Learning Methods for DNA Binding in Computational Biology

Mark Kon

Dustin Holloway

Yue Fan

Chaitanya Sai

Charles DeLisi

Boston University

IJCNN
Orlando
August 16, 2007

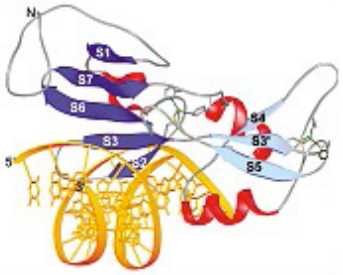
Outline

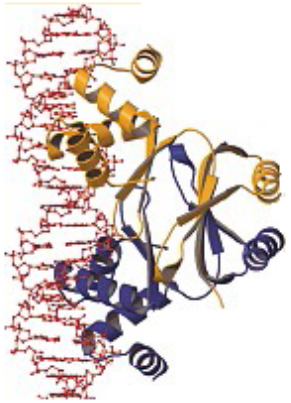
- Background on Transcription Factors and Regulation
- Motivation
- Use of SVM
- Inferences
- Regulatory pathway prediction
- Human genome work

Wealth of Sequence and Biochemical Data

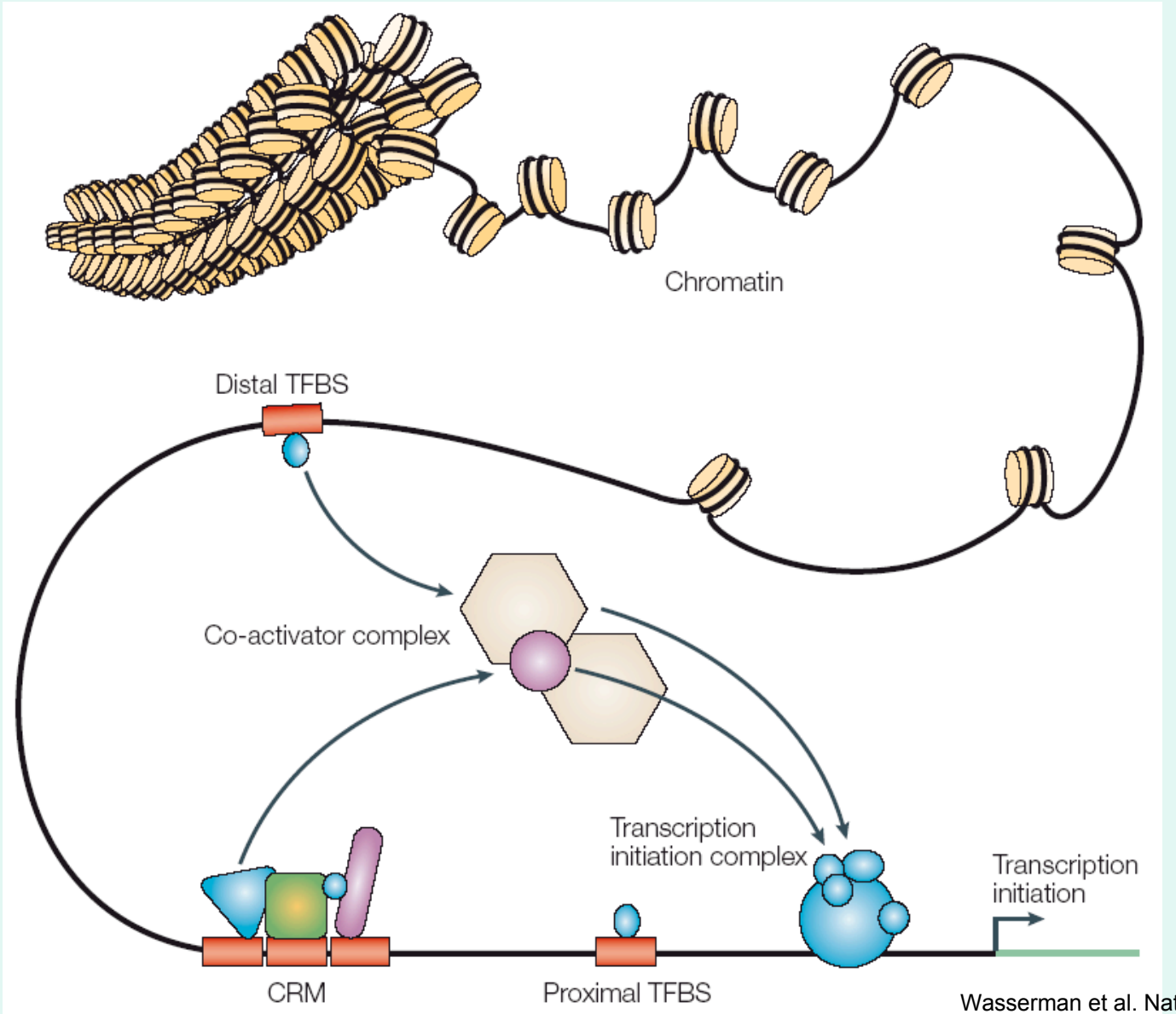
The amount of sequence data available is rapidly increasing. Over 1,500 genome projects are ongoing.

There is a need for techniques that can rapidly determine which sequences in a genome are functional.



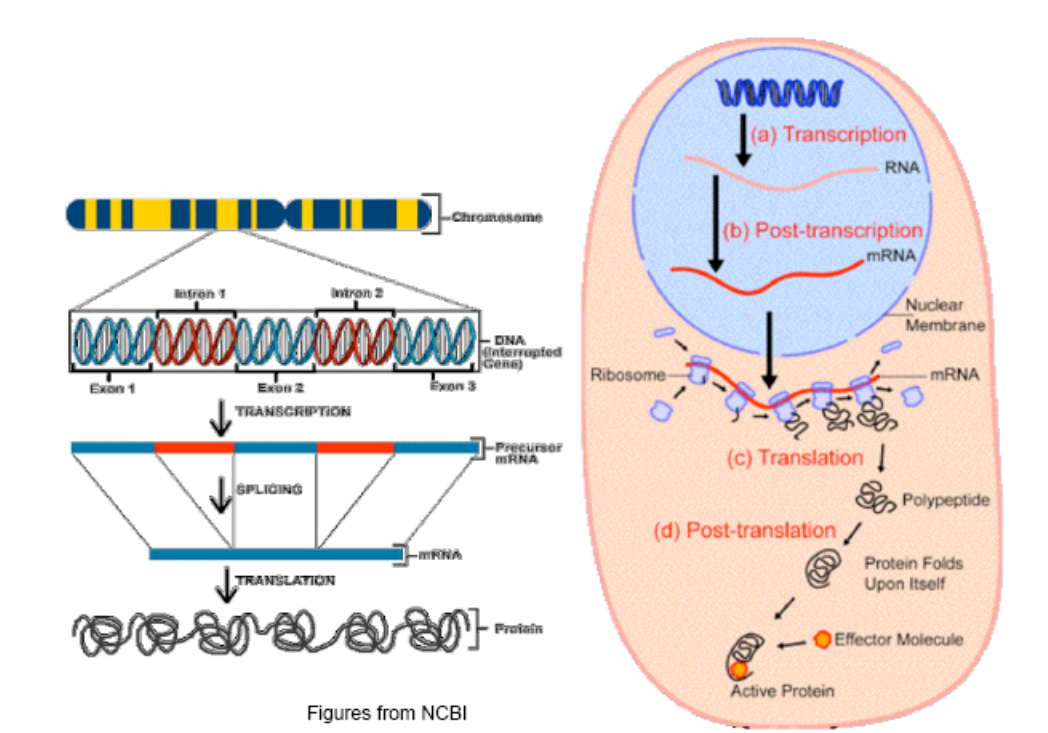


Biology: Transcription and Regulatory Control



Wasserman et al. Nature Reviews Genetics 2004, 5 276-287

Transcription: key to gene expression

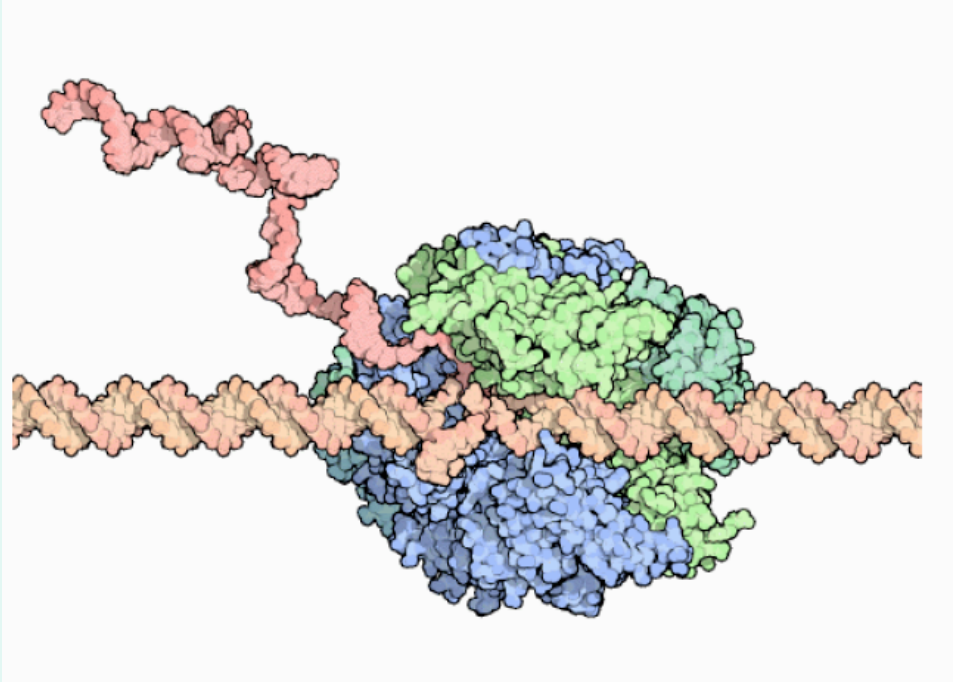


Figures from NCBI

DNA is transcribed into RNA and eventually proteins

Our concern: the first step - initiation of transcription

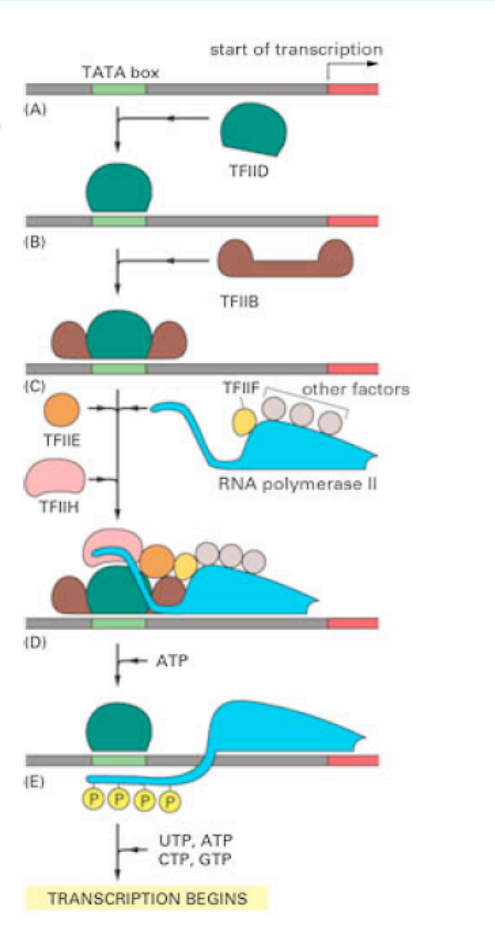
The transcription process



RNA Polymerase runs along DNA to produce RNA copy

Initiation of this process occurs when a TF binds to DNA at the start of transcription

