TOWARD AN ALGEBRAIC UNDERSTANDING OF HAPLOTYPE INFERENCE BY PURE PARSIMONY

Daniel G. Brown and Ian M. Harrower
David R. Cheriton School of Computer Science, University of Waterloo,
200 University Avenue W.,
Waterloo, Ontario, Canada N2L 3G1

Haplotype inference by pure parsimony (HIPP) is known to be NP-Hard. Despite this, many algorithms successfully solve HIPP instances on simulated and real data. In this paper, we explore the connection between algebraic rank and the HIPP problem, to help identify easy and hard instances of the problem. The rank of the input matrix is known to be a lower bound on the size an optimal HIPP solution. We show that this bound is almost surely tight for data generated by randomly pairing p haplotypes derived from a perfect phylogeny when the number of distinct population members is more than $\left(\frac{1+\varepsilon}{2}\right) p\log p$ (for some positive ε).

Moreover, with only a constant multiple more population members, and a common mutation, we can almost surely recover an optimal set of haplotypes in polynomial time. We examine the algebraic effect of allowing recombination, and bound the effect recombination has on rank. In the process, we prove a stronger version of the standard haplotype lower bound. We also give a complete classification of the rank of a haplotype matrix derived from a galled tree. This classification identifies a set of problem instances with recombination when the rank lower bound is also tight for the HIPP problem.