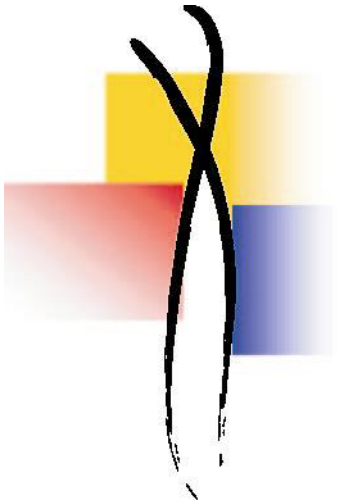
Discussion, Software Demos and the Details



Analysis of regulatory sequences

Wyeth Wasserman

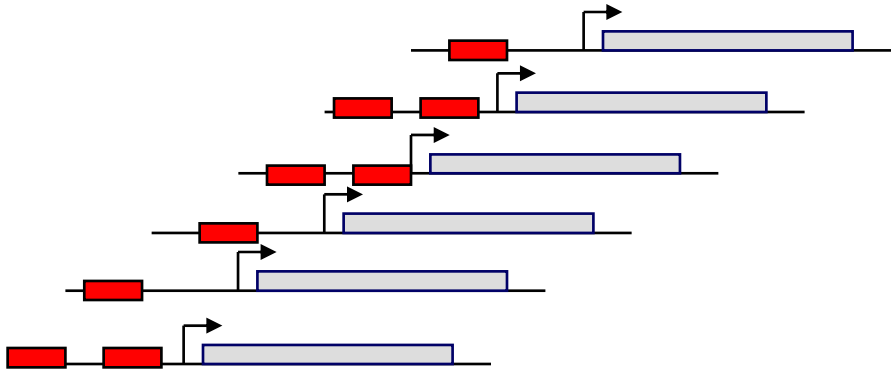
Regulatory regions problem space

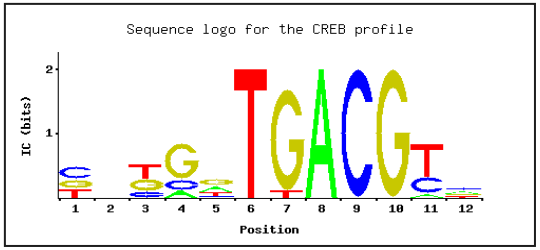
Sets of binding sites

AATCACCA
AATCACCA
AATCACCA
AATCACCA
AATCTCCC
AATCTCCG
AATCACAC
AATCATCA
AATCTCAC
AATCTCTG
AGTCCCCA
AATCCCGG
AATCTGAG
AATCCATA
ATTCAGCC
AATAACTT
GATAACCT
AATTAGAC
GATTACAG
GATTAGCG
ATTCTTCC
TATGAACA
GATTAAAA
AGACCCCA

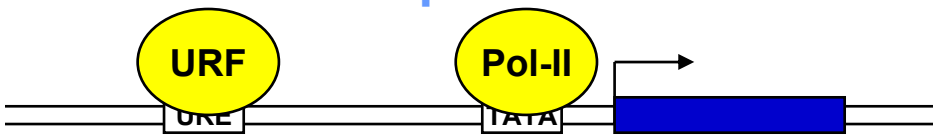
Specificity profiles for binding sites

```
A  [   -2     0    -2 -0.415  0.585    -2    -2 2.088    -2    -2    -1  0.585 ]
C  [    1 0.585     0      0    -1    -2    -2    -2 2.088    -2 0.585  0.807 ]
G  [0.585 0.322 0.807  1.585     1    -2     2    -2    -2 2.088    -2     0 ]
T  [0.319 0.322     1     -2     0 2.088    -1    -2    -2    -2 1.459 -0.415 ]
```



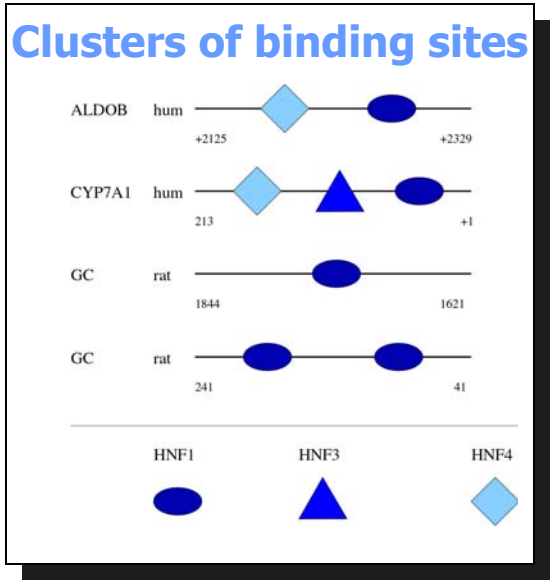


Transcription factors



Transcription factor binding sites
Regulatory nucleotide sequences

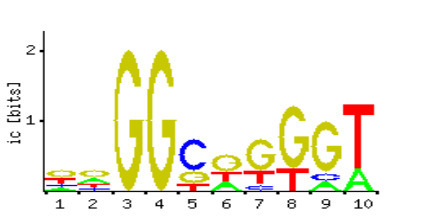
Clusters of binding sites



Detecting binding sites in a single sequence

Scanning a sequence against a PWM

Sp1



ACCCTCCCCAGGGGCGGGGGGCGGTGGCCAGGACGGTAGCTCC

A	[-0.2284	0.4368	-1.5	-1.5	-1.5	0.4368	-1.5	-1.5	-0.2284	0.4368]
C	[-0.2284	-0.2284	-1.5	-1.5	1.5128	-1.5	-0.2284	-1.5	-0.2284	-1.5]
G	[1.2348	1.2348	2.1222	2.1222	0.4368	1.2348	1.5128	1.7457	1.7457	-1.5]
T	[0.4368	-0.2284	-1.5	-1.5	-0.2284	0.4368	0.4368	0.4368	-1.5	1.7457]