The beginning: TF binds to DNA

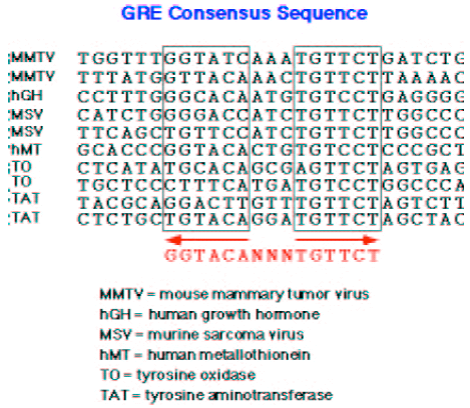


Basics of Transcription

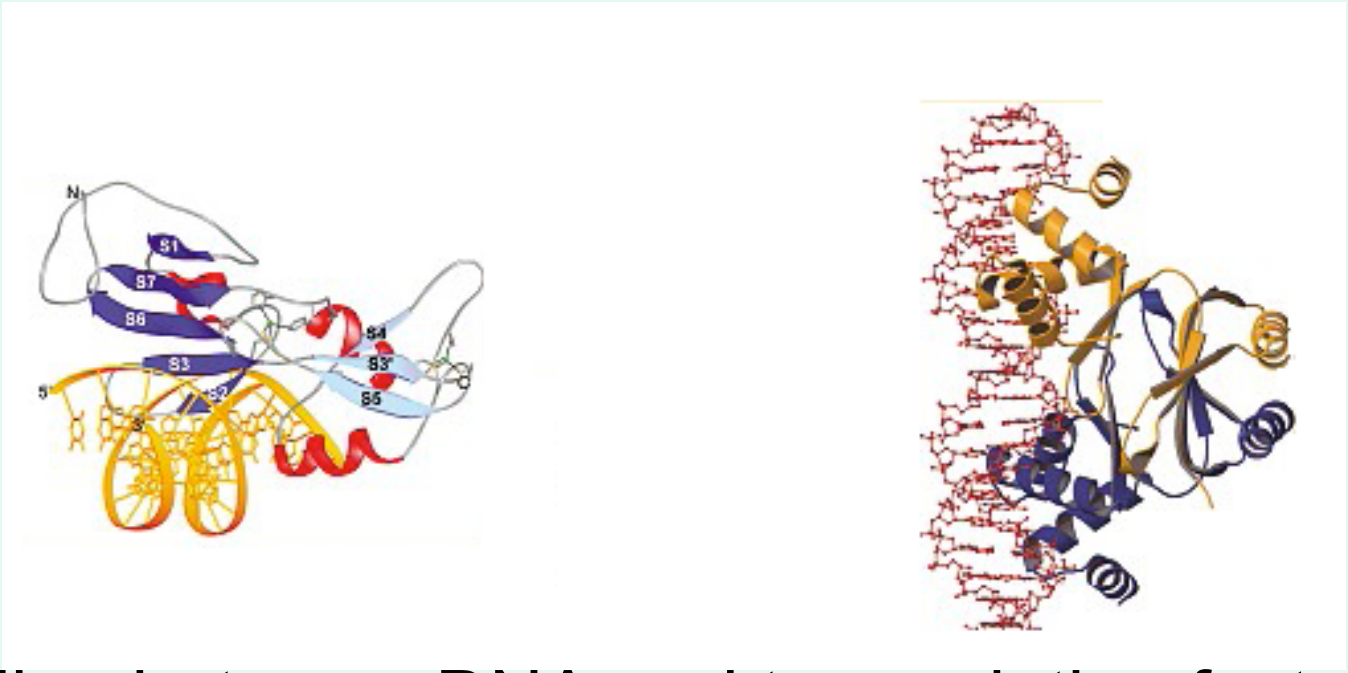
- 1. Promoter is a region of the DNA which tries to attract RNA polymerase so that transcription can be initiated. When cell transcribed it is expressed as a protein.
- 2. Differences between cells are determined by which proteins they produce, which are determined by which genes are expressed.
- 3. Promoter region of DNA contains regulatory sequences which attract proteins called transcription factors (TF). The presence of these proteins is required for transcription with RNA polymerase to begin.
- 4. Regulatory sequences consist of inexactly repeating patters (motifs)
- 5. Motifs stand out as highly similar patterns across species - their function is to attract very specific transcription factors.

- Regulatory sequences on the DNA attract the TF
- Recurring attracting sequences are *motifs* or *consensus sequences*

Regulatory sequences



Transcription Factor Binding



Binding between DNA and transcription factors (TF's) is hard to predict chemically

Goal: For a given TF in yeast or human, determine which genes' promoters it binds to, and where.

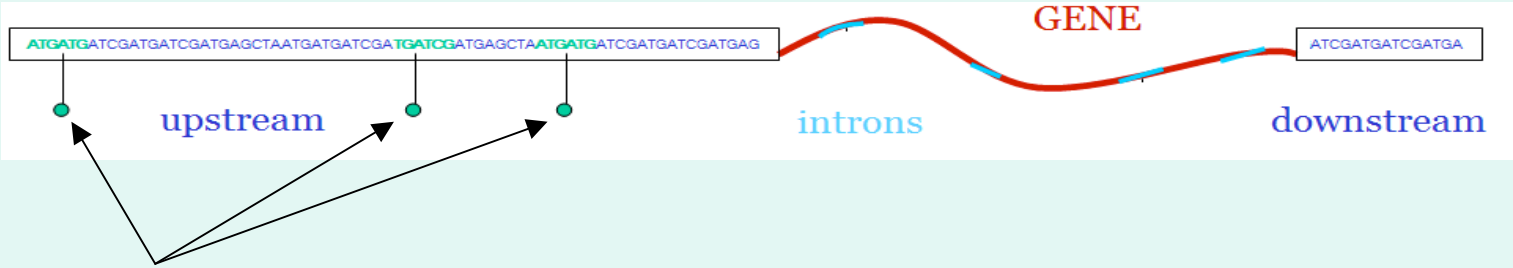
Summary

- High throughput technologies, including ChIP-chip data, are rapidly increasing experimental information about transcription factor binding to DNA
- Identification of TF binding sites in the genome remains difficult and incomplete
- Machine learning approaches have potential to supplant difficult experimental methods
- SVM methods studied here have sensitivity of 70% and positive predictive value of 90% on the average.

Summary

- Applications to inferences on biochemical pathway information are given

Binding Site Representation



GCN4 binding sites

AGACCAA
GGACGCA
TGACTCA
TGACTCA
TGACTCA
TGACTCA
TGACTCA
TGACTCA
TGACTCA
TGACTCA
TGACTCA
TGACTCC
TGAGTCC
TGAGTCG
TGAGTCT
TGAGTCT
TGAGTCT
TGTGTGT

TGAsTCa

PSSM = Position Specific Scoring Matrix

Probability matrix for GCN4

pos	0	1	2	3	4	5	6
A	0.059	0.036	0.927	0.029	0.044	0.101	0.697
C	0.017	0.022	0.008	0.662	0.027	0.827	0.043
T	0.908	0.077	0.054	0.058	0.911	0.043	0.214
G	0.015	0.866	0.012	0.251	0.018	0.029	0.045

Sequence logo

•G. D. Stormo, DNA Binding Sites: Representation and Discovery., *Bioinformatics* 16 16-23,2000
•W. W. Wasserman and A. Sandelin, Applied Bioinformatics for the Identification of Regulatory Elements, *Natue Reviews Genetics* 5 276-287,2004.

Support Vector Machines

Assume a fixed species $\mathcal{S}$ (e.g. baker's yeast, s. *cerevisae*) has genome $\mathcal{G}$ (full set of genes).

Gene g begins transcription (for protein production) when a *transcription factor (TF)* t (a protein) chemically binds to it.

Question: given a fixed TF t, which genes $g \in \mathcal{G}$ does it bind to?

Chemically hard to solve -

Machine learning approach

Consider training data set

$$\mathcal{D}_0 = \{(g_i, y_i)\}_{i=1}^{n},$$

where $g_i \in \mathcal{G}$ and $y_i \in \mathbb{B} = \{-1, 1\}$.

Assume

$$y_i = \begin{cases} 1 & \text{if } g_i \text{ attaches the TF} \\ -1 & \text{otherwise} \end{cases}.$$

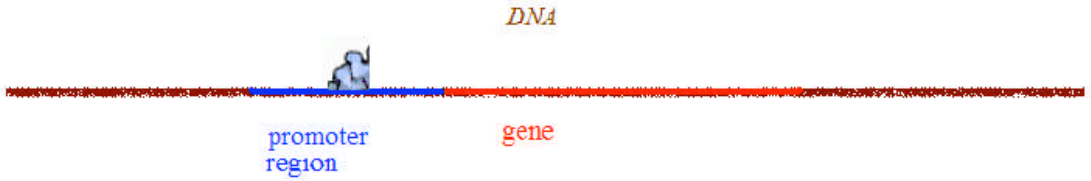
How to learn $f_0 : \mathcal{G} \to \mathbb{B}$ from examples?

First define $g \in \mathcal{G}$ formally by its promoter sequence

Machine learning approach

$$\mathbf{p} = \mathbf{p}(g) = ACGGTCTGGT...CGT$$

= DNA sequence of promoter region of gene



(promoter region is where TF will attach; in yeast it has ~1000 bases).

Effectively

$$f_0 \colon \mathcal{A}^{1000} \to \mathbb{B},$$

with $\mathcal{A} = \{A, G, C, T\}$

Use of feature maps

Remark: If we map $\mathcal{A}$ into numbers (e.g., $g = 0211323113...213$), f_0 difficult to guess from examples $\mathcal{D}$.

Feature maps

A solution: map $\mathcal{G}$ into a space where it's easier to classify.

Sample feature maps

Example: Feature map

$$\Phi_1(g) = \mathbf{x}(\mathbf{p}(g)) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{104} \end{bmatrix}$$

with x_i = # hits in $\mathbf{p}(g)$ by PSSM for TF t_i (e.g., from a list of 104 TF's for yeast).

Example: Microarray expression data for gene g

$\Phi_2(g) = \mathbf{x}$ = vector of expression levels of g in 25 microarray experiments

Sample feature maps

Example:

$$\Phi_3(g) = \mathbf{x}(\mathbf{p}(g)) = \text{vector of string counts in } \mathbf{p}(g)$$

Consider ordered list

string1	AAAAAA
string2	AAAAAC
string3	AAAAAG
string4	AAAAAT
string5	AAAACA
⋮	⋮

of all strings of 6 base pairs.

Sample feature maps

Note $\mathbf{x} = \mathbf{x}(g)$ has components:

$$x_i = \text{\# appearances of string } i \text{ in (upstream region of) } g.$$

$\mathbf{x}$ has $4^6 = 4{,}096$ components; $F = \mathbb{R}^{4{,}096}$.

We thus have a set of *feature maps* $\Phi_i : \mathcal{G} \to F_i =$ *feature spaces* :

$$\Phi_i(g) = \mathbf{x}_i \in F_i;$$

For yeast, we use $k = 26$ such feature maps (some of them highly discriminatory).

Concatenation of feature spaces

Now form full feature space as direct sum:

$$F = F_1 \oplus F_2 \oplus \ldots \oplus F_k,$$

i.e.,

$$\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k)$$

Concatenation of feature spaces

Note: The kernel (thus geometry) of the *full* feature space F is the *sum* of the individual kernels of F_i:

$$K(\mathbf{x}, \mathbf{y}) = \sum_i K_i(\mathbf{x}_i, \mathbf{y}_i),$$

with $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_k)$.

In particular, combining information contained in a collection of kernels $K_i(\,\cdot\,,\,\cdot\,)$ is obtained from just taking their sum.

Basic SVM setup: the discriminating function *f*

With data $\mathcal{D}$, can we find function $f_1 : F \to \mathcal{B}$ which generalizes above examples, so $f_1(\mathbf{x}) = y$ (i.e., correct prediction) for all feature vectors **x**?

