

Lab 6 – *Cis* regulatory element mapping and prediction

[Software needed: web access]

In this lab we will examine *cis*-acting transcriptional regulation in the model plant *Arabidopsis thaliana*, and in human. This will include using online resources and repositories, as well as several prediction programs to examine the promoters of sets of coexpressed *Arabidopsis* and human genes.

One of the principle means of coordinating transcription spatially and temporally is through the presence of *cis*-regulatory elements. These short DNA sequence motifs, proximal to the transcriptional start site (TSS), are bound by transcription factors responsible for the recruitment of the transcriptional initiation machinery. *De novo* prediction of these binding sites has become an active area of research in the field of functional genomics. Global elucidation of gene-specific *cis*-regulatory control will permit the understanding of global transcriptional networks and facilitate our understanding of the “blueprint of life” and what is required for a properly functioning biological system.

In this lab we ask the question, “**Which (potential) *cis*-elements are present in the promoters of *Arabidopsis* AP3-coexpressed genes and human *INSULIN*-coexpressed genes?**”

Answering this question would help elucidate potential regulatory mechanisms for these genes. We’ll use several online prediction and mapping tools with the promoters of these coexpressed genes in the hope of characterizing the *cis*-elements common to each set of promoters.

Cis-element identification and prediction in *Arabidopsis* and human can be broken down into: i) databases of literature-documented and *in vivo*-tested *cis*-elements, PLACE and JASPAR; ii) graphical tools for the visualization of these elements; and iii) prediction tools for identifying which sequence motifs are significantly enriched within the promoter regions of sets of coexpressed genes.

Box 1. *cis*-element prediction methods

There are two main types of intrinsic *cis*-element prediction strategies, and a third one that is extrinsic, involving the use of comparative genomics. In the first two strategies, promoter sequences from groups of coexpressed genes are analyzed for the over-representation of certain subsequences, which could be potential *cis*-elements. At a basic level, these two strategies may be subdivided into word counting methods and probabilistic methods.

Word Counting

In this method, promoters from coexpressed genes are split into “*k*-mer words” (e.g. 6 nt chunks) and the number of instances of each word (e.g. AACTAC) is counted in the set of words generated from all of the coexpressed promoters. A background distribution of the number of times these words show up in sets of randomly chosen promoter is also computed. It is then possible to calculate the difference between these two values and apply a statistical test to see if the count seen for a word in the coexpressed set of promoters is significantly different from the count seen in the randomly selected set of promoters. This is the principle

behind Promomer. *Cis*-elements identified by this method are deterministic, i.e. no wobble nucleotides are allowed. Promomer will return all possible statistically overrepresented words in one run, and these will be consistent between runs.

Gibbs Sampling

The Gibbs sampling method (implemented in Motif Sampler) involves specifying the k -mer word size that should be found. The start position for each k -mer is randomly chosen in each promoter sequence. If we are dealing with 50 sequences, 50 k -mers would thus be “identified”. Of course, most of them will be different because their start positions were randomly chosen. A frequency matrix is then calculated for the As, Cs, Gs and Ts at each position, starting at position 1 and continuing up to the length of the k -mer. In the first sequence, the k -mer start position is then adjusted so that it better matches this frequency matrix (i.e. all the possible k -mers from that promoter are compared to the frequency matrix and the best match is chosen). The frequency matrix is updated with new values based on the better k -mer. Then the program moves to the 2nd sequence and adjusts the k -mer start so that it better matches the frequency matrix, and so on. The whole process is repeated a 50-100 times, after which the k -mers will converge towards a “consensus” sequence, which could represent a potential *cis*-element. It is important to note that Gibbs sampling will **always** identify a “consensus” sequence, but this may or may not be biologically relevant. It is thus important to compare the frequency of the “consensus” sequence in the set of coexpressed genes versus its frequency in a randomly selected set. The Gibbs sampling method is applied in MotifSampler. Note that *cis*-elements identified by Gibbs sampling will contain “wobble” nucleotides, i.e. are probabilistic in nature. Each run of a Gibbs sampling query will often return a different predicted *cis*-element, due to the random (stochastic) nature of where the k -mers are initially chosen to start.

