(Tutorial Proposal for ISMB2002)

Alexander Kel  
Senior Scientist, Lab. of Molecular Genetic Systems, Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia  
mail: kel@bionet.nsc.ru

Edgar Wingender  
Research Group Bioinformatics, GBF, Gesellschaft fuer Biotechnologische Forschung mbH, Mascheroder Weg 1, 38124 Braunschweig, Germany.  
President & CSO, BIOBASE GmbH, Halchersche Strasse 33, D-38304 Wolfenbuettel, Germany  
mail ewi@biobase.de

Tutors:

Dr. Alexander Kel is the Senior Scientist of the Laboratory of Molecular Genetic Systems of the Institute of Cytology and Genetics in Novosibirsk, Russia. He is also the Vice President of the Bioinformatic company BIOBASE GmbH in Germany. Over many years he has been working in the field of computational analysis of gene regulation. His group has developed one of the first databases on gene regulation and the first program of promoter recognition. Alexander Kel has been invited to several conferences for oral presentations. He also has been teaching genetics and bioinformatics at the Novosibirsk State University since 1988.

Dr. Edgar Wingender is the head of Research Group Bioinformatics at the German Research Centre for Biotechnological. He is also one of the founders, President and Chief Scientific Officer of the company BIOBASE GmbH that is maintaining and distributing the TRANSFAC database, the main database on transcription factors and gene regulation, as well as several other databases and software tools. Edgar Wingender is one of the worldwide leading scientist in the field of development of databases on gene regulation. His book on transcription factors is one of the most comprehensive survey on transcription regulation of eukaryotic genes. His teaching experience is rather broad and includes several series of lectures and tracks at several Universities.

Tutorial presentation:

Regulation of gene expression becomes the key problem of the era of “Functional Genomics”. Now we know that the activity of genes in genomes of higher eukaryotic organisms are regulated mainly by the means of huge class of regulatory proteins (transcription factors, TF), through specific regulatory sequences – TF binding sites that are located usually in a proximity of the genes. Having available genomic sequences on one side and the massive though purely phenomenological gene expression data on the other side the challenge is to understand regulatory mechanisms of all and every gene in a genome by computer analysis of the gene regulatory sequences and by integrating this data with biological knowledge of signal transduction, metabolic and physiological networks. Sophisticated computational regulatory sequence analysis tools that employ powerful statistical and machine learning algorithms driven by the rich databases that collect known biological facts enable us to make profound in silico predictions and formulate experimentally testable hypothesis. Such in silico driven
experiments can greatly speed up the process of our understanding of the gene regulatory mechanisms and identification of new regulatory target genes. The understanding of how gene regulation mechanisms are encoded in the genomic regulatory sequences will give us a powerful means for deciphering causes of major human diseases. The goal of this tutorial is to provide a comprehensive introduction to the up-to-date knowledge and dominating hypotheses on gene regulatory mechanisms, to give a survey of computational apaches of in silico analysis of gene regulatory sequences and to give practical tips on using available software tools and databases.

**Intended audience:**

The tutorial is made to marry needs and interests of both computer scientists who are looking for better understanding of molecular mechanisms of gene regulation and biologists who seek for useful computational approaches and tools that can advance their experimental work in the field of gene regulation. The tutorial requires only basic knowledge in statistics and molecular biology. During the tutorial several illustrative examples of analysis of different regulatory sequences and identification of new target genes will be given. There will be an on-line demonstration of the major databases and analysis tools.

**Length:**

Half day (approximately 4 hours).

**Detailed outline of the presentation:**

1. Introduction:
   1.1. Different levels of regulation of gene expression (transcription, translation splicing, degradation).
   1.2. Modern concepts of the hierarchical structure and function of the gene regulatory regions in genome. Combinatorial regulation. Concept of fuzzy puzzle
   1.3. Transcription factors. Structure, classification and evolution.

2. Databases on gene regulation
   2.1. Gene regulatory sequences in different sequence databases (transcription: promoters, enhancers, LCR, S/MARs; translation: 5'UTR, 3'UTRs, translation enhancers; splicing: regulated splice sites)
   2.2. Databases on transcription factors and their binding sites
   2.3. Databases on gene regulatory networks and signal transduction pathways.
   2.4. Organisation of gene expression data. Databases on organs, tissues, developmental stages.
   2.5. Functional classifications of genes.
   2.6. Gene regulatory circuits of some common human diseases

3. Regulatory sequence analysis tools and approaches.
   3.1. Motif analysis
      3.1.1. Motif finding algorithms (pattern and sequence driven)
      3.1.2. Gibbs sampling
      3.1.3. Clusterisation, local consensi and Kernel functions
      3.1.4. Limitations of the motif finding in the analysis of regulatory sequences
   3.2. Recognition of DNA functional sites
3.2.1. Search by consensus/pattern. Search by homology. GCG, ConsInspector.
3.2.2. Weight matrices (mono-, di-, oligonucleotide approach). Theory and practice. MatInspector, Match, TRANSPLORER,
3.2.3. Application of modern statistic and machine learning approaches: Fast transformations, pattern recognition, utility theory. SITEVIDEO system.
3.2.4. Influence of site flanks. Local context.
3.2.5. Recognition of composite elements (consisting of two sites).

3.3. Recognition of promoters.
3.3.1. Types of promoters. TATA-less promoters. CpG islands.
3.3.2. Brief survey of promoter recognition programs. Over-fitting problem.
3.3.3. Does a general promoter structure exist?

3.4. Functional classification of promoters and prediction of gene regulation.
3.4.1. Classification functions based on combinations of binding sites. Composite clusters.
3.4.2. Decision tree.
3.4.3. AI approaches. Genetic algorithm. ClusterScan tool.

3.5. Phylogenetic footprinting. Analysis of regulatory sequences by comparison of genomic sequences from different species. Human/mouse map of conservative regulatory regions. What common and what is different between human and mouse gene regulation?

3.6. Promoter engineering.

4. Analysis of gene expression data
4.1. Analysis of sequence motifs in the promoters of the clusters of co-regulated genes.
4.3. Search for common patterns of TF binding sites in the co-regulated gene promoters. ModelInspector tool.

5. Regulatory network and pathways analysis.
5.1. Enrichment of gene regulatory network by sequence analysis and reverse engineering. Upstream and downstream modelling.
5.2. Identification of common regulatory nodes. Application of graph algorithms.

6. Practical examples of applying computational tools for analysis of gene regulatory circuits and prediction of new target genes.

7. Practice of using databases and applying tools in inter- and intranet for analysis of regulatory sequences.