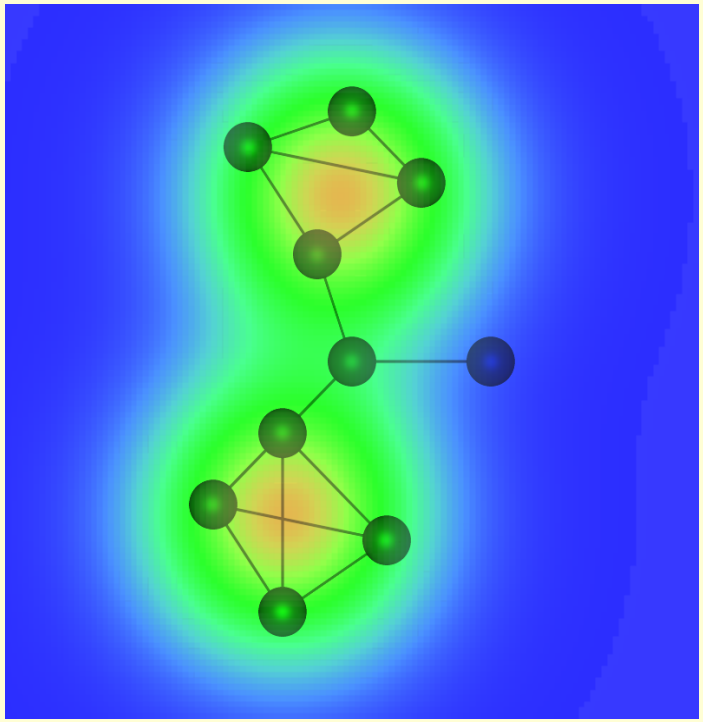
Computational Molecular Biology
Biochem 218 – BioMedical Informatics 231

http://biochem218.stanford.edu/

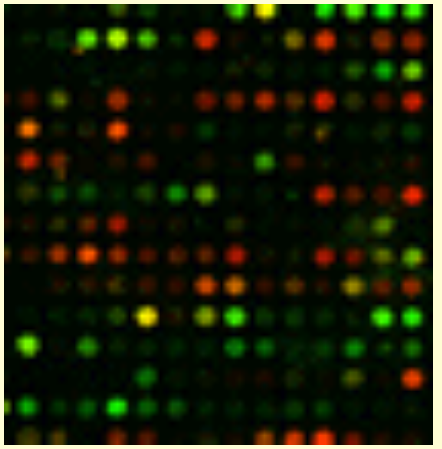
Discovering Transcription Factor Binding Sites in Co-Regulated Genes



Doug Brutlag
Professor Emeritus
Biochemistry & Medicine (by courtesy)

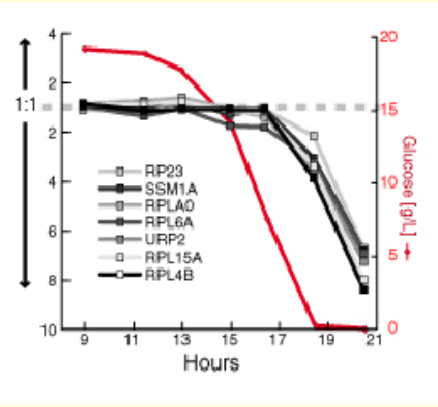
Motivation

MicroArray analysis of whole genome gene expression



↓

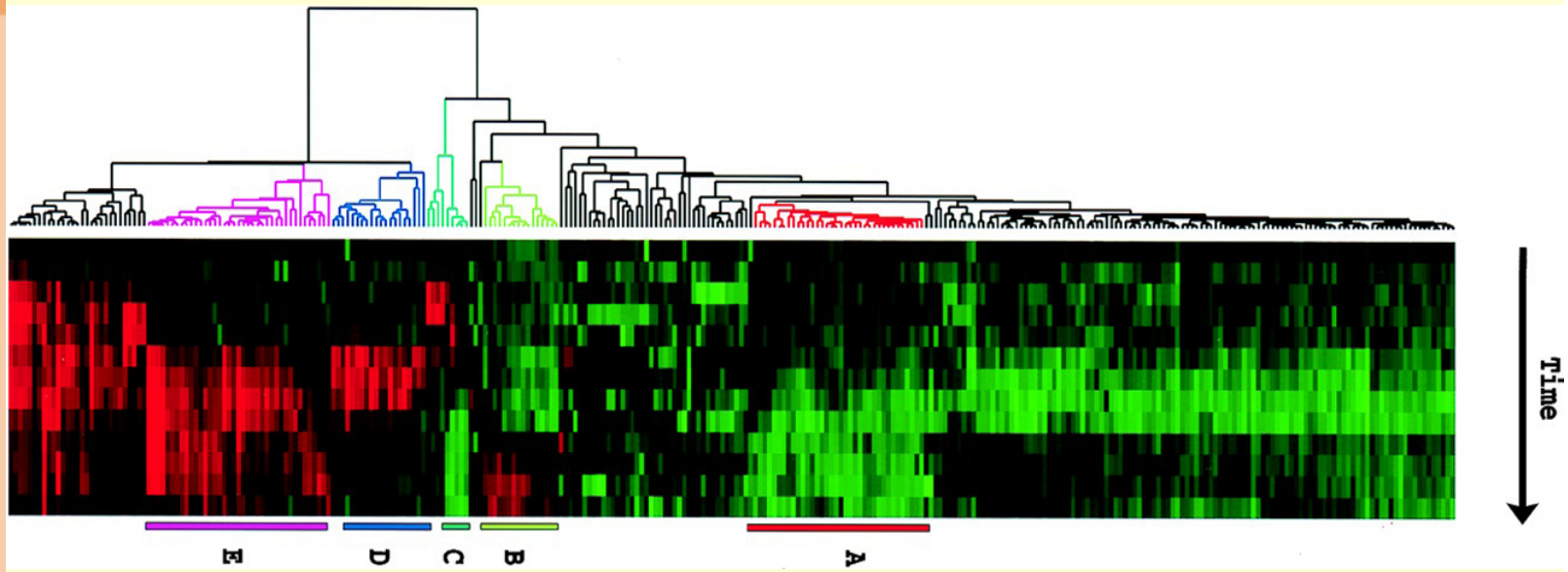
Clustering of genes based on their expression pattern



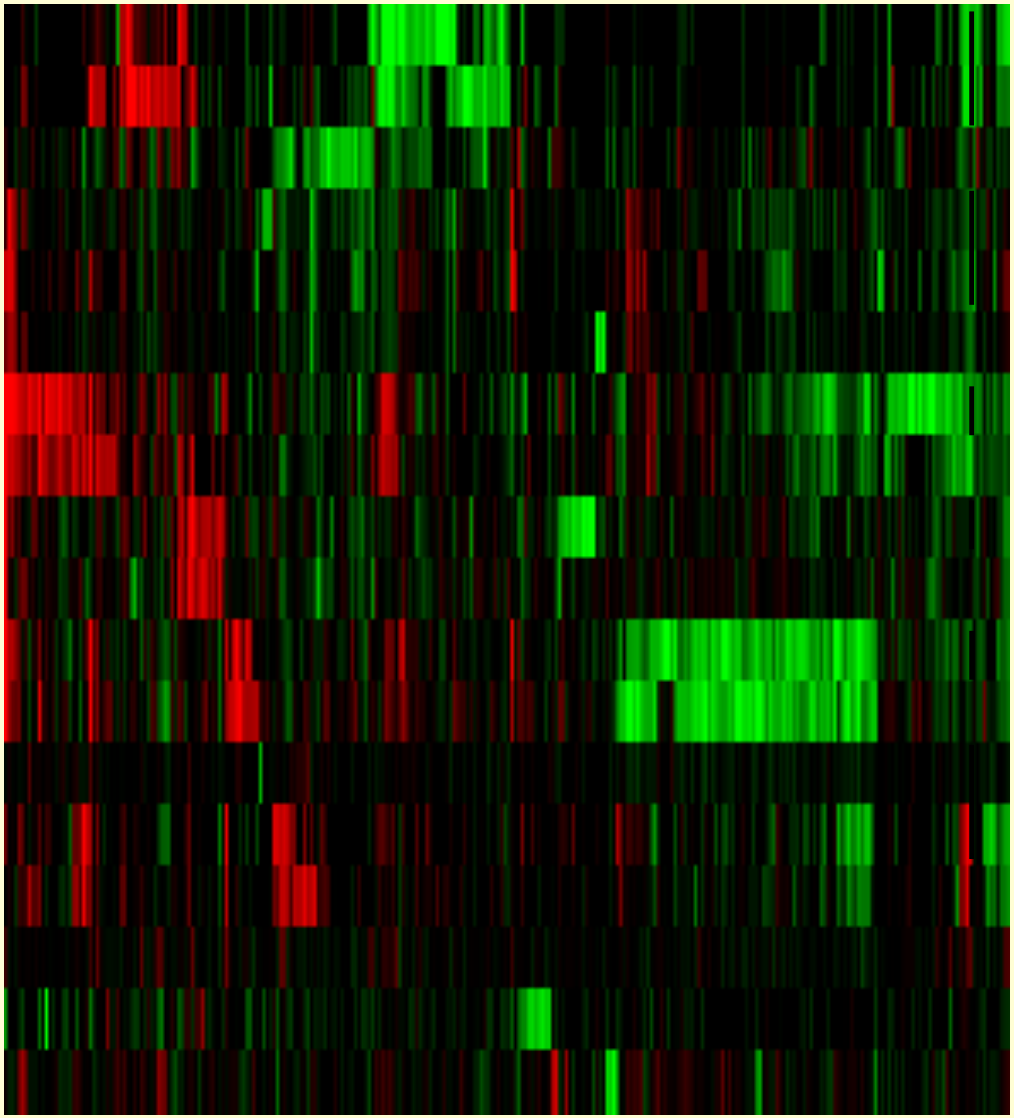
↓

Searching for conserved sequence motifs regulating the expression

Megacluster of Yeast Gene Expression



Human Gene Expression Signatures



T Cells Signaling

DNA Damage

Fibroblast Stimulation

B Cells Signaling

CMV Infection

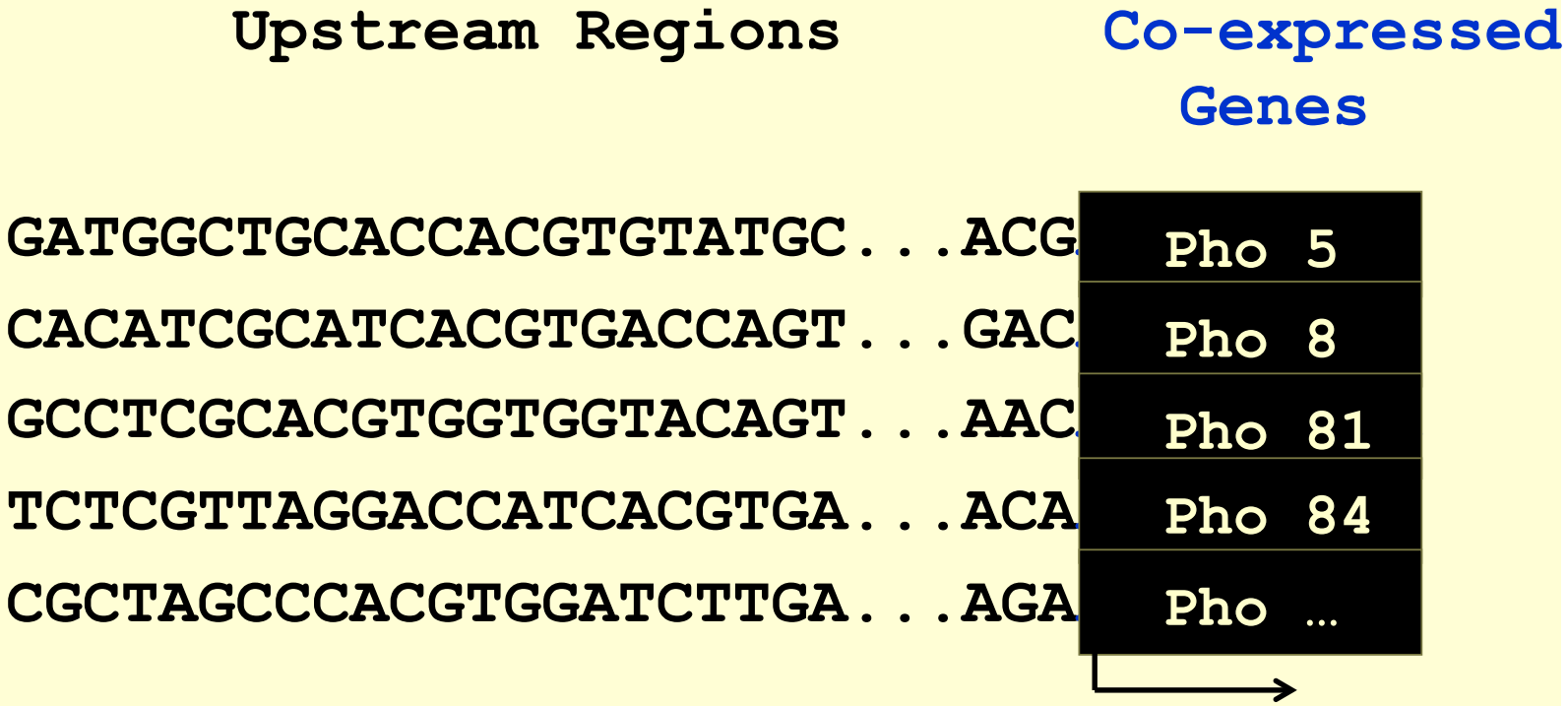
Anoxia

Polio Infection

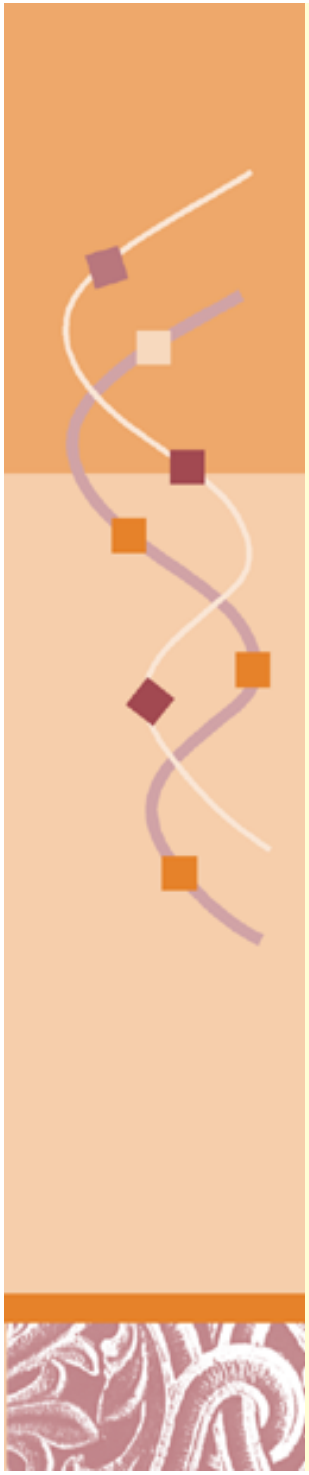
Monocytes Signaling IL4

Hormone

Finding Transcription Factor Binding Sites



Finding Transcription Factor Binding Sites



Upstream Regions	Co-expressed Genes
GATGGCTGCAC**CACGTG**TATGC...ACG	**ATGTCTCGC**
CACATCGCAT**CACGTG**ACCAGT...GAC	**ATGGACGGC**
GCCTCG**CACGTG**GTGGTACAGT...AAC	**ATGACTAAA**
TCTCGTTAGGACCAT**CACGTG**A...ACA	**ATGAGAGCG**
CGCTAGCC**CACGTG**GATCTTGT...AGA	**ATGGCCTAT**

Finding Transcription Factor Binding Sites

Three Algorithms

- BioProspector
  - Presented in 2000
  - Extends Gibb's sampling (stochastic method)
  - For any cluster of sequences
- MDScan
  - Deterministic approach
  - Enumerative
  - Very fast
  - For sequences with some ranking information
- MotifCut and MotifScan
  - Graph-based
  - Does not use PSSMs
  - Novel and sensitive

Representing Ambiguous DNA Motifs

- Sequence Patterns (Regular expressions)

Consensus motif: CACAAAA

Degenerate motif: CRCAAAW
A/G A/T

- IUPAC nomenclatures for DNA ambiguities

A G	Adenine Guanine	C T	Cytosine Thymine
R (A, G)	puRine	Y (C, T)	pyrimidines
W (A, T)	Weak hydrogen bond	S (C, G)	Strong hydrogen bond
M (A, C)	common aMino group	K (G, T)	common Keto group
B (C, G, T) H (A, C, T)	not A not G	D (A, G, T) V (A, C, G)	not C not T or U
N (A, C, G, T)	aNy		

Weight Matrix for Transcription Factor Binding Sites

A DNA Motif as a position specific frequency weight matrix

Sites

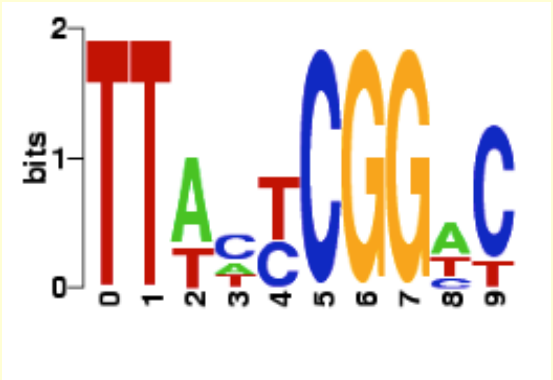
ATGGCATG
AGGGTGCG
ATCGCATG
TTGCCACG
ATGGTATT
ATTGCACG
AGGGCGTT
ATGACATG
ATGGCATG
ACTGGATG

Alignment Matrix

Pos	A	C	G	T
1	9	0	0	1
2	0	1	2	7
3	0	1	7	2
4	1	1	8	0
5	0	7	1	2
6	8	0	2	0
7	0	3	0	7
8	0	0	8	2

Frequency weight Matrix

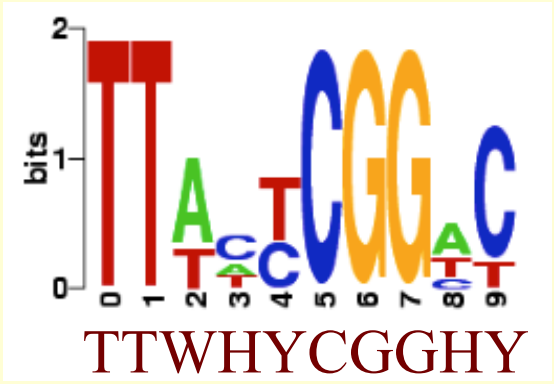
Pos	A	C	G	T	Con
1	0.9	0	0	0.1	A
2	0	0.1	0.2	0.7	T
3	0	0.1	0.7	0.2	G
4	0.1	0.1	0.8	0	G
5	0	0.7	0.1	0.2	C
6	0.8	0	0.2	0	A
7	0	0.3	0	0.7	T
8	0	0	0.8	0.2	G



Weight Matrix with Consensus Sequence & Logotype with Degenerate Consensus

Weight Matrix or Position Specific Scoring Matrix

Positions	A	G	C	T	Consensus
1	0.05	0.85	0.07	0.03	G
2	0.87	0.05	0.01	0.07	A
3	0.03	0.12	0.7	0.15	C
4	0.1	0.03	0.02	0.85	T
5	0.6	0.02	0.35	0.03	A
6	0.01	0.03	0.9	0.06	C
7	0.02	0.05	0.9	0.03	C
8	0.8	0.05	0.03	0.12	A

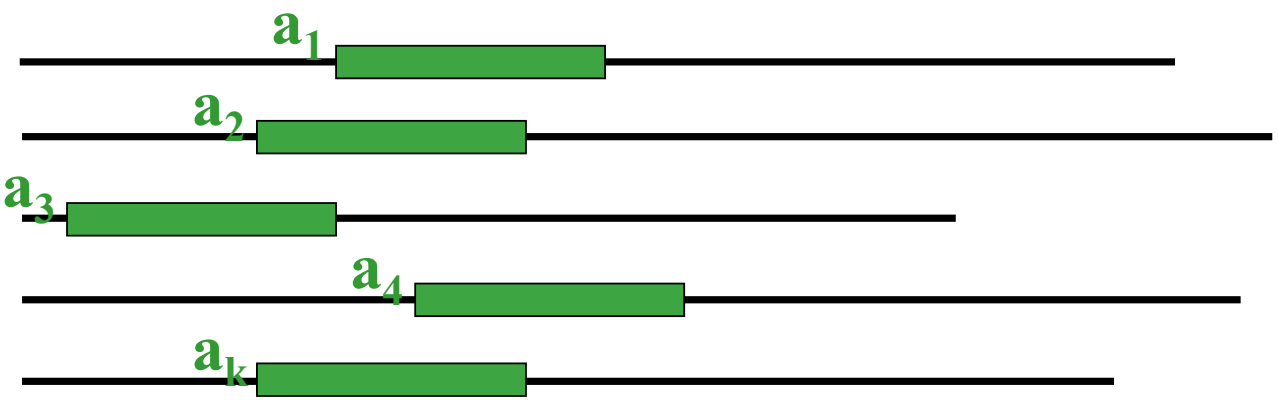


BioProspector Initialization

Gather together upstream regulatory regions

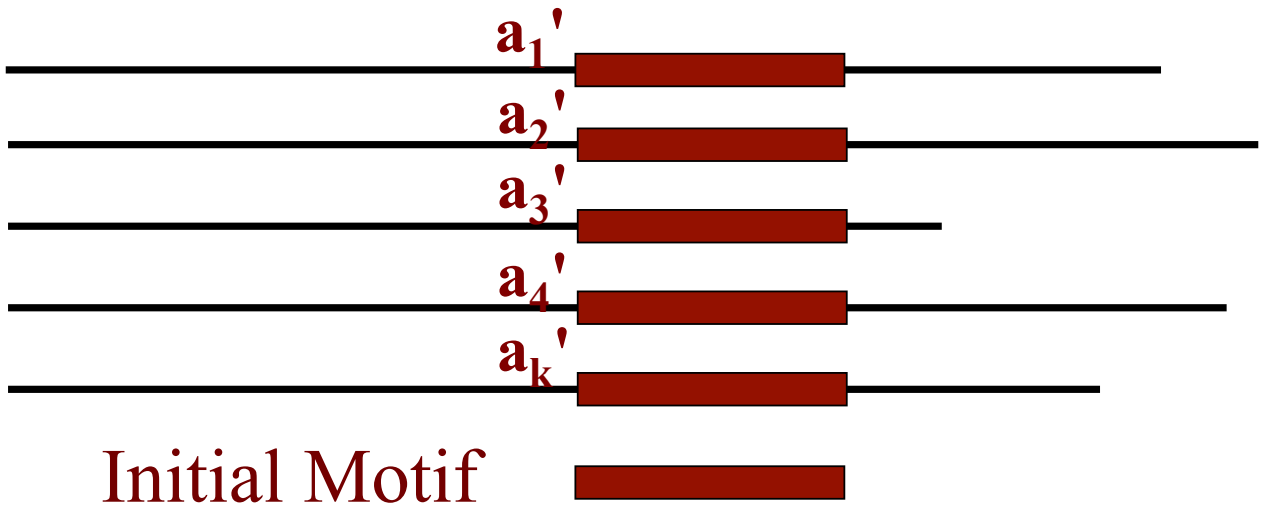
BioProspector Initialization

Actual Location of Regulatory Motifs is Unknown



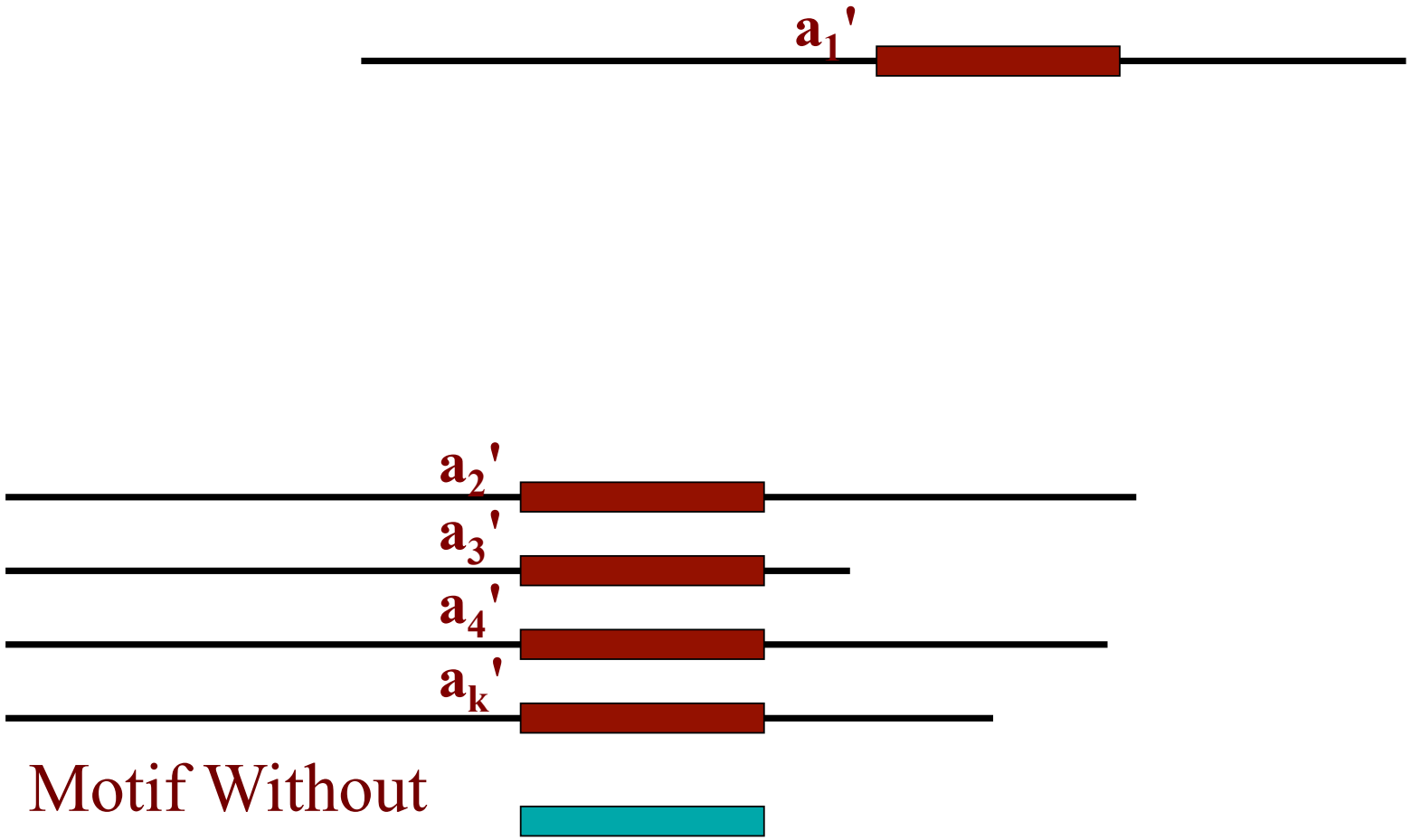
BioProspector Initialization

Randomly initialize the beginning motif



BioProspector Iterative Update

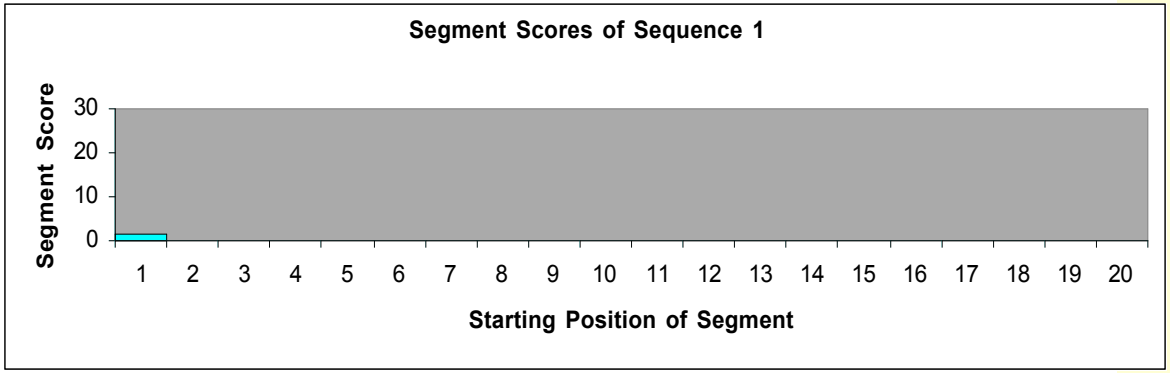
Take out one sequence at a time with its segment