Phylogenetic Footprinting/Comparative Genomics

In this case, orthologous promoters are obtained (this assumes that orthologous promoter sequences are available), and then conserved regions between orthologous promoters are identified. In practice, this is somewhat tricky as the evolutionary distance between the orthologs should not be too great that potential conserved regions have degenerated, but also not be too close that there is no difference observed. In the case of Arabidopsis, 90000 conserved non-coding sequences in the Brassicaceae were recently identified by Mathieu Blanchette and colleagues at McGill University (see Haudry et al. 2013, Nature Genetics, <http://dx.doi.org/10.1038/ng.2684>), and rapidly falling sequencing prices have enabled the sequencing of the genomes of several close human relatives, thereby enabling such an approach (although just a few nucleotide changes are enough to change expression patterns, see Wittkopp and Kalay, 2012, Nature Reviews Genetics, <http://dx.doi.org/10.1038/nrg3095>).

1. Promomer

Promomer is an online tool that identifies short, statistically over-represented nucleotide sequences within the promoters of coexpressed genes. Algorithmically, the program exhaustively identifies all possible k -mers within the coexpressed promoters and then, for each possible k -mer: 1) generates 1000 random promoter clusters of equal size to the coexpressed cluster, 2) generates 1000 clusters by resampling the coexpressed promoters themselves, 3) counts the number of k -mer instances in all clusters and 4) compares the frequency distributions between the

coexpressed promoters and randomly chosen promoters clusters for each k -mer. The length of the k -mer can vary from 4nt to 10nt and the number of promoters in which the k -mer must occur to be considered for further analysis can range from 50-100% of the promoters. Promomer can identify short k -mers that function as core sequences for larger motifs occurring as general transcriptional enhancers.

Go to the Promomer application at http://bar.utoronto.ca/ntools/cgi-bin/BAR_Promomer.cgi. In the Promomer window, click on the radio button for the 2nd application, and paste the 25 AGI IDs for *AP3*-coexpressed genes from **List 1** (as determined using the BAR's Expression Angler algorithm in the previous lab) into the input box. Set the search for 6mers (default is 4) in 75% of the genes and click 'Submit Query'. When Promomer finishes running, click on "View graphical interpretation of results" and read the output description at the top of the page.

At3g54340
At5g20240
At1g59640
At2g38110
At2g41540
At4g32460
At1g16705
At2g34340
At5g20270
At1g35170
At5g40350
At3g20820
At3g27810
At4g35110
At2g15530
At3g18850
At3g01980
At1g75290
At1g23060
At1g78170
At1g55510
At4g34588
At1g70720
At4g24130
At2g42900

List 1. *AP3*-coexpressed AGI IDs

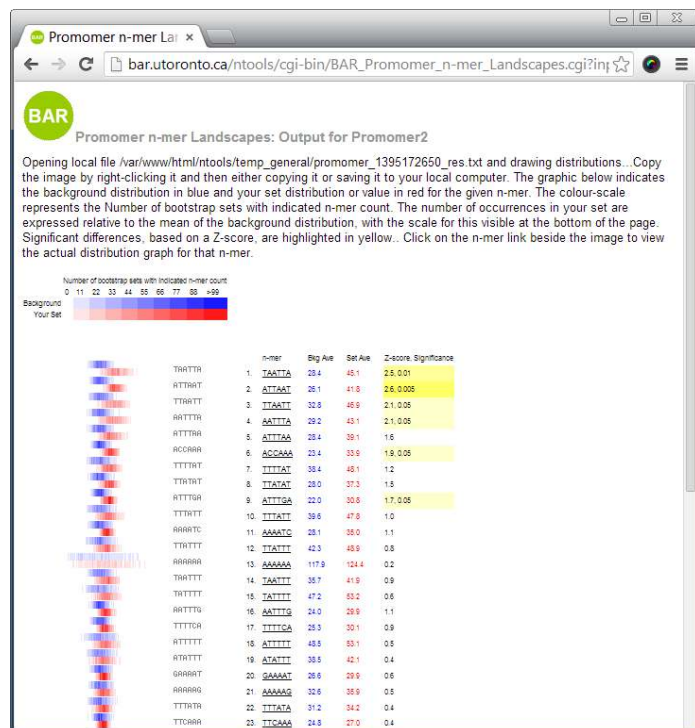


Figure 1. Output of Promomer Program

a. What colour is the background distribution? And the coexpressed cluster distribution?

Look through the list, noting those pairs with significantly disparate distributions (highlighted yellow) and their associative critical values (z-scores and p-values). Here the z-score is a critical statistic applied to a Gaussian probability distribution. (At a $z = 2.5$, $p = 0.05$, and a $z = 2.95$ translates to $p < 0.001$).

