Machine Methods for Identifying DNA Binding Sites

Mark Kon, Yue Fan, Dustin Holloway, and Charles DeLisi

Overview

I. TF binding

Goal:

Overview

We are developing tremendous amounts of biological and DNA sequence information.

Large numbers of genomic and proteomic projects are ongoing.

There is an **EXPLOSION** of biological sequence data.

We need to understand basics: how to determine if a string of DNA is functional?

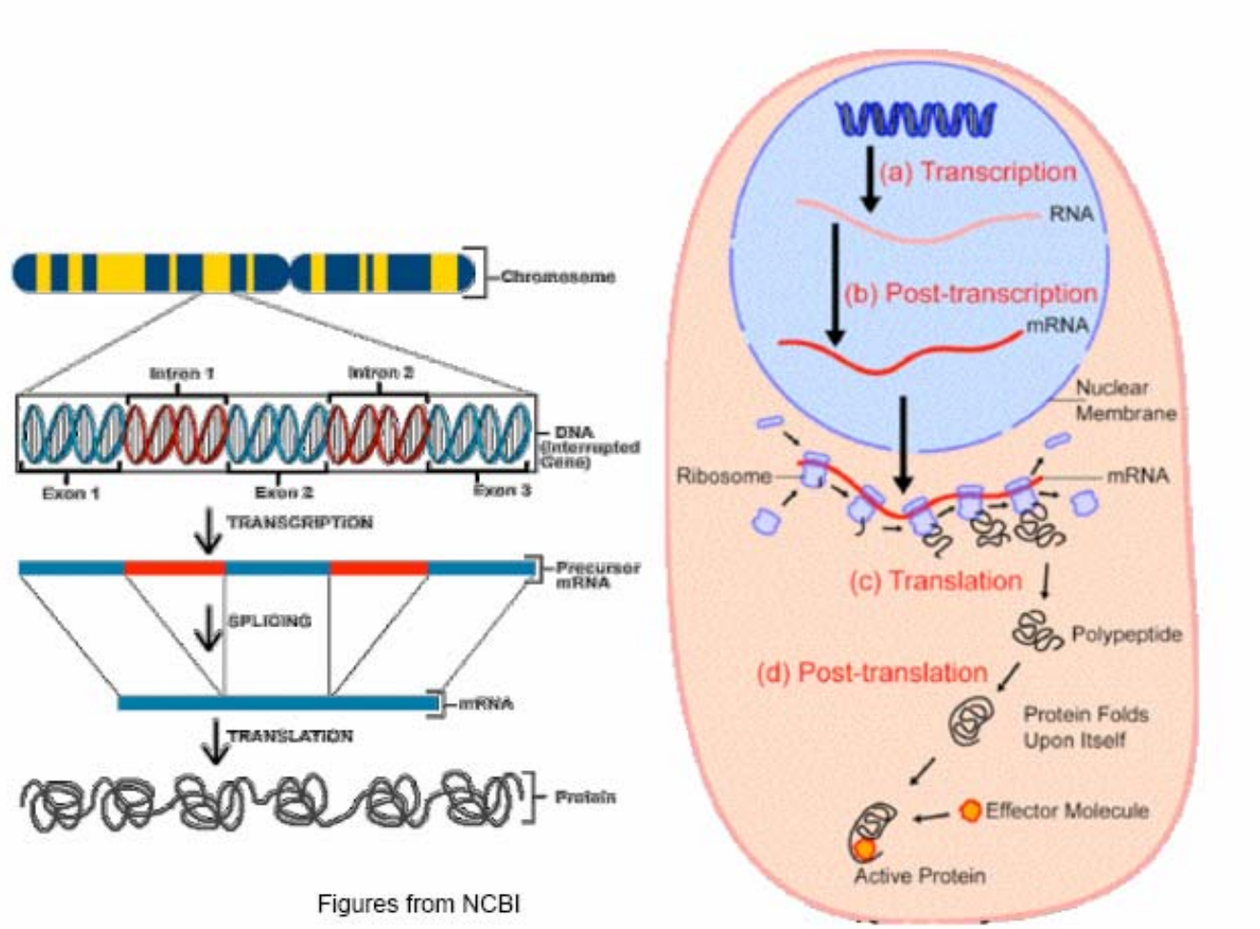
And if functional, what is its function?

- Code a protein?
- Act as a regulatory site (initiating transcription of genetic information)?
- Act as a break point during genetic recombination?

These are very **High Dimensional Problems**, and often a decision (classification) must be made.

Identification of transcription promoter motifs

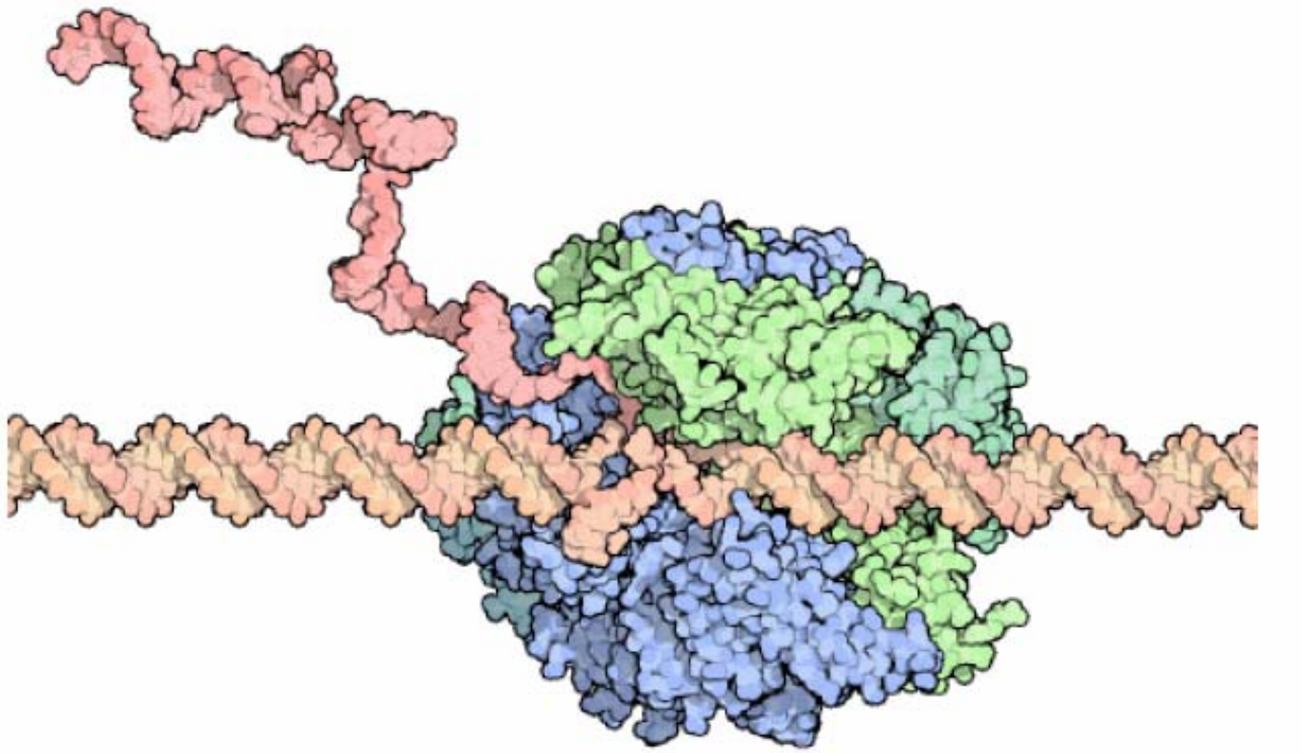
The cellular process:



Figures from NCBI

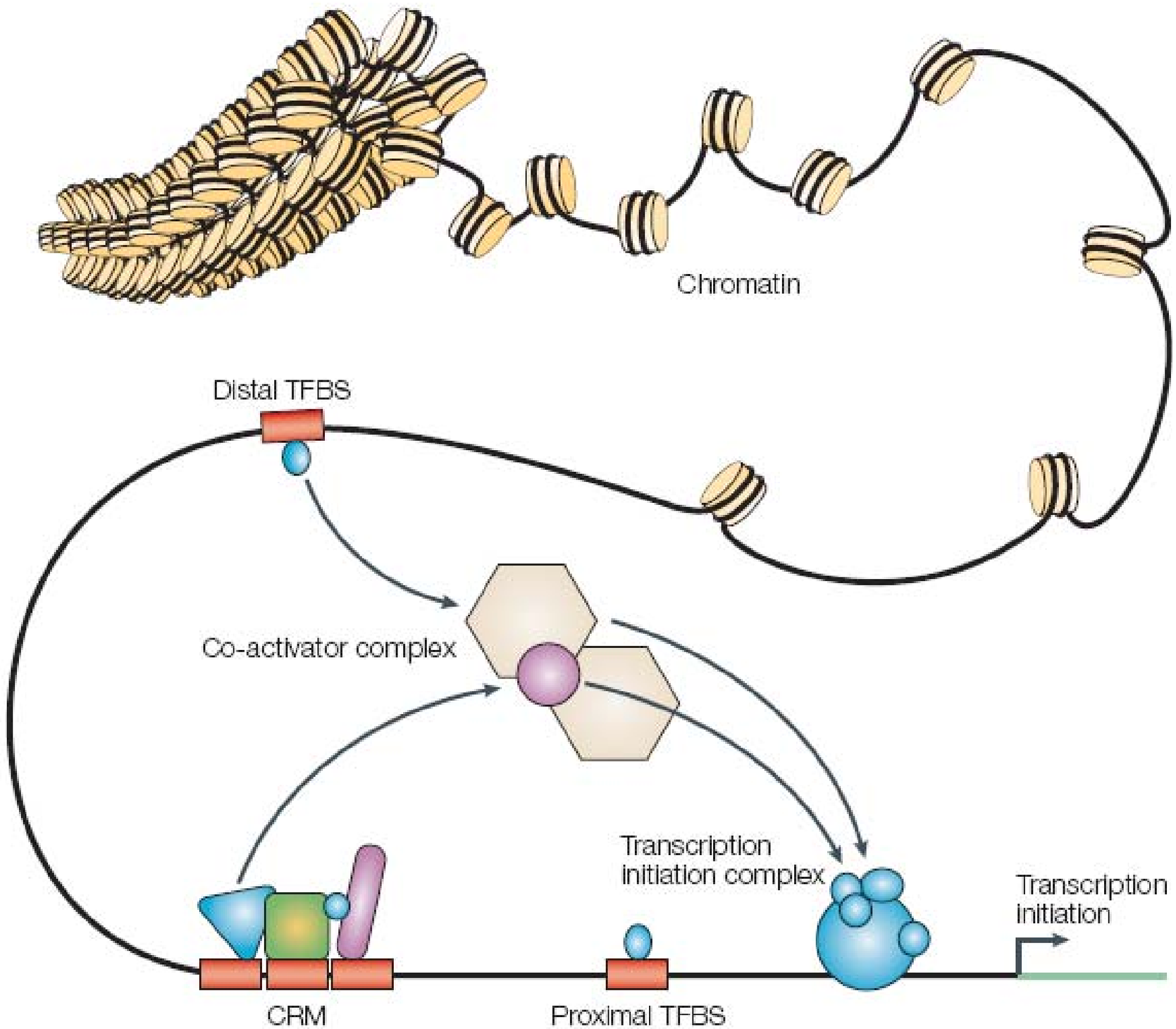
The transcription process: RNA Polymerase moves down DNA molecule creating RNA copy:

