

Inference with Gene Expression and Sequence Data

BMI/CS 776

www.biostat.wisc.edu/~craven/776.html

Mark Craven

craven@biostat.wisc.edu

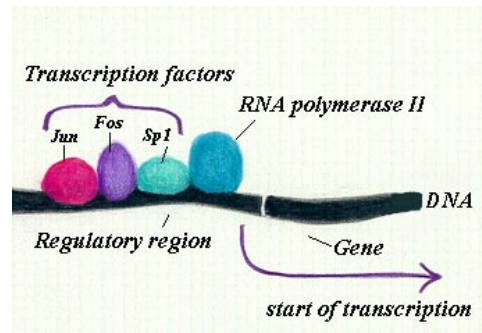
April 2002

Announcements

- HW #3 ready
 - HMMs for recognizing genes in *E. coli*
 - due April 29
- exam
 - 6pm Monday 5/6 or Tuesday 5/7
- plan
 - text analysis (on Wednesday)
 - Thomas Anantharaman guest lecture (next Monday)
 - SCFGs for RNA modeling (2 lectures)
 - modeling cellular systems (1 lecture)
 - ontologies and data integration (1 lecture)
- reading for Wednesday
 - Genes, Themes and Microarrays, Shatkay et al., *Proc. of ISMB* 2000

Gene Expression + Sequence Data

- Brazma et al., *Genome Research*, 1998
- task: predict transcription factor binding sites in DNA



- can do this using only sequence data
- but can get additional evidence using expression data

Describing TF Binding Sites

- various representations have been used to describe these motifs
 - n-mers (fixed length words)
 - position-specific probabilities (e.g. MEME)
 - regular expressions
- Brazma et al. use restricted class of regular expressions
 - . matches any character
 - [] matches any character inside of brackets
- example: **A**[**TG**].**C** matches all strings that
 - start with **A**
 - followed by either **T** or **G**
 - followed by any character
 - followed by **C**

Predicting TF Binding Sites

- given:
 - a cluster C of genes with similar expression profiles
 - 300 base upstream sequences for all genes
- do:
 - find motifs that occur more frequently in sequences upstream from genes in C than upstream from other genes

Extracting Reg-Ex Patterns

- Brazma et al. build a *suffix trie* for all upstream sequences
- here is a suffix trie for **ATACATAS**



- like a suffix tree but each edge represents the addition of just one character

Extracting Reg-Ex Patterns

- represent all sequences in suffix trie
- at all nodes maintain a list of matching subsequences
- during construction, prune trie at a node if
 - pattern has fewer than t occurrences
 - pattern matches fewer than p positive sequences
 - etc.



Scoring TF Binding Sites

- the scoring method used by Brazma et al. is essentially a likelihood ratio

$$\text{Score}(P, S_+, S_-) = \frac{\Pr(P/S_+)}{\Pr(P/S_-)}$$

P a given pattern

S_+ a sequence from the positive class

S_- a sequence from the negative class

Experiment

- Brazma et al. did this for one clustering of the yeast diauxic shift expression time series
- patterns extracted in two restricted reg-ex representations
 - at most 3 wild cards, no *group* characters
 - at most one group character (two alternatives), plus wild cards
- validated patterns by comparing against TRANSFAC, a DB of known transcription factor binding sites

High-Scoring Patterns for One Cluster

Table 4. Highest Scoring Patterns for the Cluster (05,2,40)0000022)

Pattern	N ^a	Total ^b	Score ^c	TRANSFAC (exact matches)
A. Highest score in experiment allowing patterns to have at most 3 wild cards and no group characters				
CCCC.T	22	27	1.09	YSDOR2_01, YSDOR2_02, YSTP1_02
A..GGGG	22	27	1.09	-
GGGGC	20	27	4.09	YSGAL2_02, YSSUC2_02, YSRBNA_01, YSERG11_01
CCCC	20	27	4.09	YSCYB2_02
G..GGGG	19	28	3.73	YSCYC1_04, YSCYC1_05, YSCYC1_06
CCCC..C	19	28	3.73	YSGAL3_05, YSMAL2R_01
CCCC..T	25	42	3.65	YSSUC2_01, YSDOR2_01, YSDOR2_02, YSTP1_02, YSGAL3_01, YSGAL4_01, YSMAL2R_01, YSMAL3_01, YSPDC1_02, YSHAP4_01
A..GGGG	25	42	3.65	YSSUC2_02, YSRBNA_01, YSERG11_01, YSMR11_02, YMPY1_01
CCCC	25	38	3.03	YSDOR2_01, YSDOR2_02, YSTP1_02
AAGGG	25	38	3.03	YSCAR1_02
CCCC..TT	19	22	2.95	YSDOR2_01
AA..GGGG	19	22	2.95	-
GGG..TG	20	21	2.93	-
CA.CCC	20	21	2.93	YSGAL1_04, YSCYC1_12, YSGAL1_14, YSDOR2_02, YSTP1_02
B. Highest score in experiment allowing patterns having at most one group character with two alternative letters (all pairs allowed)				
CCCC[GT]	20	28	3.96	YSDOR2_01, YSDOR2_02, YSTP1_02
CCCC[AT]	20	24	3.58	YSDOR2_02, YSTP1_02
CC[GG]CC	24	47	3.27	YSCYB2_02, YSGAL2_02, YSSUC2_02, YSRBNA_01, YSERG11_01
CCCC[CT]	29	58	2.94	YSDOR2_01, YSDOR2_02, YSTP1_02, YSSUC2_02, YSCAR1_02, YSRG11_01
[AG]CCCC	29	48	2.90	YSCYB2_02, YSDOR2_02, YSTP1_02, YSCYC1_04, YSCYC1_05, YSCYC1_06, YSGAL2_02, YSSUC2_02, YSRBNA_01, YSCAR1_02, YSERG11_01, YSGAL1_15

Discussion

- genes in a cluster may not be regulated by same TPs
 - noise in expression data
 - subjectivity of clustering
 - coincidence in expression does not necessarily imply same regulation mechanisms