cisDB: A database for *cis*-regulatory elements of eukaryotes

http://ir41151.iit.nrc.ca/cisDB.php

Youlian Pan[1], Brandon Smith[2], Hung Fang[2], Michael W. Floyd[3], Fazel A. Famili[1], Roy Walker[2], Marianna Sikorska[2]

[1]Institute for Information Technology and [2]Institute for Biological Sciences, NRC-Ottawa

[3]Department 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 *cis*DB, in early 2005. Currently, we are actively involved in the GHI phase III program “Personalized Medicine for Cancer” (BRI, IBS, IIT, IBD), and other research programs, such as Neurogenesis and Neurodegeneration (IBS), and Neuro-glycobiology (IBS), etc. This database will be a key resource in our research. Data from our research will constitute part of the database. This poster (1) provides our vision and database schema, (2) creates community awareness, and (3) seeks feedback and research requirements for the database from the GHI community.

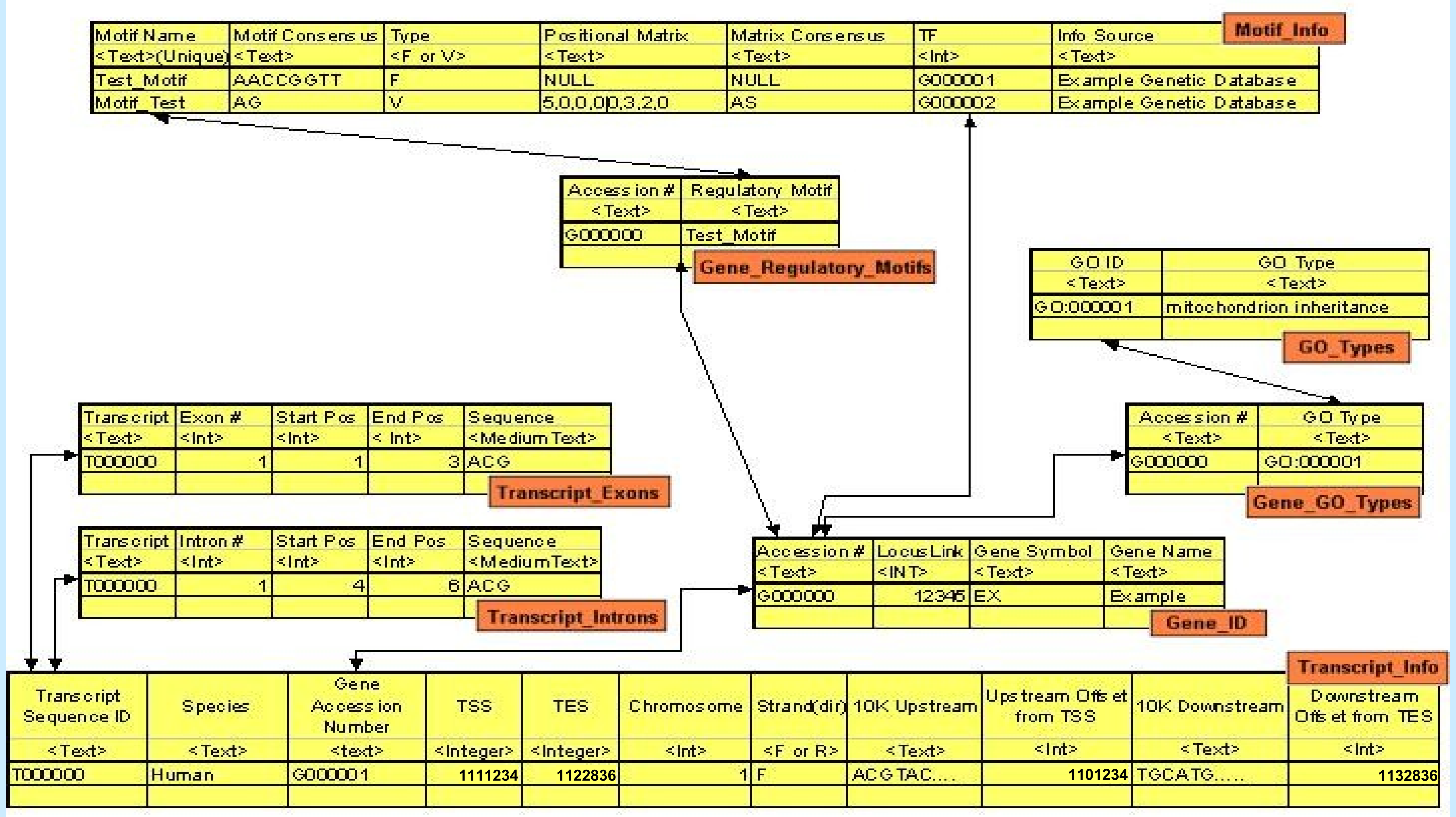
Vision

- 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://ir41151.iit.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.