Identification of transcription promoter motifs



Initiation of this process:

Identification of transcription promoter motifs



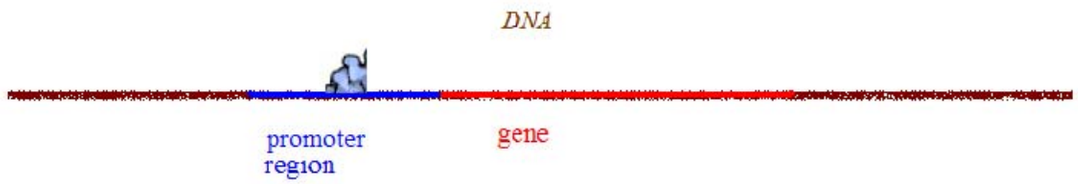
Identification of transcription promoter motifs

Figure: Biology of DNA transcription. CRM is *cis*-regulatory module, a molecular complex can co-activate transcription once a TF arrives.

We are dealing with the first step - the start of transcription

1. The Problem:

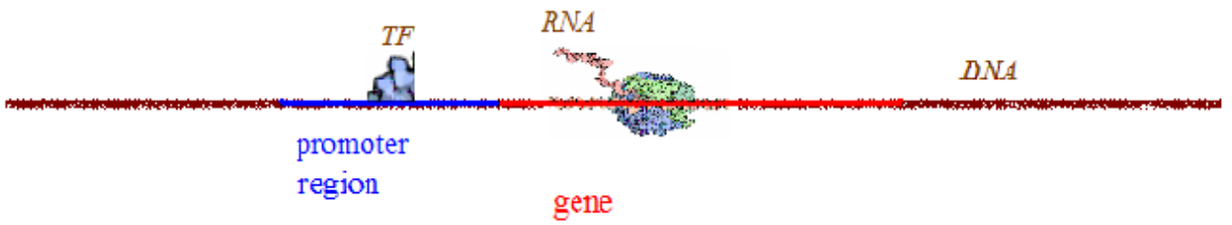
1. Transcription factor (or a combination of them) attaches to DNA



Identification of transcription promoter motifs

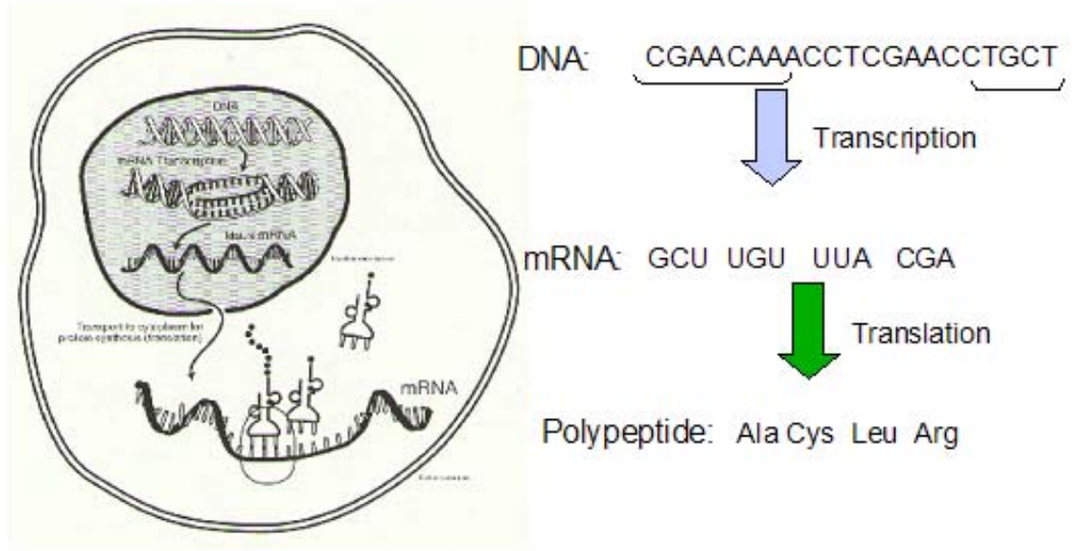
Fig. 1: Portion of chromosome string with TF attached to promoter region; this pattern repeated approximately 1000 times per chromosome

2. Signals RNA polymerase to start to copy DNA to RNA at a unique location nearby (*transcription*).



Identification of transcription promoter motifs

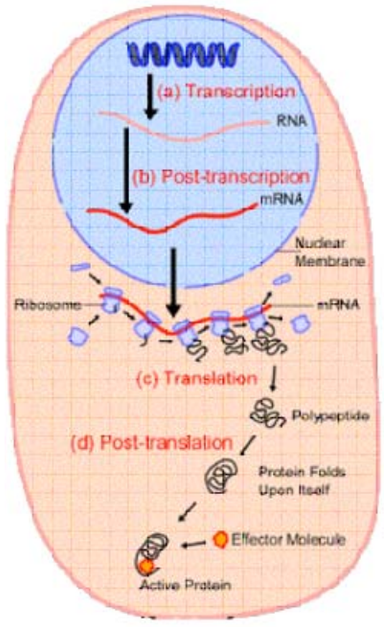
3. mRNA copy of DNA is transcribed and outside necleus to ribosome



4. tRNA (transfer RNA) matches amino acids to codons in mRNA. Each amino acid has own tRNA that binds only it.

Identification of transcription promoter motifs

5. Ribosome carries out construction of the protein in the exact sequence coded by the RNA.



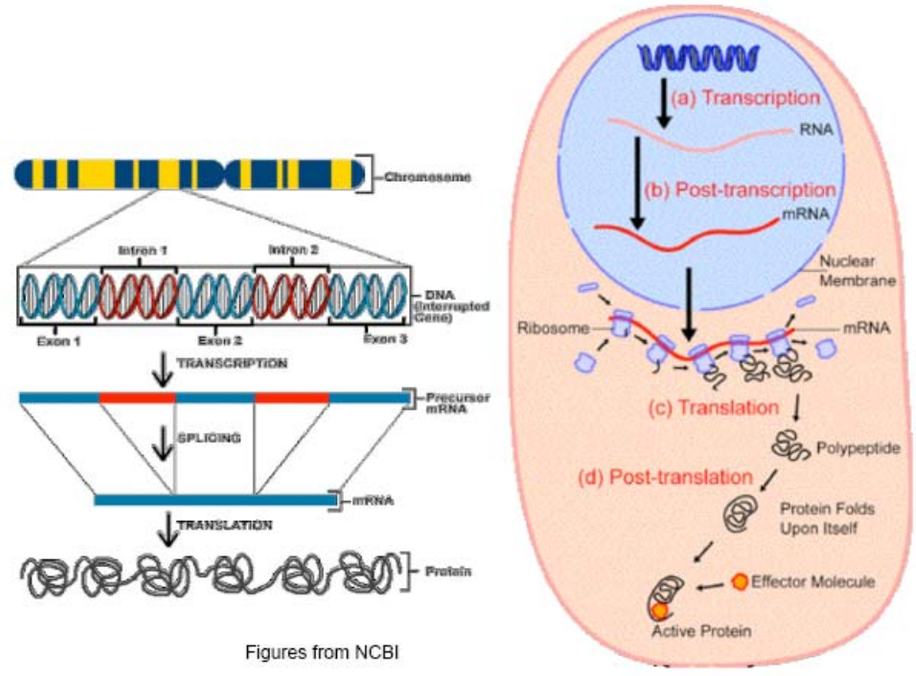
More succinctly:

$$\text{DNA} \xrightarrow{\text{Transcription}} \text{pre-mRNA} \xrightarrow{\text{Splicing}} \text{mRNA} \xrightarrow{\text{Translation}} \text{protein}$$

Note: translation involves transformation of

Identification of transcription promoter motifs

mRNA $\rightarrow$ tRNA $\rightarrow$ protein.



Figures from NCBI

Human: $\sim$ 25,000 genes $\rightarrow$ 25,000 proteins $\rightarrow$ determination of cellular and body function and structure

Identification of transcription promoter motifs

Goal:

Develop statistical and math methods for identifying DNA locations where transcription is initiated through attachment of transcription factors (TFs).

Important biologically, expensive experimentally.

Use statistical learning methods to identify new binding sites from examples based on DNA sequences of experimentally known ones.

