

ILP Methods for Family Trio Phasing

D. Brinza¹ J. He¹ W. Mao¹ K. Westbrook¹ M. Fraser¹ R. Harrison^{1,2} A. Zelikovsky^{1*}

¹Department of Computer Science, Georgia State University, Atlanta, GA 30303

²Department of Biology, Georgia State University, Atlanta, GA 30303

1 Phasing Family Trios of Genotypes

In population genotyping, it is common to genotype family trios consisting of the two parents and their child since that allows to recover haplotypes with higher confidence. Interestingly, the available software tools are primarily intended to phase only unrelated genotypes. In this section we first formulate the problem and describe specificity of family trio phasing and then analyze existing computational tools and discuss the pure parsimony objective. In the following section we give three integer linear program formulations and compare their runtime for the Daly et al [18] data.

The haplotypes of children is much harder to recover than haplotypes of parents since we are not aware of recombinations which may happened when parents haplotypes are inherited by a child. Therefore, for simplicity, we assume no recombinations in child chromosomes and that exactly one child chromosome is inherited from one parent and another from the other parent. Formally, given a set of genotypes partitioned into family trios, the Trio Phasing Problem (TPP) requires to find for each trio a quartet of parent haplotypes which agree with all three genotypes.

A simple logical analysis allows to substantially decrease uncertainty of phasing. For example, for two SNP's in a trio with parent genotypes $f = 22$ and $m = 02$, and the child genotype $k = 01$, there is a unique feasible phasing of the parents: $f_1 = 10$, $f_2 = 01$, $m_1 = 01$, $m_2 = 00$ such that the haplotypes f_2 and m_1 are inherited by the child. In fact, it is not difficult to check that logical ambiguity exists only if all three genotypes have 2's in the same SNP site.

Another source of ambiguity is in missing data – certain SNP's for certain individuals may be not available due to failures during genotyping. Although in the most recent data the missing data rate decreases, still they constitute a substantial part of the entire data (as large as 16% of the genotype data in Daly et al [5] data and 10% in Gabriel et al [19]).

As mentioned above, one of the goals of these study has been to design and verify discrimination algorithms for the data [5] which is one of infrequent publicly available large-scale case/control genotype data. We have tried several well-known computational methods for phasing this data trying to find feasible solution for the TPP since this data are given in family trios. Surprisingly, all the methods which we have tried give infeasible solutions with high inconsistency rate. The error rate has been measured as the ratio of the number of inconsistently phased SNP's over the total number of ambiguous SNP which are either missed or cannot be logically inferred. Note that the error rate does not rely on the assumption that no recombinations happen in the children.

The Phamily tool based on well-known phasing tool PHASE is intended to phase the trio families [2]. It first uses the logical method described above to infer the SNP's in the parental haplotypes. Then children genotypes are discarded while the parental genotypes and known haplotypes are passed

*The corresponding author, e-mail: alexz@cs.gsu.edu, ph.(404) 651-0676, FAX (815) 642-0052.

to PHASE. Because the children genotypes are discarded, PHASE no longer can maintain parent-child trio constraints resulting in 8.02% error rate for phasing Daly et al [5] data. A recent tool HAP [10] also does not deliver a feasible solution having 9.8% error rate.

An interesting phenomenon have been discovered for the greedy method for missing data recovery [11]. The authors replace each genotype in Daly et al [5] data with a pair of logically resolved haplotypes referring to each ambiguous SNP value as a ?. The ?'s constitute 15% of all data. Then extra 10% of data are erased (i.e., replaced with ?'s) and the resulted 25% of ambiguous SNP values are inferred by the greedy algorithm minimizing haplotype variability within blocks. When measured on the additionally erased 10% of data, the error rate of the greedy algorithm is 2.8% [11] which has been independently confirmed in our computational experiments. Unfortunately, the error rate of the original 15% of ?'s is at least 25% which has been measured by the number of inconsistently phased SNP's. This may lead to a conclusion that the complexity of missing genotype data is considerably higher than the complexity of the successfully genotyped data.

