

Haplotype Inference by Entropy Minimization

Ion Mandoiu¹ and Bogdan Paşaniuc¹

Keywords: haplotype inference, entropy minimization

1 Introduction.

The completion of the Human Genome Project and the identification of millions of SNPs in the human population has opened the way for large-scale association studies between genetic variation and susceptibility to common diseases. Constructing a haplotype map of the human population has become the next major target in genomics (<http://www.hapmap.org/>). The *population haplotyping problem* (PHP) is to infer the haplotypes from the genotypes of a large population; see [1, 2, 3] for recent surveys on computational methods for solving this problem. Recently, Halperin and Karp [4] introduced a novel approach to PHP based on the entropy minimization principle, which suggests that most common haplotypes are evenly distributed in large populations. They also proposed a simple greedy entropy minimization algorithm for missing allele imputation and haplotype inference. In this poster we show that a local improvement entropy minimization algorithm yields significantly higher phasing accuracy than the algorithm of [4], coming close to the phasing accuracy of the widely used but much slower PHASE algorithm [5].

2 Background and Problem Definition

A *Single Nucleotide Polymorphism*, or SNP, is a position in the genome at which exactly two of the possible four nucleotides occur in a large percentage of the population. SNPs account for most of the genetic variability between individuals, and mapping SNPs in human population has become the next high-priority in genomics after the completion of the Human Genome project. In diploid organisms such as humans, there are two non-identical copies of each chromosome. A description of the SNPs in each chromosome is called a *haplotype*, which can be viewed as a 0/1 vector, e.g., by representing the most frequent (dominant) SNP allele as a 0 and the alternate (minor) allele as a 1. At present, it is prohibitively expensive to directly determine the haplotypes of an individual, but it is possible to obtain rather easily the conflated SNP information in the so called *genotype*. A genotype can be conveniently represented as a 0/1/2 vector, where 0 (1) means that both chromosomes contain the dominant (respectively minor) allele, and 2 means that the two chromosomes contain different alleles.

We say that a haplotype h is *compatible* with a genotype g if $g(i) = h(i)$ whenever $g(i) \in \{0, 1\}$. A pair of haplotypes (h_1, h_2) *explains* g if $h_1(i) = h_2(i) = g(i)$ whenever $g(i) \in \{0, 1\}$, and $h_1(i) \neq h_2(i)$ whenever $g(i) = 2$. For a given pair (h_1, h_2) that explains g we say that h_2 is the complement of h_1 with respect to g . A *phasing* of a set of genotypes $\mathcal{G}$, each of length k , is a function $f : \mathcal{G} \rightarrow \{0, 1\}^k \times \{0, 1\}^k$, such that, for every $g \in \mathcal{G}$, $f(g)$ is a pair of haplotypes that explain g . For a haplotype h and a phasing f , the *coverage of h under f* , denoted by $cov(h, f)$, is the number of genotypes $g \in \mathcal{G}$ such that $f(g) = (h, h')$ or $f(g) = (h', h)$ plus twice the number of of genotypes $g \in \mathcal{G}$ such that $f(g) = (h, h)$. Following [4], the *entropy* of a phasing f is defined as $ENT(f) = \sum_{h: cov(h, f) \neq 0} -\frac{cov(h, f)}{2|\mathcal{G}|} \log \frac{cov(h, f)}{2|\mathcal{G}|}$.

Minimum Entropy PHP: Given a set of genotypes, find a phasing with minimum entropy.

3 Implemented Algorithms and Results

We have implemented the greedy algorithm of Halperin and Karp as described in [4]. At each step, the algorithm chooses the haplotype h that explains the maximum number of unexplained genotypes; as proved in [4], this algorithm gives a solution whose entropy is within an additive factor of 3 of the optimum entropy. We have also implemented a local optimization algorithm for entropy minimization. Our algorithm, which we refer to as 1-OPT, starts from a random phasing, then, at each step, finds the genotype whose re-explanation yields the largest decrease in phasing entropy. We also included in our comparison (a) the well-known PHASE algorithm [5], (b) a random phasing, and (c) the solution constructed by 1-OPT when starting from the greedy algorithm phasing.

In order to test the entropy approach, we used a program developed by R.Hudson [6] to generate populations of between 50 and 200 individuals, and between 10 and 30 SNP sites. For each population size we generated 10

¹Computer Science and Engineering Department, University of Connecticut, 371 Fairfield Rd., Unit 2155, Storrs, CT 06269-2155. E-mail: {ion,bogdan}@engr.uconn.edu. Partially supported by a "Large Grant" from the University of Connecticut's Research Foundation.

