Template Matrices for Perfect Phylogeny Haplotyping and Site Consistency

Tetsuo Asano, [1] Francesc Rosselló, [2] Gabriel Valiente [3]

Keywords: haplotype, genotype, perfect phylogeny, site consistency, phylogenetic network with recombination, galled-tree, conflict graph, boolean matrix multiplication

1 Introduction.

The problem of inferring haplotype phase from a population of genotypes has received much attention recently, especially on the light of current large-scale efforts to characterize populations in terms of hyplotypes. The site consistency problem in perfect phylogeny has also received much attention lately, motivated in part by the characterization of a particular form of phylogenetic networks under mutation and constrained recombination that admit a polynomial-time solution to the site consistency problem.

Perfect phylogeny haplotyping is the problem of resolving a given set of n diploid sequences into a set of $2n$ haploid sequences that form a perfect phylogeny [9]. The bottleneck of current algorithms for the perfect phylogeny haplotyping problem [3, 5] lies in the efficient computation of the genotype graph [2], and the usual algorithm for computing the genotype graph for a given $n \times m$ genotype matrix takes $O(nm^2)$ time. However, the genotype graph for a given genotype matrix can be obtained by a simple combination [2, Lemma 2] of eight template matrices over the alphabet $\Sigma = \{0, 1, 2\}$.

The notion of template matrix was introduced in [2] and, independently, in the proof of [1, Lemma 11].

Definition 1 *The template matrix derived from a matrix M over an alphabet Σ for the template ab, where $a, b \in \Sigma$, is the boolean matrix $\mathcal{M}_{ab}$ with $\mathcal{M}_{ab}[j, k] = 1$ if and only if there exists some row i such that $M[i, j] = a$ and $M[i, k] = b$.*

Site consistency in perfect phylogeny is the problem of resolving a given genomic matrix into another one that admits a perfect phylogeny, by removing the least possible number of sites (columns). While the site consistency problem is NP-hard in phylogenetic networks under mutation alone [4], it becomes polynomial-time solvable in a particular form of phylogenetic networks under mutation and constrained recombination, called galled-trees [6, 8], because of the additional structure of the conflict graph [7]. Thus, the bottleneck of current algorithms for perfect phylogeny under mutation and constrained recombination lies in the efficient computation of the conflict graph, and the usual algorithm for computing the conflict graph for a given $n \times m$ genomic matrix takes $O(nm^2)$ time. However, the conflict graph for a given genomic matrix can be obtained by a simple combination of three template matrices over the alphabet $\Sigma = \{0, 1\}$.

[1] Japan Advanced Institute of Science and Technology, School of Information Science, Asahidai 1-1, Nomi, Ishikawa 923-1292, Japan, E-mail: `t-asano@jaist.ac.jp`

[2] Department of Mathematics and Computer Science, Research Institute of Health Science, University of the Balearic Islands, E-07122 Palma de Mallorca, E-mail: `cesc.rossello@uib.es`

[3] Algorithms, Bioinformatics, Complexity and Formal Methods Research Group, Technical University of Catalonia, E-08034 Barcelona, E-mail: `valiente@lsi.upc.edu`

2 Results and Discussion.

In this poster, we show that the efficient computation of template matrices can be done by compressing the columns of the input genotype or genomic matrix.

Theorem 1 *Let M be an $n \times m$ matrix over an alphabet Σ. Then, for any $a, b \in \Sigma$, the set P_{ab} of pairs (j, k) for which there exists a row i such that $M[i, j] = a$ and $M[i, k] = b$, can be enumerated in $O(nm^2/\log n)$ time.*

The previous result has a simple constructive proof. The algorithm itself is not complicated to implement, and it only requires the use of standard data structures.

Corollary 1 *Let M be an $n \times m$ genotype matrix over the alphabet $\Sigma = \{0, 1, 2\}$. Then, the genotype graph for M can be computed in $O(nm^2/\log n)$ time.*

Corollary 2 *Let M be an $n \times m$ genomic matrix over the alphabet $\Sigma = \{0, 1\}$. Then, the conflict graph for M can be computed in $O(nm^2/\log n)$ time.*

Notice that the efficient computation of template matrices entails the fast multiplication of boolean matrices. The proof of the following result is similar to [2, Lemma 5].

Lemma 1 *Let $T(n, m)$ be the time needed to construct a template matrix $\mathcal{M}_{ab}$ for any $a, b \in \Sigma$, given an $n \times m$ matrix M over Σ. Two boolean matrices with dimensions $m_1 \times n$ and $n \times m_2$, respectively, can be multiplied in $O(T(n, m_1 + m_2))$ time.*

In summary, we present an algorithm for computing template matrices for any matrix over an arbitrary alphabet in $O(nm^2/\log n)$ time. This provides an $O(nm^2/\log n)$ time solution to the perfect phylogeny haplotyping problem and to the perfect phylogeny problem under mutation and constrained recombination, as well as an $O(n^3/\log n)$ time solution to the boolean matrix multiplication problem.

References

[1] Asano, T., Evans, P., Uehara, R., Valiente, G.: Site consistency in phylogenetic networks with recombination. In Iliopoulos, C.S., Park, K., Steinhöfel, K., eds.: Algorithmics in Bioinformatics. Volume 6 of Texts in Algorithmics. College Publications (2006) 15–26

[2] Bafna, V., Gusfield, D., Hannenhalli, S., Yooseph, S.: A note on efficient computation of haplotypes via perfect phylogeny. J. Comput. Biol. **11** (2004) 858–866

[3] Bafna, V., Gusfield, D., Lancia, G., Yooseph, S.: Haplotyping as perfect phylogeny: A direct approach. J. Comput. Biol. **10** (2003) 323–340

[4] Day, W.H.E., Sankoff, D.: Computational complexity of inferring phylogenies by compatibility. Syst. Zool. **35** (1986) 224–229

[5] Eskin, E., Halperin, E., Karp, R.M.: Large scale reconstruction of haplotypes from genotype data. In: Proc. 7th Annual Int. Conf. Research in Comput. Mol. Biol., ACM Press (2003) 104–113

[6] Gusfield, D., Eddhu, S., Langley, C.: Efficient reconstruction of phylogenetic networks with constrained recombination. In: Proc. 2003 IEEE Comput. Syst. Bioinformatics Conf. (2003) 1–12

[7] Gusfield, D., Eddhu, S., Langley, C.: The fine structure of galls in phylogenetic networks. INFORMS J. Comput. **16** (2004) 459–469

[8] Gusfield, D., Eddhu, S., Langley, C.: Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. J. Bioinformatics Comput. Biol. **2** (2004) 173–213

[9] Gusfield, D.: Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In: Proc. 6th Annual Int. Conf. Research in Comput. Mol. Biol., ACM Press (2002) 166–175