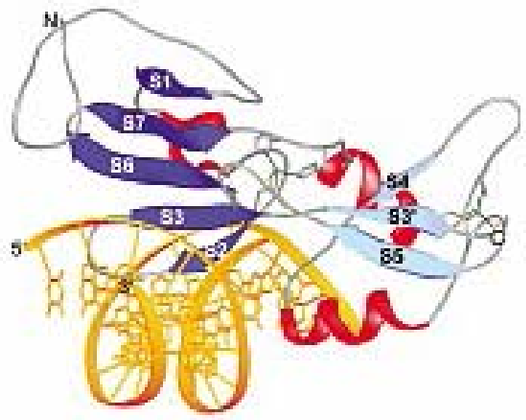
Identification of transcription promoter motifs

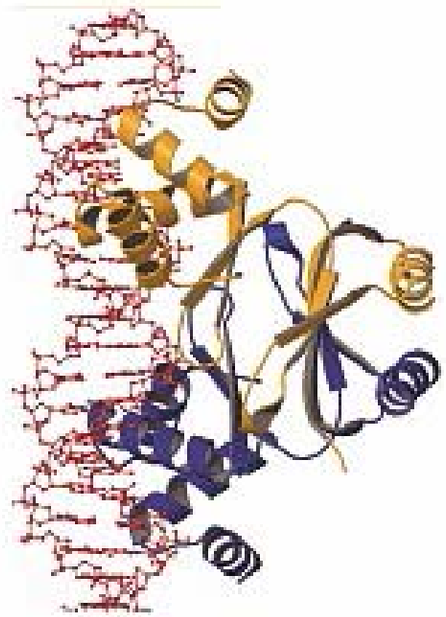
The problem:

Of the 25,000 human genes over all 23 chromosome pairs, find promoters $\mathbf{p}_i$ and the locations on them which which a specfic TF t attaches to.

- The mechanism of TF-gene binding is hard to solve chemically, because of its complexity;
- Means for explicitly solving bindings (fig. 1) computationally still far off

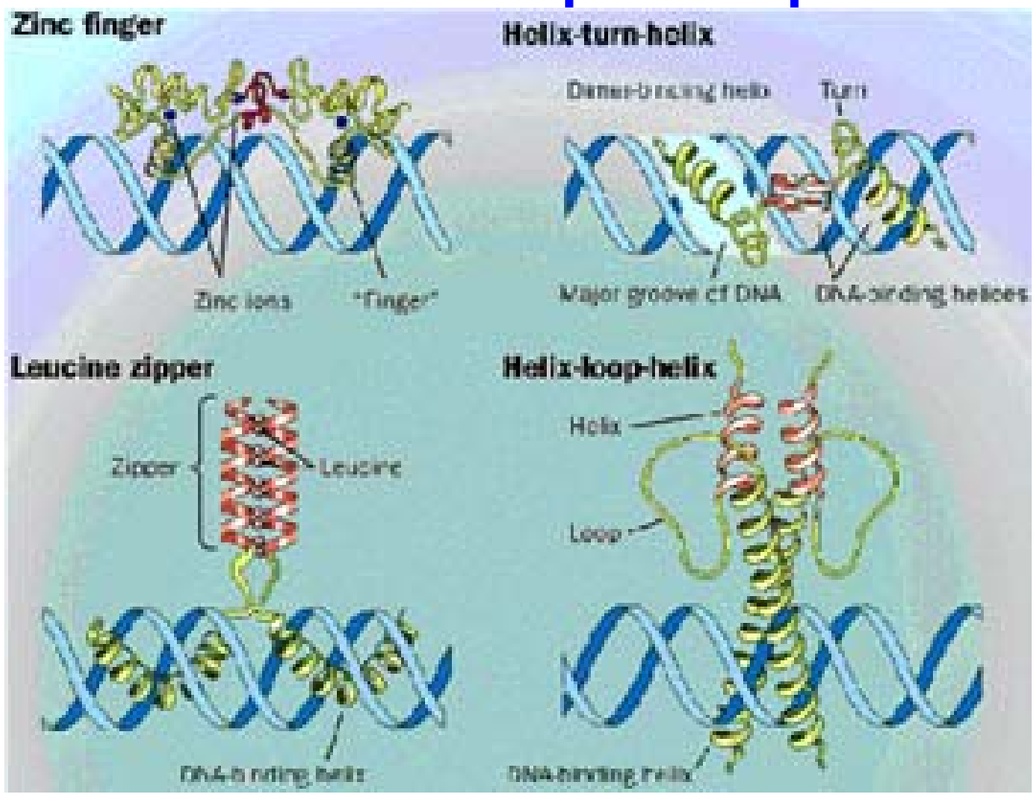
Identification of transcription promoter motifs





Left: DNA binding of GCM; right: binding of Fur (C. Ehmke, E. Pohl, EMBL, 2005)

Identification of transcription promoter motifs



Stanford University

Typical binding modes

- basic uncertainty principle trades off between gene-TF identification accuracies vs. throughputs.

Identification of transcription promoter motifs

More Detail: Promoters

1. **Promoter** is a region of DNA which attracts RNA polymerase for the initiation of transcription.

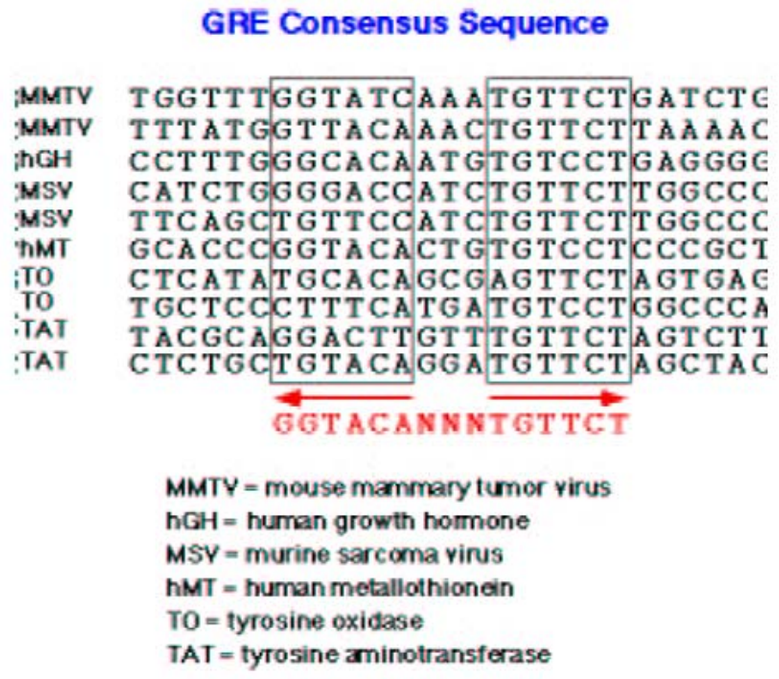
2. Promoter region of DNA contains *regulatory sequences* which attract transcription factors (TF's)

3. Regulatory sequences consist of inexactly repeating patterns (*motifs*).

4. Motifs are very similar across species - they attract specific transcription factors.

Identification of transcription promoter motifs

Regulatory sequences



Goals:

Identification of transcription promoter motifs

- Decrease cost and increase speed of TF-gene binding and binding site identification.
- Extend methodologies to other bioinformatic areas

- **(Modeling)** How should an integrated mathematical model of TF-gene interactions look?

We are extending toolboxes for gene-TF information, focusing on information integration.

- **(Applications)** How can such a model be used biologically to improve knowledge of systems and pathways?

- **(Ergonomics)** How to allow biologists to use such inferences?

Implications:

Gene-TF associations studied here can integrate information on biochemical pathways

Identification of transcription promoter motifs

using machine learning methods

Protocol for organized replacement of experimental methods with mathematical and statistical ones.

2. Approach

Problem: find a pattern occuring in vectors in a high-dimensional space S

Here

$$S = \text{space of} \left\{ \begin{array}{c} \text{DNA} \\ \text{RNA} \\ \text{Protein} \end{array} \right\} \text{sequences}$$

Introduction

Initial goal: discover a pattern-sensitive map

$$f : S \to \left\{ \begin{array}{c} \mathbb{B} \\ \mathbb{R} \end{array} \right\}$$

where:

$$\mathbb{B} = \{1, -1\}$$

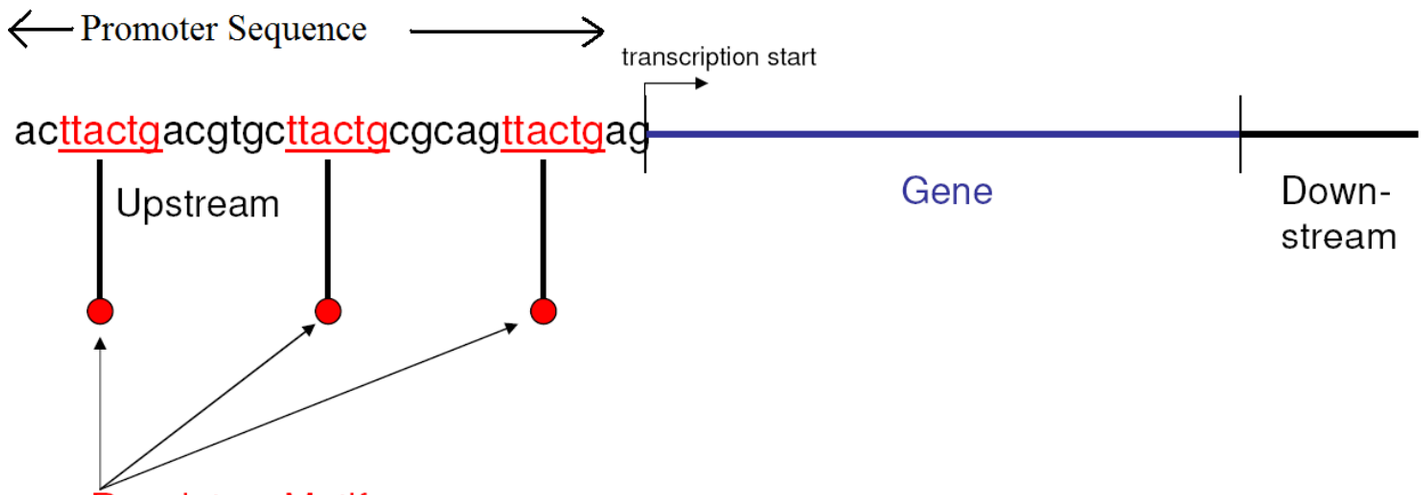
$$\mathbb{R} = \text{real numbers}$$

Ex: For fixed transcription factor (TF) t:

Given gene g, let

$$\mathbf{p} = \text{promoter sequence of gene} = \text{...acttact...}$$

Introduction



extrapolate

$$f(s) = \begin{cases} 1 & \text{if } g \text{ binds } t \\ -1 & \text{otherwise} \end{cases},$$

from examples to determine which genes g are targets of t.

