Contents

Abstracts

Combinatorial Problems of Strings and Graphs and Their Applications in Bioinformatics Part 2

4–15 April 2022

1 Md. Shamsuzzoha Bayzid

Bangladesh University of Engineering and Technology, Bangladesh
Genome-scale species tree estimation from quartets

Abstract

A *species* tree represents the evolutionary history of a group of organisms, while a *gene tree* represents the evolutionary pathways of a particular gene within a group of species. Estimations of species trees are typically based on multiple genes, in some cases from throughout the genome. However, species tree estimation from genes sampled from throughout the whole genome is complicated due to the *gene tree-species tree discordance*. Incomplete lineage sorting (ILS) is one of the most frequent causes for this discordance. Quartet-based methods for estimating species trees from a collection of gene trees are becoming popular due to their high accuracy and statistical guarantee under ILS. Generating quartets with appropriate weights, where weights correspond to the relative importance of quartets, and subsequently amalgamating the weighted quartets to infer a single coherent species tree allows for a statistically consistent way of estimating species trees. In this talk, I will discuss quartet-based species tree estimation methods. I will particularly present wQFM, a new method that matches or improves on the accuracy of ASTRAL, one of the most accurate and widely used coalescent-based species tree estimation methods.

2 Sankardeep Chakraborty

The University of Tokyo, Japan
Succinct data structures for automata with optimal membership query

Abstract

Finite automata are basic and fundamental combinatorial objects with myriad applications. In this talk, we show how one can represent these objects optimally. Furthermore, given a string x, a very basic and important query is to figure out if x belongs to the language accepted by the automata. We show that it is possible to support this membership query efficiently. If time permits, we also show some extensions of the basic data structures for some special cases.

3 Roberto Grossi

Università di Pisa, Italy
Fast assessment of Eulerian trails in graphs, and its relation to sequences

Abstract

An Eulerian Trail (ET) traverses all the edges of a directed graph G = (V,E) exactly once. Given an integer z, we want to assess whether there are at least z node-distinct ETs in G, where two ETs are node-distinct if their sequences of nodes are different. This problem has been formalized by Bernardini et al. [ALENEX 2020] as the core computational problem in several string processing applications. We show how to design a combinatorial algorithm for assessing ETs that requires O(z —E—) time and does not need the well known BEST theorem.

Joint work with A. Conte, G. Loukides, N. Pisanti, S. P. Pissis and G. Punzi

4 Wing-Kai Hon

National Tsing Hua University, Hsinchu
FM-Indexing grammars induced by suffix sorting

Abstract

Run-length compressed Burrows–Wheeler transform (RLBWT) used in conjunction with the backward search introduced in the FM index is the centerpiece of most compressed indexes working on highly-repetitive data sets like biological sequences. Compared to grammar indexes, the size of the RLBWT is often much bigger, but queries like counting the occurrences of long patterns can be done much faster than on any existing grammar index so far. In this talk, we combine the virtues of a grammar with the RLBWT by building the RLBWT on top of a special grammar based on induced suffix sorting. Our experiments reveal that our hybrid approach outperforms the classic RLBWT with respect to the index sizes, and with respect to query times on biological data sets for sufficiently long patterns, which could be interesting for aligning long reads in bioinformatics.

This is a joint work with Jin-Jie Deng, Dominik Köppl, and Kunihiko Sadakane.

5 Yu Lin

Australian National University, Australia
Binning metagenomic sequences

Abstract

Metagenomics studies have provided key insights into the composition and structure of microbial communities found in different environments. Among the techniques used to analyze metagenomic data, binning is considered as a crucial step in order to characterize the different species of microorganisms present in microbial communities. We propose two binning methods, GraphBin and MetaBCC-LR, to bin metagenomic sequences. GraphBin makes use of the assembly graph to refine binning results and to enable detecting shared sequences among multiple species. MetaBCC-LR is a reference-free binning method which directly clusters long reads based on their k-mer coverage histograms and oligonucleotide composition.

6 Bernard Moret

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Randomized algorithms in computational biology

Abstract

Randomized algorithms are a mainstay of algorithmic development in Computer Science. They are typically much simpler than deterministic algorithms with comparable performance and more robust thanks to their stochastic component. However, they have yet to see much use in computational biology. In this talk, I will briefly present the general idea on two trivial problems, then illustrate two applications to phylogenetics, both algorithms to measure tree-to-tree distances: one creating randomized fingerprints for a pair of trees and the other combining fingerprints, bounded-error projection into a random subspace (Johnson-Lindenstrauss lemma), and hashing, to compute all pairwise tree-to-tree distances in a forest faster than any deterministic algorithm.