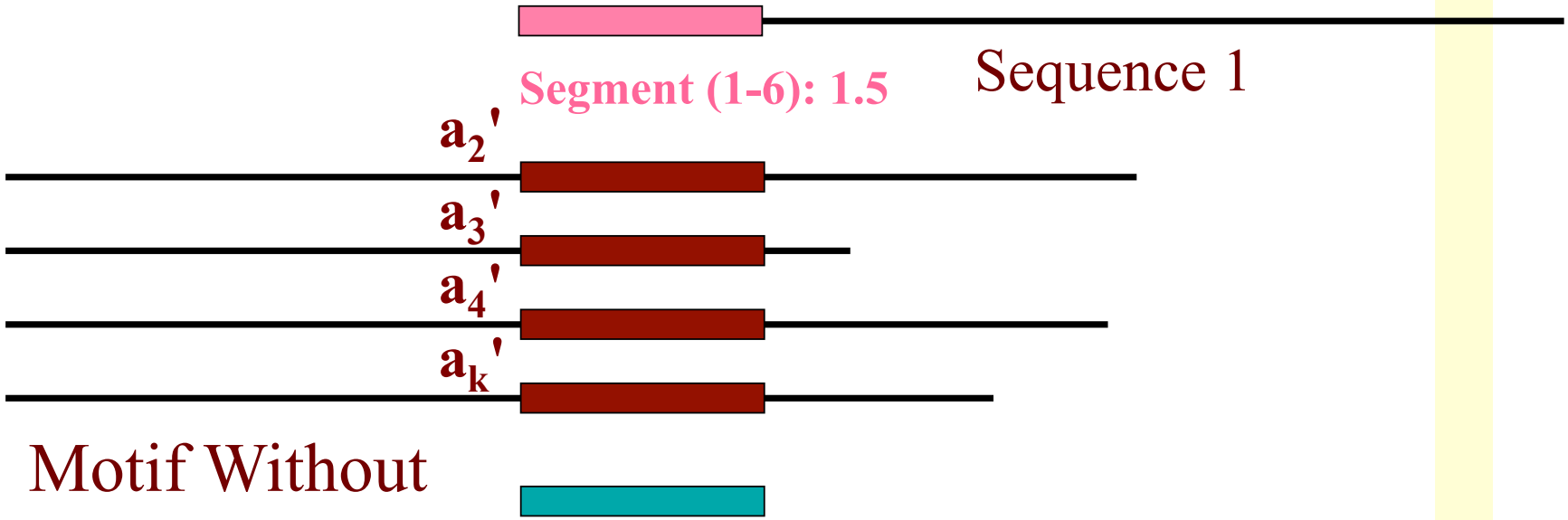


Motif Without a_1' Segment

BioProspector Iterative Update

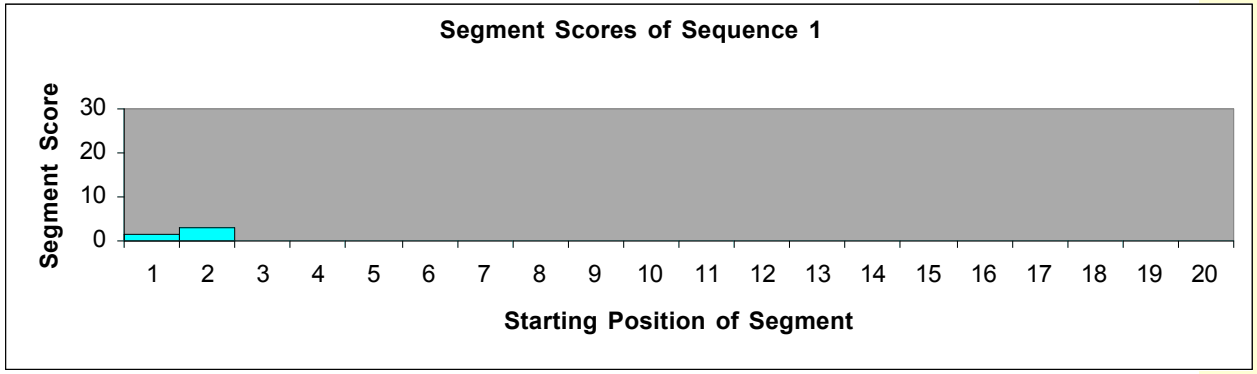
Score each segment with the current motif

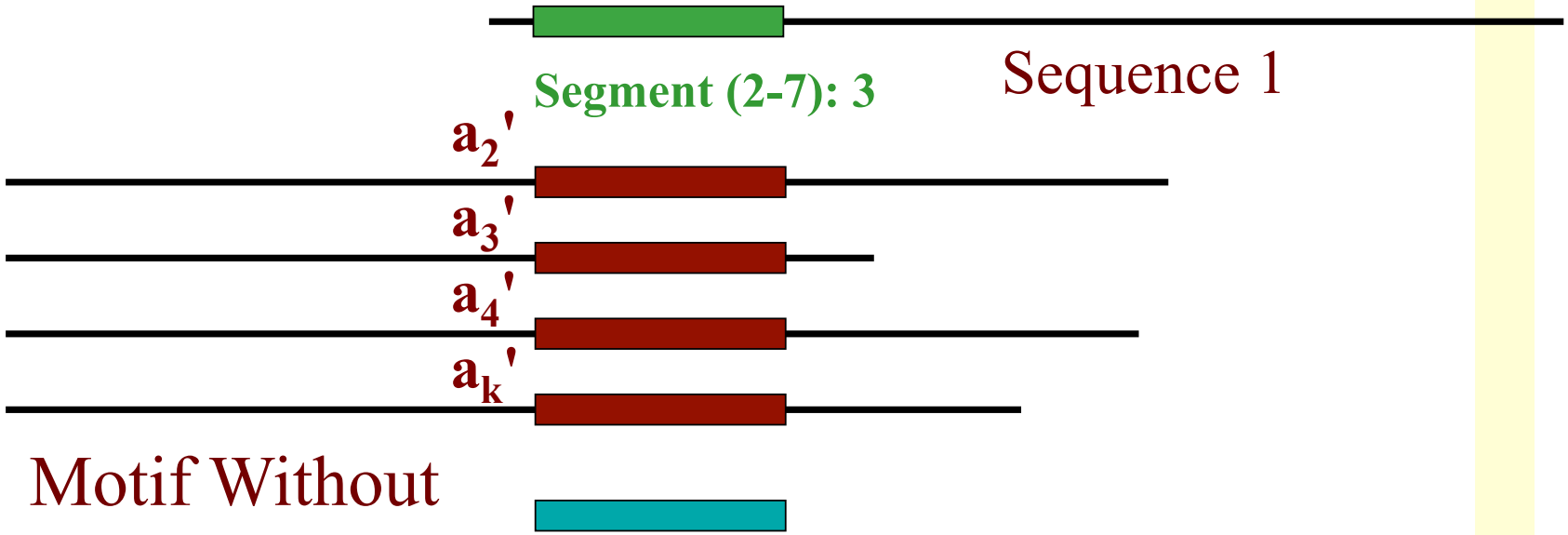




BioProspector Iterative Update

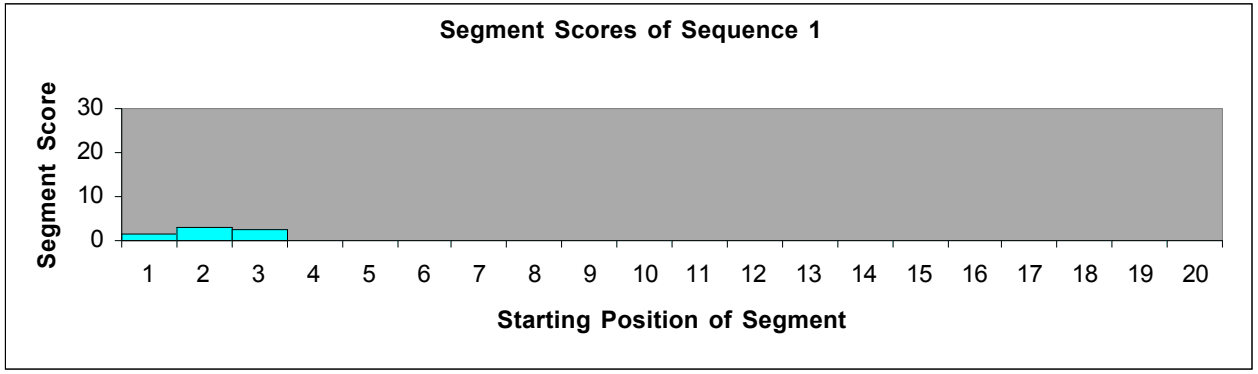
Score each segment with the current motif

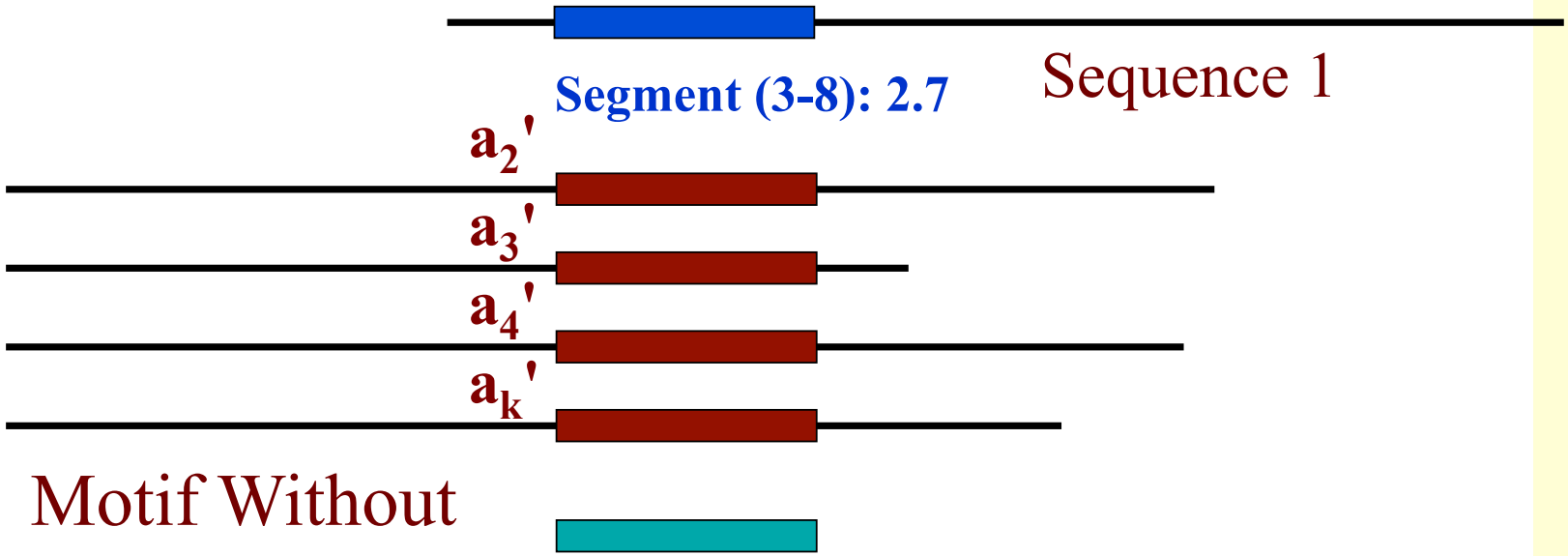




BioProspector Iterative Update

Score each segment with the current motif

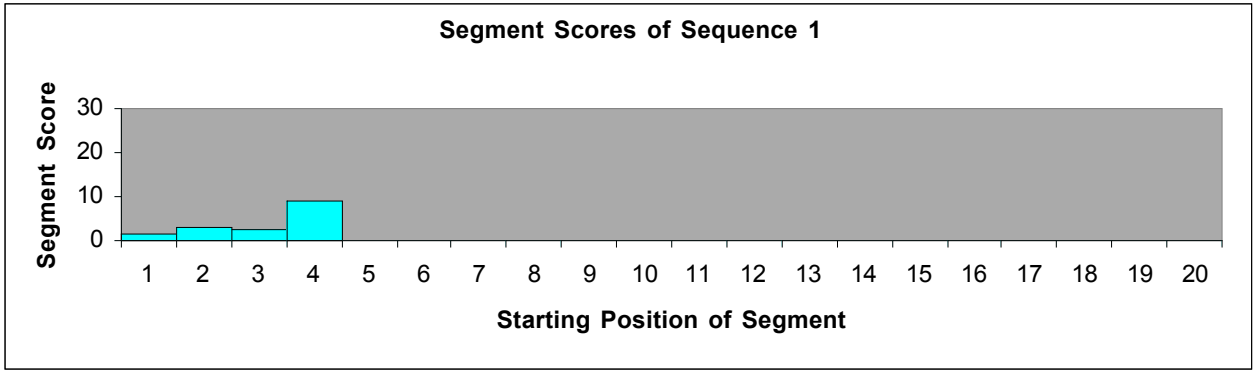


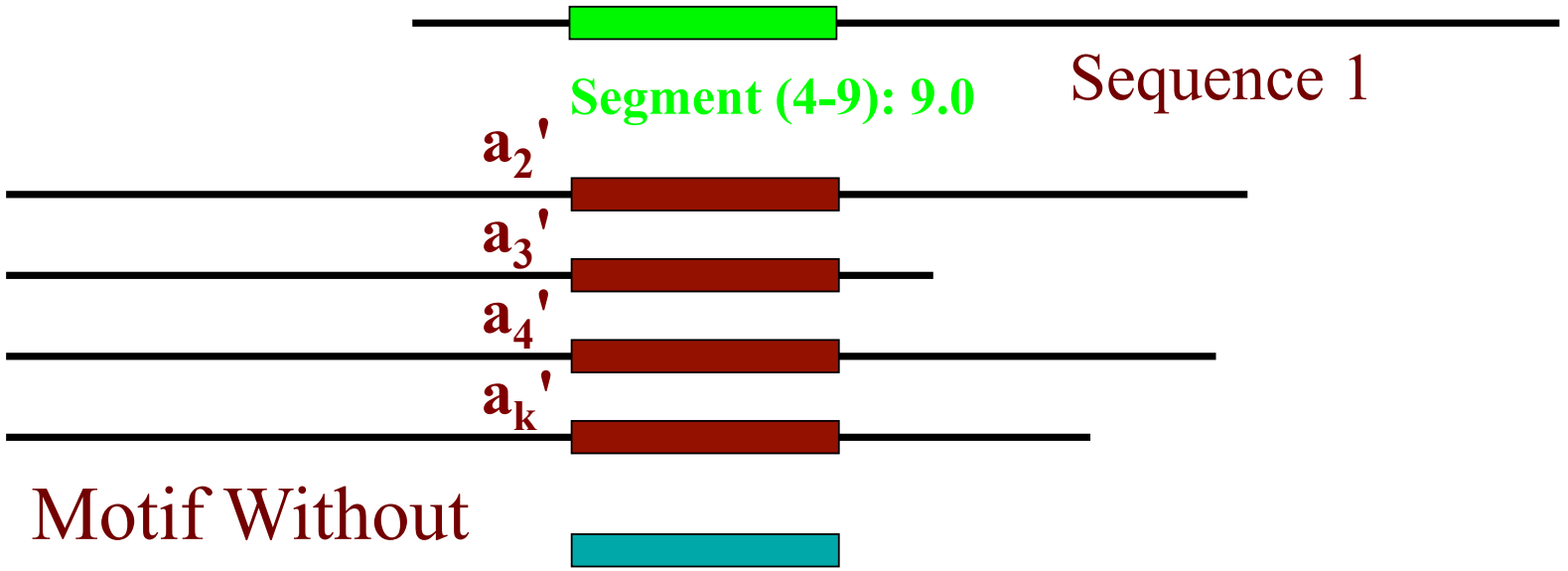


Motif Without a_1' Segment

BioProspector Iterative Update

Score each segment with the current motif

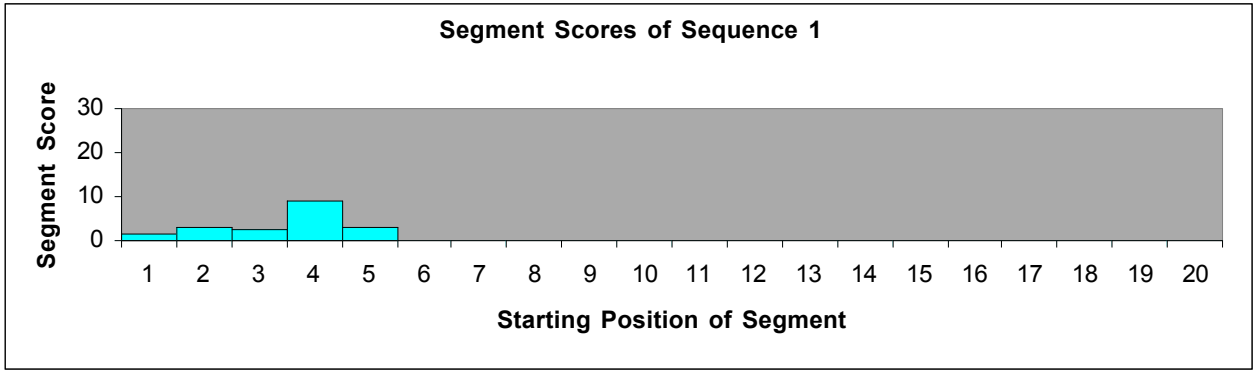


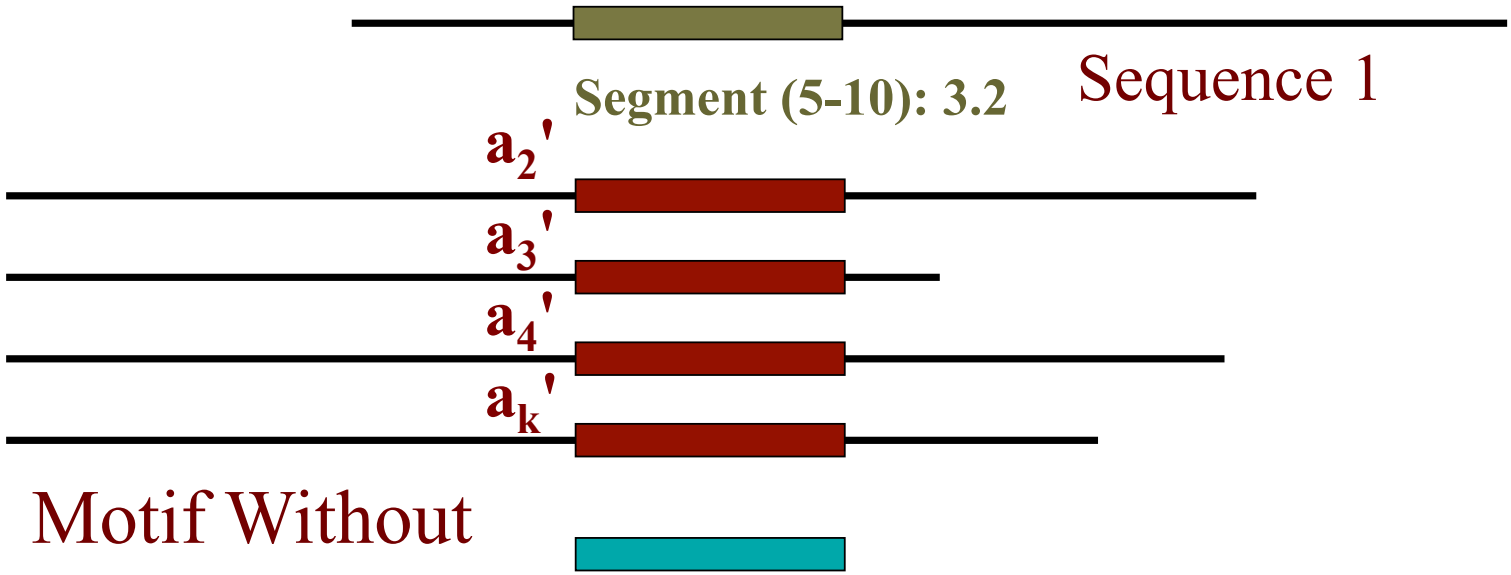


Motif Without a_1' Segment

BioProspector Iterative Update

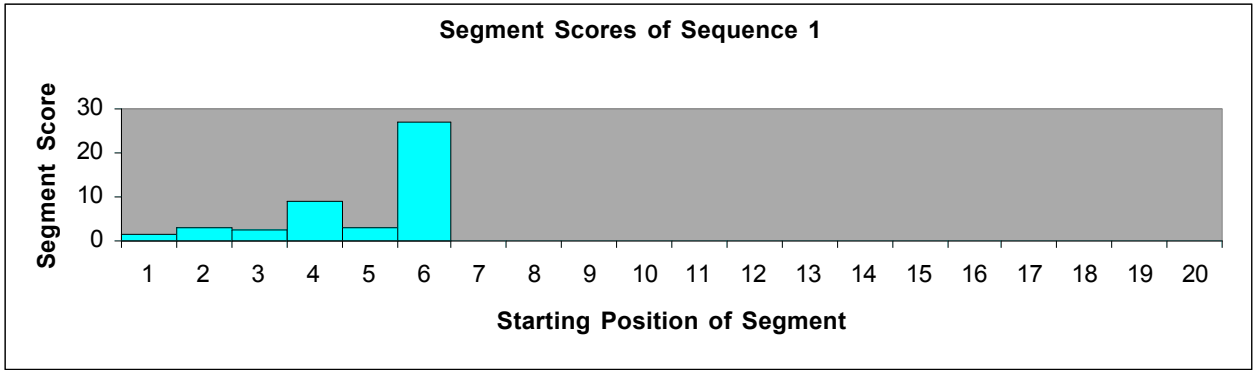
Score each segment with the current motif