Or: we may want to extrapolate unknown f defined by

Introduction

$$f(\mathbf{p}) = \text{location of binding sites}$$

(Binding sites occur as motifs of length $\sim$ 10 bases)

Such information traditionally experimental.

Can also be a classical learning problem -

Information about f:

Data = $D = \{s_i, y_i\}_{i=1}^{n}$

where

Introduction

$\mathbf{p}_i =$ promoter sequence of gene g_i

$$y_i = \begin{cases} +1 & \text{if } g_i \text{ binds with } t \\ -1 & \text{if not} \end{cases}$$

Extrapolate

$$f(\mathbf{p}) : S \to \mathbb{B}.$$

We seek new learning methods for such gene classification and transcription factor binding site (TFBS) identification.

Thus problem is:

1. Given fixed TF t and experimentally known gene set $\{g_i\}_i$ which it targets (i.e., to whose promoters p_i it attaches).

Introduction

Can we extrapolate (e.g., through learning or Gibbs sampling) to computationally find *new* targets g?

2. Given a known set of target genes $\{g_i\}_i$, can we determine the *binding sites* on promoters p_i? Typically 10-mers (10-strings) with little variation, known as *motifs*.

Here we consider problem 2

For a given TF t, our goal here is to find genes it binds and its binding site motifs.

Equivalently: Given a known gene set $\{g_i\}_i$ with promoter sequences $\{\mathbf{p}_i\}_i$ known to bind t, what is an overrepresented set of 10-mers (DNA subsequences of length 10) in the set $\{\mathbf{p}_i\}$?

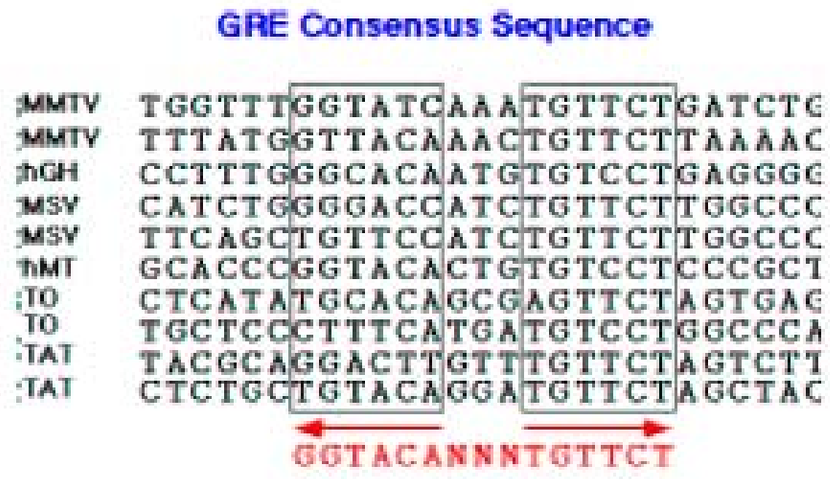
Rationale: promoters generally have multiple copies of a motif if the are functional.

Introduction

This 'overrepresented set' of 10-mers should contain *binding motif* of t.

Introduction

Regulatory sequences



MMTV = mouse mammary tumor virus
hGH = human growth hormone
MSV = murine sarcoma virus
hMT = human metallothionein
TO = tyrosine oxidase
TAT = tyrosine aminotransferase

Introduction

Position weight matrix (PWM):

$$
\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}
\begin{array}{c}
\text{Position number} \\
\begin{bmatrix}
0 & 0 & & 0 \\
.1 & 0 & & 0 \\
.6 & .8 & \dots & 0 \\
.3 & .2 & & 1
\end{bmatrix}
\end{array}
$$

represent preferred binding motif for t.

Introduction

Typical test data: often more than 1 motif per sequence:

J. Liu

Q: How to find overrepresented 10-mers?

Introduction

Current algorithms

- Largely optimization, taking genes known to bind specific TF, and identifying DNA subsequences (sequential strings of base pairs typically of length 5 to 15) overrepresented in gene promoters.

- Typically use Gibbs sampling, with objective of maximal alignment of sequences in known binding genes.

- Machine learning methods for this are new

These include: both types of TF problems described above (identifying genes and gene locations)

Have capacity for processing high dimensional information-

Introduction

Introduction

We consider:

- Support Vector Machine (SVM)
- Random forests (RF)
- SVM Ensembles (SVME)

All based on maps into feature spaces.

Plan: head-to-head comparison of substring-based kernel methods with standard motif algorithms.

- BioProspector
- AlignACE
- MDSCAN

Claim: SVME and RF can find subtle motifs (binding DNA patterns) in humans (challenging task).

Introduction

Important aspect of both Gibbs and learning methods: ranking of promoter substrings by statistical correlation with genes whose promoters bind to given TF t.

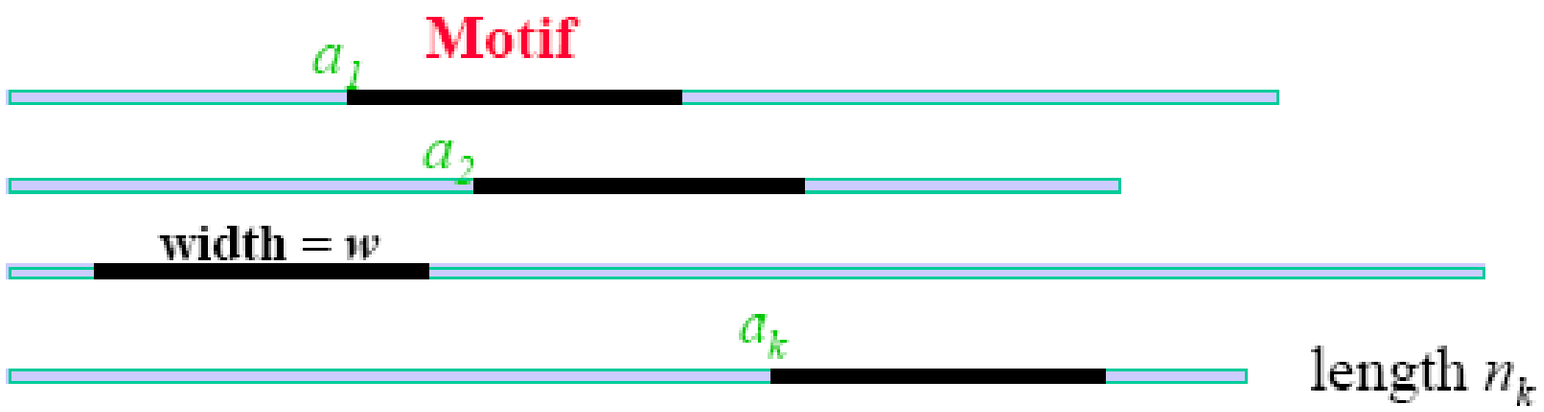
Introduction

Gibbs Sampling:

Optimally align all $\mathbf{p}_i$ which bind t (previous diagram).

i.e., score alignments of all (split) 12-mers (in this case) $\mathbf{p}_i$ and find alignment with highest score with Gibbs sampling and simulated annealing [Lawrence, Liu, 1993].

Typical pre-alignment:



Learning appoaches

3. Learning approaches

Important advantage of learning methods for TF binding site location:

Can use negative examples as well as positive examples, i.e., promoters which do and do not bind t.

Machine learning approach:

In species $\mathcal{S}$ for fixed t, obtain set of gene promoters

$$\mathbf{p}_1, \ldots, \mathbf{p}_n$$

likely to bind t (experimental information, etc.). Note for gene g_i we have promoter

$$\mathbf{p}_i = \mathbf{p}(g_i)$$

is the promoter sequence of gene g_i.

Learning appoaches

In addition, find negatives, i.e., promoters $\mathbf{p}_{n+1}, \ldots, \mathbf{p}_{n+m}$ to which t probably does not bind.

Use as *learning examples* for extrapolating

$$f : \mathbf{P} \to \mathbb{B} = \{-1, 1\}$$

on the space **P** of promoters **p** with

$$f(\mathbf{p}) = \begin{cases} 1 & \text{if } t \text{ binds } \mathbf{p} \\ -1 & \text{otherwise} \end{cases}.$$

Key in learning methods:

Feature map Φ from genes g into feaures $\mathbf{x} \in F =$ feature space :

$$\Phi(g) = \Phi(\mathbf{p}(g)) = \mathbf{x} \in F$$

Learning appoaches

$\mathbf{x} = \Phi(g)$ gives relevant information now any learning algorithm can be applied.

Feature space and learning methods now common in computational biology:

- protein analysis [Leslie, et al.]
- TF binding prediction [Holloway, Kon, et al.]
- Motif finding [Vert, et al.].

