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].

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]. 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

Publications

Journal Articles


Articles in Conference Proceedings


