Conference Program

Fourth RECOMB Satellite Conference on Bioinformatics Education (RECOMB-BE 2012)

August 26, 2012

RECOMB Satellite Conference on Open Problems in Algorithmic Biology (RECOMB-AB 2012)

August 27-29, 2012

St. Petersburg, Russia

Contents

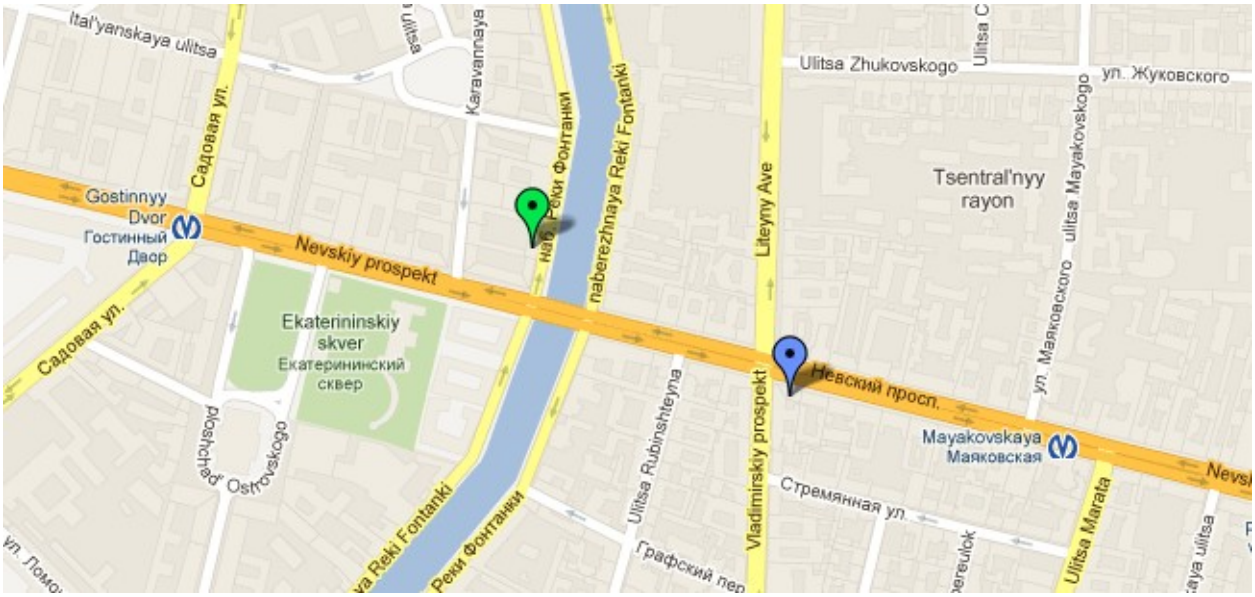
Directions

Venue

RECOMB-BE will be held in the St. Petersburg Department of the V.A. Steklov Institute of Mathematics of the Russian Academy of Sciences (Fontanka 27, green marker on the map below).

RECOMB-AB will be held at the Radisson Royal Hotel (49/2 Nevsky Prospect, blue marker).

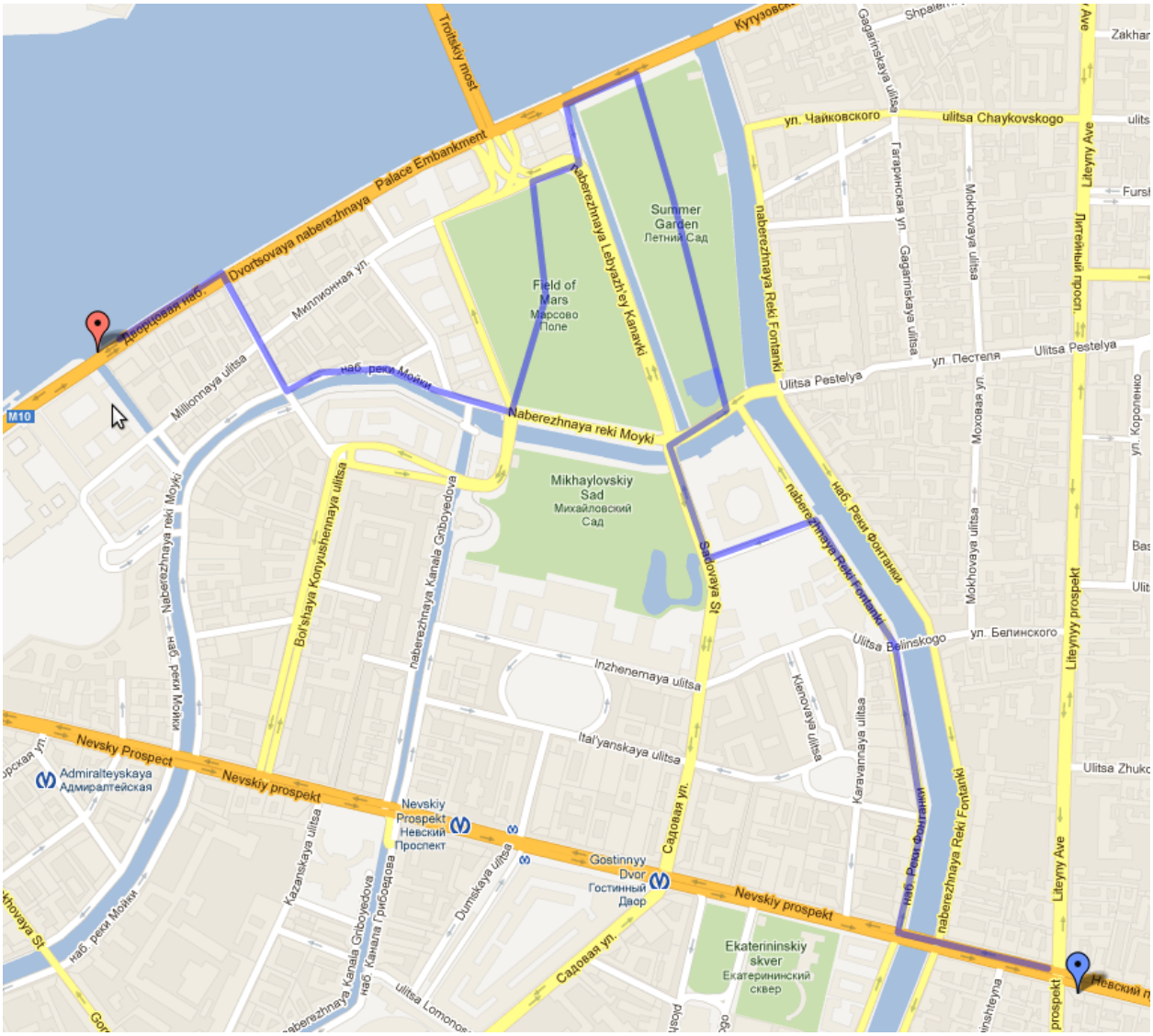
Both venues are located in the city center, between underground stations Gostinnyy Dvor (Гостинный двор) and Mayakovskaya (Маяковская).



RECOMB-AB social program

RECOMB-AB participants are invited to an evening of music on a jazz boat that will leave from Zimnyaya kanavka (red marker) on August 28 at 19:45.

