Bioinformatics

In the following, we will highlight recent achievements of the Department in Computational Biology.

Genetics We developed methods for the problem of *two-locus mapping*, that is to detect pairs of bases in the human genome whose state correlates with complex diseases. This problem is complicated by an enormous search space, with up to 10^{14} candidate pairs. We overcame this difficulty in two ways. First, we developed implementations of two-locus search on graphical processing units (GPUs) [10, 11, 9], which are extremely optimized for performing thousands of matrix operations in parallel. We could bring down the runtime for two-locus search using 4 GPUs to less than a day for current studies. Large genetics consortia for migraine, lung diseases and psychiatric diseases are now using our GPU implementations for two-locus mapping. Second, we approached the problem mathematically and developed an algorithm for two-locus mapping on binary phenotypes, which is provably subquadratic in the number of SNPs [16]. tooltip In our future work, we aim to discover networks of SNPs that are jointly associated with a phenotype, and to exploit our results from the field of network comparison [19, 14] for this task.

Genomics We are most interested in applying discriminative machine learning techniques to DNA sequence analysis. Earlier, we have developed a prototype gene finder called mGene which had shown high prediction quality in an international competition[1]. In mGene, the segmentation problem was solved with hidden semi-Markov support vector machines (HSM-SVMs). Applying this technique to data sets of the size and complexity of genomic-scale sequences posed a substantial challenge. It is now possible to process thousands of sequences in a reasonable time span, while taking full advantage of the wealth of encoded information. We produced genome-wide annotations for *C. elegans* and four related nematodes, which were provided to the wormbase consortium for inclusion into the official annotation [13]. These annotations formed the basis of an in-depth comparative analysis of the five nematode gene catalogues. To aid the highly involved process of annotation, a web server was developed that allows to perform the complex process of gene prediction [12]. More recently, mGene has become the core of a new system that is aimed at gene structure reconstruction using both, DNA sequence and RNA-Seq data. In addition, we believe and have shown that various problems in the biomedical domain can be formulated as a Multitask Learning problem, allowing us employ our methods to obtain more accurate models [20, 18, 15, 17].

Systems Biology Our research in Systems Biology is concerned with biological networks, for example protein-protein interaction networks, which accumulate in high-throughput experiments. This raises the need for methods that automatically and reliably detect characteristic patterns in structured relational data. We have developed an algorithm that systematically finds all densely connected groups in large weighted networks, optionally taking into account constraints from other data sources [3, 2]. tooltip Furthermore, we have devised a general framework for extracting dense substructure patterns from higher-order association data, including symmetric or asymmetric n-ary relations, hypergraphs, and weight tensors [4]. The approach has assisted in meta-analyses of gene signatures and gene networks obtained from different experiments.

Proteomics We develop computational tools to study, model and predict aspects of protein structure [8]. Our major focus is biomolecular structure determination with challenging experimental data including sparse NMR data [5] and low-resolution data obtained with cryo-electron microscopy (cryo-EM) [6]. We are pursuing a unique approach, ISD (Inferential Structure Determination), that tackles structure determination problems by means of Bayesian probability theory[2]. This allows us to integrate heterogeneous data sets, to model different sources of uncertainty by means of probabilistic models and to combine experimental data with information from known protein structures [7]. We have used ISD to determine the first structure of a mitochondrial porin [1]. More recently, we have determined the solution structure of a MAP kinase P38/inhibitor complex by combining crystallographic with solution NMR data and the first structure of a membrane protein from solid-state NMR data exclusively.

[1] A Coghlan *et al.* nGASP – the nematod genome association assessment project. *BMC Bioinformatics* 2008, **9**:549

[2] W Rieping *et al.* "ISD: A Software Package for Bayesian NMR Structure Calculation". *Bioinformatics* 2008 Apr 15;24(8):1104-5

Publications

Journal Articles

[1] M Bayrhuber, T Meins, M Habeck, S Becker, K Giller, S Villinger, C Vonrhein, C Griesinger, M Zweckstetter, and K Zeth. Structure of the human voltage-dependent anion channel. *Proceedings of the National Academy of Sciences of the United States of America*, 105(40):15370–15375, 10 2008. 1

[2] S Dietmann, E Georgii, A Antonov, K Tsuda, and H-W Mewes. The DICS repository: module-assisted analysis of disease-related gene lists. *Bioinformatics*, 25(6):830–831, 1 2009. 1

