Statement of Ongoing Work: Extending Boolean Satisfiability Techniques for Haplotype Inference by Pure Parsimony

Eric I. Hsu and Sheila A. McIlraith

Department of Computer Science
University of Toronto
{eihsu,sheila}@cs.toronto.edu

Abstract. Here the authors overview an ongoing effort to extend satisfiability-based methods for haplotype inference by pure parsimony (HIPP). This genome analysis task, first formulated as a boolean satisfiability probelm by Lynce and Marques-Silva [1], has been performed successfully by modern SAT-Solvers. But, it is not as widely used as some better-publicized statistical tools, such as PHASE [2]. This paper presents the authors' assessment of the current situation, and a preliminary statement of intention concerning their aims in this area. Namely, the situation suggests three categories of improvements for making HIPP more widely-used within the biological community: 1) the ability to handle larger problems; 2) more detailed empirical understanding of the accuracy of the "pure parsimony" criterion; and 3) additional criteria and methods for improving on this level of accuracy.

1 Background

As detailed in a recent overview paper [3], the haplotype inference problem is defined over a population of individuals represented by their respective genotypes. Each genotype can be viewed as a sequence of nucleotide pairs, where the two values of each pair are split across the individual's two chromosomes as inherited from their two parents. Much of the genetic variation between individuals consists of point mutations at known sites within this sequence, known as single nucleotide polymorphisms (SNP's). Thus, a genotype can represented (with loss of information) as a sequence of nucleotide pairs at successive SNP sites.

In particular, at each such site, an individual might have two copies of the "minor allele"– in this case, the presumably mutated or at least rarer of the two possible nucleotide values. At this site the individual is then *homozygous (major)*. Similarly, a SNP site is realized on each chromosome by the nucleotide value that is most common for the species then the site is *homozygous (minor)* for this individual. The third possibility is that one of the individual's chromosomes has the major allele at a given site, while the other chromosome has the minor allele at that site–then the genotype is *heterozygous* at the site in question. Notationally, the first case can be represented by the character '0', the second by '1', and the third, heterozygous case, by '2'. So, the sequence "0102" indicates an individual with two minor alleles at both the first and the third SNP sites measured by a particular study, and two major alleles at the second site. At the fourth site, we know that one chromosome exhibits the major allele and the other exhibits the minor. Thus, an individual's genotype is well-defined by the sequences of its two constituent chromosomes; these two individually inherited sequences are called haplotypes. If we overload '0' and '1' to indicate a single minor or single major allele in a particular haplotype, then we can denote that the genotype "0102" arises from the two haplotypes "0100" and "0101".

Conversely, though, a genotype with more than one heterozygous site can be explained by multiple pairs of haplotypes, as the major and minor alleles at all heterozygous sites in the genotype can be permuted across corresponding sites in the two haplotypes. More concretely, the genotype "02122" could be realized the pairs "00100"/"01111", "00101"/"01110", "00110"/"01101", or "01100"/"00111". (Naturally, the term "pair" is used colloquially here, and does not signify any sense of ordering within the haplotype sets of size two that arise in this domain.) Accordingly, there are 2^{k-1} candidate haplotype pairs for explaining a genotype with k heterozygous sites.

In practice, this distinction is made relevant by the lack of any practical experimental method to measure an individual's two haplotypes instead of its genotype; in other words, the machinery can identify sites at which the two haplotypes have opposing values, but cannot tell which values are grouped together on which chromosome. The goal of haplotype inference is to guess the most likely haplotypes that generated a given set of genotypes.

2 Underlying Biological Principles

How can one answer to the haplotype inference problem be preferred to any other? Because the ultimate goal is to accurately predict haplotypes appearing in a particular subject (*i.e.* human) population, haplotype inference frameworks must apply standards that seek to model the types of phenomena that actually drove the true state of affairs in the evolution of the subject species' genome. In other words, human haplotypes are not drawn uniformly from the space of all possible pairs that could explain human genotypes. Rather, under the coalescent model of evolution there should be only a small number human haplotypes that were recombined and mutated to produce any of the genotypes within a given haplotype inference problem instance. Thus, early researchers used greedy methods to try to minimize the set of answers to a haplotype inference problem [4], while later systems integrated models based on "perfect phylogeny" [5] or other hierarchical organizations of answer haplotypes [6]. The most widely-used techniques at this point in time integrate empirical statistical measures of likely haplotypes [7–9, 2, 10], by analyzing the population in question, or consulting outside sources [11]. Such approaches may not require that the inferred set of answer haplotypes can be arranged into a particular evolutionary structure, but they all require the haplotype set to be maximally likely according to a particular statistical model that has been fit to the problem and/or outside frequency data.