Abs_score = 13.4 (sum of column scores)

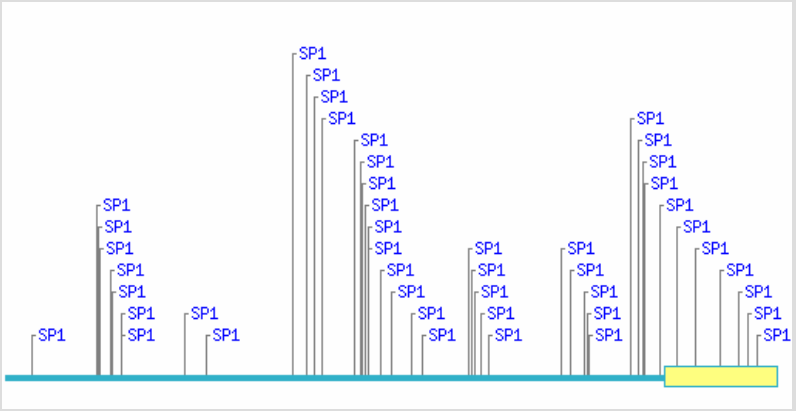
Calculating the relative score

Is 93% better than 82%?

$$\text{Rel\_score} = \frac{\text{Abs\_score} - \text{Min\_score}}{\text{Max\_score} - \text{Min\_score}} \cdot 100\,\%$$

$$= \frac{13.4 - (-10.3)}{15.2 - (-10.3)} \cdot 100\% = \mathbf{93\%}$$

Scanning 1300 bp of human insulin receptor gene with Sp1 at rel_score threshold of 75%



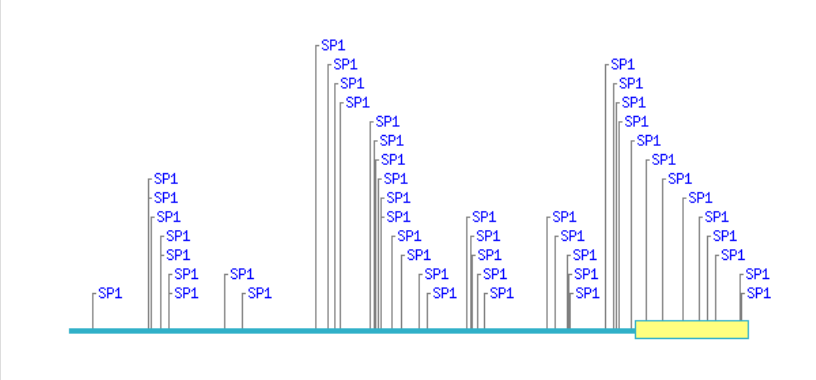
Ouch.

OnLine resources for the detection of TFBS

- TESS
- TRRD
- MatInspector (Transfac)
- ConSite (JASPAR)
  - www.phylofoot.org/consite

Phylogenetic Footprints

Scanning a single sequence

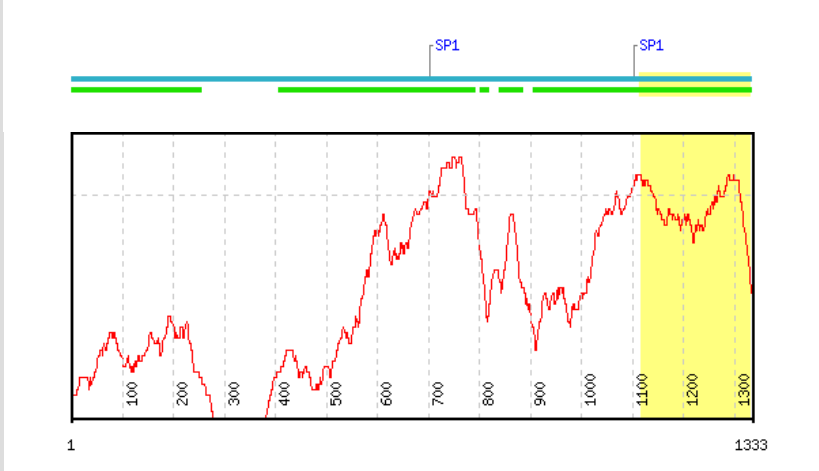


Low specificity of profiles:

- too many hits
- great majority are not biologically significant

Scanning a pair orf orthologous sequences for conserved patterns in conserved sequence regions

A dramatic improvement in the percentage of biologically significant detections



Global Progressive Alignments (ORCA, AVID, LAGAN)

