Mammary-related gene promoter database
Peterson Lab Group, Cal Poly, January 2012

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).

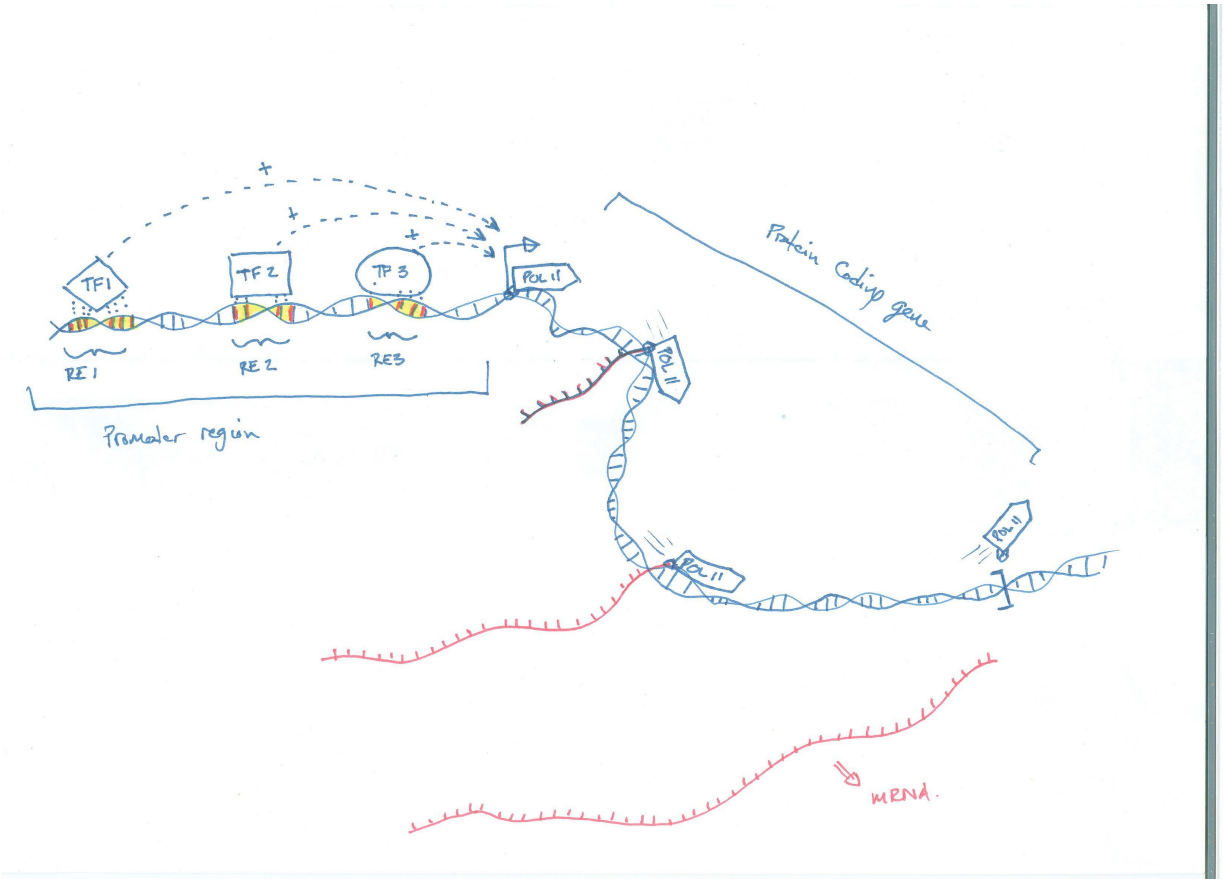


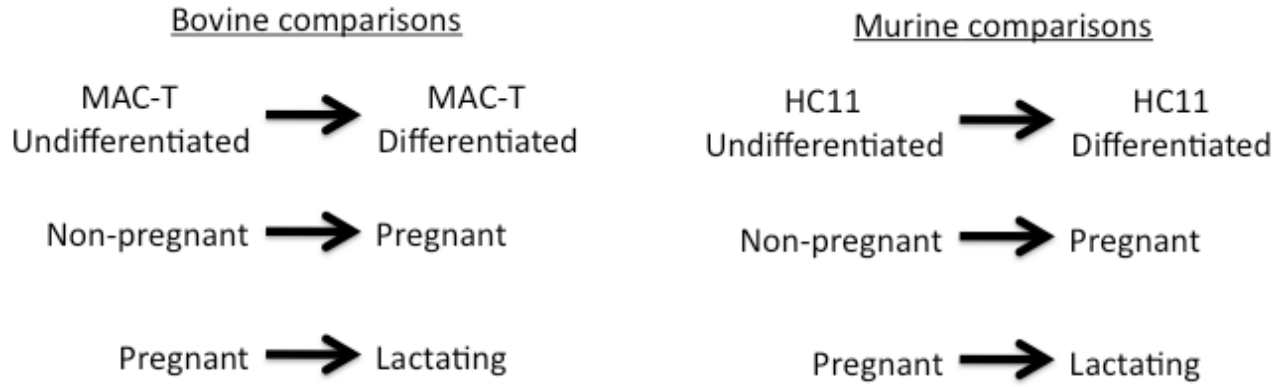
Figure 1. Schematic showing the promoter region of a gene with three regulatory elements (RE) each bound by their respective transcription factor (TF) contributing to the activation of transcription of the gene.