Dogma of learning approaches for TF binding:

1. Choose good feature space F with features $\mathbf{x}(g) = (x_1, \ldots, x_l)$ depending on promoter of gene g.

2. Re-define desired output

Learning appoaches

$$f(g) \rightarrow f(\mathbf{x}(g)) = f(\mathbf{x})$$

so it is defined on F

3. Learn f from training data

$$D = \{(\mathbf{x}_i, y_i)\}_i.$$

How to learn $f : F \rightarrow \mathbb{B}$

$$f(\mathbf{x}) = \begin{cases} 1 & \text{if } g \text{ corresponding to } \mathbf{x} \text{ is a target} \\ -1 & \text{otherwise} \end{cases}$$

from examples?

Note without feature space, f would map promoter **p** to $y \in \mathbb{B}$, so

$$f\colon \mathcal{A}^{1000} \rightarrow \mathbb{B},$$

Learning appoaches

where $\mathcal{A} = \{A, G, C, T\}$.

Remark: The structure is wrong on this space; difficult to guess f from examples $\mathcal{D}$ (telephone directory problem).

String map

4. The string feature map

Given list

$string_1$	AAAAAA
$string_2$	AAAAAC
$string_3$	AAAAAG
$string_4$	AAAAAT
$string_5$	AAAACA
⋮	⋮

String map

of strings of 6 base pairs. Recall **P** is the space of promoter DNA sequences.

Consider feature map **x** : **P** $\to F$, with **x**(**p**) = **x** $\in F$, with components

$$x_i = \#\text{ appearances of string } s_i \text{ in promoter } \mathbf{p}.$$

Then **x** $\in F$ has $4^6 = 4{,}096$ components $\Rightarrow$ $F = \mathbb{R}^{4{,}096}$.

Transferring f from **p** to **x** = **x**(**p**); now

$$f : F \to \mathbb{B}.$$

Thus: f maps sequence **x** of string counts in F to yes/no in $\mathcal{B}$.

With data set

String map

$$D = \{\mathbf{x}_i, y_i\},$$

seek function $f : F \to \mathbb{B}$ which generalizes D.

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

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

SVM approach

5. SVM approach

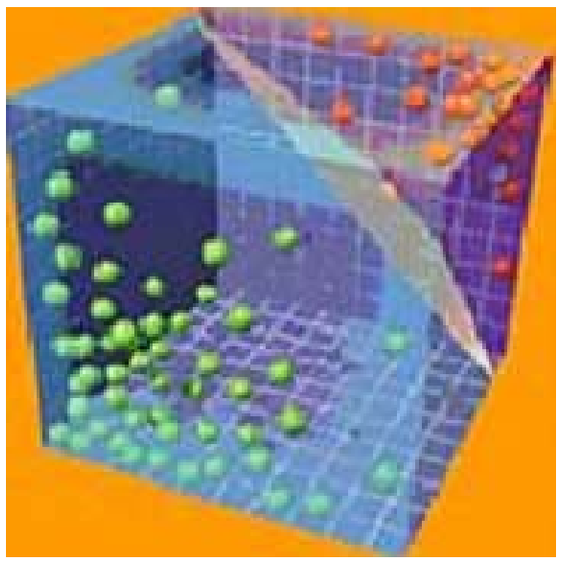
Assume SVM kernel (similarity measure) $K(\mathbf{x}, \mathbf{y})$ for string feature vectors $\mathbf{x}, \mathbf{y} \in F$

Seek

$$f : F \to \mathbb{R}$$

with $f = f(\mathbf{x})$ which predicts binding to gene g with $\mathbf{x} = \mathbf{x}(g)$.

$f > 0$ and $f < 0$ cases separated by hyperplane $H : f = 0$ in F:



SVM approach

Geometry of F encoded in $K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$ (nonlinear dot product).

Obtain

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

For linear dot product $K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$:

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

Prior SVM work

6. Prior SVM gene classification work

Spectrum (string) kernels - Vert, Noble, et al.:

$F =$ feature space of k-mer (k-string) counts ($k = 5$).

$$\Phi^{\text{spect}}(\mathbf{x}) = \text{vector of length } 4^5$$

with i^{th} position $\Phi_i(\mathbf{x}) =$ count of the i^{th} k-mer.

Conservation information: their feature vectors take phylogenetic conservation (string conservation across species) into account.

Specifically: given promoter $\mathbf{p} \in \mathcal{A}^n$ in *S. cerevisiae*, consider alignment
$\mathbf{c} \in (\mathcal{A}^5)^n =$ aligned array of five (matching) promoter regions of length n from 5 related yeast species.

Prior SVM work

Have 'marginalized' feature map

$$\Phi_i^{\text{marg}}(\mathbf{c}) = \sum_{\mathbf{h}} \Phi_i^{\text{spect}}(\mathbf{h}) p(\mathbf{h}|\mathbf{c}) = E_{\mathbf{h}}\left(\Phi_i^{\text{spect}}(\mathbf{h})\right)$$

$=$ expected value of the spectrum kernel over possible common anscestoral sequences **h** of the set.

Probability distribution of **h** conditioned on **c** obtained with phylogenetic model from [Tsuda].

Effect: reinforces k-mers consistent across related species (more likely to be functional);

More direct way: first determine vector $\mathbf{d}(g)$ which labels every site of promoter $\mathbf{p}(g)$ by whether it is likely functional or not based on conservation among species.

Prior SVM work

Specifically

$$d_i(g) = \begin{cases} 1 & \text{if site } i \text{ is conserved among 5 yeast species} \\ 0 & \text{otherwise} \end{cases}.$$

Discretizing above leaves room for errors, but resulting noise reduction is useful.

Now [Vert] replace spectrum (k-mer) kernel by the 'relevant' kernel corresponding to feature map

$$\Phi_i^{\text{rel}}(\mathbf{h}, \mathbf{d}) = \text{count of 'relevant' occurrences of } k\text{-mer } i$$

$$= \#\,\text{occurrences at sites all of whose locations are relevant.}$$

for any string **h**.

SVM Classifiers

7. SVM-based classifiers for targets of t

Before we locate motifs, we will classify genes (wrt binding/nonbinding by t).

Given fixed TF t:

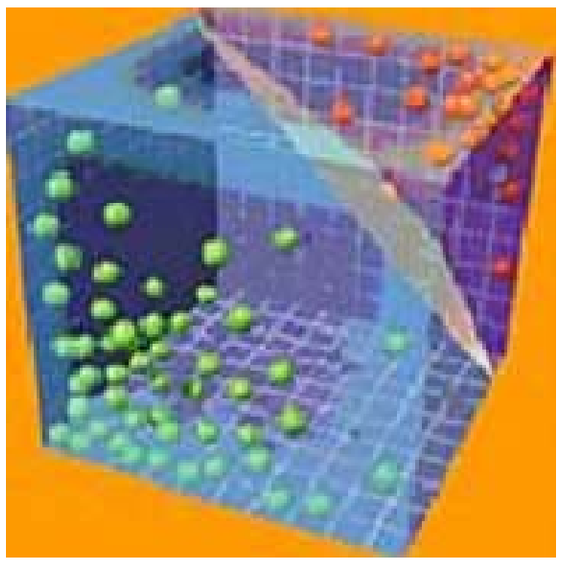
$$f : \mathbf{X} \to \mathbb{R}$$

$$\mathbf{x} = \mathbf{x}(g) = \text{feature vector of gene } g$$

$$\mathbf{x} \in F = \text{feature space}$$

SVM Classifiers

$y = 1$ and $y = -1$ cases separated by a hyperplane $H : f(\mathbf{x}) = 0$:



Recall for linear $K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y}$,

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = \begin{cases} > 0 & \text{if } y = 1 \\ < 0 & \text{if } y = -1 \end{cases}.$$

SVM Classifiers

Procedure:

1. **Train SVM**

$$D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n+m}$$

[use n positive $\mathbf{x}_i$, m negative $\mathbf{x}_i$].

2. **Find optimal $f = \mathbf{w} \cdot \mathbf{x} + b$**

f predicts new genes g which are targets of t

SVM classifier applications

Ex: Human TF WP1:

From 14 known positive genes have extrapolated much larger number of potential new targets for investigation, and have some new implications for pathways.
New high reliability targets of the WT1 are genes

RNH1, *IGF2AS*, *CD151*

Relation to Wilms' tumor:

WT1 involved in Wilms' Tumor (8% childhood cancers).

Genes in significant loci include oncogenes and tumor suppressors -- candidates for involvement in cancer progression

May explain some observed clinical and biochemical data.

SVM classifier applications

Example: chromosomal region 11p15.5 (known in Wilms' Tumor)
New targets for WT1 ($p \leq .005$) here are tumor suppressors
RNH1, *IGF2AS*, and *CD151*.

Other regions involved in Wilms' Tumor have new target predictions:

16q, 1p36.3, 16p13.3, 17q25, and 4p16.3.