- Global alignments memory = product of sequence lengths
- Progressive alignment by banding with local and running global algorithm on short banded segments
- Recursion with decreasingly stringent parameters for local

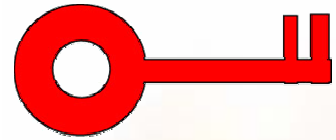
Phylogenetic Footprinting with Local Alignments

AAAAA/TTTTT	0	0
AAAAC/GTTTT	1	0
AAAAG/CTTTT	1	1
AAAAT/ATTTT	0	2
AAACA/TGTTT	3	1
…		

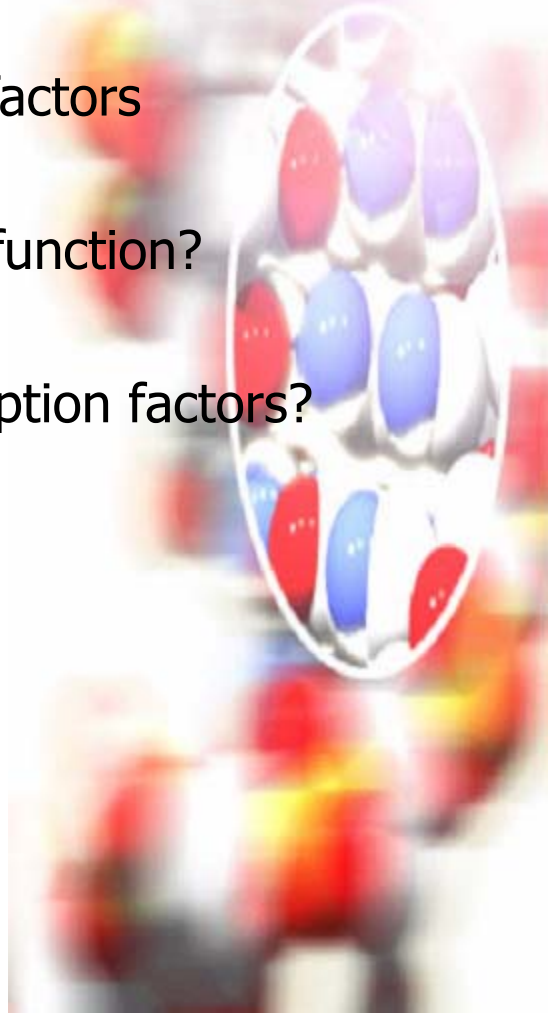
OnLine Resources for Phylogenetic Footprinting

- Alignments
  - Blastz
  - Lagan
  - Avid
  - ORCA Aligner/OrthoSeq
- Visualization
  - SymPlot
  - Vista Browser
  - PipMaker
- Linked to TFBS
  - ConSite
  - rVISTA

Considerations in Searching for Clusters of Binding Sites: Key items



- Biological motivation for grouping transcription factors
- Is there sufficient data to train a discrimination function?
- Are there binding profiles for the critical transcription factors?

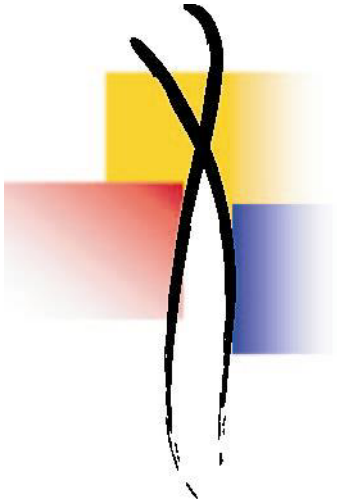


Untrained Methods

- New generation of tools to identify clusters of TFBS for user-specified set of TFs
- Identify statistically significant clusters of sites within genomes
- MSCAN Overview

OnLine Tools for Detection of Site Clusters

- MSCAN (user defined sets of TFs)
- TransRegio (liver and muscle)
- COMET/CISTER/ClusterBuster
- MCAST



Promoter Detection

Statistical Properties of Sequences

Promoter Detection

- Approaches based on detection of TFBS
- Approaches based on sequence properties
- Some considerations regarding current approaches

Promoters by Detection of Binding Sites

- Early promoter detection tools were based on promoters of small set of highly expressed genes
  - "TATA" Box at –30; CATT Box at –90
- Attempted to define the specific position at which RNA transcripts are initiated
- Benchmarking test in late 1990s
  - Most promoter prediction tools were slightly better than random guessing
  - nothing dramatically better than TATA prediction at -30

What were we doing wrong?