Easier: find $f : F \to \mathbb{R}$ where

$$f(\mathbf{x}) > 0 \text{ if } f_1(\mathbf{x}) = 1; \;\; f(\mathbf{x}) < 0 \;\; \text{if} \;\; f_1(\mathbf{x}) = -1.$$

Basic SVM setup: the kernel

Now define geometry of space F by defining dot product:

Assume we are given any *kernel function* $K(\mathbf{x}, \mathbf{y})$ which is *positive definite* and symmetric in $\mathbf{x}, \mathbf{y}$.

We then define geometry of F by defining the nonlinear dot product

$$\mathbf{x} \cdot \mathbf{y} \equiv K(\mathbf{x}, \mathbf{y}).$$

Then apply SVM algorithm using geometry induced by K to find optimized choice of f (here $\overline{\mathbf{x}}_i$ are examples)

$$f(\mathbf{x}) = \sum_i \alpha_i K(\overline{\mathbf{x}}_i, \mathbf{x}) + b = \sum_i K(\mathbf{w}, \mathbf{x}) + b.$$

Basic SVM setup: the kernel

Linear kernel case: $K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ (linear dot product). Then

$$f(\mathbf{x}) = \sum_i \alpha_i \overline{\mathbf{x}}_i \cdot \mathbf{x} + b = \left(\sum_i \alpha_i \overline{\mathbf{x}}_i \right) \cdot \mathbf{x} + b \equiv \mathbf{w} \cdot \mathbf{x} + b.$$

Final classification rule: $f(\mathbf{x}) > 0 \Rightarrow y = 1$ (TF binds gene); $f(\mathbf{x}) < 0 \Rightarrow y = -1$ (TF does not bind).

Learning from training data:

$$Nf = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)) = (y_1, \ldots, y_n).$$

Consider separating hyperplane $H : f(\mathbf{x}) = 0$:

Basic SVM setup: diagram

Geometric interpretation

Recall:

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b;$$

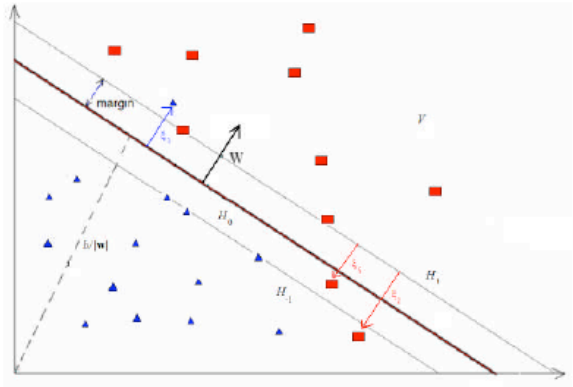


Fig 2: SVM geometry (2 dimensions)

Feature spaces

MOT: Motif hits in S.cerevisiae
CON: Motif hits conservation 18 organisms
PHY: Phylogenetic profile
EXP: Expression correlation
GO: GO term profile
KMER: K-strings – 4,5,6-mers
S1: Split 6-string 1 gap kkk_kkk

Feature spaces

S2:	Split 6-string 2 gaps kkk__kkk
S3:	Split 6-string 3 gaps kkk___kkk
S4:	Split 6-string 4 gaps kkk____kkk
S5:	Split 6-string 5 gaps kkk_____kkk
S6:	Split 6-string 6 gaps kkk______kkk
S7:	Split 6-string 7 gaps kkk_______kkk
S8:	Split 6-string 8 gaps kkk________kkk

Feature spaces

M01: 6-string with 1 mismatch (count 0.1)
M05: 6-string with 1 mismatch (count 0.5)
ENT: Condition specific TF-target correlation
BIT: Nucleotide sparse binary encoding
CRV: Promoter Curvature prediction
HC: Homolog Conservation
HYD: Hydroxyl Cleavage

Feature spaces

KPo: Kmer median positions from start
KPr: Kmer Probabilities (-log pval)
MT: Promoter Melting Temperature-20bp window
DG: Promoter Melting Delta G profile-20bp win
BND: Promoter bend prediction

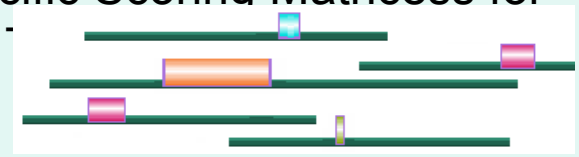
Feature spaces

Many of these methods are not so reliable on their own, but can combine using statistical inference to yield a more powerful prediction scheme.

Selection of Features: Rationale

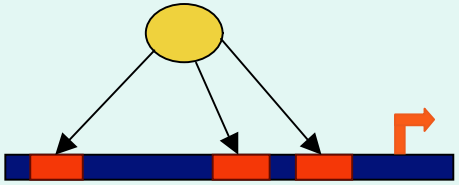
Promoter Sequences

Motif Detection using Position Specific Scoring Matricess for 163



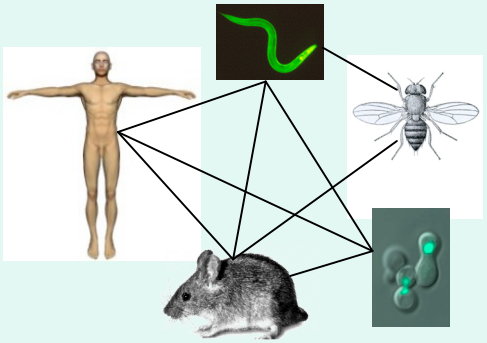
Overrepresentation (Degeneracy) Analysis

Count motifs for each TF-target pair

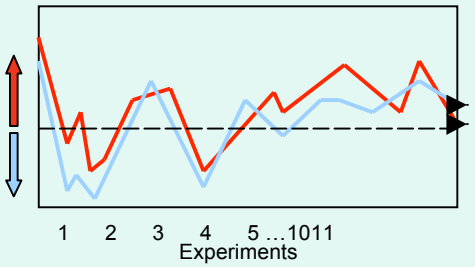


Conservation Analysis

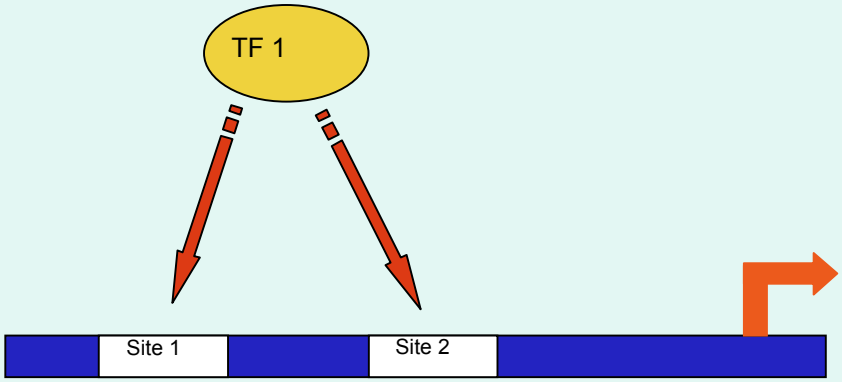
Using 18 Genomes



Expression Correlation Analysis



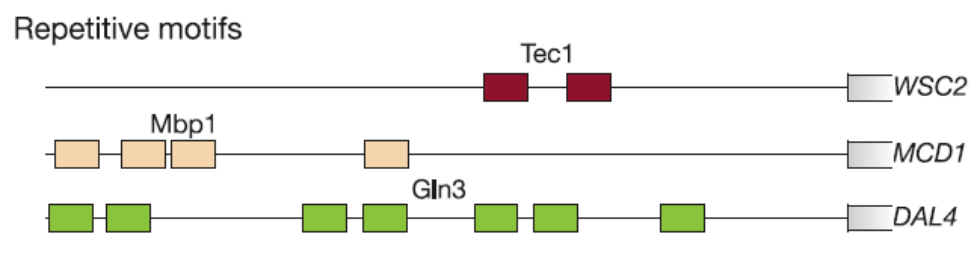
Degeneracy: Repetitive TF Binding Sites



$P(\text{True}|2\text{ hits}) = 2 \cdot P(\text{True}|1\text{ hit})$

Having more than one detected binding site for a TF in the upstream region of a gene increases the likelihood that the TF truly binds the gene.

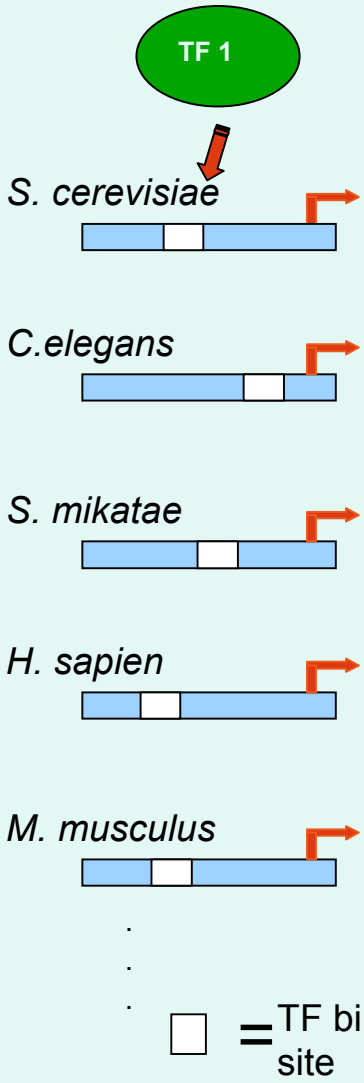
Some transcription factors have a preference for repetitive motifs.



This is Supplementary Table 5 From C. Harbison, E. Fraenkel, R. Young and e. al., Transcriptional Regulatory Code of a Eukaryotic Genome, *Nature* 431 99-104,2004.

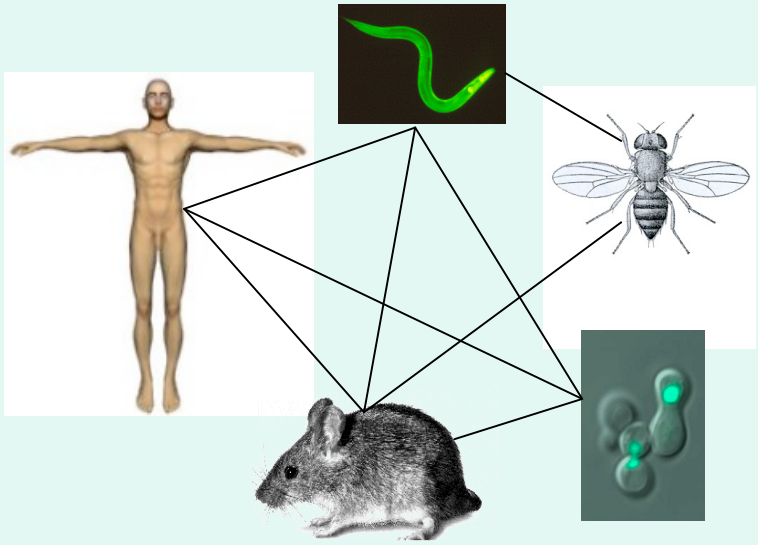
Conservation

Link: Shadowing



18 genomes

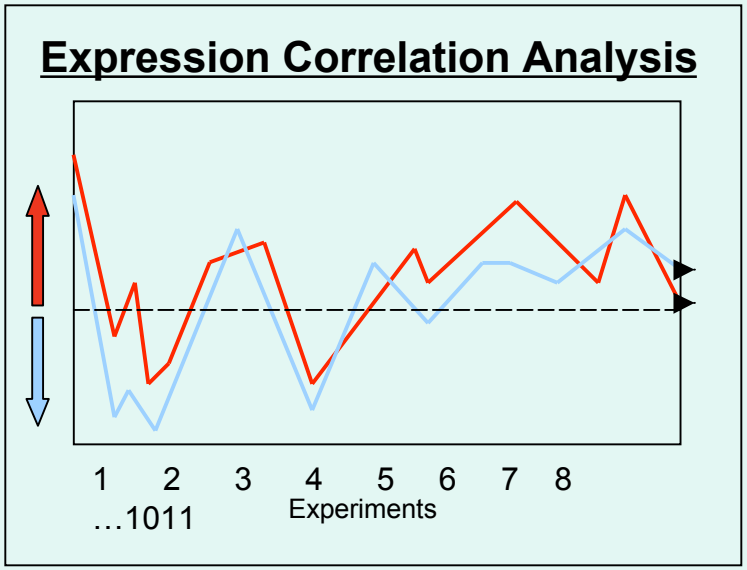
Conservation of a TF binding site in several orthologous upstream regions increases the likelihood that a potential site is a True site



