Srinath Sridhar

Computer Science Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

Cell: 412-320-6895
Email: `srinath@cs.cmu.edu`
Web: `http://www.cs.cmu.edu/~srinath`

Education

Ph.D. Computer Science, Carnegie Mellon University, December 2007
Thesis: Algorithms for Analyzing Intraspecific Sequence Variation

B. S. (Honors) Computer Science, University of Texas at Austin, 2003
Dean's Honored Graduate (Selected as top graduate); GPA: 4.0

Work Experience

Software Engineering Intern — Summer 2007
Google Inc — Mountain View, California
Project: Algorithms for improving google search result quality

Graduate Internship — Summer 2006
Advisor: Eran Halperin — International Computer Science Institute
Berkeley, California
Project: Design, implementation and testing of algorithms to detect population sub-structure using SNP data

Graduate Internship — Summer 2005
Advisors: Mike Wigler, Ira Hall, Jonathan Sebat — Cold Spring Harbor Labs
Cold Spring Harbor, New York
Project: Computational analysis of copy number polymorphisms (CNPs)

Teaching Experience

Teaching Assistant, Computational Methods for Biological Modeling and Simulation, Graduate Course, Spring 2005; taught by Russell Schwartz.

Teaching Assistant, Design and Analysis of Algorithms, Undergraduate Course, Fall 2006; co-taught by Manuel Blum and Avrim Blum.

Selected Software Packages and Patents

`http://www.cs.cmu.edu/~imperfect`
A webserver that can construct maximum parsimony phylogenies on SNP data and display (minimum) number of recurrent mutations that the most widely used human genome variation database (HapMap) has at any specific location provided by the user. The webserver works by solving several million problem instances off-line and compiling results at run-time. Algorithms were implemented in C++ using the `Cplex Concert` libraries. The software was self-developed.

`http://www.cs.cmu.edu/~triplets`
A webserver that clusters individuals based on SNP data into two sub-populations and returns the significance of the clustering. The primary application is to detect population sub-structure prior to conducting disease association testing. The algorithm was implemented in C++. The software was self-developed.

Google Patent Application, "RANKING SEARCH RESULTS", Inventors: S. Sridhar, etc., filed on 8/27/07.

Research Overview

Reconstruction of Evolutionary Trees: This project involved maximum parsimony phylogeny reconstruction specialized for single nucleotide polymorphism (SNP) data. For experimentation, we used data from HapMap, coalescent simulations and smaller published data-sets from varying sources.

Theory: We considered both the problem of maximum parsimony phylogeny reconstruction from haplotypes (phase known) and from genotypes (unphased). The second problem is significantly harder since it optimizes over many instances of the first. We developed two novel algorithms with running times that are currently the best for each problem. The algorithms were implemented in C++.

Applied: We developed practical Integer Linear Programming formulations of the same problems with extensive pre-processing techniques that guarantee optimality (maximum parsimony) while reducing the computational problem sizes. The algorithms were implemented in C++ using `Concert` library of CPLEX 10.0.

Practice: The algorithms were tested on both real and simulated data. The algorithms solved very large instances to optimality. The efficiency also enabled us to solve millions of smaller instances that cover the whole genome. Specifically, we computed piece-wise maximum parsimony phylogenies for all of $\approx$4 million SNPs produced by the International HapMap consortium. As an immediate application, we can provide the minimum number of recurrent mutations that any region has undergone (under no recombination assumption). A webserver interface for the software is also available.

Population Substructure: This project involved detection of population stratification using SNP data. No added information such as ancestry-specific allele frequency is needed. We used SNP genotypes from HapMap with some added simulations for experiments.

No Admixture: We assumed that the underlying sequences were produced under no-admixture. We designed and implemented an algorithm that has the theoretical guarantee of finding the correct clustering given enough SNPs. Experiments showed that the algorithm was significantly better than two prior methods (STRUCTURE and EIGENSTRAT) in either or both run-time and accuracy. Implementation in C++ with a webserver interface available.

Admixed Subpopulations: We designed and implemented algorithms to detect admixture. The algorithm was specifically designed to scale to large number of SNPs and skewed sizes of sub-populations. Experiments showed the algorithm to be significantly faster than other methods (STRUCTURE and SABER) while being very accurate. Implementation in C++.

Copy Number Polymorphisms(CNPs): For this project, the associated data was generated by comparative genome hybridization (array CGH) resulting in copy number variation observed as noisy fractional values. The data used was generated by collaborators.

Segmentation: Designed and implemented a Hidden Markov Model based algorithm to computationally detect copy number variation in the inbred mouse. The algorithm overcomes noise at single probes by leveraging on the correlation of values on the adjacent probes. Most of the experiments contained $\approx$80K probes while some $\approx$400K. The algorithm was efficient to scale to such large sizes. The algorithm was implemented in C++ and linked into Splus.

Mutation Rate: The primary advantage of this project is that the inbreeding history for mice is well known. However, mutation rate estimate is not immediate since most of the ancestral mutations will be lost due to segregation and hence will not be observed at the descendents. We designed and implemented a dynamic programming algorithm to estimate the mutation rate given the phylogeny and the set of mutations observed at the leaves. The algorithm was implemented in C++.

Quality: The array CGH experiments sometimes contained periodic noise of unknown origin. We implemented Fast Fourier Transform(FFT) to detect and remove such trends. The code was written in Splus.

Other problems: I worked on a few other projects which are not mentioned here in detail. These include improving statistical power in disease association testing using SNP data, finding recombination points using SNP data, shortest paths algorithms, sorting algorithms and priority queue data structures.