Note that it is easy to find a feasible solution to TPP but the number of feasible solutions is exponential and it is necessary to choose a criteria for comparing such solutions. In [?] for haplotyping pedigree data, the objective is to minimize recombinations. That objective is not suitable for TPP since the trios are not full-fledged pedigree data and contain no clues to evidence recombination reconstruction. Thus, following [?, 8], we have decided to pursue parsimonious objective, i.e., minimization of the total number of haplotypes.

The drawback of pure parsimony is that when the number of SNP's becomes large (as well as the number of recombinations), then the quality of pure parsimony phasing is diminishing [8]. Another important drawback of large number of SNP's is huge runtime of finding the smallest number of haplotypes caused by inherent complexity of the problem [8]. Therefore, following the approach in [10], we suggest to partition the genotypes into blocks, i.e., substrings of bounded length, and find solution for the pure parsimony problem for each block separately. Note that in case of family trios we have great advantage over the method of [10] since we do not need to solve the problem of combining blocks. Indeed, for each family trio we can make four haplotype templates (partially resolved by logic means haplotypes) that imply unique way of gluing together blocks to arrange complete haplotypes for the entire sequence of SNP's. It is not clear what should be the size of a block. We suggest to minimize the variability (i.e., the minimum number of haplotypes) over the block length. Empirically, for Daly et al data, such block has on average 6 SNP's.

Formally, let *genotype* be a vector with m coordinates each corresponding to an SNP and having one of the following values: 0 (homozygote with major allele), 1 (homozygote with minor allele), 2 (heterozygote), or ? (missing SNP value). Let *haplotype* be a vector with m coordinates where each coordinate is either 0 or 1. We say that two haplotypes *explain* a genotype if

- for any 0 (resp. 1) in the genotype vector, the corresponding coordinates in the both haplotype vectors are 0's (resp. 1's),
- for any 2 in the genotype vector, the corresponding coordinates in the two haplotype vectors are 0 and 1,
- for any ? in the genotype vector, the corresponding coordinates in the haplotypes are unconstrained (can be arbitrary).

We say that four haplotypes h_1, h_2, h_3, h_4 *explain* a family trio of genotypes (f, m, k) , if h_1 and h_2 explain the genotype f , h_3 and h_4 explain the genotype m , and h_1 and h_3 explain the genotype k .

Pure-Parsimony Trio Phasing Problem (PPTPP). Given $3n$ genotypes corresponding to n family trios find minimum number of distinct haplotypes explaining all trios.

2 Integer Linear Programs for Pure-Parsimony Trio Phasing

The first two ILP formulations for the PPTPP in this section are inspired by known ILP formulations for phasing from [8] and [4] and the third ILP improves the first two in all three parameters - number of variables, number of constraints and the runtime. We conclude the section with empirical comparison of all three ILP.

The ILP phasing formulation from [8] uses 0-1 variable x_i for each possible haplotype with the minimization objective:

$$\text{Minimize } \sum x_i \quad (1)$$

The main drawback of this approach is in exponential number of possible haplotypes which becomes less critical for blocks of limited size. Indeed, if the block size is b , then regardless of the number of genotypes, there are at most 2^b distinct haplotypes. Although depending on block definition one can find lengthy blocks, e.g. of length 22 [18], b is rarely exceeds 11. Anyway, we can always enforce an appropriate limit on the block size.

The constraints forcing haplotypes to explain given genotypes can be expressed as follows [8]. For any genotype g we introduce a constraint $\sum_{h_i, h_j \text{ explain } g} p_{ij} \geq 1$, where the 0-1 *pair* variable $p_{ij} = 1$ if g can be explained with h_i and h_j in the resulting phasing. That can happen only if the corresponding variables x_i and x_j are set to 1, i.e., $x_i \geq p_{ij}$ and $x_j \geq p_{ij}$.

