OPTIMAL IMPERFECT PHYLOGENY RECONSTRUCTION AND HAPLOTYPING (IPPH)

Srinath Sridhar
Computer Science Department, Carnegie Mellon University
Pittsburgh, PA 15213. USA.

Guy E. Blelloch
Computer Science Department, Carnegie Mellon University
Pittsburgh, PA 15213. USA.

R. Ravi
Tepper School of Business, Carnegie Mellon University
Pittsburgh, PA 15213. USA.

Russell Schwartz
Department of Biological Sciences, Carnegie Mellon University
Pittsburgh, PA 15213. USA.

The production of large quantities of diploid genotype data has created a need for computational methods for large scale inference of haplotypes from genotypes. One promising approach to the problem has been to infer possible phylogenies explaining the observed genotypes in terms of putative descendants of some common ancestral haplotype. The first attempts at this problem proceeded on the restrictive assumption that observed sequences could be explained by a perfect phylogeny, in which each variant locus is presumed to have mutated exactly once over the sampled population's history. Recently, the perfect phylogeny model was relaxed and the problem of reconstructing an imperfect phylogeny (IPPH) from genotype data was considered. A polynomial time algorithm was developed for the case when a single site is allowed to mutate twice, but the general problem remained open. In this work, we solve the general IPPH problem and show for the first time that it is possible to infer optimal q-near-perfect phylogenies from diploid genotype data in polynomial time for any constant q, where q is the number of \extra" mutations required in the phylogeny beyond what would be present in a perfect phylogeny. This work has application to the haplotype phasing problem as well as to various related problems in phylogenetic inference, analysis of sequence variability in populations, and association study design. Empirical studies on human data of known phase show this method to be competitive with the leading phasing methods and provide strong support for the value of continued research into algorithms for general phylogeny construction from diploid data.