Examine the 6mers found to be highly significant.

b. Do any of the k -mers appear similar? I.e., could they partially overlap?

Click on the sequence for the top two significant k -mers and view the k -mer-specific output page. Note the distance between the two distributions and read the annotation information for the graph to the right of the distribution plot. Browse through the PLACE annotation list below the distribution graph.

c. Does these motifs seem to match any previously characterized motifs? If so, which one(s)?

Interestingly, it has been shown using reporter constructs that AGAMOUS (AG) binds to the AP3 promoter (see Hill et al., 1998, Development <http://dev.biologists.org/content/125/9/1711.short>). Examine the annotation pages for several other significantly enriched k -n to see if any other overlap with known binding motifs.

2. Athena

Athena maps *literature identified motifs* from the PLACE DB to user-provided promoters IDs and uses a hypergeometric test to assess the enrichment of each motif (similar to the test we used to assess GO enrichment in a couple of the other labs). Three things to note about Athena: first, PLACE is a database of consensus sequences (e.g. ACGT[TG][CA]), not position specific matrices, so it is hard to ascertain how good a match is – it is either a match or it is not; second, PLACE hasn't been updated for a while; and third, the hypergeometric test might be appropriate for this kind of database, but a better test would be to generate random promoter data sets equal in size to an input set and ask how often an identified pattern occurred in those. Paste the AGI IDs from **List 1** into <http://www.bioinformatics2.wsu.edu/Athena/> - click on the Visualization tab.

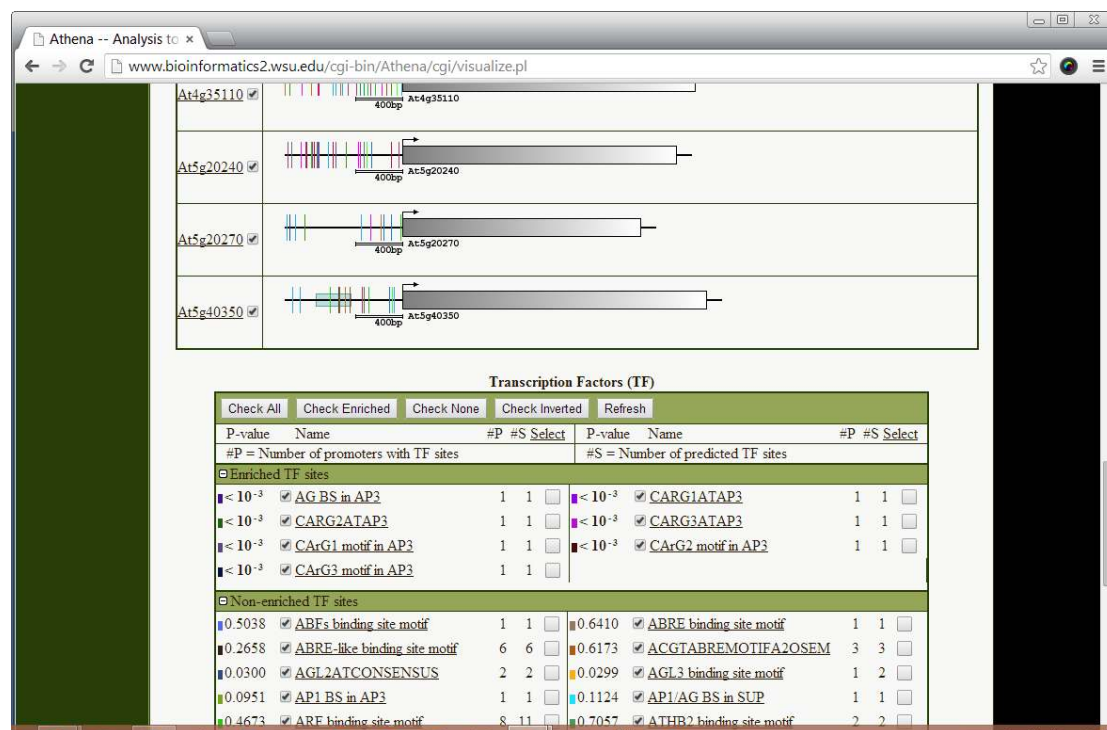


Figure 2. Output of Athena.

Once there, make sure you have your AGI IDs listed in the Accessions box and click on “Display”, keeping the default settings. When Athena finishes running, notice the gene model for all your inputted AGI IDs. The vertical coloured lines correspond to *cis*-element consensus sequences that have matched in that gene’s promoter region. Mouse over some of the *cis* sites in the At3g54340 promoter (i.e., the *AP3* promoter), keeping your mouse still for a second, to see the annotation information pop-up boxes.

d. Comment on the number of cis-element matches in each promoter.