Genomes

S.cerevisiae	*A.thalania*	*D.melanogaster*
S.pombe	*R.norvegicus*	*P.falciparum*
H.sapien	*C.elegans*	*A.gambiae*
N.crassa	*M.musculus*	*S.paradoxus*
S.bayanus	*S.kudriazevii*	
S.mikatae	*S.castelli*	
S.kluyveri	*M.grisea*	

Expression Analysis



Two methods can be used to explore expression relationships:

1. Transcription factors that are highly correlated with potential targets are more likely to regulate those targets.

2. Pairs of genes with highly correlated expression are more likely to be regulated by the same TF.

SVM Algorithm

- 26 feature spaces lead to 26 kernels
- SVM forms hyperplane

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = 0$$

- Kernel

$$K_{ij} = \mathbf{x}_i \cdot \mathbf{x}_j$$

(generalized inner product)

Kernel Choices

Kernel	Parameters	Description
linear	none	$K(\mathbf{x},\mathbf{y})=\mathbf{x}\cdot\mathbf{y}$
polynomial	poly degree d	$K(\mathbf{x},\mathbf{y})=(\mathbf{x}\cdot\mathbf{y}+1)^d$
Gaussian radial basis function (RBF)	σ	$K(\mathbf{x},\mathbf{y}) = \exp\left(\frac{-\lvert \mathbf{x}-\mathbf{y} \rvert^2}{2\sigma^2}\right)$
Gaussian	σ	$K(\mathbf{x},\mathbf{y}) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$

Probabilistic Intepreptation (Platt)

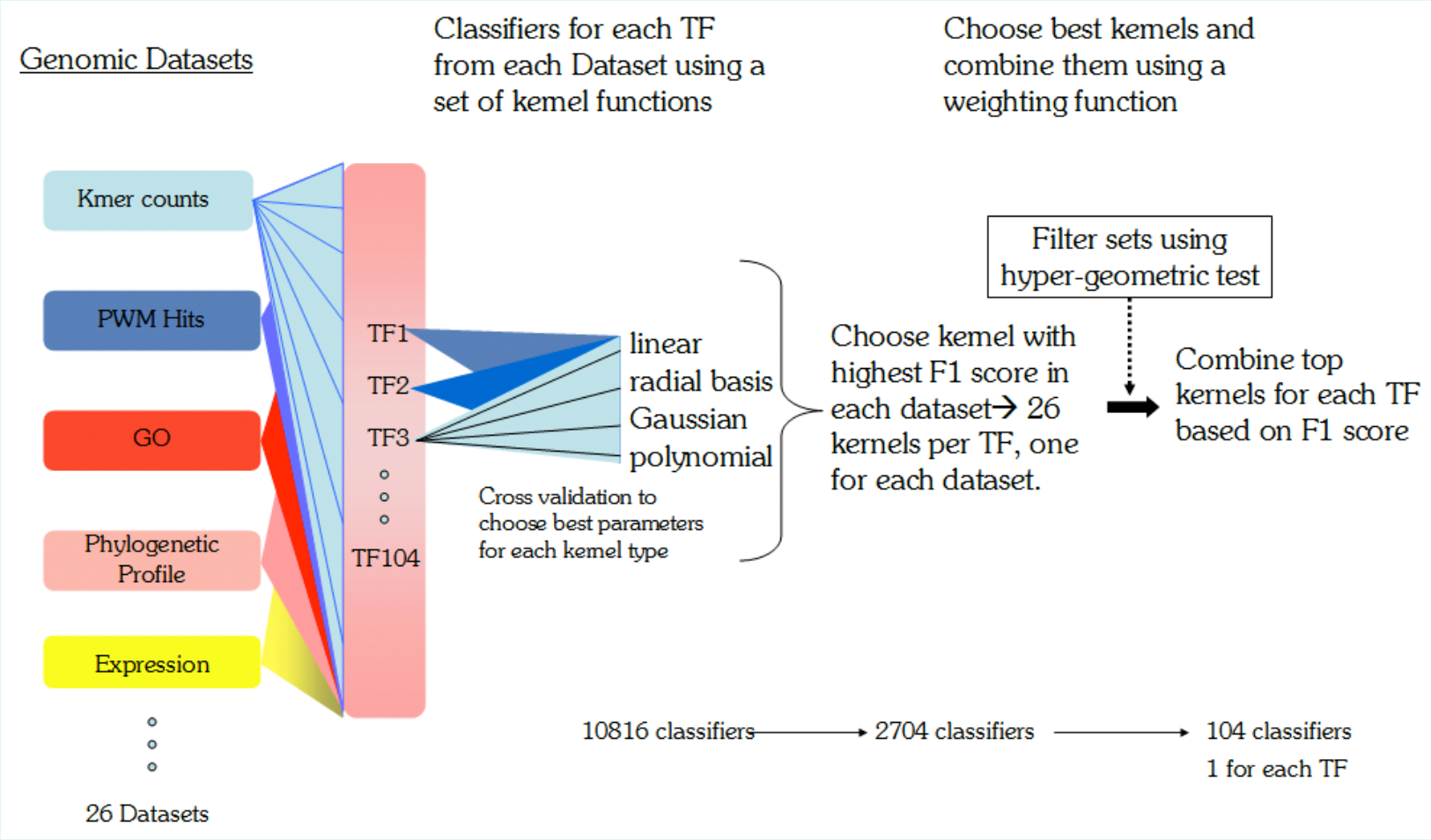
- Rank the data by

$$P(y_i = 1 | \mathbf{w} \cdot \mathbf{x}_i + b)$$

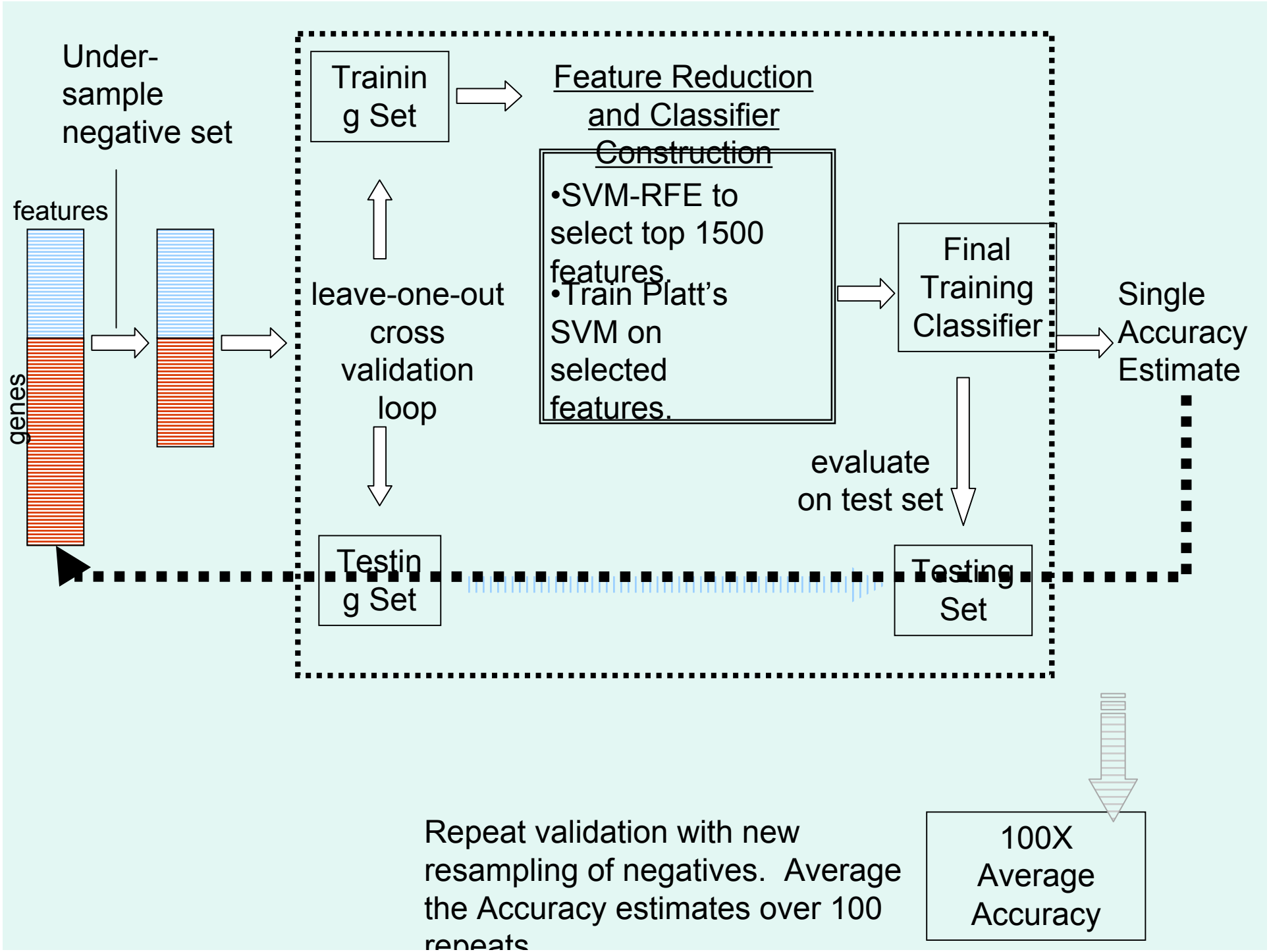
= posterior probability of positive classification given distance of **x** from hyperplane.

Result: empirically based confidence levels given to SVM predictions.

