Universität Bielefeld

Computational Methods for the Analysis of the
Diversity and Dynamics of Genomes | GRK 1906

Application deadline: 8 May 2016 — start of projects: October 2016

For application instructions, see: `http://www.didy.uni-bielefeld.de`

Project Descriptions

Project 1: Applying succinct data structures to massive sequence clustering

Supervisor: Markus E. Nebel (Bielefeld University)

Background

The advent of high-throughput sequencing (HTS) technologies has drastically changed how research is conducted in the life sciences. Today, massive data sets comprising several million sequences provide the basis for new insights in various medical and biological disciplines. A central task in many bioinformatics pipelines is clustering in order to, e.g., denoise the data or assign sequences to biologically interpretable groups. Intense research on related problems such as similarity joins and approximate dictionary matching led to a variety of tools, most of which focus on speeding up the computation in order to handle the continuously growing data sets. However, they usually fall under the restriction that all data (structures) have to be held in the faster, but relatively small and expensive main memory (compared to disk space). While today even small laboratories can produce data on HTS-scale, they very often do not have the technical or financial capabilities to constantly upgrade their computational infrastructure in order to process these data sets. Cloud computing is an increasingly popular approach to relieve this problem, but the cost and time of transferring the data as well as concerns regarding data control and privacy can exclude this option. Therefore, the goal of this project is to facilitate the processing of HTS-scale data sets on current lab infrastructures by augmenting state-of-the-art clustering tools with succinct data structures.

Project Description

Within this project, we aim at examining the effect of introducing succinct data structures into clustering tools on their runtime and space consumption. Since succinct data structures often hide their dominant costs in lower-order terms and constants, we are particularly interested in the trade-off between space and runtime. More precisely, we will analyze the effect of relaxing the space restriction of succinct data structures and study how much additional space we have to spend on top of the information-theoretical minimum in order to obtain practically superior tools. To that end, existing data structures have to be adapted or fine-tuned, and, potentially, new advanced data structures have to be developed. While the focus will be on tools working with the commonly accepted edit distance, we will also consider alternative, biologically motivated notions of similarity based on, e.g., base-base correlations, chaos game representation or compression. These try to avoid some of the drawbacks and computational bottlenecks of the edit distance and some are based on dinucleotide (or even trinucleotide) composition, which is known to be biologically meaningful. Furthermore, we will investigate the accessibility of the used succinct data structures for practical optimizations involving, e.g., cache utilization and parallelization (broadword programming, GPU computations, threading etc.).

Required Skills

Strong background in algorithmics and data structures as well as programming skills (e.g. C++ or Java) are required. Knowledge in biology is beneficial.

Project 2: Biomedical network analysis using higher Petri nets

Supervisors: Ralf Hofestdt (Bielefeld University), Martin Ester (Simon Fraser University)

Background

Determining the biological function of genes and understanding the interaction of metabolism has become a major task in the post-genomic era. Huge data sets from high throughput techniques are available to analyze and to use for creating models. In this project, we will focus on models using data of medical background, e.g. related to cardio diseases. Such models could represent gene regulatory systems of cardiovascular diseases or cancer. In general, the role of individual components (e.g. RNA-molecules, genes) and processes in the reversible or irreversible changes of networks or pathways will be investigated. Beside the investigation of the aforementioned networks, new ways to explore and extract useful data to enrich the network should be developed. For further simulation, the network should be translated into Petri net formalism.

Project Description

The main aim of this project is to extend the methods of systems biology by the application of higher Petri nets for modelling, analysis and simulation of biological networks. This work should be based on previous methods of already published higher Petri nets like: colored Petri nets, Fuzzy Petri nets, Hybrid Petri nets and object oriented Petri nets. The application part should be focused on medical networks and the analysis of those data. The simulation tool should use the already existing data stored in our data warehouse DAWIS-M.D. Moreover medical and biological data from experiments could be integrated into the data warehouse. For the implementation of the system, the network and analysis application VANESA should be taken into account.

Required Skills

Experiences in relational database systems (Hibernate, MySQL) and information systems as well as high skills in Java programming are required. Knowledge in modeling and analysis of networks as well as knowledge in the field of biological and medical data is of advantage.