Scroll down to the bottom of the screen and examine the box titled “Transcription Factors (TF)”. Notice the upper selection box for “Enriched TF sites”. These sites are statistically more prevalent in the coexpressed promoters than in all promoter regions of the genome, as determined using a hypergeometric test. Click on each of the motifs in the “Enriched TF sites” box to see the consensus sequence and annotation information for that motif.

e. What can you infer about the identified significantly enriched motifs?

3. JASPAR and Cistome

Let’s use another tool to map these motifs to the promoters of our set of *AP3*-coexpressed genes – this additional information will help us decide if the significant motifs overlap or not. Further, we will test the hypothesis that AG binding sites are enriched in the promoters of our *AP3*-coexpressed genes by retrieving position-specific scoring matrices for AG and AG-like transcription factors from an open-access online database called JASPAR.

Go to the Cistome tool at http://bar.utoronto.ca/cistome/cgi-bin/BAR_Cistome.cgi. Check the “Consensus sequences” button and enter the first two motifs from your Promoter analysis as follows into the first input box:

```
>motif1
TAATTA
>motif2
ATTAAT
```

Check “Significance testing with Bootmer2” and change the Z-score cutoff to 2 from the default setting of 3. Paste the AGI IDs from **List 1** on Page 3 into the second input box in the second section. Select 1000 bp as the sequence length. Click Map at the bottom of the page.

f. Do you think that these two motifs are part of a larger motif?

You can easily assess whether motifs overlap by checking the Merge option towards the bottom of the page with both motifs selected then clicking redraw.

Let’s test the hypothesis that AGAMOUS binding sites are enriched in the promoters of our *AP3*-coexpressed genes. JASPAR is an expertly curated, open access database of transcription factor binding sites (TFBSs) for many transcription factors (TFs) for many species. Unlike other TF databases, such as TransFac, JASPAR is not under any usage restrictions.



Figure 3. Output of Cistome.

Go to <http://jaspar.genereg.net/> and scroll down to the search area towards the bottom of the page. In the **Search by** section, select Species for the first box and type 3702 (this is the NCBI's taxonomic identifier for *Arabidopsis thaliana*) and click Search. You'll see a list of the PSSMs for all *Arabidopsis* transcription factors in the JASPAR database. You could actually select all of them and scan the promoter of *AP3* to see if any of the PSSMs match to sequences in the *AP3* promoter but you'd find that a lot (more than 100) of them do, just as we saw for the Athena PLACE scan. Instead, search through the list for AG or AGL (AGAMOUS-like) PSSMs.

g. What are the JASPAR identifiers for the two AG PSSMs, and the AGL15 PSSM?

If you click on the SeqLogo for these three PSSMs, you'll see detailed information as to the specificity of the binding, the experimental method used to identify the binding specificity, and a literature reference.

h. How was the specificity for the second AG PSSM determined (ID ending in .2)?

Lab Quiz
Question 2



Figure 4: JASPAR PSSMs for *Arabidopsis thaliana*, with PSSM for AG highlighted.

We'll redo the analysis we did on Page 5, except we'll use the PSSM option instead of the Consensus sequence option. Go to the Cistome tool at http://bar.utoronto.ca/cistome/cgi-bin/BAR_Cistome.cgi. Check the "PSSMs" button and enter these reformatted motifs from JASPAR for AG and AGL15 as follows into the first input box:

```
> AG-1
0 2 38 61 73 31 23 20 22 1 27
87 86 9 5 1 3 16 13 0 0 25
2 0 7 2 1 4 17 26 64 82 9
1 2 36 22 15 52 34 31 4 7 29
> AG-2
21 9 10 29 0 0 31 47 52 25 17 19 7 2 22 45 40 15
20 3 0 8 66 65 3 2 0 0 15 8 0 0 17 4 6 10
6 3 1 8 0 0 6 0 1 1 11 20 57 54 5 5 9 16
19 51 55 21 0 1 26 17 13 40 23 19 2 10 22 12 11 25
> AGL15
9 4 18 0 4 63 15 33 3 15 14 18 2 54 80
5 10 0 144 99 20 23 4 0 9 14 0 0 30 10
3 6 11 2 1 18 16 3 0 8 27 124 110 30 11
133 130 121 4 46 49 96 110 147 118 95 8 38 36 49
```

Check “Significance testing with Bootmer2” and change the Z-score cutoff to 2 from the default setting of 3. Paste the AGI IDs from **List 1** on Page 3 into the second input box in the second section. Select 1000 bp as the sequence length. Click Map at the bottom of the page.

i. *Would it appear that the promoters of our AP3-coexpressed genes are enriched in AG or AGL15 binding sites?*

Hmmm...not really! So let's try to predict some novel motifs for this set of a gene promoters. We'll also try to predict motifs for a set of INSULIN-coexpressed genes from human.

