A NOTE ON SOLUTION SIZES IN THE HAPLOTYPE TAGGING SNPS PROBLEM

Staal Vinterbo
Decision Systems Group, Brigham and Women's Hospital
Harvard Medical School
Boston, MA, 02115, USA
E-mail: staal@dsg.harvard.edu

Stephan Dreiseitl
Upper Austria University of Applied Sciences
Dept. of Software Engineering
A-4232 Hagenberg, Austria
E-mail: sdreisei@fh-hagenberg.at

KEYWORDS

Haplotype tagging, single nucleotide polymorphism, minimal hitting set, combinatorial optimization

ABSTRACT

The problem of finding a minimal set of single nucleotide polymorphisms that can distinguish between given haplotypes is known to be NP-hard. In this paper, we summarize the properties of a simple polynomial-time approximation algorithm. We investigate which regions of the solution space contain a phase transition, i.e., the range of solution sizes which are neither trivially easy nor impossibly hard to solve. We give empirical results showing that the phase transition region is an interval of length $O(\log(n))$ centered around $O(\log(n))$.

INTRODUCTION

With the conclusion of the Human Genome Project in 2003, it is now possible to analyze the genetic blueprint of *homo sapiens*. One of the surprising findings was that at the genetic level, humans are much more alike than previously imagined: out of the about 3 billion base pairs in the human genome, only about 1 in 1200 is different for two humans. This makes the individual genomic makeup of humans 99.9% identical. The interesting information—which makes us unique as individuals—is contained in the remaining 0.1%, the *single nucleotide polymorphisms (SNPs)*. Being able to identify SNPs is of paramount importance for personalized medicine, because the effect of a medication targeted at individuals is directly dependent on the unique genetic makeup of these individuals. SNPs also play an important role in developing an understanding of the genetic causes of disease (Carlson et al., 2004): they allow the measurement of *linkage disequilibrium*, a term that refers to the fact that the combination of alleles at SNP locations sometimes occurs more often than would be expected in case they were independent of one another (Reich et al., 2001; Goldstein and Weale, 2001).

Figure 1: Simplified example showing four haplotypes comprising the alleles at eight SNP locations. The arrows indicate possible htSNP locations.

The correlation of SNPs in close proximity can be attributed to the fact that these SNPs tend to be inherited together. Such a collection of correlated SNPs is known as a *haplotype*, and research efforts undertaken by the *HapMap project* aim to establish an understanding of the haplotypes of human beings. Since all information about a person's individual genetic composition is contained in the SNPs, and thus in the haplotypes, it is advantageous to find minimal combinations of SNP locations that allow the unique identification of a haplotype. Such combinations of SNPs are called *haplotype tagging SNPs (htSNPs)*. An schematic example of htSNPs is shown in Figure . In this example, one can see that while no single SNP location (column in the data table) can uniquely identify every haplotype (row in the data table), two such locations are sufficient to do so. The arrows indicate only one choice of htSNPs combination; others, such as the first and second column, are also possible.

In this paper, we will build upon previous results (Vinterbo et al., 2004, 2006) to investigate how many htSNPs are required to uniquely identify every haplotype in a given collection. We will provide empirical evidence about the structure of the "phase transition" between over- and underconstrained problem instances. These results can be used to better understand the computational complexity of finding htSNPs in the average case. Theoretical results present only worst-case analyses, which may not be representative of genuine problem instances encountered in real-world settings.

PREVIOUS WORK

The problem of determining a minimum-cardinality set of htSNPs for a given collection of SNPs is known to be NP-hard, via the association with the minimum hitting set and minimum set cover problem. Nevertheless, it is possible to give an easily implementable polynomial-time approximation algorithm that is asymptotically tight, and which shows that the problem is approximable with $1 + \ln((n^2 - n)/2)$ for n haplotypes (Vinterbo et al., 2006). Other algorithms consider the problem of haplotype tagging via Boolean algebra (Wiuf et al., 2003); that paper and the work by Bafna et al. (Bafna et al., 2003) also give polynomial-time algorithms, but only for restricted problem statements. Special search strategies and algorithms for finding htSNPs are presented in (Ke and Cardon, 2003; Sebastiani et al., 2003; Hao et al., 2005); a stochastic algorithm is given in (Johnson et al., 2001). Principal component analysis is used for htSNP identification in (Liu and Altman, 2004; Horne and Camp, 2004). Comparisons of algorithms are given in (Burkett et al., 2005; Ke et al., 2005; Stram, 2005).