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 line from the bovine species (cow), though we will be adding to this with multiple cell sources 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.

What we've done

The experiments: Our approach is to study what changes occur in mammary cells from different sources during the transformation from one state to another state (e.g. pregnancy to lactation in mice, or undifferentiated to differentiated in the MAC-T bovine cell line). We do this by comparing global protein expression in the two states and identifying a subset of the proteins that change in abundance with the transformation from one state to the other. The data that we will begin with is from a cell line (MAC-T) from one species (cow, or "bovine"). We will later be adding more species (mouse or "murine", as well as pig or "porcine" and possibly human). For each species, we will also be including more than one comparison based on the different sources of cells we are comparing. For example, within the murine species, we will be comparing pregnant to lactating, nonpregnant to pregnant, and undifferentiated to differentiated HC11 cell line (3 separate comparisons). In the figure below, each arrow corresponds to a comparison. From each comparison, we generate a list of proteins that change in abundance as well as their direction and magnitude of change (though we are not including magnitude in our database).



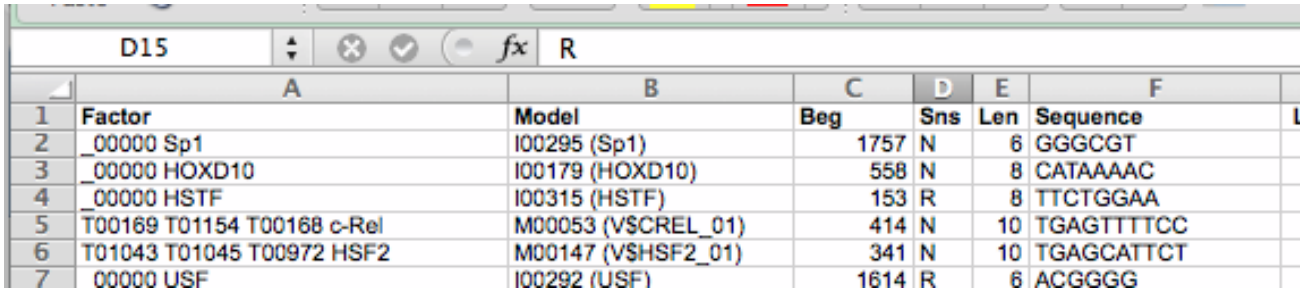
The gene promoters: Remember that each protein is the result of a gene being "expressed", and that the gene expression is regulated in part by sequences in the gene promoter interacting with activating proteins or "factors". Once we have the list of proteins that changed in abundance (up or down) in our comparison, we search for the gene for each protein in a genome database and find its promoter sequence. We then copy 2000 bases of that promoter sequence for analysis of the sequence to identify potential regulatory sequence elements within the 2000 base sequence.

Regulatory sequence search (TESS): We use an online program called Transcription Element Search System (TESS) that analyzes our input DNA sequence (the 2000 base promoter) and identifies short sequences that have been shown somewhere before to interact with a transcription factor to activate gene expression. The results of this search can be downloaded in a Microsoft Excel for each gene promoter that we used as a query.

TESS output to be managed: Each Excel file contains four worksheets, two with data that are user-provided. The first of these is **Job Parameters** and includes the TESS search parameters which should be stored only for posterity for each gene for future reference, but also includes the identity and date of the experiment that led to these data (**Experiment**), the specific cell types being compared (**Comparison**), the species from which the cells are derived (**Species**), the name of the gene (**Gene Name**) and its abbreviation (**Gene Abbreviation**), the location within the species genome that the promoter (2000 bases used in the query) exists (**Chromosome**, **Begin Site**, **End**

Site), and whether the protein corresponding to this gene was up- or down- regulated in the comparison (**Regulation**). The **Sequences** worksheet includes the 2000 bases that was used in the TESS query .

