CGCG

Comparative Genomics in Campo Grande

a full day of Comparative Genomics with
professors and researchers from Germany and Brazil

December 19th

Sala de Videoconferência, Faculdade de Computação, UFMS

Program

- 09:00–09:10 Welcome
- 09:10–09:50 The *Xylella* genome project: First Brazilian whole-genome sequencing project
  João Meidanis, IC–UNICAMP
- 09:50–10:30 Computing the rearrangement distance of natural genomes
  Marília Dias Vieira Braga, TechFak–Universität Bielefeld
- 10:30–11:00 Coffee break
- 11:00–11:40 TBA
  Luis Antonio Kowada, IC–UFF
- 11:40–12:20 Finding all maximal perfect haplotype blocks in linear time
  Jens Stoye, TechFak–Universität Bielefeld
- 12:20–14:30 Lunch break
- 14:30–15:10 Hierarchical organization of syntenic blocks in large genomic datasets
  Daniel Dörr, TechFak–Universität Bielefeld
- 15:10–15:50 Counting scenarios, intermediate genomes, and dealing with missing genes with the rank distance
  João Meidanis, IC–UNICAMP
- 15:50–17:00 Coffee and discussion

Abstracts

- Title: The *Xylella* genome project: First Brazilian whole-genome sequencing project
- Speaker: João Meidanis, IC–UNICAMP
- Abstract: Next year we celebrate 20 years of the publication of the complete genome sequence of bacterium *Xylella fastidiosa*, the first phytopathogen ever sequenced, a feat acomplished by a Brazilian network of 35 labs, including biological and computational teams. For many participants a turning point in their careers, this project positioned Brazil among the selected group of nations having finished a free-living organism's genome. This talk highlights the best and worst moments of the project, by an insider.

- Title: Computing the rearrangement distance of natural genomes
- Speaker: Marília Dias Vieira Braga, TechFak–Universität Bielefeld
- Abstract: The computation of genomic distances has been a very active field of computational comparative genomics over the last 25 years. Substantial results include the polynomial-time computability of the inversion distance by Hannenhalli and Pevzner in 1995 and the introduction of the double-cut and join (DCJ) distance by Yancopoulos, Attie and Friedberg in 2005. Both results, however, rely on the assumption that the genomes under comparison contain the same set of unique markers (syntenic genomic regions, sometimes also referred to as genes). In 2015, Shao, Lin and Moret relax this condition by allowing for duplicate markers in the analysis. This generalized version of the genomic distance problem is NP-hard, and they give an ILP solution that is efficient enough to be applied to real-world datasets. A restriction of their approach is that it can be applied only to balanced genomes, that have equal numbers of duplicates of any marker. Therefore it still needs a delicate preprocessing of the input data in which excessive copies of unbalanced markers have to be removed. In this talk we present an algorithm solving the genomic distance problem for natural genomes, in which any marker may occur an arbitrary number of times. Our method is based on a new graph data structure, the multi-relational diagram, that allows an elegant extension of the ILP by Shao, Lin and Moret to count runs of markers that are under- or over-represented in one genome with respect to the other and need to be inserted or deleted, respectively. With this extension, previous restrictions on the genome configurations are lifted, for the first time enabling an uncompromising rearrangement analysis. Any marker sequence can directly be used for the distance calculation. The evaluation of our approach shows that it can be used to analyze genomes with up to a few ten thousand markers, which we demonstrate on simulated and real data. (*joint work with Leonard Bohnenkämper, Daniel Doerr, and Jens Stoye*)

- Title: TBA
- Speaker: Luis Antonio Kowada, IC–UFF
- Abstract: TBA

- Title: Finding all maximal perfect haplotype blocks in linear time
- Speaker: Jens Stoye, TechFak–Universität Bielefeld
- Abstract: Recent large-scale community sequencing efforts allow at an unprecedented level of detail the identification of genomic regions that show signatures of natural selection. Traditional methods for identifying such regions from individuals' haplotype data, however, require excessive computing times and therefore are not applicable to current datasets. In 2019, Cunha et al. (Proceedings of BSB 2019) suggested the maximal perfect haplotype block as a very simple combinatorial pattern, forming the basis of a new method to perform rapid genome-wide selection scans. The algorithm they presented for identifying these blocks, however, had a worst-case running time quadratic in the genome length. It was posed as an open problem whether an optimal, linear-time algorithm exists. In this paper we give two algorithms that achieve this time bound, one conceptually very simple one using suffix trees and a second one using the positional Burrows-Wheeler Transform, that is very efficient also in practice. (*This is joint work with Jarno Alanko, Hideo Bannai, Bastien Cazaux and Pierre Peterlongo.*)

- Title: Hierarchical organization of syntenic blocks in large genomic datasets
- Speaker: Daniel Dörr, TechFak–Universität Bielefeld
- Abstract: Many methods in comparative genomics require prior identification of syntenic blocks, making synteny deduction a fundamental task in genomic analyses. Yet, a recent study suggests that different synteny tools yield very different collections of syntenic blocks on the same data. We present a new method for constructing synteny hierarchies with measurable objectives that support analyses in comparative genomics on multiple levels of granularity. This method is part of a new workflow that facilitates the identification of synteny blocks from raw genomic sequences, requiring no assumptions on genes and gene families (although its pipeline can also be entered at a later stage with predefined markers). Our method is robust against segmental duplications, insertions and deletions of one or few markers.

- Title: Counting scenarios, intermediate genomes, and dealing with missing genes with the rank distance
- Speaker: João Meidanis, IC–UNICAMP
- Abstract: The rank distance model, introduced by Zanetti et al. in 2016, represents genome rearrangements in multi-chromosomal genomes looking at them as matrices. In this talk, we show how to count the number of optimal sorting scenarios and the number of intermediate genomes between any two given genomes, under the rank distance. In addition, we'll show how to generalize the model, allowing for genomes with different gene content. We approach such a generalization from two different angles, both using the same representation of genomes, and leading to simple distance formulas and sorting algorithms for genomes with different gene contents, but without duplications.

Organization

Comparative Genomics TeAm in Campo Grande–CGTACG (FACOM–UFMS)

- Diego Padilha Rubert
- Francisco Eloi Soares Araujo
- Fábio Henrique Viduani Martinez