SVMs and Probabilistic Approaches for Classifying Promoters

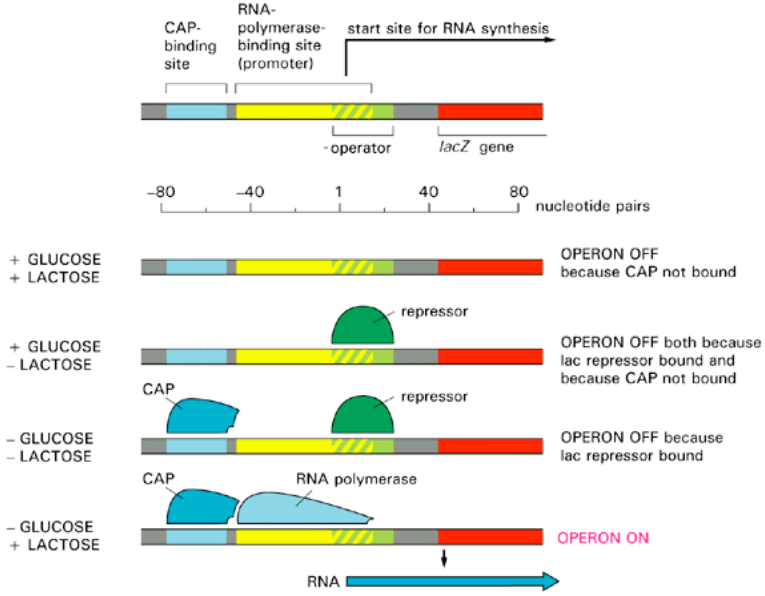
Anirvan M. Sengupta
Physics Dept./ BioMaPS Inst.
Rutgers University

Plan of the Talk

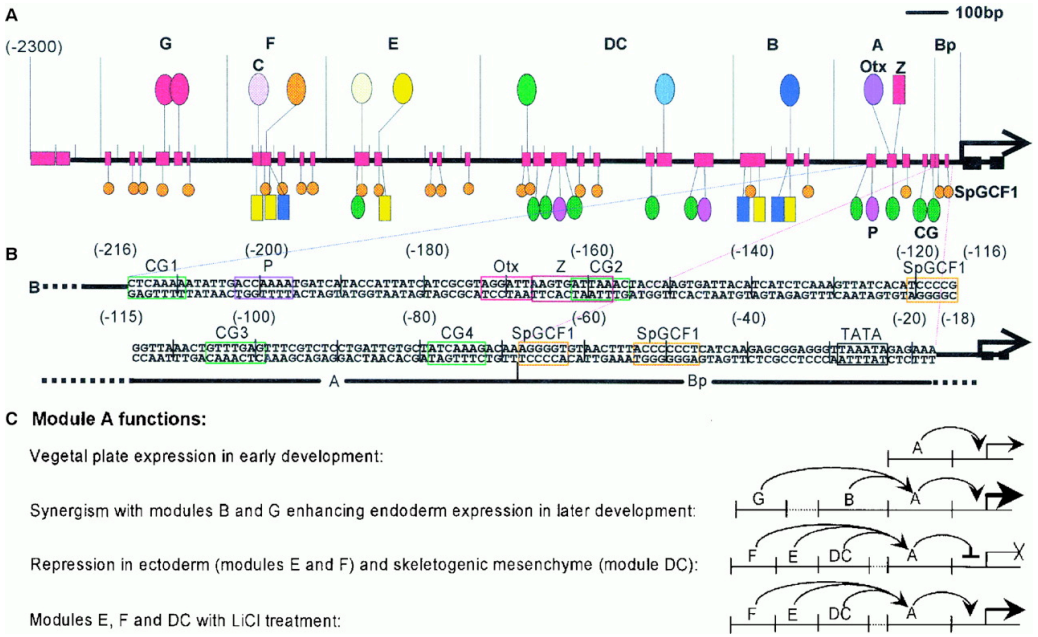
- The problem
- “Philosophical” issues
- Probability models and support vector machines
- Issues in combining heterogeneous input data

Transcriptional Regulation: Promoters as Computing Devices

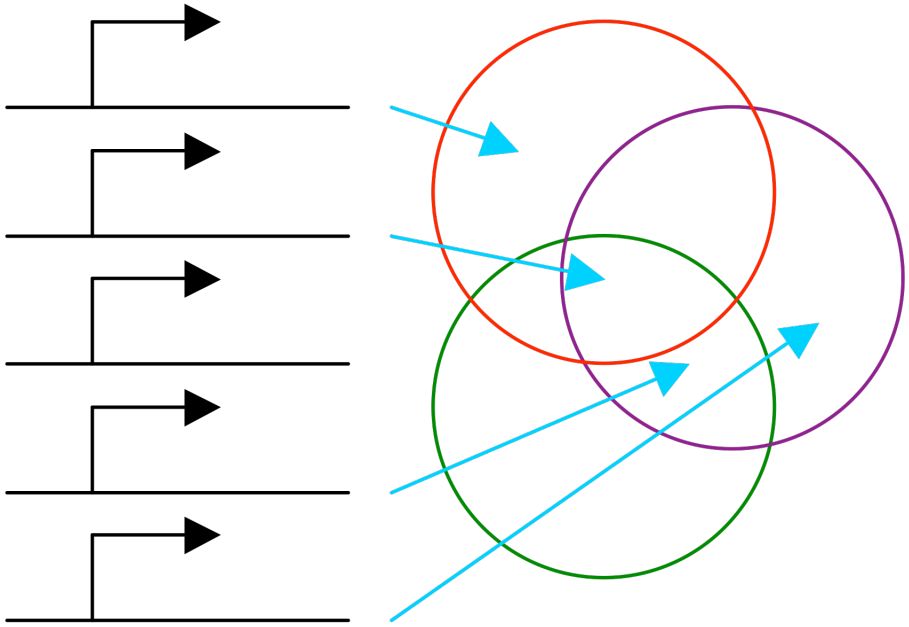
E. coli



Sea Urchin



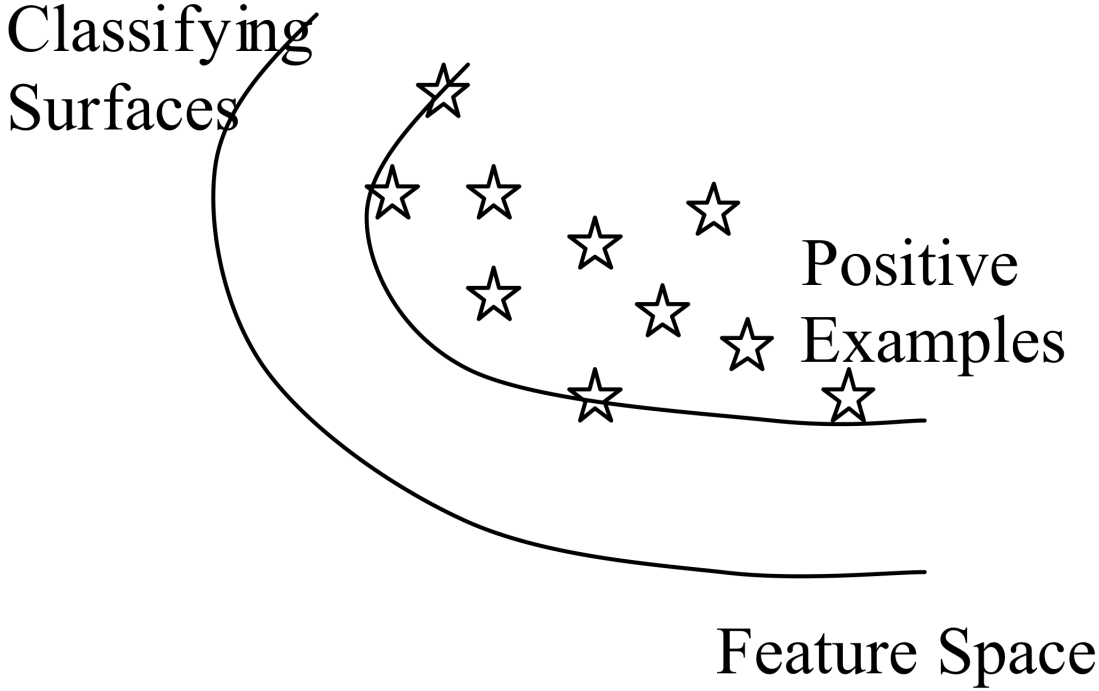
Classifying Promoters



Classifiers for Biological Problems

- Accuracy versus interpretability
- Off-the-shelf versus domain-specific tools
- Perfect versus imperfect labeling

Positives-only Data: Big picture



Wrap the “tightest” surface around the known examples!

An Example: Identification of Targets of a Transcription Factor

Supervised learning problem:
Find all functional targets of a factor in the genome from the knowledge a few examples.

Some Known Targets of lacI and of crp

lacI binding sites:

AATTGTGAGCGGATAACAATT
AAATGTGAGCGAGTAACAACC
GGCAGTGAGCGCAACGCAATT

some crp binding sites:

.....
TAATGTGACGTCCTTTGCATAC
GAAGGCGACCTGGGTCATGCGA
GGTG TTAAATTGATCACGTTTC
GATG CGAGGCGGATCGAAAAA
AAA TTCAATATTCATCACACTT
.....
TTTTGCGATCAAAATAACACTT
AAACGTGATCAACCCCTCAATT
TAATGTGAGTTAGCTCACTCAT
AATTGTGAGCGGATAACAATTT
.....

Consensus:

AAATGTGATCTAGATCACATTT

Describing Fuzzy Motifs

Weight Matrix

[Berg, vonHippel,Studen,Stormo,…]

Given a set of known factor binding motifs, like,

TAATGTGACGTCCTTTGCATAC
GAAGGCGACCTGGGTCATGCTG
CGATGCGAGGCGGATCGAAAAA
……
……
ATTTGAACCAGATCGCATTA
AAATGTAAGCTGTGCCACGTTT

construct a frequency matrix n_{ib}

Position	1	2	3	……	22
A	3	4	5	……	3
C	2	1	1	……	2
G	2	2	2	……	2
T	2	2	2	……	2

Fuzzy Motifs

Weight Matrix Continued…

Calculate weights by taking logarithm : $w_{ib} = \log(n_{ib} / n_s)$

For any sequence S, the score W is given by: $W = \sum_{ib} w_{ib} S_{ib}$

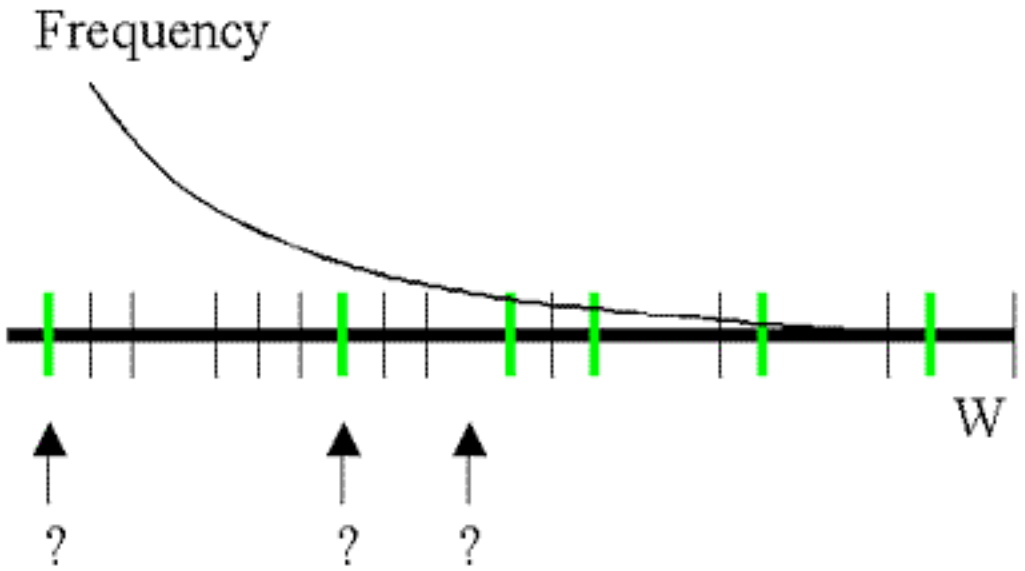
For example:

$W(TTAGCA\ldots..) = w_{1T} + w_{2T} + w_{3A} + w_{4G} + w_{5C} + w_{6A} + \cdots\cdots$

Sequences with higher W are better binders.

Precise relationship with binding energy in certain limits.