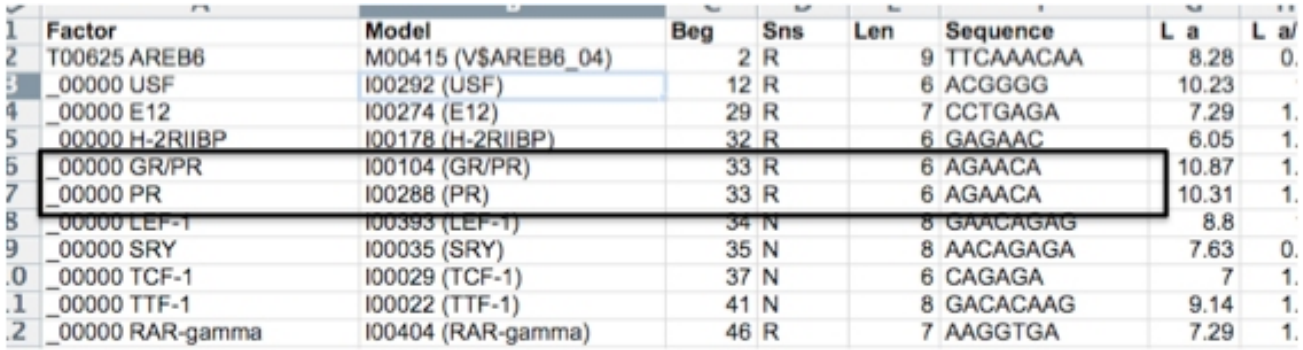
The majority of the results we are interested in at present are contained in the **Hits 1** worksheet which is a list of the potential regulatory element sequences and includes the name of the factor that should interact with the identified sequence (**Factor**), the identified sequence itself (**Sequence**), the start location of the sequence within the 2000 base input sequence (**Beg**), the direction that the sequence reads (i.e. forward (normal) "N" or reverse "R"; **Sns**), the length of the sequence in number of nucleotides (**Len**), identifiers that correspond to the studies that implicated the sequence as being involved in gene expression (**Model**) as well as several numerical values that are measures of likelihood that the sequence is "real" (**L a**, **L a/**, **L q**, **L d**, **L pv**, **S c**, **S m**, **S pv**, **P pv**). While the values associated with these parameters are important, their mathematical derivation is not important, so these can be stored just as numerical values. Any given sequence might appear in multiple locations within the 2000 base promoter, as is true of any given factor. Also, a single factor may have multiple sequences associated with it.



D15 | R

	A	B	C	D	E	F
1	Factor	Model	Beg	Sns	Len	Sequence
2	_00000 Sp1	I00295 (Sp1)	1757	N	6	GGGCGT
3	_00000 HOXD10	I00179 (HOXD10)	558	N	8	CATAAAAC
4	_00000 HSTF	I00315 (HSTF)	153	R	8	TTCTGGAA
5	T00169 T01154 T00168 c-Rel	M00053 (V$CREL_01)	414	N	10	TGAGTTTTCC
6	T01043 T01045 T00972 HSF2	M00147 (V$HSF2_01)	341	N	10	TGAGCATTCT
7	00000 USF	I00292 (USF)	1614	R	6	ACGGGG

A unique event is a particular Beg, Sns, Len and Model, however if the same Beg, Sns and Len occur more than once and differ only by the Model, then only the "best" of these repeats should be kept (discussed below under "use cases").



	Factor	Model	Beg	Sns	Len	Sequence	L a	L a/
2	T00625 AREB6	M00415 (V$AREB6_04)	2	R	9	TTCAAACAA	8.28	0.
3	_00000 USF	I00292 (USF)	12	R	6	ACGGGG	10.23	
4	_00000 E12	I00274 (E12)	29	R	7	CCTGAGA	7.29	1.
5	_00000 H-2RIIBP	I00178 (H-2RIIBP)	32	R	6	GAGAAC	6.05	1.
6	_00000 GR/PR	I00104 (GR/PR)	33	R	6	AGAACA	10.87	1.
7	_00000 PR	I00288 (PR)	33	R	6	AGAACA	10.31	1.
8	_00000 LEF-1	I00393 (LEF-1)	34	N	8	GAACAGAG	8.8	
9	_00000 SRY	I00035 (SRY)	35	N	8	AACAGAGA	7.63	0.
10	_00000 TCF-1	I00029 (TCF-1)	37	N	6	CAGAGA	7	1.
11	_00000 TTF-1	I00022 (TTF-1)	41	N	8	GACACAAG	9.14	1.
12	_00000 RAR-gamma	I00404 (RAR-gamma)	46	R	7	AAGGTGA	7.29	1.

The "PSG 1" worksheet contains a Poisson analysis and summarizes the data on the "Hits 1" tab. Most of this sheet is derivative of what is on the Hits 1 sheet and does not have immediate utility for us. However, we would like these data to be stored in the database "for posterity" so that they can be retrieved for a given gene if desired in the future. We are not entirely sure at this point how the numbers here are derived, so these can all be treated as numerical values. 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 factor and how it was found, 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 (if convenient, this entire column could be deleted).
FACs: Only present if Model begins with M. Transcription factor accession numbers - used on the TESS website to locate the information on the Factors 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)

Duplicate deletion: Within any gene, identify duplicate Factors as defined by identical Beg, Sns and Len but different Model 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/sort functions:

- 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 L a
  - Higher L a/
  - Higher L q
  - Lower L d
- Ability to limit search/sort functions to specific sets of Comparisons or Species

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)