Overall Algorithm



Synthesizing a single classifier from various data sources


Under-sample negative set
features
genes
Training Set
Feature Reduction and Classifier Construction
•SVM-RFE to select top 1500 features.
•Train Platt's SVM on selected features.
leave-one-out cross validation loop
Final Training Classifier
Single Accuracy Estimate
evaluate on test set
Testing Set
Testing Set
Repeat validation with new resampling of negatives. Average the Accuracy estimates over 100 repeats
100X Average Accuracy

Weighting schemes for kernel sums

- Weighted sums of kernels are taken:

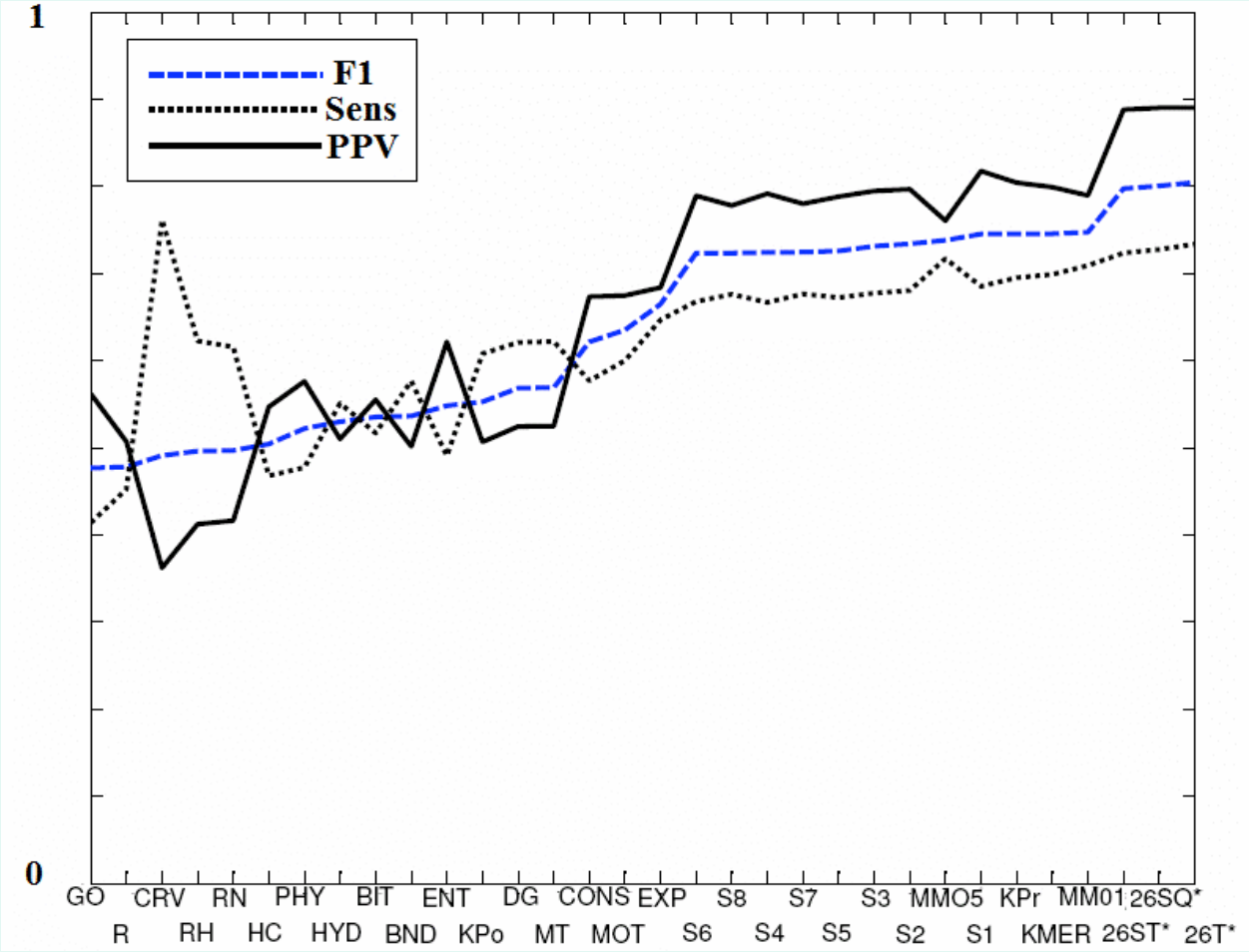
$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{26} \alpha_i K_i(\mathbf{x}, \mathbf{y})$$

Scale with α_i =

- Scaled *F1* score
- Square of scaled F_1 score
- Squared tangent of F_1 score

(note latter have effect of emphasizing higher and better F_1 values)

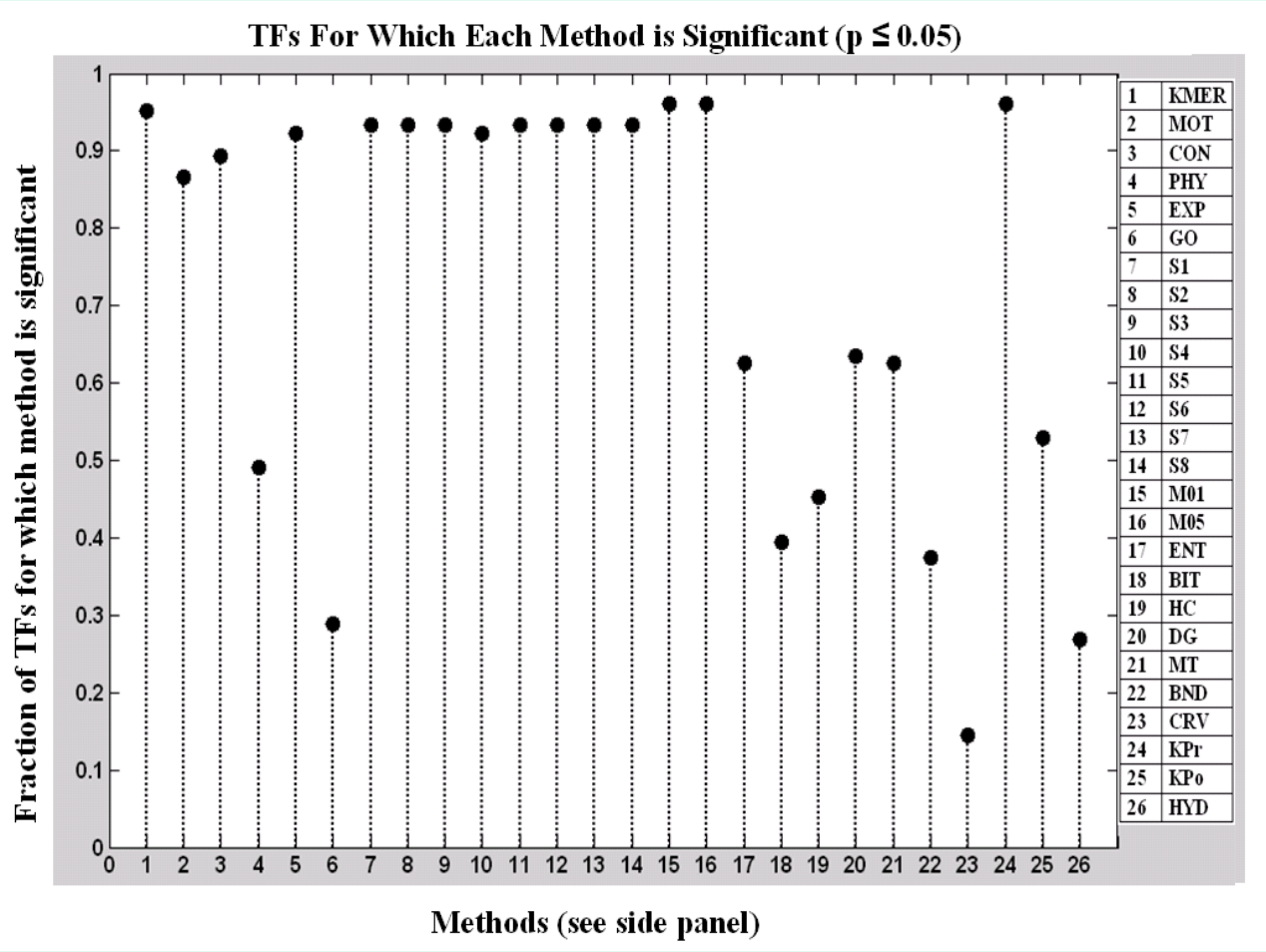
Kernels: accuracy scores



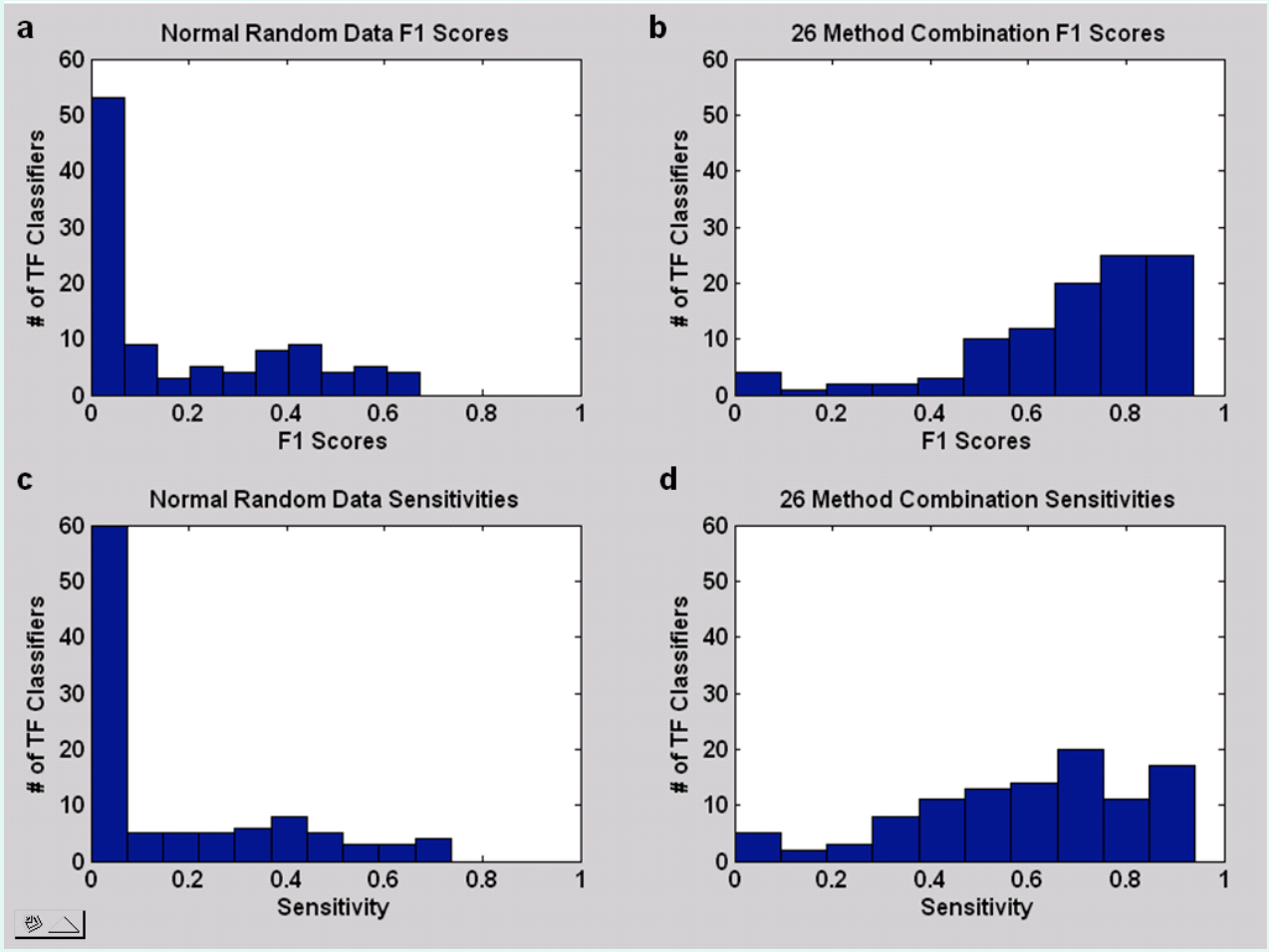
Summary of accuracy

- Best single kernel has sensitivity of .71 and PPV of .82
- Squared-tan weighting gives sensitivity .73 and PPV of .89