PROBLEM STATEMENT

In this section, we give a concise statement of the problem of selecting a minimal set of SNPs that allow the unique identification of given haplotypes.

Consider given an $n \times m$ data matrix M, i.e., n haplotypes each containing m SNPs. Without loss of generality, we assume that there are no duplicate rows in M. Given M, the objective is to find a minimal cardinality set H of column indices with the property

$$\forall\, i, j \in \{1, \ldots, n\} \left(\exists\, k \in \{1, \ldots, m\} \right. \\ \left. M_{ik} \neq M_{jk} \Rightarrow\ \exists k' \in H\, M_{ik'} \neq M_{jk'}\right).$$

This means that whenever two rows in M differ, they differ at least in the column indices in H. The desired H is the smallest index set that allows to distinguish between any two rows in the data table, and thus to uniquely identify all rows, given only the entries in the columns in H.

A simple greedy algorithm to select column indices for H is as follows: Let d be the function that returns all pairs of row indices that are distinguishable by a given column, i.e.,

$$d(i) := \{(j, k) \mid M_{ji} \neq M_{ki}\}.$$

The greedy search algorithm given below iteratively chooses the column index that allows the most rows to be distinguished.

$H \leftarrow \emptyset$
cols $\leftarrow \{1, \ldots, m\}$
$D \leftarrow d(\text{cols})$
while $D \neq \emptyset$
  select $c \in$ cols that maximizes $|D \cap d(c)|$
  $D \leftarrow D \setminus d(c)$
  $H \leftarrow H \cup \{c\}$
return H

It is known that this algorithm approximates the optimal solution to within $1 + \ln((n^2 - n)/2)$ (Johnson, 1974). A worst-case run-time analysis is given in (Vinterbo et al., 2006): the run-time complexity of this algorithm is $O(m(n^2 - n)/2)$.

EMPIRICAL RESULTS

While the theoretical results presented above give bounds on run time and approximation properties of the greedy htSNP calculation algorithm given above, they provide little understanding of the average-case situation. To this end, we considered various aspects of solution size properties, and obtained simulation results for randomly generated problem instances.

The motivation for this can be found in recent work investigating the "phase transitions" found in combinatorial optimization problems (Cheeseman et al., 1991; Cook and Mitchell, 1997; Martin et al., 2001). For a simple explanation, consider as parameter the size $r = |H|$ of a column index set H in the htSNP optimization problem. If r is too small, it may be simple to see that H cannot possibly be a solution; similar arguments could show that a large r cannot be the size of the *smallest* solution. The interesting cases can thus be found in the intermediate region, where the problem undergoes a so-called phase transition. In this work, we use simulation experiments to locate the phase transition region of the htSNP problem.

For the following, let r be the size of a potential solution to the htSNP problem. We can then give bounds for the probability that there is a solution of size r, if we assume that the haplotype entries have a uniform random distribution in some set $\{0, \ldots, l-1\}$. This is not an exact model of the true situation, where SNPs in haplotypes are correlated, but a reasonable approximation for the sake of the following analysis. Details and proofs for the derivations below can be found in (Vinterbo et al., 2004).