Project 3: Causal data analysis

Supervisors: Barbara Hammer (Bielefeld University), Martin Ester (Simon Fraser University)

Background

For any type of measurement such as e.g. SNP analysis or the analysis of medical measurements, the crucial question about which factors determine an observed outcome (such as a disease) occurs. While it is often possible to identify strongly correlated and highly predictive features automatically from given measurements, the decision whether some feature constitutes the cause for an observed effect is much harder to answer. Typically, this question requires a manipulation of the observation, i.e. further tests. Recently, quite a number of technologies have been developed in the machine learning domain which allow practitioners to identify causal relations based on observed data only, or, provided this is not possible from the observed data, can recommend minimum test settings to decide this question. The goal of the project is to integrate popular technologies in this context into a platform which enables the automated analysis of causal relations in the context of biomedical data analysis.

Project Description

Within the project, two different methods should be integrated into an advanced data analysis tool for the identification of causal relations, namely causal inference for linear models based on non-Gaussianity and independent component analysis, and causal inference based on Markov blanket analysis. The goal is to adapt these methods to the conditions and probability distributions as present in biomedical data analysis, to integrate these novel machine learning techniques into an interactive data analysis and visualization tool, and to exemplarily test the efficiency and effectiveness of the suite in benchmarks from the biomedical domain.

Required Skills

Excellent programming skills (e.g. Python, Matlab), good mathematical skills, basic knowledge in machine learning, and basic knowledge in bioinformatics are required.

Project 4: Duplication-aware ancestral genome reconstruction

Supervisors: Jens Stoye (Bielefeld University), Cedric Chauve (Simon Fraser University)

Background

Reconstructing ancestral genomes by analyzing the diversity of extant genomes is key to explore the dynamics of evolution. It is a well studied computational biology problem, especially motivated by the increasing number of sequenced and assembled genomes of extant species, but also by the development of a sequencing protocol for ancient genomes (paleogenomes). This allows to integrate both extant and extinct sequencing data within a phylogenetic framework for genome evolution analysis. Several models and methods have been developed, some of which aim at minimizing a rearrangement distance between all genomes in the tree. Others reconstruct conserved structures of co-located genes along the phylogeny. In a recent project, we successfully combined both approaches — minimizing the SCJ-distance as well as the number of gains and losses of adjacencies (Luhmann et al., Proc. of BSB 2014).

Project Description

An intricate subproblem in ancestral genome reconstruction is to guarantee consistencyöf the predicted gene order, i.e., each gene can have at most two neighbors. If genes appear only once, this linearity constraint relates to interesting computational questions like the consecutive-ones-problem or graph matchings. In contrast, gene trees provide an opportunity to integrate a duplication/loss history of duplicated genes with the reconstruction of the genome structure. However, previous approaches do not consider consistency (Anselmetti et al., Proc. of RECOMB-CG 2015).

Goal of this project is to develop a gene-tree-based and consistent reconstruction method. The new objective should combine consistency, rearrangement distance, conservation of co-localization, and agreement with duplication/loss history of genes or their reconstruction. Immediate applications would be scaffolding of extant and ancient genomes, and resolving duplication/loss histories of genes.

Required Skills

A solid background in algorithm design and analysis, discrete mathematics, and genomics is required. In particular, background in dynamic programming, linear programming, phylogenetics and probabilistic models would be a welcome addition.

Project 5: Efficient determination of accessible motifs in RNA sequences

Supervisor: Markus E. Nebel (Bielefeld University)

Background

It is well known that, in order to respond to different stimuli, the synthesis of proteins needs to be highly regulated. One mechanism being in charge relies on cis-regulatory motifs within mRNAs to which trans-regulatory proteins and microRNAs bind. Since the chemical recognition is based on an interaction between amino acids residing in the protein and the corresponding nucleotides in the cis-regulatory motif residing in the mRNA, it is of importance that the nucleotides constituting the motif are accessible, i.e. are not bonded within the mRNA. However, in order to decide this property it is no longer sufficient to work on the sequence level, but the 2D conformation of the mRNA, i.e. its secondary structure, needs to be taken into account. Thus, for the computational search for cis-regulatory motifs in RNA, structure prediction algorithms are needed. However, if one aims for identifying mRNA motifs on a genome-wide level, the classical $\Theta(n^3)$ time algorithms are not appropriate. This becomes even more problematic if one aims at comparing different genomes. There was hope to overcome this problem by sparsification of the DP recursion as proposed by Wexler et al., aiming at a quadratic time complexity on average. However, subsequent publications hinted at a mistake within the underlying analysis, proving that only a large constant factor is saved.