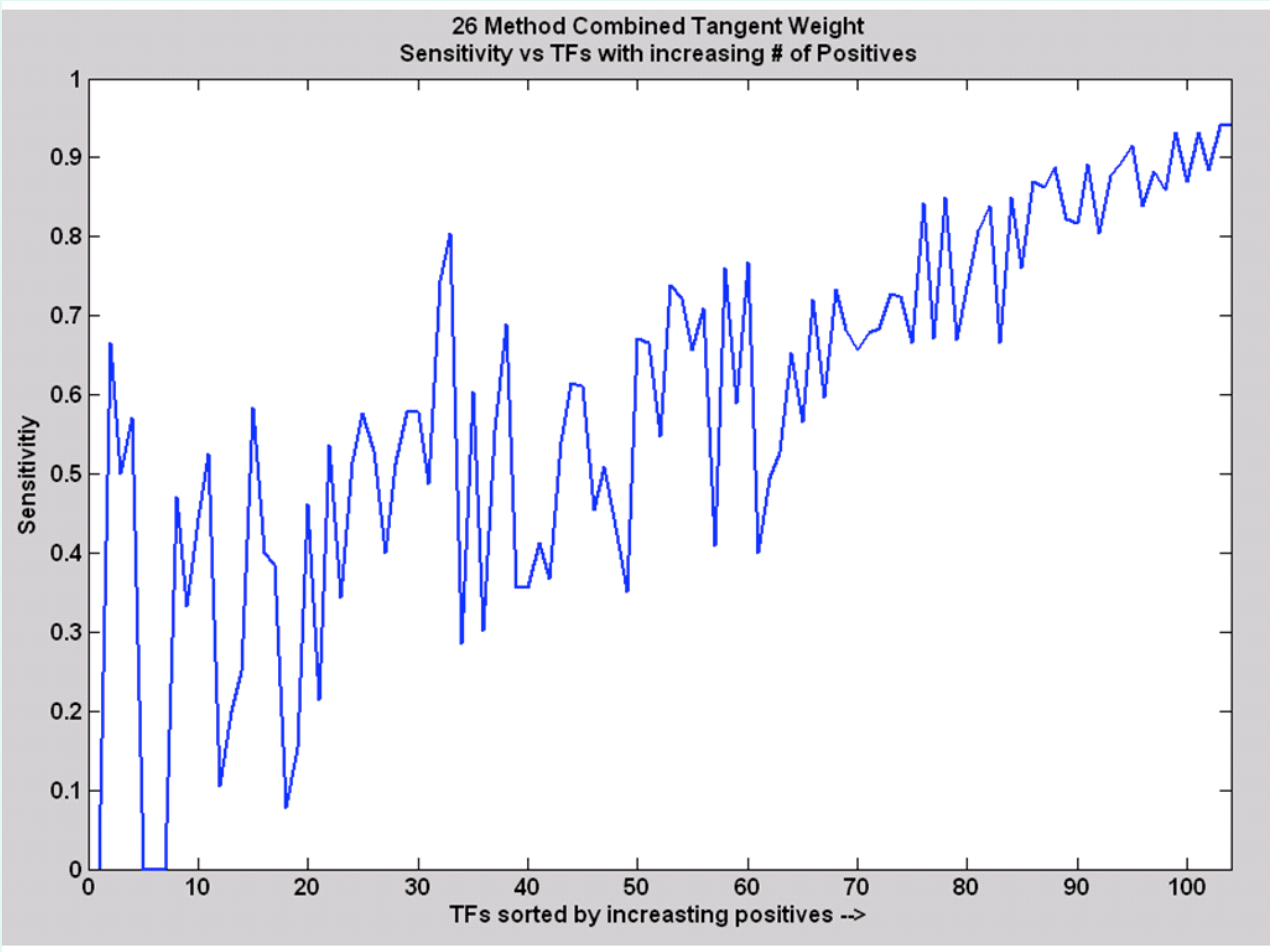
Summary of accuracy



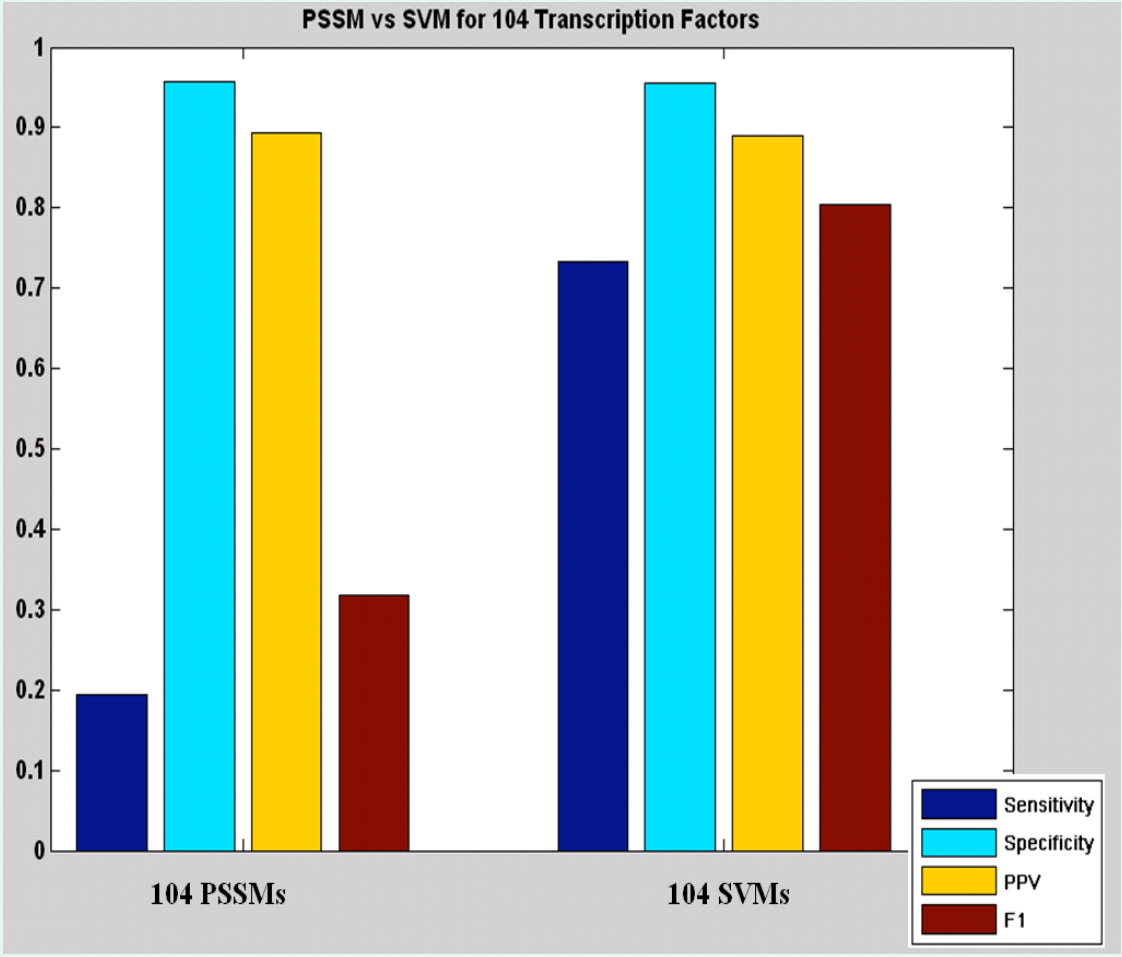
F1 Scores: Random vs. Genomic Data



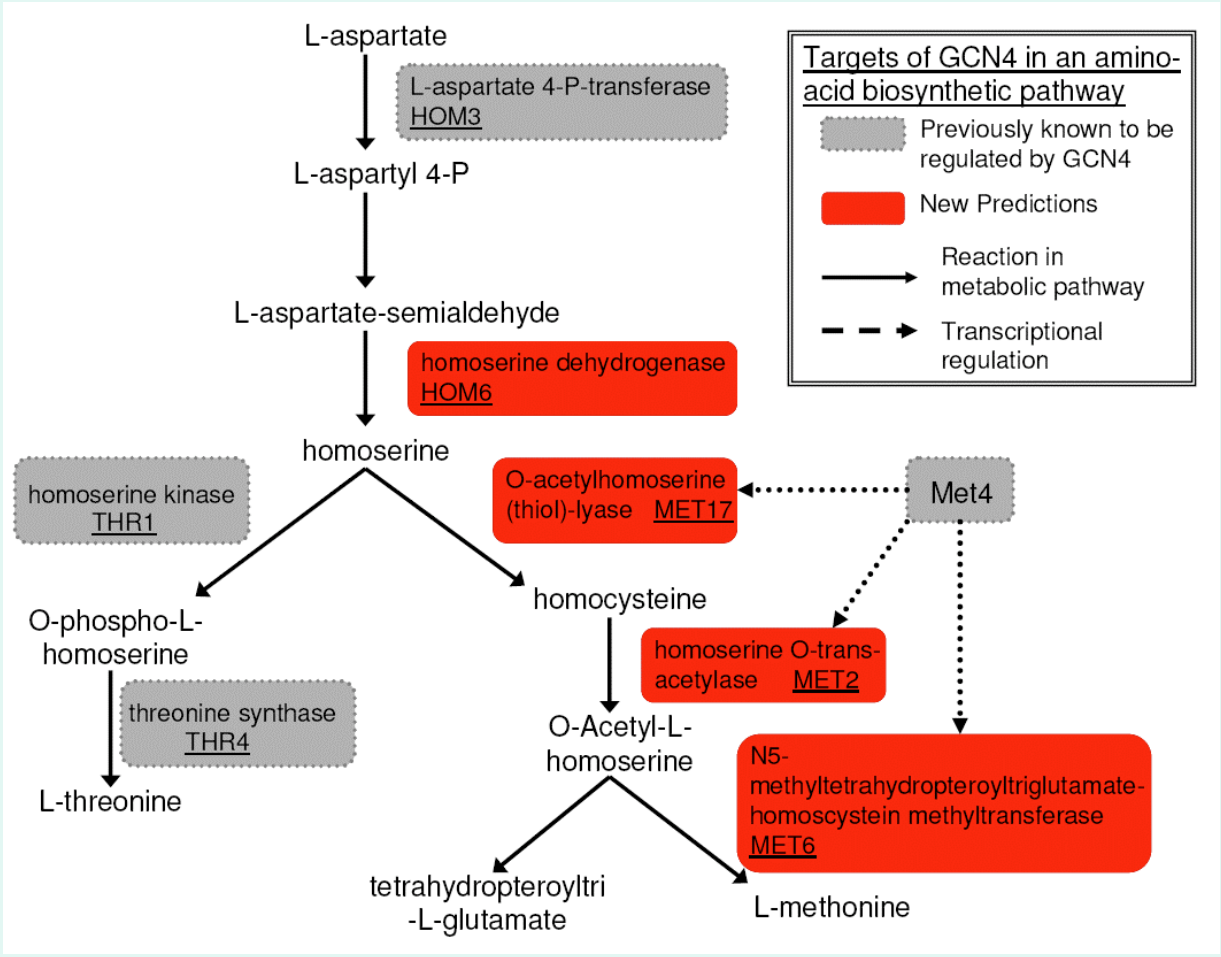
Sensitivity vs. Example Size



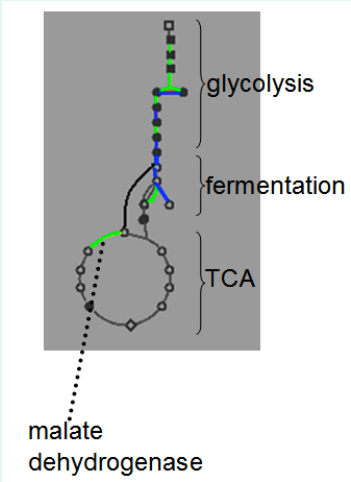
SVM vs. PSSM Scan

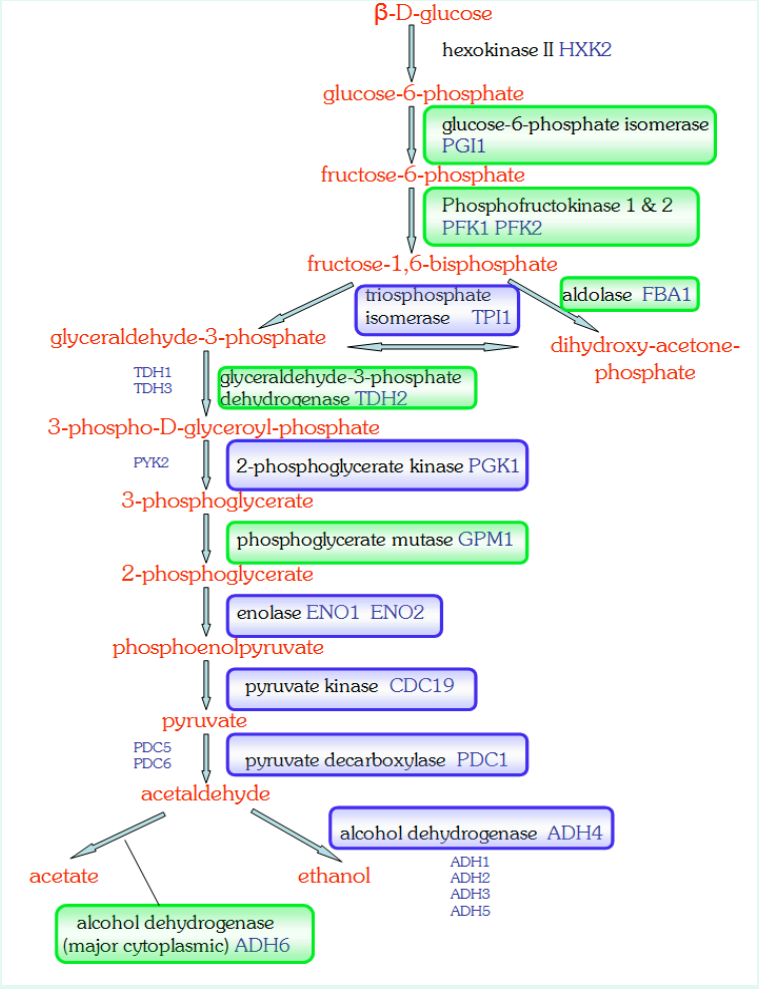


Implications for Pathways: GCN4 and Amino Acid Biosynthesis

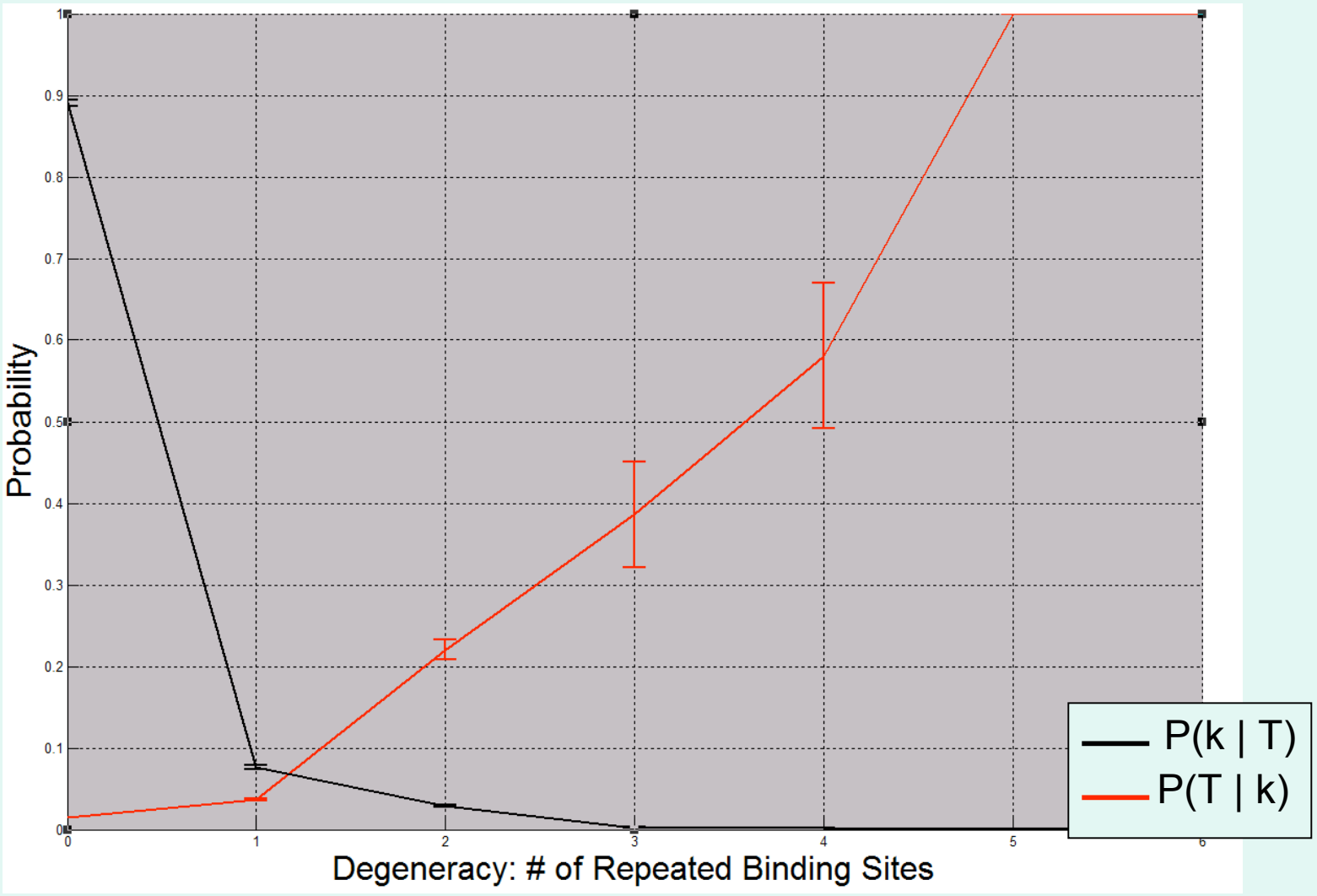


Implications for Pathways: RAP1 and Glycolytic/TCA Cycle



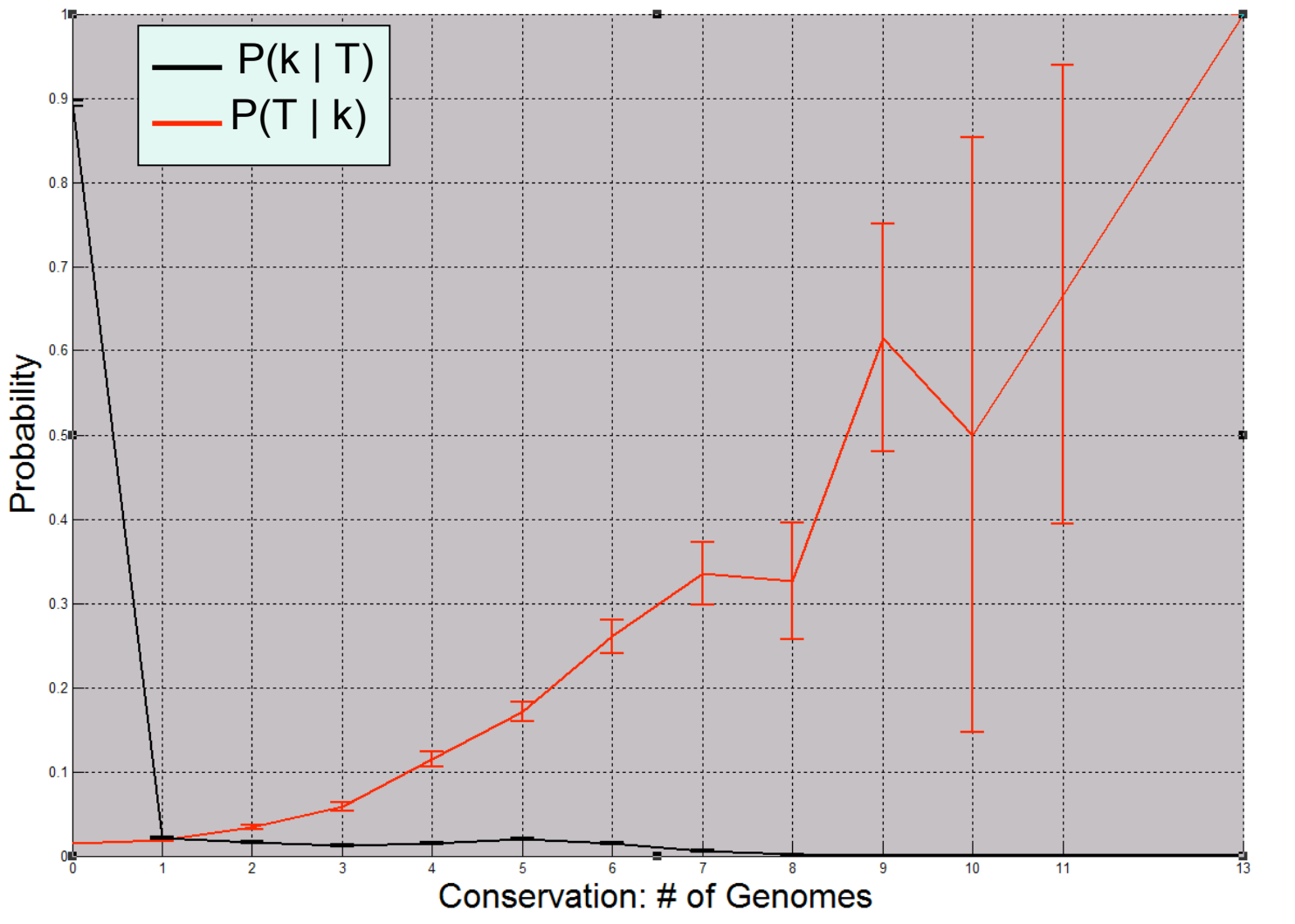


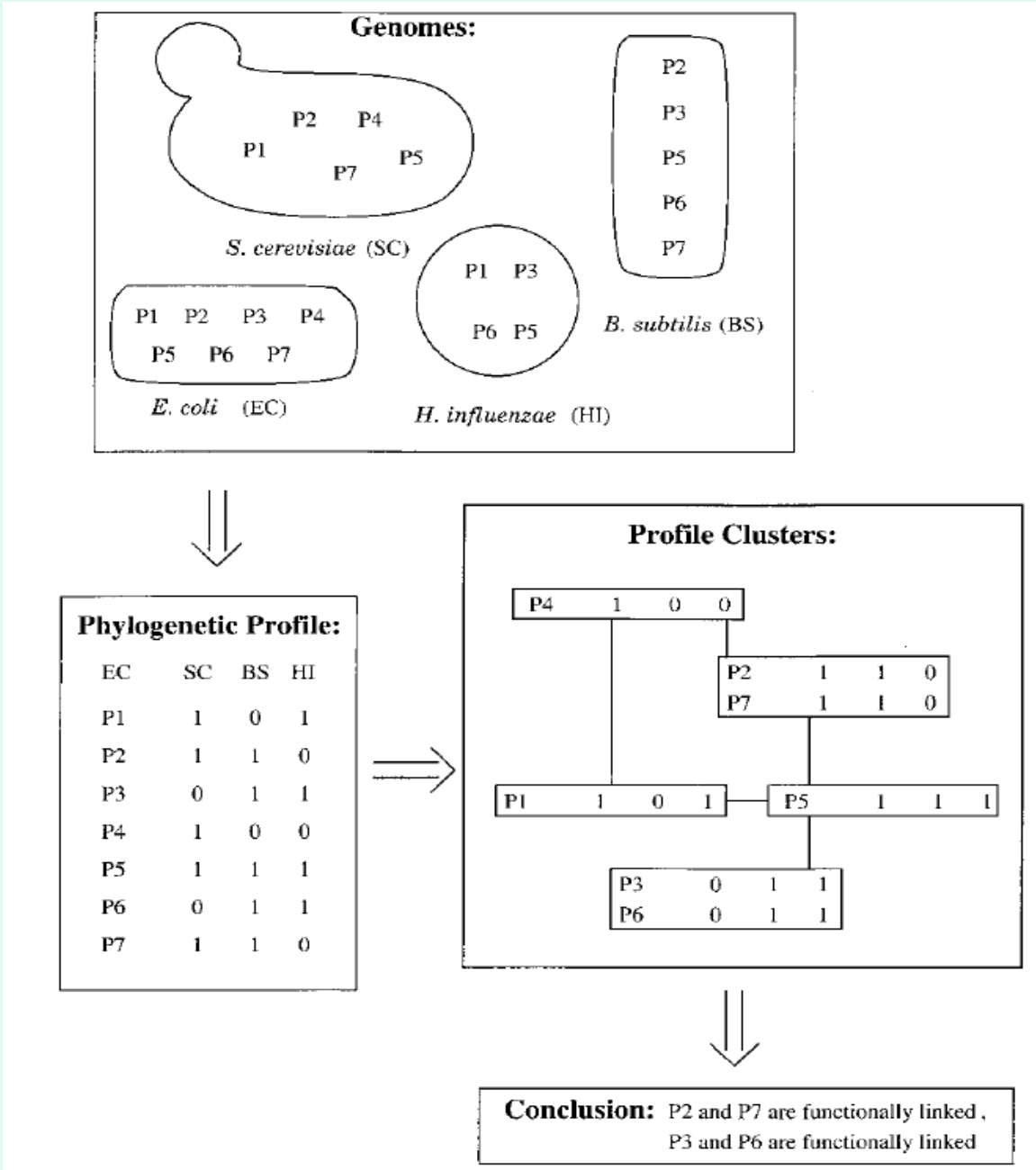
Degeneracy Significance



Degeneracy 0 means not detected by Motifscanner

Conservation Results




Genomes:
P2
P4
P1
P5
P7
S. cerevisiae (SC)
P2
P3
P5
P6
P7
B. subtilis (BS)
P1 P3
P6 P5
H. influenzae (HI)
P1 P2 P3 P4
P5 P6 P7
E. coli (EC)
Phylogenetic Profile:
EC SC BS HI
P1 1 0 1
P2 1 1 0
P3 0 1 1
P4 1 0 0
P5 1 1 1
P6 0 1 1
P7 1 1 0
Profile Clusters:
P4 1 0 0
P2 1 1 0
P7 1 1 0
P1 1 0 1
P5 1 1 1
P3 0 1 1
P6 0 1 1
Conclusion: P2 and P7 are functionally linked,
P3 and P6 are functionally linked

Degeneracy

	Gene1	Gene2	Gene3
Motif1	2	1	2
Motif2	0	0	1
Motif3	1	0	1

Conservation

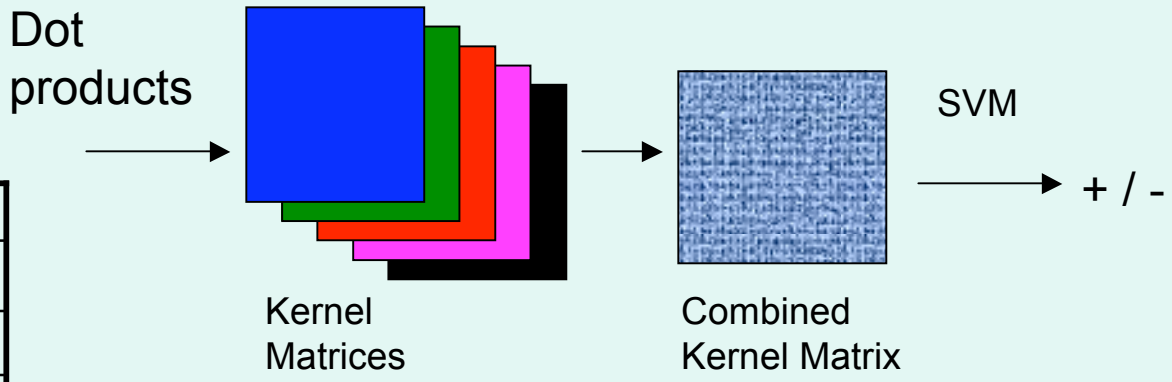
	Gene1	Gene2	Gene3
Motif1	4	8	0
Motif2	5	0	0
Motif3	0	2	2

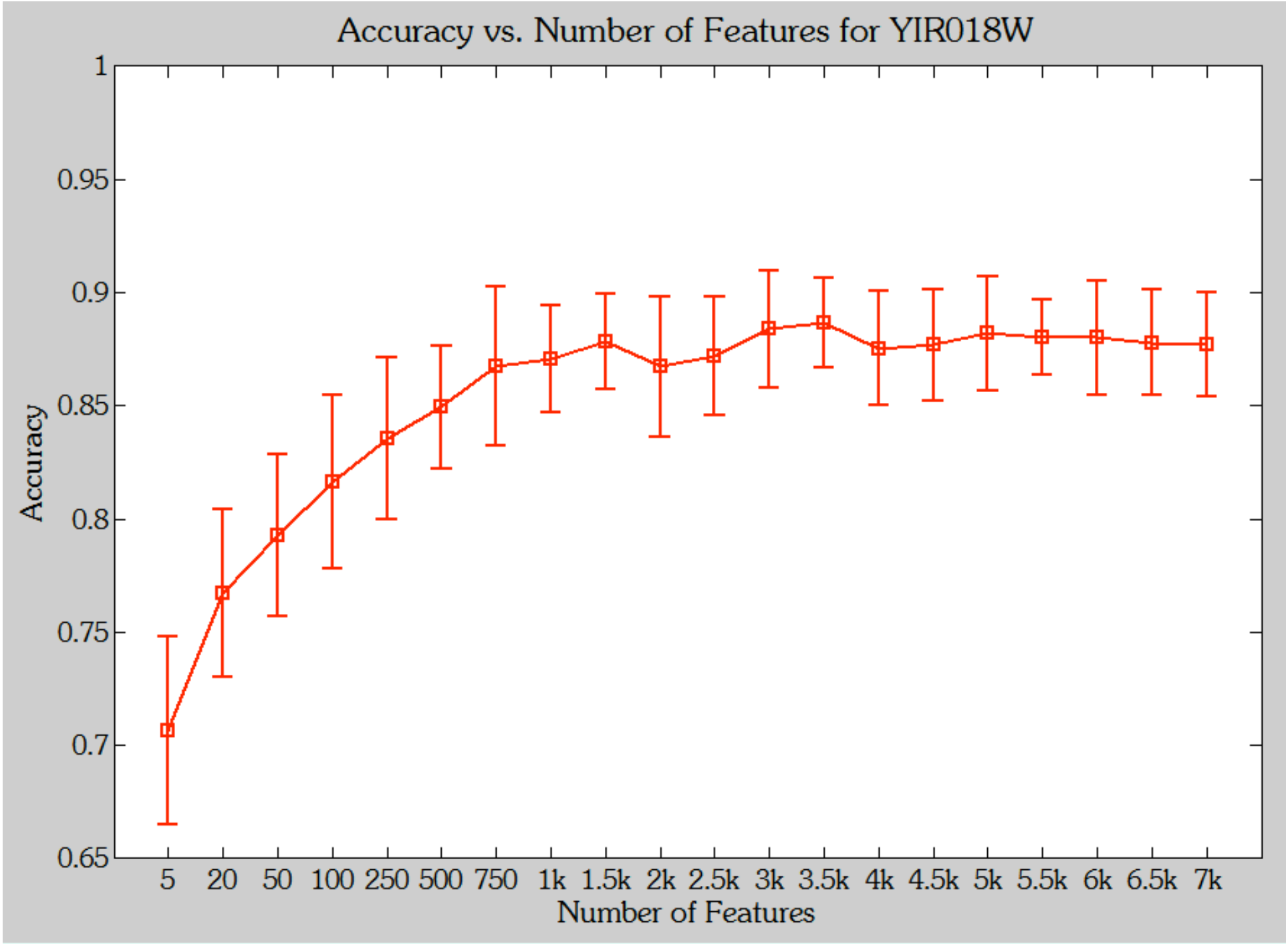