Problem of Threshold Selection



Independent Base Model

$$E(TTAGCAA) = \varepsilon_{1T} + \varepsilon_{2T} + \varepsilon_{3A} + \varepsilon_{4G} + \varepsilon_{5C} + \varepsilon_{6A} + \varepsilon_{7A}$$
$$= \Sigma_{ib}\, \varepsilon_{ib}\, S_{ib} = \varepsilon \bullet S$$

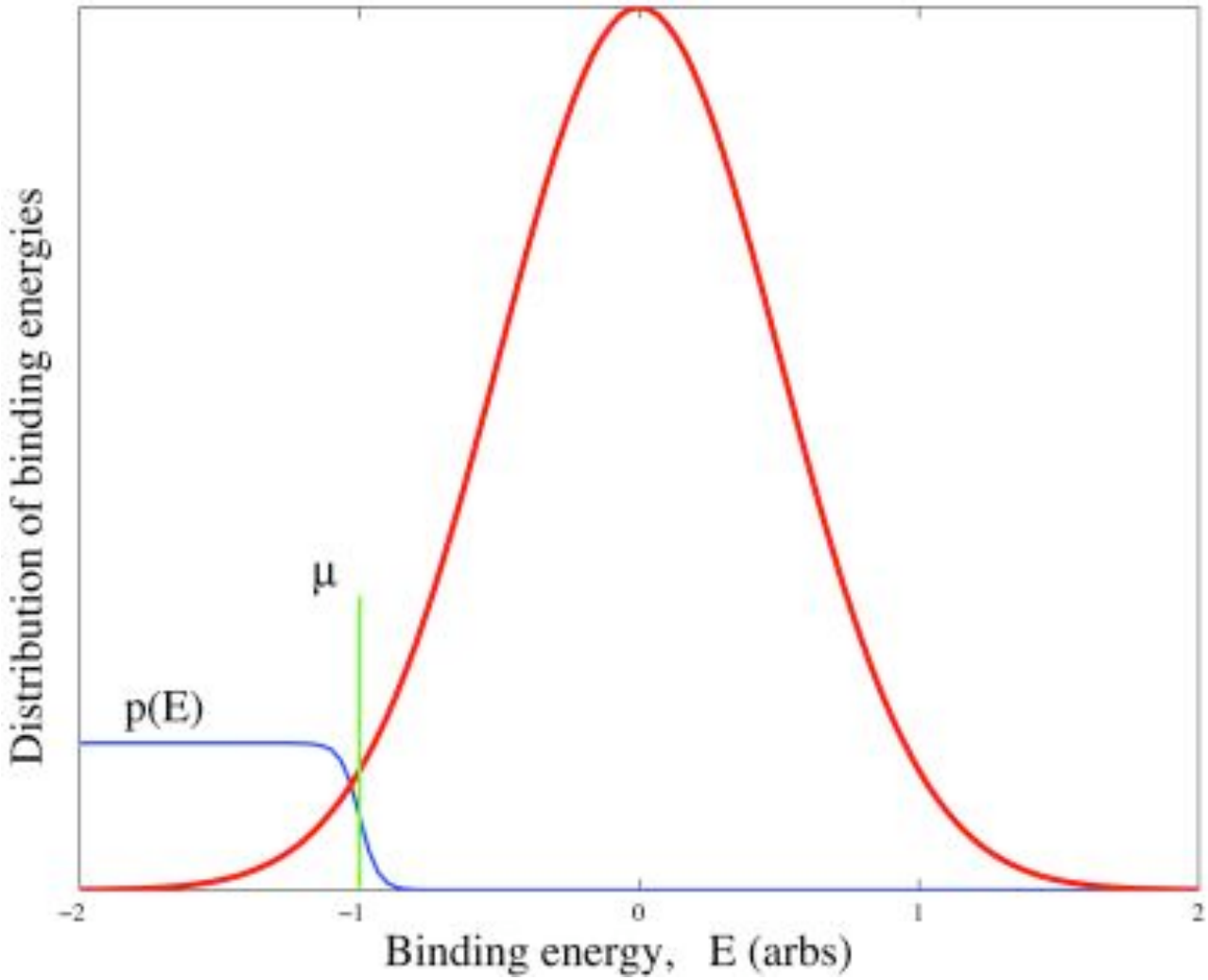
Binding Probability

$$f(E(S))=n/(n+Ke^{\beta E(S)})=1/(e^{\beta(E(S)-\mu)}+1)$$

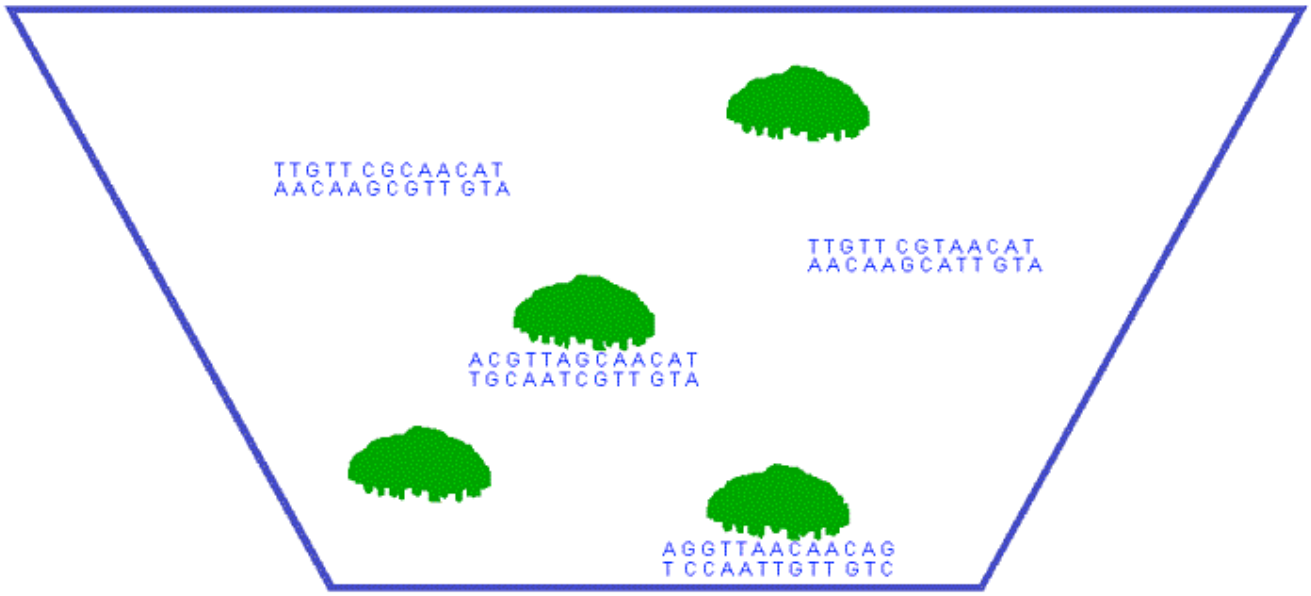
Remember $\mu=k_B T \ln(n/K)$

It is the Fermi (Logistic) function!

Threshold Set by Concentration of Transcription Factor



The Probability Model for Data: Low Stringency SELEX



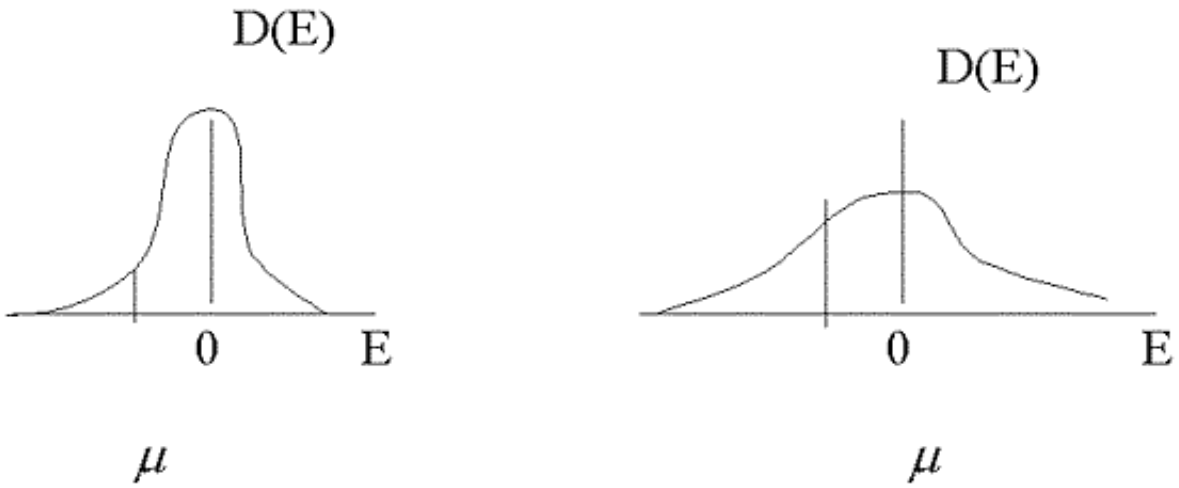
Maximum Likelihood Method for Estimating ε and μ

$$e^{\mathcal{L}(\varepsilon,\mu|O)}=\prod_{S\in O}[\gamma f(E(S))]\prod_{S'\notin O}[1-\gamma f(E(S'))]$$

Non-degenerate Limit
Low μ
$f(E(S))\rightarrow e^{\beta\mu}e^{-\beta E(S)}$
Info. Theory Weight Matrix

Zero Temperature Limit
Low T
$f(E(S))\rightarrow \Theta(\mu-E(S))$
Support Vector Machine (QPMEME)

Increasing Width of D(E) increases number of 'False Positives'/ Random Background



Minimize the Variance!

Quadratic Programming Method for Energy Matrix Estimation

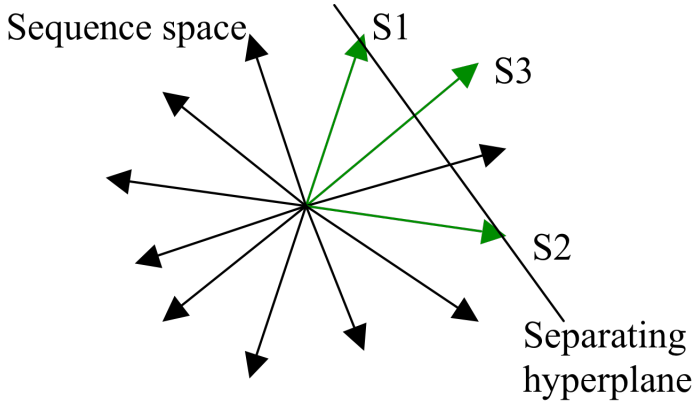
Minimize variance ε^2
Subject to constraints
$E(S_a) = \varepsilon$. $S_a < \mu = -1$
for each example a.

Solvable by Quadratic Programming.

Similar to Support Vector Machine (SVM) pattern finder.

Applied to ~50 *E. coli* TFs in the DPInteract Database

The hyperplane farthest from the origin consistent with data



$$\vec{\varepsilon} = \sum_a \alpha_a \vec{S}_a = \alpha_1 \vec{S}_1 + \alpha_2 \vec{S}_2$$

Sengupta et al., PNAS (2002), Djordjevic et al., Genome Res. (2003)

Parallel Work in Machine Learning Community

Probability models and SVM
Platt (1999)

One class SVM
Schoelkopf et al. (2001)
Manevitz and Yousef (2001)
Tax and Duin (2002)

Results from QPMEME

Djordjevic, Sengupta, Shraiman, Genome Research (2003)

http://biomaps.rutgers.edu/bioinformatics/QPMEME.htm

A biophysical approach to transcription factor binding site discovery

Summary

Identification of transcription binding sites within the regulatory segments of genomic DNA is an important step towards understanding of regulatory circuits that control expression of genes. It is also a task where methods of bio-informatics can be very effective. A powerful general approach to bio-informatic identification of binding sites is based on a "weight matrix" which assigns a position dependent value to each of the possible bases of a sequence segment and combines them into a "score" used for classification. Currently, the widely used method for defining the weight matrix is based on the information theoretic considerations and assigns each sequence an "information score" (for review see Stormo G.D. (2000), "DNA binding sites: representation and discovery", Bioinformatics *16*, 16-23). Here we describe a novel method, which is based on the bio-physical considerations and defines the weight matrix by estimating the sequence dependent (free) energy of binding, which is then used for site classification. The new method also provides for each transcription factor an estimate of the chemical potential which acts as a "binding threshold". Although derived from physical considerations, our method is algorithmically related to the 'support vector machine' approach to pattern recognition (Cristianini, N. and Shawe-Taylor, J., (2001), Intro to support vector machines, Cambridge Univ. Press). The new method for binding site discovery provides a significant improvement over the information score based weight matrix approach, particularly in the ubiquitous case of low specificity factors where it allows to reduce the expected number of false positives without sacrifice in the number of false negatives. The new method is used to identify likely genomic binding sites for the E.coli transcription factors collected in DPInteract database.