- Grouping diverse promoters into a single mega-class
- Attempting to pinpoint a specific start position when biochemical system is ambiguous
- Ignoring a common observation in the laboratory-based literature…

Sequence Properties in Regions containing Promoters

- Long recognized (in labs) that a significant subset of promoters are situated within or adjacent to regions rich in CG dinucleotides (What %?)
  - Without selection CG dinucleotides are modified
  - CpG islands believed to favor "open" chromatin

- A new generation of promoter detection tools (CpG-island detectors) are based on the detection of C/G-rich regions containing over-represented strings/motifs (generally A/T-rich) identified in training data

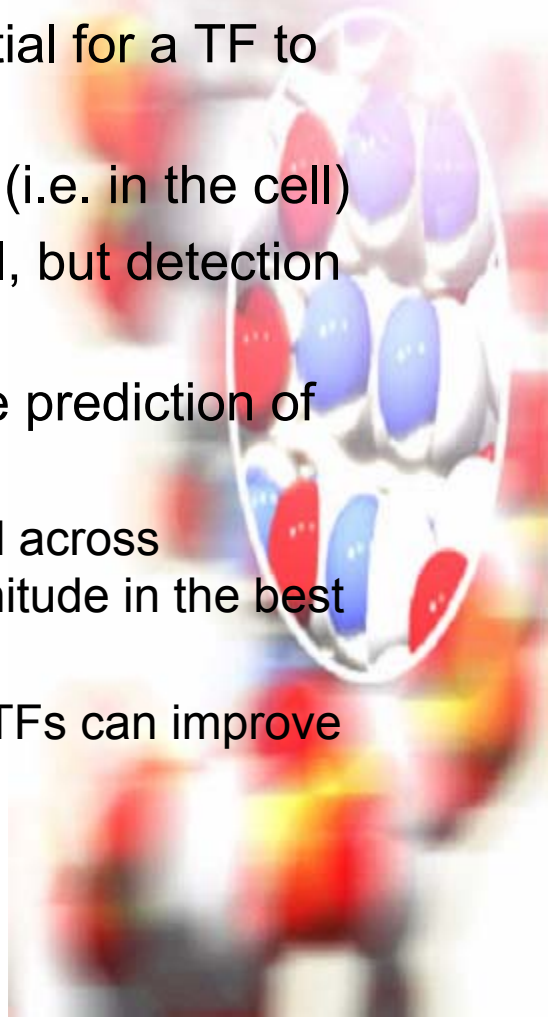
OnLine Tools for Promoter Detection

- EpoNine
- Promoter Inspector
- FirstEF

- Others?
  - Defining the likely TSS with NNPP

Looking back at part 1: Key items

- Profiles provide reasonable estimate of the potential for a TF to bind to a sequence *in vitro* (i.e. in the lab)
- In vitro binding is not predictive of *in vivo* function (i.e. in the cell)
- Prediction of promoters with CpG islands is useful, but detection of the other 50% of promoters is poor
- There are two reasonable methods to improve the prediction of individual TF binding sites
  - Phylogenetic Footprinting identifies sites conserved across evolution, improving specificity by an order of magnitude in the best cases
  - Analysis of clusters of TFBS for biologically linked TFs can improve specificity by two orders of magnitude



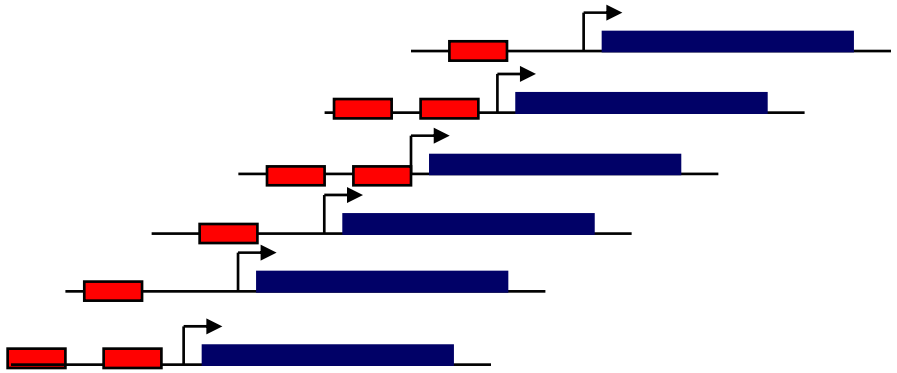
The problem

Given a set of "co-regulated" genes, define motifs over-represented in the regulatory regions

Definitions

Co-regulation:

Genes with similar expression patterns resulting from the influence of one or more common control mechanisms



Selection of Promoter sequences for analysis

Expression Profiling

Litterature-based selection

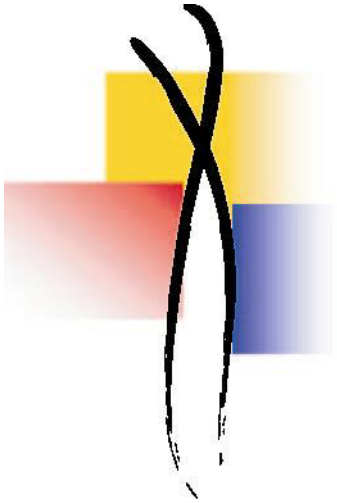
Chromatin immuno-precipitation

In vivo profiling: Green Fluorescent Protein-based approaches

Selection of Promoter sequences for analysis

Online Resources

General :	NCBI Gene Expression Omnibus EMBL ArrayExpress Stanford Microarray Database dbEST
Emerging:	UCLA Microarray Tissue Profiles Promoter Pickers



Methods for Pattern Discovery

- Word-based vs matrix-based
- Exhaustive
- Probabilistic
- Enhancements

Methods for Pattern Discovery

AAGTTAAWSAWTAAC

- Word-based

TFBS are words

Words are easily counted

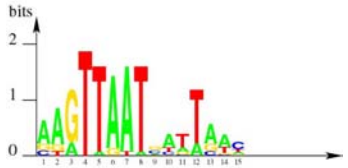
Pros

Realistic complexity

Based on well-understood statistics

Cons

TF binding properties are unevenly degenerate



- Matrix-based

TF:s do not bind to words

Pros

Matrix models are more accurate descriptions of binding preferences

Cons

Large computation time

Many local maxima (in significance)

Exhaustive methods

Exhaustive algorithm: All possible solutions are evaluated

In this context

Count all possible motifs/words. Analyze over-representation

Exhaustive methods

Word-based methods: How likely are *X* words in a set of sequences, given sequence characteristics?

