	0	136	3	0	0	3	0
	D	39	1	1	2	0	0	52	1	2	3	0	0
	E	39	3	3	-	4	1	193	7	3	-	6	1
(b)	F	39	4	11	-	6	4	186	5	12	-	8	4
	G	39	14	20	-	7	5	257	27	44	-	8	6

Table 1: Results for genotype sets with (a) little or no recombination and (b) high recombination rates. Number of genotypes in the set (n); the number of genotypes incorrectly phased by Pure Parsimony (Pars), HAP, EM-DeCODER (EM), PHASE, and Splitting Heirs (Split); total number of heterozygous sites in set of genotypes; and number of sites incorrectly phased by each method are tabulated.

haplotypes that are similar. It is biologically intuitive that clusters of haplotypes are similar, not just pairs. Splitting Heirs effectively incorporates this intuition.

We have tested the biological accuracy of various haplotype inference methods using seven sets of true haplotype data derived experimentally (i.e. the individual haplotypes were identified, not the melded pairs). The data came from two sources. The first source of data used for comparisons is a set of 80 human *ApoE* haplotype pairs, each with nine SNPs, that was experimentally found by Orzack et al. [5]. These SNPs are drawn from the *apolipoprotein E* locus. Data set A in Table 1 is composed of these 80 pairs of haplotypes.

The second source of data was experimentally collected by Andrés et al. [1]. It contains 39 pairs of human haplotypes, each with 411 sites, in a 48 kb region containing the *KLK13* and *KLK14* genes. There is a substantial amount of missing data in this set. Pure Parsimony, EM-DeCODER, and the current implementation of Splitting Heirs all require complete data. Six regions of complete data from this set are used for this study and correspond to data sets B through G in Table 1. They range from 5 sites to 47 sites in length. The 17 sites of set D have no recombination and are combined with 9 additional sites, which have a low recombination rate, to make set F.

We compare Splitting Heirs with several popular haplotype inference methods: Pure Parsimony [2], HAP [3], EM-DeCODER [4], and PHASE [6]. Two of these implementations use combinatorial methods and the other two use statistical approaches. Table 1(a) shows the results for the data sets with little or no recombination. These results list the number of genotypes incorrectly phased as well as the total number of sites that were incorrectly phased by each method. As shown in the table, Splitting Heirs did better than, or as well as, all of the other solvers in every case. Table 1(b) contains the results for data with high recombination rates. Splitting Heirs outperformed the other algorithms on both data sets.

Due to consequences for vitally important genome-wide association studies and population genetics studies, the benefits of accuracy for the haplotype inference problem cannot be measured by mere financial gains. Splitting Heirs finds globally optimal solutions for this problem that favor low cardinality of unique haplotypes as well as similarities across clusters of haplotypes. Favoring cluster-wide similarities is biologically intuitive and this assumption is experimentally validated using true haplotype data.

References

[1] A. M. Andrés, A. G. Clark, E. Boerwinkle, C. F. Sing, and J. E. Hixson. Assessing the accuracy of statistical haplotype inference with sequence data of known phase. *Genet. Epi.*, 31:659–671, 2007.

[2] D. Gusfield. Haplotype inference by pure parsimony. In *14th Annual Symposium on Combinatorial Pattern Matching (CPM'03)*, pages 144–155, 2003.

[3] E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20:1842–1849, 2004.

[4] T. Niu, Z. Qin, X. Xu, and J. Liu. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *The American Journal of Human Genetics*, 70:157–169, 2002.

[5] S. H. Orzack, D. Gusfield, J. Olson, S. Nesbitt, L. Subrahmanyan, and V. P. Stanton Jr. Analysis and exploration of the use of rule-based algorithms and consensus methods for the inferral of haplotypes. *Genetics*, 165:915–928, 2003.

[6] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68:978–989, 2001.