Reference: Marko Djordjevic, Anirvan M. Sengupta and Boris I. Shraiman, "A biophysical approach to transcription factor binding site discovery" (Genome Research 2003, submitted)

Summary of binding sites found in *E. coli* genome search:

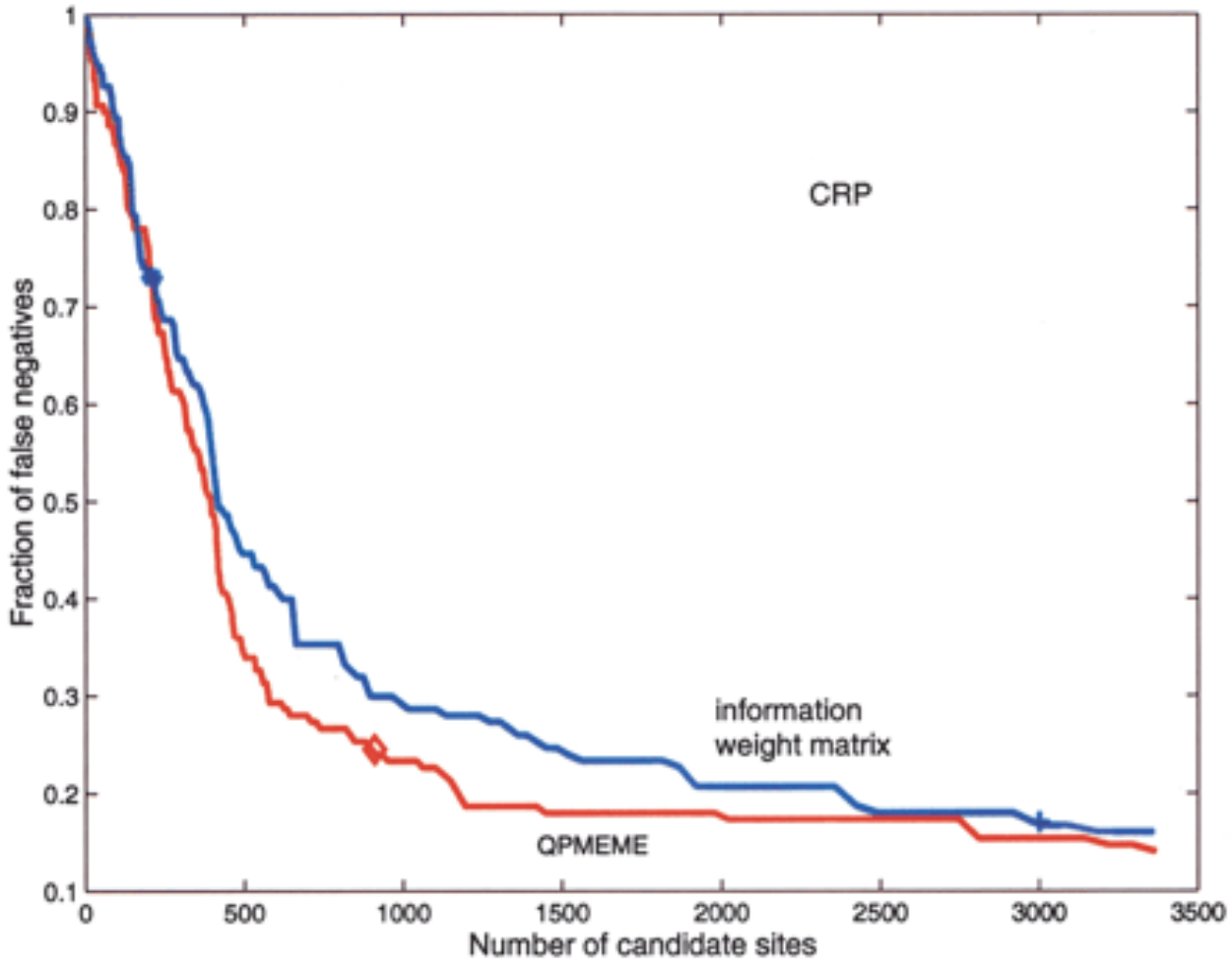
The table below summarizes search results for *E. coli* transcription factors compiled in the DPInteract database, and compares with the information score search results (Robison et al, (1998) J. Mol. Biol. 284, 241-254.) Transcription factor names link to search parameters (energy matrices and binding thresholds) and complete lists of candidate sites.

Explanation of file formats

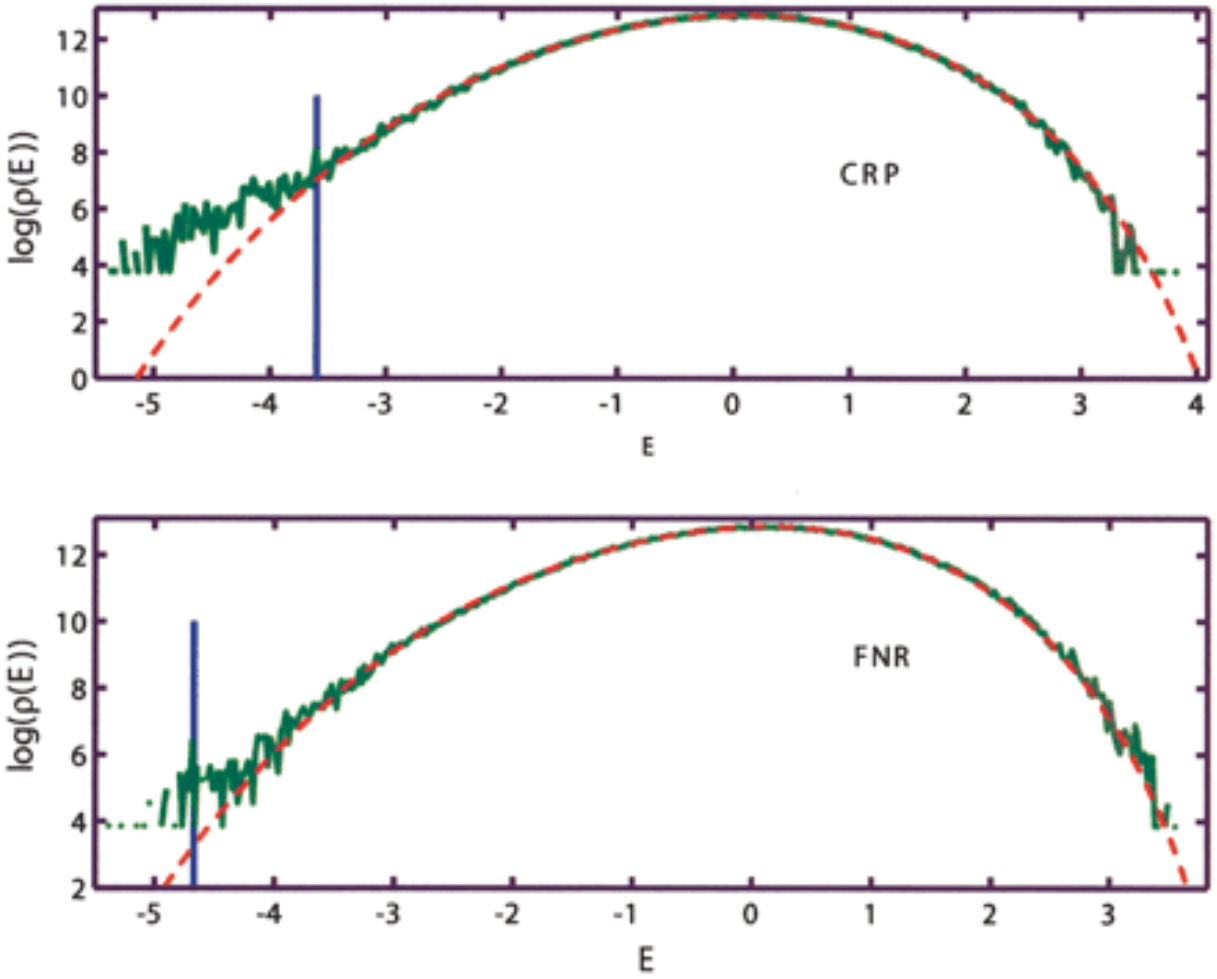
Name	Length	Number of examples	Information score "hits"	QPMEME "hits"	Significance
AraC	48	6	6	6	7*10^5
ArcA	15	14	391	52	6.4
ArgR	18	17	320	79	8.9
CarP	25	2	2	2	1*10^5
Crp	22	49	3093	796	27.2
CspA	20	4	15	4	2*10^3
CynR	21	2	2	2	3*10^4

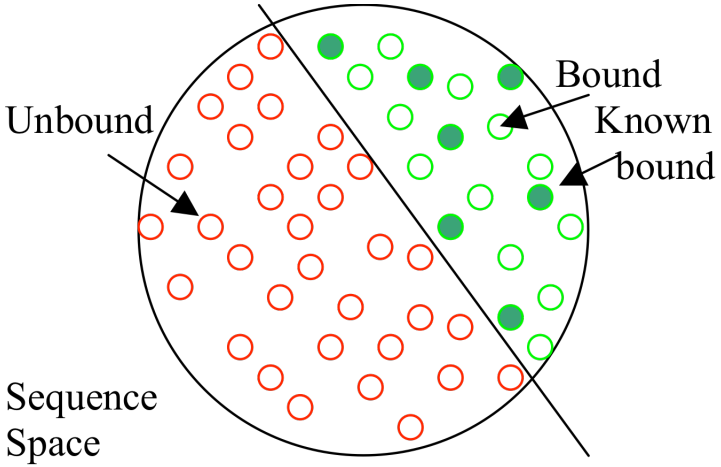
- Solves the problem of threshold selection
- Better sensitivity/specificity tradeoff than conventional methods
- False negative rate 25% (for CAP)
- Positive predictive value 60-70% (for CAP)

Comparison with Conventional Weight Matrix Results

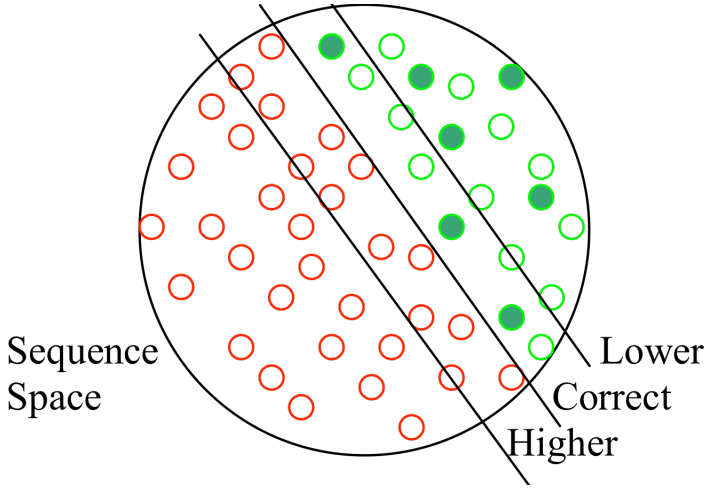


Significant over-abundance of *E. coli* sites under the threshold



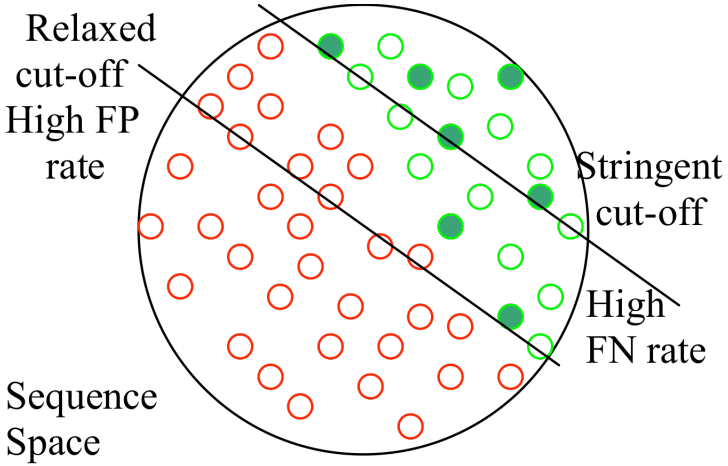


Binding at physiological concentration:
Separation of bound sequences from the rest