3. MEME

Meme applies an Expectation Maximization algorithm to the prediction of *cis*-regulatory elements. It can be installed for local uses on a Linux platform, but here we'll be using an online implementation to predict *cis*-elements in two sets of coexpressed genes.


Go to <http://meme.ebi.edu.au/meme/cgi-bin/meme.cgi>, which is the submission page for the MEME program (there are several programs that are part of the MEME Suite).

Use the *AP3*-coexpressed gene promoters file we've provided on the Week Navigator page and enter this into the “actual sequences” box of the MEME Data Submission Form. Change the distribution of motifs to **Any number** of repetitions, enter your email address, and provide a description for your sequences, if desired. Click Start Search at the bottom of the page.

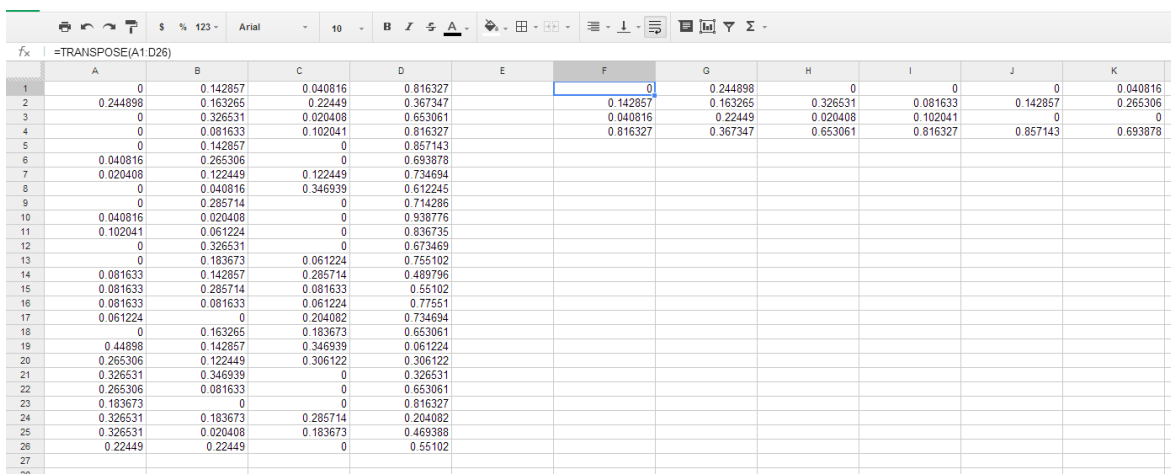
Figure 5: MEME input page.

Wait a few minutes (you'll receive an email when the job is done, or click on the link on the submission page, and the results will appear there when ready). On the output page, you'll see an overview of some motifs, with statistics associated with them, and then detailed information on these motifs, such as where they are located, the sequence composition, and so on.

j. Are there any “good” motifs discovered by MEME in the promoters of the AP3-coexpressed genes? You'd be looking for a high number of occurrences, a low p-value and a relatively long motif.

Let's see if this motif matches anything in JASPAR. We'll need to generate a matrix of nucleotide frequencies for the first motif, which is the one with the best score and which looks most promising. On the output HTML page in the **Data Formats** section, click on the  PSPM option. You'll see a section pop up with rows of values. These represent the frequencies of the nucleotides A, C, G, and T (the four values, respectively, per row) at each position in the motif. We'll need to do a bit of text editing to convert these into a format that we can use to search the matrices in JASPAR.