The goal of this project is to design an algorithm with subcubic runtime that allows for the prediction of unpaired (accessible) regions within RNA secondary structure on stochastic grounds. The underlying model will not only consider structural features of RNA molecules but should also take the knowledge of the accessible regions for already studied homologous sequences into account.

Project Description

To identify unpaired regions within a secondary structure it is not necessary to reliably predict the entire folding. A probability profile which (based on sampling secondary structures from an appropriate distribution) assigns each nucleotide the probability of being paired is sufficient. Thus, our heuristic $\Theta(n^2)$ sampling strategy for secondary structure prediction (Nebel and Scheid, Proc. of BIOINFORMATICS 2012), is a good starting point to search for a corresponding algorithm. Aforementioned heuristic builds on a fuzzy model that replaces inside-outside probabilities (which depend on two positions of the sequence) derived from a stochastic context-free grammar by corresponding weights that only use the length of the underlying subsequence thus saving a linear factor. In cases where we only want to decide a nucleotide to be paired or not (instead of predicting the full folding) it seems reasonable to hope for improved models which still allow for a speed-up. Similar to stochastic approaches where sequence and alignment are combined (covariance models, see, e.g., the Infernal tool), our grammar should be extended in order to incorporate knowledge of accessible regions for a set of sequences homologous to the one processed.

Required Skills

Strong background in formal languages, especially stochastic context-free grammars, algorithms and data structures as well as programming skills.

Project 6: Enhance the DCJ model with positional constraints

Supervisors: Jens Stoye (Bielefeld University), Cedric Chauve (Simon Fraser University)

Background

Many models have been proposed to represent structural evolution between two genomes. Most successful in recent years has been the Double Cut and Join (DCJ) model by Yancopoulos et al. (Bioinformatics 2005). It uniformly models chromosome fusions, fissions, translocations and inversions. Shortest scenarios transforming one genome into another one can be computed efficiently. One disadvantage is, though, that the solution space is of exponential size, and the result may be quite arbitrary in biological terms.

To this end Swenson and Blanchette (Proc. of WABI 2015) restrict the solution space using positional constraints, which can be obtained, for example, from chromosome conformation capture (Hi-C) experiments. They give a polynomial-time algorithm for a version of the DCJ sorting problem respecting these constraints. The constraints are very simplistic, however, because the considered weight function w considers only binary information, where two DNA regions are either considered close ($w = 0$) or not ($w = 1$). The real picture is more complicated, though.

Project Description

In this project, two extensions of the existing DCJ model with positional constraints by Swenson and Blanchette shall be studied.

First, we suggest to include genome transformation scenarios with a slightly higher number of DCJ steps, if this allows to reduce the overall weight w. Clearly, this will increase the underlying search space and therefore potentially make the problem harder. However, other extensions of DCJ, adding operations like indels or substitutions, did not lead to a substantially harder problem complexity. Therefore we hope that a similar, efficient extension may be possible here as well, while producing much more realistic results.

Secondly, the range of the weight function w shall be generalized to better reflect the underlying (non-binary) Hi-C contact data. Again, it is not yet clear whether this will increase the problem complexity or not.

Required Skills

Knowledge of algorithms and data structures for bioinformatics, basic knowledge of genome rearrangement problems, programming in Python, Java or C/C++.

Project 7: Hybrid assembly strategies for microbial and viral metagenomes

Supervisor: Alexander Sczyrba (Bielefeld University)

Background

The advent of long-read sequencing technologies such as IlluminaÕs Synthetic Long Reads, PacBio SMRT sequencing, or Oxford Nanopore sequencing, have been shown to be highly advantageous for genome assemblies. Steady increase in throughput of these technologies render possible their application to metagenomics studies, which require a much higher sequencing depth.

Project Description

The CAMI challenge (www.cami-challenge.org) has identified several shortcomings of current assembly approaches. Especially, populations with a high number of closely related species pose a challenge on existing assemblers. The goal of the project is to enhance microbial and viral metagenome assemblies by hybrid strategies incorporating long reads into the assembly process, providing an unprecedented view of the diversity of microbial and viral genomes.

Required Skills

Excellent programming skills, knowledge of algorithms and data structures for bioinformatics, especially sequence assembly are required.