CCCG**CCGGAA**TGAAATCTGATTGACATTTTCC	>EP71002 (+) Ce[IV] msp-56 B; range -100 to -75
TTCAAATTTTAACG**CCGGAA**TAATCTCCTATT	>EP63009 (+) Ce Cuticle Col-12; range -100 to -75
TCGCTGTAA**CCGGAA**TATTTAGTCAGTTTTTG	>EP63010 (+) Ce Cuticle Col-13; range -100 to -75
TATCGTCATTCTCCGCCTCTTTTCTT	>EP11013 (+) Ce vitellogenin 2; range -100 to -75
GCTTATCAATGCGC**CCGGAA**TAAAACGCTATA	>EP11014 (+) Ce vitellogenin 5; range -100 to -75
CATTGACTTTATCGAATAAATCTGTT	>EP11015 (-) Ce vitellogenin 4; range -100 to -75
ATCTATTTACAATGATAAAACTTCAA	>EP11016 (+) Ce vitellogenin 6; range -100 to -75
ATGGTCTCTA**CCGGAA**AGCTACTTTCAGAATT	>EP11017 (+) Ce calmodulin cal-2; range -100 to -75
TTTCAAAT**CCGGAA**TTTCCAC**CCGGAA**TTACT	>EP63007 (-) Ce cAMP-dep. PKR P1+; range -100 to -75
TTTCCTTCTTC**CCGGAA**TCCACTTTTTCTTCC	>EP63008 (+) Ce cAMP-dep. PKR P2; range -100 to -75
ACTGAACTTGTCTTCAAATTTCAACA**CCGGAA**	>EP17012 (+) Ce hsp 16K-1 A; range -100 to -75
TCAATG**CCGGAA**TTCTGAATGTGAGTCGCCCT	>EP55011 (-) Ce hsp 16K-1 B; range

Exhaustive methods(3)

$$P[w \ begins \ in \ i] = \prod_{j=1}^{k} p(a_j)$$

$$E[X_w] = (n-k+1)\prod_{j=1}^{k} p(a_j)$$

$$Z_w = \frac{X_w - E[X_w]}{Var[X_w]}$$

Over-representation

How many words of type '*AGGAGTGA*' are found in our sequences?

How likely is this result?

Exhaustive methods(4)

Background properties

Simple:

How likely are single nucleotides? (extended Bernoulli)

Complex:

Neglect certain words

Locations of TFBS

Higher-order descriptions of DNA

Exhaustive methods(5)

Find all words of length 7 in the yeast genome

GTCTTTATCTTCAAAGTTGTCTGTCCAAGATTTGGACTTGAAGG
ACAAGCGTGTCTTCTCAGAGTTGACTTCAACGTCCCATTGGAC
GGTAAGAAGATCACTTCTAACCAAAGAATTGTTGCTGCTTTGC
CAACCATCAAGTACGTTTTGGAACACCACCCAAGATACGTTGT
CTTGTTCTCACTTGGGTAGACCAAACGGTGAAAGAAACGAAAA
ATACTCTTTGGCTCCAGTTGCTAAGGAATTGCAATCATTGTTG
GGTAAGGATGTCACCTTCTTGAACGACTGTGTCGGTCCAGAA
GTTGAAGCCGCTGTCAAGGCTTCTGCCCCAGGTTCCGTTATTT
TGTTGGAAAACTGCGTTACCACATCGAAGAAGAAGGTTCCAGA
AAGGTCGATGGTCAAAAGGTCAAGGCTCAAGGAAGATGTTCA
AAAGTTCAGACACGAATTGAGCTCTTTGGCTGATGTTTACATC
ACGATGCCTTCGGTACCGCTCACAGAGCTCACTCTTCTATGGT
CGGTTTCGACTTGCCAACGTGCTGCCGGTTTCTTGTTGGAAAA
GGAATTGAAGTACTTCGGTAAGGCTTTGGAGAACCCAACCAG
ACCATTCTTGGCCATCTTAGGTGGTGCCAAGGTTGCTGACAAG
ATTCAATTGATTGACAACTTGTTGGACAAGGTCGACTCTATCAT
CATTGGTGGTGGTATGGCTTTCCCTTCAAGAAGGTTTTGGAAA
ACACTGAAATCGGTGACTCCATCTTCGACAAGGCTGGTGCTG
AAATCGTTCCAAAGTTGATGGAAAAGGCCAAGGCCAAGGGTG
TCGAAGTCGTCTTGCAGTCGACTTCATCATTGCTGATGCTTTC
TCTGCTGATGCCAACACCAAGACTGTCACTGACAAGGAAGGT
ATTCCAGCTGGCTGGCAAGGGTTGGACAATGGTCCAGAATCT
AGAAAGTGTTTGCTGCTACTGTTGCAAAGGCTAAGACCATTGT
CTGGAACGGTCCACCAGGTGTTTTCGAATTCGAAAAGTTCGCT
GCTGGTACTAAGGCTTTGTTAGACGAAGTTGTCAAGAGCTCTG
CTGCTGGTAACACCGTCATCATTGGTGGTGGTGACACTGCCA

Make a lookup table:

TTTTTTT/aaaaaaa	57788
GATAGGCA/tgcctatc	589
AAACCTTT/aaaggttt	456

Etc...