For a given $n \times M$ data matrix M, let $X_H(i, j)$ denote the random variable that counts at how many positions in the column index set H the two rows i and j differ, i.e.,

$$X_H(i, j) = \sum_{k \in H} \mathbb{1}[M_{ik} \neq M_{jk}],$$

with $\mathbb{1}[\cdot]$ being the Boolean indicator function. Since the probability that any two entries in M are the same is $\frac{1}{l}$, we have that the probability that H can distinguish between rows i and j in M is

$$P(X_H(i, j) \geq 1) = 1 - \left(\frac{1}{l}\right)^{|H|}.$$

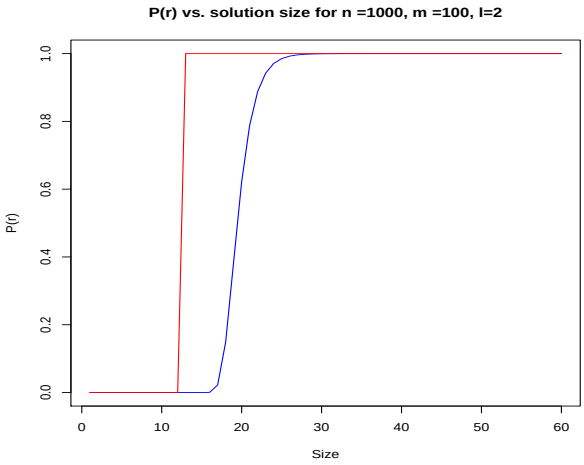

Figure 2: Bounds on $P(r)$ versus solution size.

Note that since all entries in M are independently generated, the above probability does not depend on the choice of i and j. We denote the probability that two rows in M can be distinguished by columns in H as

$$p_{\text{hit}}(|H|) = P(X_H(i,j) \geq 1).$$

Extending the analysis to *all* rows of M, we can easily see that the probability $p_{\text{tag}}(|H|)$ that all $(n^2-n)/2$ combinations of rows can be distinguished by columns in H is

$$p_{\text{tag}}(|H|) = p_{\text{hit}}(|H|)^{(n^2-n)/2}.$$

We can now use $p_{\text{tag}}(r)$ to bound the probability $P(r)$ of there existing an htSNP solution of size r from below:

$$p_{\text{tag}}(r) \leq P(r).$$

This bound holds because $p_{\text{tag}}(r)$ is the probability for a randomly chosen H of size r, and $P(r)$ denotes the probability over *all* such sets.

Similarly, we can bound $P(r)$ from above by noting that there exists an r-cardinality solution if at least one of the $b = \binom{m}{r}$ choices for H is a solution. If we use the Boolean random variable E_k to denote that the k-th choice of H is a solution, then

$$P(r) = P(E_1 \vee \cdots \vee E_b) \leq \sum_{i=1}^{b} P(E_k)$$

by Boole's inequality. It thus follows that

$$p_{\text{tag}}(r) \leq P(r) \leq \binom{m}{r} p_{\text{tag}}(r).$$

To get some understanding of these bounds, we ran some simulations using the parameter values of $l = 2$, $n = 1000$, and $m = 100$. Figure 2 shows the lower bound $p_{\text{tag}}(r)$ and upper bound $\min(\binom{m}{r}p_{\text{tag}}(r), 1)$ for $P(r)$. This figure shows the regions of likely interesting solution sizes, in this case the region of about $15 \leq r \leq 30$.

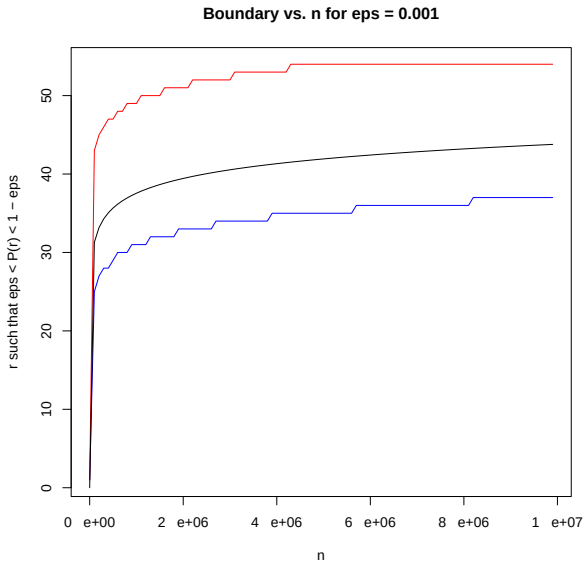

Figure 3: Upper and lower bounds for r such that $0.001 < P(r) < 0.999$ versus instance size n. The middle line is the function $e\log(n)$. The maximum gap between the upper and lower bounds is 19.