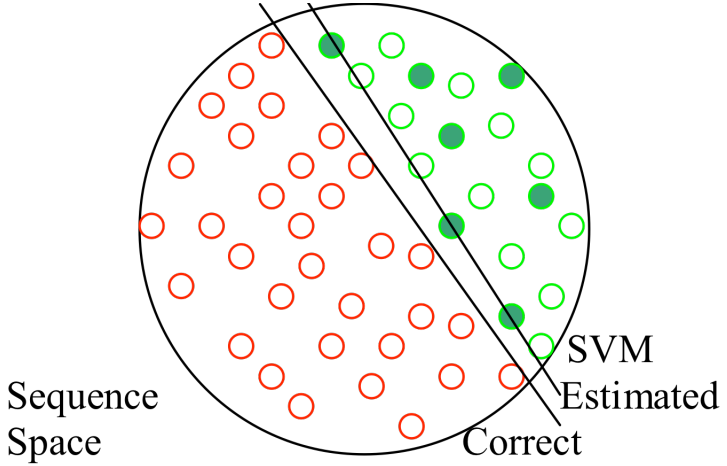


Effect of TF concentration:
Lower->Stringent, Higher->Relaxed

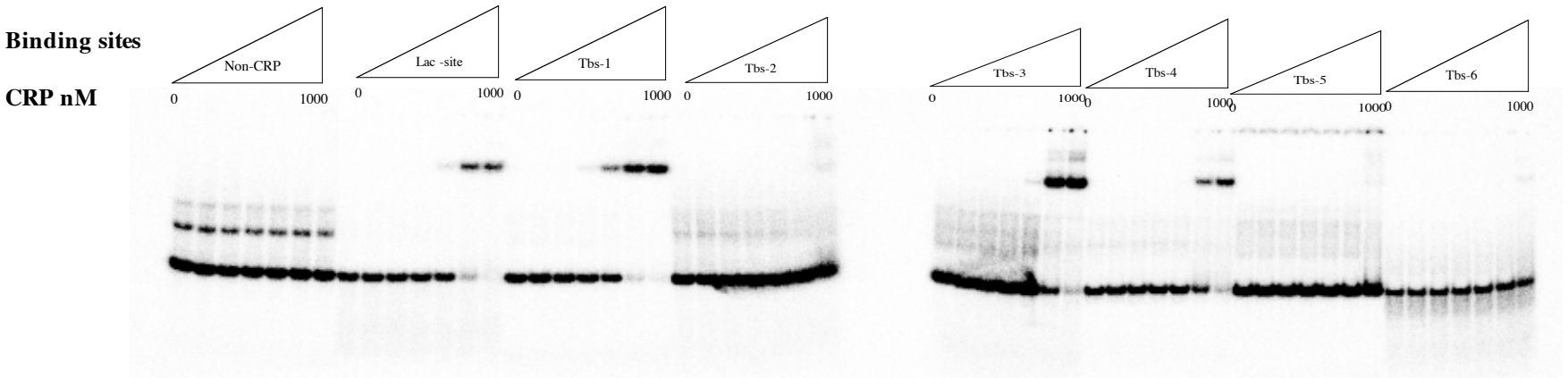
Result of weight matrix misestimating
the orientation of the separating plane



QPMEME estimates the orientation
and the location from marginal examples



EMSA for Predicted Sites



Positive predictive value=TP/(TP+FP) ~60-70%

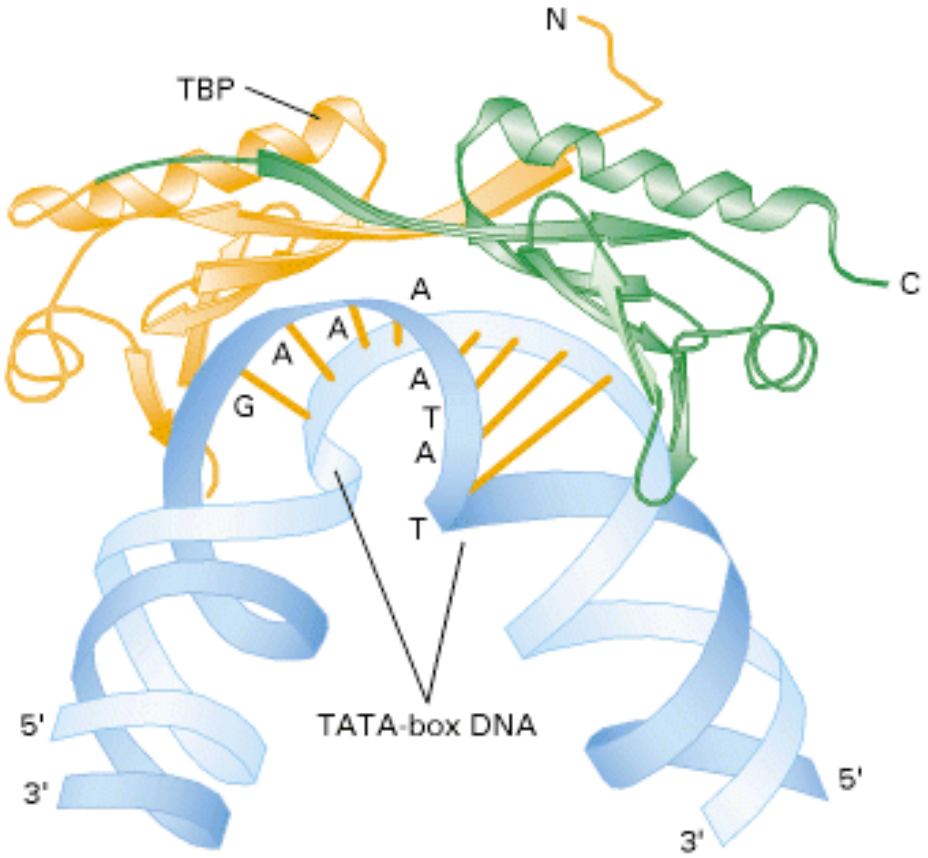
Improvements?

- Corrections to independent base model:
  adding nearest neighbor terms
  (with O'Flanagan, Paillard, Lavery)
- "Unbiased" datasets: high-throughput SELEX
  experiments on CAP
  (with Nagaraj, O'Flanagan, Shraiman)
- Incorporation other type of information
  (gene expression, inter-species comparison,..)

DNA Deformation

Protein-DNA complex free energy
= Direct Protein DNA terms
+ DNA deformation terms
$=\Sigma_{ib}\varepsilon_{ib}S_{ib} + \Sigma_{ib}J_{ii+1;ab}S_{ia}\,S_{i+1\,b}$
+……

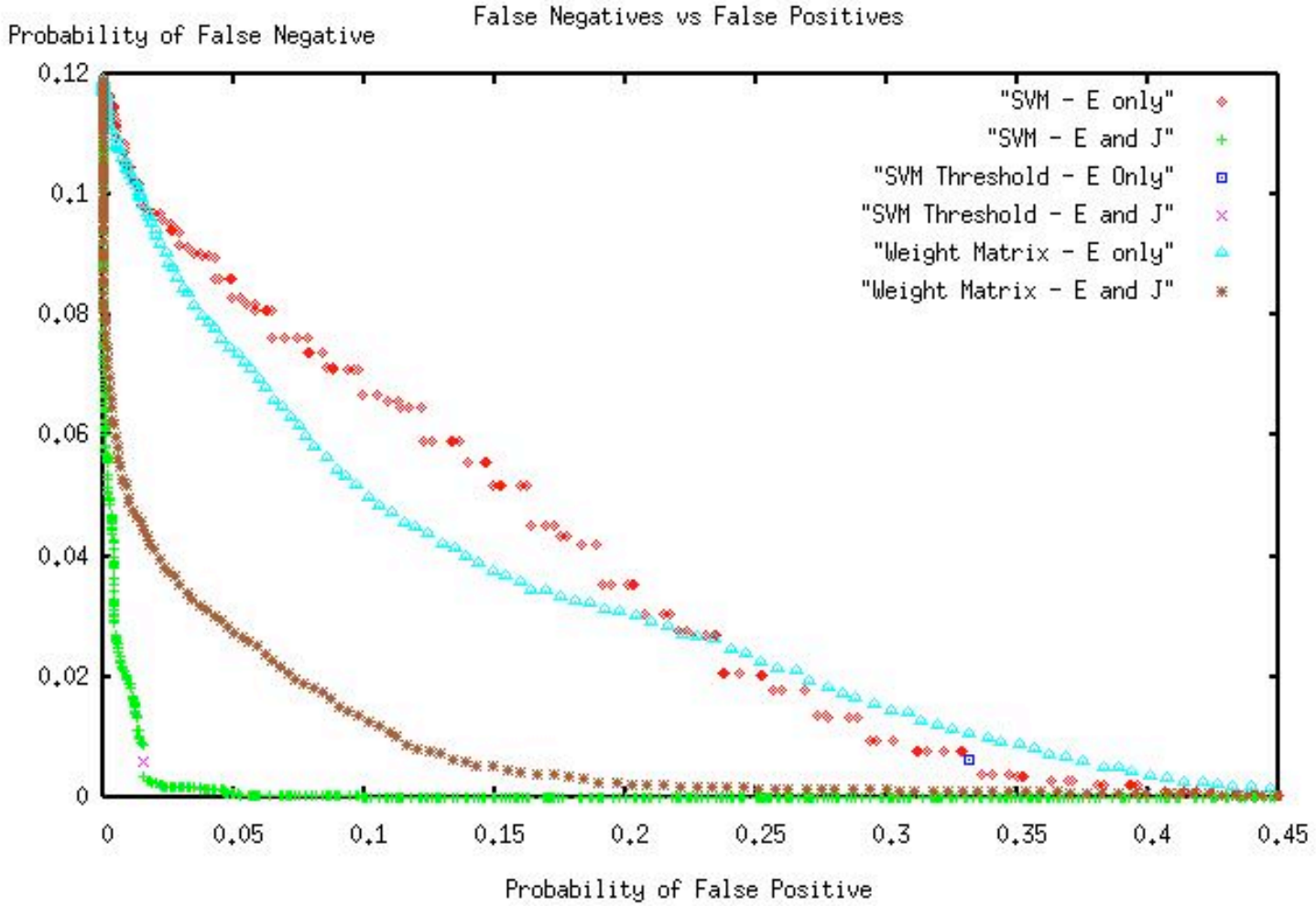
TBP Binding TATA Box



Problems of Generalizing QPMEME to Include Deformation

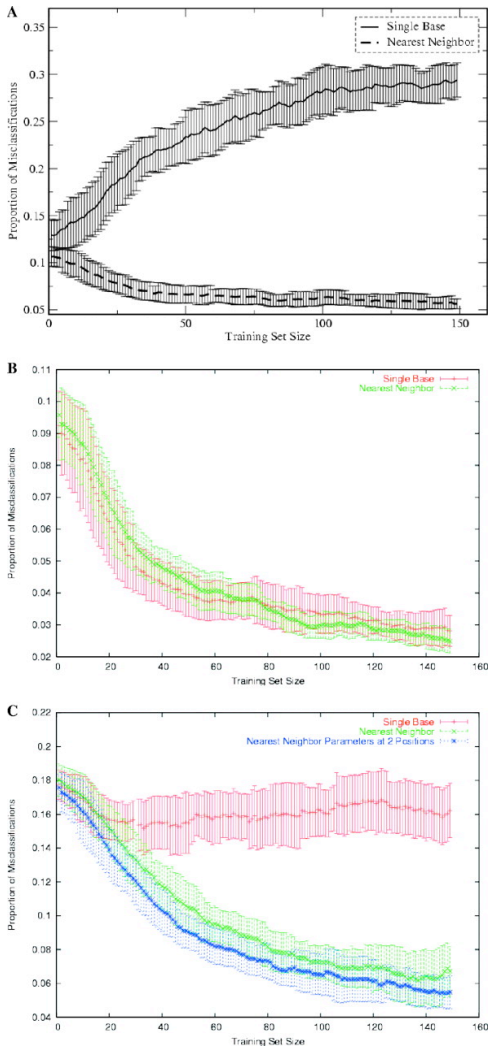
- Four times more parameters.
- Number of sequences needed to train in hundreds
- Can possibly be done with SELEX SAGE
- However, for the time being, why not use atomistic calculation to get the best binders use that to test ideas? (O'Flanagan, Paillard, Lavery, Sengupta, Bioinformatics, to appear)