An obvious adaptation to PPTPP of the above phasing ILP is to have a constraint for each trio t

$$\sum_{h_i, h_j, h_k, h_l \text{ explain } t} q_{ijkl} \geq 1 \quad (2)$$

where the 0-1 *quartet* variables $q_{i,j,k,l} = 1$ if t can be explained with h_i, h_j, h_k, h_l in the resulting phasing. That can happen only if the corresponding variables x_i, x_j, x_k, x_l are set to 1, i.e.,

$$x_i, x_j, x_k, x_l \geq q_{ijkl} \quad (3)$$

The simple ILP (1-3) has too many variables and constraints and can handle only blocks of size at most 4 for Daly et al data (see Table 1).

Our second ILP is based on templates of haplotypes. For each trio we introduce four template haplotypes, i.e., haplotypes with the coordinates 0,1,2 and ?. The values of 0 and 1 correspond to fully resolved SNP's which can be found via logical resolution from the previous section, while 2 corresponds to the fact that there is another template with 2 in the same position such that feasible phasing requires these two values be complementary (0 and 1). The ?'s corresponds to free positions. For each 2 in each template we introduce a 0-1-variable y and constraints connecting each pair of complementary 2's:

$$y + y' = 1 \quad (4)$$

For each ? in each template we also introduce a 0-1 variable z .

Finally, we need to express dependencies between x -variables and y, z -variables. Assume that the template T is resolved by the haplotype h if the variables $y_i, i \in I_0$, are set to 0, $y_i, i \in I_1$, are set to 1, $z_j, j \in J_0$, are set to 0, $z_j, j \in J_1$ are set to 1. Then the x -variable corresponding to h is constrained as follows

$$x \geq 1 + \sum_{i \in I_1} y_i - |I_1| + \sum_{i \in I_2} (1 - y_i) - |I_2| + \sum_{j \in J_1} z_j - |J_1| + \sum_{j \in J_2} (1 - z_j) - |J_2| \quad (5)$$

Indeed, if all y 's and z 's are set as in the haplotype h , then all elements of all four sums are 1's and they will be canceled by subtraction of the number of elements in these sums; thus the right hand side is 1. Otherwise right hand side is at most 0 and x is not constrained.

		Daly Data		
Block Size		LP1	LP2	LP3
4	rt (s)	-	0.103	0.0738
	var	131783	383.73	68.76
	cons	131896	1332.12	1428.08
5	rt (s)	-	0.225	0.121
	var	-	485.28	97.5
	cons	-	1740.76	1663.0
6	rt (s)	-	3.778	0.693
	var	-	620.12	145.82
	cons	-	2396.75	2036.76
7	rt (s)	-	4688.26	2.161
	var	-	746.6	202.13
	cons	-	3193.6	2503.4
8	rt (s)	-	-	3.507
	var	-	-	5142.85
	cons	-	-	598.769

Table 1: The comparison of the running times, number of variables, number of constraints of three linear programs. Each value is averaged over all blocks. All phasing block sizes are uniform.

The ILP (1)-(4)-(5) has considerably less variables and constraints than the first ILP but in case of many ?'s there may be too many variables which can slow down ILP solver (see Table 1).

Our third ILP takes advantage of the fact that ?'s are really not constrained. Instead of completely resolving templates as in constraint (5), we can partially resolve templates, i.e., resolve only 2's. Then several haplotypes can fit partially resolved templates and at least one of the corresponding x -variables should be set to 1, i.e., for any y -assignment of 2's in each template T ,

$$\sum_{x \text{ fits all } y\text{'s in template } T} x \geq 1 + \sum_{i \in I_1} y_i - |I_1| + \sum_{i \in I_2} (1 - y_i) - |I_2| \quad (6)$$

The last constraint is not completely equivalent to (5) since now we should guarantee that each template is resolved. This is guaranteed by the following constraint. For each template T ,

$$\sum_{x \text{ fits template } T} x \geq 1 \quad (7)$$

The ILP (1)-(4)-(6)-(7) can resolve longer blocks (see Table 1).

