Göttingen Bioinformatics – Institute of Microbiology and Genetics
Main research fields

June 13, 2008

This handout outlines the research fields at the Department of Bioinformatics (IMG), Göttingen. For each topic, a general description and some selected publications of methods developed in our group are given. Further publications are available at http://gobics.de/department/publications.

Multiple Sequence Alignment

Contact: Prof. Dr. Burkhard Morgenstern

Burkhard Morgenstern, head of the bioinformatics group, developed the well known program DIALIGN for multiple alignment of DNA and protein sequences. The algorithm is consequently enhanced in the Department of Bioinformatics and in cooperation with Amarendran R. Subramanian from the University of Tübingen.

- A.R. SUBRAMANIAN, M. KAUFMANN, B. MORGENSTERN (2008) **DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment**, *Algorithms for Molecular Biology*, 3:6
- B. MORGENSTERN, S.J. PROHASKA, D. PÖHLER, P.F. STADLER (2006) **Multiple sequence alignment with user-defined anchor points**, *Algorithms for Molecular Biology* 1,6.
- A.R. SUBRAMANIAN, J. WEYER-MENKHOFF, M. KAUFMANN, B. MORGENSTERN (2005) **DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment**, *BMC Bioinformatics* 6, 66.
- B. MORGENSTERN, N. WERNER, S.J. PROHASKA, R. STEINKAMP, I. SCHNEIDER, A.R. SUBRAMANIAN, P.F. STADLER, J. WEYER-MENKHOFF (2005) **Multiple sequence alignment with user-defined constraints at GOBICS**, *Bioinformatics* 21, 1271-1273

▷ M. BRUDNO, R. STEINKAMP, B. MORGENSTERN (2004) **The CHAOS/DIALIGN WWW server for Multiple Alignment of Genomic Sequences**, *Nucleic Acids Res.* 32, W41-W44.

Phylogenetic reconstruction

Contact: Prof. Dr. Burkhard Morgenstern & Jun.-Prof. Gert Wörheide, Fabian Schreiber

Phylogeny is the study of evolutionary relatedness among various groups of organisms. Scientist try to visualize this relatedness of all living organisms by a phylogenetic tree, called the tree of life. In our group we focus on the construction of a software pipeline for their use in phylogenetic analysis. This software pipeline will construct datasets by searching given databases for homologous genes and further process them to build a reliable basis for the subsequent tree inference.

HIV subtyping / classification

Contact: Dr. Mario Stanke, Dr. Ingo Bulla, Anne-Kathrin Schultz

The Human Immunodeficiency Virus (HIV) has developed into many genomic subtypes with different geographical associations. Frequently, recombinant forms of the virus are detected that stem from multiple infections of an individual with different strains. These recombinants involve two or more "pure" subtypes or other recombinant forms and can involve multiple recombination breakpoints along the HIV genome.
Our group develops bioinformatics tools for detecting recombination and for the identification of parental subtypes and breakpoints for HIV and other viruses. Further, we are developing a model using coalescent theory for reconstructing the genealogy of HIV with special consideration of recombination.

▷ A.-K. SCHULTZ, M. ZHANG, T. LEITNER, C. KUIKEN, B. KORBER, B. MORGENSTERN, M. STANKE (2006) **A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes**, *BMC Bioinformatics* 7, 265.

▷ M. ZHANG, A.-K. SCHULTZ, C. CALEF, C. KUIKEN, T. LEITNER, B. KORBER, B. MORGENSTERN, M. STANKE (2006) **jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1**, *Nucleic Acids Res.* 34, W463 - W465.

Gene prediction

Eukaryotic gene prediction

Contact: Dr. Mario Stanke

We develop mathematical models and algorithms for the identification of protein-coding genes on a genome-wide scale. We are currently or were recently involved in the genome projects of about a dozen eukaryotes (e.g. the green algae *Chlamydomonas reinhardtii*, the yellow wasp *Nasonia vitripennis*, the beetle *Tribolium castaneum*, the nematode *Brugia malayi*). In order to provide high-quality annotation we are working on the integration of data from a variety of external data sources that help to improve genome annotation. These data sources include peptide mass spectrometry, transcript alignments, genomic conservation and known protein sequences.

- M. STANKE, M. DIEKHANS, R. BAERTSCH, D. HAUSSLER (2007) **Using native and syntenically mapped cDNA alignments to improve de novo gene finding**, *Bioinformatics* 24, 637-644.
- M. STANKE, A. TZVETKOVA, B. MORGENSTERN (2006) **AUGUSTUS+ at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome**, *Genome Biology* 7, S11.
- M. STANKE, O. KELLER, I. GUNDUZ, A. HAYES, S. WAACK, B. MORGENSTERN (2005) **AUGUSTUS: ab initio prediction of alternative transcripts**, NUCLEIC ACIDS RES. 34, W435 - W439.
- M. STANKE, R. STEINKAMP, S. WAACK, B. MORGENSTERN (2004) **AUGUSTUS: a web server for gene finding in eukaryotes**, *Nucleic Acids Res.* 32, W309-W312.