Potential binding sites can be extracted using standard motif finding algorithms (e.g., Weeder)

SVM classifier applications

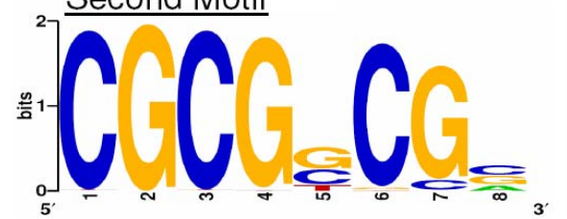
Potential Binding sites for WT1

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

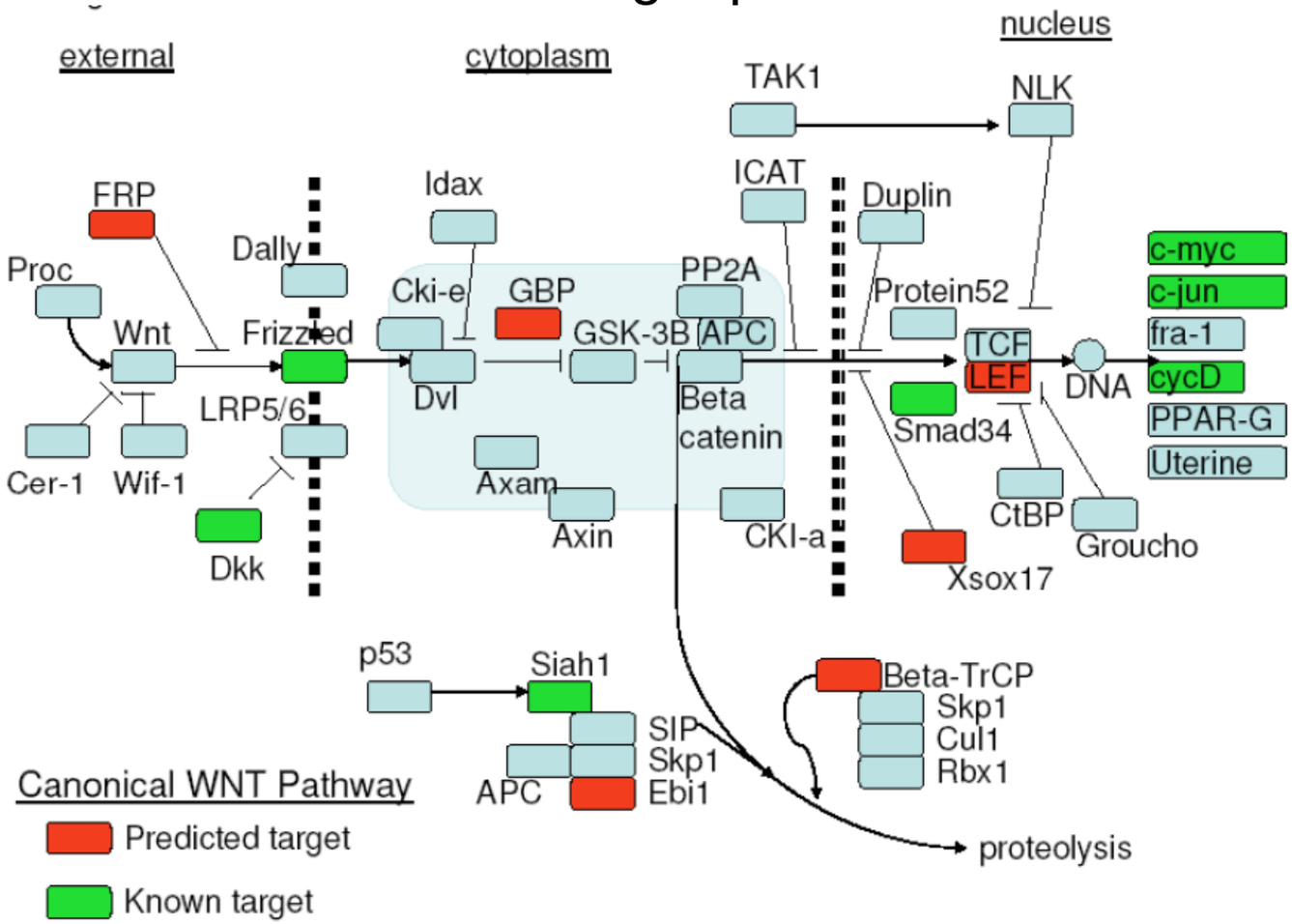
C Weeder Top Motif

Second Motif



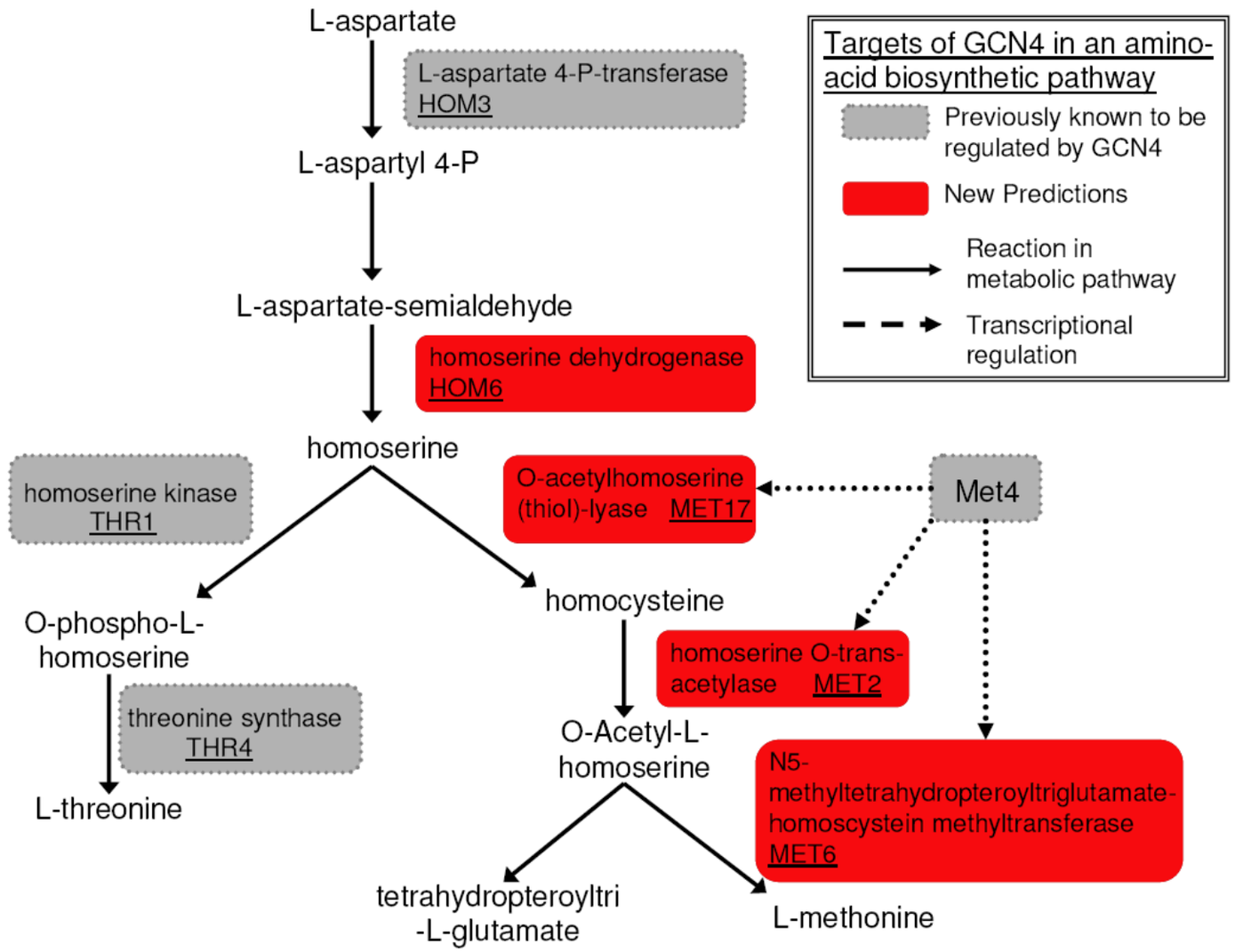
SVM classifier applications

Example: Human TF Oct4 - new target predictions fit into WNT pathway:



Example: Yeast TF's:

SVM classifier applications



Finding binding motifs

8. Finding binding site motifs

Feature space $F =$ string counts:

$$\Phi(\mathbf{p}) = \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_j \\ \vdots \\ x_q \end{bmatrix} \begin{matrix} AAAAAA \\ \\ ATGCTG \\ \\ GCCGTA \\ \\ TTTTTT \end{matrix}$$

$x_i =$ count of 6-mer i

Finding binding motifs

Better choice:

$$x_i = \text{count of } 6\text{-mer } i \text{ weighted by conservation}$$

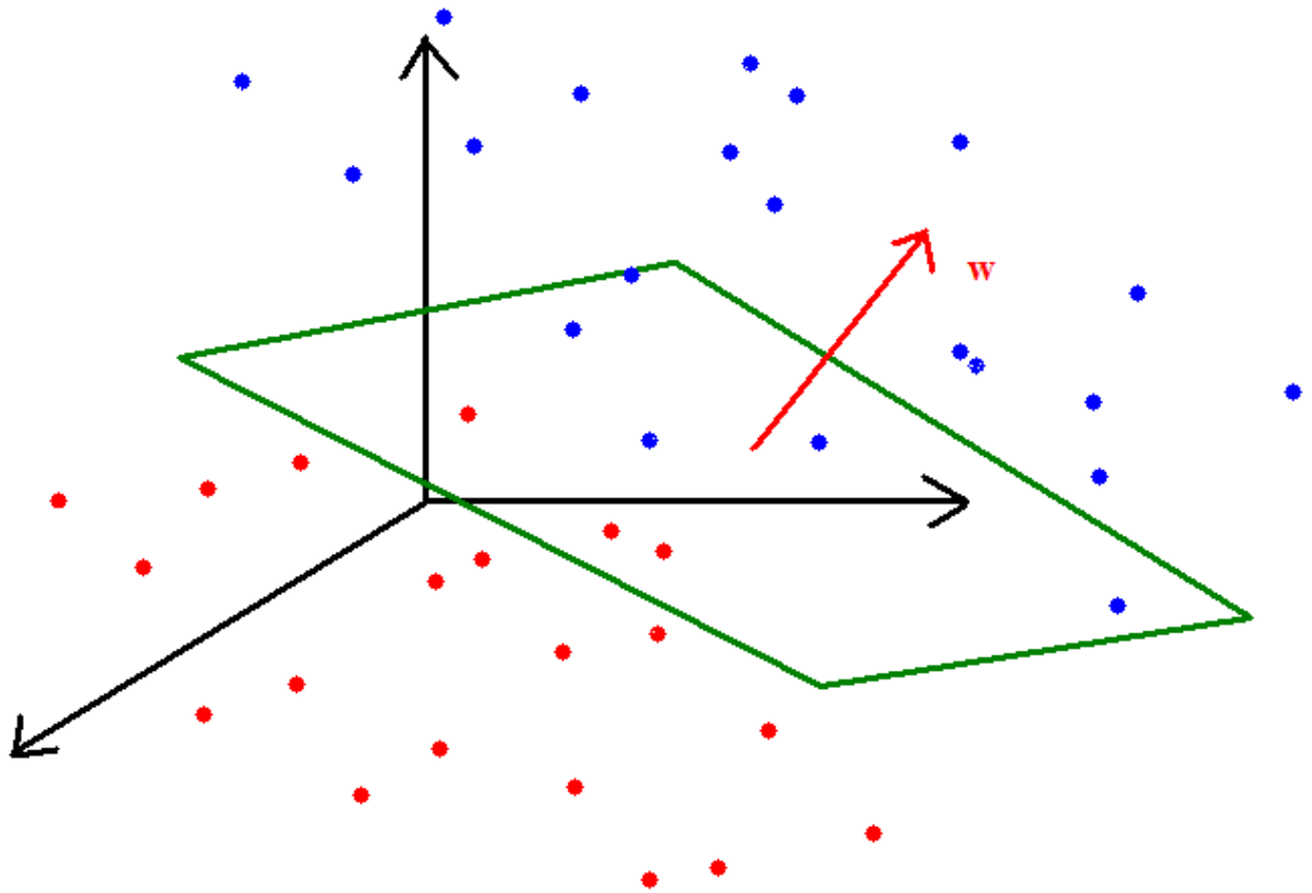
i.e., 6-mer M is weighted by $c \in [0, 1]$ which determines level of conservation of average position in M in closely related species.