Project 8: Interactive relevance determination for metabolomics analysis

Supervisors: Barbara Hammer (Bielefeld University), Ryan Brinkman (BC Cancer Agency)

Background

High throughput metabolomics constitutes a powerful technology for the analysis of complex processes such as the degree of toxicity of compound substances. One of the crucial questions in this context concerns the identification of relevant factors which trigger an observed process and which constitute potential parameters for an explicit model or biomarkers. While many techniques exist which allow the identification of strong signals or relevant linear factors, relevance determination becomes much more problematic as it concerns weak signals and local or nonlinear factors. The project will center around the development of new, interactive techniques which enables the analysis of weak signals in methabolomic data analysis.

Project Description

The goal of the project is to integrate novel methods from machine learning which enable an advanced analysis of metabolomics data. A particular focus will be laid on the identification of relevant factors which can explain observed phenomena. Thereby, emphasis will be laid on nonlinear methods which all relevance determination, methods which can identify weak signals in complex settings as opposed to strongly relevant features, and interactive tools which are based on nonlinear data visualization.

Required Skills

Excellent programming skills (e.g. Python, Matlab), good mathematical skills, basic knowledge in machine learning, and basic knowledge in bioinformatics are required.

Project 9: Large-scale storage, analysis and integration of metabolomics data

Supervisors: Stefan P. Albaum (Bielefeld University), Tim W. Nattkemper (Bielefeld University), Karsten Niehaus (Bielefeld University)

Background

Recent improvements in technologies and experimental methods have led to rapidly increased amounts of data in all areas of life science. Also in the field of metabolomics, in which mass spectrometry is the key technology to investigate the metabolites that are abundant in an organism or a tissue, state-of-the-art methods allow to gain ten thousands of mass spectra within a few minutes which in turn characterize hundreds of potential compounds. Nowadays, more and more, the processing of these large amounts of data has emerged as a bottleneck and requires new ways of data handling, processing and interpretation.

Project Description

This project aims at the development and evaluation of novel methods for the storage and analysis of mass spectrometry-based metabolomics data based on so called Big Data frameworks which allow for the distributed processing of large data sets across clusters of computers in a scalable manner. Examples for this are the Apache Hadoop software libraries and Apache Spark as a fast and general engine for large-scale data processing. Goal of the project is to provide users a bioinformatics platform for the efficient and user-friendly handling of own experimental data. Key objectives of the platform are, at first, the automated processing of large numbers of chromatographic datasets in terms of untargeted metabolite profiling, quantification and de-novo identification, and at second, their integration with other omics data to finally enable a target oriented and time efficient interpretation of the data. This project will be done in close cooperation with experimental metabolome researchers at the Center for Biotechnology (CeBiTec) at Bielefeld University.

Required Skills

Knowledge in bioinformatics data handling and processing, preferably but not essentially (also) in the field of mass spectrometry-based omics are required. At least a basic understanding of molecular biology as well as the ability to communicate with potential users will be necessary.

Project 10: RNA secondary structure prediction with base triples

Supervisor: Markus E. Nebel (Bielefeld University)

Background

The classic RNA secondary structure model considers only Watson-Crick base pairs together with GU pairs. The analysis of 3D structures, however, shows that much more interactions are possible. In detail, 12 families of base pairs are characterized. The diversity of interactions stems from the possibility to form H bonds along three different edges. Thus, a single base can interact with up to three other bases simultaneously. It has been discovered that about 40% of all bases in structured RNA take part in edge-to-edge interactions other than canonical Watson-Crick base pairs. Base triples are a commonly found non-canonical interaction. They occur in RNA motifs such as sarcin-ricin loops and support tertiary RNA interactions. Base triples are even part of a highly conserved, thus universal packaging mode of RNA helices. Accordingly, the introduction of base triples into existing RNA models has the potential to bring them closer to nature. First prediction tools exist that build on models of the free energy for foldings with base triples. In this project, we want to start a line of research which extends well-known stochastic approaches for the prediction of RNA secondary structures to incorporate base triple interactions. Furthermore, we want to improve our stochastic models by incorporating knowledge on conserved structural features of sets of homologous sequences.