References

- [1] The International HapMap Project. <http://www.hapmap.org>
- [2] H.Ackerman et al(2003). Haplotypic analysis of the TNF locus by association efficiency and entropy, *Genome Biology*, 4:R24.
- [3] M. Anderson , Crohn's: An Autoimmune or Bacteria-Related Disease? *The Scientist* 15[16]:22, Aug. 20, 2001

- [4] D.G.Brown and I.M. Harrower. A new integer programming formulation for the pure parsimony problem in the haplotype association. WABI, 2004.
- [5] M. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander. High resolution haplotype structure in the human genome. *Nature Genetics*, 29:229–232, 2001.
- [6] Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296:2225-2229.
- [7] Goldestein, D. and Weale, M. Population genomics:Linkage disequilibrium holds the key. *Current Biology*, 11, R-576-R-579, 2001
- [8] D. Gusfield. Haplotype inference by pure parsimony. In R. Baeza-Yates, E. Chavez, and M. Chrochemore, editors, 14'th Annual Symposium on Combinatorial Pattern Matching (CPM'03), volume 2676 of Springer LNCS, pages 144–155, 2003.
- [9] L. Helmuth. Genome research: Map of the human genome 3.0. *Science*, 293(5530):583–585, 2001.
- [10] E. Halperin and E.Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*. Advance Access published on February 26, 2004.
- [11] E. Halperin and Richard M. Karp. Perfect phylogeny and haplotype assignment. RECOMB, 2004.
- [12] J. He and A. Zelikovsky. Linear Reduction for Haplotype Inference. Proc. Workshop on Algorithms in Bioinformatics (WABI'04), September 2004, Lecture Notes in Bioinformatics (LNBI) 3240, 242-253.
- [13] J. He and A. Zelikovsky. Linear Reduction Methods for Tag SNP Selection. Proc. International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'04), September 2004, 2840-2843.
- [14] R. Hudson. Gene genealogies and the coalescent process. *Oxford Survey of Evolutionary Biology*, 7:1–44, 1990.
- [15] N. Patil et al(2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294, 1719-23.
- [16] D. Reich et al(2001).Genetic variation in the 5cytokine gene cluster confers susceptibility to crohn disease. *Nature Genetics*, 29, 223-8.
- [17] H Zhao, R Pfiffer, and MH Gail. Haplotype analysis in population genetics and association studies, *Pharmacogenomics*, 4:171-178, 2003.
- [18] Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, and Lander ES High-resolution haplotype structure in the human genome. *Nat Genet* 29:229-232, 2001.
- [19] Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M,Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D The structure of haplotype blocks in the human genome. *Science* 296:2225–22, 2002.

- [20] Johnson, G.C.L., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Genova, G.D., Ueda, H., Cordell, H.J., Eaves, I.A., Dubbridge, F., Twells, R.C.J., Huges, W., Stevens, H., Phillipa, C., Tuomilehto-Wolf, E., Tuomilehto, J., Gough S.C.L., Clayton D.G., Todd, J.A. Haplotype tagging for the identification of common disease genes. *Nature Genetics* 29: 233–237, 2001.
- [21] Zhang, K., Calabrese, P., Nordborg, M., Sun, F. Haplotype block structure and its applications in association studies: power and study design. *The American Journal of Human Genetics*, 71: 1836–1894, 2002.
- [22] Clark AG. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr Opin Genet Dev.* (2003) Jun;13(3):296-302.
- [23] M. Stephens, Smith, N.J., and P. Donnelly . A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*,(2001) 68:97898, 2001.
- [24] Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered.* 2003;56(1-3):18-31, 2003.
- [25] Zhang W, Collins A, Morton NE. Does haplotype diversity predict power for association mapping of disease susceptibility? *Hum Genet.* 2004 Jul;115(2):157-64. Epub 2004 Jun 04, 2004.
- [26] J. Bell. Predicting disease using genomics. *Nature* 429, 453–456 (27 May 2004)
- [27] SVMlight. http://www.cs.cornell.edu/People/tj/svm_light/
- [28] Basic Classification Trees. <http://www.ece.wisc.edu/~nowak/ece901/lecture11.pdf>
- [29] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.