Schematic relations among the *cis*DB tables



Motif_Info

Motif Name	Motif Consensus	Type	Positional Matrix	Matrix Consensus	TF	Info Source
<Text>(Unique)	<Text>	<F or V>	<Text>	<Text>	<Int>	<Text>
Test_Motif	AACCGGTT	F	NULL	NULL	G000001	Example Genetic Database
Motif_Test	AG	V	5,0,0,0,0,3,2,0	AS	G000002	Example Genetic Database

Gene_Regulatory_Motifs

Accession #	Regulatory Motif
<Text>	<Text>
G000000	Test_Motif

GO_Types

GO ID	GO Type
<Text>	<Text>
GO:000001	mitochondrion inheritance

Transcript_Exons

Transcript	Exon #	Start Pos	End Pos	Sequence
<Text>	<Int>	<Int>	< Int>	<Medium Text>
T000000	1	1	3	ACG

Gene_GO_Types

Accession #	GO Type
<Text>	<Text>
G000000	GO:000001

Transcript_Introns

Transcript	Intron #	Start Pos	End Pos	Sequence
<Text>	<Int>	<Int>	<Int>	<MediumText>
T000000	1	4	6	ACG

Gene_ID

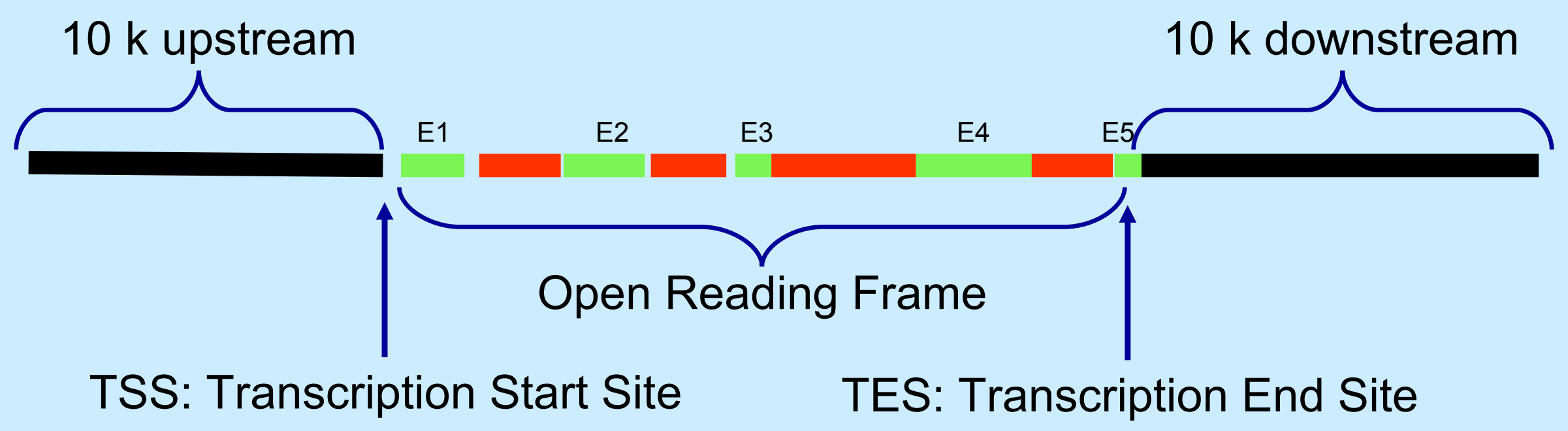
Accession #	LocusLink	Gene Symbol	Gene Name
<Text>	<INT>	<Text>	<Text>
G000000	12345	EX	Example

Transcript_Info

Transcript Sequence ID	Species	Gene Accession Number	TSS	TES	Chromosome	Strand(dir)	10K Upstream	Upstream Offset from TSS	10K Downstream	Downstream Offset from TES
<Text>	<Text>	<text>	<Integer>	<Integer>	<Int>	<F or R>	<Text>	<Int>	<Text>	<Int>
T000000	Human	G000001	1111234	1122836	1	F	ACGTAC....	1101234	TGCATG.....	1132836

Content

- ***Sequences:***



- ***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

Acknowledgements

- This database is being developed by the BioMine/BioIntelligence research team. Other team members include: Alan Barton, Junjun Ouyang, Ziying Liu, Sieu Phan, Julio Valdes, and Zuojian Tang.
- Contribution to the tool development also includes former students Weiling Xu, Ganming Liu (Carleton Univ.), Nan Zhang (UBC), and Anne Marie Simmie (Univ. Waterloo).
- This research is partially funded by the Genomics and Health Initiative (GHI) of the National Research Council Canada.

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 informative 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, Neuro-glycobiology, 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

Bray N, Pachter L (2004). MAVID: Constrained ancestral alignment of multiple sequences. Genome Res 14: 693-9.

Brudno M, D CB, Cooper GM, et al (2003). LAGAN and Multi-LAGAN: Effient tools for large scale multiple alignment of genomic DNA. Genome Res 13: 721-31.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32: D91–4

Schwarts S, Kent WJ, Smit A, et al (2003). Human-Mouse alignments with BLASTZ. Genome Res 13: 103-7.

Wingender E, Chen X, Fricke E, et al (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29: 281-3.