A Heuristic Greedy Algorithm for Haplotype Inference from Genotype by Pure Parsimony

Mehrdad Azadi[a]*, Hadi Poormohammadi[b]

[a] Department of mathematics, Central Tehran Branch, Islamic Azad University, Tehran, Iran
[b] Computer Engineering Department, Haeri University, Meibod, Iran

Abstract. Haplotypes provide important information in the study of complex diseases and drug design. However, due to technical limitations, genotype rather than haplotype data are usually obtained. Thus, haplotype inference from genotype data using computational methods is of interest. There are several models in the literature for inferring haplotypes. One of the most important models is haplotype inference by pure parsimony (HIPP), consisting of finding the minimum number of haplotypes that can resolve all given genotypes. It has been shown that HIPP is an NP-hard problem. In this paper, we propose a new heuristic greedy algorithm for this problem. The greedy algorithm predicts an efficient haplotype for inferring the remaining genotypes in each step. By applying our algorithm on a variety of biological and simulated data we show that the proposed method is more effective and accurate compared to other available algorithms. Additionally, we introduce a new measure for evaluating the effectiveness of the algorithms. This measure is based on the pure parsimony approach which finds the minimum number of haplotypes for resolving the input genotypes.

Keywords: Haplotype; Genotype; Gene; Greedy algorithm.

1. Introduction

Sequencing of the human genome is certainly one of the important milestones at the beginning of the 21st century in biology and genetics. Since then, characterization of genetic variations among human populations has attracted increasing attention and become one of the most important topics in genomics.

Single Nucleotide Polymorphisms (SNP's) are the most frequent form of human genetic variations. The genomes of two individuals are the same in 99.9% of nucleotide positions and the differences occur in SNP sites that are about 0.1% of nucleotide positions [1]. In human genome, all nucleotide sites except SNP positions are the same and SNP sites are biallelic, i.e., exactly two different nucleotides are observed at each SNP site [2]. In human genome there are two copies of each chromosome. A description of SNP›s from a single copy is called a haplotype, while a description of the SNP's from the two copies is called a genotype. Haplotype information is more informative than genotype data for a variety of purposes such as disease gene mapping, drug design and inferring population history [3]. Obtaining haplotype information via experimental methods is both expensive and time consuming whereas genotypes information are much easier to obtain. Current sequencing technologies typically determine genotypes rather than haplotypes. Such restrictions in experimental methods and technologies force us to infer haplotypes from genotype data, which is called haplotype inference. In more words haplotype inference problem (HIP) is as follows: given a set of genotypes with the same length, find a set of haplotypes such that each given genotype can be expressed as a combination of a pair of haplotypes.

Clark was the first who proposed an algorithm for HIP [4]. Clark algorithm is applied widely and is an efficient method for HIP. Finding effective methods for solving haplotype inference problem is very attractive and important to researchers in the field of biology. So a number of different methods have been developed for HIP.

In general, there are two major types of methods for solving HIP: combinatorial and statistical methods. Combinatorial methods are usually based on some optimization criteria [5], whereas the statistical methods are based on haplotype evolution [6]. In contrast to combinatorial methods, statistical methods are usually time consuming.

One of the most popular combinatorial approaches to HIP is Haplotype Inference by Pure Parsimony (HIPP). For a set of given genotypes with the same lengths, finding the minimum number of haplotypes is the goal of HIPP in which each given genotype can be expressed as a combination of a pair of haplotypes.

Gusfield was formulated HIPP problem and proposed an

* Corresponding author.
E-mail address: Meh.azadi@iauctb.ac.ir

integer linear program (ILP) algorithm to solve it [7]. It should be noted that in natural populations the number of observed distinct haplotypes is much smaller than the number of combinatorial possible haplotypes that can be inferred from all given genotypes. Moreover, the minimum recombination principle is satisfied in natural population. The minimum recombination principle states that the genetic recombination are rare and thus haplotypes with fewer recombinations should be preferred in a haplotype reconstruction [8, 9]. Therefore, the pure parsimony criterion for HIP is reasonable and biologically meaningful [7, 10].