Project Description

In this project, we will use the grammar from (Mller and Nebel, J Comp Biol, 2015) in order to design algorithms for the stochastics-based prediction of RNA structures with base triples. We will consider the determination of the most probable folding as well as sampling-based approaches (probability profiles, centroid structure, ...). One challenge comes with the training of the models since only a small amount of training data is available. This problem, however, might be resolved by a two-stage approach, where first a core grammar without base triples is trained on traditional secondary structure data and afterwards the additional parameters for base triples are estimated. A special case is given, if we predict the folding of a sequence known to be homologous to a set of sequences with known structure. Here we plan to fine tune our model parameters by taking this additional information into account. We furthermore aim for a generalization of shape abstraction which takes base triples into account.

Required Skills

Strong background in formal languages, especially stochastic context-free grammars, algorithms and data structures as well as programming skills.

Project 11: Spatial metabolomics: Analysis of multimodal bioimage data in medical research

Supervisors: Tim W. Nattkemper (Bielefeld University), Karsten Niehaus (Bielefeld University)

Background

In collaboration between the Faculty of Physics (Prof. T. Huser) and Biology (Prof. K. Niehaus) and the Faculty of Technology (Prof. T. W. Nattkemper) a new imaginig platform is under developed. Using a sophisticated protocol, high-dimensional image data is recorded for one sample by combining Raman imaging with imaging mass spectrometry (IMS). These two technologies visualize totally different biochemical information with spatial resolution and the analysis of such high-dimensional mutimodal data is still an open issue. In this project, we will address the specific data analysis problems when analyzing such data from tissue section provided by our collaborators from medical research institutes, e.g. tumor samples. In this project we will address the specific data analysis problems when analysig such data from tissue sections provided by our collaborators from medical research institutes, e.g. tumor samples, with a spatial focus to comaparative analysis of n data sets and the specific algorithmic questions related to that (such as signal alignment, normalization, visualization, etc).

Project Description

In the project the candidate will develop new algorithmic approaches to track down hidden regularities and interesting novelties in the data. The final aim is an algorithmic pipeline to support the researchers in the process of rendering a mental model of the image data and to discover new relationships or molecular markers. In the first part of the project, the candidate conentrates on the question of spatial alignment / registration between the two image domains. Afterwards, a platform for visually exploring the data is implemented that provides the basis for the development and application of spatially focused analytical / data mining tools.

Required Skills

Good programming skills, experience in object oriented programming, good background in mathematics, interest in statistics, and background in image processing are required.

Project 12: Using de Bruijn Graphs for DNA Read Quality Control

Supervisors: Jens Stoye (Bielefeld University), Cenk S. Sahinalp (Simon Fraser University)

Background

Modern DNA sequencing technologies produce longer reads, but at the cost of higher error rates. By merging sets of short (second-generation) and long (third-generation) reads, it is possible to improve the sequence quality, while keeping the context information.

Recently a possible such strategy has been proposed under the name Jabba (Miclotte et al., Proc. of WABI 2015), using an underlying de Bruijn graph data structure. It first creates a de Bruijn graph for a high number of short sequencing reads, which is then corrected by intrinsic graph correction methods, removing small bubbles and tips. Finally, long reads are mapped into the graph (allowing for small errors), thus finding a path that most likely represents the corrected sequence.

Other basic tasks in DNA quality control include clustering of highly similar DNA sequences, identification of decontamination, or the prediction of chimeric sequences.

Project Description

Recently, we developed a very efficient implementation of a colored de Bruijn graph, the Bloom Filter Trie (BFT), which was published in (Holley et al., Proc. of WABI 2015). So far it has been applied to the storage and analysis of bacterial pangenomes and to the compression of large sets of DNA sequences. In this project, various algorithmic tasks in DNA quality control shall be devised and implemented using the BFT as underlying data structure.

The first task will be to re-implement the idea of Jabba. Since the BFT is a colored de Bruijn graph, it may be possible to store both short and long reads together in the BFT, using different colors. Then the graph correction step can alrady take some information from the long reads into account, while the higher quality of the short reads will still dominate, due to their much larger number.

Later in the project, other DNA quality control tasks shall also be considered and, if successful, be implemented using the BFT.

Required Skills

Knowledge of data structures and algorithms for bioinformatics sequence analysis is mandatory, including sequence alignment algorithms, suffix trees/arrays, Burrows Wheeler Transformation, and de Bruijn graphs. It will be necessary to work in a Unix/Linux environment. Experience in the implementation of optimized software in C/C++ is required.