Performance of Algorithms on Computationally Generated Data

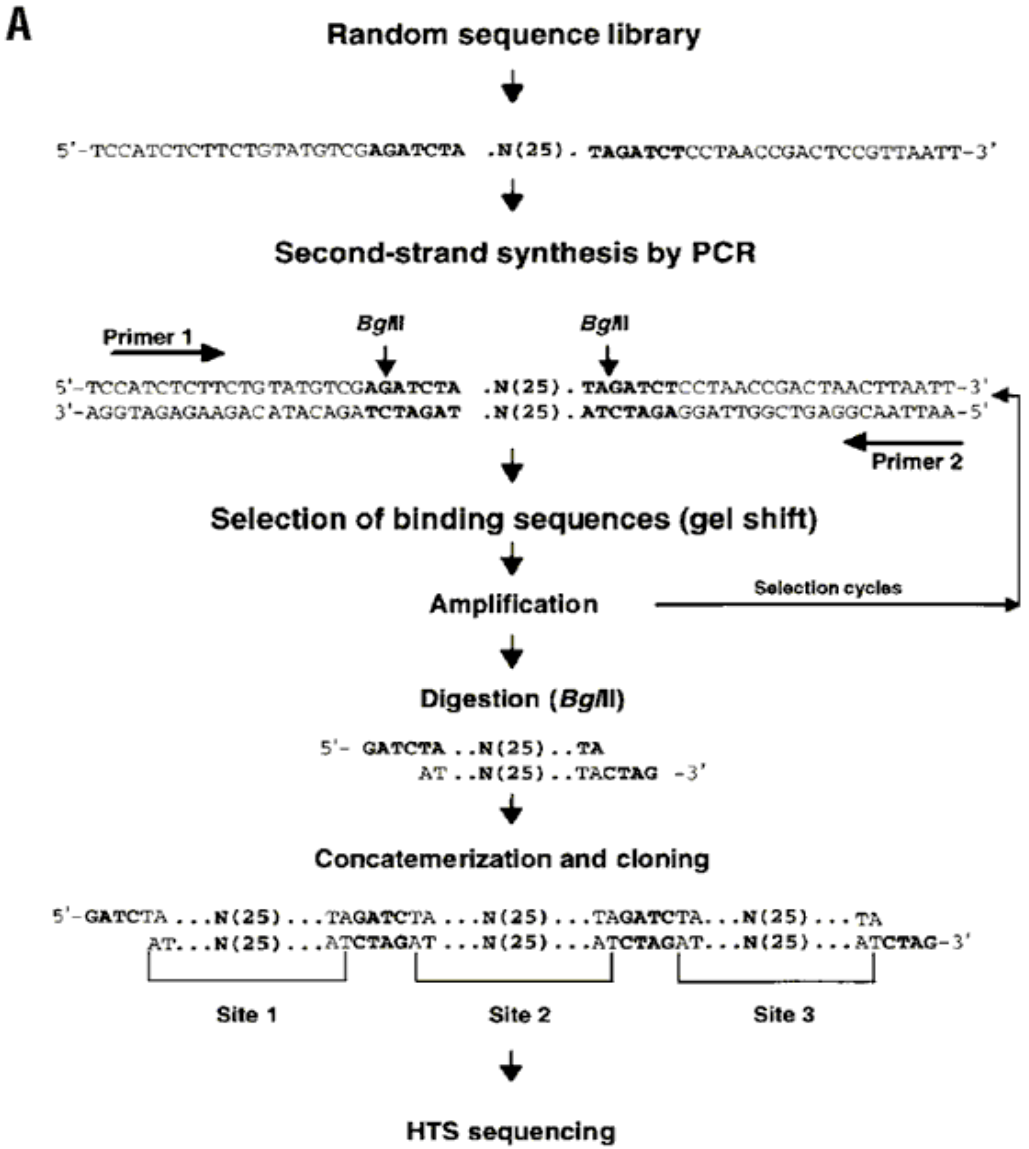


Performance vs # of Examples

- One could estimate optimal number of sites necessary (e. g. 60-70 for TBP)
- Could use structural insights to make it more sparse (informative priors).

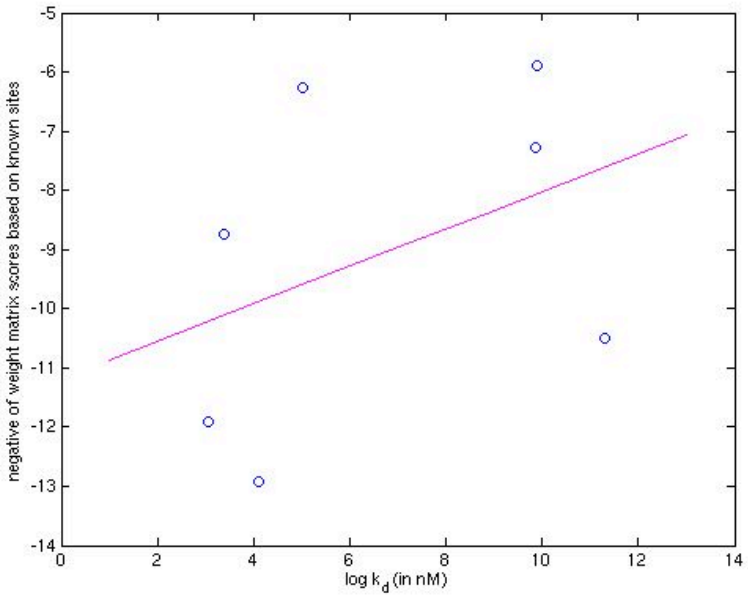


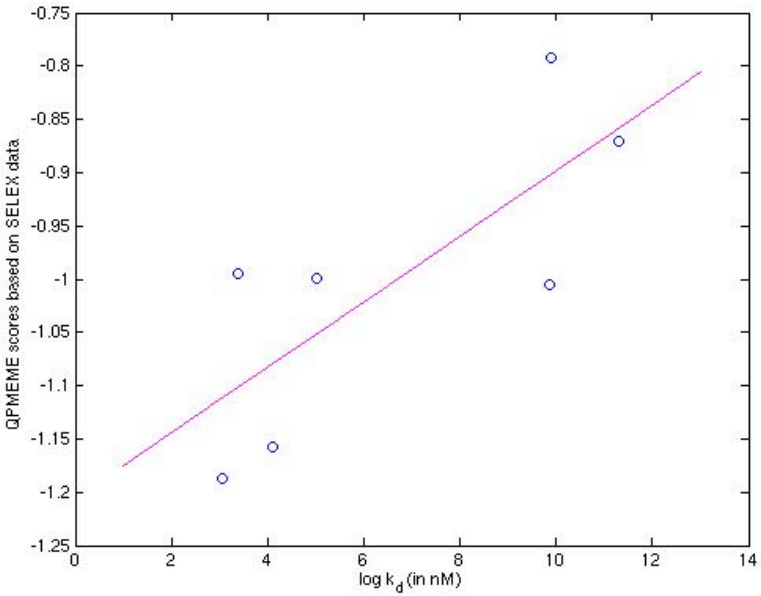
Many Sequences: SELEX SAGE



Roulet et al. Nat. Biotech. 2002

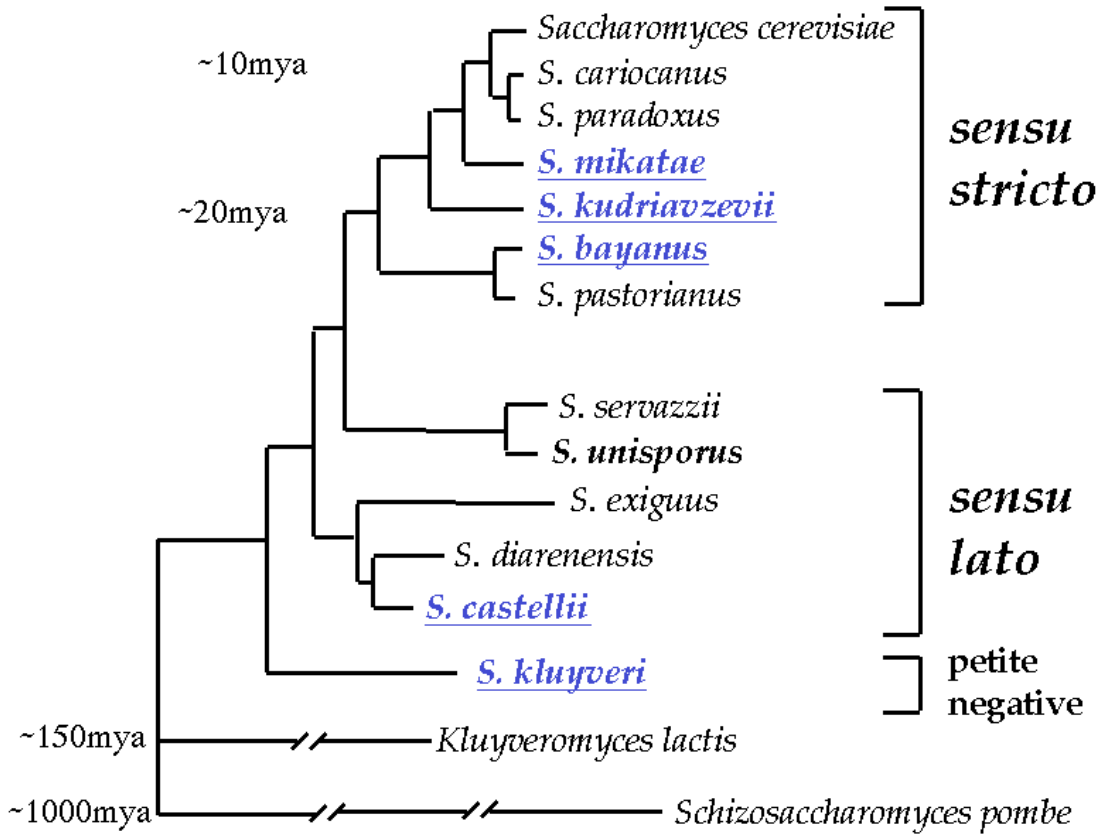
Improvement of Correlation with Affinity





Comparison between E. Coli and Salmonela:
Fraction of conserved predicted sites improves: 55-60%-->75-80%
(Nagaraj, O'Flanagan, Shraiman, Sengupta, ms. in preparation)

Phylogenetic Footprinting



From http://www.genetics.wustl.edu/saccharomycesgenomes/yeast_phylogeny.html

Functional weak sites

HO(10) Strong site, highly conserved

```
Scer --------GTTTTGCCGCGTTAAAACCTACATC-AAAAAAGG-CGGATCA
Spar gtcaaTACGTTTTGCCGCGTTAAAACCTACATC-AAAAAAAGGCGGATCA
Smik CAAt-----TTTTACCGCGTTAAAACATACATCgAAAAAAGGGCGGATCA
Skud -----TACGTTTTACCGCGTTAAAACTTACATC-AAAAAAGGGCGGATCA
Sbay AAAgtTACATTTTACCGCGTTAAAACCTACATC-AAAAAAGGGCGGATCA
     00000111266665666666777777677777707777765545555555
```

HO(7) Medium strength site, highly conserved

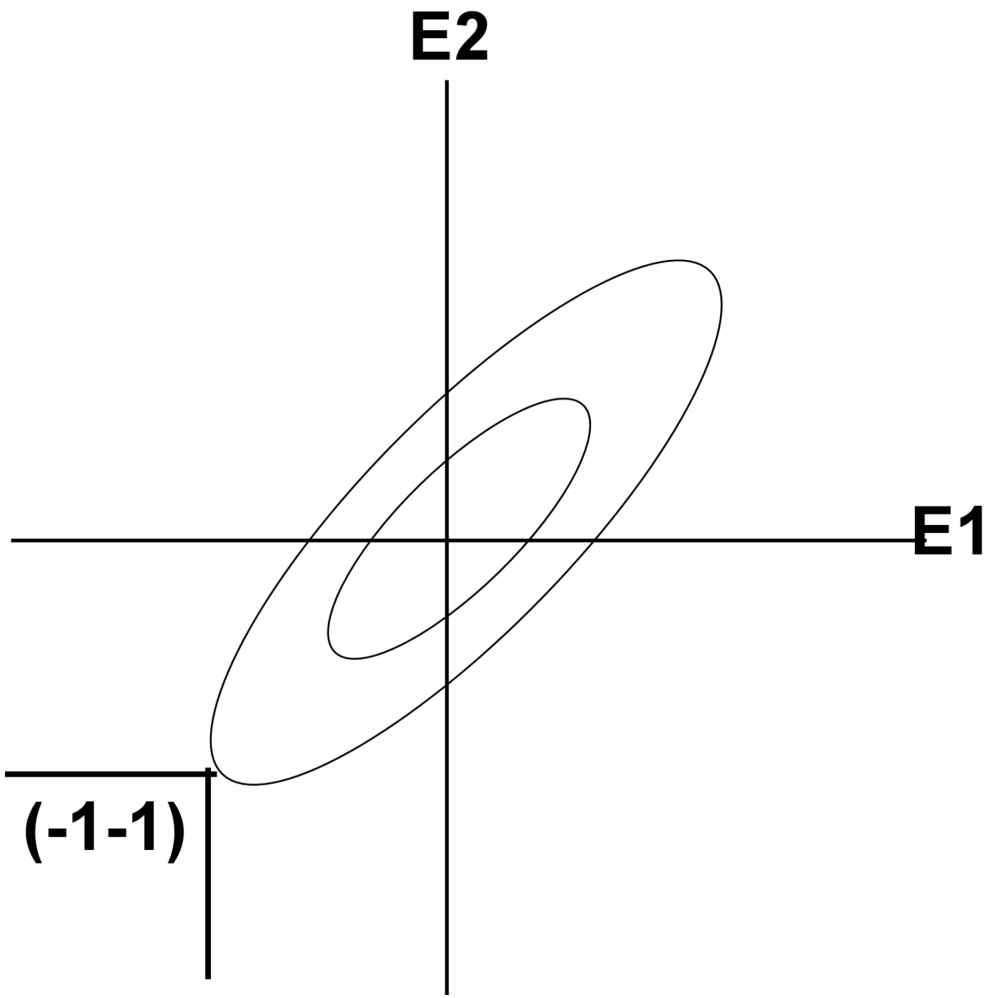
```
Scer CCCAAAGGGGTATCAAATAATCGATGTGCTTTTTCACTCTACGAATGATC
Spar CCCAAAGGGGTATGAAATAATCGATGTGCTTTTTCACTCTACGAATGATC
Smik CCCAAAAGGGTATGAAATAATCGATGTGCTTTTTCACTCTACGAATGATC
Skud CTGAAAGGGGTATCGAATAATCGATGTGCTTTTTCACTCTACGAATGATC
Sbay CTGAAAGGG--ATGAAGTAACTGATTTGTCTTTTCTCTCTGCGGATGATC
     444444444445544666666666677777777777777888988888
```

HO(2) Medium strength site, not well conserved

```
Scer AATTCA-TGTCAT-GTCCACATTAACATCATTG-CAGAGCAACAATTCAT
Spar AATTCA-TGTAAATGTTTACATTAACATCACTTGCAGGAGAACGGCTCGT
Smik AACcttaTGCGAAcGTTTACATTACTATCACTCACAGGAAAATAATAAAT
Skud AAagaA-TTTATTTGTTTACATCAACATCTCTTGTAGAGGAACAATGCAT
Sbay AACTGA-TGTAATTGTTTACATCAATATCTTCG-CAGAAGAGCAATCCAT
     221001022211112222222222222332111112222222211112 2
```

Output of Dialign, Morgenstern et al.