The HIPP problem is APX-hard (and therefore NP-hard) [11]. For this reason different kind of heuristic algorithms were proposed. One of the most famous algorithms for solving HIPP is Branch and Bound algorithm. Wang & Xu proposed an exact branch and bound algorithm called HAPAR to find the optimal solution of HIPP [10]. The branch and bound is an exponential algorithm for which the increase in the size of genotype matrix results in serious problems.

Recently Brwon & Harrower [12] proposed an integer linear program algorithm that is more efficient compared to the first algorithm proposed by Gusfield [7]. Similar to branch and bound, this algorithm is an exponential algorithm which is efficient only for small size data. When the number of genotypes or SNP sites is relatively large this method is not applicable.

Another heuristic method, GAHAP, based on genetic algorithm is proposed in [13]. PTG is another heuristic algorithm based on parsimonious tree-grow method for HIPP problem [14]. This is an effective algorithm in time and complexity in comparison with other methods. HAPINFERX is an implementation of Clark's algorithm provided by [4]. Also methods are proposed for the HIPP based on boolean satisfiability, SAT (SHIPs [15]) and pseudo boolean optimization (PolyPB, RPoly and NRPoly [16, 17]). In [18] the authors proposed a new preprocessing step for the HIPP problem and then solved this problem using GAHAP and integer linear program (ILP) algorithms.

In statistical methods, the frequencies of haplotypes are used for inferring haplotypes from genotype data based on the pure parsimony criterion. Statistical methods such as PHASE [6] and HAPLOTYPER [19] are widely used. PHASE algorithm is usually compared with combinatorial methods for HIPP problem.

Algorithms for optimization problems usually go through a series of steps, with a set of choices at each step. A greedy algorithm always selects the choice that produces the best results in the current step. In other words, it selects a locally optimal choice in the hope that this choice will lead to a global optimal result. It should be noted that greedy algorithms do not always produce optimal solutions, but for many problems they do. Thus, they are quite powerful and work well for a wide range of problems. Also, greedy algorithms are usually very fast because they make their selections locally. In this paper we propose a new greedy algorithm for HIPP problem. Our greedy algorithm predict an efficient haplotype for inferring the remaining genotypes in each step. This haplotype is selected using the information of the remaining genotypes and the haplotypes which are already added to the set of output haplotypes. Results of applying this algorithm on a variety of biological and simulated data show that it is very effective with a high accuracy in comparison with other algorithms. Moreover the order of our algorithm is O(m3n) where m is the number of genotypes and n is the number of SNP sites each genotype contains.

The rest of this paper is organized as follows: section 2 includes definitions and notation. In section 3, we present our greedy algorithm in detail. In section 4, we compare our results with the results obtained by some other methods. Concluding remarks are presented in section 5.

2. Definitions and notation

An SNP is a single nucleotide site where exactly two different nucleotides are observed. Hence each SNP site can be characterized by the elements 0 or 1. A haplotype is a sequence of SNPs. In diploids genome there are two copies of each chromosome. The SNP positions that contain the same nucleotide from the two copies of a chromosome are called homozygous, otherwise they are called heterozygous. The alphabet 2 is used for characterizing heterozygous site. Hence a haplotype is a sequence of alphabets in {0,1} and a genotype is a sequence of alphabets in {0,1,2}.

For every two elements x, y ϵ {0,1} define

$$x \oplus y = \begin{cases} 0 & if\ x = y = 0; \\ 1 & if\ x = y = 1; \\ 2 & if\ x \neq y, \end{cases}$$

Let $h_1=(h_{11},h_{12},\dots,h_{1n})$ and $h_2=(h_{21},h_{22},\dots,h_{2n})$ are two haplotypes of the same length n. Define $h_1 \oplus h_2 = (h_{11} \oplus h_{21}, h_{12} \oplus h_{22} \dots, h_{1n} \oplus h_{2n})$ given a pair (h_1, h_2) of haplotypes and a genotype g, we say that h1 and h2 resolve g if $h_1 \oplus h_2 = g$. Thus, for every given genotype g with k elements 2 in its sequence, there are 2k-1 pairs of haplotypes which can resolve g. In a genotype, an SNP position is called resolved if it has values 0 or 1 and otherwise it is called ambiguous. A genotype g that contains at most one element 2 is called resolved, and it is called ambiguous otherwise.