How to find binding sites?

Which 6-mers x_i best differentiate $+$ and $-$?

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

Finding binding motifs



w $=$ optimized gradient vector between $y > 0$ and $y < 0$ cases

Finding binding motifs

Use

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \nabla f.$$

Largest components give primary direction of gradient; these are 6-mers which best differentiate $y > 0$ and $y < 0$.

Use SVM-RFE (recursive feature elimination) to iteratively reduce **w** to most important. Typically cut number of features each time and re-calculate **w** recursively.

For yeast: approx. 50-300 positive examples per gene.

Sample selection repeated 20 times with different choices of negatives (genes with high p values in yeast ChIP-chip experiments) out of 600 available.

Finding binding motifs

In each SVM run numbers of positives and negatives are equal.

Reduce from $\sim$ 4000 to 150 to 50 features x_i.

1. a c t g t g
2. g t ca c t
3. t g a c t a

Best clustering:

a c t g t g

g t c a c t

t g a c t a

Finding binding motifs

Now: 'unrelated' 6-mer:

t c t t t a

$\longrightarrow$ start new cluster.

Result: Typically obtain $\sim$ 4 significant clusters, each with a probability weight materix (PWM) representing probabilitity distribution of bases in each cluster position.

Finding binding motifs

GCN4 binding sites

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

TGAsTCa

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

Formation of PWM

Finding binding motifs

Clusters scored using combination of

entropy scores = how atypical the string is and

hitting scores = counts of above-threshold matches to PWM along candidate promoter $\mathbf{p}_i$.

Precisely hitting ratio score for a PWM is defined as

$$\text{HR} = \frac{\text{\# hits on positive genes}}{\text{\# hits on negative genes}}$$

in a pair of positive and negative gene samples of same size.

Finding binding motifs

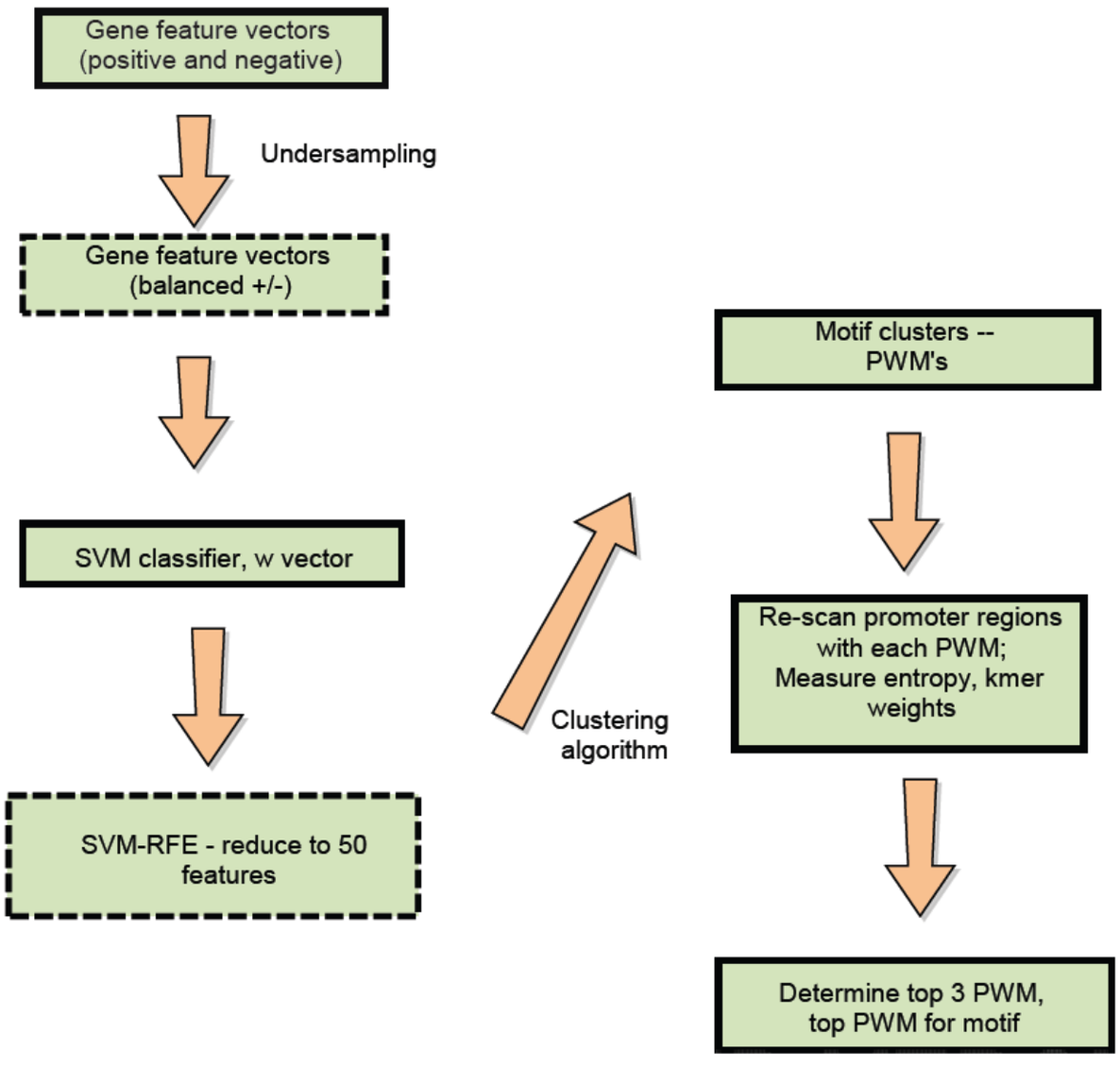
Clustering algorithm: Assume C is current set of clusters

Initial step: $S = \{s_1, s_2, \ldots, s_n\}$ = string (k-mer) set to be aligned, ordered by weight in **w**.
C = empty.
Step **0**: Form a new cluster from s_1. Delete s_1 from S.
Step **1**: If S = empty, then quit. Otherwise, pick out the string $x \in S$ with the highest weight.
Step **2**: Compute the scores of the string x from Step 1 w.r.t. each of the current clusters in C. If the highest score is greater than the addition threshold (threshold1 in the code), go to step 3 (addition step). If the highest score is less than the new cluster threshold (threshold2 in the code), go to *step* **4**. Otherwise, go to *step* **5**
Step **3** (addition step): Add string x into the cluster producing the highest score, and delete x from S. Let $S = S \cup E$. Go to *step* **3'**.
Step **3'** (deletion step): Examine each element in the cluster being updated in step 3 by computing the score of this element w.r.t. the PWM of this cluster. If the score is smaller than the deleting threshold, move this string back into S.
Step **4**: Form a new cluster in C from string x, and delete string x from S. Let $S = S \cup E$. Go to step 1.
Step **5**: Move string x into the Exceptional set E. Go to step 1.

Clustering algorithm

Finding binding motifs

Overall strategy:



Finding binding motifs

Typical results for 4 TF's:

Name	#pos	YeastGenome (Transfac)	BioProspector	SVM
YBR049C	186	CGGGTRR TTACCCG	CGGGTAA TACCCGG	AAGAAGARG CYTCTTCTT GGCGGGTAAC
YDL020C	134	GGTGGCAAA	GGCGGGTAA GTTTCCCCG GTTTCCCCG	CCGGTGGCRG TSGCCACCSG AAGAAGAGG
YDL056W	207	ACGCGT	TACATA GCGACT GGTTGG	ARACGCGTYT GYYTTCTTS SAAGAARRC
YDL106C	69	SGTGCGSYGYG	ATCCTCGAGTT GACTCACAATC GCACTTACAAC	CSCCACGTGGG CCGCTGCAGCG CCCGGG

Table: A sample of yeast transcription factors analyzed.
pos is number of positive examples for TF. Selected motifs to the right are in order of priority.

Finding binding motifs

Example: Results in yeast -

	BioProspector		SVM	
	Top	1 to 3	1	1 to 3
Ungapped (34)	15	16	16	24
Gapped (10)	4	5	3	5

Table: Motif performance on YeastGenome TF's (Transfac + random) between BioProspector and SVM.

'Top' $=$ # hits of top motif PWM in yeastgenome
'1 to 3' $=$ hits in top 3 PWM

Remark: We note there are cross-validated multiple clusters (motifs) correlated with a given TF t.

Finding binding motifs

Biological interpretation: Note there is only one motif (with small variations) per TF.