Exhaustive methods(6)

Matrix based methods

cagagcgat**AGGTCA**acgataatat

gcgatagca**AGGTCG**ccccgtatag

aacttggtt**AGGTCA**ttagcgagta

ggggatggg**CCCTCA**aatacgcgga

aaccggaag**GGTTCA**acgatctatt

A	3	0	0	0	0	4
C	1	1	1	0	5	0
G	1	4	3	0	0	1
T	0	0	1	5	0	0

= local multiple alignment

Few current exhaustive methods, due to NP-completeness (small widths -> extension)

Exhaustive methods(7)

Resources

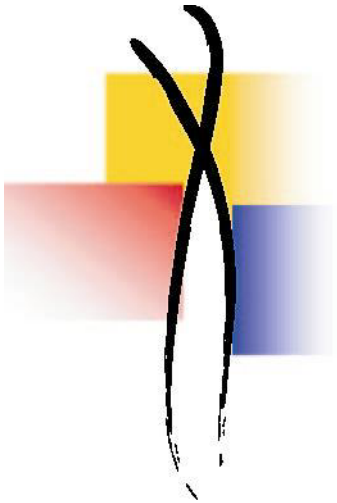
Moby Dick (*Bussemaker et al*) (not online)

RSA/Dyad analysis (*van Helden et al*)

YMF (Sinha and Tompa)

Exhaustive methods: Key items

- Algorithms with high complexity - Large sequences and/or many possible word lengths not possible
- Often word-based
- TFBS are not words (’fuzzy’ binding)
- Sensitivity susceptible to noisy indata (e.g. microarrays)



Probabilistic Methods for Pattern Discovery

- What is a probabilistic method?
- The Gibbs sampler algorithm
- Improving background models

Probabilistic Methods for Pattern Discovery(1)

Computer science:
Probabilistic algorithm: uses randomness

Bioinformatics:
Probabilistic algorithm often the same as *Monte Carlo algorithm:*
an approximation algorithm that always is fast but does not always give the best solution

Probabilistic Methods for Pattern Discovery(2)

Overview:

Find a local alignment of width *x* of sites that maximizes information content in reasonable time

Usually by Gibbs sampling or EM methods

Motivation:

TFBS are not words

Efficiency

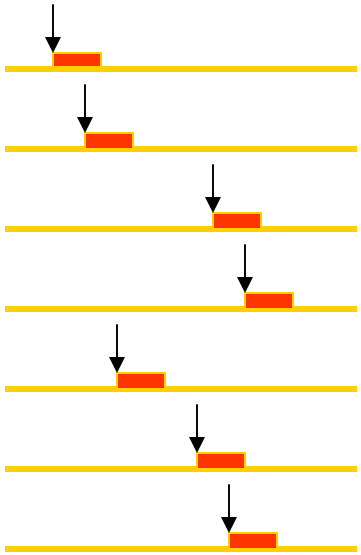
Can be intentionally influenced by biological data

The Gibbs Sampling algorithm

tgacttcc
tgatctct
agacctca
tgacctct

Two data structures used:

1) Current pattern nucleotide frequencies $q_{i,1},\ldots, q_{i,4}$ and corresponding background frequencies $p_{i,1},\ldots, p_{i,4}$



2) Current positions of site startpoints in the N sequences $a_1, \ldots, a_N$, i.e. the alignment that contributes to $q_{i,j.}$
One starting point in each sequence is chosen randomly initially.

Probabilistic Methods for Pattern Discovery(4)

Iteration step

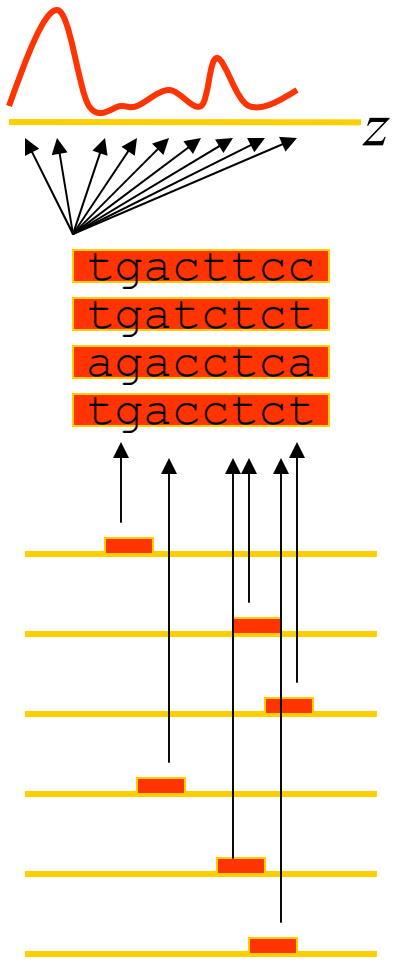
A Remove one sequence z from the set. Update the current pattern according to

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B}$$

Pseudocount for symbol j

Sum of all pseudocounts in column

B 'Score' the current pattern against each possible occurence a_k in z. Draw a new a_k with probabilities based on respective score divided by the background model



Building in biological knowledge in pattern finding - priors

How do priors work?

