ELSEVIER

Discrete Applied Mathematics 71 (1996) 1–4

DISCRETE APPLIED MATHEMATICS

Introduction

With this volume, *Discrete Applied Mathematics* starts a series of special issues devoted to combinatorial and algorithmic techniques in computational molecular biology. This series will publish novel research results on the mathematical and algorithmic foundations of the inherently discrete aspects of computational biology.

The traditional partnership of mathematics and physics has advanced and enriched both disciplines. In a similar partnership, mathematics and computing are becoming crucial tools in the rapid advancement of molecular biology. At the same time, computational challenges in biology raise exciting new problems in discrete mathematics and theoretical computer science. To paraphrase Stan Ulam, those challenges reflect "not [only] what mathematics can do for biology but what biology can do for mathematics."

The diversity of the papers in this special volume demonstrates the richness of this new and exciting area. We have taken the liberty of partitioning the papers into seven areas: sequence comparison, mapping, molecular evolution, protein structure, genome rearrangements, DNA computing, and DNA statistics and computational support for biological experiments.

Sequence comparison. The area of sequence comparison is arguably the most thoroughly-studied area of computational biology, as it has attracted researchers for over 20 years now. Nevertheless, many practical and theoretical challenges still remain, and some of them are addressed here. In their paper, "Alignment networks and electrical networks", Vingron and Waterman study the problem of attributing weights to overlapping sets and their elements, which has important applications for DNA sequence alignment. The authors reveal an interesting connection between alignment weighting problems and electrical circuits. "Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method", the note by Schneider and Mastronarde, reports on a fast search-based method for the multiple alignment problem and on its implementation and performance. In the paper "Local multiple alignment via subgraph enumeration", Zhang, He, and Miller study the problem of computing multiple alignment from given pairwise alignments.

Mapping. Physical mapping is an essential tool for gene hunting projects and for the advent of the Human Genome Project into the large-scale sequencing phase. As it has a major computational component and it ties to mainstream graph-theoretical models, it has already attracted a lot of attention in the computer science community over the last five years. Most problems which were modeled in this area

PII S0166-218X(96)00073-X

are computationally hard, as exemplified by the papers in this issue: In the paper "On physical mapping and the consecutive ones property for sparse matrices", Atkins and Middendorf refer to the mapping problem for probe-clone incidence data. They strengthen previous results by showing the NP-hardness of physical mapping of sparse data under (1) chimericism restricted to two, and (2) non-unique probes where the probe multiplicities are at most two. Bodlaender and de Fluiter, in their paper "On intervalizing k-colored graphs for DNA physical mapping", refer to the mapping problem for clone–clone incidence data. They prove that the problem of completing a k-colored graph to a properly colored interval graph is NP-hard for any fixed number $k \geqslant 4$, but polynomial on biconnected graphs when $k = 3$.

Molecular evolution. This is another area that has been studied for many years from a biological perspective, and in recent years has drawn interest from the computer science community, leading to many new results of both practical and theoretical interest. Dress, Huson and Moulton review the split-decomposition technique for analyzing DNA sequences and describe a software tool called SplitsTree for producing splits graph. In the cases when it is hard to reconstruct unambiguously the true phylogenetic tree, splits graphs are effective alternative tools in molecular evolution. Goldberg, Goldberg, Phillips, Sweedyk, and Warnow, in the paper "Minimizing phylogenetic number to find good evolutionary trees" introduce a new approach to evaluating evolutionary trees. They show that the related l-phylogeny problem is NP-hard for any $l > 1$, but in the case of fixed topology it is polynomial for $l = 2$ and when the number of states is bounded. In the paper "On the complexity of comparing evolutionary trees", Hein, Jiang, Wang, and Zhang, study several problems arising in the comparison of rooted unordered trees with uniquely labeled leaves. They present a polynomial-time algorithm and some NP-hardness and non-approximability results for variants of the maximum agreement subtree and the maximum refinement subtree problems. They also prove the NP-hardness of computing the subtree-transfer distance, and give an approximation algorithm with performance ratio 3. Phillips and Warnow discuss the problem of inferring a consensus of a set of evolutionary trees. They propose a new model for consensus tree which is called asymmetric median tree and show that the asymmetric median tree combines desirable features of different consensus tree models.

