Introduction

Physiological function in multicellular organisms requires different cells to have different specific functions and characteristics, and the attainment of these specialized characteristics is called differentiation. Differentiation involves myriad regulatory events that lead to structural and functional organization of the genome to allow expression and regulation of the appropriate set of proteins for the specialized functions of that cell type.

The proteins that characterize a cell’s phenotype are strings of amino acids bound together whose order is coded by the sequence of nucleotides in a messenger RNA (mRNA) that is transcribed from DNA in the nucleus. The process of transcription (production of the mRNA from DNA sequence) of these protein coding genes is regulated in part by the interaction of regulatory proteins called transcription factors (TF) with specific sequences (called regulatory elements) within a region of DNA adjacent to the protein coding gene called the promoter (Figure 1).

Figure 1. Schematic showing the promoter region of a gene with three regulatory elements (RE) each bound by their respective transcription factor (TF) contributing to the activation of transcription of the gene.
Understanding the regulatory events that lead to differentiation of different cell types is a critical step in understanding normal biological function as well as understanding what errors underlie dysfunctions such as cancer. We have been studying the differentiation of mammary epithelial cells by identifying a subset of the proteins whose abundance changes during the transition from a mammary precursor to a functional mammary cell. We have also analyzed the promoter region of each of the genes coding for these proteins to identify many potential regulatory elements (hundreds) that may have contributed to their differential expression. It is likely that few, if any, of these elements were actually involved, making critical analysis of the probability of their involvement very important. The sheer number of potential elements makes analysis “by hand” impractical, and is the reason for constructing a searchable database.

At present, we have a list of roughly 40 genes, with up to ~700 elements per gene promoter, with many of these elements redundant within a gene promoter. This list is all from one cell model from Bos taurus (cow), though we will be adding to this with multiple cell models from multiple species in the future. We will also likely be expanding the gene list markedly with more experiments of the type we have been conducting, but also with different approaches that will lead to vastly larger data sets (on the order of thousands of genes as opposed to tens or hundreds). Before the end of Winter quarter 2012, we will be interested in adding ~120 more genes to the database from an experiment already in progress.

**Data to be managed**

The promoter analysis outputs are from an online program called Transcription Element Search System (TESS) that analyzes input DNA sequence (up to 2000 bp; user provided) and identifies short sequences within that larger sequence that have been identified as involved in transcription regulation in some experimental system and described in the scientific literature. The data to be added to the database are in the form of individual Microsoft Excel files for each gene named with the lower case designation for specie (b=bovine, m=murine) followed by the abbreviation for the gene name as listed by GenBank (e.g. CAPG=gelsolin-like capping protein). Within each Excel file are four worksheets. The “Job Parameters” worksheet describes the parameters of the search (user provided). The “Sequences” worksheet lists the user provided gene promoter sequence (the up to 2000 bp user provided sequence). The “Hits 1” worksheet contains most of the data to be managed, under the following headings:

- **Factor**: names/ID#s of the actual individual transcription factors themselves that have been identified associating with the particular model number (specific element sequence)
- **Beg**: Start of the site in the query sequence. Numbered from 1
- **Sns**: Sense of the site: N - normal, R - reverse complement
- **Len**: Length of the site
- **Sequence**: Model element sequence appearing in query sequence
- **La**: Log-likelihood score, higher is better.
- **La/**: La / Len, higher is better, maximum is 2.0.
- **Lq**: La / L_M, where L_M is the maximum La possible for the site model, higher is better, best is 1.0
- **Ld**: L_M - La, 0 is best, higher is worse.
- **LPv**: Approximate p-value for La score
Sc: Core similarity as reported at TRANSFAC site  
Sm: Matrix similarity as reported at TRANSFAC site  
SPv: Approximate p-value for Sm score  
P-v: Poisson-model p-value  
Model: Which site strings or weight matrix was used to pick this site

The “PSG 1” worksheet contains a Poisson analysis and summarizes the data on the “Hits 1” tab. This worksheet contains the following headings:

P-value: Poisson-model p-value  
N: Number of times the MAC was repeated in the provided promoter sequence  
Rate: (we don’t know what this is…)  
La: Log-likelihood score, higher is better.  
MAC: Model accession number; each MAC pertains to one specific sequence and is used to ID the model on the TESS website  
MID: Model ID number; used on the TESS website to identify a specific model  
FACs: Transcription factor accession numbers - used on the TESS website to locate the information on the TFs that associate with sequence of a particular model  
FIDs: Transcription factor IDs - names the transcription factors that associate with the sequence of a particular model

Desired functions of the database (use cases)

Within any gene, identify duplicate Factors as defined by identical Sequence and Beg, and using the “Division” distinction on the linked out Model information website, throw out duplicates based on “Division” value, keeping only Division:Mam entry if present, or if no Mam is present, Bir. If neither, don’t throw out any (user can then manually review).

Desired search outputs:
- All the genes that have a specific Factor or Factors
- List of Factors sorted by number of occurrences across all genes
- All genes with Factor or Factors in a certain defined region or regions
- Genes that share more than one specified Factor
- Genes that share over a certain number or percentage of Factors
- Ability to sort lists by “quality” of Factor identified using:
  - Higher La
  - Higher La/
  - Higher Lq
  - Lower Ld

For any given factor entry, link out to Model website from I, M or R number
I: http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=IMD-FRMREQ-Search  
M: http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=MTX-FRMREQ-Search  
R: http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=SIT-FRMREQ-Search
For Models having an R or M, link out to the T number listed under factor
http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=FCT-FRMREQ-Search

For a given gene (with its factor list), sort the rest of the genes by degree of similarity based on presence of the same Factors and Beg similarity for each of those common factors. This should have user defined constraints for “quality” measures for each Factor (La, La/, Lq, Ld)