1. Copy the contents of the text box (excluding the header line) into a text editor, such as Crimson. Search and replace the spaces (enter a space character into the replace box) with a tab character, denoted by \t. Check the regular expression box, and replace all instances of spaces with the tab character.
2. Then, open a new spreadsheet in Google Docs at <http://docs.google.com>. **CREATE** a new spreadsheet, then copy the text, now with tab characters, into the spreadsheet. You should see 4 columns of values, with 26 rows as in Columns A-D of **Figure 6**.
3. As JASPAR requires the rows to be the frequencies of each nucleotide, instead of the columns, use the TRANSPOSE function to transpose the matrix. Type =TRANSPOSE(into a field beside the original data (e.g. F1 in **Figure 6**) and then select the initial four columns and 26 rows. Execute the function by closing the parenthesis “)” and hit ENTER on your keyboard. You'll see the rows and columns transposed. You can also do this in Excel with the Paste Special – Transpose checkbox.
4. Then copy the newly created matrix of 4 rows and 26 columns onto your clipboard.



	A	B	C	D	E	F	G	H	I	J	K
1	0	0.142857	0.040816	0.816327		0	0.244898	0	0	0	0.040816
2	0.244898	0.163265	0.22449	0.367347		0.142857	0.163265	0.326531	0.081633	0.142857	0.265306
3	0	0.326531	0.020408	0.653061		0.040816	0.22449	0.020408	0.102041	0	0
4	0	0.081633	0.102041	0.816327		0.816327	0.367347	0.653061	0.816327	0.857143	0.693878
5	0	0.142857	0	0.857143							
6	0.040816	0.265306	0	0.693878							
7	0.020408	0.122449	0.122449	0.734694							
8	0	0.040816	0.346939	0.612245							
9	0	0.285714	0	0.714286							
10	0.040816	0.020408	0	0.938776							
11	0.102041	0.061224	0	0.836735							
12	0	0.326531	0	0.673469							
13	0	0.183673	0.061224	0.755102							
14	0.081633	0.142857	0.285714	0.489796							
15	0.081633	0.285714	0.081633	0.55102							
16	0.081633	0.081633	0.061224	0.77551							
17	0.061224	0	0.204082	0.734694							
18	0	0.163265	0.183673	0.653061							
19	0.44898	0.142857	0.346939	0.061224							
20	0.265306	0.122449	0.306122	0.306122							
21	0.326531	0.346939	0	0.326531							
22	0.265306	0.081633	0	0.653061							
23	0.183673	0	0	0.816327							
24	0.326531	0.183673	0.285714	0.204082							
25	0.326531	0.020408	0.183673	0.469388							
26	0.22449	0.22449	0	0.55102							
27											
28											

Figure 6. Transposing rows/columns in Google Spreadsheet with the TRANSPOSE function.

Now we can search the matrices in JASPAR: Go to <http://jaspar.genereg.net/> and paste the PSPM matrix into the **Align to custom matrix or IUPAC string** box, and click Align. After a few seconds, you'll see an output like that of **Figure 7**. The matches are sorted in descending order of degree of match to the query matrix. Keep in mind that the NCBI Species code for *Arabidopsis* is 3702.