Two genotypes $g_1 = (g_{11}, g_{12}, \dots, g_{1n})$ and $g_2 = (g_{21}, g_{22}, \dots, g_{2n})$ are in conflict if there exists $1\leq i \leq n$ such that g_{1i}, g_{2i} ϵ {0,1} and $g_{1i} \neq g_{2i}$. Two genotypes g_1 and g_2 are compatible if they are not in conflict.

Let $G=\{g_1, g_2, \dots, g_m\}$ be a set of m genotypes of the same length n and $g_i=\{g_{i1}, g_{i2}, \dots, g_{in}\}$. The genotype matrix corresponding to G is defined by $M = (g_{ij})$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. Now HIPP problem is described as follows.

HIPP problem: Given a genotype matrix M, find the minimum number of haplotypes in a way that for each genotype g (rows of M) there exist two haplotypes h_1 and h_2 such that they resolve g, i.e. $h_1 \oplus h_2 = g$.

3. Methods

Suppose there is a set G = {g_1, g_2, …, g_m} of genotypes and assume that M is its corresponding genotype matrix. We describe our greedy algorithm in 6 steps. In each step some genotypes are removed from G. This algorithm will continue until G becomes empty. Consider an empty set H as representative set of haplotypes. In each step at least one haplotype is added to H.

Step 1: Genotypes that do not have homozygous positions are in fact haplotypes. In this step all such genotypes are added to H and removed from G.

Step 2: Consider all rows of M which have exactly one 2 in their entries. Let g_i = {g_{i1}, g_{i2}, …, g_{in}} be such a row. Define the haplotypes h = {h_{i1}, h_{i2}, …, h_{in}} and h´ ={h'_{i1}, h'_{i2}, …, h'_{in}} that satisfy h ⊕ h´ = g_i. For 1 ≤ j ≤ n, h_{ij} = h'_{ij} = g_{ij} if g'_{ij} {0, 1} and h_{ij} = 1, h'_{ij} = 0 if g_{ij} = 2. Now h ⊕ h´ = g_i. Add h, h´ to H and remove g_i from G.

Step 3: Each remaining genotype of G which can be resolved with at least one pair of the elements of H is removed from G.

Step 4: Assume H = {h_1, h_2, …, h_t}. Starting from step 1 and ending up with step 3 we have G = {g_1, g_2, …, g_k}. Therefore none of g_i, 1 ≤ i ≤ k can be resolved by a pair of elements of H. For 1 ≤ i ≤ t, define A(h_i)={h | ∃ g ϵ G, h ⊕ h_i = g}. It is obvious that for 1 ≤ i ≤ t A(h_i) ∩ H = Ø. Define A = $U^t_{i=1}$A(h_i). If A = Ø go to step 5 of the algorithm.

Let A ≠ Ø . For each h ϵ A define B_h = {g ϵ G | h_i ⊕ h = g for some 1 ≤ i ≤ t} and C_h = {g ϵ G | ∃ h´ ϵ H, h ⊕ h´ = g}. Let b_h and c_h be the size of B_h and C_h respectively. Define d_h = (b_h +1) × (c_h +1). Let h_0 be an element of A such that d_{h0} = Max {d_h ϵ | h ϵ A}. We add h_0 to H and remove B_{h0} from G. Now go to step 3.

Step 5: If each pairs of genotypes are in conflict then go to step 6. Otherwise consider all pairs of compatible genotypes and randomly choose one of these pairs such as g_1 and g_2. So there exists a haplotype h and two haplotypes h_1 and h_2 such h_1 ⊕ h = g_1 and h_2 ⊕ h = g_2. Add h to H. Go to step 4.

Step 6: For each genotype g ϵ G add two haplotypes h_1 and h_2 at random to H such that h_1 ⊕ h_2 = g and the algorithm halts.