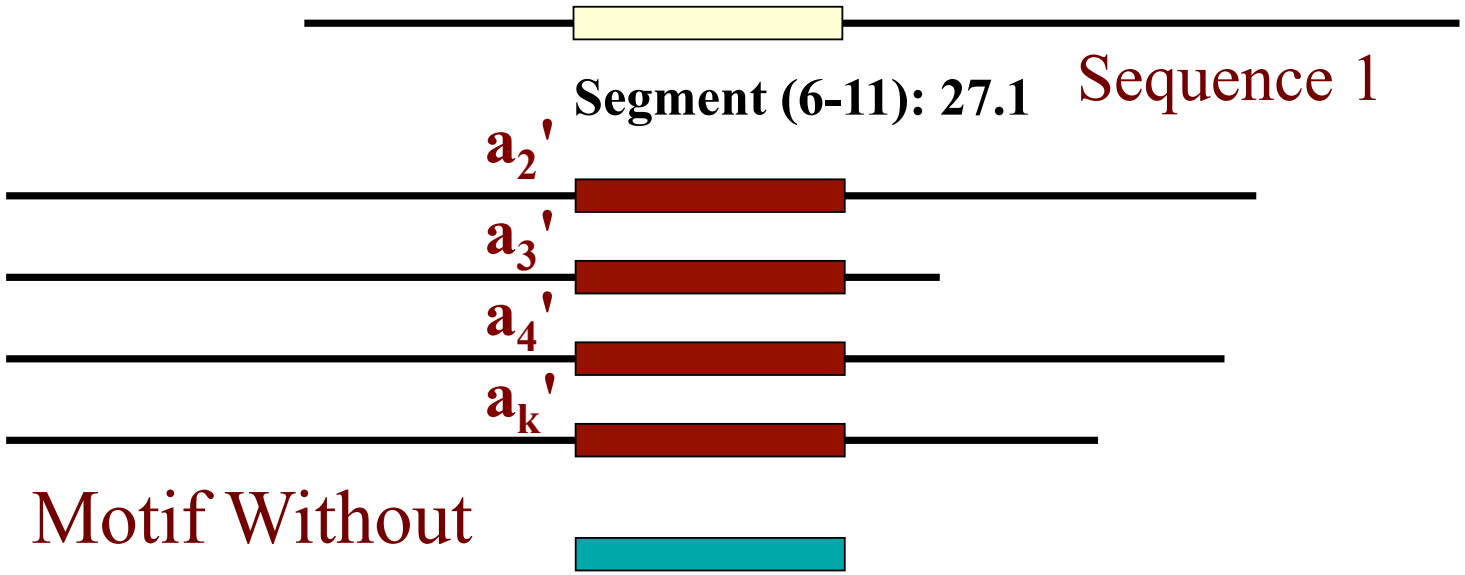




BioProspector Iterative Update

Score each segment with the current motif

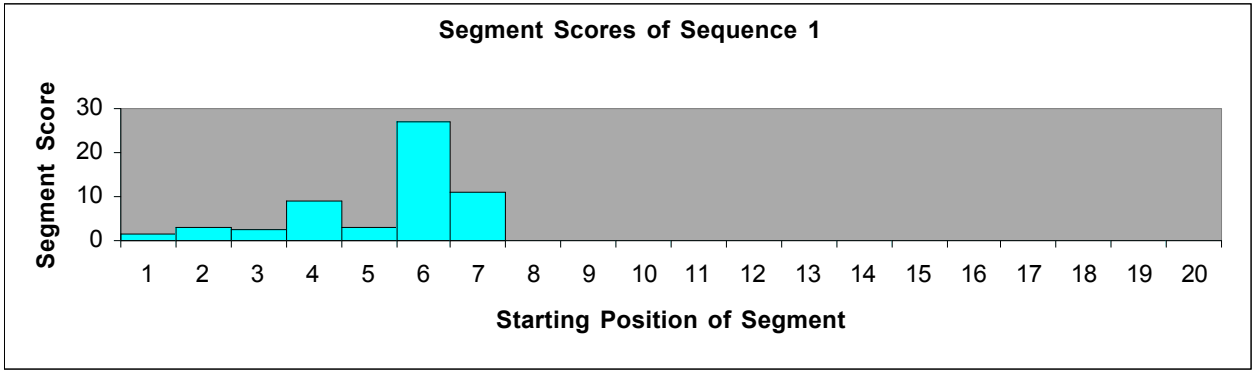


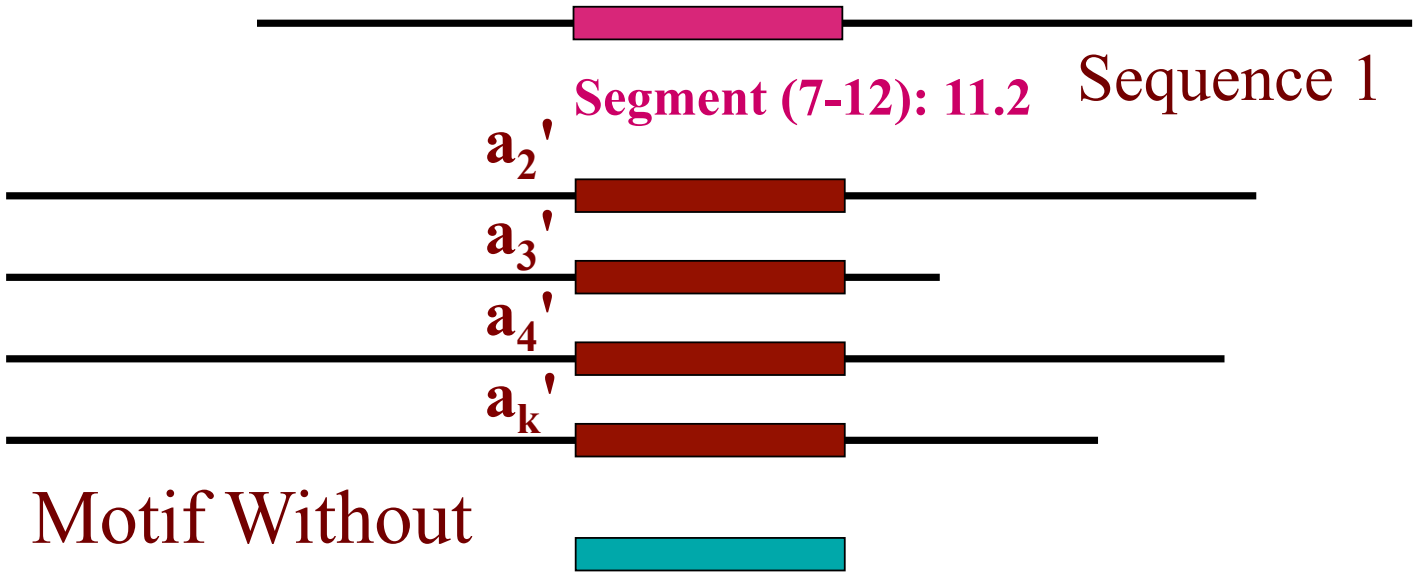


Motif Without a_1' Segment

BioProspector Iterative Update

Score each segment with the current motif

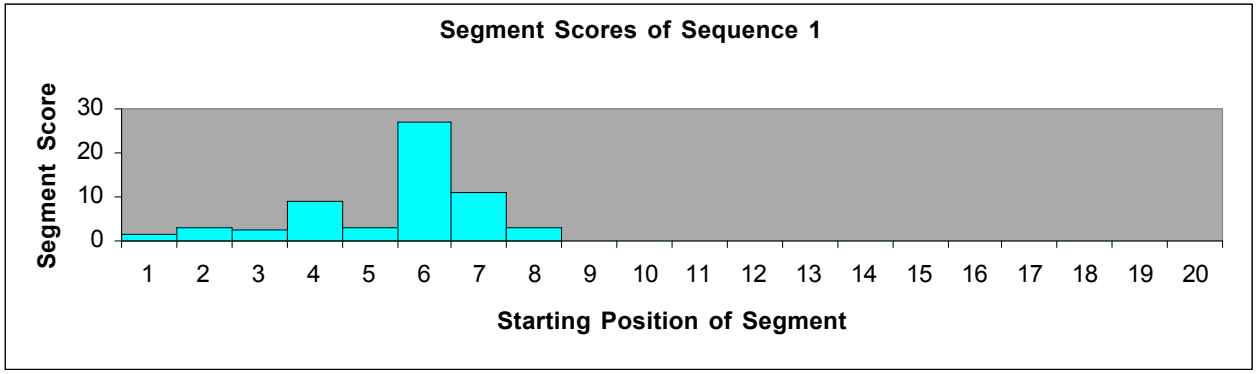


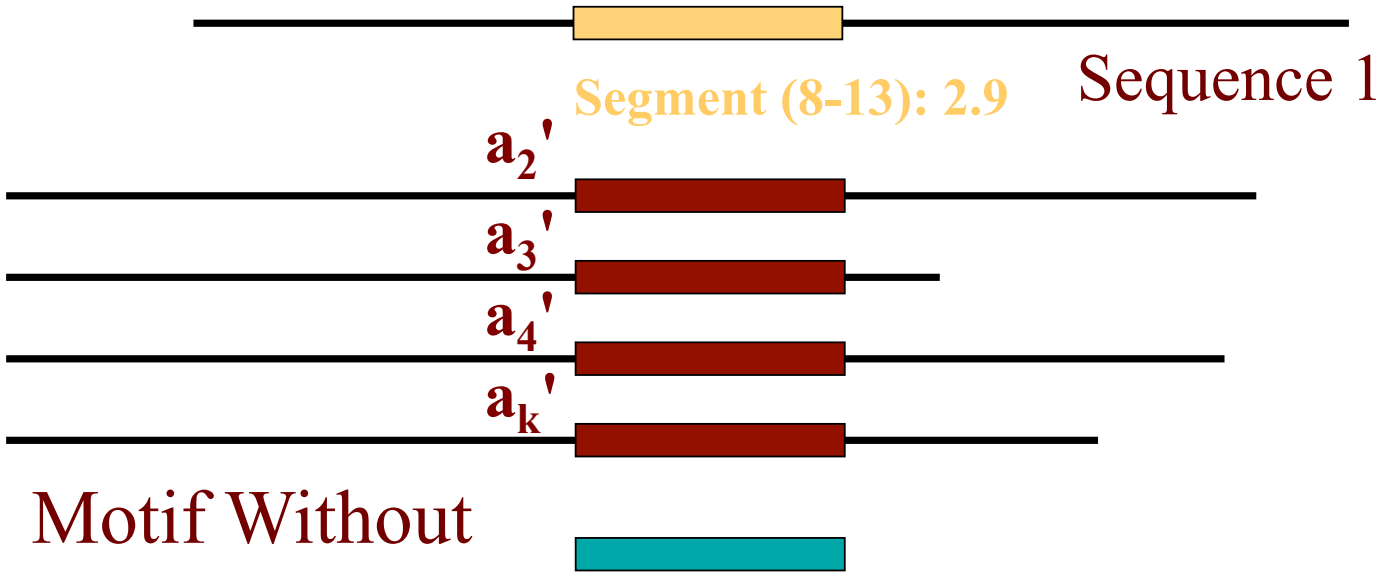


Motif Without a_1' Segment

BioProspector Iterative Update

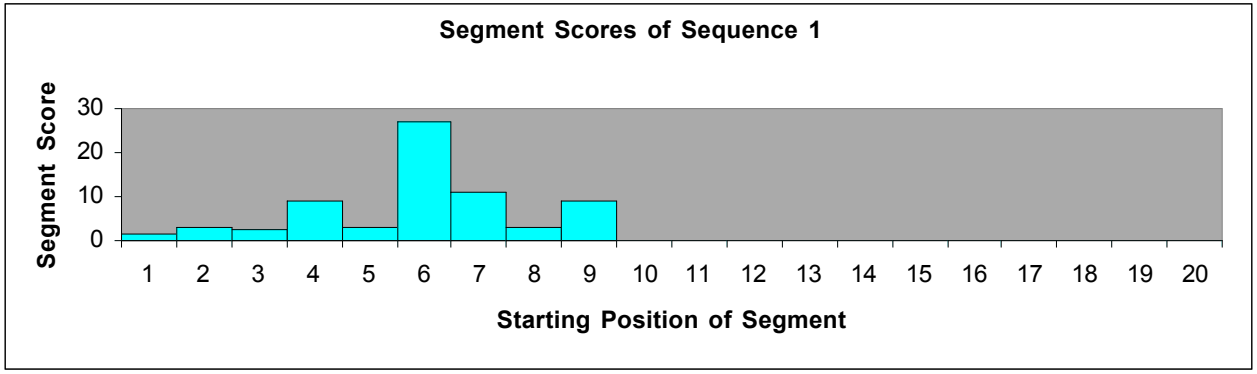
Score each segment with the current motif

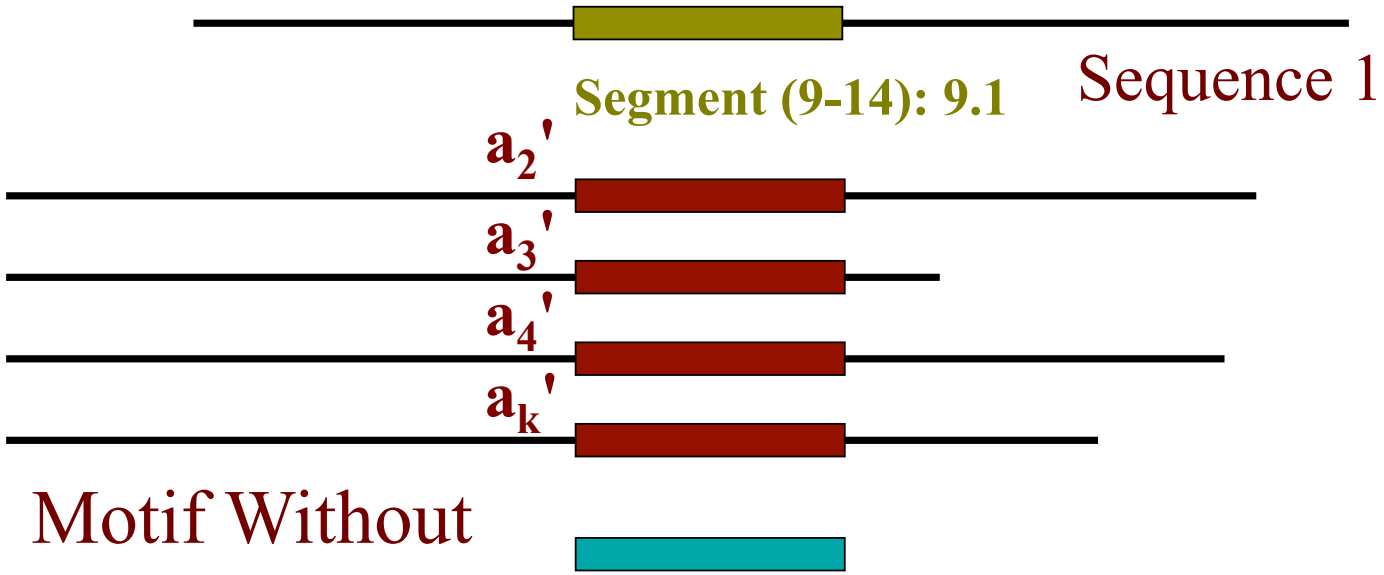




BioProspector Iterative Update

Score each segment with the current motif

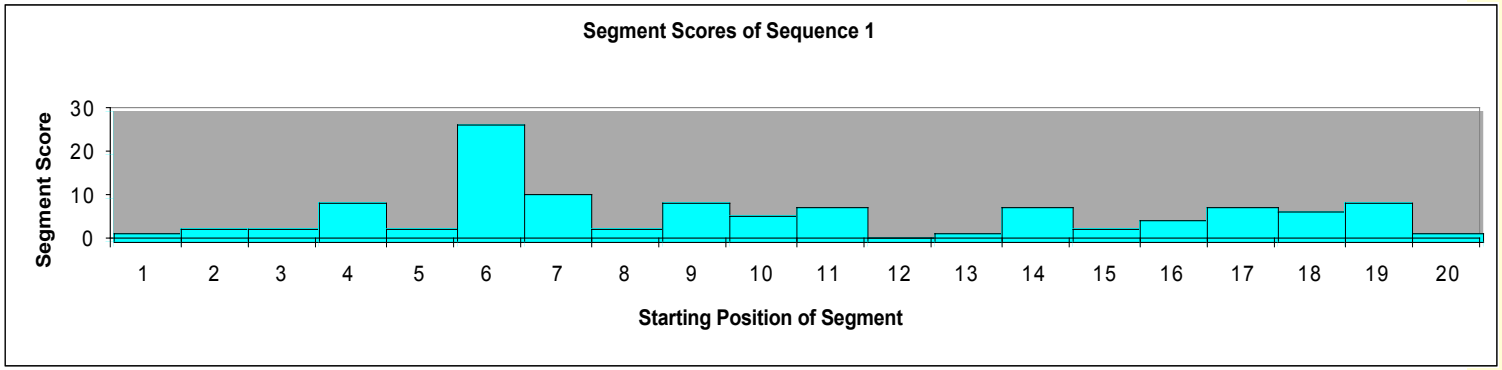


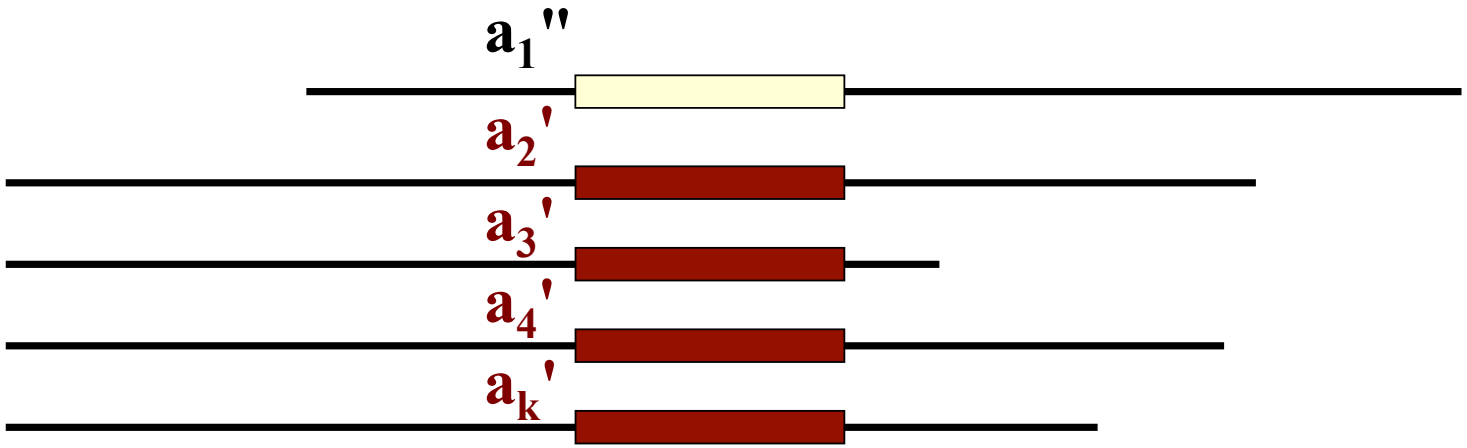


Motif Without a_1' Segment

BioProspector Iterative Update

Score sequence 1 in all possible alignments

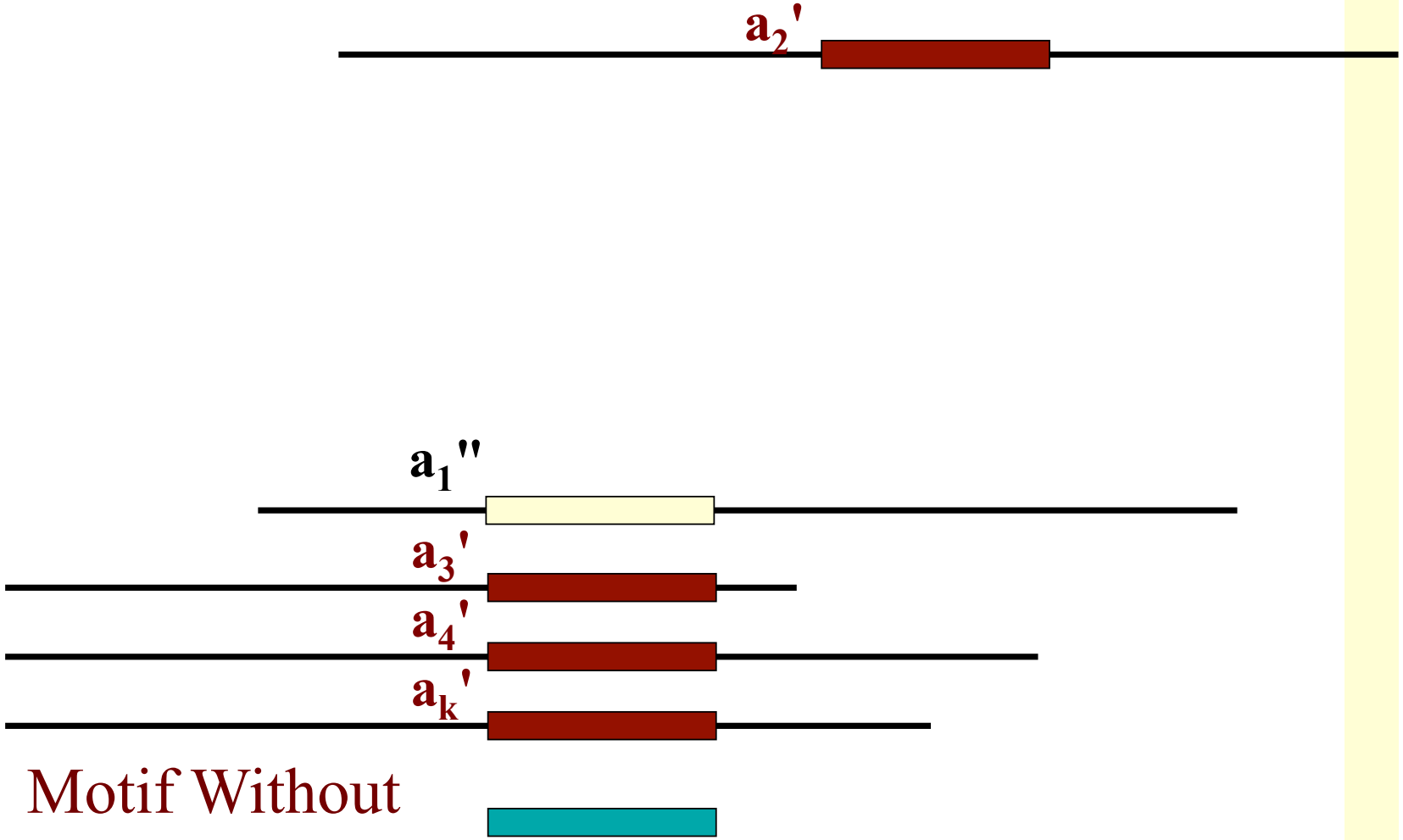




Candidate Motif

BioProspector Iterative Update

Repeat the process until convergence



Challenges for BioProspector

http://bioprospector.stanford.edu/

- Variable (0-n) motif sites per sequence
- Motif enriched only in upstream sequences, not in the whole genome
- Some motifs could have two conserved blocks separated by a variable length gap
- Motifs are not highly conserved (~50%)
- Some motifs show a palindromic symmetry
- Assign motifs a measure of statistical significance

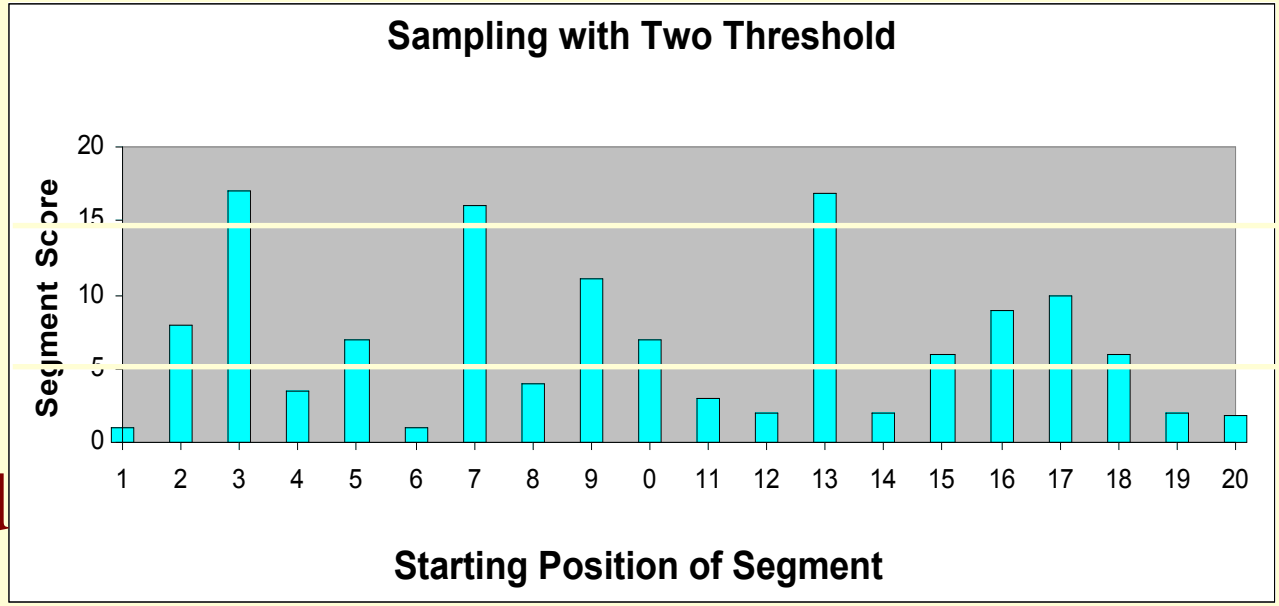
Thresholds Allow for Variable Motif Copies

