
Bioinformatics
(Org: **Miklos Csuros** (University of Montreal))

CEDRIC CHAUVE, Simon Fraser University

Mapping ancestral genomes with massive gene loss: a matrix sandwich problem

Reconstructing ancestral genomes generally requires markers that are present in all extant descendants. Motivated by the reconstruction of plant and vertebrate genomes, we introduce a new method that handles non-universal markers, by combining the consecutive-ones property with the classical notion of sandwich graph. This leads to a notion of sandwich consecutive-ones matrix. We initiate a study of this problem and apply our results on real datasets. Collaboration with Eric Tannier, Jerome Salse and Haris Gavranovic.

LUCIAN ILIE, University of Western Ontario

Fast combinatorial computation of good seeds for genome alignment

Aligning entire genomes requires high sensitivity and speed. Whereas other methods trade one for the other, multiple spaced seeds increase both. They have many applications to homology search, alignment of next generation sequencing reads, oligonucleotide design, etc. Finding optimal seeds is a hard problem and even heuristic algorithms are exponential. We present the only polynomial-time algorithm, whose ideas stemmed from stringology. It is much faster and computes better seeds than all the other programs.

BIN MA, University of Waterloo

Algorithms for Protein/Peptide Sequencing with Mass Spectrometry

The talk reviews some computational problems arising in mass spectrometry data analysis for protein identification and sequencing. In particular, the peptide de novo sequencing problem and its algorithm will be introduced. The possible sequencing errors also affect the traditional sequence homology search model (the Smith-Waterman algorithm). In the talk we will also examine such effect and introduce a new homology search model that tolerates both the de novo sequencing errors and the evolutionary mutations.

PAUL MEDVEDEV, University of California, San Diego

Paired de Bruijn Graphs: a Novel Approach for Incorporating Mate Pair Information into Genome Assemblers

Genome assembly can be viewed as a problem of reconstructing contiguous portions of a string (contigs) given a random sampling of its substrings. Algorithms based on deBruijn graphs are some of the most powerful and practical methods for assembly. Mate pairs, which are pairs of reads between which the distances are approximately known, have been algorithmically incorporated into most assemblers as various heuristic post-processing steps to correct the deBruijn graph or to link contigs into scaffolds. Such methods have allowed the identification of longer contigs in the presence of long repeating regions; however, they can still fail to resolve complex repeats.

Here, we introduce the paired de Bruijn graph, a generalization of the deBruijn graph that incorporates mate pair information into the graph structure itself instead of analyzing mate pairs at a post-processing step. This graph has the potential to be used in place of the deBruijn graph in any deBruijn graph based assembler, maintaining all other assembly steps such as error-correction and repeat resolution.

This is joint work with SonPham, MarkChaisson, GlennTesler, and PavelPevzner.

JURAJ STACHO, Caesarea Rothschild Institute, University of Haifa

Unique perfect phylogeny is NP-hard

We answer, in the affirmative, the following question proposed in 1992 by Mike Steel as a \$100 challenge: "*Is the following problem NP-hard? Given a ternary phylogenetic X -tree \mathcal{T} and a collection \mathcal{Q} of quartet subtrees on X , is \mathcal{T} the only tree that displays \mathcal{Q} ?*" As a particular consequence of this, we show that the unique chordal sandwich problem is also NP-hard.

Joint work with Michel Habib.