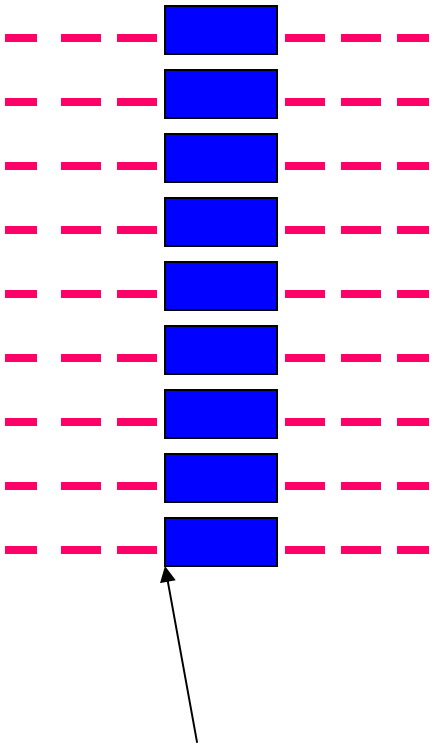
Essentially by *increasing the psudocounts* by some fraction submitted in the prior

Example: A certain residue is according to our prior knowledge an A in 47/100 cases.

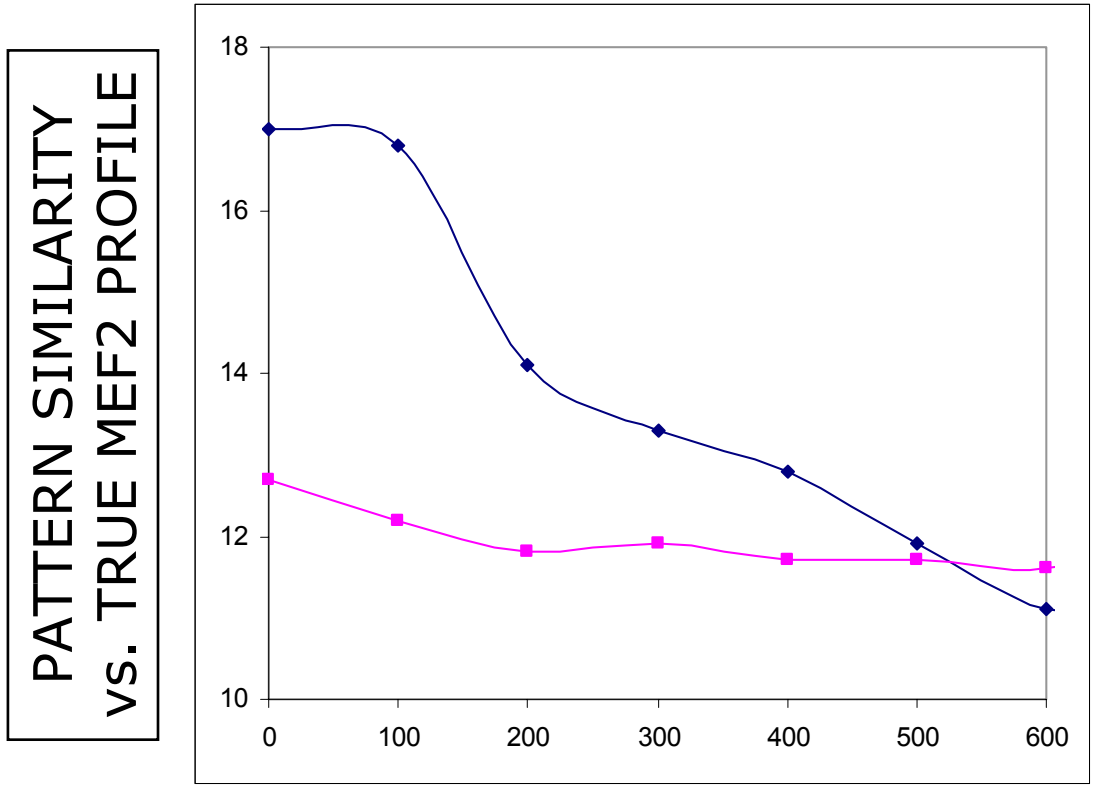
New pseudocount for first residue, A:
50/100 x k x#number of sites

Probabilistic Methods for Pattern Discovery(5)

Sensitivity weaknesses: 'Pattern drowning'



True Mef2 Binding Sites



Probabilistic Methods for Pattern Discovery(6)

Correction for background properties

Workman & Stormo (ANN-Spec)

– Train on background set as well to find 'commonly occuring' patterns. Maximization of probabililty of finding pattern in positive sequences and not in background seqsequences

In effect: Try to discriminate between 'common' and 'novel' patterns

Thijs et al, Bailey and Elkan

Markov background model describing DNA in *m*:th order

Probabilistic Methods for Pattern Discovery(7)

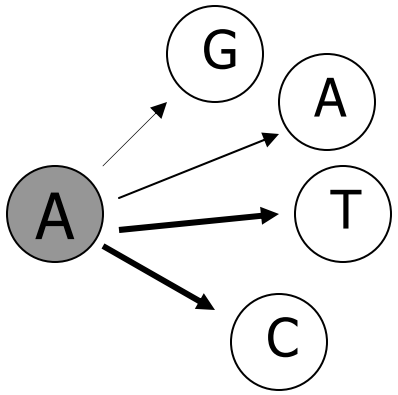
What is a higher-order background model?