- Sequences that do not have the motif
- Sequences with multiple copies of motif

Keep T_H

Sample

T_L **Discard**



BioProspector Finds Motif With Two Blocks

Two-block motifs:

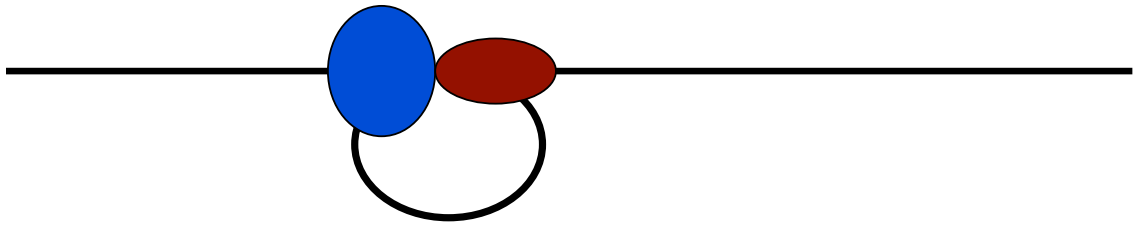
GACACATTACCTATGC TGGCCCTACGACCTCTCGC

CACAATTACCACCA TGGCGTGATCTCAGACACGGACGGC

GCCTCGATTACCGTGGTA TGGCTAGTTCTCAAACCTGACTAAA

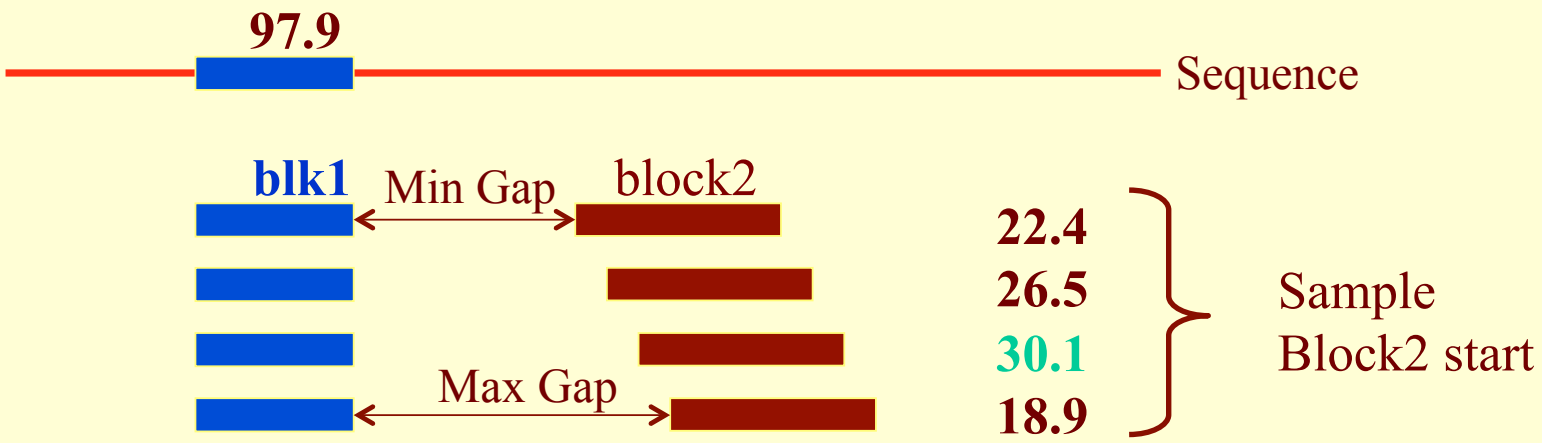
TCTCGTTAGATTACCACCCA TGGCCGTATCGAGAGCG

CGCTAGCCATTACCGAT TGGCGTTCTCGAGAATTGCCTAT



BioProspector Finds Motifs With Two Blocks

Two-block motifs



BioProspector Finds Motif With Inverse Complementary Blocks

Two-block motifs

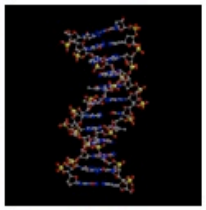
Palindrome motifs:

BioProspector Results: *B. subtilis* two-block promoter

- *B. subtilis* transcription best studied
- 136 σ^A-dependent promoter sequences [-100, 15]
- Look for $w_1 = w_2 = 5$, gap[15, 20] two-block motif
- Correctly identified motif [TTGACA, TATAAT] and 70% of all the sites
- Occasionally predicted two promoters

	"Correct" site	Second site
abrB	TTGACG	TACAAT
veg	TTGACA	TATAAT
f105	TTTACA	TACAAT

BioProspector Web Server: http://bioprospector.stanford.edu/



Overview
BioProspector finds enriched sequence motifs

Motif Finding
Search for interesting motifs in your sequences on our server

Input Format
How to specify the input parameters

Output Explanations
How to understand the output email we send you

Reference
Proc Pac Symp Biocomput 2001;:127-38

Contacts
People behind the project

SeqMotifs
See other motif finding algorithms we have developed.

BioProspector

Discovering Conserved DNA Motifs in Upstream Regulatory Regions of Co-Expressed Genes

Xiaole Liu, Jun S. Liu, Douglas L. Brutlag
Stanford Medical Informatics, Stanford University

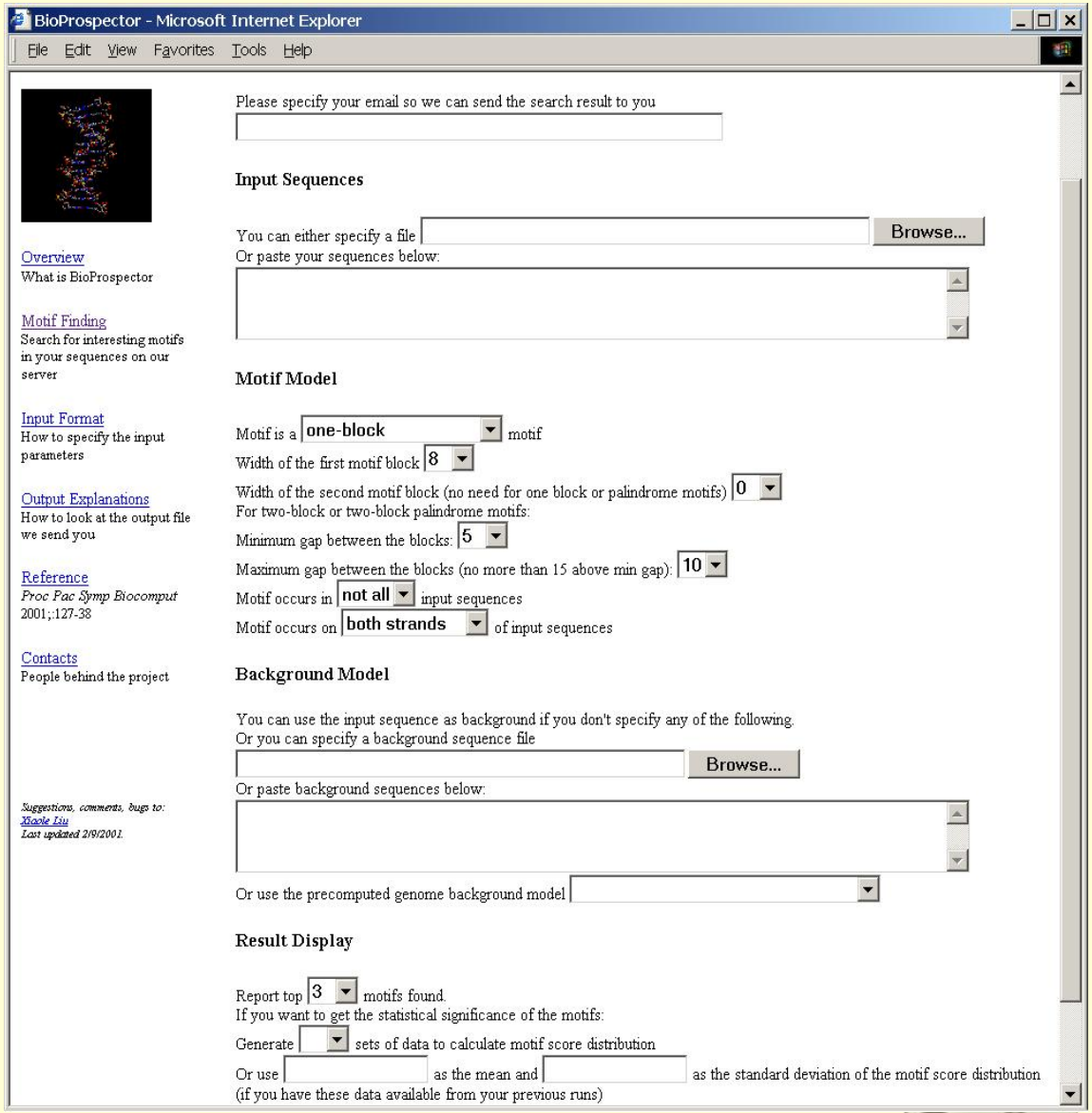
The development of high throughput genome sequencing and gene expression techniques gives rise to the demand for data-mining tools. BioProspector, a C program using a Gibbs sampling strategy, examines the upstream region of genes in the same gene expression pattern group and looks for regulatory sequence motifs. BioProspector uses Markov background to model the base dependencies of non-motif bases, which greatly improved the specificity of the reported motifs. The parameters of the Markov background model are either estimated from user-specified sequences or pre-computed from the whole genome sequences. A new motif scoring function is adopted to allow each input sequences to contain zero to multiple copies of the motif. In addition, BioProspector can model gapped motifs and motifs with palindromic patterns, which are prevalent motif patterns in prokaryotes. All these modifications greatly improve the performance of the program. Besides showing preliminary success in finding the binding motifs for *S. cerevisiae* RAP1, *B. subtilis* RNA polymerase, and *E. coli* CRP, we have used BioProspector to find s54 motif from *M. xanthus* genome, many *B. subtilis* motifs from DBTBS collection of promoters, and motifs from yeast expression data.

BioProspector requires the user to specify a motif width. Recently, JS Liu and his student have developed an algorithm BioOptimizer to automatically adjust a user-specified motif width to optimize the motif's information. The program can be downloaded from: http://www.people.fas.harvard.edu/~junliu/BioOptimizer/.

Obtaining a local copy of BioProspector:
BioProspector is free-of-charge to academia. Please check out:
Brutlag Bioinformatics Group Software Download and
Academic License Instructions for details.

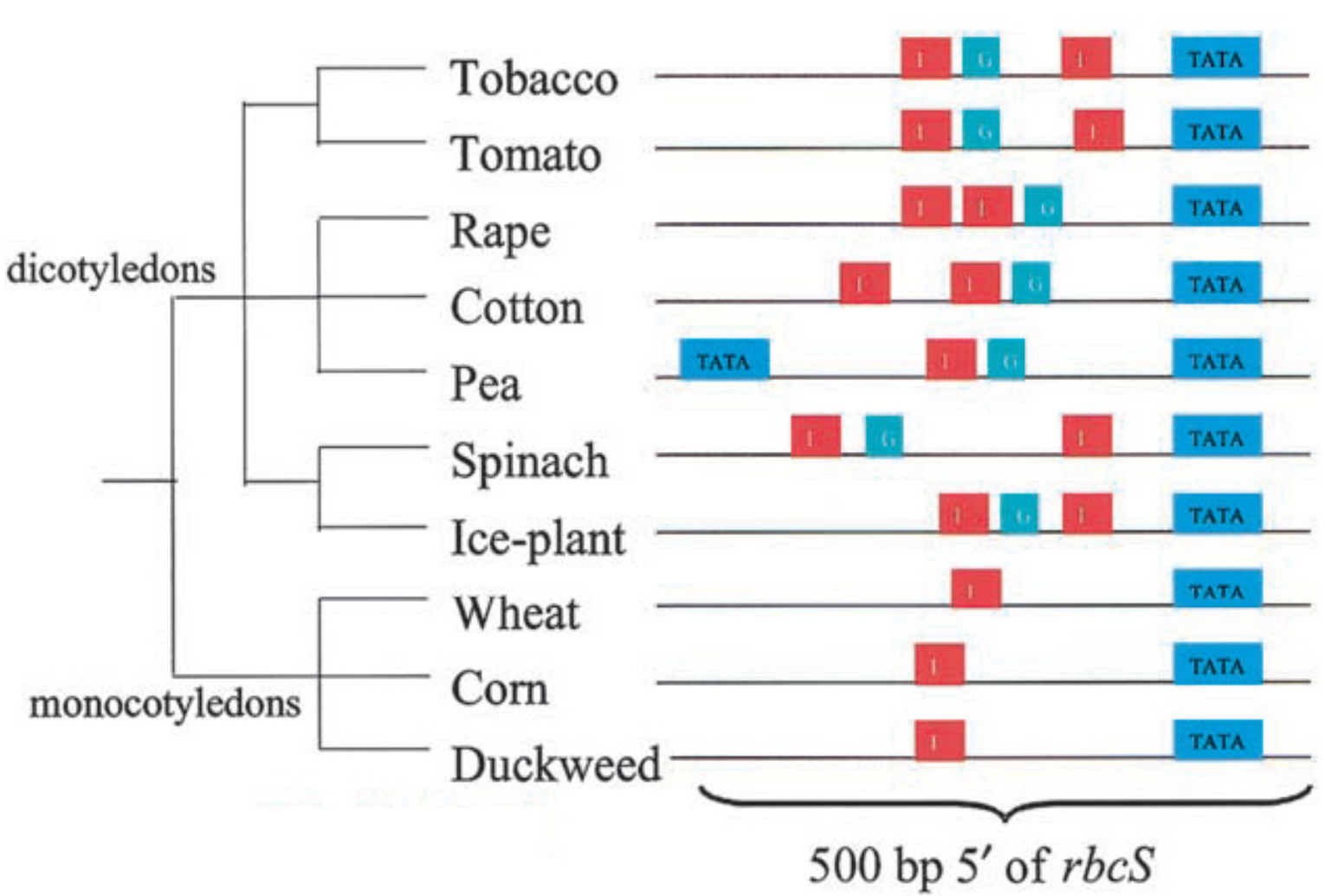
Reference:
Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*. 2001;:127-38.

BioProspector Web Server: http://bioprospector.stanford.edu/



Compare Prospector

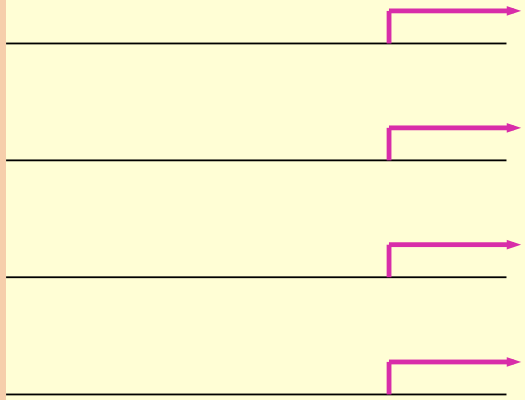
http://compareprospector.stanford.edu/

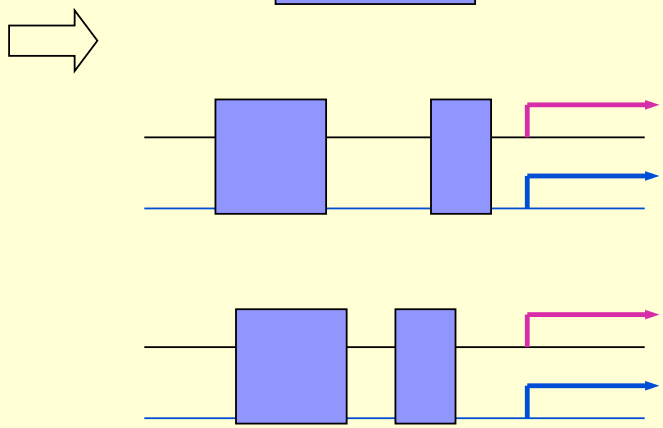


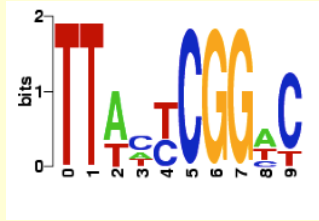
Liu et al, 2004, Genome Res 14(3): 451-458.

Compare Prospector

http://compareprospector.stanford.edu/







Motif

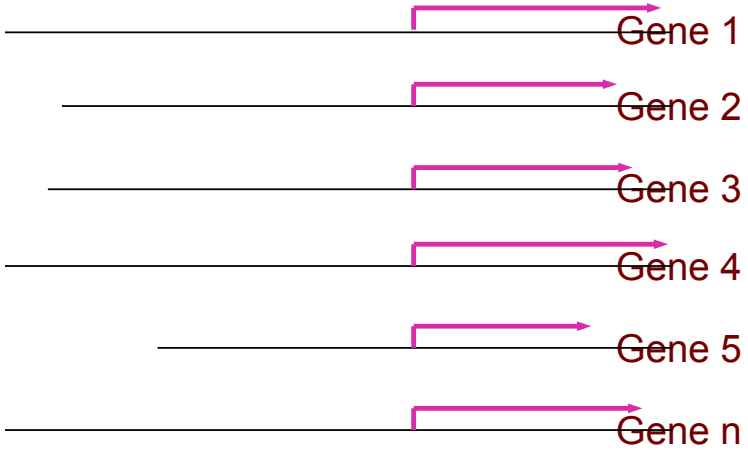
1 kb

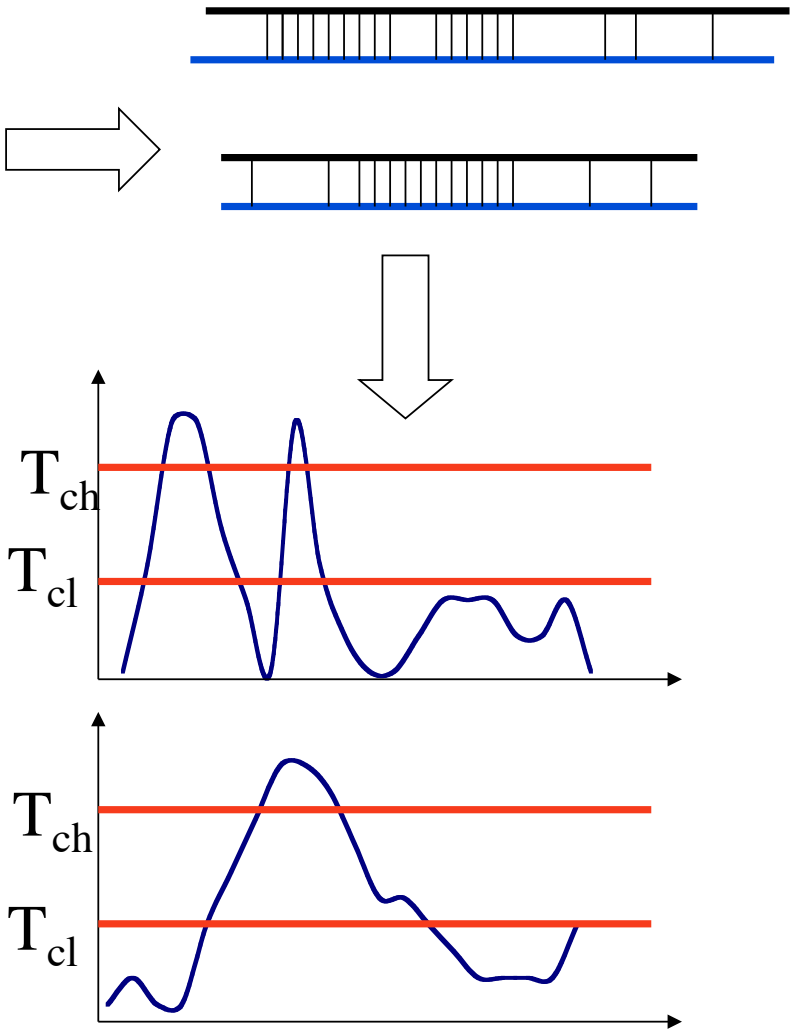
Regions conserved between two species

Liu et al, 2004, Genome Res 14(3): 451-458

Compare Prospector

http://compareprospector.stanford.edu/





Biased sampling:

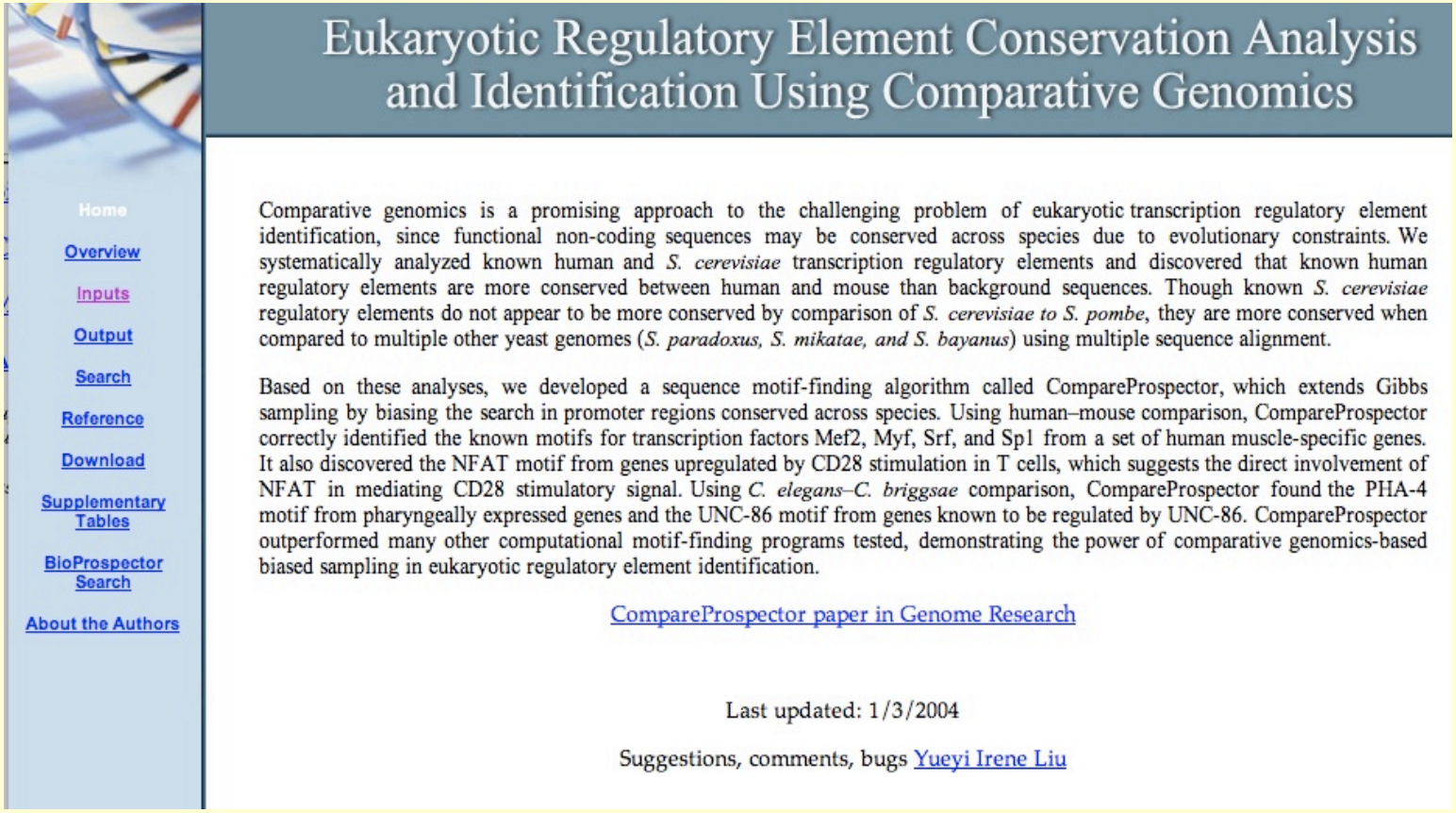
Initial iterations: T_{ch}

Later iterations: T_{cl}

Liu et al, 2004, Genome Res 14(3): 451-8,

Compare Prospector

http://compareprospector.stanford.edu/



Home
Overview
Inputs
Output
Search
Reference
Download
Supplementary Tables
BioProspector Search
About the Authors

Eukaryotic Regulatory Element Conservation Analysis and Identification Using Comparative Genomics

Comparative genomics is a promising approach to the challenging problem of eukaryotic transcription regulatory element identification, since functional non-coding sequences may be conserved across species due to evolutionary constraints. We systematically analyzed known human and *S. cerevisiae* transcription regulatory elements and discovered that known human regulatory elements are more conserved between human and mouse than background sequences. Though known *S. cerevisiae* regulatory elements do not appear to be more conserved by comparison of *S. cerevisiae to S. pombe*, they are more conserved when compared to multiple other yeast genomes (*S. paradoxus, S. mikatae, and S. bayanus*) using multiple sequence alignment.