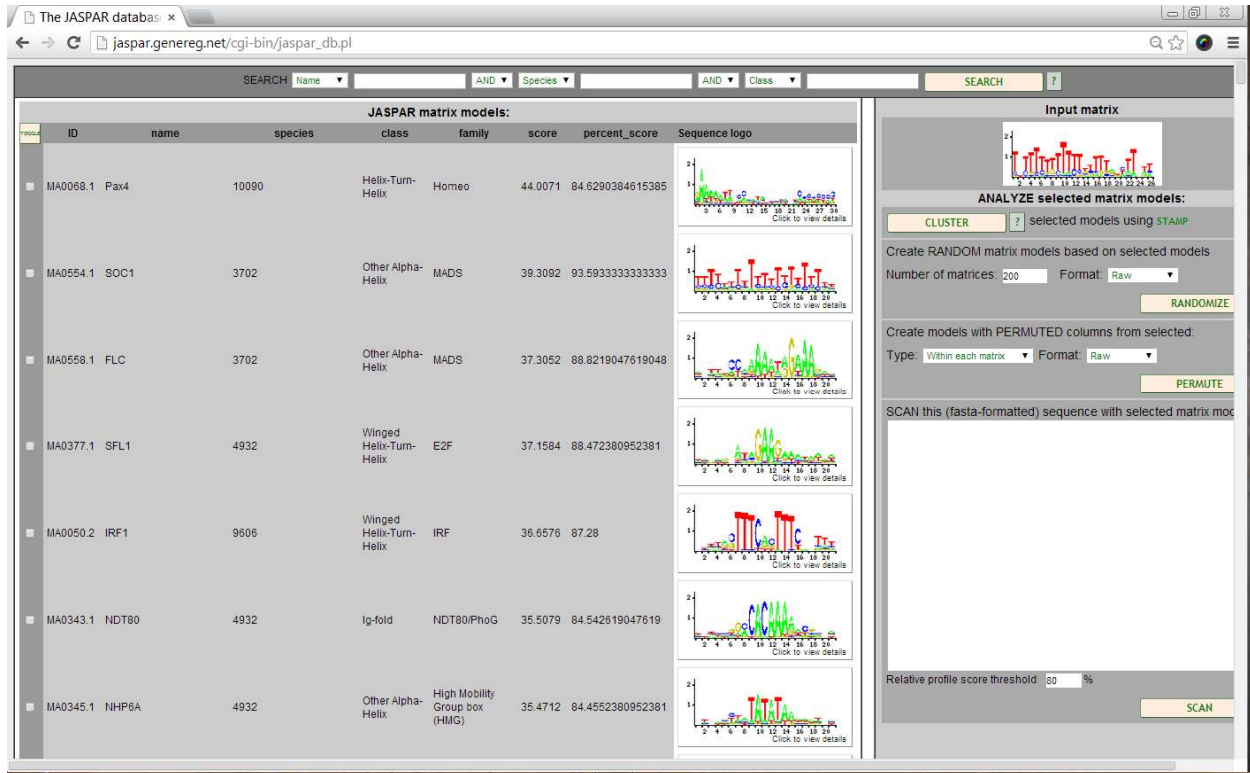


Figure 7: Output of a JASPAR custom matrix search.

k. What is the identity of the first *Arabidopsis* matrix match?

There's a considerable amount of literature on SOC1 (search for *SOC1 AP3 Arabidopsis* at <http://scholar.google.com>), which is a positive regulator of flowering. It looks like the promoters of our *AP3*-coexpressed genes contain SOC1-like binding element, which would make sense biologically.

Now we'll explore the promoters of genes that are coexpressed with *INSULIN*. We can use the BioGPS.org to search for and explore expression patterns for the insulin gene, *INS*, as described last class. To generate a list of insulin-coexpressed genes, the correlation tab was used at a cut-off of 0.9 to identify genes that are coexpressed with the *INSULIN* gene. This list was downloaded and the gene identifiers were used to retrieve 1000 bp of the region surrounding the start of transcription for 8 genes coexpressed with the insulin gene and that of the insulin gene itself, using Michael Zhang's promoter retrieval tool at <http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=searchPromForm>. 700 bp upstream and 300 bp downstream of the transcriptional start site was retrieved for each of the 9 genes, and these are available in the file on the Week Navigator.

Confirm that these genes are in fact mostly expressed just in the islet cells of the pancreas (those are the cells where insulin is produced) using the Human eFP Browser tool at http://bar.utoronto.ca/efp_human/cgi-bin/efpWeb.cgi. Select the “Skeletal Immune Digestive” Data Source, and type the gene name from the promoter file (e.g. INS) into the Primary Gene ID box and click Go.

Follow the instructions above the figure on Page 8 to submit these promoters to a MEME analysis (choose **any number** of repetitions of a motif per sequence, with the number of motifs set to 5 this time, otherwise use the default values).