The boat is an approximately 20 minute walk from the Radisson (blue marker). The map also shows a non-optimal but particularly nice walk from the Radisson to Zimnyaya kanavka. This alternate route takes about an hour and passes by many beautiful St. Petersburg landmarks.



Contacts

- Alexander Kulikov: +7 911 240 9485
- Nikolay Vyahhi: +7 904 645 65 73
- Radisson: +7 812 322 50 00
- English language tourist helpline: +7 812 300 33 33

Fourth RECOMB Satellite Conference on Bioinformatics Education (RECOMB-BE 2012) August 26, 2012

08:30 - 09:00	Registration, Welcome Coffee		
Rosalind (Chair: Ron Shamir)			
09:00 - 09:15	Rosalind Introduction		
09:15 - 10:00	Bioinformatics Learning 2.0: proposing an open source consortium for bioinformatics teaching materials	Christopher Lee	Invited
10:00 - 10:10	Sequence Composition	Gabriel Valiente	Rosalind
10:10 - 10:20	Three Problems Illustrating Bioinformatics Concepts In a Standard Spreadsheet	Tomáš Vinař	Rosalind
10:20 - 10:50	**Coffee Break**		
10:50 - 11:10	From Sequence to Structure and Function: Inspiring Students with Bioinformatics Problems	Brian Tjaden	Rosalind
11:10 - 11:20	The Number of Reversing Substitutions	Sergey Naumenko	Rosalind
11:20 - 11:30		Jennifer McDowall	Rosalind
11:30 - 12:10	Discussion Panel 1: Rosalind and online problem solving		
12:10 - 13:40	**Lunch (Radisson)**		
Bioinformatics for Biologists (Chair: Nikolay Vyahhi)			
13:40 - 14:20	Comparing sequence motifs	Uri Keich	Invited
14:20 - 14:35	Machine learning methods for protein sorting prediction	Henrik Nielsen	B4B
14:35 - 14:50	Evolutionary History of Repeats	Sergey Nikolenko	B4B
14:50 - 15:05	Identifying the Microbiomic Basis of Disease	Gabriel Valiente	B4B
15:05 - 15:35	**Coffee Break**		
15:35 - 16:15	Alignment Beyond Sequences: Forwards and Backwards in Colors and DAGs	Michael Brudno	Invited
16:15 - 17:05	Discussion Panel 2: How do we teach bioinformatics to 10K biology students at a time?		
17:05 - 17:10	Closing Remarks		

Invited

Bioinformatics Learning 2.0: proposing an open source consortium for bioinformatics teaching materials

Christopher Lee
University of California, Los Angeles

I propose an open source consortium for bioinformatics teaching materials, including textbook chapters, slides, concept tests, homework and exam questions and answers, programming problems and data analysis projects, and software tools for using these materials in class and out.

To seed this effort I am contributing materials from two courses: a bioinformatics theory course (for Computer Science students) emphasizing probabilistic models and methods; and a genomics and computational biology course (for Life Science students).

This effort is based on several principles. First, bioinformatics is highly interdisciplinary, yet bioinformatics textbooks tend to each reflect only one disciplinary part of that. Furthermore, both available textbooks and the traditional lecture method fall far short of giving students adequate exercises to truly learn the concepts and skills. In effect, the job of writing all these teaching materials is too big for any one person. Instead, every teacher should be enabled to focus on writing materials in areas where they are expert, while drawing whatever materials they want from everyone else, via an open source consortium for sharing teaching materials.

Second, bioinformatics teaching should draw lessons from other fields such as physics teaching, where it has been shown that traditional lecturing (passive learning) is far less effective than active learning, where students answer and discuss problems in class. Specifically, I have developed teaching materials and software tools for in-class concept tests, defined as a question that challenges the students' understanding of a specific concept.

Whereas ROSALIND computational problems may be viewed as empirical (implicit) tests of mastery of a concept or skill, in-class concept testing explicitly teaches such mastery by challenging students to think about how to use a concept, and rapidly exposing the most common errors for all to see and understand. I illustrate with examples from the approximately 300 bioinformatics concept tests I have written for this effort.

I also present software tools for in-class concept testing, and for selecting and "re-compiling" content in flexible ways. Finally, I will discuss critical issues for such a consortium, such as automatic authorship tracking, sharing, and security.

For more details, see http://thinking.bioinformatics.ucla.edu/teaching.

Rosalind talk

Sequence Composition
Gabriel Valiente
Technical University of Catalonia

A genomic or proteomic sequence can be seen as composed of a number of possibly overlapping words of a certain length, and the composition of a sequence is given by the frequency with which each possible word occurs within the sequence. In this talk, we review the biological significance of sequence composition and discuss efficient methods to obtain the word composition of a sequence, along with their implementation in the framework of the ROSALIND programming and testing environment for bioinformatics problems.

Rosalind talk

Three Problems Illustrating Bioinformatics Concepts In a Standard Spreadsheet

*Tomáš Vinař, Brona Brejova
Comenius University

Here, we describe three problems that we have previously used in the context of a bioinformatics class taught at the Comenius University in Bratislava. The class is targeted at both computer science and biology students. Students with both backgrounds attend the same lectures, while tutorials and assignments are provided separately for biologists and computer scientists.

One particular challenge in teaching this course is to design assignments for biology students, illustrating basic algorithmic and mathematical concepts used in bioinformatics without requiring prior programming experience. The class does not require any previous programming courses, nor it is the goal of the class to teach programming. We have found that many concepts can be illustrated in a standard spreadsheet (MS Excel or one of its open-source equivalents) to which most of the students have been exposed previously.

Rosalind talk

From Sequence to Structure and Function: Inspiring Students with Bioinformatics Problems
Brian Tjaden
Wellesley College

Recent advances in sequencing technology have enabled scientists to gather large amounts of DNA and RNA sequence data. One of the bioinformatics challenges is extracting new insights about the structure and function of biomolecules from the wealth of sequence data. In this talk, we look at two problems in the field of computational molecular biology designed to stimulate and challenge students. The first problem relates to understanding the secondary structure of an RNA molecule based on its primary sequence. The second problem relates to processing large amounts of DNA sequence data so as to capture the internal structure in the data and support a range of queries on the data efficiently. Applications will be discussed for aligning high-throughput sequencing reads to a genome and for screening a genome for interesting genetic elements such as CRISPRs.

Rosalind talk

The Number of Reversing Substitutions

Sergey Naumenko
Institute for Information Transmission Problems, Russian Academy of Sciences

The basic task for molecular evolution studies is to calculate the frequency of a particular event in the evolutionary history. Reversing substitution is an example of such molecular event. At some moment in the past the direct amino acid substitution $A \rightarrow B$ occured. And after a certain period of time, we observe the reversing substitution $B \rightarrow A$. Unfortunately, in most cases, with the possible exception of experimental evolution in bacteria, we don't know the intermediate (ancestral) state of a protein. We can observe proteins in human, mouse, dog, elephant and other species in their current state in the form of the multiple alignement of orthologous protein-coding genes. But we can restore the ancestral states in the internal nodes of the phylogenetic tree using the knowledge of amino acids on the terminal branches of the tree and the tree topology itself. There are a variety of methods (maximum parsimony, maximum likelihood, bayesian methods) and programs (PAML, Phylip, PAUP) to do so. Using the ancestral and terminal aminoacids at a site we can infer the substitutions. Problem. Given the multiple alignment with internal states restored and the phylogenetic tree it is necessary to calculate the number of reversing substitution for different distances between the direct and reversing subsitution.