4. Results and discussion

The greedy algorithm is applied on both real and simulated data sets and the performance of the algorithm is compared with the existing algorithms for which some results are reported. To analyze the performance of greedy algorithm, we use the error rate factor which is frequently used in HIP [14]. The error rate factor is the proportion of genotypes whose original haplotypes are inferred incorrectly by the algorithm. For example, suppose that the number of input genotypes are y, and the number of genotypes that are incorrectly inferred are x. Thus, the error rate is x/y. Since we use the pure parsimony approach, the goal is to find the minimum possible number of haplotypes for inferring the input genotypes. Thus, we define another criterion for measuring the performance of the proposed algorithm. This criterion is the number of haplotypes that the algorithm produces as the output. If it is less than or equal to the number of input haplotypes, then the results are close to the optimal results which can be produced by pure parsimony.

4.1. Real datases

4.1.1. AR gene data

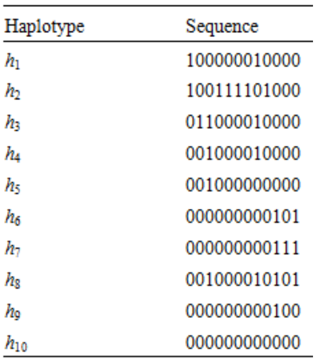
β$_2$-adrenergic receptors (β$_2$AR) are G protein-coupled receptors that mediate the actions of catecholamine in multiple issues. 13 variable sites within a span of 1.6kb were reported in the human β$_2$AR gene. Among 121 individuals, there are 18 distinct genotypes and 10 haplotypes resolve all the 18 genotypes. 10 haplotypes and 18 genotypes are illustrated in Tables 1 and 2 [10].

Implementation of the exact algorithm HAPAR [10] for the HIPP problem shows that the minimum number of haplotypes needed to resolve the 18 genotypes is 10 and the haplotype set output by HAPAR is exactly the original one. It means that the error rate of HAPAR is 0, but it takes a very long time (about 150 seconds) to run the algorithm. GAHAP also returns the same set of haplotypes like HAPAR in several seconds as they mentioned in their paper [13]. In [14] the authors ran PTG on β$_2$AR gene data 100 times; in 80 runs, they found 10 distinct haplotypes to resolve all 18 genotypes, where 9 of the 10 haplotypes correctly resolve 17 genotypes. The average error rate in 100 runs was 0.056. In particular, in 10 of 100 runs, they found all 10 correct haplotypes to resolve all 18 genotypes. The average running time was ~ 0.016 s. The number of haplotypes which is found by our greedy algorithm is exactly 10, the error rate is 0 and the running time is very small (about 0.019 seconds). The results of the algorithms are shown in Table 3.

4.1.2. Angiotensin converting enzyme (ACE) gene data

Angiotensin converting enzyme (ACE) is encoded by the gene DCP1. Information about genomic sequencing of DCP1 from 11 individuals in 22 chromosomes are considered [20]. There are 52 SNP sites and 11 genotypes, which are resolved by 13 distinct haplotypes. Two genotypes from dataset have the same sequences. Thus only 10 different genotypes exist as input data. 10 genotypes and 13 haplotypes are illustrated in Tables 4 and 5. The results of the error rate

Table 1. Ten haplotypes of β$_2$AR genes.



Haplotype	Sequence
h_1	100000010000
h_2	100111101000
h_3	011000010000
h_4	001000010000
h_5	001000000000
h_6	000000000101
h_7	000000000111
h_8	001000010101
h_9	000000000100
h_{10}	000000000000

of different algorithms are shown in Table 6. The error rate of PTG and HAPLOTYPER are smaller than other algorithms. But the number of output haplotypes which are produced by greedy algorithm is exactly 13 which is equal to the number of original haplotypes. This shows that the greedy algorithm does not produce extra haplotypes for this dataset. Thus, the results obtained by the greedy algorithm satisfy the purpose of parsimony approach. The number of haplotypes for other methods are not available.