Based on these analyses, we developed a sequence motif-finding algorithm called CompareProspector, which extends Gibbs sampling by biasing the search in promoter regions conserved across species. Using human–mouse comparison, CompareProspector correctly identified the known motifs for transcription factors Mef2, Myf, Srf, and Sp1 from a set of human muscle-specific genes. It also discovered the NFAT motif from genes upregulated by CD28 stimulation in T cells, which suggests the direct involvement of NFAT in mediating CD28 stimulatory signal. Using *C. elegans–C. briggsae* comparison, CompareProspector found the PHA-4 motif from pharyngeally expressed genes and the UNC-86 motif from genes known to be regulated by UNC-86. CompareProspector outperformed many other computational motif-finding programs tested, demonstrating the power of comparative genomics-based biased sampling in eukaryotic regulatory element identification.

CompareProspector paper in Genome Research

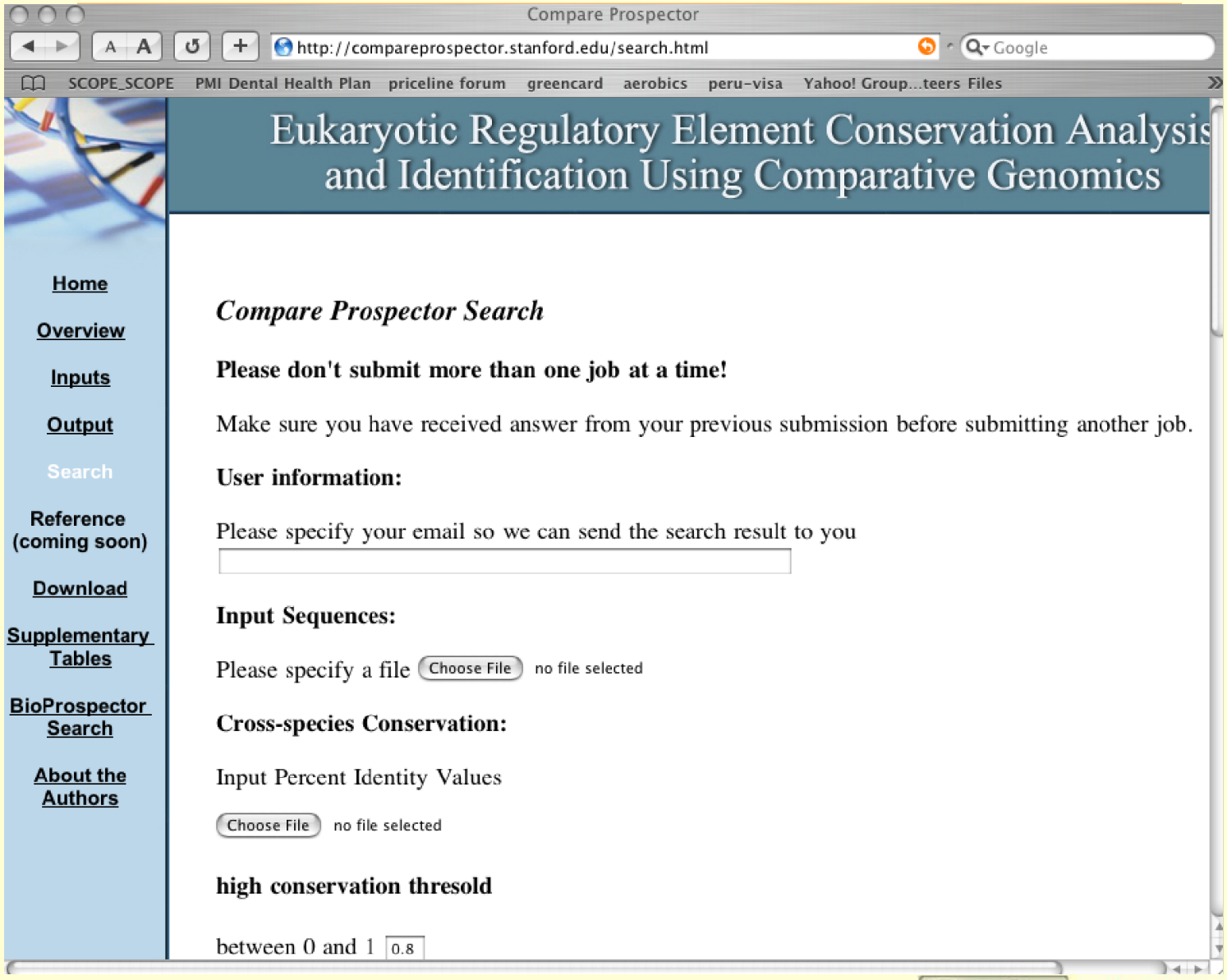
Last updated: 1/3/2004

Suggestions, comments, bugs Yueyi Irene Liu

(Liu Y et al, Nucleic Acids Res 32:W204-7)

Compare Prospector

http://compareprospector.stanford.edu/



(Liu Y et al, Nucleic Acids Res 32:W204-7)

Yeast Rap1 Sequences

- Chromatin immunoprecipitation + microarray (ChIP-on-chip, ChIP-array, IP) experiment

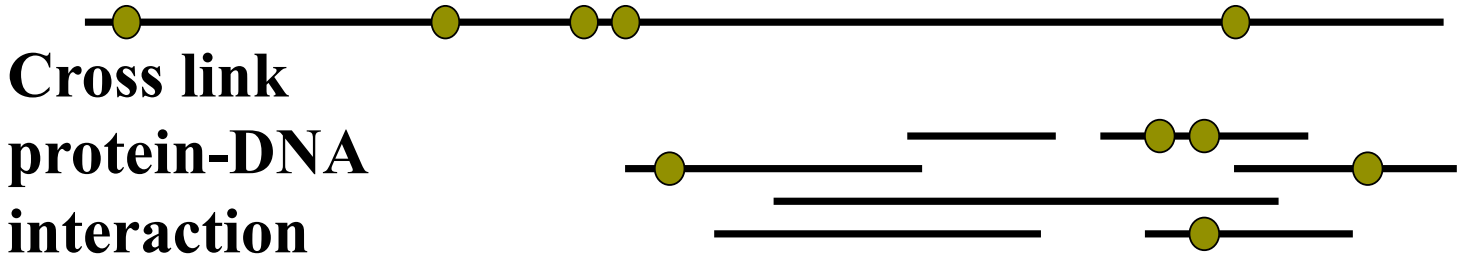
Yeast Rap1 Sequences

- Chromatin immunoprecipitation + microarray (ChIP-on-chip, ChIP-array, IP) experiment

Cross link protein-DNA interaction

Yeast Rap1 Sequences

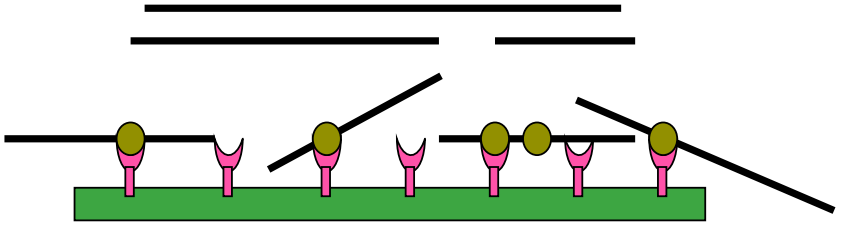
- Chromatin immunoprecipitation + microarray (ChIP-on-chip, ChIP-array, IP) experiment



Yeast Rap1 Sequences

- Chromatin immunoprecipitation + microarray (ChIP-on-chip, ChIP-array, IP) experiment

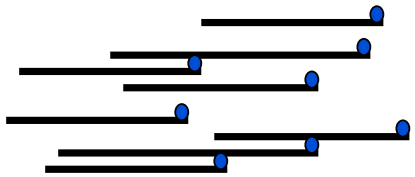
Immunoprecipitation



Yeast Rap1 Sequences

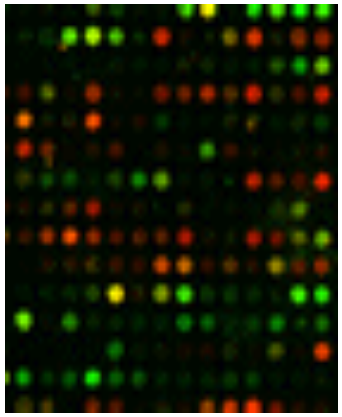
- Chromatin immunoprecipitation + microarray (ChIP-on-chip, ChIP-array, IP) experiment

PCR amplify and label DNA



Yeast Rap1 Sequences

- Chromatin immunoprecipitation + microarray (ChIP-on-chip, ChIP-array, IP) experiment

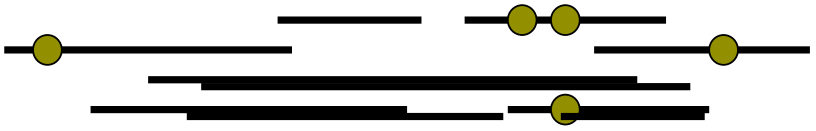


Hybridize with microarray and measure reading

Yeast Rap1 Sequences

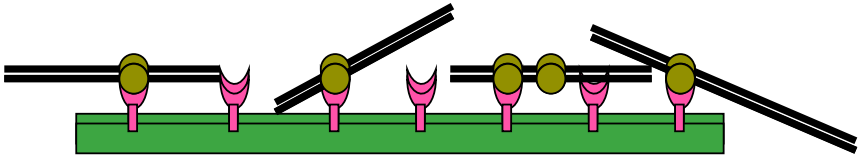
- Chromatin immunoprecipitation + microarray (ChIP-on-chip, ChIP-array, IP) experiment

Cross link protein-DNA interaction



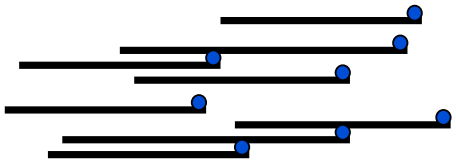
Shear DNA

Immunoprecipitation

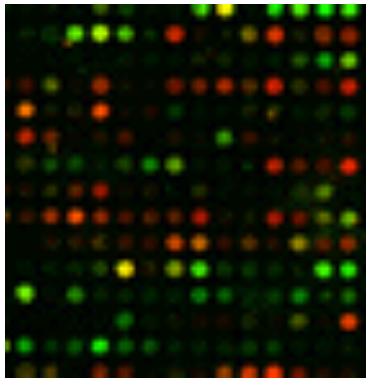


Purify DNA

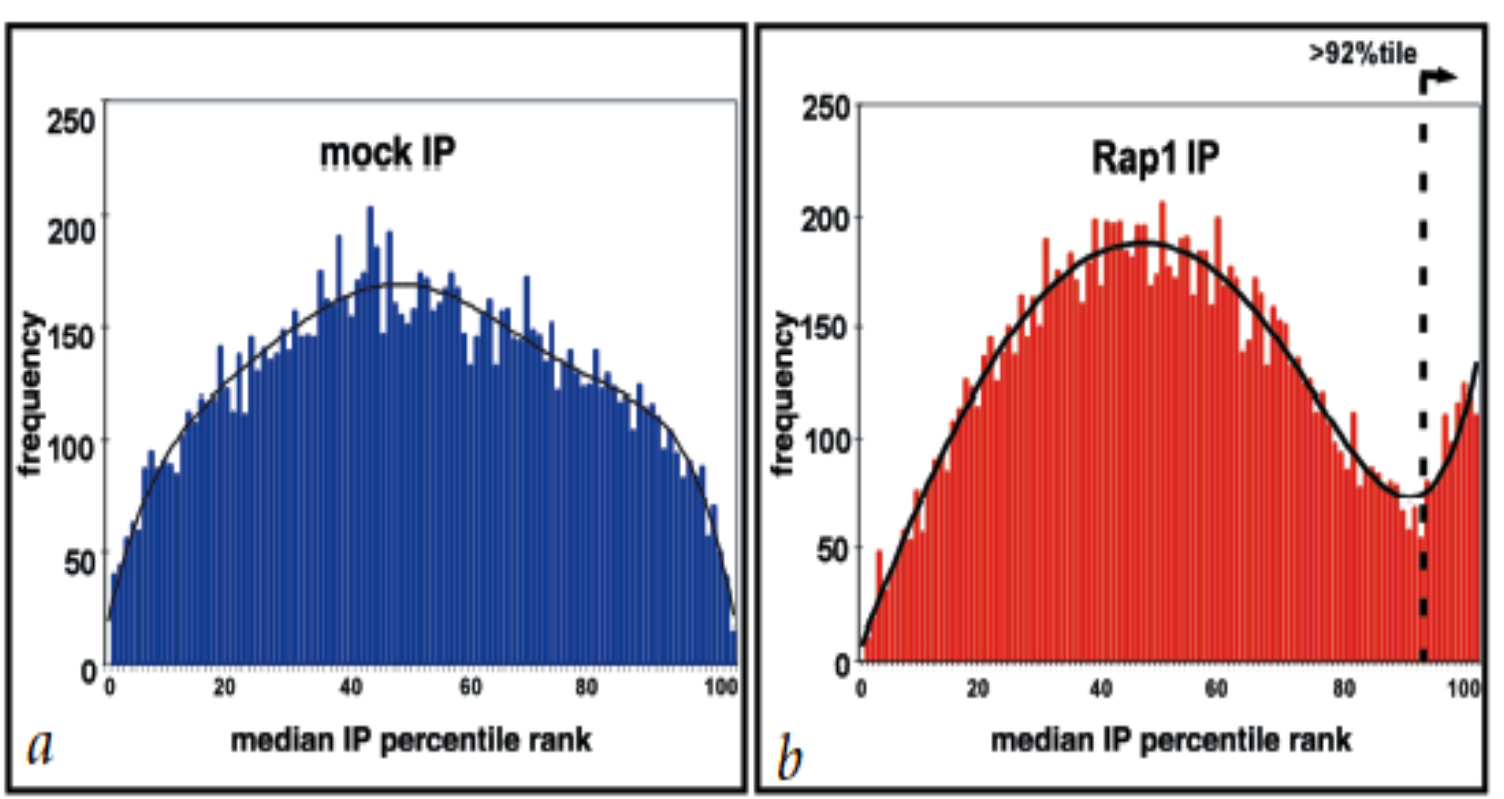
PCR amplify and label DNA



Hybridize with microarray and measure reading



Chromatin Immune Precipitation



Yeast Rap1 Sequences

- Chromatin immunoprecipitation + microarray (ChIP-on-chip, ChIP-array, IP) experiment
- Rap1 IP Enriched 727 DNA fragments
  - 45% are intergenic
  - Average length 1-2 KB
  - Some are false positives
  - Some have multiple Rap1 sites

Useful Insights

- In ChIP-array experiments, highly enriched sequences are usually the real targets
- Transcription factor binding sites occurs more abundantly in these real targets
- Search TF sites from high-confidence sequences first before examine the rest sequences?

Motif Discovery Scan (MDscan)

MDscan Algorithm: Define *m*-matches

For a given *w*-mer and any other random *w*-mer

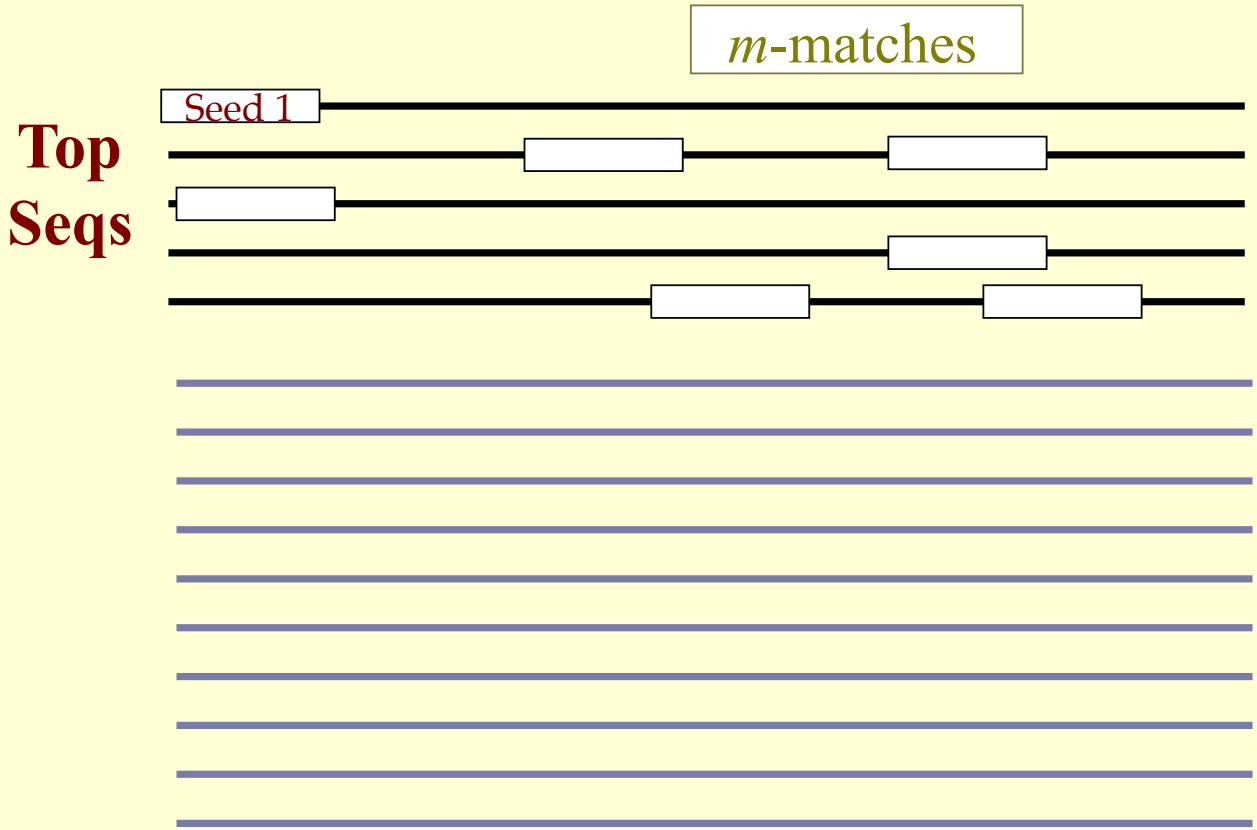
```
TGTAACGT      8-mer
TGTAACGT      matched 8
AGTAACGT      matched 7
TGCAACAT      matched 6
TGACACGG      matched 5
AATAACAG      matched 4
```

m-matches for an 8-mer

Pick a reasonable *m*, e.g. in yeast

w	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
m	5	6	6	7	7	8	8	9	9	10	10	10	11	11	12

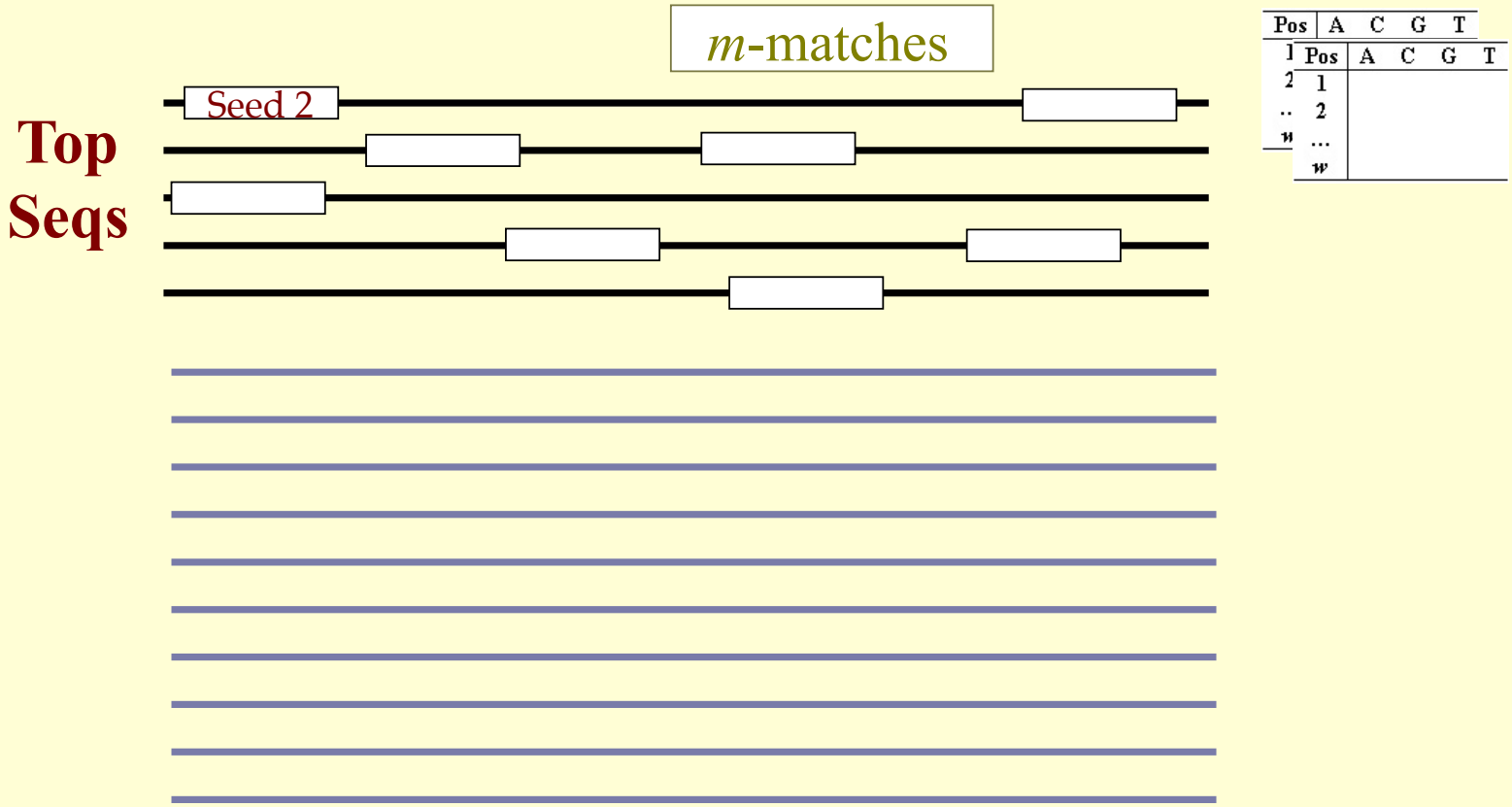
MDscan Algorithm: Finding candidate motifs