For a given probability bound $0 < \epsilon < 1$ we can determine the minimum number m of SNPs required to tag n haplotypes with probability at least $1 - \epsilon$:

$$p_{\text{tag}}(m) = \left(1 - \left(\frac{1}{l^m}\right)\right)^{(n^2-n)/2} \geq 1 - \epsilon,$$

which leads to the lower bound of

$$\left\lceil -\log_2\left(1 - 2^{\left(\frac{2\log_2(1-\epsilon)}{n^2-n}\right)}\right)\right\rceil \leq m.$$

For fixed ϵ, this bound grows very slowly in n. For $n = 1000$ and $\epsilon = 0.001$ the value is 29, which agrees with Figure 2.

Figure 3 contains a plot of the values of r that bound $P(r)$ from below and above, respectively, as functions of n. The middle line is the function $e\log(n)$. In fact, as n grows we can envelop the upper and lower bounds on r as $s\log(n)$ and $t\log(n)$ meaning that the r gap grows approximately as $(s-t)\log(n)$ around $(s+t)/2\log(n)$. Specifically, for the plot in Figure 3 we can choose $s = 3.64$ and $t = 2.18$, which then results in $s - i = 1.46$.

For $l > 2$ we have very similar behavior. As l increases, the length of the r "phase transition" interval decreases. A plot of $a = s - t$ vs. l, where as before, $s\log(n)$ and $i\log(n)$ are the upper and lower envelopes of the r-interval respectively, is given in Figure 4.

DISCUSSION

The considerations and observations outlined in the previous section lead us to the following conjecture: For the

Figure 4: $s - t$ (the factor of the envelope width $(s - t)\log(n)$) as a function of l.

decision problem of whether a uniformly sampled $n \times m$ data matrix has an htSNP solution set of size r, we have a "phase transition" interval for r of $O(\log(n))$ length centered around $O(\log(n))$. This interval length decreases exponentially as l increases.

If we assume that this conjecture is true, then according to the assumption that under- and over-constrained instances are easier to solve, we have that overall instance difficulty decreases as the range of the features increases. However, our analysis is not sufficient to make statements about relative hardness of boundary instances.

From the point of view of molecular diagnosis and the promise it holds for personalized medicine, the algorithm and analysis in this paper represent a contribution to the important problem of identifying haplotype tagging SNPs. Polymorphisms that occur within gene coding regions of the genome can be used for association studies, i.e., to check which diseases are associated with which haplotypes. Due to the high costs involved in determining SNP values, it is advantageous to restrict attention to those SNPs that allow a unique identification of a person's haplotype. The results in this paper show how the locations of these haplotype tagging SNPs can be calculated efficiently.

Note also that the applicability of the considerations given in this paper is not limited to the htSNP problem, but can be applied in any problem domain where a set of data vectors should be distinguised using the smallest number of features. It is also easy to extend the analysis to consider the case of weighted features, i.e., when the relative importance of features/SNPs is different. More details on these considerations are given in (Vinterbo et al., 2006).

CONCLUSION

We presented an easy to implement greedy algorithm for determining the smallest number of SNPs required to uniquely tag a set of haplotypes. We outlined the approximation properties of this algorithm, and investigated the question of which problems may be under- or overdetermined.

REFERENCES

Bafna, V., Halldorsson, B. V., Schwartz, R., Clark, A. G., and Istrail, S. (2003). Haplotypes and informative SNP selection algorithms: don't block out information. In *Proceedings of the seventh annual international conference on Computational molecular biology*, pages 19–27. ACM Press.

Burkett, K., Ghadessi, M., McNeney, B., Graham, J., and Daley, D. (2005). A comparison of five methods for selecting tagging single-nucleotide polymorphisms. *BMC Genetics*, 6 Suppl 1(S71).