4.1.3. GHI gene promoter data

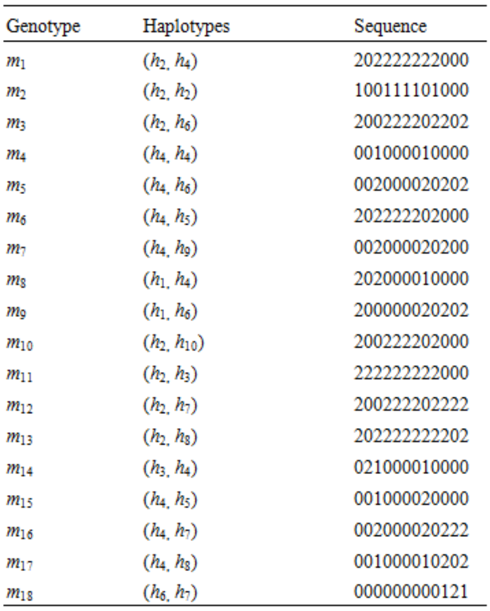
We use the empirical dataset of Horan [21] which contains 16 SNP sites and 308 genotypes. This dataset is also used by Adkins [22]. It contains 27 original haplotypes and 54 genotypes. The error rate of greedy algorithm on this dataset is equal to 0.18. The greedy algorithm found 21 haplotypes as the output. Thus, the number of output haplotypes is less than the number of original haplotypes and it is closer to the pure parsimony purpose. These data are not used by other algorithms and we applied the data to show the efficiency of our greedy algorithm.

4.2. Random datasets

4.2.1. Maize dataset

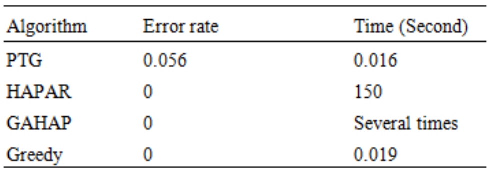
The Maize data were used in [10, 23] is one of the benchmarks to evaluate haplotyping programs. The locus 14 of maize profile containing 18 SNP sites and 4 different haplotypes (with frequencies 9 (27%), 17 (47%), 8 (23%) and 1 (3%)) were identified. We randomly generated four samples as input data. Each sample contains some genotypes that were constructed by randomly picking 2 haplotypes according to their frequencies and conflating them as shown in Table 7. The simulation results for different algorithms are shown in Table 8. These results suggest that the error rate of the greedy algorithm is smaller than or equal to that of other algorithms for all Maize dataset.

Table 2. Eighteen genotypes of β_2AR genes.



Genotype	Haplotypes	Sequence
m_1	(h_2, h_4)	202222222000
m_2	(h_2, h_2)	100111101000
m_3	(h_2, h_6)	200222202202
m_4	(h_4, h_4)	001000010000
m_5	(h_4, h_6)	002000020202
m_6	(h_4, h_5)	202222202000
m_7	(h_4, h_9)	002000020200
m_8	(h_1, h_4)	202000010000
m_9	(h_1, h_6)	200000020202
m_{10}	(h_2, h_{10})	200222202000
m_{11}	(h_2, h_3)	222222222000
m_{12}	(h_2, h_7)	200222202222
m_{13}	(h_2, h_8)	202222222202
m_{14}	(h_3, h_4)	021000010000
m_{15}	(h_4, h_5)	001000020000
m_{16}	(h_4, h_7)	002000020222
m_{17}	(h_4, h_8)	001000010202
m_{18}	(h_6, h_7)	000000000121

Table 3. Error rate of β_2AR dataset.



Algorithm	Error rate	Time (Second)
PTG	0.056	0.016
HAPAR	0	150
GAHAP	0	Several times
Greedy	0	0.019

4.2.2. Simulated dataset

In this section simulated data are used. A well-known haplotype generator, ms, in [24] is applied for generating genotype matrix as input for HIPP problem. This program is based on the coalescent model. In that model we assume each genotype is determined exactly by a pair of haplotypes and the rate of recombination is zero. Software ms generates 2×m haplotypes of the same length n (n is the number of SNP sites) and then randomly pairs them to obtain \$m\$ genotypes. These m genotypes are used as an input data for HIPP problem. In this section we generate 4 model samples based on ms program. The number of SNP sites is fixed and the number of genotypes varies. Results of comparisons of the efficiency of our algorithm with other methods are shown in Table 9. As shown in table 9, the results obtained by the greedy algorithm is better than other algorithms except for one sample size in which the HAPLOTYPER error rate is better than the greedy error rate.