The solution of this problem does not require the intelligent algorithm, but it is an example (simplified) of the real world problem in molecular evolution. It contains the basic concepts: the site, the phylogenetic tree, the multiple alignment, the correspondance between these two, the inference of substitution events.

Invited

Comparing sequence motifs
Uri Keich
University of Sydney

The identification of transcription factor binding sites is an important step in understanding the regulation of gene expression. To address this need, many motif-finding tools have been described that can find short sequence motifs given only an input set of sequences. Somewhat surprisingly, development of the significance analysis of the motifs reported by those motif finders has lagged considerably behind the extensive development of the finders themselves. Nevertheless, this analysis is often crucial in helping scientists decide whether or not to carry the predicted motifs to the next stage of their analysis. We will discuss the problem of evaluating the statistical significance of sequence motifs in the general context of evaluating the statistical significance of an observed result.

Bioinformatics for Biologists talk

Machine learning methods for protein sorting prediction

Henrik Nielsen
Technical University of Denmark

Prediction of "protein sorting", i.e. the subcellular location of proteins, has become a major task in bioinformatics. The problem is easy to formulate and understand from a biological point of view, yet the computational solutions are often complex and involve several machine learning methods. Thus, protein sorting is a well suited case for introducing sequence-based machine learning methods for biologists.

Methods for predicting protein sorting from the amino acid sequence can roughly be divided into three types: Homology-based methods that rely on alignment to proteins with known location; signal-based methods that attempt to recognize the actual sorting signals; and global property methods that utilize the fact that proteins from different subcellular compartments differ in amino acid composition or other global properties of the sequence.

In my presentation, I will focus on two very important sorting signals, the signal peptide and the transmembrane helix, and show how two machine learning methods, artificial neural networks and hidden Markov models, have been successfully applied in their recognition. In addition, I will briefly mention issues of training set / test set division and overfitting, which apply to all types of machine learning and are important to understand even for the casual user of such methods.

Bioinformatics for Biologists talk

Evolutionary History of Repeats

Sergey Nikolenko
St. Petersburg Academic University, Russian Acedemy of Sciences

Mobile elements constitute large portions of eukaryotic genomes. They are sequences that are often replicated, and replicated copies then undergo evolution separately; thus, by considering a family of mobile (repeat) elements one can deduce evolutionary history, leading to important insights on species relations, population structure etc. We consider probabilistic modeling of mobile elements phylogeny, starting from the simplest statistical considerations and then proceeding to more complicated models.

Bioinformatics for Biologists talk

Identifying the Microbiomic Basis of Disease
Gabriel Valiente
Technical University of Catalonia

Human genetic and metabolic diversity is heavily influenced by complex microbial communities that inhabit the human body. The microbiota is highly variable both within and between people in body habitats such as the gut, skin, and oral cavity, and changes in the microbiota can cause or prevent disease. In this talk, we discuss the biological problem of comparing microbial communities across people, body habitats, health conditions, and time, along with the related computational problems of designing taxonomically universal PCR primers, determining and quantifying the composition of environmental samples, and comparing abundance profiles across microbial communities.

Invited

Alignment Beyond Sequences: Forwards and Backwards in Colors and DAGs
Michael Brudno
University of Toronto

In this lecture I will present the algorithmic challenges presented by two novel types of sequencing technologies: the SOLiD system, which generates color-space reads, and Single-Molecule Sequencing systems, which have an extremely high indel error rate, but can read each piece of DNA two or more times. I will then explain how classical string alignment algorithms must be adopted to deal with this type of data, in particular explaining the generalization of sequence alignment to the Weighted Sequence Graph abstraction, and showing how this can be further adopted to work with color-space data.

RECOMB Satellite Conference on Open Problems in Algorithmic Biology
(RECOMB-AB 2012)
August 27-29, 2012

Monday August 27

08:00 - 08:45	Registration		
08:45 - 09:00	Introductory Remarks		
Assembly (Chair: Alexander Kulikov)			
09:00 - 09:45	SPAdes: a New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing	Sergey Nikolenko	Invited
09:45 - 10:00	De Bruijn Superwalk with Multiplicities Problem	Paul Medvedev	Open Problem
10:00 - 10:15	Mate-pair consistency and Generating Problems	Son Pham	Open Problem
10:15 - 10:45	**Coffee Break**		
Genome Rearrangements (Chair: Kira Vyatkina)			
10:45 - 11:30	Genome rearrangements: when intuition fails	Max Alekseyev	Invited
11:30 - 12:15	The combinatorics of the breakage fusion bridge and its application to cancer genomics	Vineet Bafna	Invited
12:15 - 13:30	**Lunch (included)**		
Phylogenetic Trees (Chair: Kira Vyatkina)			
13:30 - 14:15	Following Dobzhansky: Extending The Reach of Phylogenies	Bernard Moret	Invited
14:15 - 14:30	Algorithms for constructing k-articulated network: a more powerful classification of phylogenetic networks	S.M. Yiu	Open Problem
14:30 - 14:45	Phylogenetic tree reconstruction with protein linkage	Francis Chin	Open Problem
14:45 - 15:00	Reconciliation of Gene and Species Trees with Polytomies	Yu Zheng	Research Abstract
15:00 - 15:30	**Coffee Break**		
Alignment (Chair: Max Alekseyev)			
15:30 - 16:15	Alignment-Free Local Sequence Comparison: Statistics and Algorithms	Michael Waterman	Invited
16:15 - 16:30	Alignment Plots: Local Sequence Comparison in Sliding Windows	Alexander Tiskin	Research Abstract

Tuesday August 28

Systems Biology (Chair: Sergey Nikolenko)			
09:00 - 09:45	Three Optimization Problems in Molecular Biology and Genetics	Richard Karp	Invited
09:45 - 10:30	Dissecting inner structure in disease regulatory networks using differential co-expression	Ron Shamir	Invited
10:30 - 11:00	**Coffee Break**		
11:00 - 11:15	Metabolic stories: uncovering all possible scenarios for interpreting metabolomics data	Paulo Vieira Milreu	Open Problem

Genetics / Statistics (Chair: Nikolay Vyahhi)			
11:15 - 12:00	The Linkage Disequilibrium Measures Unification Problem	Sorin Istrail	Invited
12:00 - 12:15	Reconstructing Pedigrees from Data	Bonnie Kirkpatrick	Open Problem
12:15 - 12:30	Analysis of single-cell RNA-seq data using probabilistic mixture models	Peter Kharchenko	Research Abstract
12:30 - 14:00	**Lunch (on your own)**		

Sequencing (Chair: Nikolay Vyahhi)			
14:00 - 14:45	Efficient communication and storage vs. accurate variant calls in high throughput sequencing: two sides of the same coin	Cenk Sahinalp	Invited
14:45 - 15:45	**Coffee Break**		