Programming Languages

C, C++, Java, Matlab/Octave, Perl, Splus/R.

Awards

Co-author of one of the contributing papers to **Science magazine's breakthrough of the year 2007**

Best paper award, International Symposium of Bioinformatics Research and Applications (ISBRA) 2007

Travel grant for attending HapMap conference, Oxford, UK 2005

Graduate Fellowship at Computer Science Department, Carnegie Mellon University (2003-present)

Dean's honored graduate: elected by the faculty to be the top graduate among the graduating computer science class of 2003, University of Texas, Austin

Nortel Networks, Cisco Systems and Proctor and Gamble fellowships for undergraduate research

Sun certified Java 2 programmer

Manuscripts Under Review

A Human Genome-Wide Library of Local Phylogeny Predictions for Whole-Genome Inference Problems. S. Sridhar and R. Schwartz.

Peer-Reviewed Manuscripts

Mixed Integer Linear Programming for Maximum Parsimony Phylogeny Inference. S. Sridhar, F. Lam, G. E. Blelloch, R. Ravi and R. Schwartz. To appear in *ACM/IEEE Transactions on Computational Biology and Bioinformatics (TCBB)* 2008.

Estimating Local Ancestry in Admixed Subpopulations. S. Sankararaman, S. Sridhar, G. Kimmel, E. Halperin. In *American Journal of Human Genetics* 2008.

Direct Maximum Parsimony Phylogeny Reconstruction from Genotype Data. S. Sridhar, F. Lam, G. E. Blelloch, R. Ravi and R. Schwartz. In *BMC Bioinformatics* 2007.

Recurrent DNA copy number variation in the laboratory mouse. C. Egans, S. Sridhar, M. Wigler and I. Hall. To Appear in *Nature Genetics* 2007.

Efficiently Finding the Most Parsimonious Phylogenetic Tree via Linear Programming. To appear in proc *International Symposium on Bioinformatics Research and Applications (ISBRA) 2007.* S. Sridhar, F. Lam, G. E. Blelloch, R. Ravi and R. Schwartz. **Won the best paper award.**

An Efficient and Accurate Graph-Based Method to Detect Population Substructure. To appear in proc *Research in Computational Molecular Biology (RECOMB) 2007.* S. Sridhar, S. Rao and E. Halperin. Gave an oral presentation at the conference.

Optimal Imperfect Phylogeny Reconstruction and Haplotyping. In Proc *Computational Systems Bioinformatics (CSB) 2006.* S. Sridhar, G. E. Blelloch, R. Ravi and R. Schwartz. Gave an oral presentation at the conference.

Algorithms for Efficient Near-Perfect Phylogenetic Tree Reconstruction in Theory and Practice. In *ACM/IEEE Transactions on Computational Biology and Bioinformatics(TCBB) 2007.* S. Sridhar, K. Dhamdhere, G. E. Blelloch, R. Ravi and R. Schwartz.

Fixed Parameter Tractability of Binary Near-Perfect Phylogenetic Tree Reconstruction. In Proc *International Colloquium on Automata, Languages and Programming (ICALP) 2006.* G. E. Blelloch, K. Dhamdhere, E. Halperin, R. Ravi, R. Schwartz and S. Sridhar. Gave an oral presentation at the conference.

Simple Reconstruction of Binary Near-Perfect Phylogenetic Trees. In Proc *International Workshop on Bioinformatics Research and Applications (IWBRA) 2006.* S. Sridhar, K. Dhamdhere, G. E. Blelloch, E. Halperin, R. Ravi and R. Schwartz. Gave an oral presentation at the conference.

Relaxing Haplotype Block Models for Association Testing. In proc *Pacific Symposium on Biocomputing (PSB) 2006.* N. Castellana, K. Dhamdhere, S. Sridhar and R. Schwartz.

Experimental Evaluation of a New Shortest Paths Algorithm. In proc *Algorithm Engineering and Experiments (ALENEX) 2002.* S. Pettie, V. Ramachandran and S. Sridhar.

Technical Reports

Evaluation of the Haplotype Motif Model using the Principle of Minimum Description. *Carnegie Mellon University, Computer Science Tech Report.* S. Sridhar, K. Dhamdhere, G. E. Blelloch, R. Ravi and R. Schwartz.

A Heap-Based Inversions-Sensitive Sorting Algorithm. *University of Texas, Austin, Computer Science Tech Report.* V. Ramachandran and S. Sridhar

Presentations Besides Above Listed Conferences

Evaluation of Halplotype Motif Model using the Principle of Minimum Description. Poster presentation at *HapMap and the Genomics*, University of Oxford, UK 2005. **Awarded travel grant**.

Near-Perfect Phylogenetic Tree Reconstruction from Genotypes and Haplotypes. Chalk talk at University of California, Davis, 2006.

Reconstruction and Applications of Near-Perfect Phylogenetic Trees. Oral presentation at International Computer Science Institute, Berkeley, California, 2006.

Efficiently Finding the Most Parsimonious Phylogenetic Tree via Linear Programming. Oral presentation at *Annual Meeting Institute of Operations Research and Management Sciences (INFORMS)*, Pittsburgh, Pennsylvania, 2006.

Faculty References

Guy E. Blelloch
Computer Science Department
Carnegie Mellon University.
Email: blelloch@cs.cmu.edu

Russell Schwartz
Dept of Biological Sciences
Carnegie Mellon University
Email: russells@andrew.cmu.edu

R. Ravi
Tepper School of Business
Carnegie Mellon University
Email: ravi@cmu.edu

Eran Halperin
International Computer Science Institute
Berkeley, CA
Email: heran@icsi.berkeley.edu