5. Conclusion

In this paper, we proposed a new heuristic greedy algorithm for the HIPP which is an NP-hard problem [11]. The results show that the proposed algorithm works efficiently compared to other algorithms. One of the most important features of our greedy algorithm is its ability to predict an efficient haplotype for inferring the remaining genotypes in each step. The information for predicting an efficient haplotype is calculated using the remaining genotypes and the haplotypes which are already added to the set of the haplotypes. Since we use pure parsimony approach, a new measure is introduced for evaluating the effectiveness of algorithms. This new measure is the number of haplotypes that the algorithm returns as the output. If this number is less than or equal to the number of original haplotypes, the results are very close to the results of pure parsimony. The running time of the algorithm is O(m3n) where m is the number of given genotypes and n is the number of SNP sites.

Acknowledgments

This section should come before the References. Funding information may also be included here.

Table 4. Thirteen haplotypes of ACE genes.

Haplotype	Sequence
h_1	0001000010010000000000000000000010000010000000101100
h_2	0000000100000001100000000000000000000000101000010000
h_3	0000000100000001100000000000000000000000111000010000
h_4	0110100001010000000000000000000010000010000000101100
h_5	0000010000000000010100000000001000000000000010101101
h_6	0000000000001101100000000000000000000000000000000001
h_7	0000100000000010011111011011111101111101010111000000
h_8	0000100000000010011111011111111101111101000110000110
h_9	0001000010100010011111111011111101111101000110000001
h_{10}	0000010000001100000000000000000010000010000000101100
h_{11}	0110100001010010001011011011101111111111010101000000
h_{12}	1100101000000010011111011011111101111101010111000011
h_{13}	1100101000000010011111011111111101111101010111000000

Table 5. Ten genotypes of ACE genes.

Genotype	Haplotypes	Sequence
m_1	(h_4, h_9)	0222200022220020022222222022222222222222000220202202
m_2	(h_4, h_4)	0110100001010000000000000000000010000010000000101100
m_3	(h_1, h_9)	0001000010220020022222222022222222222222000220202202
m_4	(h_9, h_{10})	0002020020202220022222222022222222222222000220202202
m_5	(h_1, h_4)	0222200022010000000000000000000010000010000000101100
m_6	(h_2, h_3)	0000000100000001100000000000000000000000121000010000
m_7	(h_{12},h_{13})	1100101000000010011111011211111101111101010111000022
m_8	(h_5, h_6)	0000020000002202220200000000002000000000000020202201
m_9	(h_5, h_{11})	0220220002020020022222022022202222222222020222202202
m_{10}	(h_7, h_8)	0000100000000010011111011211111101111101020112000220

Table 6. Error rate of ACE dataset.

Algorithm	Error rate
PTG	0.182
HAPAR	0.273
HAPLOTYPER	0.182
HAPINFERX	0.273
PHASE	0.273
Greedy	0.273

Table 7. Haplotypes for Maize dataset.

Haplotype	Sequence	Frequency
h_1	011001001100100101	0.03
h_2	000000000000000000	0.47
h_3	000010001000000000	0.23
h_4	101101110111011010	0.27

Table 8. Error rates of Maize dataset.

Sample size	3	4	7	10
PTG	0.02	0	0	0
HAPAR	0.51	0.10	0.05	0
HAPLOTYPER	0.47	0.14	0.05	0
HAPINFERX	0.86	0.64	0.43	0.28
PHASE	0.53	0.15	0.07	0
Greedy	0.02	0	0	0

Table 9. Error rates of simulated dataset.