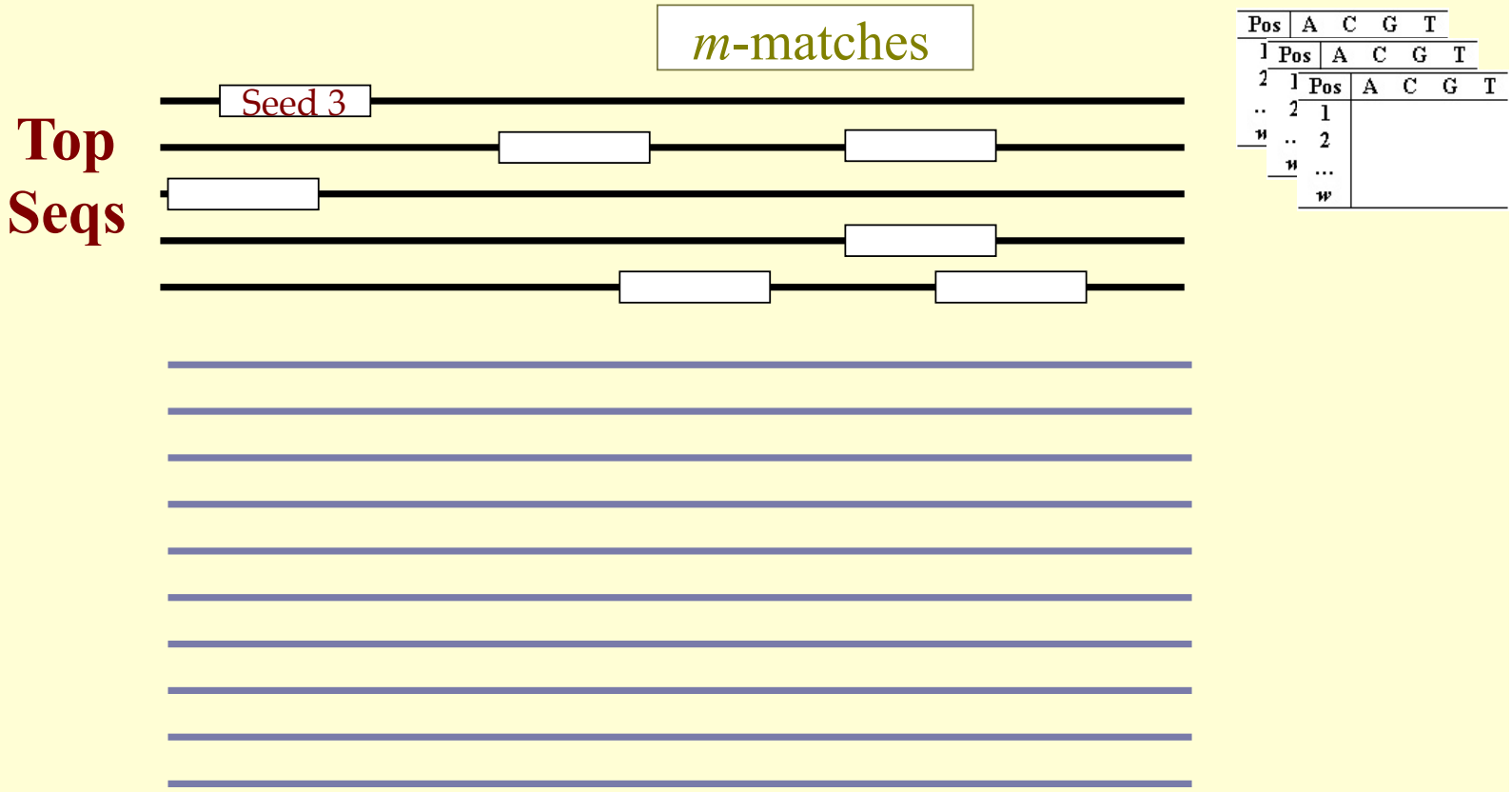
Pos	A	C	G	T
1				
2				
…				
w				

All IP enriched sequences

MDscan Algorithm: Finding candidate motifs



MDscan Algorithm: Finding candidate motifs



All IP enriched sequences

MDscan Algorithm: Scanning sequences with top motifs

- Keep 30-50 top scoring candidate motifs:

$$\frac{\log(x_m)}{w} \times [\sum_{i=1}^{w} \sum_{j=A}^{T} p_{ij} \log p_{ij} - \frac{1}{x_m} \sum_{every\ segment\ s} \log(p_0(s))]$$

Motif Signal Abundance

Conserved Positions

Specificity (unlikely in genome)

MDscan Algorithm: Scanning sequences with top motifs

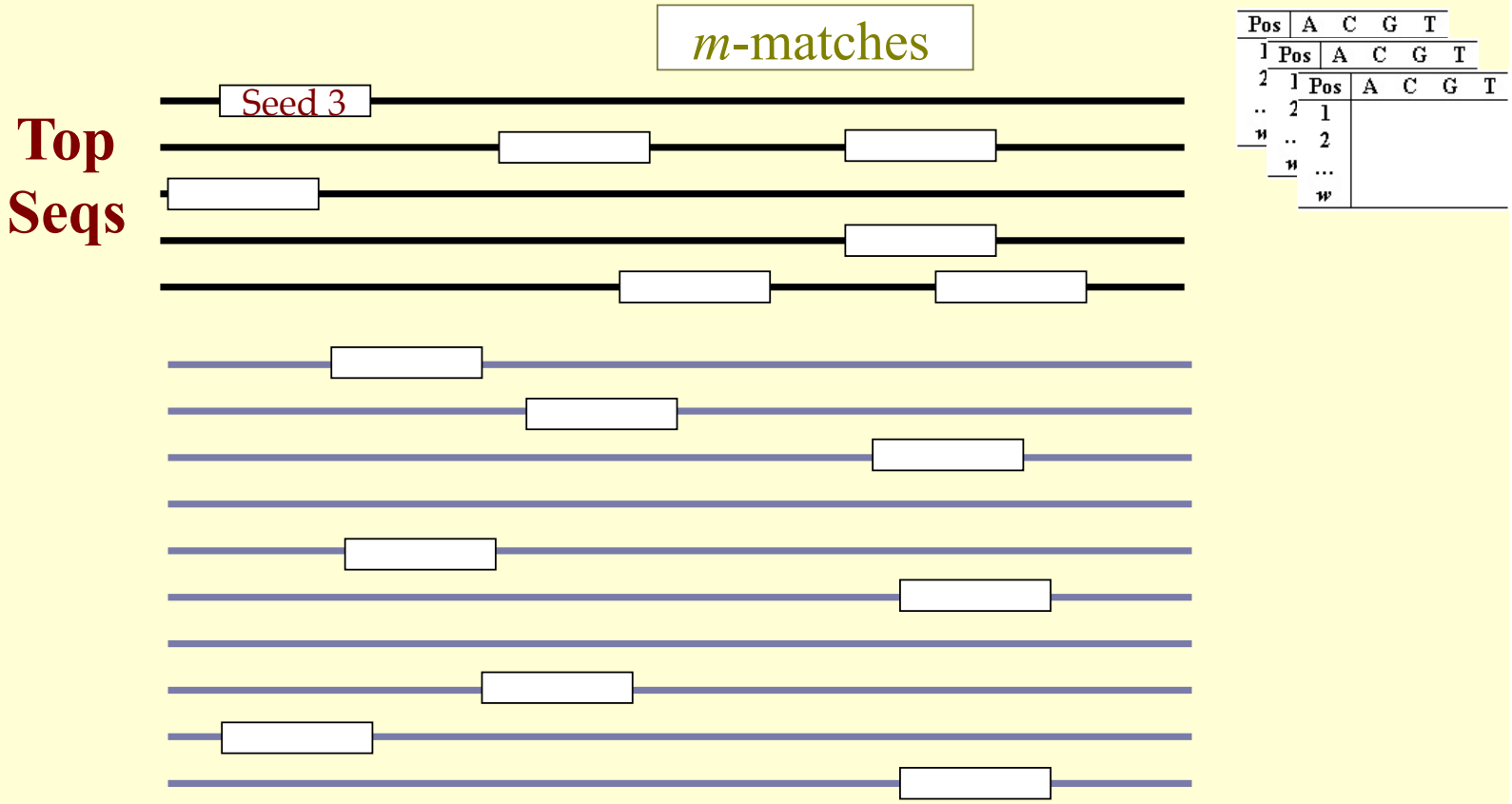
- Keep 30-50 top scoring candidate motifs:

$$\frac{\log(x_m)}{w} \times [\sum_{i=1}^{w} \sum_{j=A}^{T} p_{ij} \log p_{ij} - \frac{1}{x_m} \sum_{every\ segment\ s} \log(p_0(s))]$$

Motif Signal Abundance	Conserved Positions	Specificity (unlikely in genome)

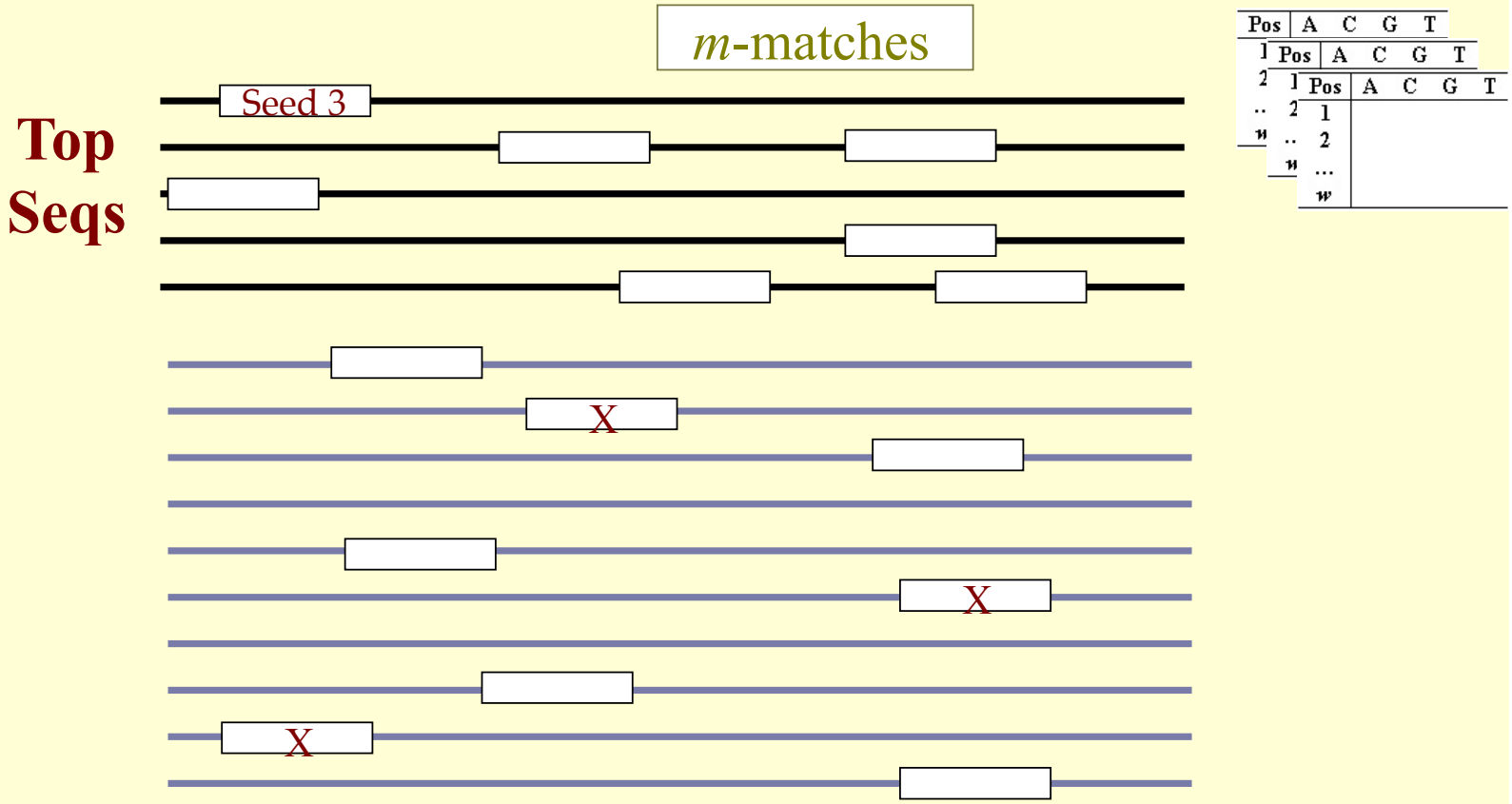
- Scan the rest of the sequences with the candidate motifs

MDscan Algorithm: Finding All Motif Instances



All IP enriched sequences

MDscan Algorithm: Refine the motifs



All IP enriched sequences

MDscan Simulation

- Nine motif matrix models with 3 widths and 3 degeneracy

Motif width, (consensus)	Motif Information Content (strength)		
	S1	S2	S3
W8, (GACTACCA)	0.772	0.667	0.551
W12, (GACTACCATGGA)	0.654	0.582	0.522
W16, (AGGATCTAATGATCCT)	0.577	0.520	0.461

W8S1 More Conserved

GACTCCCA
GATTGCCT
GGCTACCT
GACTACCA
GAGTACCA
GACTATCT
GAGTACCA
GGCTCCCA
GACTCCCA

W8S3 Less Conserved

GACTCCGA
GGGAACCA
GCTTCCAA
GACTACCA
CAGTACGA
GGCTAGCA
GACTGCCG
GACTACCA
GACTCCCG

MDscan Simulation

Each test set:

- 100 sequences of 600 bases from yeast intergenic
- Motif segments generated and inserted according to the following abundance:

Expected copies of motif segments	A1	A2	A3	A4
Among top 5 sequences	3	2.5	2	1.5
Among middle 35 sequences	1.4	1.1	0.8	0.5
Among last 60 sequences	0.4	0.3	0.2	0.1
Total expected motif segments	88	69	50	31

Higher confidence
Motif more abundant

MDscan Simulation

- 100 tests for
  - 3 widths
  - 3 strengths
  - 4 abundances

} 3600 tests

MDscan Simulation

- 100 tests for
  - 3 widths
  - 3 degeneracy
  - 4 abundance

  } 3600 tests

- MDscan speed
  - 3 X Consensus
  - 14 X BioProspector
  - 27 X AlignACE

MDscan

MDscan Simulation Accuracy
$w = 8$

100 Tests in	MDscan		BioProspector		Consensus		AlignACE	
	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif
W8S1A1	100	1.01	100	1.00	78	1.10	86	1.08
W8S1A2	100	1.01	100	1.00	58	1.34	52	1.08
W8S1A3	89	1.15	95	1.04	32	1.94	10	1.10
W8S1A4	47	1.89	54	1.72	6	2.00	0	N/A
W8S2A1	91	1.05	99	1.01	38	1.16	34	1.03
W8S2A2	88	1.12	91	1.16	22	1.64	6	1.00
W8S2A3	62	1.68	66	1.42	7	2.14	0	N/A
W8S2A4	25	1.92	21	1.62	2	2.50	0	N/A
W8S3A1	82	1.24	84	1.32	10	2.50	3	1.00
W8S3A2	60	1.65	64	1.31	5	1.60	0	N/A
W8S3A3	28	2.07	30	1.90	4	1.50	0	N/A
W8S3A4	6	1.33	5	1.80	1	3.00	0	N/A

MDscan

MDscan Simulation Accuracy
$w = 12$

100 Tests in	MDscan		BioProspector		Consensus		AlignACE	
	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif
W12S1A1	100	1.00	100	1.00	100	1.00	99	1.00
W12S1A2	100	1.00	100	1.00	98	1.06	97	1.06
W12S1A3	99	1.00	98	1.00	81	1.17	77	1.16
W12S1A4	85	1.07	76	1.18	32	1.56	34	1.24
W12S2A1	100	1.00	100	1.00	95	1.02	99	1.07
W12S2A2	95	1.01	100	1.00	82	1.20	82	1.13
W12S2A3	88	1.05	82	1.05	39	1.56	42	1.21
W12S2A4	62	1.24	29	1.31	14	1.71	6	1.50
W12S3A1	99	1.00	100	1.00	83	1.23	88	1.22
W12S3A2	89	1.02	97	1.04	44	1.43	63	1.27
W12S3A3	70	1.21	73	1.25	15	2.40	17	1.24
W12S3A4	32	1.87	13	1.85	4	3.25	2	1.50

MDscan

MDscan Simulation Accuracy
$w = 16$

100 Tests in	MDscan		BioProspector		Consensus		AlignACE	
	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif	Correct Found as Top 5	Avg Rank of Correct Motif
W16S1A1	100	1.00	100	1.00	100	1.00	100	1.00
W16S1A2	100	1.00	100	1.00	100	1.00	94	1.00
W16S1A3	100	1.00	100	1.00	92	1.07	71	1.20
W16S1A4	97	1.00	83	1.06	50	1.44	24	2.04
W16S2A1	100	1.00	100	1.00	99	1.00	96	1.00
W16S2A2	100	1.00	100	1.00	94	1.07	86	1.06
W16S2A3	95	1.00	100	1.02	64	1.09	51	1.76
W16S2A4	91	1.03	68	1.26	24	1.54	22	2.05
W16S3A1	100	1.00	100	1.00	94	1.05	89	1.12
W16S3A2	100	1.05	98	1.00	76	1.24	72	1.54
W16S3A3	92	1.00	86	1.06	40	1.43	35	2.03
W16S3A4	63	1.17	27	1.52	13	1.92	2	2.00

MDscan

MDscan Biological Tests

- Gal4 & Ste12 [Ren *et al. Science* 2000]
  - Gal4: galactose metabolism

Biological test 1 [Ren *et al.* 2001]	Published motif consensus	MDscan results and ranks
Gal4 (23 sequences)	$CGGN_{11}CCG$ [Marmorstein *et al.* 1992]	* CGGAGCACTCTGGTCCG 1 * CGGTCCACTGTGGTCCG 2 * CGGAGCACTCTGGCCCG 3 * CGGACCACTGTGGTCCG 4 * CGGAGGACTGTCCTCCG 5 * CGGAGCAGTCTGCCCCG 6 * CGGAGCACTGTCGCCCG 7, 8 * GGAGCACTGTTGACCGA 9 * CGGAGCAGTGCGGCCCG 10
Ste12 (26 sequences)	TGAAACA [Dolan *et al.* 1989]	* TGAAACA 1, 2, 5, 9 AAACCAA 3 * GAAACAA 4 * CTGAAAC 6 * TTGAAAC 7 * TGCAACA 8 AAACAAA 10

MDscan

MDscan Biological Tests

- SBF & MBF [Iyer *et al. Nature* 2001]
  - SBF: Swi4 + Swi6 budding, membrane, cell wall

Biological test 2 [Iyer *et al.* 2001]	Published motif consensus	MDscan results and ranks
SBF (163 sequences)	CACGAAA [Spellman *et al.* 1998] CGCGAAAA [Iyer *et al.* 2001]	GACGCGA 1 AACGCGA 2 # ACGCGTA 3 # GACGCGT 4 * CGCGAAA 5, 8-10 * ACGCGAA 6 # CGCGTAA 7
MBF (87 sequences)	ACGCGT [Spellman *et al.* 1998]	* AACGCG 1 * ACGCGT 2-5, 8 * GACGCG 4, 7 ACACAC 6 ACCTAC 9 GGGTAA 10
SBF (120 SBF-non-MBF sequences)	CACGAAA CGCGAAAA	TACGCGA 1 * CGCGAAA 2-4, 6, 7, 9, 10 * TCGCGAA 5 * ACGCGAA 8
MBF (44 MBF-non-SBF sequences)	ACGCGT	* ACGCGT 1, 2, 4-6, 8 * AACGCG 3, 7, 9, 10

MDscan

MDscan Biological Tests

- Rap1 [Lieb *et al. Nature Genetics* 2001]

Biological test 3 [Lieb *et al.* 2001]	Published motif consensus	MDscan results and ranks	
Rap1 (727 sequences)	Same as below	* CACACACACACAC	1–10
Rap1 (719 sequence, excluding the 8 sequences with CA repeats)	RTRCACCCANNCMCC [Graham & Chambers 1994] WACAYCCRTACATY [Lascaris *et al.* 1999] RMAYCCRMNCAYY [Buchman *et al.* 1988] RMACCCANNCAYY [Buchman *et al.* 1988] ACACCCAYACAYYY [Idrissi & Pina 1999]	~ GGCACTTGCATCA * ACCCATATCTCAC * ACCCATACCTCAC ^ ACCCTTACACTAC ^ CACTTACCCTACC ^ ACTTACCCTACCA ^ CTTACCCTACCAC ^ CTTACCCTACCCT	1, 3 5 6 2 4, 10 7 8 9
Rap1 (577 non-telomere sequences, excluding the 3 sequences with CA repeats)	Same as above	* ACACCCATACATC * GACACCCATACAT * TACACCCATACAT * CACCCATACATCT	1–3, 7 4–6 8 9, 10
Rap1 (142 telomere sequences, excluding the 5 sequences with CA repeats)	Same as above	~ TGCACTTGCATCA ~ GGCACTTGCCTCA ~ GCACTTGCCTCAG ^ CACTTACCCTACC ^ AACTTACCCTACC ^ ACTTACCCTACCA ^ CTTACCCTACCAT	1 2 4 3, 10 5 6, 8 7, 9

TAMO: Tools for the Analysis of Motifs

http://fraenkel.mit.edu/TAMO/

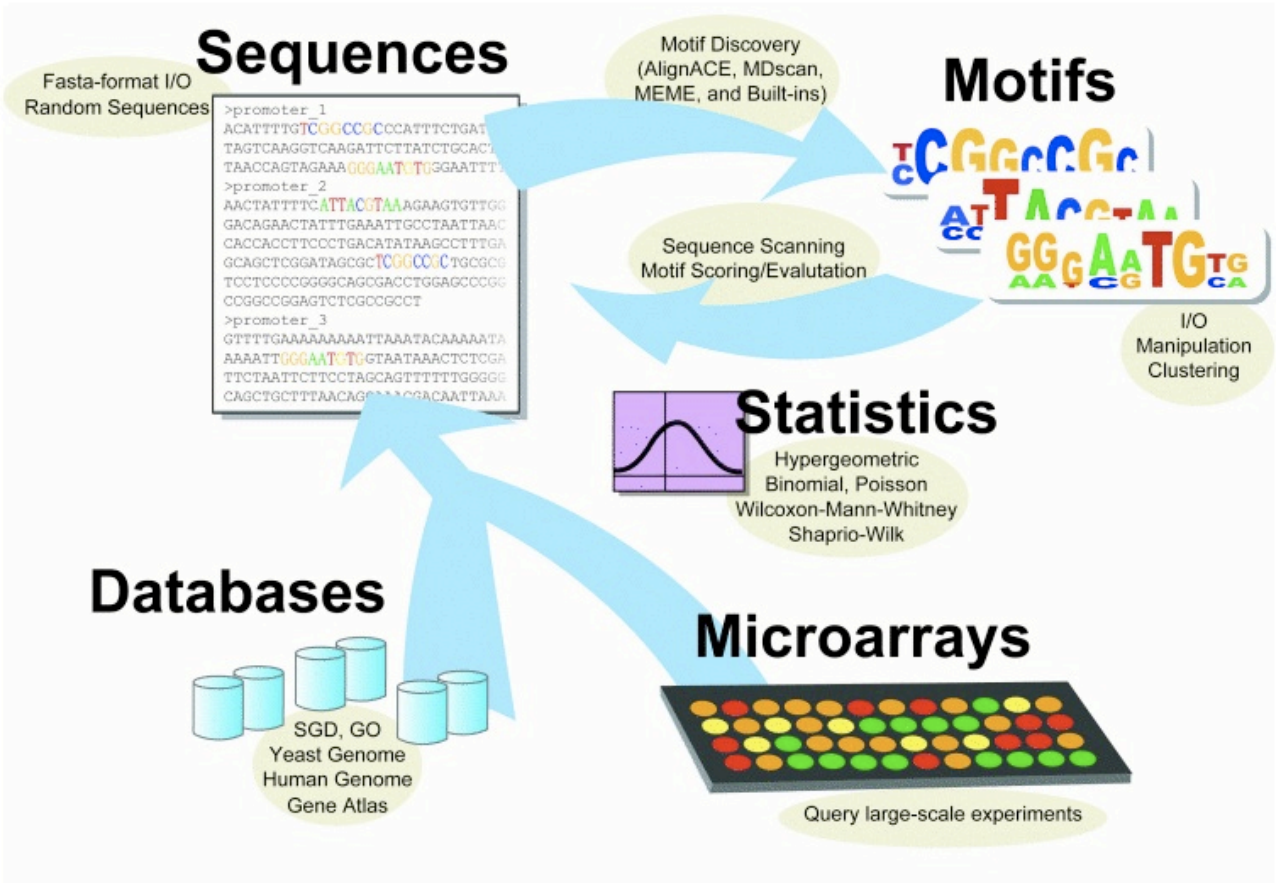
The Fraenkel Lab

Biological Engineering MIT

TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs.

Bioinformatics. 2005 Jul 15;21(14):3164-5.