Table 1: Haplotype (H) and SNP (S) phasing accuracy for recombination rate 0.

#Gen	#SNP	ENT OrigSol	PHASE			Local Opt.	Greedy			Random		
			H(%)	S(%)	ENT		H(%)	S(%)	ENT	H(%)	S(%)	ENT
100	10	1.24	99.20	99.93	1.24	None	75.70	97.11	1.44	69.50	95.97	1.72
						1OPT	94.10	99.24	1.29	98.90	99.89	1.24
100	20	1.81	98.00	99.89	1.80	None	70.00	98.26	1.96	53.00	94.93	2.63
						1OPT	92.00	99.52	1.83	94.90	99.68	1.80
100	30	2.24	97.56	99.91	2.24	None	75.22	98.65	2.27	43.33	95.65	3.27
						1OPT	83.89	99.19	2.21	86.44	99.37	2.21
200	10	1.53	97.30	99.98	1.37	None	80.75	98.20	1.50	71.20	96.08	1.83
						1OPT	96.70	99.91	1.37	96.80	99.91	1.37
200	20	1.94	95.25	99.97	1.70	None	78.15	98.59	1.84	44.65	94.04	2.77
						1OPT	93.85	99.87	1.70	95.00	99.93	1.70
200	30	2.28	96.35	99.96	2.13	None	59.65	97.68	2.44	37.50	94.35	3.50
						1OPT	91.70	99.72	2.15	84.65	98.74	1.94

Table 2: Haplotype (H) and SNP (S) phasing accuracy for recombination rate 40.

#Gen	#SNP	ENT OrigSol	PHASE			Local Opt.	Greedy			Random		
			H(%)	S(%)	ENT		H(%)	S(%)	ENT	H(%)	S(%)	ENT
100	10	2.07	89.00	98.86	2.03	None	75.20	97.29	2.09	70.20	96.61	2.31
						1OPT	89.30	98.84	2.01	88.30	98.74	2.02
100	20	3.09	86.10	99.11	3.05	None	47.50	96.03	3.17	30.70	94.25	4.13
						1OPT	67.70	97.61	3.01	70.30	97.80	3.02
100	30	3.29	85.70	99.27	3.27	None	39.80	96.02	3.42	18.20	93.91	4.64
						1OPT	65.00	97.76	3.25	78.10	98.65	3.19
200	10	1.96	91.00	99.37	1.77	None	81.35	98.36	1.85	72.50	97.28	2.04
						1OPT	88.65	99.14	1.79	90.40	99.30	1.78
200	20	3.19	86.45	99.38	2.99	None	52.25	96.64	3.22	31.65	94.94	4.13
						1OPT	80.60	98.92	2.98	82.35	98.95	2.97
200	30	3.47	91.80	99.67	3.39	None	41.35	96.76	3.65	29.30	96.30	4.51
						1OPT	79.95	99.10	3.37	85.10	99.33	3.34

instances; average haplotype and SNP phasing accuracies for the compared methods are reported in Table 1 and Table 2 for recombination rates of 0, respectively 40.

The results we obtained show that the local improvement entropy minimization algorithm yields higher phasing accuracy than the greedy algorithm, coming close to the phasing accuracy of the PHASE algorithm.

References

- [1] P. Bonizzoni, G. Della Vedova, R. Dondi, and J. Li. The haplotyping problem: An overview of computational models and solutions. *Journal of Computer Science and Technology*, 18:675–688, 2003.
- [2] D. Gusfield. An overview of combinatorial methods for haplotype inference. In *Proc. of the DIMACS/RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotype Inference*, volume 2983 of *Lecture Notes in Bioinformatics*, pages 9–25, Berlin, 2004. Springer-Verlag.
- [3] B.V. Halldorsson, V. Bafna, N. Edwards, R. Lippert, S. Yooseph, and S. Istrail. A survey of computational methods for determining haplotypes. In *Proc. of the DIMACS/RECOMB Satellite Workshop on Computational Methods for SNPs and Haplotype Inference*, volume 2983 of *Lecture Notes in Bioinformatics*, pages 26–47, Berlin, 2004. Springer-Verlag.
- [4] E. Halperin and R. Karp. The Minimum-Entropy Set Cover Problem. *International Colloquium on Automata Languages and Programming 2004*
- [5] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics* 68:978–989, 2001
- [6] R. Hudson. Generating samples under the Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338, 2002