Sample size	10	15	20	25
PTG	0.12	0.10	0.05	0.03
HAPAR	0.18	0.10	0.08	0.05
HAPLOTYPER	0.20	0.05	0.03	0.02
HAPINFERX	0.80	0.60	0.38	0.31
PHASE	0.28	0.18	0.05	0.03
Greedy	0.11	0.09	0.03	0.02

References

1. Terwilliger, Joseph D., and Kenneth M. Weiss. "Linkage disequilibrium mapping of complex disease: fantasy or reality?." Current Opinion in Biotechnology 9.6 (1998): 578-594.
2. Hoehe, Margret R., et al. "Sequence variability and candidate gene analysis in complex disease: association of μ opioid receptor gene variation with substance dependence." Human molecular genetics 9.19 (2000): 2895-2908.
3. Clark, Andrew G., et al. "Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase." The American Journal of Human Genetics 63.2 (1998): 595-612.
4. Clark, Andrew G. "Inference of haplotypes from PCR-amplified samples of diploid populations." Molecular biology and evolution 7.2 (1990): 111-122.
5. Gusfield, D., and S. Orzach. "Handbook on Computational Molecular Biology, volume 9 of Chapman and Hall/CRC Computer and Information Science Series, chapter Haplotype Inference." (2005).
6. Stephens, Matthew, Nicholas J. Smith, and Peter Donnelly. "A new

statistical method for haplotype reconstruction from population data." The American Journal of Human Genetics 68.4 (2001): 978-989.

7. Gusfield, Dan. "Haplotype inference by pure parsimony." Annual Symposium on Combinatorial Pattern Matching. Springer Berlin Heidelberg, 2003.
8. Qian, Dajun, and Lars Beckmann. "Minimum-recombinant haplotyping in pedigrees." The American Journal of Human Genetics 70.6 (2002): 1434-1445.
9. O'Connell, Jeffrey R. "Zero-recombinant haplotyping: Applications to fine mapping using SNPs." Genetic Epidemiology 19.S1 (2000).
10. Wang, Lusheng, and Ying Xu. "Haplotype inference by maximum parsimony." Bioinformatics 19.14 (2003): 1773-1780.
11. Lancia, Giuseppe, Maria Cristina Pinotti, and Romeo Rizzi. "Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms." INFORMS Journal on computing 16.4 (2004): 348-359.
12. Brown, Daniel G., and Ian M. Harrower. "A new integer programming formulation for the pure parsimony problem in haplotype analysis." International Workshop on Algorithms in Bioinformatics. Springer Berlin Heidelberg, 2004.
13. Wang, Rui-Sheng, Xiang-Sun Zhang, and Li Sheng. "Haplotype inference by pure parsimony via genetic algorithm." Operations Research and Its Applications: the Fifth International Symposium (ISORA'05), Tibet, China, August. 2005.
14. Li, Zhenping, et al. "A parsimonious tree-grow method for haplotype inference." Bioinformatics 21.17 (2005): 3475-3481.
15. Lynce, Inês, and João Marques-Silva. "Efficient haplotype inference with Boolean satisfiability." Proceedings of the National Conference on Artificial Intelligence. Vol. 21. No. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
16. Graça, Ana, et al. "Efficient haplotype inference with combined CP and OR techniques." International Conference on Integration of Artificial Intelligence (AI) and Operations Research (OR) Techniques in Constraint Programming. Springer Berlin Heidelberg, 2008.
17. Graça, Ana, et al. "Efficient haplotype inference with pseudo-Boolean optimization." International Conference on Algebraic Biology. Springer Berlin Heidelberg, 2007.
18. Irurozki, Ekhine, Borja Calvo, and Jose A. Lozano. "A preprocessing procedure for haplotype inference by pure parsimony." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 8.5 (2011): 1183-1195.
19. Niu, Tianhua, et al. "Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms." The American Journal of Human Genetics 70.1 (2002): 157-169.
20. Rieder, Mark J., et al. "Sequence variation in the human angiotensin converting enzyme." Nature genetics 22.1 (1999): 59-62.
21. Horan, Martin, et al. "Human growth hormone 1 (GH1) gene expression: Complex haplotype-dependent influence of polymorphic variation in the proximal promoter and locus control region." Human mutation 21.4 (2003): 408-423.
22. Adkins, Ronald M. "Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset." BMC genetics 5.1 (2004): 22.
23. Ching, A. D. A., et al. "SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines." BMC genetics 3.1 (2002): 19.
24. Hudson, Richard R. "Generating samples under a Wright–Fisher neutral model of genetic variation." Bioinformatics 18.2 (2002): 337-338.