Posters			
14:45 - 15:45	Registration and analysis of array images of general layout	Vitaly Galinsky	Open Problem
	The problems of reconciling gene and species trees, mapping a gene tree into a species tree and gene tree inference	Konstantin Gorbunov	Open Problem
	Finding a Combinatorially Optimal Model for RNA Folding Pathways	Bonnie Kirkpatrick	Open Problem
	Non-Identifiable Pedigrees and a Bayesian Solution	Bonnie Kirkpatrick	Research Abstract
	Open Problems in Metagenomics	Gabriel Valiente	Open Problem

Panel	
15:45 - 16:45	Discussion panel on problem formulation
19:45	**Jazz Boat**

Wednesday August 29

Systems Biology (Chair: Sergey Nurk)			
09:00 - 09:45	Lasting relations	Marie-France Sagot	Invited

Mass Spectrometry (Chair: Sergey Nurk)			
09:45 - 10:30	Sequencing antibiotics: bioinformatics meets nuclear physics	Pavel Pevzner	Invited
10:30 - 11:00	**Coffee Break**		
11:00 - 11:15	Combinatorial problems for SNP and mutation discovery using base-specific cleavage and mass spectrometry	Xin Chen	Open Problem

Evolution (Chair: Alla Lapidus)			
11:15 - 12:00	Evolution of bacterial genomes	Mikhail Gelfand	Invited
12:00 - 12:45	A Moving Landscape for Comparative Genomics in Mammals	Steve O'Brien	Invited
12:45 - 13:00	Concluding remarks		
13:00	**Lunch (on your own)**		

Monday August 27, 2012

Invited

SPAdes: a New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing
Glenn Tesler
University of California, San Diego

The lion's share of bacteria in various environments cannot be cloned in the laboratory and thus cannot be sequenced using existing technologies. A major goal of single-cell genomics is to complement gene-centric metagenomic data with whole-genome assemblies of uncultivated organisms. Assembly of single-cell data is challenging because of highly non-uniform read coverage as well as elevated levels of sequencing errors and chimeric reads. We describe SPAdes, a new assembler for both single-cell and standard (multicell) assembly, and demonstrate that it improves on the recently released E+V-SC assembler (specialized for single-cell data) and on popular assemblers Velvet and SoapDeNovo (for multicell data). SPAdes generates single-cell assemblies, providing information about genomes of uncultivatable bacteria that vastly exceeds what may be obtained via traditional metagenomics studies.

This is a joint work with Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey Gurevich, Mikhail Dvorkin, Alexander Kulikov, Valery Lesin, Sergey Nikolenko, Son Pham, Andrey Prjibelski, Alexey Pyshkin, Alexander Sirotkin, Nikolay Vyahhi, Max Alekseyev, and Pavel Pevzner.

Software: http://bioinf.spbau.ru/spades

Article: Bankevich and Nurk et al. (2012), Journal of Computational Biology, 19(5): 455-477, doi:10.1089/cmb.2012.0021.

Open Problem talk

De Bruijn Superwalk with Multiplicities Problem

*Paul Medvedev[1] and Michael Brudno[2]
[1]*University of California, San Diego*
[2]*University of Toronto*

Whole-genome shotgun sequencing is an experimental technique used for obtaining information about a genome's sequence, whereby it is broken up into many short segments (called reads) whose sequence is then determined. The problem of assembly is to reconstruct the sequence of the genome from these reads. Many common approaches rely on the de Bruijn graph, where the problem of genome assembly was formulated as the De Bruijn Superwalk Problem (Pevzner, Tang, and Waterman, 2001) and shown to be NP-hard (Medvedev et al., 2007). However, today's sequencing technologies provide a high amount of coverage, allowing us to make estimates regarding the multiplicities of the various k-mers (strings of length k) in the genome. It is clear that knowing these multiplicities should help to improve assembly, but it is not clear whether introducing this information as constraints to the formal problem makes it tractable. We formulate this question as the De Bruijn Superwalk with Multiplicities Problem.

Open Problem talk

Mate-pair consistency and Generating Problems

*Son Pham and Paul Medvedev
University of California, San Diego

Fragment assembly is among the most important and long standing problems in bioinformatics. It is often formulated as the problem of reconstructing a string from the set of its substrings. However, since most current sequencing platforms produce mate-pairs — pairs of reads whose separation is approximately known — a more useful formulation for the genome assembly problem should consider the reconstruction of a genome from the set of its mate-pairs. Unfortunately, while there have been a few methods proposed for utilizing mate-pairs in genome assemblers (Medvedev et al., 2011; Pevzner and Tang, 2001; Pham et al., 2012), there has been a lack of theoretical formulation for the problem of reconstructing a string from its mate-pair reads. In this work, we introduce two open problems for genome reconstruction from mate-pair reads.

Invited

Genome rearrangements: when intuition fails
Max Alekseyev
University of South Carolina

I shall describe two rather unexpected phenomena in genome rearrangements analysis. First, weighted genomic distance designed to bound the proportion of transpositions in rearrangement scenarios between two genomes does not actually achieve this goal. Second, while the median score of three genomes can be approximated by the sum of their pairwise genomic distances (up to a constant factor), these two measures of evolutionary remoteness of genomes are not as well-correlated as one's intuition may suggest.

Invited

The combinatorics of the breakage fusion bridge and its application to cancer genomics
Vineet Bafna
University of California, San Diego

The breakage-fusion-bridge (BFB) mechanism was proposed over seven decades ago and is a source of genomic variability and gene amplification in cancer. Here we formally model and analyze the BFB mechanism, to our knowledge the first time this has been undertaken. We show that BFB can be modeled as successive inverted prefix duplications of a string. Using this model, we show that BFB can achieve a surprisingly broad range of amplification patterns. We find that a sequence of BFB operations can be found that nearly fits most patterns of copy number increases along a chromosome. We conclude that this limits the usefulness of methods like array CGH for detecting BFB.

Invited

Following Dobzhansky: Extending The Reach of Phylogenies

Bernard Moret
École Polytechnique Fédérale de Lausanne

One of the most cited articles in biology is a 1973 piece by Theodosius Dobzhansky in the The American Biology Teacher entitled "Nothing in biology makes sense except in the light of evolution." Since then, phylogenetic tools have seen fast increasing use throughout biological and biomedical research – over 10,000 citations to such tools appear every year. Yet, in spite of the fame of the article and the spread of phylogenetic analysis, the use of methods grounded in evolutionary biology is not nearly as pervasive as it ought to be. We illustrate some of the potential of phylogenetic approaches through two projects carried out in our laboratory. The first, phylogenetic transfer of knowledge, leverages known phylogenetic relationships to improve inference of data about modern structures. We have successfully applied our ProPhyC model to the refinement of regulatory networks and are currently using it for the refinement and prediction of protein contact networks in protein complexes. The second is a two-level phylogenetic model and associated inference algorithm for the evolution of isoforms in alternative transcription. We have applied our TrEvoR tool to the entire ASPIC database of alternative transcripts, resulting in much enhanced accuracy in the classification of transcripts. Our contention is that these are but two of the numerous further applications of phylogenetic methods in the life sciences, applications that will require both modelling and algorithmic research.