Prediction of prokaryotic translation initiation sites

Contact: Dr. Peter Meinicke, Dr. Maike Tech

Prediction of protein coding genes for prokaryotic genomes has reached a high level of detection sensitivity and specificy. Nevertheless, most classical gene finders lack accuracy for predicting the exact translation initiation site (TIS), because prokaryotic genes generally implicate several putative start sites. Our group developed the tool TICO to correct possibly false TIS predictions. Reannotation by TICO has been estimated to yields an improvement about 10 to 30 %.

- C. IGEL, T. GLASMACHERS, B. MERSCH, N. PFEIFER, P. MEINICKE (2007) **Gradient-based Optimization of Kernel-Target Alignment for Sequence Kernels Applied to Bacterial Gene Start Detection**, *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4, 216-226.

- ▷ M. TECH, B. MORGENSTERN, P. MEINICKE (2006) **TICO: a tool for post processing the predictions of prokaryotic translation initiation sites**, *Nucleic Acids Res.* 34, W588 - W590.
- ▷ M. TECH and P. MEINICKE (2006) **An unsupervised classification scheme for improving predictions of prokaryotic TIS**, *BMC Bioinformatics* 7, 121.
- ▷ P. MEINICKE, M. TECH, B. MORGENSTERN, R. MERKL (2004) **Oligo Kernels for datamining on biological sequences: A case study on prokaryotic translation initiation sites**, *BMC Bioinformatics* 5, 169

Gene prediction in metagenomics

Contact: Dr. Peter Meinicke & Dr. Maike Tech, Katharina J. Hoff

Direct sequencing of DNA samples from a habitat allows the culture independent analysis of organisms. Bioinformatics focuses in this field on two issues: 1.) Gene prediction for detection of novel proteins and for functional profiling of the habitat; 2.) Classification of sequences for phylogenetic analysis of the habitat. Our group develops new methods for large-scale gene prediction on metagenomic sequences.

- ▷ K. J. HOFF, M. TECH, T. LINGNER, R. DANIEL, B. MORGENSTERN, P. MEINICKE (2008) **Gene prediction in metagenomic fragments: a large scale machine learning approach**, *BMC Bioinformatics* 9:217

Prediction of small functional RNAs

Contact: Prof. Dr. Burkhard Morgenstern, Isabelle Schnneider

Functional RNA (fRNA) play a fundamental role in many regulatory processes in the cell. While for some RNA families like transfer-RNA (tRNA) and ribosomal RNA (rRNA) reliable methods exist to identify the corresponding genes, the detection of other kinds of functional RNA is still a difficult task. A combined approach of comparative sequence analysis and structure prediction seems to be most promising. At present, we use a strategy based on reliable tools for the detection of fRNA in complete genome sequences.

Protein classification

Contact: Dr. Peter Meinicke, Thomas Lingner

In the last few years, the number of known protein sequences has rapidly increased. In order to characterize these sequences, we develop alignment-free approaches for protein function prediction and remote homology detection. Our research focuses on methods that allow ultra-fast analysis and interpretation of features that have been learned from the sequences. Currently, discriminative machine learning approaches provide state-of-the-art prediction performance, therefore the design and development of suitable evaluation setups is a key element of our research.

- T. LINGNER and P. MEINICKE (2006) **Remote homology detection based on Oligomer Distances**, *Bioinformatics* 22, 2224 - 2231.
- T. LINGNER and P. MEINICKE (2008) **Word Correlation Matrices for Protein Sequence Analysis and Remote Homology Detection**, *BMC Bioinformatics*, in press
- T. LINGNER and P. MEINICKE (2008) **Fast Target Set Reduction for Large-scale Protein Function Prediction: a Multi-class Multi-label Machine Learning Approach**, 8th Workshop on Algorithms in Bioinformatics (WABI) 2008, in preparation

Data Mining in metabolomics

Contact: Dr. Peter Meinicke, Alexander Kaever

One of the central goals of global metabolomic analysis is to identify metabolic markers that are hidden within huge amounts of mass spectrometry measurements. These markers reflect the concentration variation of relevant intracellular metabolites under varying conditions, like disease or environmental and genetic perturbations. We develop methods for data mining on large sets of marker candidates using machine learning techniques. In particular we built a tool for clustering and visualization of complex marker profiles which involve several experimental conditions.

- P. MEINICKE, T. LINGNER, A. KAEVER, K. FEUSSNER, C. GÖBEL, I. FEUSSNER, P. KARLOVSKY AND B. MORGENSTERN (2008) **Metabolite-based clustering and visualization of mass spectrometry data using one-dimensional self-organizing maps**, *Algorithms for Molecular Biology*, submitted