1. Is there anything peculiar about the first motif?

Lab Quiz
Question 3

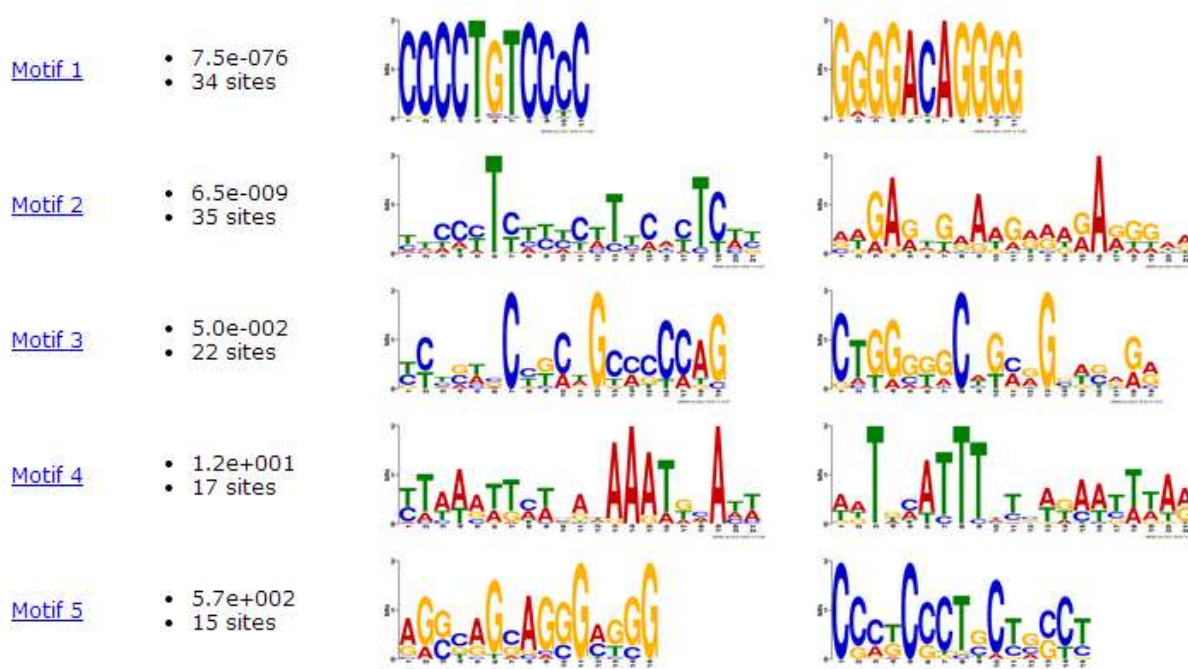


Figure 8: Motifs discovered in the promoters of *INS*-coexpressed genes.

Explore the top three motifs using JASPAR to see if they are similar to any known motifs.

m. Does the 1st motif resemble a motif present in JASPAR? (The NCBI code for human is 9606)

There are many known motifs present in the promoter of the *INSULIN* gene, as summarized by Kay and Docherty in 2006 in the journal *Diabetes* (see <http://diabetes.diabetesjournals.org/content/55/12/3201.full#sec-6>, summarized in Figure 1 of that paper), but none of these seem to be the same as the ones we’ve discovered in the *INS*-coexpressed gene set. Thus these elements maybe be required for tissue specificity instead of response to glucose levels or other metabolites.

We could design a construct containing several copies of our putative *cis*-element of interest upstream of a minimal promoter, driving the expression of a reporter gene, then create transgenic

Copyright © 2014 by D.S. Guttman and N.J. Provart

plants or animals to test whether or not the putative *cis*-element is in fact able to direct gene expression in the manner we expect. This would be necessary to confirm that our putative *cis*-element is active *in vivo*. Absence of reporter gene expression doesn't mean, however, that the *cis*-element is not functional, as another nearby element might be necessary to recruit a co-factor required for expression.

Some closing notes: here we have seen different ways of searching for elements in promoters, and how some databases store representations of these elements as regular expressions (similar to the way described in the first lecture of this course for patterns in protein sequences), while others, like JASPAR, store representations that include the binding specificity, that is, profiles. You might imagine that the latter is a better method. But while this is an improvement, we are still a long way from understanding the complex events that occur at the promoters of genes to bring about specific spatio-temporal patterns of gene expression.

End of lab!

Lab 6 Objectives

By the end of Lab 6 (comprising the labs including their boxes, and the lectures), you should:

- know the main technologies for identifying transcription factor binding sites *in vitro*;
- be able to name some methods for identifying potential transcription factor binding sites *in silico*;
- understand how word-count and Gibbs sampling methods work and the differences between these two methods;
- understand why we would want to generate results which referenced a background set of all promoters;
- know how to use online tools, such as Athena or JASPAR to identify known *cis*-elements in promoters of coexpressed genes;
- be able to use online versions of MEME.

Do not hesitate check with the online forums if you do not understand any of the above after reading the relevant material.

Further Reading

Bailey T.L., Elkan C (1995). Unsupervised Learning of Multiple Motifs In Biopolymers Using EM. *Mach. Learn.* 21:51–80.

Bailey TL (2007). Discovering sequence motifs. *Methods Mol Biol.* 395:271-92.

Mathelier A ~ Wasserman W (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, <http://dx.doi.org/10.1093/nar/gkt997>.

O'Connor TR, Dyreson C, and Wyrick, JJ (2005) Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics*, <http://dx.doi.org/10.1093/bioinformatics/bti714>.

Toufighi K, Brady S, Austin R, Ly E, Provart N (2005). The Bio-Analytic Resource: e-Northerns, Expression Angling, and promoter analyses. *The Plant Journal.* 43, 153-163.