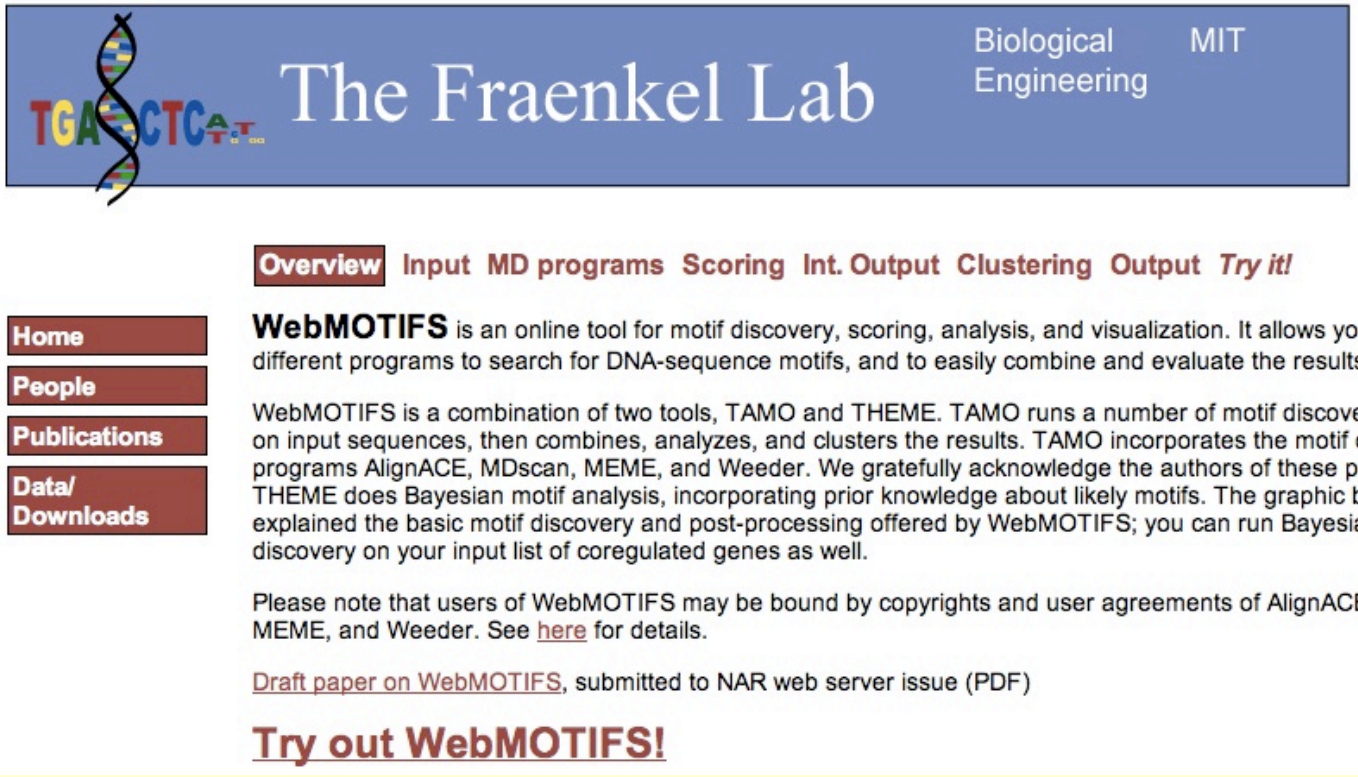
- Home
- People
- Publications
- Data/ Downloads
- Online Motif Discovery with WebMOTIFS



010

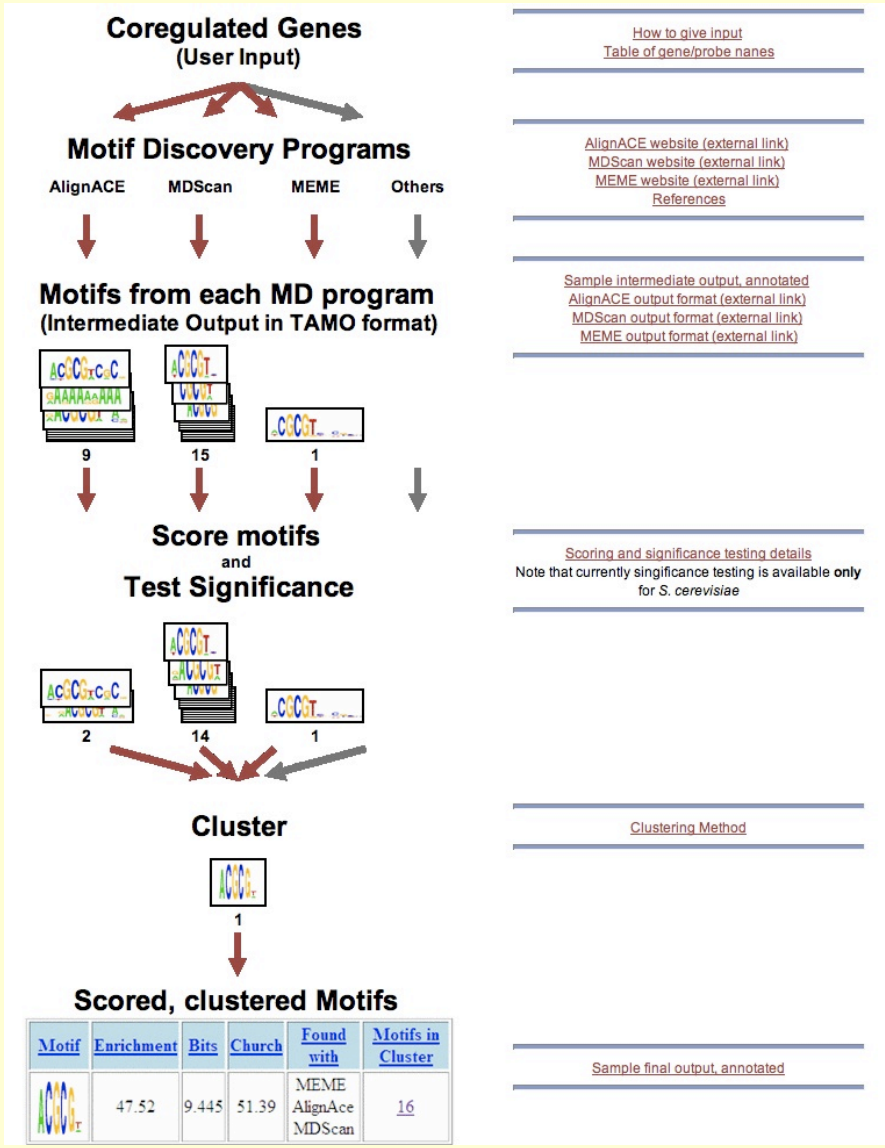
WebMotifs

http://fraenkel.mit.edu/webmotifs/



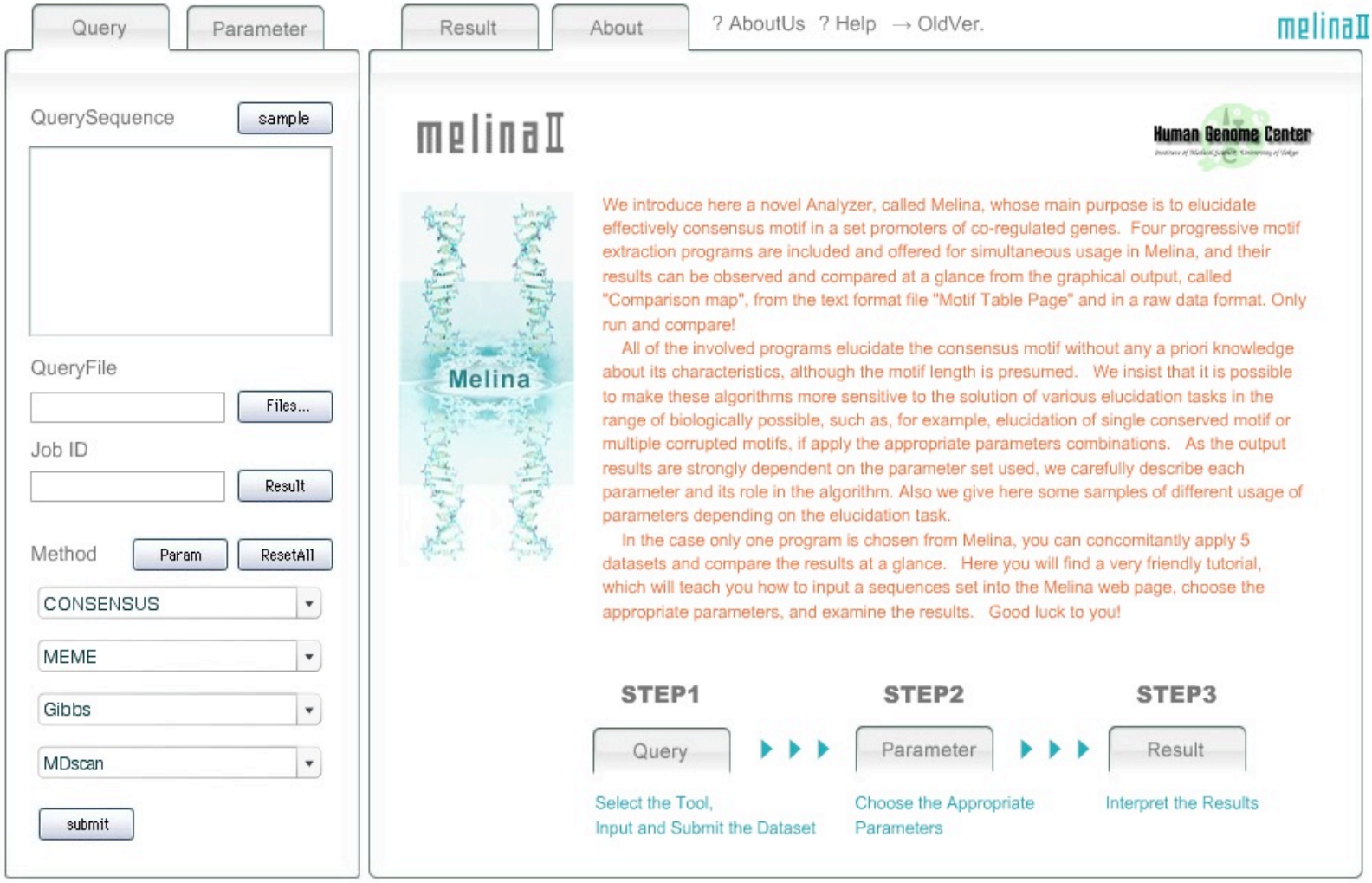
WebMotifs

http://fraenkel.mit.edu/webmotifs/



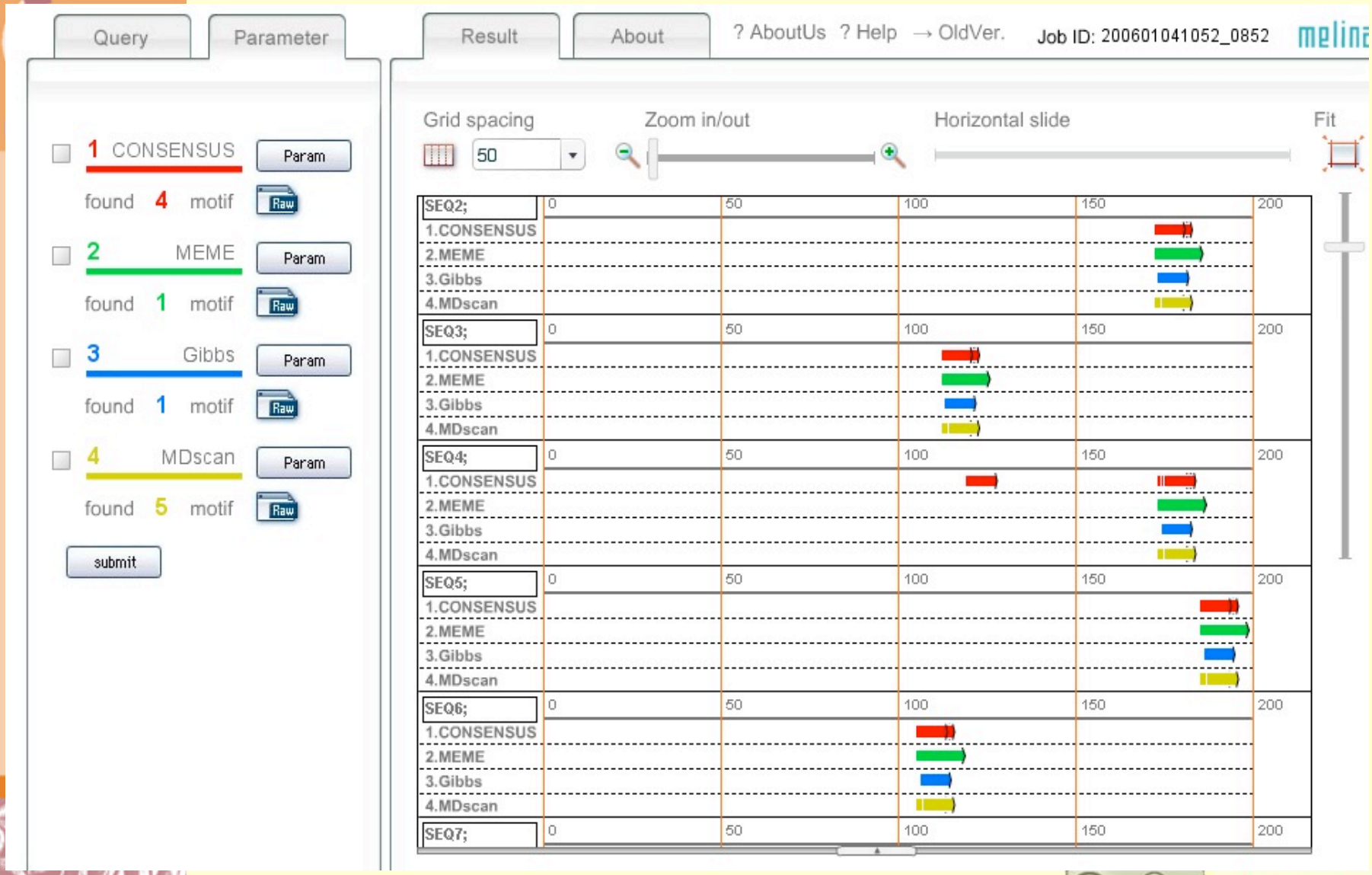
Melina: Comparing Motifs

http://melina1.hgc.jp/



Melina: Comparing Motifs

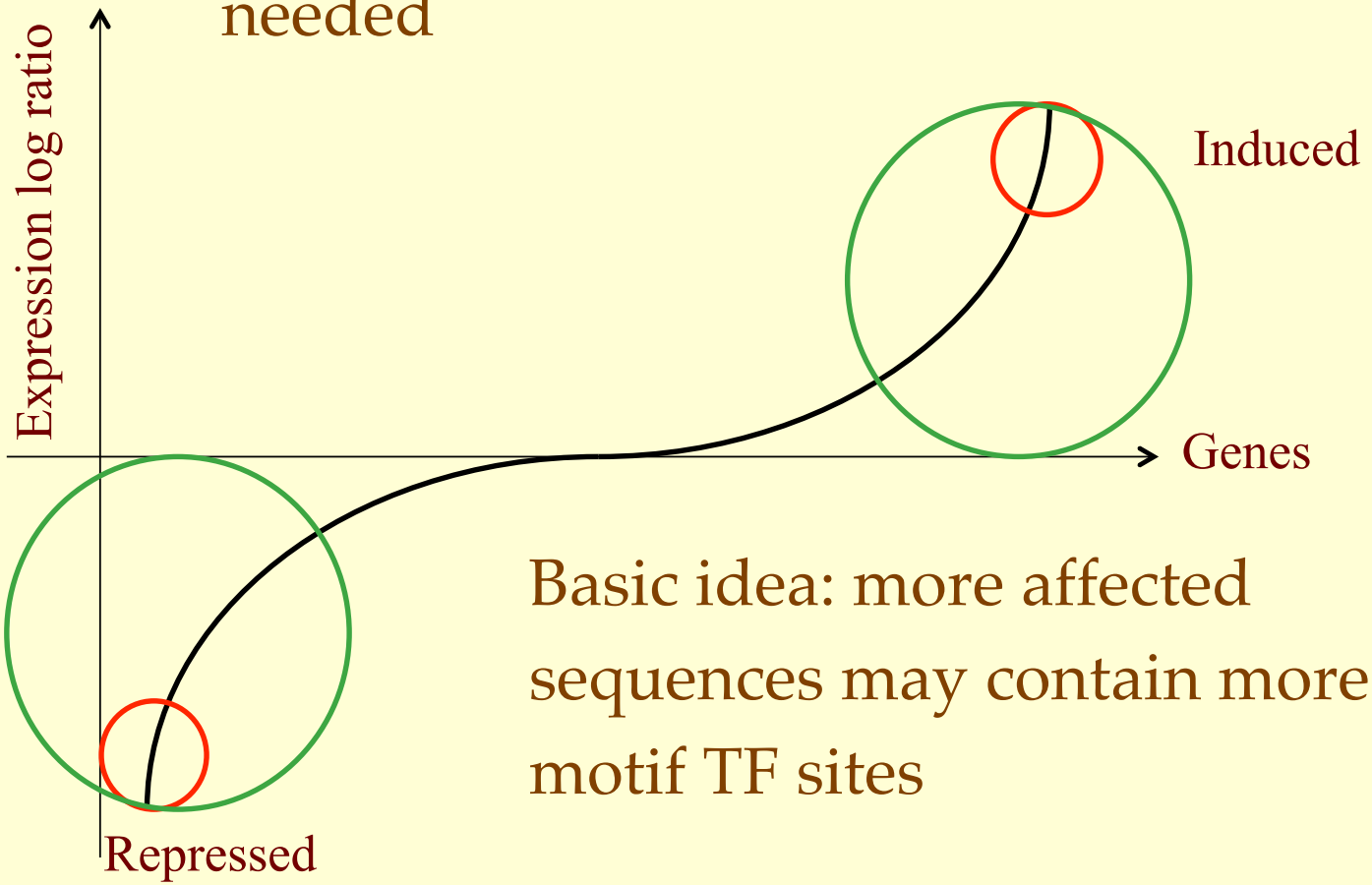
http://melina1.hgc.jp/



MDscan

Single Microarray Determination of Transcription Factor Motifs

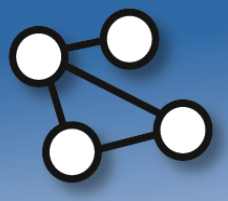
One microarray experiment, no clustering needed



Basic idea: more affected sequences may contain more motif TF sites

Summary

- BioProspector is stochastic
- BioProspector can get trapped in local maxima
- BioProspector must be run multiple times to discover the true globally optimal motif
- BioProspector is slow
- MDScan is deterministic
- MDScan always gives the same answer with the same data
- MDScan is fast
- MDScan uses rank order data to accelerate the search process and to allow it to be deterministic
- MDScan is fast enough to search intergenic regions from entire genomes.
- MDScan is not as sensitive as BioProspector

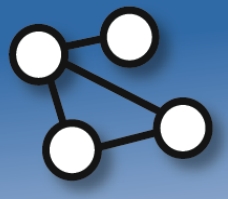


Graph-Based Methods for Representing DNA Regulatory Sites

[1]Naughton, B., E. Fratkin, S. Batzoglou and D. L. Brutlag. 2006. MotifScan - A non-Parametric Algorithm for DNA motif detection. Nucleic Acids Res 34:5730-5739.

[2]Fratkin, E., B. Naughton, D. L. Brutlag and S. Batzoglou. 2006. MotifCut: An Algorithm for Finding Regulatory Motifs. Bioinformatics:150-157.

[3]Naughton, B. SEQUENCE ANALYSIS METHODS FOR THE DETECTION OF PROMOTERS AND TRANSCRIPTION FACTOR BINDING SITES, Thesis, Biomedical Informatics Stanford University. 2006, 142 Pages.



Problems with Current Representations of DNA Motifs

- All current methods for representing DNA motifs involve either consensus sequences or probabilistic models (such as PSSMs) of the motif.

- Consensus sequences do not adequately represent the variability seen in promoters or transcription factor binding sites.

- Both consensus sequences and PSSM models assume positional independence. Neither method can accommodate correlations between positions.

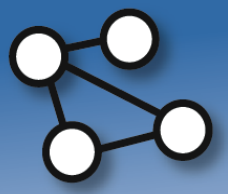
- Probabilities calculated from PSSM models can be highly misleading.

Parametric methods: a PSSM

AAA
AAA
AGG

→

	1	2	3
A	1.00	0.67	0.67
C	0.00	0.00	0.00
G	0.00	0.33	0.33
T	0.00	0.00	0.00



Parametric methods: a PSSM

AAA
AAA
AGG

→

	1	2	3
A	1.00	0.67	0.67
C	0.00	0.00	0.00
G	0.00	0.33	0.33
T	0.00	0.00	0.00

P(AAA) = 1 * 0.67 * 0.67 = 0.44

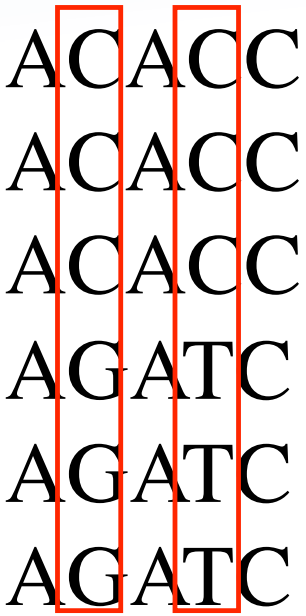
P(AGG) = 1 * 0.33 * 0.33 = 0.11

P(AAG) = 1 * 0.67 * 0.33 = 0.22

Yeast motifs

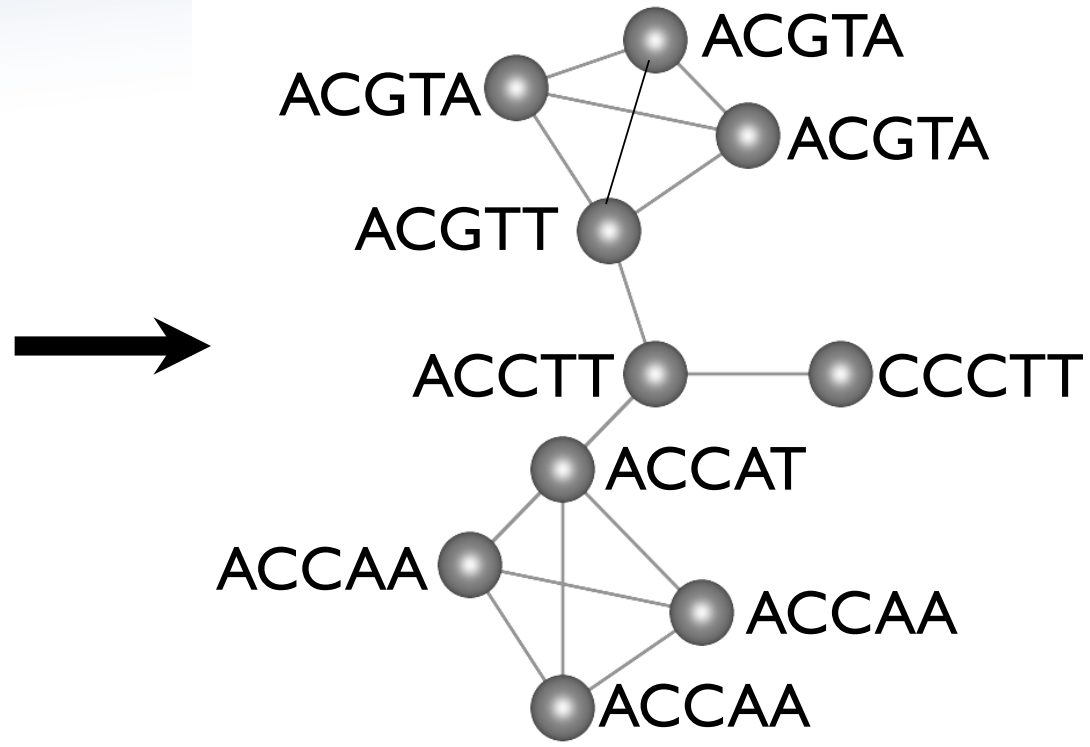
We analyzed yeast motifs for pairwise dependencies. We used a chi-square statistic to find whether two positions were correlated or not.

We found that **25%** of motifs have significantly correlated positions.