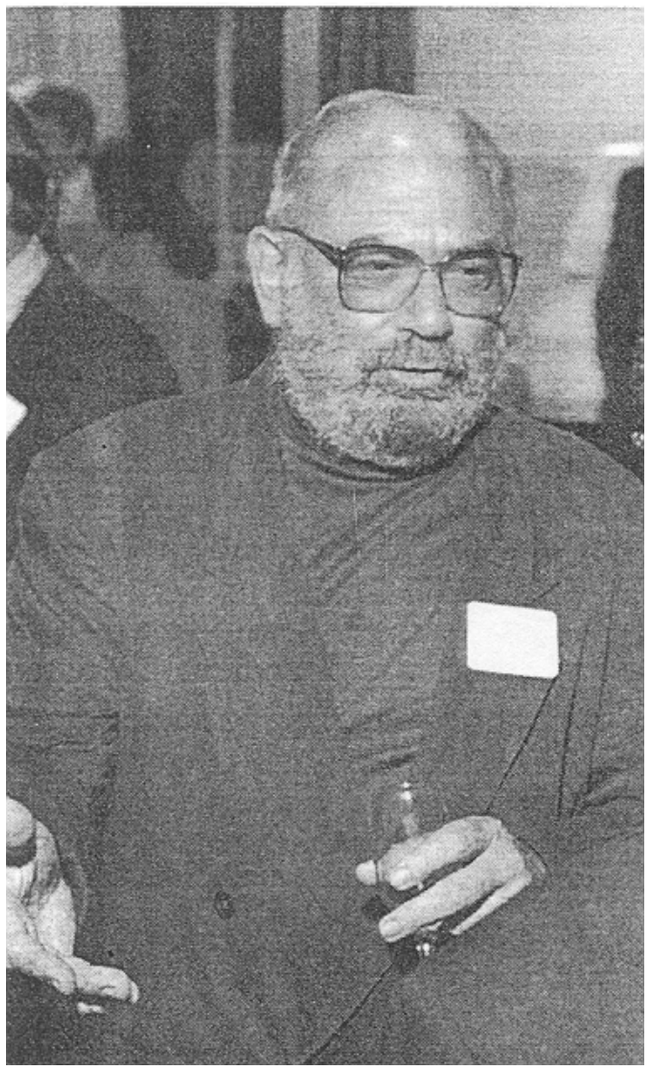
Additional motifs must represent targets of other TF's in the same gene 'transcription module', i.e., cascade of TF activations and resulting new TF productions.

Thus a multiple set $t_1, \ldots, t_r$ of TF's involved in a single transcription module. Genes in this module are activated in a coordinated way by the TF's.

Result: Confounding of motif finding - positive genes often share more than one TF.

Random forests

9. Random Forests



Random forests

Portraits of Statisticians

Same learning approach - use feature space

$$F = \text{string count space.}$$

Again F high-dimensional (around 4,000 dimensions if use 4-, 5-, and 6-mer counts with pruning).

Note: It is not necessary or useful to use *all* k-mers. Some just confound the counts, e.g., simple but highly repetetive k-mers such as $AAAAAA$. Thus pruning of k-mers is appropriate.

Best to remove these initially.

Random forests

Have:

$$F \ni \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}.$$

Goal: again find significant features x_i differentiating $+$ and $-$ examples.

Strategy: As usual - first form a classifier which predicts $+$ and $-$; then find what variables (strings) it actually uses and make the largest difference.

Such strings are motif candidates.

Here: form $k \approx \sqrt{d}$ decision trees to form a random forest:

Random forests

Forest consists of decision trees $T_1, \ldots, T_k$.
Train trees on bootstrap samples from the dataset

$$D = \{\mathbf{x}_i, y_i\}_i$$

Then provide new feature vector **x**.

Random forests

Classification

$$f_1(\mathbf{x}) \approx y \in \mathbb{B}$$

determined by a vote of the k trees (bagging classifier).

Advantages: accurate, easy to use (Breiman software), fast, robust

Disadvantages: difficult to interpret

How to combine results?

RF is a bagging algorithm: Take a vote of trees: majority rules

Random forests

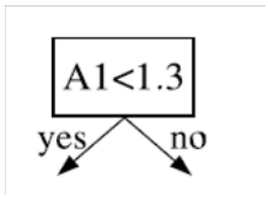
General features:

If original feature vector **x** $\in \mathbb{R}^d$ has d features $A_1, \ldots, A_d$, forming feature space F :

♦ Random selection of $m \approx \sqrt{d}$ features $\{A_{i_j}\}_{j=1}^m$ made from all features A_1, $A_2, \ldots, A_d$; the associated feature space is F_k, $1 \leq k \leq K$.

(Often $K = \#$ trees is large; e.g., $K = 500$).

♦ For each split in a tree based on a given variable, choose the variable by information content, e.g.,



RF: information content

Information content for the above node N :

$$I(N) = |S|H(S) - |S_L|H(S_L) - |S_R|H(S_R),$$

where

$|S| =$ input sample size; $\quad |S_{L,R}| =$ size of L, R subclasses of S

$$H(S) = \text{Shannon entropy of } S = -\sum_{i=\pm 1} p_i \log_2 p_i$$

with

$$p_i = \widehat{P}(C_i|S) = \text{proportion of class } C_i \text{ in sample } S.$$

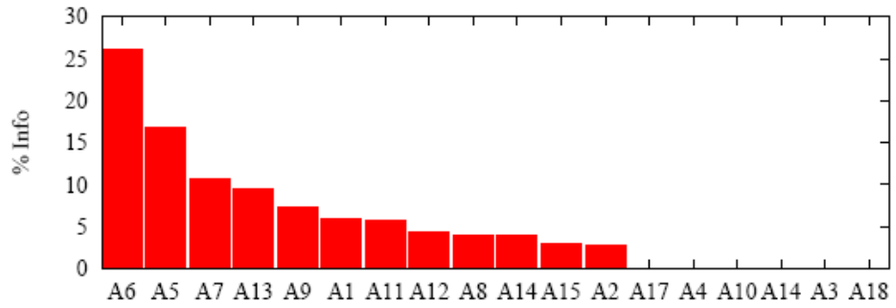
[can also split according to *Gini index*, another criterion]

Thus $H(S) =$ "variablity" or "lack of full information" in the probabilities p_i.

RF: information content

$I(N) =$ "information from node N".

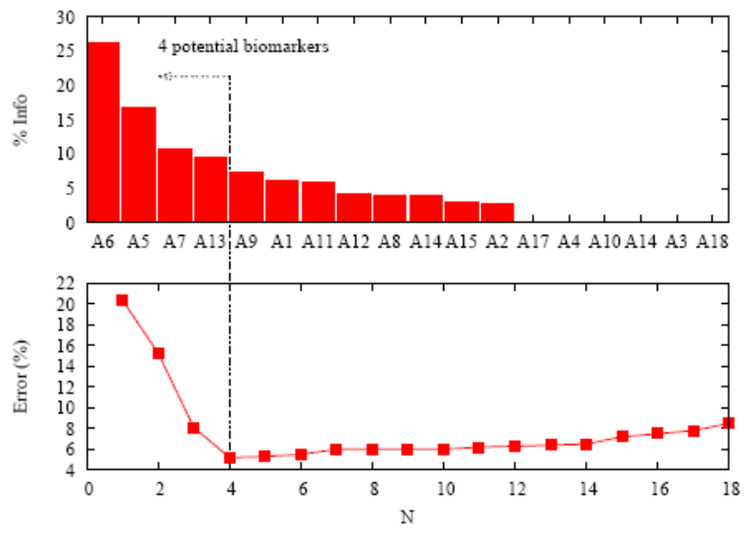
For each variable A_i, average over all splits in all trees involving this variable to find average information content $H_{\text{av}}(S)$; use this to determine value of A_i.



(a) Rank all variables A_i according to information content

RF: information content

(b) For each $n_1 < n$ use only the first n_1 variables. Select n_1 which minimizes error:



Geurts, et al.

RF: importance scores

Now use cross-validation to independently determine importances of variables:

Use RF variable importance score:

Each tree T_i has 1/3 variables left 'out of bag' (i.e., unused in bootstrap training sample).

Use out of bag variables (different group for each tree) to test variable importance independently:

Add noise to each variable and check decrement in classification accuracy.

The 'cross-validation' aspect of this sampling gives much better accuarcy:

RF: importance scores

Table - motif recognition for 26 Transfac motifs in MacIsaac database

	Top 1	Top 3
RF	13	14
BioProspector	11	13

Total: 26 TF's

Data: K. MacIsaac, T. Wang, D. B. Gordon, D. Gifford, G. Stormo, and E. Fraenkel, "An improved map of conserved regulatory sites for Saccharomyces erevisiae," BMC Bioinformatics, 2006

SVM Ensembles

10. SVM Ensemble (SVME) - a new bagging algorithm

Ensemble (forest) of SVM - same bagging principle and variable imporance sampling principles.

Performance: Similar to RF but with differing strengths in different sample elements.

Thus complementation of RF by SVME can be useful.

Here on same 26 TF's:

	Top 1	Top 3
RF	13	14
SVME	10	15
SVME - Poisson Normalization	15	16
BioProspector	11	13

Poisson normalization in addition to SVME works best.

SVM Ensembles

Poisson normalization: normalize k-mer counts x_i to lie between 0 and 1 by composing with cumulative distribution function of Poisson distribution.

Based on Premise: There will be more uniform variation (in fact close to a uniform distribution) on $[0, 1]$ for the normalized k-mer counts, assuming original counts are Poisson.

Further work

11. Further TF-gene binding work

As mentioned, functional components of DNA (e.g., TF binding sites) are conserved among closely related species. Combination of conservation information into machine learning methods is planned.

Plan also to include physical chemistry information on the promoters, including numbers for promoter twisting, curvature, and melting temperature, all of which are correlated with motif locations.

Further work

Reference:

Portraits of Statisticians: http://www.york.ac.uk/depts/maths/histstat/people/welcome.htm