7 Md. Saidur Rahman

Bangladesh University of Engineering and Technology, Bangladesh
Research on pairwise compatibility graphs: current state and open problems

Abstract

Let T be an edge weighted tree and d_{min}, d_{max} be two non-negative real numbers where $d_{min} \leq d_{max}$. The pairwise compatibility graph (PCG) of T for d_{min}, d_{max} is a graph G such that each vertex of G corresponds to a distinct leaf of T and two vertices are adjacent in G if and only if the weighted distance between their corresponding leaves lies within the interval $[d_{min}, d_{max}]$. A graph G is a PCG if there exist an edge weighted tree T and suitable d_{min}, d_{max} such that G is a PCG of T. The class of pairwise compatibility graphs was introduced to model evolutionary relationships among a set of species. Since not all graphs are PCGs, researchers become interested in recognizing and characterizing PCGs. In this talk I review the results regarding PCGs and some of its variants. I also discuss some open problems which we

are considering over the years. The talk is based on my joint works with M. N. Yanhaona, M. S. Bayzid, K. S. M. T. Hossain, S. A. Salma, S. Durocher, D. Mondal, S. Ahmed, B. B. Papan, P. B. Pranto and S. A. Hakim.

8 Sohel Rahman

Bangladesh University of Engineering and Technology, Bangladesh
Phylogeny-aware MO optimization approach for computing MSA

Abstract

Multiple sequence alignment (MSA) is an important early step in the pipeline of inferring the phylogenetic tree where the given sequences are arranged according to evolutionary history. The characteristics, as well as the quality of the obtained MSA dramatically influence the accuracy of the estimated tree. Usually, the MSAs are inferred by optimizing a single function or objective. The alignments estimated under one criterion may be different from the alignments generated by other criteria, inferring discordant homologies and thus leading to different hypothesized evolutionary histories relating the sequences. Therefore, multi-objective (MO) optimizations, where multiple conflicting objective functions are being optimized simultaneously to generate a set of alternative alignments, seem appealing and have been considered in the literature. However, no theoretical or empirical justification with respect to a real-life application has been shown for a particular MO formulation. In this talk, we examine whether a phylogeny-aware metric can guide us in choosing appropriate MO formulations that can result in better phylogeny estimation. Employing MO metaheuristics, we demonstrate that (a) trees estimated on the alignments generated by MO formulation are substantially better than the trees estimated on the alignments generated by the state-of-the-art MSA tools and (b) highly accurate alignments with respect to popular measures do not necessarily lead to highly accurate phylogenetic trees.

9 Mingfu Shao

Pennsylvania State University USA
SubseqHash and its applications in error-prone long reads analysis

Abstract

We present SubseqHash, a hashing function that maps a k-mer to one of its subsequence of length b with highest rank according to an ordering. We analyze the probability of hash collision and propose a linear-time algorithm to construct such a hashing function. Two variants of SubseqHash are proposed with constraints added to speed up its construction and to reduce false positives. Through experimental studies we show that SubseqHash outperforms strobemers and k-mer methods in analyzing long reads with high error-rate, including chaining, and overlap detection.

10 Tetsuo Shibuya

The University of Tokyo, Japan
Differentially private methods in bioinformatics

Abstract

Bioinformatics is a research area where we analyze various biomedical data, but many of these data, especially the increasing individual genome data, contain highly sensitive data. When we publish even the simplest statistics of biomedical databases, we need to be very careful so that (preferably) no individual sensitive information will be revealed. Differential privacy is one of the most important measures to evaluate how safe the published data are from the viewpoint of privacy. In this talk, we will give several efficient differential private methods for publishing key statistics in genome association study, such as chi-square test, Fisher's test, Cochran-Armitage's trend test and TDT test.

11 Krister Swenson

LIRMM and CNRS, France
Using quadragulations and planar trees for weighted genome rearrangement

Abstract

Genome rearrangement has inspired many remarkable algorithmic results in the last 30 years. The applicability of these results has been somewhat limited due to many factors, however, including an inability to inform the choice of rearrangements based on biological information. We present techniques, developed with Pijus Simonaitis, for the weighting of genome rearrangements based on properties of the adjacencies that they act upon. Such weightings facilitate the use of chromatin conformation information, as represented by Hi-C data. The edge-pair weighted rearrangement problem is reduced to the computation of a planar tree linking vertices embedded on a convex polygon. More general weighted rearrangement problems are reduced to the computation of a quadragulation of a convex polygon. We present polynomial time dynamic programs for each of these cases.