Open Problem talk

Algorithms for constructing k-articulated network: a more powerful classification of phylogenetic networks

*S.M. Yiu and Francis Y.L. Chin
The University of Hong Kong

To capture the evolutionary history of a set of species with the presence of articulation events such as horizontal gene transfer of a set of species, it is well-known that using a phylogenetic network is more appropriate. A common classification for phylogenetic networks is called level-x networks. However, existing algorithms for constructing a level-x network from a given set of trees are all exponential in x. The bad news is that for viruses, since the mutation rate is high, the evolutionary history can only be modeled by a high-level network, say $x = 4$. Thus, the algorithms are not efficient enough to solve the problem. We propose a new classification of phylogenetic network, called k-articulated network. The merit of this classification is that every level-x network is also an x-articulated network while some level-x network can be modeled by a k-articulated network with $k < x$. Thus, even solving the problem for k-articulated network with k as small as 2, some of the meaningful high level networks can be constructed efficiently. In this article, we highlight some of the interesting and important open problems related to constructing a k-articulated network from a given set of trees.

Open Problem talk

Phylogenetic tree reconstruction with protein linkage

*Francis Chin and S.M. Yiu
The University of Hong Kong

Phylogenetic tree reconstruction for a set of species is an important problem for understanding the evolutionary history of the species. When reconstructing a phylogenetic tree, one common representation for a species is a binary string indicating the existence of some selected genes/proteins. Up until now, all existing reconstruction methods have assumed the existence of these genes/proteins to be independent. However, in most cases, this assumption is not valid. From an evolutionary point of view, some functionally dependent proteins should usually be present (or absent if the function is no longer needed) in the same generation. The Phylogenetic Tree Reconstruction with Protein Linkage (*PTRPL*) problem considers the reconstruction problem by taking into account the dependency of proteins, i.e. protein linkage, in addition to the hamming distance that represents the number of insertions/deletions of genes/proteins.

Research talk

Reconciliation of Gene and Species Trees with Polytomies

*Yu Zheng, Taoyang Wu, and Louxin Zhang
National University of Singapore

Millions of genes in the modern species belong to only thousands of gene families. A gene family includes instances of the same gene in different species and duplicate genes in the same species. Two genes in different species are ortholog if they diverged when the most recent common ancestor of the species speciated. Orthologs are used to infer signaling pathway evolution and correspondence between genotype and phenotype and hence ortholog identification is a basic task in comparative genomics. Because of complex gene evolutionary history, however, ortholog identification is extremely difficult. One key method for it is to use an explicit model of the evolutionary history of the genes subject to study, called the gene (family) tree. It compares the gene tree with the evolutionary history of the species in which the genes reside, called the species tree, using a procedure known as the tree reconciliation. Tree reconciliation presents challenging problems when species trees are not binary in practice. Here, non-binary gene and species tree reconciliation is studied in a binary refinement model, which unifies gene duplication inference through tree reconciliation with reconstruction of species tree from gene trees.

The problem of reconciling gene and species trees is proved NP-hard when the input species tree is not binary even for the duplication cost. We then present the first efficient method for reconciling a non-binary gene tree and a non-binary species tree. The method attempts to find a binary refinement of the given gene and species trees that minimizes the given reconciliation cost if they are not binary. Our algorithms have been implemented into an automatic tool for inferring gene duplication events through tree reconciliation and for reconstructing species tree from gene trees. The tool supports quick automated analysis of large data sets.

Invited

Alignment-Free Local Sequence Comparison: Statistics and Algorithms

Michael Waterman
University of Southern California

The sum of the products of the k-word counts from each of two sequences (D_2) has been used to test if the sequences have a significant similarity. Important for large-scale applications, the statistic can be rapidly computed. In this talk some known results will be summarized where superior statistics (D_2^*) are available. Local versions of the statistics are considered in this talk. Although local behaves well statistically, a naive linear comparison increases the time to quadratic. We designed a geometric framework for identifying pairs of sequence windows that maximize the dot product. The D_2 and D_2^* measures are similar to cosine similarity but with the additional information using the norm of the word count vectors, rather than only the angle between them. Our framework discretizes the norms of vectors, and transforms our dot-product similarity into Euclidean distance, but at loss of accuracy. We bound the error our transformation introduces and relate this error to the efficiency of algorithms based on this transformation.

This is joint work with Fengzhu Sun, Ehsan Behmanghader, and Andrew D. Smith.

Research talk

Alignment Plots: Local Sequence Comparison in Sliding Windows

Peter Krusche and *Alexander Tiskin
University of Warwick

Numerous methods of local sequence comparison have been proposed in the past. We introduce a new algorithmic technique that allows highly detailed, loss-free local comparison via the computation of an *alignment plot*, i.e. an exhaustive set of alignments for pairs of fixed-length windows in both sequences. Our approach improves on a number of existing local sequence comparison algorithms in terms of accuracy and flexibility. The underlying algorithmic technique is based on a deep and surprising connection between string comparison and algebraic semigroup theory. An implementation of the alignment plot and related methods has been developed for a recent study of evolutionary conservation in DNA sequences. A similar approach has also been used recently as a key subroutine in a software tool finding approximate repeats in DNA sequences. We believe that the alignment plot method has the potential of gaining wider use in applications that currently employ either lossy techniques, or less efficient loss-free methods of local alignment.

Invited

New (and old) problems in phylogeny and multiple sequence alignment estimation

Tandy Warnow
Department of Computer Sciences, The University of Texas at Austin

Molecular sequences evolve under processes that include substitutions, insertions, and deletions (jointly called "indels"), as well as other mechanisms (e.g., duplications and rearrangements). The inference of the evolutionary history of these sequences has thus been performed in two stages: the first estimates the alignment on the sequences, and the second estimates the tree given that alignment. While such methods seem to work well on relatively small datasets, these two-stage approaches can produce highly incorrect trees and alignments when applied to large datasets, or ones that evolve with many indels. With the advent of next generation sequencing technologies, which produce extremely large datasets, many of which contain fragmentary sequences, phylogeny and alignment estimation has become even more challenging.

In this talk, I will present three methods, SATe, DACTAL, and SEPP for estimating phylogenies for large datasets. Each of these methods employs divide-and-conquer (and in some cases also iteration) in order to obtain highly accurate alignments and/or phylogenies for large datasets. SATe, or Simultaneous Alignment and Tree Estimation (Liu et al., Science 2009 and Systematic Biology 2012), iterates between alignment estimation and tree estimation, with each iteration beginning with a new tree, re-aligning on that tree (using a novel divide-and-conquer technique), and then recomputing a maximum likelihood tree on the new alignment. SATe has been shown to produce highly accurate alignments and trees on datasets with up to tens of thousands of sequences. DACTAL, which stands for Divide-and-Conquer Trees (almost) without ALignments (Nelesen et al., ISMB 2012 and Bioinformatics 2012) estimates trees by computing alignments and trees on carefully selected small subsets of the taxa, and combines the trees using a new supertree method developed in my lab. DACTAL matches the accuracy of SATe without needing an alignment on the entire dataset. Finally, SEPP, which stands for SATe-enabled phylogenetic placement (Mirarab et al., PSB 2012), can compute alignments and trees on large datasets containing fragmentary sequences, and achieves improved accuracy compared to other phylogenetic placement methods.

Open problems in phylogenetic estimation will also be discussed, focusing on the challenges caused by that model violations (i.e., biological data do not evolve under the simple models that are used in phylogeny estimation).