Phylogenetic Profile

	Genome1	Genome2	Genome3
Gene1	1	1	1
Gene2	0	0	1
Gene3	1	1	0

Expression

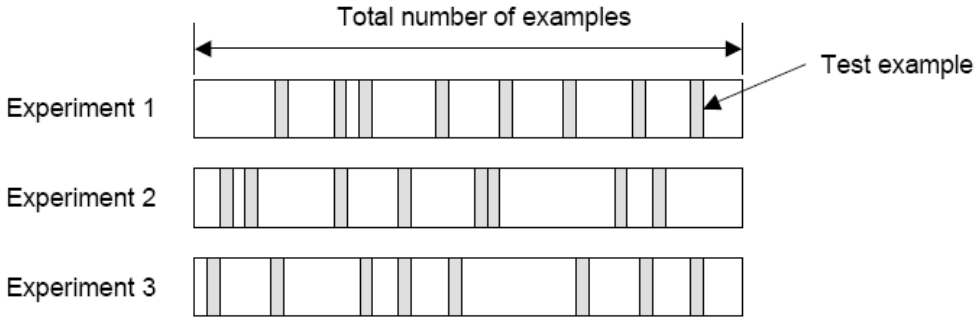
	Exp1	Exp2	Exp3
Gene1	0.32	0.001	0.5
Gene2	-0.2	0.04	-0.001
Gene3	-0.6	0.4	-0.3




Accuracy vs. Number of Features for YIR018W
Accuracy
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
5
20
50
100
250
500
750
1k
1.5k
2k
2.5k
3k
3.5k
4k
4.5k
5k
5.5k
6k
6.5k
7k
Number of Features

K-fold Random Resampling

- **Random Subsampling performs K data splits of the dataset**
  - Each split randomly selects a (fixed) no. examples without replacement
  - For each data split we retrain the classifier from scratch with the training examples and estimate E_i with the test examples



- **The true error estimate is obtained as the average of the separate estimates E_i**
  - This method is significantly better than simple split sample techniques

$$E = \frac{1}{K}\sum_{i=1}^{K} E_i$$

Some human target predictions

WT1 - a TF involved in Wilms' Tumor - makes up 8% of childhood cancers.

SVM predictions for WT1 targets suggest new Wilms tumor models.

Genes in significant loci include several oncogenes and tumor suppressors which are candidates for involvement in cancer progression.

Some human target predictions

Example: chromosomal region 11p15.5

