A machine learning method for the prediction of regulatory interactions

Heidelberg bioinformaticians have developed a novel method for the automated prediction of regulatory interactions. The regulatory interaction predictor, a machine-learning based approach for predicting interactions between DNA-binding transcription factors and their target genes and obtaining important insights into the gene regulatory networks in complex cells.

The different characteristics and function of the 200 or so different cell types in the human body depends on the proteins that are expressed in the cells. The availability of these proteins depends on the activity of the protein-coding genes that can be identified by analysing the expression of genes using mRNA microarrays. The diversity of gene expression patterns is a lot greater than cell type diversity due to the broad range of different functional and differentiation states and pathological alterations. Complex, finely tuned regulation patterns determine the expression of these genes. Gene regulatory proteins, i.e. transcription factors (TF) that bind to the promoters of genes and initiate or suppress the process of RNA-polymerase mediated transcription, play a crucial role in the expression of genes.

There are one hundred known DNA-binding TFs that control the transcription of the 20,000 – 22,000 human protein-encoding genes. The regulatory systems are closely linked with one another. However, detailed reconstruction of the regulatory networks from the whole genome is impossible due to the huge number of potential TF/target gene combinations and due to the lack of standardised experimental data and techniques.

Integrative bioinformatics and systems biology (iBioS)

The “Network Modelling” group, a team of Heidelberg bioinformaticians and systems biologists at the DKFZ in Heidelberg, has developed an effective bioinformatic tool for the analysis of gene expression data. This tool allows a better understanding of regulatory networks. The RIP classifier (“regulatory interaction predictor”) is a novel machine-learning based approach for predicting the regulatory interaction between TFs and the DNA sequences of the target genes in higher eukaryotes such as humans. The scientists also hope that the tool will help them identify disease-specific drug targets and put them to good use in the individualised therapy of diseases such as cancer.

The “Network Modelling” team is part of the “iBioS – Integrative Bioinformatics and Systems Biology” research group which was jointly established by the Department of Theoretical Bioinformatics at the German Cancer Research Center and the Department of Bioinformatics and Functional Genomics at the Institute of Pharmacy and Molecular Biotechnology at the University of Heidelberg. iBioS, which is led by Professor Dr. Roland Eils, is also part of the BioQuant research network, which is the systems biology centre at the University of Heidelberg.

PD Dr. Rainer König, Institute of Pharmacy and Molecular Biotechnology (IPMB) – BioQuant
© Universität Heidelberg

iBioS is developing computer-assisted methods for the analysis of the huge and complex quantities of data generated by modern high-throughput technologies and the life sciences. The group is also developing mathematical models with which the behaviour of key processes of cellular systems and alterations caused by disease can be simulated.

PD Dr. Rainer König is project leader of the Network Modelling group. The RIP classifier developed by König and his group of researchers searches for promoter sequence features (motives) that can act as TF binding sites and correlates these features with a set of experimentally verified regulatory interactions (RIs) between TFs and target genes retrieved from a public repository (TRANSFAC) and with microarray gene expression data. The researchers selected a number of motives from the TRANSFAC database based on experimentally confirmed RIs. These motives served as the gold standard to train so-called support vector machines (SVMs: algorithms or mathematical programmes for machine-based learning) with the goal to predict automatically new regulatory interactions for a large number of TFs and target genes. The 2,000 support vector machines form the basis of the RIP classifier and are applied to infer new RIs.

The RIP classifier

König and his colleague Dr. Tobias Bauer describe the RIP classifier development process as follows: “For our approach we searched for binding site motives of 303 transcription factors in the promoter sequences of 13,069 genes. We subsequently integrated comprehensive gene expression analyses. Genes that are involved in the same biological functions are often co-regulated and therefore co-expressed. The strength of the correlation of gene expression was therefore determined for each gene pair of the 13,069 genes analysed (Laborwelt 6/2011, p. 32). The more similar the cells and their differentiation states, the greater the co-expression and co-regulation of the genes.
The gene expression data of more than 4,000 mRNA microarrays were analysed. The majority of these mRNAs were derived from human tumours. The Heidelberg researchers reported in the journal Bioinformatics that their RIP classifier inferred nearly 74,000 regulatory interactions for 301 transcription factors and more than 11,000 genes with a reliability that surpasses similar prediction methods on the genome-wide level. In addition to identifying known associations between TFs and genes, the results permit the generation of hypotheses for TF-mediated regulations that can be experimentally tested. For example, transcription factors have been shown to control signal transduction and metabolic pathways that are of fundamental importance for the cell cycle, cell proliferation and cell transformation in the pathogenesis of cancer.

In order to test the predicted regulatory interactions, König and his co-workers examined the IFNα-induced (interferon-alpha) signal transduction in human blood monocytes and compared their results with a published mRNA microarray gene expression analysis. They found that the target genes, all of which were associated with 13 key TFs, belonged exclusively to the IFNα-activated genes described in the literature. The most prominent TF was the interferon-stimulated gene factor 3, which regulated the transcriptional response, and hence differential expression, of 70% of the predicted 28 target genes upon the stimulation of monocytes with interferon-alpha. The case study also showed that the predicted TF modules were closely associated with the signalling and metabolic pathways that were related to their functions.

The machine-learning based method presented by the iBioS Network Modelling group does not require any specific conditions. It can basically be employed for any cell type to automatically predict the regulatory interactions between transcription factors and their target genes. If necessary, it can also be extended to other genes and transcription factors. The method also works when the transcription factor is not co-expressed with the target gene, but instead is regulated on the protein level. This is due to the fact that the features used are only deduced from the analysis of the co-regulated target genes. The RIP classifier software is available free of charge to the scientific community and can be further developed to integrate other features from new high-throughput technologies into the correlation analysis, for example.

References: