A database for cis-regulatory elements of eukaryotes

Youlian Pan1, Brandon Smith1, Hung Fang1, Michael W. Floyd1, Fazel A. Famili2, Roy Walker1, Marianna Sikorska1

1Institute for Information Technology and 2Institute for Biological Sciences, NRC-Ottawa
2Department of System and Computer Engineering, Carleton University, Ottawa, (MWF was a co-op student at NRC)

Introduction
Gene expression in eukaryotic cells is regulated by a group of proteins called transcription factors, which bind to specific fragments of a DNA sequence known as Transcription Factor Binding Sites (TFBSs) or more generally “motifs”. These motifs are typically 5-15 bp in length (but can be as long as 30 bp) and usually appear in the upstream region of a gene. Once bound to their motifs, the transcription factors (TFs) may activate, enhance, or repress transcription and often as a part of a multi-factor mechanism.

As many completed genomes became available over the past decade, more gene annotations, which include genomic mapping and functional annotations (GO), are added to various databases. During the GHI phase II program “Systems Biology of Brain Cell Interactions”, we performed genome-wide searches of transcription factor binding sites in a wide range of eukaryotes from yeast to human and identified the need for a database to assist the search and analysis of cis-regulatory elements such as transcription factor binding sites (TFBSs). In order to provide support to various programs under NRC’s Genomics and Health Initiative (GHI) and as part of bioinformatics research at NRC, we started developing a database, called cisDB, in early 2005. Currently, we are actively involved in the GHI phase III program “Personalized Medicine for Cancer” (BRI, IBS, IIT, IBID), and other research programs, such as Neurogenesis and Neurodegeneration (IBS), and Neuroglycobiology (IBS), etc. This database will be a key resource in our research programs, such as Neurogenesis and Neuro-glycobiology, and other disease-related work involving transcriptional regulation.

Vision
- Develop a key resource for research in gene expression regulation at NRC and collaborating laboratories.
- Gradually expand service to the entire research community including Canadian universities and industry.

Availability
- A preliminary version of the database is available in a central server at the Institute for Information Technology (Ottawa) and accessible through the NRC intranet (http://it41151.it.nrc.ca/cisDB.php).
- It can also be accessed via an interface within the most recent version (Version 2.2.3, September, 2005) of our genomics data mining software, BioMiner.

Content
- **Sequences:**
  - 10k upstream
  - 10k downstream
  - Open Reading Frame

- **Transcription factors:** Known transcription factors and their binding sites.

- **Motifs:** motif pattern, and the TF that binds to.

- **Motifs Found:** location in the promoter of a gene, TF that binds to, and motif pattern.

- **Gene Ontology:** function, process, and cellular location.

- **Orthologues:** human, mouse, rat.

Data Sources
- UCSC genome database (http://genome.ucsc.edu)
  - DNA sequences:
    - TransFac professional (http://www.biobase.de/pages/products/transfac.html)
    - known transcription factors
    - transcription factors binding sites
    - position weight matrices (PWMs)
    - consensus motifs

Tools
- Currently, search tools are available in the BioMiner, a recently developed biological data mining tool suite. Some web-based tools will be available soon.
  - Transcription factor binding sites
    - Exact match
    - Hidden Markov models
  - Available soon:
    - Gibbs sampling
    - MotifFilters
    - Motif search based on phylogenetic footprinting
  - Orthologues
    - Known orthologues
  - Available soon:
    - Search for orthologues through alignment of various regions by BlastZ, AVID, and LAGAN, etc.
    - One-to-one (pair-wise) alignment vs. multiple alignment.

Future perspectives
- Incorporation of other public databases, such as JASPAR, an annotated and matrix-based transcription factor binding site profiles database for multicellular eukaryotes (Sandelin et al., 2004).
- Contribution to integrated data mining which involves microarray gene expression data, motif search, and phylogenetic conservation to discover information about the genes, novel cis-regulatory elements and their functional modules.
- Contribution to discovery of transcriptional regulatory networks.
- Contribution to research in cancer genomics, Neurogenesis and Neurodegeneration, Neuroglycobiology, and other disease-related work involving transcriptional regulation.

Suggestions for requirements
Please fill the form with your suggestions and comments and put it in the envelope provided. Feel free to send additional comments and suggestions through e-mail (youlian.pan@nrc-cnrc.gc.ca).

References
James I, N. (2002). cis-regulatory elements and their...