Tuesday August 28, 2012

Invited

Three Optimization Problems in Molecular Biology and Genetics
Richard M. Karp
Simons Institute for the Theory of Computing
University of California, Berkeley

We present statistical evidence that, within the protein interaction networks of *H. sapiens* and *D. melanogaster*, conserved connected subgraphs with low conductance tend to have high functional coherence. We present efficient algorithms for finding such subgraphs (joint work with Luqman Hodgkinson).

A protein regulatory network can be described by an acyclic digraph in which each node corresponds to a protein that may be in either an active or an inactive state, depending on the states of the nodes immediately preceding it. Given the graph-theoretic structure of the network and examples of its input-output behavior under various perturbations of the states of some of its proteins, we use integer programming to describe the dependencies at the nodes by Boolean functions that come closest to producing the observed input-output behavior (joint work with Roded Sharan).

Given a set of coins with unknown heads probabilities we consider the problem of performing a sequence of coin tosses to efficiently identify the coin with highest heads probability. This is a simplified variant of the following problem in genetics: given a set of mutations that may be associated with a disease, select a sequence of experiments, each of which observes the state of a mutation in a case and a control, to efficiently identify the mutation most associated with the disease (joint work with Karthik Chandrasekaran).

Invited

Dissecting inner structure in disease regulatory networks using differential co-expression

Ron Shamir
School of Computer Science, Tel Aviv University

Novel approaches to gene expression analysis seek differential co-expression patterns, wherein the level of co-expression of a particular set of genes differs markedly between disease and control samples. Such patterns can arise from a disease-related change in the regulatory mechanism governing that set of genes. Here we present a new method for detecting differentially co-expressed gene sets. We introduce a novel probabilistic score for differential correlation, and use it to detect pairs of modules whose intra-module correlation is consistently high but whose inter-module correlation differs markedly between disease and normal samples. The algorithm outperforms the state of the art methods in terms of significance and interpretability of the detected gene sets. Moreover, the discovered gene sets are enriched with disease-specific microRNA families. In a case study on Alzheimer's disease, the method dissected biological processes into functional subunits that are differentially co-expressed, thereby revealing inner structures in disease regulatory networks.

Joint work with David Amar and Hershel Safer.

Open Problem talk

Metabolic stories: uncovering all possible scenarios for interpreting metabolomics data

*Paulo Vieira Milreu[1], Andrea Marino[2,5], Vicente Acuña[1,2], Etienne Birmelé[1,2,3], Ludovic Cottret[4], Fabien Jourdan[8], Vincent Lacroix[1,2], Alberto Marchetti-Spaccamela[6], Leen Stougie[7], Pierluigi Crescenzi[5], and Marie-France Sagot[1,2]

[1]*Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, LBBE, Villeurbanne, France*

[2]*INRIA Rhône-Alpes, 38330. Montbonnot Saint-Martin, France*

[3]*Lab. Statistique et Génome, CNRS UMR8071 INRA1152, Université d'Évry, France*

[4]*Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés (LISBP), UMR CNRS 5504 - INRA 792, Toulouse, France*

[5]*Università di Firenze, Dipartimento di Sistemi e Informatica, I-50134 Firenze, Italy*

[6]*Sapienza University of Rome, Italy*

[7]*VU University and CWI, Amsterdam, The Netherlands*

[8]*INRA UMR1331 - Toxalim, Toulouse, France*

A constrained version of the problem of enumerating all maximal directed acyclic subgraphs of a graph can be used to interpret metabolomics data, since they may correspond to possible explanations of the experimental evidence. Despite some algorithmical results already achieved, there are several interesting open problems from both the modeling and the theoretical points of view.

Invited

The Linkage Disequilibrium Measures Unification Problem

Sorin Istrail
Department of Computer Science, Brown University

In this talk we will present several problems related to SNPs and haplotypes genetic variation. Linkage disequilibrium (LD) is the occurrence of some combinations of alleles or genetic markers in a population more often or less often than would be expected from a random formation of haplotypes from alleles based on their frequencies. Linkage disequilibrium measures are of fundamental importance for the analysis of the empirical patterns of genetic variation in populations of individuals, although more than half a century of statistical genetics did not reached a consensus on how to measure LD. A classic survey on the topic includes in the title "proceed with caution" as a warning of the difficulties associated with this apparently non-quantitative phenomenon. Various axioms were proposed in the literature but the development of measures simultaneously satisfying desirable axioms is very limited. We will discuss two axioms related to the "curse of the pairwise" and the "interpretability of intermediate values" especially of interest for measuring LD in the context of GWAS analysis. We will present our construction of a new LD measure consistent with both axioms and an algorithm for its computation. The new measure is a "combinatorial square-root" of the LD measure r^2. We will discuss other axioms and challenges in making progress on discovering unifying principles for LD measures. Two other problems for haplotype phasing and tagging SNPs selection will be presented aiming also at consolidating and hopefully unifying collections of diverse methods.

Open Problem talk

Reconstructing Pedigrees from Data
Bonnie Kirkpatrick
University of British Columbia

Pedigrees, colloquially family trees, are of interest both to statisticians and to computer scientists. Pedigree graphical models were some of the early examples of graphical models and helped inspire the sum-product algorithm. Statisticians have long been interested in calculations on pedigrees while interest by computer scientists has been more recent. Pedigree analysis is useful for discovering the fine-scale structure of recombination maps for humans, discovering regions linked to complex diseases, discovering regions linked to rare Mendelian diseases, and finding insights into the relationship between fertility and cystic fibrosis.

The problem of pedigree reconstruction is motivated by the time that it takes to survey research participants in order to construct the pedigree for their family. It would be much easier if we could collect the genotype or sequence data of individuals and trust computational methods to piece together the correct pedigree relating the individuals. To this end, we review the problem of pedigree reconstruction, first in the general form and then in a slightly more restricted form.

Research talk

Analysis of single-cell RNA-seq data using probabilistic mixture models
Peter Kharchenko
Harvard Medical School

A variety of modern high-throughput assays are aimed at dissecting the functional and physical state of the cell. Some methods, such as analysis of RNA sequences (RNA-seq), can achieve the sensitivity required to interrogate individual cells, facilitating studies of complex tissues specialized cellular environments. The analysis of such measurements, however, is complicated by high levels of technical noise and intrinsic biological variability. To address this challenge in the context of differential expression analysis, we developed a probabilistic approach that takes into account stochastic variability and expression magnitude distortions typical of single-cell amplificates. This Bayesian approach estimates the posterior probability distribution of the expected expression magnitude for a given gene, allowing us to detect differential expression signatures and classify individual cells in a way that is tolerant of stochastic noise and systematic biases.