Protein structure. As the three-dimensional (3D) structure of a protein, together with its chemical properties, determine its function to a large extent, structural studies are of the utmost importance. Numerous geometric studies have dealt with structural representation and comparison, and even more studies have been devoted to the protein folding problem viewed as a nonlinear optimization problem. However, new insights into these problems can be obtained using discrete mathematical tools, as shown by three papers in this volume: Akkiraju and Eddelsbrunner describe in their paper a new approach to triangulating the surface of a molecule, under the space-filling diagram, the solvent-accessible surface, and the molecular surface models. The method, based on a simplicial complex dual to the molecule models, is shown to be fast, robust, and results in topologically correct triangulations. MacGregor Smith and Toppur explore in their paper the relationship between the geometric properties of 3D Steiner minimal trees, minimum energy configurations, and the weighted graphs embedding problem in 3D. These relationships are then used to find minimum energy configurations for a class of Collagen proteins. The paper "On the complexity of string folding", by Paterson and Przytycka, gives an NP-hardness proof for a new 3D grid-based simplified model for protein folding. Interestingly, some "gadgets" used in their reduction resemble real-life biological secondary structures.

Genome rearrangements. This field is relatively young, as only the accumulation of large-scale DNA and mapping data in recent years has made comparative analyses of whole genomes possible. Several exciting results have already been obtained in this area, which poses elegant combinatorial problems. Bafna, Narayanan and Ravi, in their paper "Nonoverlapping local alignments (weighted independent sets of axis parallel rectangles)", study a problem motivated by selecting fragments of high local similarity between two strings, a preceding step to studying genome rearrangements. They prove the hardness of the problem, give an approximation algorithm with constant performance ratio, and provide a tight analysis of a local-improvement algorithm for the problem. Chen and Skiena, in their paper "Sorting with fixed-length reversals", study the problem of sorting a permutation using reversals with fixed length. They characterize the equivalence classes of such permutations, and obtain upper and lower bounds on the diameter of the permutations group under such reversals. Hannenhalli describes in his paper the first polynomial-time algorithm for computing translocation distance between genomes. The paper "Conserved synteny as a measure of genomic distance", by Sankoff and Nadeau, derives the distribution of the number of sampled genes per conserved segment to be expected when comparing synteny between two genomes, and investigates the bias in estimating syntenic distance caused by yet-unknown genes and by translocations.

DNA computing. The idea of "using biology to compute for us", which was initiated and exemplified by Adleman in 1994, has excited many people inside and outside the scientific community, and is currently actively explored. The paper "On the computational power of DNA", by Boneh, Dunworth, Lipton, and Sgall, carries this idea further by showing how DNA-based computers can be used to solve the satisfiability problem for boolean circuits, and to directly solve optimization problems. The authors also show how to use their methods for random sampling of satisfying assignments and for evaluating functions in the polynomial hierarchy.

DNA statistics, and computational support for biological experiments. The paper "Shuffling biological sequences", by Kandel, Matias, Unger, and Winkler, gives two algorithms which allow efficient random shuffling of a biological sequence so that its k-let frequencies are maintained. Pearson, Robins, Wrege and Zhang address the problem of primer selection in PCR experiments. Although the problem of minimizing the number of primers in NP-complete, a branch-and-bound algorithm, suggested by the authors, works well for biological data.

The process of editing a special volume is always a community effort, and we are grateful to all the anonymous referees who evaluated the papers and have enabled the timely publication of this volume. We are extremely indebted to Dr. Peter Hammer, the chief editor of *Discrete Applied Mathematics*, for initiating the idea of this volume, for making this volume possible, and for his guidance in all the intricacies of the editorial world. We are also thankful for the efforts of Mrs. Nelly Segal, the Editorial Manager of *Discrete Applied Mathematics*, who helped us with experienced advice and performed the organizational work involved in the final preparation of the volume. We hope that this volume will indeed be the first in a long series and will help to invigorate the young but rapidly-growing community in the exciting and important new area of computational biology

Sorin Istrail
Pavel Pevzner
Ron Shamir

Sorin Istrail
Algorithms and Discrete Mathematics Department
Sandia National Laboratories
P.O. Box 5800, MS 1110
Albuquerque, NM 87185-5800, USA.
E-mail: scistra@cs. sandia.gov
homepage: http://www.cs.sandia.gov/~scistra/

Pavel Pevzner
Department of Mathematics, DRB 155
University of Southern California
Los Angeles, CA 90089-1113, USA.
E-mail: ppevzner@hto.usc.edu
homepage: http://www-hto.usc.edu/people/Pevzner.html

Ron Shamir
Department of Computer Science
School of Mathematics
Tel Aviv University
Tel Aviv 69978 Israel.
E-mail: shamir@math.tau.ac.il
homepage: http://www.math.tau.ac.il/~shamir/