Carlson, C., Eberle, M., Kruglyak, L., and Nickerson, D. (2004). Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–452.

Cheeseman, P., Kanefsky, B., and Taylor, W. M. (1991). Where the Really Hard Problems Are. In *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, pages 331–337.

Cook, S. A. and Mitchell, D. G. (1997). Finding hard instances of the satisfiability problem: A survey. In Du, Gu, and Pardalos, editors, *Satisfiability Problem: Theory and Applications*, volume 35, pages 1–17. American Mathematical Society.

Goldstein, D. and Weale, M. (2001). Population genomics: Linkage disequilibrium holds the key. *Current Biology*, 11:R576–R579.

Hao, K., Liu, S., and Niu, T. (2005). A sparse marker extension tree algorithm for selecting the best set of haplotype tagging single nucleotide polymorphisms. *Genetic Epidemiology*, 29(4):336–352.

Horne, B. and Camp, N. (2004). Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. *Genetic Epidemiology*, 26(1):11–21.

Johnson, D. S. (1974). Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9:256–278.

Johnson, G., Esposito, L., Barratt, B., Smith, A., Heward, J., Genova, G. D., Ueda, H., Cordell, H., Eaves, I., Dudbridge, F., Twells, R., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S., Clayton, D., and Todd, J. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics*, 29(2):233–237.

Ke, X. and Cardon, L. (2003). Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, 19(2):287–288.

Ke, X., Miretti, M., Broxholme, J., Hunt, S., Beck, S., Bentley, D., Deloukas, P., and Cardon, L. (2005). A comparison of tagging methods and their tagging space. *Human Molecular Genetics*, 15(18):2757–2767.

Liu, Z. and Altman, R. (2004). Finding haplotype tagging SNPs by use of principal components analysis. *American Journal of Human Genetics*, 75(5):850–861.

Martin, O. C., Monasson, R., and Zecchina, R. (2001). Statistical mechanics methods and phase transitions in optimization problems. *Theoretical Computer Science*, 265(1–2):3–67.

Reich, D., Cargill, M., Bolk, S., Ireland, J., Sabeti, P., Richter, D., Lavery, T., Kouyoumijian, R., Farhadian, S., Ward, R., and Lander, E. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Sebastiani, P., Lazarus, R., Weiss, S., Kunkel, L., Kohane, I., and Ramoni, M. (2003). Minimal haplotype tagging. *Proceedings of the National Academy of Sciences*, 100(17):9900–9905.

Stram, D. (2005). Software for tag single nucleotide polymorphism selection. *Human Genomics*, 2(2):144–151.

Vinterbo, S., Dreiseitl, S., and Ohno-Machado, L. (2004). A testing procedure for htSNP approximation algorithms. In *Proceedings of the Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP 2004)*, pages 101–105.

Vinterbo, S., Dreiseitl, S., and Ohno-Machado, L. (2006). Approximation properties of haplotype tagging. *BMC Bioinformatics*, 7(8).

Wiuf, C., Laidlaw, Z., and Stumpf, M. (2003). Some notes on the combinatorial properties of haplotype tagging. *Mathematical Biosciences*, 185:205–216.

AUTHOR BIOGRAPHIES

STAAL A. VINTERBO received the Ph.D. degree in computer science from the Norwegian University of Science and Technology (2000). He is a research scientist at the Decision Systems Group, Brigham and Womens Hospital in Boston. He is also an assistant professor at Harvard Medical School, and is a member of the affiliated faculty at the Harvard-MIT Division of Health Sciences and Technology. His research interests include machine learning, knowledge discovery, and computational disclosure control.

STEPHAN DREISEITL received the M.Sc. and Ph.D. degrees in mathematics from the University of Linz, Austria, in 1993 and 1997, respectively. He is currently a professor of mathematics at the Upper Austria University of Applied Sciences in Hagenberg, Austria. His research interests include pattern recognition and machine learning as well as the development of medical decision support systems.