Efficient and Practical Algorithms for Deducing the History of Recombination in Populations

(Keynote Talk Abstract)

Dan Gusfield

University of California, Davis
gusfield@cs.ucdavis.edu

A phylogenetic network (or Ancestral Recombination Graph) is a generalization of a tree, allowing structural properties that are not tree-like. With the growth of genomic and population data, much of which does not fit ideal tree models, and the increasing appreciation of the genomic role of such phenomena as recombination (crossing-over and gene-conversion), recurrent and back mutation, horizontal gene transfer, and mobile genetic elements, there is greater need to understand the algorithmics and combinatorics of phylogenetic networks.

In this talk I will survey a range of our recent results on phylogenetic networks with recombination, and show applications of these results to association mapping; finding recombination hotspots in genotype sequences; imputing the values of missing haplotype data; determining the extent of recombination in the evolution of LPL sequences; distinguishing the role of crossing-over from gene-conversion in Arabidopsis; and characterizing some aspects of the haplotypes produced by the program PHASE.

I will discuss the fundamental problem of constructing a phylogenetic network for a given set of binary (SNP) sequences derived from a known or unknown ancestral sequence, when each site in the sequence can mutate at most once in the network (the infinite sites model in population genetics), but recombination between sequences is allowed. The goal is to find a phylogenetic network that generates the given set of sequences, minimizing the number of recombination events used in the network. When all the “recombination cycles” are disjoint (which is likely with low recombination rates), we have developed efficient provably correct algorithms that find a network minimizing the number of recombinations, and have proven that the optimal solution is “essentially unique”. I will also mention a network decomposition theory that shows the extent that these results can be generalized to arbitrary networks. For general phylogenetic networks (when the cycles are not constrained) we have developed algorithms that are efficient in practice and that empirically obtain close upper and lower bounds on the number of recombinations needed. In real data and simulations, these practical (heuristic) computations often produce bounds that match, demonstrating that a minimum recombination solution has been found. In small-size data we can guarantee that an optimal solution will be found by running an exponential-time algorithm to completion. For small-size data, we can also (provably) sample uniformly from the set of optimal solutions, and can also determine the phase of genotypic data so as to minimize the number of recombinations needed to derive

the resulting haplotypes. More recently, we have extended the lower and upper bound algorithms to incorporate gene-conversion as an allowed operation. Those new algorithms allow us to investigate and distinguish the role of gene-conversion from single-crossover recombination.

Various parts of this work are joint work with Satish Eddhu, Chuck Langley, Dean Hickerson, Yun Song, Yufeng Wu, V. Bansal, and Zhihong Ding. Support for this work was provided by NSF grant IIS-0513910. All the papers and associated software can be accessed at [wwwcsif/cs.ucdavis.edu/~gusfield](http://wwwcsif.cs.ucdavis.edu/~gusfield).

References

1. Gusfield, D., Eddhu, S., Langley, C.: Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology* **2**(1) (2004) 173–213
2. Gusfield, D., Eddhu, S., Langley, C.: The fine structure of galls in phylogenetic networks. *INFORMS J. on Computing*, Special issue on Computational Biology **16**(4) (2004) 459–469
3. Gusfield, D.: Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained recombination. *J. Computer and Systems Sciences*, Special issue on Computational Biology **70** (2005) 381–398
4. Gusfield, D., Bansal, V.: A fundamental decomposition theory for phylogenetic networks and incompatible characters. In: *Proc. 9th Annual International Conference on Computational Biology*. Volume 3500 of *Lecture Notes in Bioinformatics*. (2005) 217–232
5. Gusfield, D., Hickerson, D., Eddhu, S.: A fundamental, efficiently computed lower bound on the number of recombinations needed in a phylogenetic history. *Discrete Applied Math*, Special issue on Computational Biology (to appear)
6. Song, Y., Wu, Y., Gusfield, D.: Efficient computation of close lower and upper bounds on the minimum number of needed recombinations in the evolution of biological sequences. *Bioinformatics* **21**(Supplement 1, Proceedings of the ISMB 2005 Conference) (2005) 413–422
7. Song, Y., Gusfield, D., Ding, Z., Langley, C., Wu, Y.: Algorithms to distinguish the role of gene-conversion from single-crossover recombination in the derivation of snp sequences in populations. In: *Proc. 10th Annual International Conference on Computational Biology*. Volume 3909 of *Lecture Notes in Bioinformatics*. (2006) 231–245 (to appear)