On the other hand, the "pure parsimony" methodology seeks to capture such phenomena implicitly by asking directly for the *smallest* possible set of haplotypes that as a whole can explain a given population of genotypes [12]. Methods that are based on this criterion thus perform "HIPP", indicating haplotype inference by pure parsimony. This pure parsimony principle has been achieved optimally and efficiently by employing a variety of satisfiability-based techniques on small to moderately-sized data sets of about 200 sites and 100 individuals [1, 13–15]. However, aside from a preliminary and limited evaluation of pure parsimony, which did not include such satisfiability-based techniques [16], the overall accuracy and general feasibility HIPP does not seem to be well-understood within the biological community.

That is, satisfiability-based HIPP methods must demonstrate two forms of feasibility in order to achieve wider adoption within the biological community: the empirical accuracy of the pure parsimony principle itself, and the efficiency and scalability of SAT methods for achieving this parsimony criterion. While SAT-based techniques can achieve optimal parsimony, on reasonably large data sets, they still cannot be applied to the massive collections characteristic of more popular biological applications that may require on the order of half a million sites. Addressing the issue of scalability will not only enable the assessment of model accuracy and solver efficiency, but can additionally lead to more accurate results. This is because multiple minimum haplotype sets can explain the same population of genotypes [3], but some of them can be safely considered more likely *a priori*. For instance, solution sets whose members are more similar to each other are more realistic with respect to evolutionary theory [2, 10], and certainly *one* must certainly be preferred *a posteriori* with respect to the truth–in the real world there was a single haplotype set that produced a set of genotypes (and this set may or not be minimum.) So while the initial goal is evaluation of the HIPP principle and HIPP solvers, and the prerequisite goal is improved solver scalability, in pursuing these we would like to synergistically attain the overarching goal of making the most accurate predictions possible, as opposed to merely minimal ones.

We propose to do this by making finer-grained use of biological principles in designing SAT-based methodologies for HIPP. One prominent example of such phenomena would be "linkage disequilibrium", or correlation between sites within a sequence [17]. The sequential and mostly non-random nature of genotypes and haplotypes are at core of many of the optimizations and models that underly statistical alternatives to HIPP; finding a way to exploit them within a discrete reasoning framework would make HIPP competitive in efficiency and accuracy. Three proposals for doing so are outlined in the next section.

3 Proposals for Extending the Satisfiability-Based HIPP Framework

In this section we propose three general types of improvements to the HIPP framework that can make it competitive to the current methods of choice within the biological community.

- **Exploiting sequential structure to improve scalability.** The best-performing (and most widely-used) statistical systems [7, 8, 2] for haplotype inference utilize explicit or implicit variants of the "partition ligation" scheme of Niu *et al* [18]. The basic idea is to escape the combinatorial explosion of considering the entire space of possible haplotypes for a given sequence, and instead break the sequence into blocks. Each such block is small enough to be solved efficiently and to high accuracy; they are then recombined heuristically through a polynomial-time merging scheme. With the merging scheme comes a loss of optimality; in the case of HIPP we may not get the smallest possible explaining haplotype set by means of this scheme. But, the insight of the partition ligation scheme is that evolution does not create haplotypes uniformly at random over all possible explanations for the human genotype, and in practice the loss in optimality has proved negligible in comparison to the gains in accuracy for statistical methods [18, 7, 2]. For beyond being biologically justifiable, block partitioning can actually produce more accurate overall results because each block can be solved to higher standards of likelihood using more computationally expensive methods. For instance, a statistical approach based on MCMC sampling can choose to perform vastly more iterations on a more complex model within the confines of a small block, while realistically hoping to retain much of the benefit during the merging process [2]. Applying this framework to SAT-based approaches can provide similar gains, whether using parsimony alone or integrating statistical information in solving blocks.