Scores of Evolutionarily Conserved Sites



Constrained Optimization

$$E_a^{(A)} = \varepsilon \bullet S_a^{(A)} \leq -1, \forall a, A$$

$$c_{AB}(\varepsilon) = \mathrm{cov}(E^{(A)}, E^{(B)})$$

$$\max \sum_{AB} c(\varepsilon)^{-1}{}_{AB}$$

Ugly!

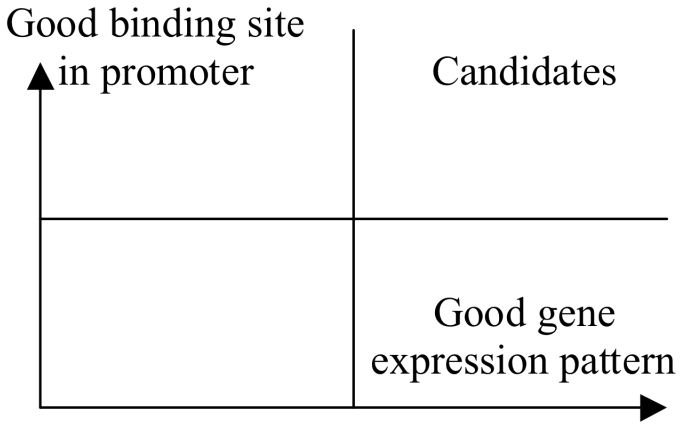
Kinder approach:
Optimization of a quartic function.
Soluble, by iterated QP.

$$E_a^{(A)} = \varepsilon \bullet S_a^{(A)} \leq -1, \forall a, A$$

$$c_{AB}(\varepsilon) = \mathrm{cov}(E^{(A)}, E^{(B)})$$

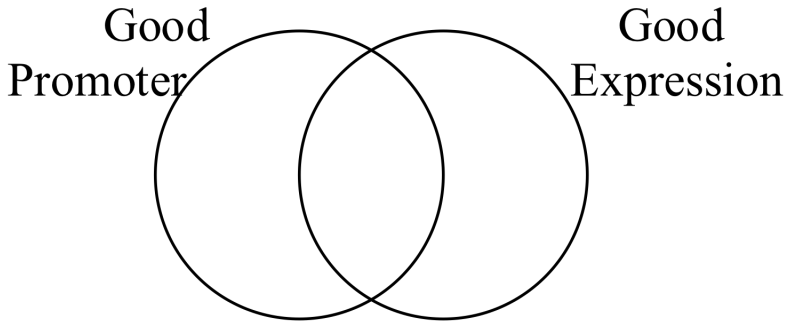
$$\min \sum_{AB} \gamma_A c_{AB}(\varepsilon) \gamma_A + \sum_A \gamma_A$$

Integration of Expression Data and Sequence Analysis: Too many Thresholds to choose?



=>

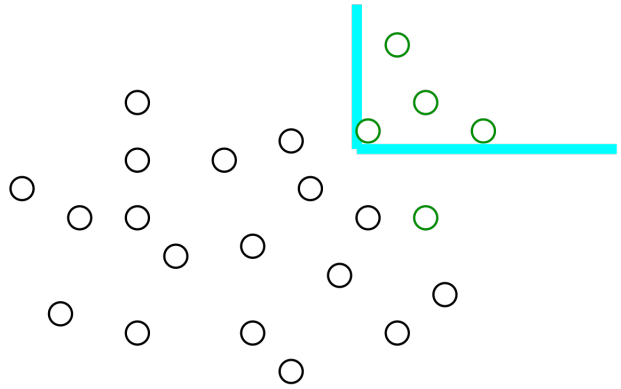
Venn Diagrams



Combining Multiple Scores

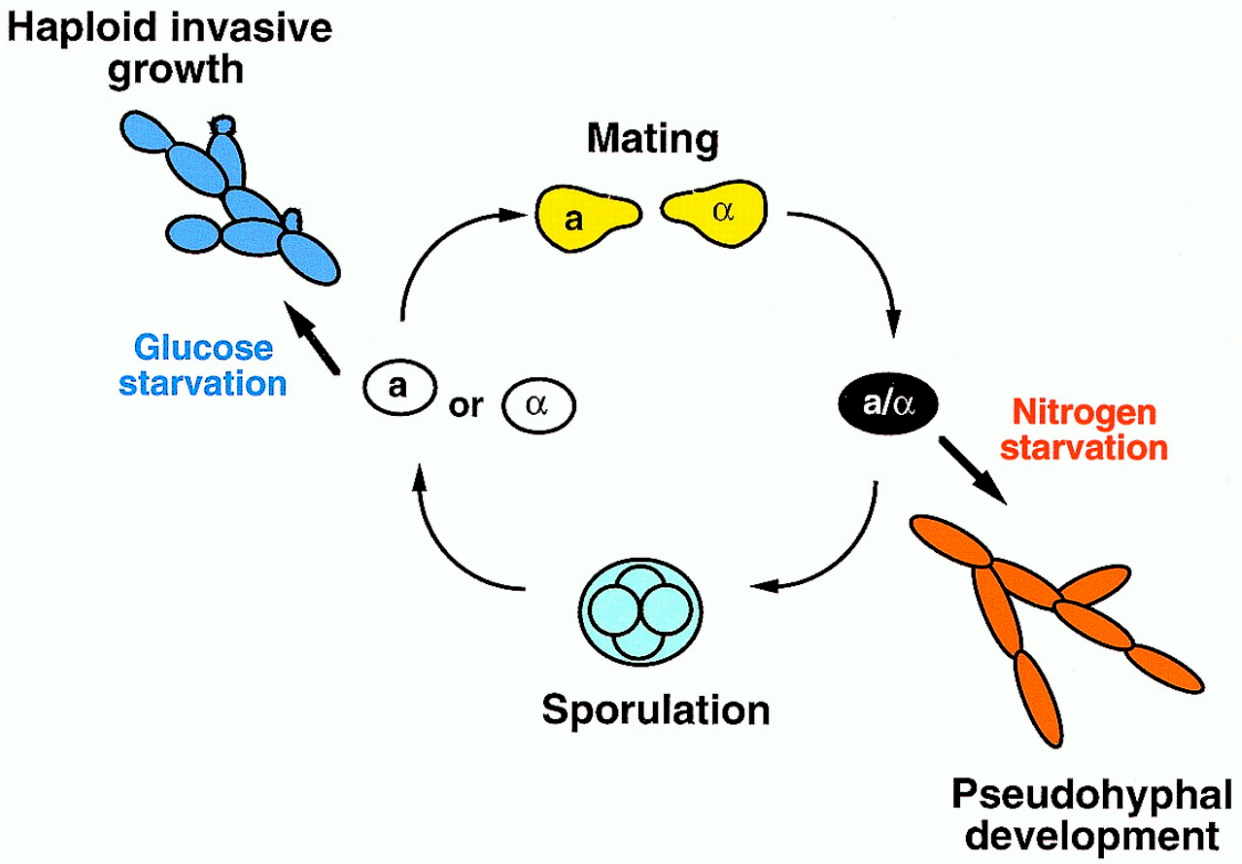
Suppose we have multiple scores (x_{1g}, …, x_{ng}) for a gene g that are uncorrelated for the "generic" gene, but correlated for the regulated ones.

$$p_g = \prod_i \Pr_i(x_i > x_{ig})$$

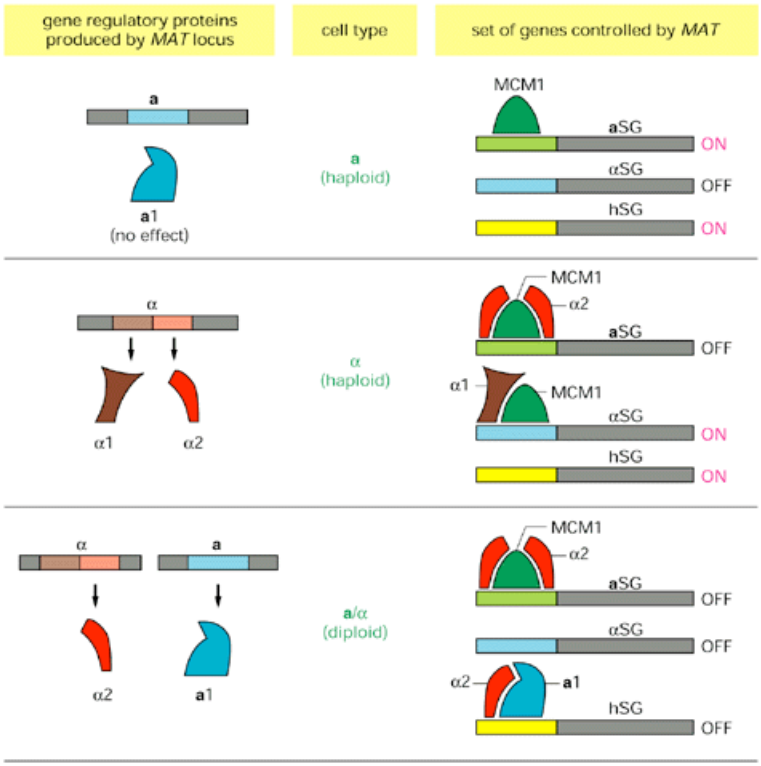


Caveat: Need some feature selection method not to throw in irrelevant (or very weakly predictive) scores.

Yeast life cycle



Combinatorial Control of Yeast Mating Type Identity



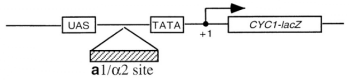
- Combinatorial control by three regulated factors (and a constitutive one) regulate cell type identity.
- Detection of direct targets by combining sequence analysis and microarray data (Nagaraj, O'Flanagan, Bruning, Mathias, Vershon, Sengupta, BMC Genomics 2004)

Mutational data

A

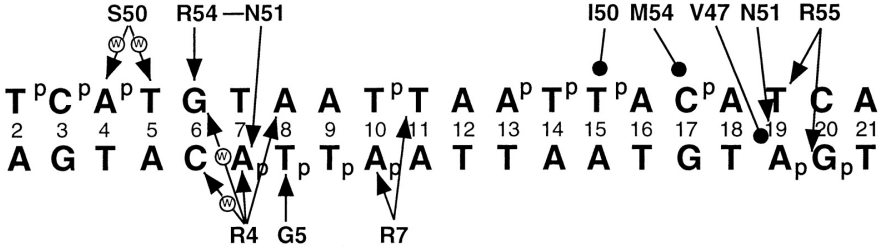
	α2 site															a1 site						
MATα1	C	A	A	T	G	T	A	G	A	A	A	A	G	T	A	C	A	T	C	A
RME1	A	G	A	T	G	T	C	A	C	A	G	A	T	T	A	C	A	T	C	A
STE5(1)	G	C	T	T	G	T	T	A	A	T	T	T	A	C	A	C	A	T	C	A
STE5(2)	T	C	A	T	G	T	A	C	T	T	T	T	C	T	G	C	A	T	C	A
AXL1	G	C	A	T	G	T	A	A	A	A	T	A	C	C	G	C	A	T	C	A
FUS3	C	C	G	T	G	T	A	A	A	A	A	A	C	C	G	C	A	T	C	A
GPA1	G	C	A	T	G	T	T	A	A	A	A	A	G	C	A	C	A	T	C	A
HO(1)	G	C	G	T	T	T	A	G	A	A	C	G	C	T	T	C	A	T	C	A
HO(2)	T	C	A	T	G	T	C	C	A	C	A	T	T	A	A	C	A	T	C	A
HO(3)	T	C	A	T	G	T	T	A	T	T	A	T	T	T	A	C	A	T	C	A
HO(4)	T	C	G	T	G	T	A	T	T	T	A	G	T	T	A	C	A	T	C	A
HO(5)	A	C	A	T	G	T	C	T	T	C	A	A	C	T	G	C	A	T	C	A
HO(6)	T	C	A	T	G	T	A	T	T	C	A	T	T	C	A	C	A	T	C	A
HO(7)	T	A	G	A	G	T	G	A	A	A	A	A	G	C	A	C	A	T	C	G
HO(8)	G	C	C	T	G	C	G	A	T	G	A	G	A	T	A	C	A	T	C	A
HO(9)	T	T	A	T	G	T	T	A	A	A	A	G	T	T	A	C	A	T	C	C
HO(10)	C	C	G	C	G	T	T	A	A	A	A	C	C	T	A	C	A	T	C	A
consensus	T	C	A	T	G	T	A	A	T	T	A	A	T	T	A	C	A	T	C	A
crystal	A	C	A	T	G	T	A	A	T	T	T	A	T	T	A	C	A	T	C	A