12 Yanni Sun

City University of Hong Kong, China
Charactering viral haplotypes using long reads

Abstract

Most RNA viruses lack strict proofreading during replication. Coupled with a high replication rate, some RNA viruses can form a virus population containing a group of genetically-related but different haplotypes. Characterizing the haplotype composition in a virus population is thus important to understand viruses' evolution. Many attempts have been made to reconstruct viral haplotypes using next-generation sequencing (NGS) reads. However, the short length of NGS reads cannot cover distant single-nucleotide variants (SNVs), making it difficult to reconstruct complete or near-complete haplotypes. Given the fast developments of third-generation sequencing (TGS)

technologies, a new opportunity has arisen for reconstructing full-length haplotypes with long reads.
In this talk, I will present our recent contributions to viral haplotype reconstruction using TGS data. The long reads can cover multiple SNVs, making accurate reconstruction of viral haplotypes feasible. We use statistical and learning methods to distinguish real SNVs from sequencing errors from long reads. We tested our method rigorously on multiple datasets and the results show that it can reconstruct viral haplotypes with higher accuracy than available tools.

13 Yufeng Wu

University of Connecticut, USA
Probability computation for trees and networks under coalescent theory

Abstract

Coalescent theory is a fundamental theory in population and evolutionary genetics. Applying coalescent theory to inference problems requires computation of probability under various settings. However, accurate computation of coalescent probability in many settings is computationally difficult. In this talk, I will focus on the so-called multispecies coalescent model. Multispecies coalescent concerns genes (lineages) that originate from multiple related populations (species), where coalescent may occur outside the population boundary. Suppose we are given a gene tree (which represent the evolutionary history of a gene of lineages from several populations) and a population tree (for the evolutionary history of populations). Given a gene tree topology and a population tree, the likelihood of the gene tree (called gene tree probability in the literature) is the probability of observing the gene tree topology on the population tree under coalescent theory. In the first part of the talk, I will describe an algorithm (published in a paper in Evolution, 2012) for computing this coalescent probability, which is much faster than a previous algorithm. I will then present an extension (published in a paper in ISMB 2020) which computes the gene tree probability under a more complex network model. I will briefly explain the applications of these two algorithms. Research partly supported by US National Science Foundation under grants CCF-1718093 and IIS-1909425.

14 Han Xu

MD Anderson Cancer Center, USA
Modeling and prediction of CRISPR/Cas9 sensitivity and specificity

Abstract

The CRISPR/Cas9 genome editing technology has revolutionized biomedical research. In a CRISPR/Cas9 system, a Cas9 protein and a single guide RNA (sgRNA) form a complex to bind and cleave DNA at targeted locus. The efficiency of cleavage (sensitivity) and the off-target effect (specificity) are highly dependent on the sequence context of sgRNA. In this talk, I will introduce a series of statistical and machine learning methods for the modeling and prediction of CRISPR/Cas9 sensitivity and specificity. These methods collectively address the needs for rational design of sgRNAs and boost the performance of CRISPR/Cas9 experiments. Moreover, the modeling of CRISPR specificity led to a new strategy for robust allele-specific genome editing.

15 Louxin Zhang

National University of Singapore, Singapore
The Sackin index of simplex networks

Abstract

A phylogenetic network is simplex if the child of every reticulation node is a network leaf and every node has at least one child that is not reticulate. Simplex networks are a superclass of phylogenetic trees and a subclass of tree-child networks. The Sacking index of a phylogenetic tree is defined to be the sum of the depths of all the leaves in the tree and is used to measure the balance of the tree. We first present a simple proof that the expected Sackin index of phylogenetic trees with n leaves is $2^{n-1}n!/a_n - n$, which is asymptotically $\Omega(\mathrm{n}^{3/2})$, in the uniform model, where a_n is the number of phylogenetic trees with n leaves. Generalizing the Sackin index to phylogenetic networks, we further prove that the expected Sackin index of a random simplex network is asymptotically $\Omega(\mathrm{n}^{7/4})$ in the uniform model.

16 Xiuwei Zhang

Georgia Institute of Technology, USA

Reconstructing cell lineage tree using single cell lineage barcode and gene expression data

Abstract

Understanding how single cells divide and differentiate into different cell types in developed organs is one of the major tasks of developmental biology. Recently, lineage tracing technology using CRISPR/Cas9 genome editing have enabled simultaneous readouts of gene expressions and lineage barcodes, which allows for the reconstruction of the cell division tree from the barcode data, and makes it possible to reconstruct ancestral cell types and trace the origin of each cell type at the whole organism level. The following two types of data are available for each single cell: the single cell gene expression and the lineage barcode of each cell. Computational tools to reconstruct the cell division history from the lineage barcodes have been developed, similarly to the practice of using genome data from multiple species to reconstruct the phylogenetic tree. However, the challenges of reconstructing a cell division tree are: (1) there is a lot of missing data in the barcodes of cells; (2) there are a large number of cells. One promising strategy to improve the reconstructed lineage barcode data is to incorporate the gene expression data. In this talk, I will present existing methods using both the barcode and gene expression data for cell division tree reconstruction, as well as a new method developed in our group.