Zero-order:

p(A)=0.29,
p(C)=0.21,
p(G)=0.21,
p(T)=0.29

$$P(seq) = \prod_{i=1\ldots N} P(nucleotide_i)$$

First-order:



m:th-order:

The chance of drawing base *x* is dependant on the identity of the previous *m* bases

Probabilistic Methods for Pattern Discovery(8)

Online resources

Gibbs Motif Sampler(*Lawrence et al*)

MEME(*Bailey and Elkan*)

AnnSpec(*Workman and Stormo*)

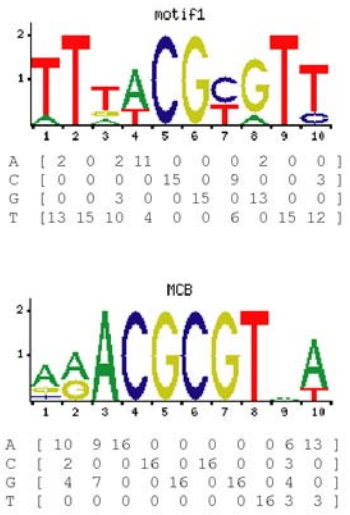
AlignAce(*Roth et al*)

Let's Try RSA (exhaustive) AND YRSA (probabilistic)

DNA-damage response partially mediating by MCB

MMS induced DNA repair genes

name	*systematic name*
PLM2	YDR501W
DIN7	YDR263C
DUN1	YDL101C
RNR2	YER070W
RNR4	YGR180C
-	YBR070C
RAD54	YGL163C
-	YER004W
RAD51	YER095W



YDR501W
YDR263C
YDL101C
YER070W
YGR180C
YBR070C
YGL163C
YER004W
YER095W

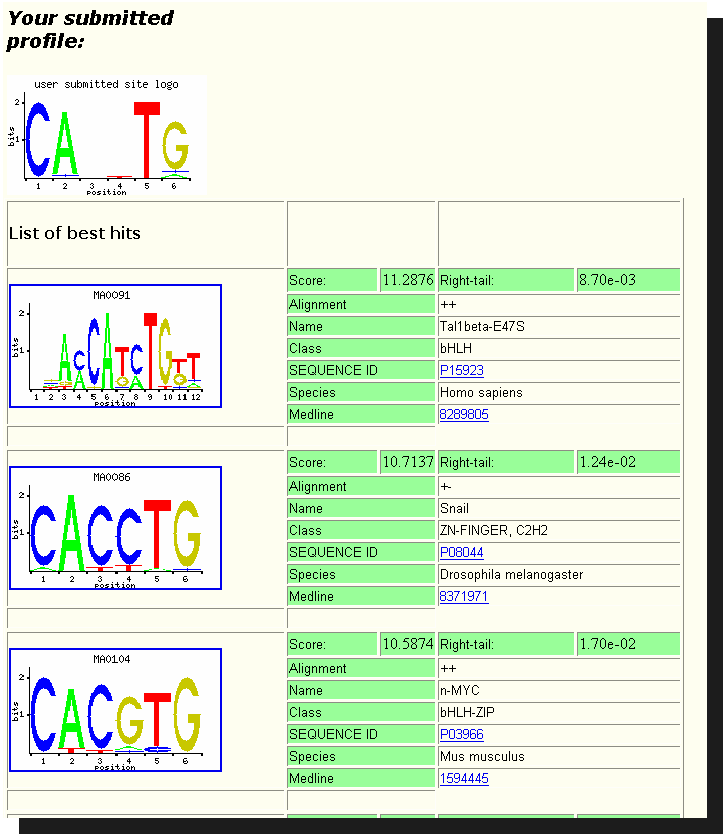
Probabilistic Methods for Pattern Discovery: Key items

Gibbs Sampling/EM algorithms

- Complexity is moderate. Optimality not guaranteed.
- Low sensitivity: patterns 'drown' in large sequences (~>500 bp)
- Sensitivity susceptible to noisy input data (e.g. microarrays)

Evaluation of patterns(2)

Algorithms for pattern comparison



Hughes et al

Based on protein BLOCKS alignment algorithm (*Pietrokowski*)

Sandelin & Wasserman

Needleman-Wunsch variant

THANKS FOR YOUR PARTICIPATION

- Analysis of regulatory sequences is not a highly-defined process
- Each method has one or more limitations that you should understand prior to relying on the results
- Cross-species comparisons help when expectation that regulation is conserved
- Largest problem in pattern discovery is usually the quality of the initial set of genes