B



Site	β-Gal. Units	Repression Ratio
None	422 ± 20	
MATα1	8.6 ± 0.3	50 X
Crystal	6.2 ± 0.2	70 X
Consensus	5.3 ± 0.2	80 X

A



B

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	62	35	**80**	35	4	3	**80**	**80**	90	44	**80**	**80**	123	70	**80**	70	**80**	7	30	**80**
G	52	10	78	10	**80**	2	11	34	19	35	40	55	160	8	70	11	5	6	28	nd
C	nd	**80**	40	23	3	5	17	10	80	75	25	50	nd	93	75	**80**	70	15	**80**	130
T	**80**	20	66	**80**	5	**80**	60	70	**80**	**80**	70	nd	**80**	**80**	8	40	45	**80**	70	150

Measurement of fold repression caused by single base mutations of the “consensus” sequence in a heterologous promoter (Jin, Zhong and Vershon, MCB,1999) allows us to score other sequences.

Microarray data for polyploids

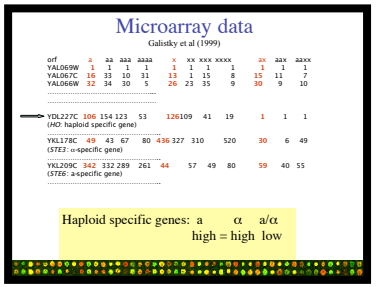
orf	a	aa	aaa	aaaa	x	xx	xxx	xxxx	ax	aax	aaxx
YAL069W	**1**	1	1	1	**1**	1	1	1	**1**	1	1
YAL067C	**16**	33	10	31	**13**	1	15	8	**15**	11	7
YAL066W	**32**	34	30	5	**26**	23	35	9	**30**	9	10
.....											
.....											
YDL227C	**106**	154	123	53	**126**	109	41	19	**1**	1	1
(*HO*: haploid specific gene)											
.....											
YKL178C	**49**	43	67	80	**436**	327	310	520	**30**	6	49
(*STE3*: α-specific gene)											
.....											
YKL209C	**342**	332	289	261	**44**	57	49	80	**59**	40	55
(*STE6*: a-specific gene)											
.....											
.....											

(Galitski, Saldhana, Styles, Lander and Fink, Science, 1999)

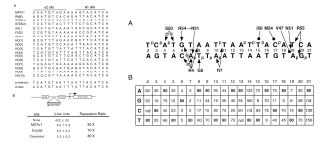
Computational Approach

Microarray data

Galistky et al (1999)



Mutational data



Measurement of fold repression caused by single base mutations of the "consensus" sequence in a heterologous promoter (Jin, Zhong and Vershon, MCB,1999) allows us to score other sequences.

Assigning Scores for Expression

Assign high scores to genes which are repressed in the presence of a1/α2, but genes should have similar expression in a- and α-type cells.

$$\text{Score} = \text{Sign}(\overline{a,\alpha} - \overline{a/\alpha})(\overline{a,\alpha} - \overline{a/\alpha})^2 - A(\overline{a} - \overline{\alpha})^2$$

Where:

$\overline{a,\alpha}$ = Average Log expression in absence of a1/α2

$\overline{a/\alpha}$ = Average Log expression in presence of a1/α2

$\overline{a}$ = Average Log expression in a-type cells

$\overline{\alpha}$ = Average Log expression in α-type cells

A = Weighting factor

Independent base model

Each nucleotide gives a single contribution to the score:

$$E(TTAGCAA) = \varepsilon_{1T} + \varepsilon_{2T} + \varepsilon_{3A} + \varepsilon_{4G} + \varepsilon_{5C} + \varepsilon_{6A} + \varepsilon_{7A}$$

Combining Information

Expression p-value = The fraction of genes with a better score

Binding p-value = Probability of finding a better binding site in a "random" promoter

Combined p-value = p-value for high expression score × p-value for good binding site in the promoter

Significant expression p-value and significant binding p-value

Significant expression p-value without significant binding p-value

Table 1. List of Potential a1-α2 Binding Sites in Haploid-specific Genes

ORF	Gene	Expression P-val	Binding P-val	Combined P-val	a1-α2 ChIP
YDL227C	***HO***	0.0006	0.0017	1.1e-6	+
YLR265C	***NEJ1***	0.0003	0.0053	1.7e-6	+
YBL016W	***FUS3***	0.0001	0.0991	1.6e-5	+
YOR212W	***STE4***	0.0020	0.0082	1.7e-5	+
YJR086W	***STE18***	0.0008	0.0218	1.7e-5	+
YHR005C	***GPA1***	0.0005	0.0437	2.1e-5	+
YDR103W	***STE5***	0.0017	0.0298	5.2e-5	+
YBR073W	***RDH54***	0.0053	0.0116	6.2e-5	+
YGR044C	***RME1***	0.0009	0.0720	6.8e-5	+
YGL248W	***PDE1***	0.0182	0.0040	7.3e-5	+
YPL038W	***MET31***	0.0292	0.0027	8.0e-5	+
YDR088C	***SLU7***	0.0303	0.0038	1.2e-4	-
YGL052W		0.0117	0.0109	1.3e-4	-
YJL157C	***FAR1***	0.0013	0.1141	1.4e-4	+
YPR122W	***AXL1***	0.0091	0.0163	1.5e-4	+
YIL099W	***SGA1***	0.0063	0.0267	1.7e-4	-
YLR233C	***EST1***	0.0226	0.0090	2.0e-4	-
YKL182W	***FAS1***	0.0578	0.0035	2.1e-4	-
YMR053C	***STB23***	0.0028	0.0884	2.5e-4	-
YNL319W		0.0123	0.0222	2.7e-4	-
YFR012W		0.0088	0.0125	2.8e-4	-
YNL188W	***KAR1***	0.0019	0.1890	3.6e-4	-
YGL193C		0.0014	0.2654	3.8e-4	-
YMR157C		0.1557	0.0026	4.0e-4	-
YIL117C	***PRM5***	0.0011	0.3689	4.1e-4	-

Table 2. Haploid-specific Genes that do not contain a1-α2 target sites

ORF	Gene	Expression p-val	a1-α2 ChIP
YIL117C	***PRM5***	0.0011	-
YLR159W		0.0015	-
YBR051W		0.0022	-
YLR080W		0.0024	-
YIR039C	***YPS6***	0.0025	-
YCL014W	***BUD3***	0.0027	-
YPL189W	***GUP2***	0.0030	-
YJL077C	***ICS3***	0.0032	-
YML042W	***CAT2***	0.0033	-
YHR004C	***NEM1***	0.0035	+
YFR046C		0.0036	-
YDR220C		0.0038	-
YPL025C		0.0041	-
YBR006W	***UGA2***	0.0043	-
YBR108W		0.0044	-
YFL034W		0.0046	-
YCL027W	***FUS1***	0.0047	-
YLR308W	***CD42***	0.0049	
YGR014W	***MSB2***	0.0050	
		0.0052	

Candidate direct targets

Candidate indirect targets

Significant Combinations

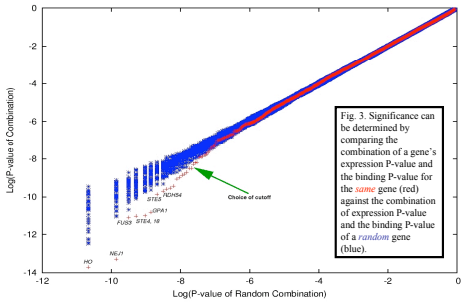


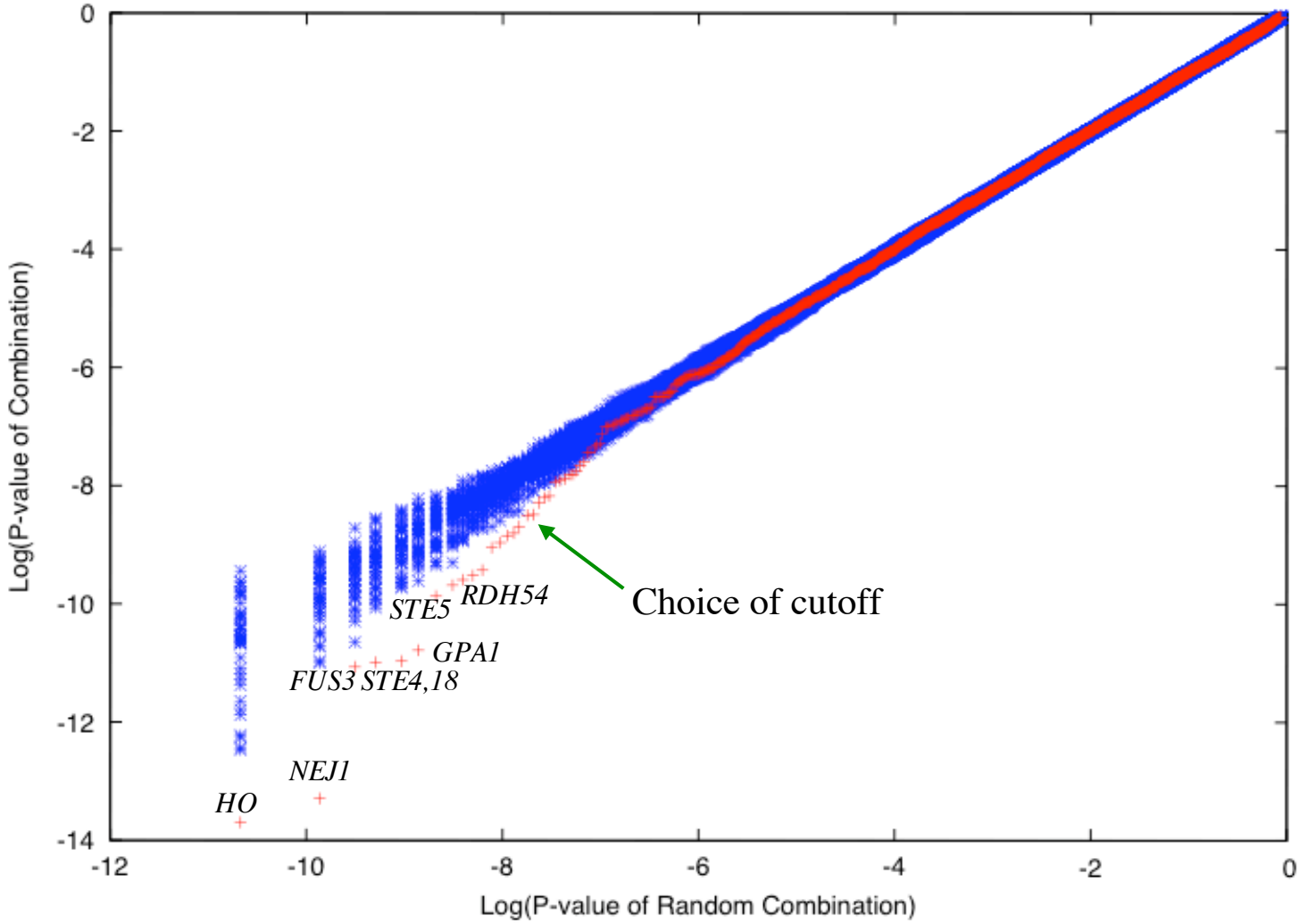
Fig. 3. Significance can be determined by comparing the combination of a gene's expression P-value and the binding P-value for the *same* gene (red) against the combination of expression P-value and the binding P-value of a *random* gene (blue).