Invited

Efficient communication and storage vs. accurate variant calls in high throughput sequencing: two sides of the same coin

Cenk Sahinalp
Simon Fraser University

Given two strings A and B from the DNA alphabet, the Levenshtein edit distance between A and B, LED(A, B), is defined to be the minimum number of single character insertions, deletions and replacements to transform A to B (equivalently B to A). If in addition to the single character edits, one is permitted to perform segmental (block) edits in the form of (i) moving a block from any location to another, (ii) copying a block to any location, and (iii) uncopying (i.e. deleting one of the two occurrences of) a block, the resulting "block edit distance", BED(A, B), captures much of our current understanding of the relation between individual genome sequences. If among two communicating parties, Alice (holding genome sequence A) and Bob (holding genome sequence B), Alice wants to compute B, then, theoretically, the total number of bits Bob needs to send to Alice is $\tilde{O}(\text{BED}(A, B) \cdot \log \text{BED}(A, B))$ [Cormode et al., SODA 2000]. Considering that between a typical donor genome B and a reference genome A, the number of single character differences are in the order of a few million and the number of structural (i.e. blockwise) differences are in the order of tens of thousands, it should be possible to communicate genomes by exchanging only a few million bytes! Yet, today, the most effective way of communicating genome sequence data involves physically exchanging hard disks.

In this talk we will try to explain the wide gap between theoretical expectations and the current reality in genome communication, as well as storage, and pose some theoretical and practical problems on the way to the Google-ization of genome search and analysis. We will also try to explore the extent our theoretical predictions for genome sequences hold for the RNA-Seq data. Finally we will briefly go through some of the recent developments in transcriptome sequence analysis, especially in the context of disease studies.

Contributed

DTRA Algorithm Prize
Christian Whitchurch
Defense Threat Reduction Agency

As nth generation DNA sequencing technology moves out of the research lab and closer to the diagnostician's desktop, the process bottleneck will quickly become information processing. The Defense Threat Reduction Agency (DTRA) and the Department of Defense are interested in averting this logjam by fostering the development of new diagnostic algorithms capable of processing sequence data rapidly in a realistic, moderate-to-low resource setting. With this goal in mind, DTRA is sponsoring an algorithm development prize.

The Challenge: Given raw sequence read data from a complex diagnostic sample, what algorithm can most rapidly and accurately characterize the sample, with the least computational overhead?

The Stakes: $1,000,000

Prize details and sequencing datasets will be made available this Fall.
Monitor http://www.dtra.mil/Business.aspx for updates on this program.

Panel

Panel discussion of problem formulation

Wednesday August 29, 2012

Invited

Lasting relations
Marie-France Sagot
INRIA, Université Lyon 1

Whether at the level of molecules, cells, organisms or species, the mechanisms underlying any functional aspect of living systems, whether in sickness or health, is almost never the affair of a single entity, but instead the affair of interactions. Many of those may involve different biological species in a close and often lasting relationship. Mitochondria and chloroplasts, which originated from bacteria who went living within the cells of other organisms, may be seen as an extreme example of such relationships, in this case leading to what may be seen as a loss of identity as an organism. More in general, the idea of individuals of a given species — this includes humans — as "supraorganisms" with an internal ecosystem of diverse other species has been advanced — indeed, humans are thought to contain 10 times more bacterial cells than human. It is also estimated that some 50% of all species parasitise another and that close to 100% of all plants and animals are parasitised as individuals, in general by more than one species.

Such pervasive and lasting relations will be the topic explored in this talk, while discussing also of another relationship, the one motivating this satellite meeting, namely the relationship between computer science (and mathematics) on one side, and biology on the other.

Invited

Sequencing antibiotics: bioinformatics meets nuclear physics

Pavel Pevzner
University of California, San Diego

Proliferation of drug-resistant diseases raises the challenge of searching for new, more efficient antibiotics. However, sequencing peptide antibiotics, once a heroic effort, remains time-consuming and error-prone.

Many antibiotics such as daptomycin or vancomycin (the modern antibiotics of last resort), represent cyclic Non-Ribosomal Peptides (NRPs) that do not follow the central dogma "DNA produces RNA produces Protein". They are assembled by nonribosomal peptide synthetases that represent both the mRNA-free template and building machinery for the peptide biosynthesis.

Thus, NRPs are not directly inscribed in genomes and cannot be inferred with traditional DNA sequencing. NRPs are of great pharmacological importance as they have been optimized by evolution for chemical defense and communication. NRPs include antibiotics, antitumor agents, immunosuppressors, toxins, and many peptides with still unknown functions.

Most NRPs contain nonstandard amino acids that are notoriously difficult to sequence. Moreover, the dominant technique for sequencing antibiotics (NMR) requires large amounts (milligrams) of highly purified materials that, for most compounds, are nearly impossible to obtain. Therefore, there is a need for NRPs sequencing by tandem mass spectrometry from picograms of material. Since nearly all NRPs are produced as related analogs by the same microorganism, we develop a mass spectrometry approach for sequencing all related peptides at once (in difference from the existing approach that analyzes individual peptides). Our results suggest that instead of attempting to isolate and NMR-sequence the most abundant compound, one should acquire spectra of many related compounds and sequence all of them at once using mass spectrometry. We illustrate applications of this approach by sequencing new variants of antibiotics from Bacillus brevis as well as sequencing a previously unknown family of NRPs (named Reginamides) which are isolated from a bacterial strain that produce natural products with anti-asthma activities.

This is a joint work with Hosein Mohimani and Pieter Dorrestein (UCSD).

Open Problem talk

Combinatorial problems for SNP and mutation discovery using base-specific cleavage and mass spectrometry

*Xin Chen[1], Qiong Wu[1], Ruimin Sun[1], and Louxin Zhang[2]
[1]*School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore*
[2]*Department of Mathematics, National University of Singapore, Singapore*

Single nucleotide polymorphisms (SNPs) and mutations are among the most important genetic factors that contribute to human evolution, disease and biological functions. One approach used to discover these sequence variations is based on base-specific cleavage and matrix assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry. To aid in developing an efficient computational solution to automate this discovery process, we formulate four combinatorial optimization problems. They are all aimed at reconstructing an unknown sample sequence by minimizing either Hamming distance or edit distance to a given reference sequence. However, they are defined to search over different candidate sequence spaces, either in the space where the in silico predicted mass spectra have all their signals contained in the measured mass spectra or in the space where the in silico predicted mass spectra contain all the signals of the measured mass spectra. In our preliminary studies, three problems are shown to be NP-hard. In addition, we designed the exact dynamic programming algorithms for solving two problems, although they may adversely run in exponential time in the worst case.

Invited

Evolution of bacterial genomes
Mikhail Gelfand
Russian Academy of Sciences and Moscow State University

I shall talk about several ongoing studies on the evolution of bacterial genomes. These include evolution of pan-genomes, genome rearrangements, homologous recombination, gene strand preference, operon structure and horizontal gene transfer. The projects are at different states of completion, and, while there already are concrete results, I shall attempt to present a research program rather that detailed specifics.

Invited

A Moving Landscape for Comparative Genomics in Mammals

Steve O'Brien
Theodosius Dobzhansky Center for Genome Bioinformatics and Saint Petersburg University

Recent completion and inspection of whole genome sequence and assembly for over fity species of mammals, from platypus to panda to human, offer the prospect of a better view of the patterns of change within genome organization across the mammalian radiations. With an international community we have worked to sequence and assemble a deep (14x) coverage of the genome of domestic cat. The annotation reveals a rich assemblage of genes, repeats, SNPs, Numt, STRs, ERVs and countess genomic acronyms now quantified in that species and compared to other mammalian genomes. The genomics resource in cats has already produced a number of gene function insights as well as an archaeological view of the history of cat domestication. In addition, my colleagues and I have created *Genome-10K*, an international consortium of scientist who have set a goal of gathering, sequencing, assembling, and annotating to high quality some 10,000 vertebrate genomes with 2nd and 3rd generation sequencing technology within the coming five years. These activities provide an enormous BioInformatics challenge whose solution will provide future zoologists of every persuasion a genome sequence resource for their favourite study animal. The applications and potential for the genome sequence in several research questions will be discussed.