- **Assess the accuracy of HIPP on large-scale data.** When HIPP is able to handle larger data sets, it will be possible to directly compare its accuracy with other, more-popular approaches. This will allow an assessment of the model's strengths and weaknesses and inform any attempts to improve this accuracy. At this point, SAT methods have been highly successful at solving HIPP, but it remains to be seen whether the resulting answers are themselves successful at modeling the human genome–as mentioned previously here and elsewhere [3], there can be many minimum sets that all qualify as HIPP solutions, while some of these are much more empirically likely than others. Characterizing which types of these solutions usually turn out to be more accurate will go a long way to improving the parsimony model's fit to real populations.
- **Exploiting linkage disequilibrium to improve accuracy.** In the same spirit as the first two points, there are important correlations between various regions of the vast majority of a species' haplotypes, due to the recombination and mutation processes that drive evolution. To compete with statistically-oriented tools, HIPP can be extended to encompass the same sort of empirical information concerning such correlations [19, 20]. This may entail a weighted SAT or a MAXSAT formulation that favors solutions that adhere to stronger correlations that have been observed from the same sources of data as used by the competitor techniques.

4 Conclusion

The authors have begun to implement block decomposition within the SAT-based HIPP framework, but this description of research is decidedly preliminary and strictly for expository purposes. Once the system is able to handle large problems, the next step will be to study its accuracy, especially in terms of adding additional solution criteria to pure parsimony. The concluding step will be to use information derive from individual problem-instance and/or reference data to actually achieve such criteria. It would be a great opportunity to be able to discuss such plans with others who are working on this problem and related areas!

References

1. Lynce, I., Marques-Silva, J.: Efficient haplotype inference with boolean satisfiability. In: Proc. of 21st National Conference on Artificial Intelligence (AAAI '06), Boston, MA. (2006)
2. Stephens, M., Scheet, P.: Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. American Journal of Human Genetics **76** (2005) 449–462
3. Lynce, I., Graça, A., Marques-Silva, J., Oliveira, A.L.: Haplotype inference with boolean constraint solving: An overview. In: Proc. of 20th IEEE Int'l Conf. on Tools with Artificial Intelligence (ICTAI '08), Dayton, OH. (2008)
4. Clark, A.G.: Inference of haplotypes from PCR-amplified samples of diploid populations. Molecular Biology and Evolution **7**(2) (1990) 111–122
5. Kimmel, G., Shamir, R.: GERBIL: Genotype resolution and block identification using likelihood. PNAS **102**(1) (2005) 158–162
6. Xing, E.P., Sohn, K.A., Jordan, M.I., Teh, Y.W. In: Bayesian Multi-Population Haplotype Inference via a Hierarchical Dirichlet Process Mixture. (2006)
7. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., Donnelly, P.: A comparison of phasing algorithms for trios and unrelated individuals. American Journal of Human Genetics **78** (2006) 437–450
8. Browning, S.R.: Missing data imputation and haplotype phase inference for genome-wide association studies. Human Genetics **124** (2008) 439–450
9. Salem, R.M., Wessel, J., Schork, N.J.: A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Human Genomics **2**(1) (2005) 39–66
10. Scheet, P., Stephens, M.: A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. American Journal of Human Genetics **78** (2006) 629–644
11. The International HapMap Consortium: A second generation human haplotype map of over 3.1 million SNPs. Nature **449** (2007) 851–862
12. Gusfield, D.: Haplotype inference by pure parsimony. In: Proc. of 14th Symp. on Combinatorial Pattern Matching (CPM '03), Morelia, Mexico. (2003) 144–155
13. Graça, A., Marques-Silva, J., Lynce, I., Oliveira, A.L.: Efficient haplotype inference with pseudo-boolean optimization. In: Proc. of 2nd Int'l Conf. on Algebraic Biology (AB '07), Linz, Austria. (2007) 125–139
14. Erdem, E., Türe, F.: Efficient haplotype inference with answer set programming. In: Proc. of 23rd National Conference on A.I. (AAAI '08), Chicago, IL. (2008) 436–441
15. Lynce, I., Marques-Silva, J., Prestwich, S.: Boosting haplotype inference with local search. Constraints **13**(1-2) (2008) 155–179
16. Wang, L., Xu, Y.: Haplotype inference by maximum parsimony. Bioinformatics **19**(14) (2003) 1773–1780
17. Slatkin, M.: Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics **9** (2008) 477–485
18. Niu, T., Qin, Z.S., Xu, X., Liu, J.S.: Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. American Journal of Human Genetics **70**(1) (2002) 157–169
19. Excoffier, L., Slatkin, M.: Maximum-Likelihood estimation of molecular haplotype frequencies in a diploid population. Molecular Biology and Evolution **12**(5) (1995) 921–927
20. Kuhner, M.K.: LAMARC 2.0: Maximum likelihood and Bayesian estimation of molecular haplotype frequencies in a diploid population. Bioinformatics **22**(6) (2006) 768–770