Ordering of Candidate Targets

Table 1. List Potential a1-a2 Binding Sites in Haploid-specific Genes

ORF	Gene	Sub-class[a]	Expression P-val	Binding P-val	Combined P-val	a1-α2 ChIP
YDL227C	HO	1	0.0006	0.0017	1.1e-6	+
YLR265C	NEJ1	1	0.0003	0.0053	1.7e-6	+
YBL016W	FUS3	2	0.0001	0.0991	1.6e-5	+
YOR212W	STE4	2	0.0020	0.0082	1.7e-5	+
YJR086W	STE18		0.0008	0.0218	1.7e-5	+
YHR005C	GPA1	2	0.0005	0.0437	2.1e-5	+
YDR103W	STE5	2	0.0017	0.0298	5.2e-5	+
YBR073W	RDH54	1	0.0053	0.0116	6.2e-5	+
YGR044C	RME1	3	0.0009	0.0720	6.8e-5	+
YGL248W	PDE1	4	0.0182	0.0040	7.3e-5	+
YPL038W	MET31	4	0.0292	0.0027	8.0e-5	+
YDR088C	SLU7		0.0303	0.0038	1.2e-4	-
YGL052W			0.0117	0.0109	1.3e-4	-
YJL157C	FAR1	2	0.0013	0.1141	1.4e-4	+
YPR122W	AXL1	2	0.0091	0.0163	1.5e-4	+
YIL099W	SGA1	3	0.0063	0.0267	1.7e-4	-
YLR233C	EST1	1	0.0226	0.0090	2.0e-4	-
YKL182W	FAS1		0.0578	0.0035	2.1e-4	-
YMR053C	STB2	3	0.0028	0.0884	2.5e-4	-
YNL319W			0.0123	0.0222	2.7e-4	-
YFR012W			0.0088	0.0125	2.8e-4	-
YNL188W	KAR1	2	0.0019	0.1890	3.6e-4	-
YGL193C			0.0014	0.2654	3.8e-4	-
YMR157C			0.1557	0.0026	4.0e-4	-
YIL117C	PRM5		0.0011	0.3689	4.1e-4	-

Significant combinations

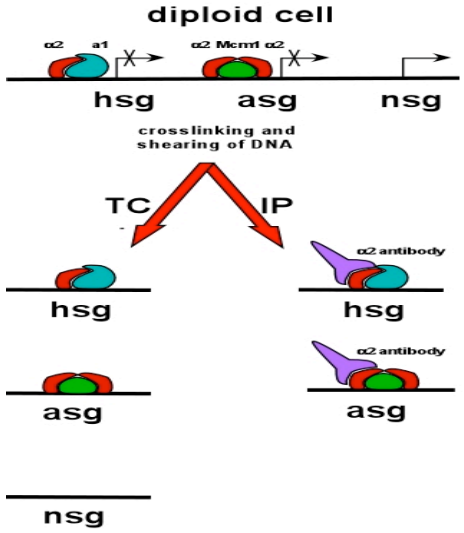


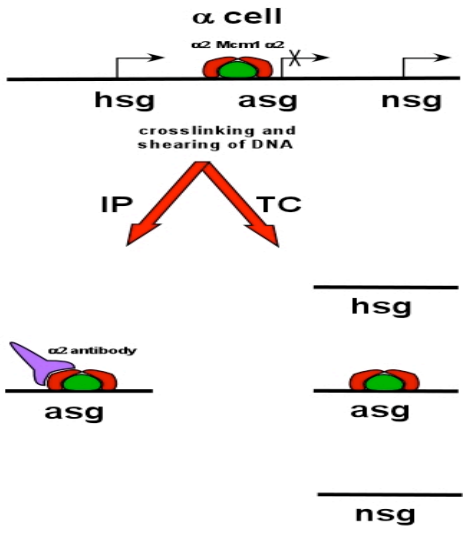
Ordering of Candidate Targets

Table 1. List Potential a1-a2 Binding Sites in Haploid-specific Genes

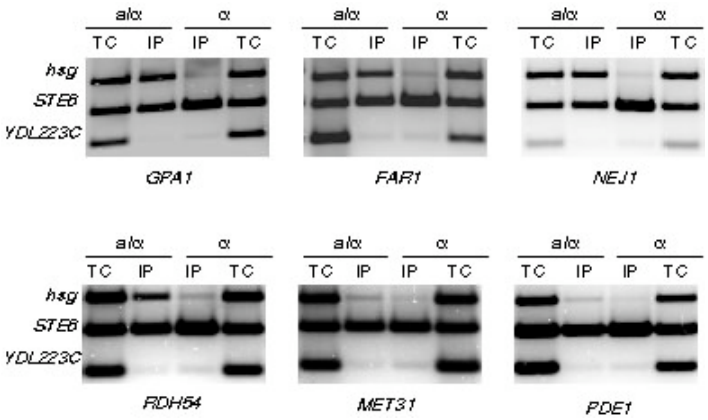
ORF	Gene	Sub-class[a]	Expression P-val	Binding P-val	Combined P-val	a1-α2 ChIP
YDL227C	HO	1	0.0006	0.0017	1.1e-6	+
YLR265C	NEJ1	1	0.0003	0.0053	1.7e-6	+
YBL016W	FUS3	2	0.0001	0.0991	1.6e-5	+
YOR212W	STE4	2	0.0020	0.0082	1.7e-5	+
YJR086W	STE18		0.0008	0.0218	1.7e-5	+
YHR005C	GPA1	2	0.0005	0.0437	2.1e-5	+
YDR103W	STE5	2	0.0017	0.0298	5.2e-5	+
YBR073W	RDH54	1	0.0053	0.0116	6.2e-5	+
YGR044C	RME1	3	0.0009	0.0720	6.8e-5	+
YGL248W	PDE1	4	0.0182	0.0040	7.3e-5	+
YPL038W	MET31	4	0.0292	0.0027	8.0e-5	+
YDR088C	SLU7		0.0303	0.0038	1.2e-4	-
YGL052W			0.0117	0.0109	1.3e-4	-
YJL157C	FAR1	2	0.0013	0.1141	1.4e-4	+
YPR122W	AXL1	2	0.0091	0.0163	1.5e-4	+
YIL099W	SGA1	3	0.0063	0.0267	1.7e-4	-
YLR233C	EST1	1	0.0226	0.0090	2.0e-4	-
YKL182W	FAS1		0.0578	0.0035	2.1e-4	-
YMR053C	STB2	3	0.0028	0.0884	2.5e-4	-
YNL319W			0.0123	0.0222	2.7e-4	-
YFR012W			0.0088	0.0125	2.8e-4	-
YNL188W	KAR1	2	0.0019	0.1890	3.6e-4	-
YGL193C			0.0014	0.2654	3.8e-4	-
YMR157C			0.1557	0.0026	4.0e-4	-
YIL117C	PRM5		0.0011	0.3689	4.1e-4	-

ChIP experiments

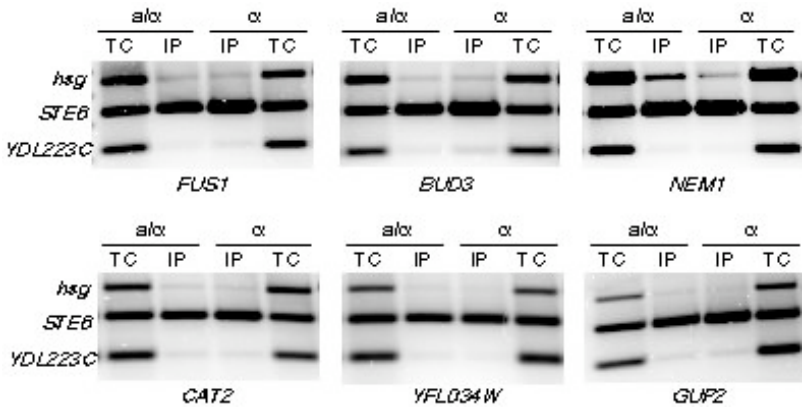




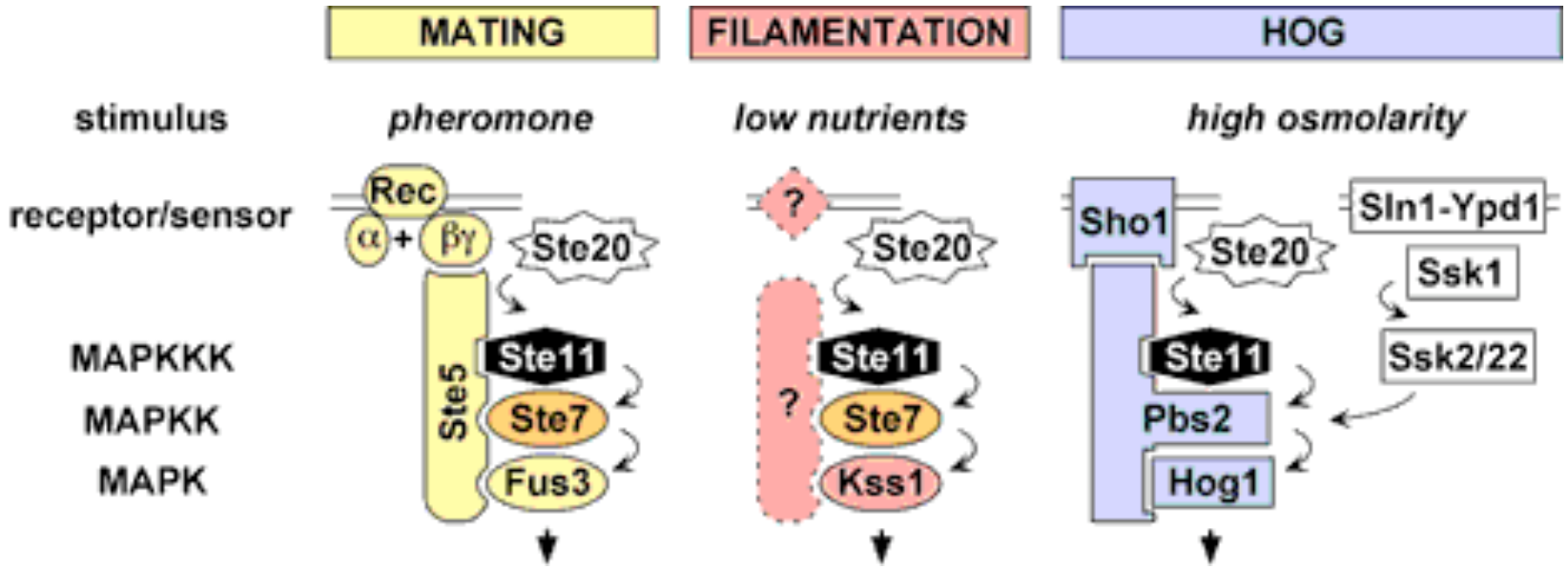
Predicted Direct Targets



Predicted Indirect Targets



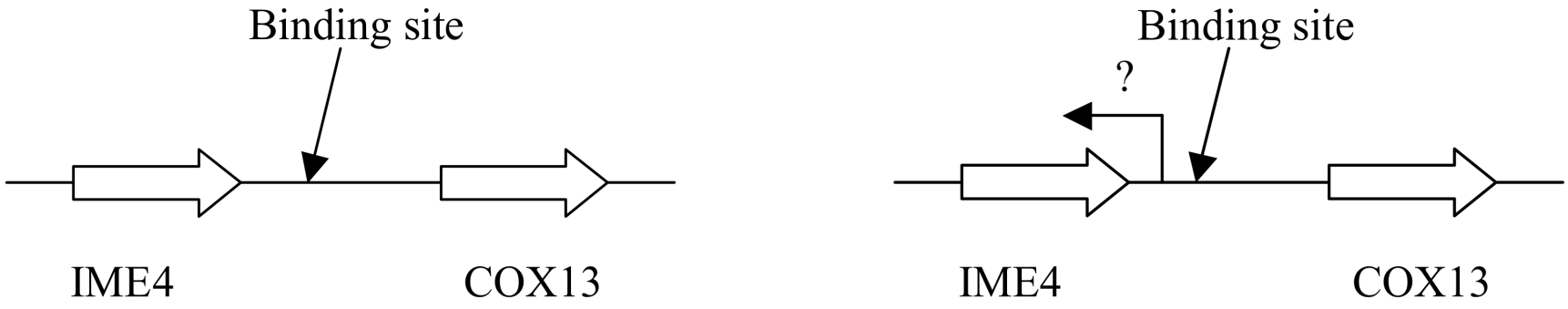
Indirect Effects via Signaling Pathways?



Ste12

Direct Targets in mating pathway: G protein components, Ste5, Fus3, but not the downstream TF Ste12. Possibly affects baseline activity.

Orphan Sites



Could it be regulating *IME4* via an antisense transcript?
Ongoing follow-up work in Vershon lab.

Conclusion

- Ability to design classifying surfaces appropriate for the problem
- Principled way of determining cutoffs
- Experimental tests encouraging
- Need studies of generalization properties

Collaborators

Viji Nagaraj
Ruadhan O'Flanagan

Richard Lavery (IBPC, Paris)
Guillaume Paillard (IBPC, Paris)

Boris Shraiman
Marko Djordjevic (Columbia)

Adrian Bruning
Sean Hanlon (UNC)
Jonathan Mathias (U. Wisc.)
Andrew Vershon