[3] E Georgii, S Dietmann, T Uno, P Pagel, and K Tsuda. Enumeration of condition-dependent dense modules in protein interaction networks. *Bioinformatics*, 25(7):933–940, 2 2009. 1

[4] E Georgii, K Tsuda, and B Schölkopf. Multi-way set enumeration in weight tensors. *Machine Learning*, 82(2):123–155, 2 2011. 1

[5] M Habeck. Statistical mechanics analysis of sparse data. *Journal of Structural Biology*, 173(3):541–548, 3 2011. 1

[6] M Hirsch, B Schölkopf, and M Habeck. A blind deconvolution approach for improving the resolution of Cryo-EM density maps. *Journal of Computational Biology*, 18(3):335–346, 3 2011. 1

[7] I Kalev and M Habeck. HHfrag: HMM-based fragment detection using hhpred. *Bioinformatics*, 27(22):3110–3116, 11 2011. 1

[8] I Kalev, M Mechelke, K Kopec, T Holder, S Carstens, and M Habeck. CSB: A Python framework for computational structural biology. *Bioinformatics*, 2012. 1

[9] T Kam-Thong, C-A Azencott, L Cayton, B Pütz, A Altmann, N Karbalai, PG Sämann, B Schölkopf, B Müller-Myhsok, and KM Borgwardt. GLIDE: GPU-based linear regression for detection of epistasis. *Human Heredity*, 73(4):220–236, 9 2012. 1

[10] T Kam-Thong, D Czamara, K Tsuda, K Borgwardt, CM Lewis, A Erhardt-Lehmann, B Hemmer, P Rieckmann, M Daake, F Weber, C Wolf, A Ziegler, B Pütz, F Holsboer, B Schölkopf, and B Müller-Myhsok. EPIBLASTER—fast exhaustive two-locus epistasis detection strategy using graphical processing units. *European Journal of Human Genetics*, 19(4):465–471, 4 2011. 1

[11] T Kam-Thong, B Pütz, N Karbalai, B Müller-Myhsok, and K Borgwardt. Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs. *Bioinformatics*, 27(13: ISMB/ECCB 2011):i214–i221, 7 2011. 1

[12] G Schweikert, J Behr, A Zien, G Zeller, CS Ong, S Sonnenburg, and G Rätsch. mGene.web: a web service for accurate computational gene finding. *Nucleic Acids Research*, 37:W312–6, 2009. 1

[13] G Schweikert, A Zien, G Zeller, J Behr, C Dieterich, CS Ong, P Philips, F De Bona, L Hartmann, A Bohlen, N Krüger, S Sonnenburg, and G Rätsch. mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Research*, 19(11):2133–43, 2009. 1

[14] N Shervashidze, P Schweitzer, EJ van Leeuwen, K Mehlhorn, and M Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 9 2011. 1

[15] C Widmer, NC Toussaint, Y Altun, and G Rätsch. Inferring latent task structure for multitask learning by multiple kernel learning. *BMC Bioinformatics*, 11 Suppl 8:S5, 2010. 1

Articles in Conference Proceedings

[16] P Achlioptas, B Schölkopf, and K Borgwardt. Two-locus association mapping in subquadratic time. In C Apté, J Ghosh, and P Smyth, editors, *17th ACM SIGKKD Conference on Knowledge Discovery and Data Mining (KDD 2011)*, pages 726–734, San Diego, CA, USA, 8 2011. ACM Press. 1

[17] N Görnitz, C Widmer, G Zeller, A Kahles, S Sonnenburg, and G Rätsch. Hierarchical multitask structured output learning for large-scale sequence segmentation. In J Shawe-Taylor, RS Zemel, P Bartlett, FCN Pereira, and KQ Weinberger, editors, *25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2690–2698, Granada, Spain, 2011. Curran Associates, Inc. 1

[18] G Schweikert, C Widmer, B Schölkopf, and G Rätsch. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *22nd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1433–1440, Vancouver, BC, Canada, 6 2009. Curran. 1

[19] N Shervashidze and KM Borgwardt. Fast subtree kernels on graphs. In Y Bengio, D Schuurmans, J Lafferty, C Williams, and A Culotta, editors, *23rd Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1660–1668, Vancouver, BC, Canada, 2009. Curran. 1

[20] C Widmer, J Leiva, Y Altun, and G Rätsch. Leveraging sequence classification by taxonomy-based multitask learning. In B Berger, editor, *Research in Computational Molecular Biology (RECOMB), LNCS, Vol. 6044*, pages 522–534, Lisbon, Portugal, 2010. Springer. 1