- known to be involved in Wilms' Tumor.

Newly predicted targets for WT1 are statistically enriched (.0005) for genes falling in this region.

Three of these are possible tumor suppressors, i.e., RNH1, IGF2AS, and CD151.

Other regions known to play a role in Wilms' Tumor also contain new target predictions (16q, 1p36.3, 16p13.3, 17q25, and 4p16.3).

Anti-apoptotic (anti-programmed cell death) effects of WT1 are possibly related several new target genes, including BAX and PDE4B - may help mediate the effect.

Some human target predictions

Motif discovery used for new candidate WT1 binding motif:

Potential Binding sites for WT1

A Previous consensus sites
GCGGGGGCG
GCGTGGGAGT
GNGNGGGNG
GNGNGGGNGNS

B Highest ranked k-mer
SVM: CGCG
Oligo-analysis: CGCGGG
Weeder best 6-mer: CGACCG
Weeder best 8-mer: CGCGTCGA
Weeder best 10-mer: CGTCGCGTCG

C Weeder Top Motif

Second Motif

Fig. 3 - Wt1 target motifs:
(A) From literature
(B) Rankings of candidate motif strings as determined by application of SVM to a string feature space str, and from another oligo-analysis.
(C) Top ranked motifs using the Weeder algorithm on SVM-based rankings.

Acknowledgments

Dustin Holloway is responsible for much of the above analysis.

Charles DeLisi initiated this project

Machine Learning Predictions:

http://visant.bu.edu/