Tuesday August 28, 2012 Poster Session

Research poster

YOABS: Yet Other Aligner of Biological Sequences — the first $O(n)$ alternative to the Smith-Waterman algorithm

Vitaly Galinsky
University of California, San Diego

This presentation outlines a new long alignment algorithm — Yet Other Aligner of Biological Sequences (YOABS) — an efficient algorithm (both memory- and performance-wise) for aligning several hundred or more base pairs query sequence to a long reference genome. It has high sensitivity and specificity (especially given a long query or a query with low error rate). The accuracy of YOABS is comparable with the most accurate long sequence aligners so far (e.g. BWA-SW or SSAHA2). By design YOABS is well suited to detect arbitrary gaps and chimeras, therefore it can be used to facilitate detection of structural variations or reference misassemblies.

In contrast to the majority of long sequence alignment algorithms YOABS does not use the seed-and-extend paradigm. Instead it records all local hits between all $l(\text{prefix}) + m(\text{suffix})$ base pairs index entries for the reference sequence (organized as a forward and a backward tries) and all $l + m$ base pairs subsequences (including l and $2l$ gapped) of the query sequence. The local hits are stored as a table of l-scaled *modulo* $2k$ query subtracted location in the reference versus *modulo* l location in the query. The algorithm avoids using the expensive dynamic programming stage (the Smith-Waterman algorithm) altogether replacing it with linearly scaling procedure.

Open Problem poster

Registration and analysis of array images of general layout

Vitaly Galinsky
University of California, San Diego

Development of algorithms suitable for registration and data analysis of array images of general layout capable to work with data produced by different array technologies using different image acquisition platforms represents substantial challenge. In order to be able to process huge and constantly growing amount of array data the algorithms used should be as efficient as possible, scaling at least not faster than linearly with the number of features M, that is the time complexity of the algorithms should not exceed $O(M)$. The algorithms should also be robust and allow processing with minimal human intervention array images with different parameters of grids, including rectangular and hexagonal grids and even grids composed from smaller subgrids, as well as with different sizes and shapes of spots, including spots with circular and rectangular shapes.

This presentation discusses a possibility of extension of powerful flexible grid registration algorithm to analysis of array images of general layout that will allow it to work with different technologies. The preliminary results of application of this flexible registration and analysis algorithm to several vastly different micro and nano array technologies used for genotyping, gene expression profiling, sequencing and optical mapping show good promise for coping with all the above challenges.

Open Problem poster

The problems of reconciling gene and species trees, mapping a gene tree into a species tree and gene tree inference

*Konstantin Gorbunov and Vassily Lyubetsky
Institute for Information Transmission Problems of the Russian Academy of Sciences

A long recognized problem is inference of a tree S that amalgamates a set of input gene trees. We further developed a traditional approach to find the tree S such that it minimizes the total cost (gene duplications and losses) of mappings of individual gene trees into S. An algorithm is novel mathematically correct and possesses the cubic running time in n and in m, where n is the number of gene trees, and m is the total number of species. Is a correct inference of the tree S possible in polynomial time with the horizontal gene transfer events? Is a correct inference of a phylogenetic net S (instead of a species tree S) possible in polynomial time at least with gene duplication and loss events?

We suggested a novel mathematically correct algorithm to map (reconcile) a gene tree into S (with time slices) that possesses the cubic running time in $|S|$. Could one does the same for phylogenetic nets?

We suggested a novel definition of a gene tree into S mapping (an evolutionary scenario) for the joint case of duplications, losses, horizontal gene transfers, gains, etc. How to define the embedding taking into account dynamics of molecular sequences?

We suggested a novel heuristic algorithm to reconstruct a gene tree on the base of alignment. Is there such mathematically correct algorithm?

Open Problem poster

Finding a Combinatorially Optimal Model for RNA Folding Pathways

*Bonnie Kirkpatrick and Anne Condon
University of British Columbia

RNA secondary structure is well explored, but there is less algorithmic work in the area of RNA folding pathways. RNA sequences have secondary structures comprised of a set of allowed base pairs. Given a particular RNA sequence, there are polynomial-time algorithms for computing both the minimum free energy (MFE) secondary structure and for enumerating over the possible secondary structures to compute the partition function. Despite the success of research into secondary structure, RNA folding pathways are less well understood; this is the dynamic behavior of the RNA molecule as the secondary structure changes over time.

RNA is known to have a role in DNA transcription, splicing, translation of mRNA, gene regulation, and other critical cellular functions. Understanding the folding pathways of RNA is key to decoding how RNA perform their functions.

Given an RNA sequence of length n, the set of possible secondary structures, U, has size $O(3^n)$. Having the possible secondary structures as its state space, computations on the full RNA folding model are intractable. The problem posed here: given fixed computational resources that allow computation on $|S|$ secondary structures, choose a subset, $S \subseteq U$, of the possible secondary structures that gives the optimal approximation to the RNA folding process and that satisfy certain technical constraints.

Research poster

Non-Identifiable Pedigrees and a Bayesian Solution
Bonnie Kirkpatrick
University of British Columbia

Some methods aim to correct or test for relationships or to reconstruct the pedigree, or family tree. These methods use a structured machine learning approach where the objective is to find the pedigree graph that maximizes the likelihood which is the probability of inheriting the data under the pedigree model. We show that these methods cannot resolve ties for correct relationships due to identifiability of the pedigree likelihood. This means that no likelihood-based method can produce a correct pedigree inference with high probability. This lack of reliability is critical both for health and forensics applications.

We present the first discussion of multiple typed individuals in non-isomorphic pedigrees where the likelihoods are non-identifiable. While there were previously known non-identifiable pairs, we give an example having data for multiple individuals.

Additionally, deeper understanding of the general discrete structures driving these non-indentifiability examples has been provided, as well as results to guide algorithms that wish to examine only identifiable pedigrees. We introduce a general criteria for establishing whether a pair of pedigrees is non-identifiable and two easy-to-compute criteria guaranteeing identifiability. Finally, we suggest a method for dealing with non-identifiable likelihoods: use Bayes rule to obtain the posterior from the likelihood and prior. We propose a prior guaranteeing that the posterior distinguishes all pairs of pedigrees.

Open Problem poster

Open Problems in Metagenomics

Gabriel Valiente
Technical University of Catalonia

Next-generation sequencing technologies allow for the genetic study of complex microbial communities, which were so far largely unknown because they cannot be cultured in the laboratory. The core problem of metagenomics is to determine and quantify the composition of a sample consisting of a mixture of different, and possibly unknown, microbial species. Solving this core biological problem involves a series of algorithmic and computational problems, ranging from the simulation of metagenomic samples to the alignment or mapping of sequence reads, the non-taxonomic assignment or binning of sequence reads, and the taxonomic assignment of sequence reads. We give a detailed definition of some of these algorithmic and computational problems.