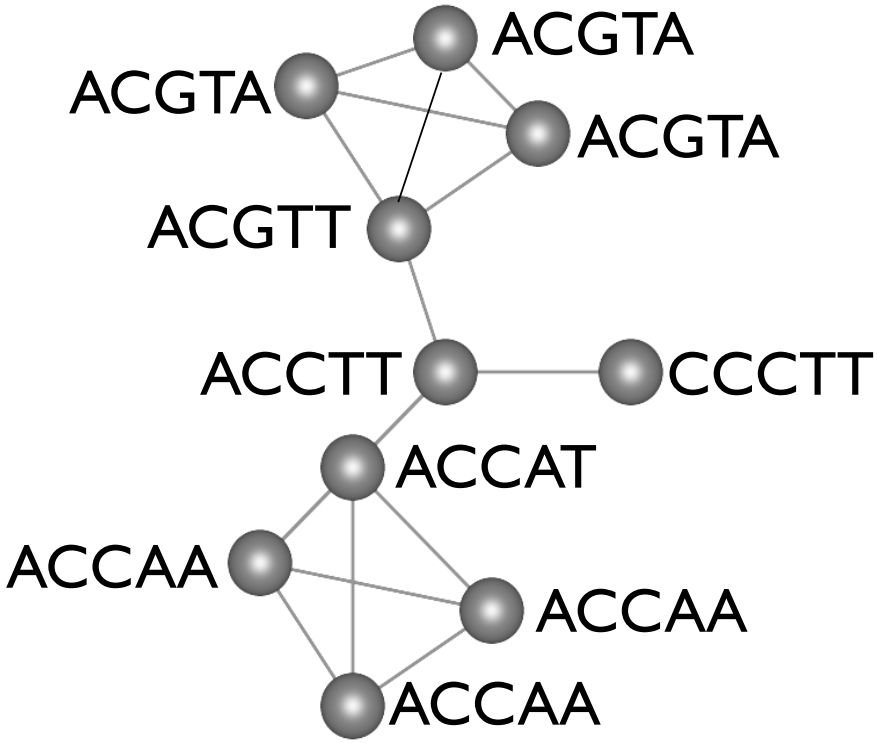


A Graph-Based Model of a Motif

ACGTA
ACGTA
ACGTA
ACGTT
ACCTT
CCCTT
ACCAT
ACCAA
ACCAA
ACCAA

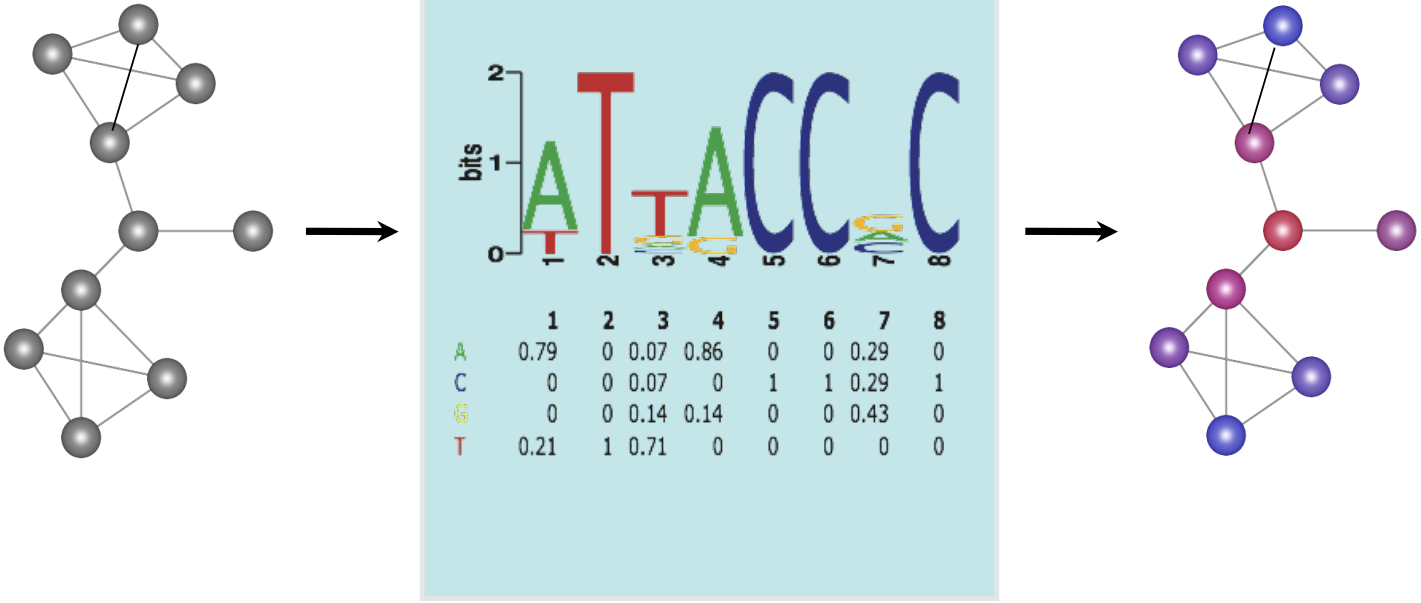


Motif Representations

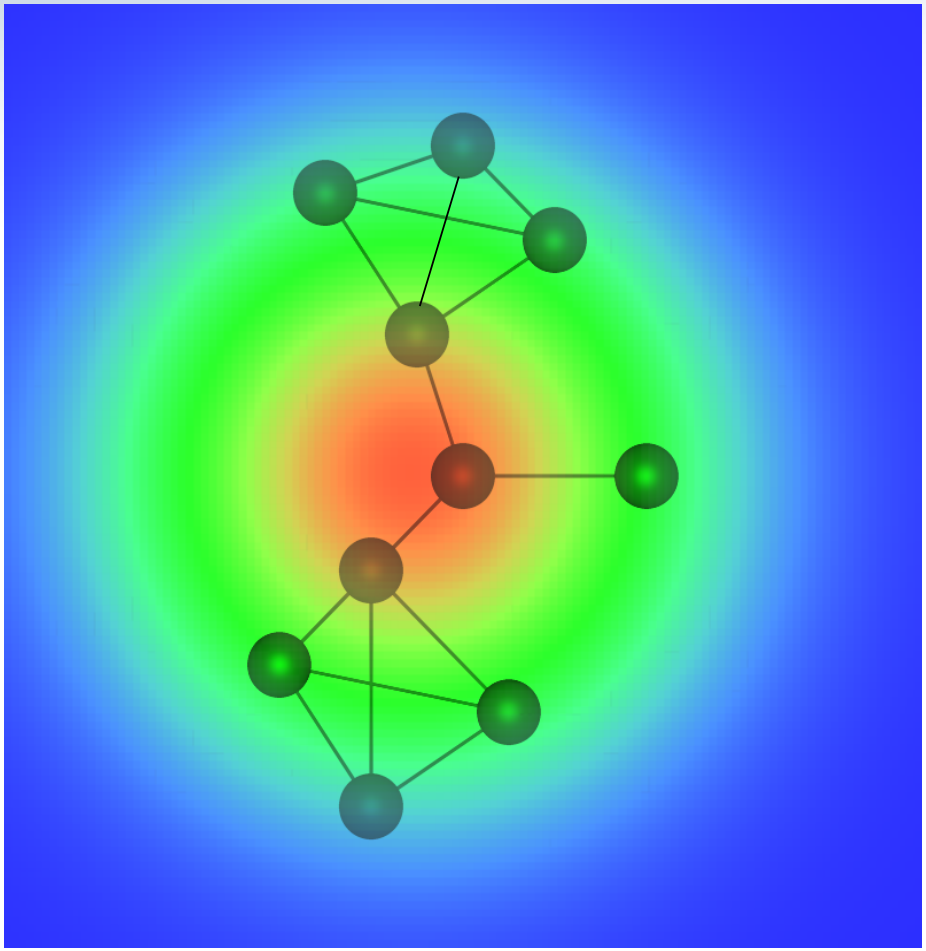


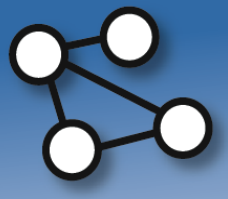
	1	2	3	4	5	6	7	8
A	0.79	0	0.07	0.86	0	0	0.29	0
C	0	0	0.07	0	1	1	0.29	1
G	0	0	0.14	0.14	0	0	0.43	0
T	0.21	1	0.71	0	0	0	0	0

How Well Does a PSSM Model the Motif?



PSSM Scores





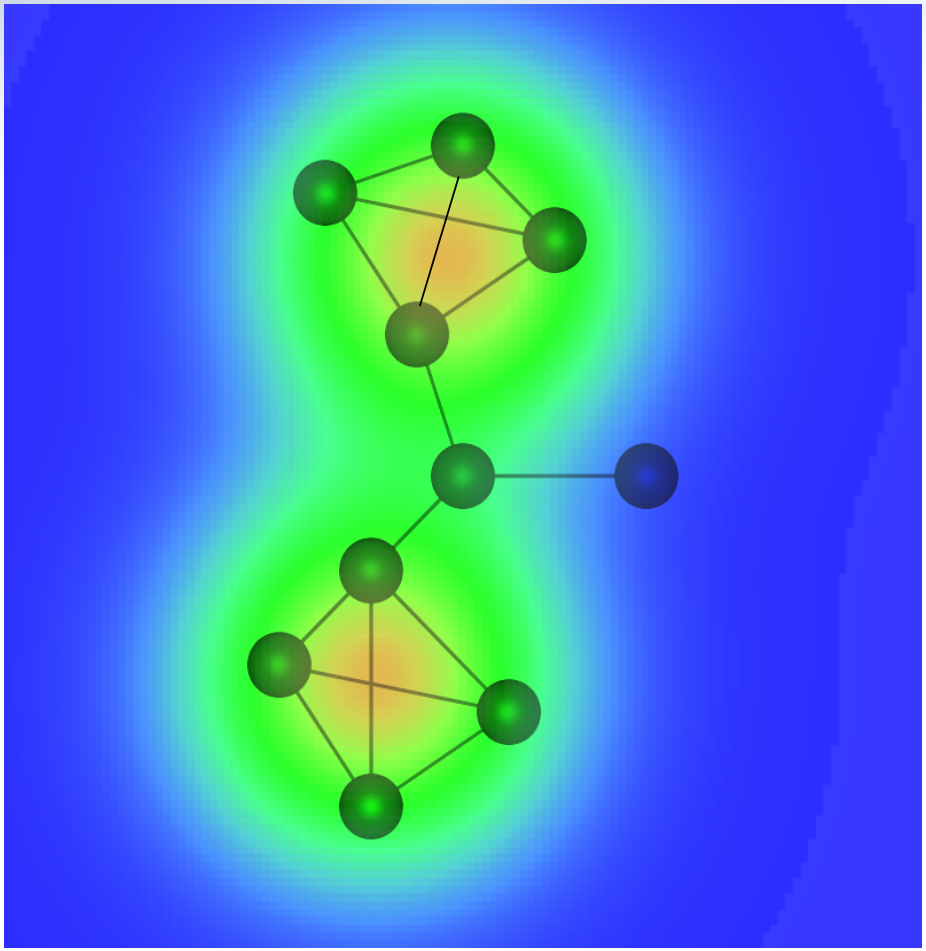
More complex models

- **Barash et al.** developed a Bayesian network model. They investigated mixtures of PSSMs, tree Bayesian networks and mixtures of trees.

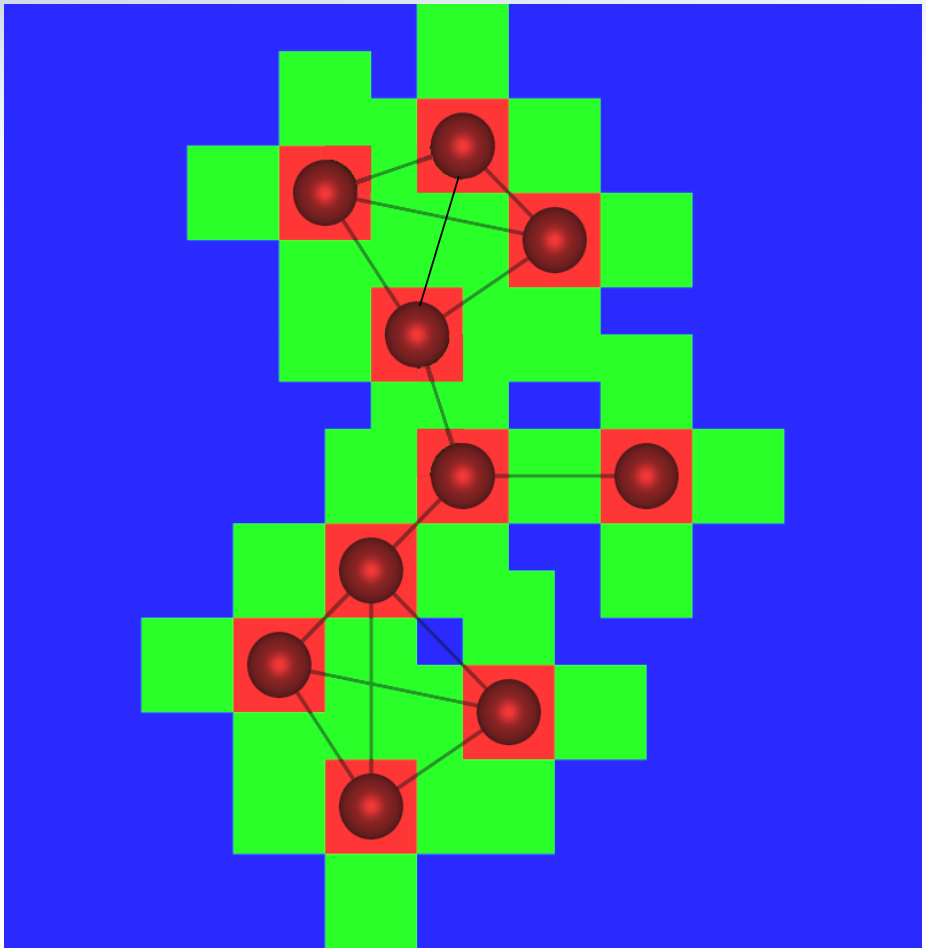
- **Zhou and Liu** developed a PSSM that includes pairs of correlated positions.

- **King and Roth** developed a PSSM-based non-parametric method. Their model interpolated between a PSSM based on all members of the motif, and a mixture model, with one PSSM for each member of the motif.

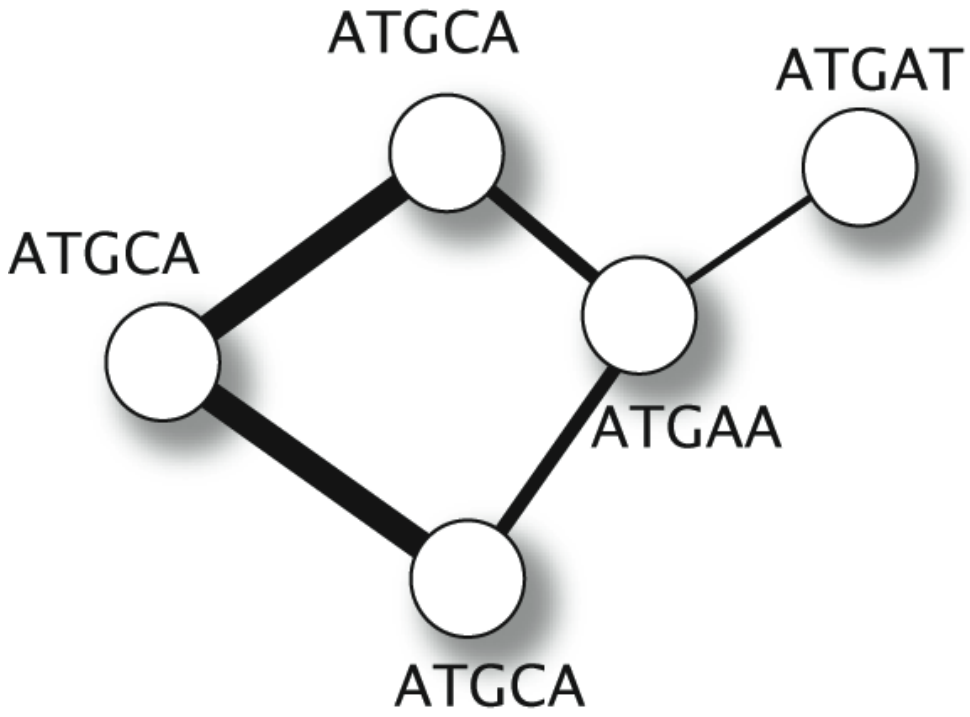
A Mixture of PSSMs



One PSSM Per Example

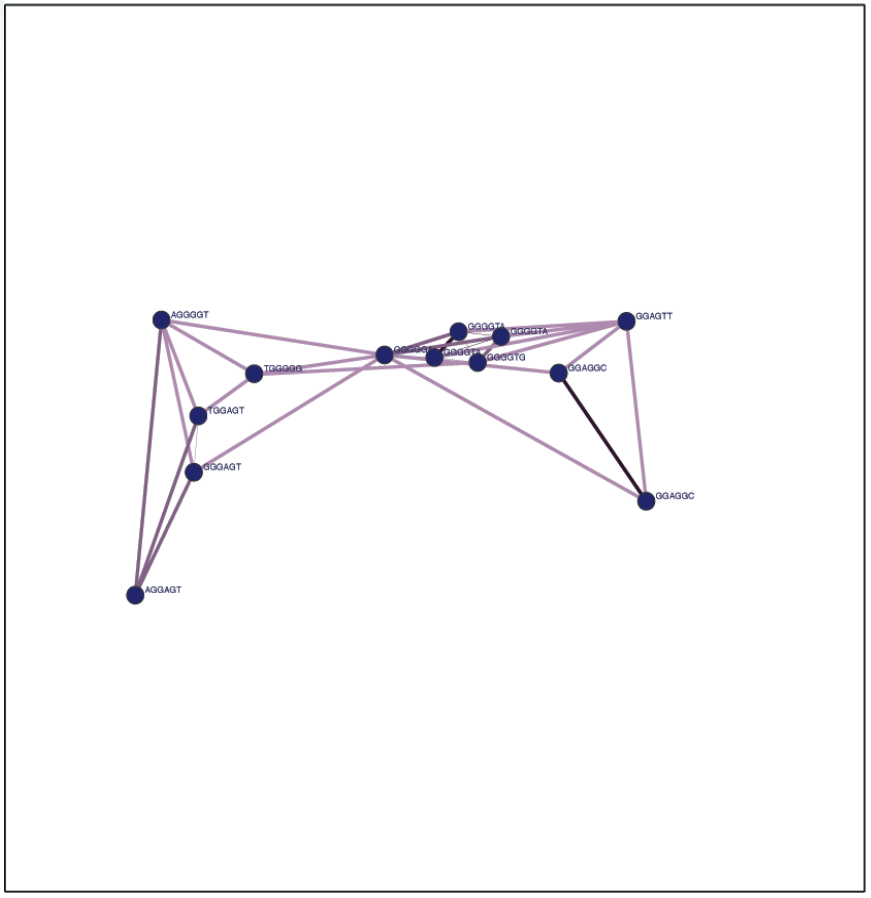


Graph Representation

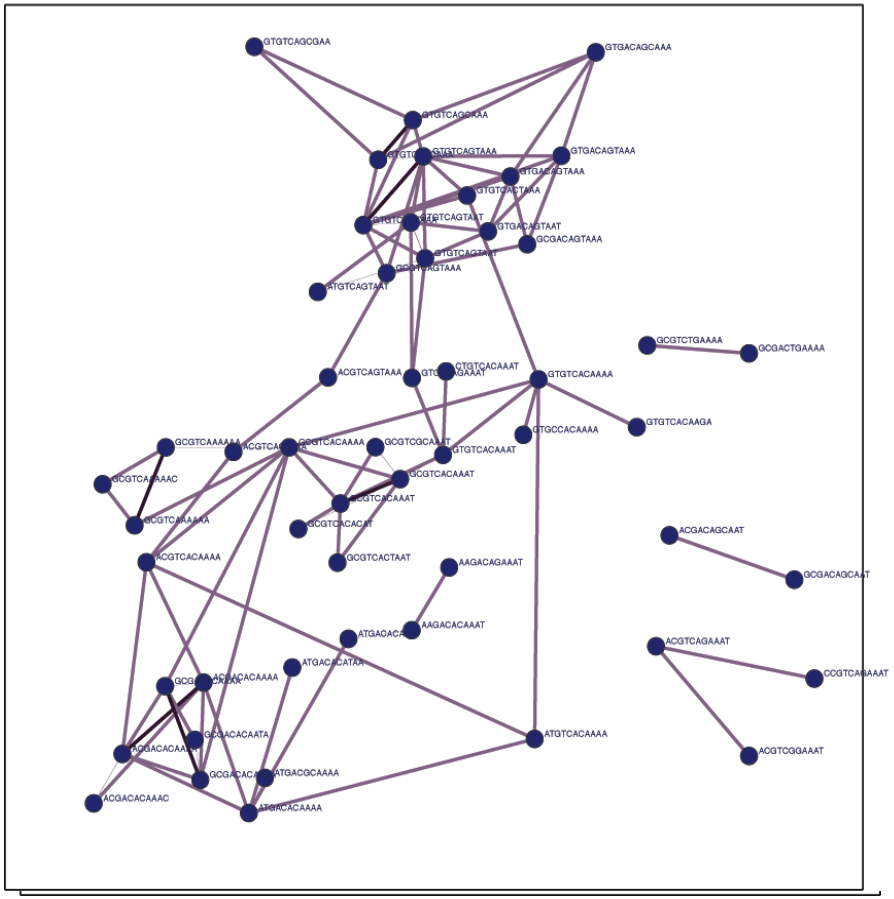


Some Yeast Motifs

ADR1

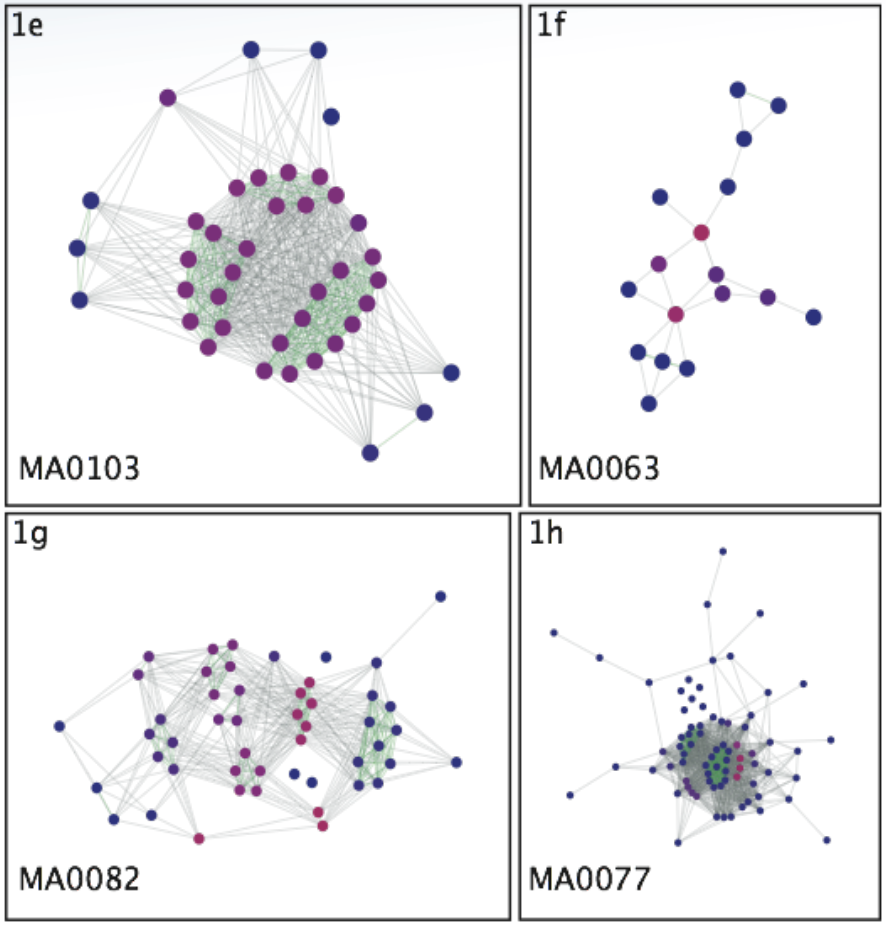


SUM1

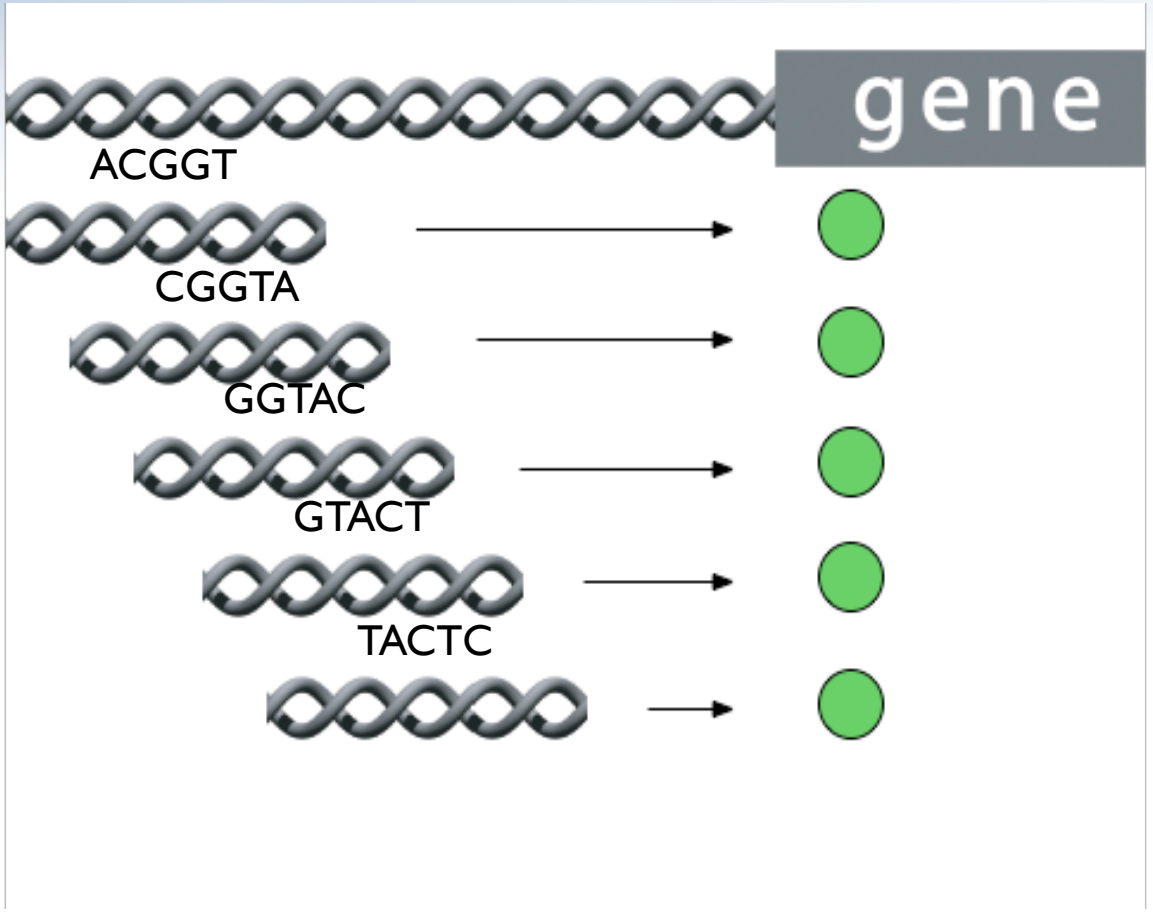


Some Eukaryotic Motifs

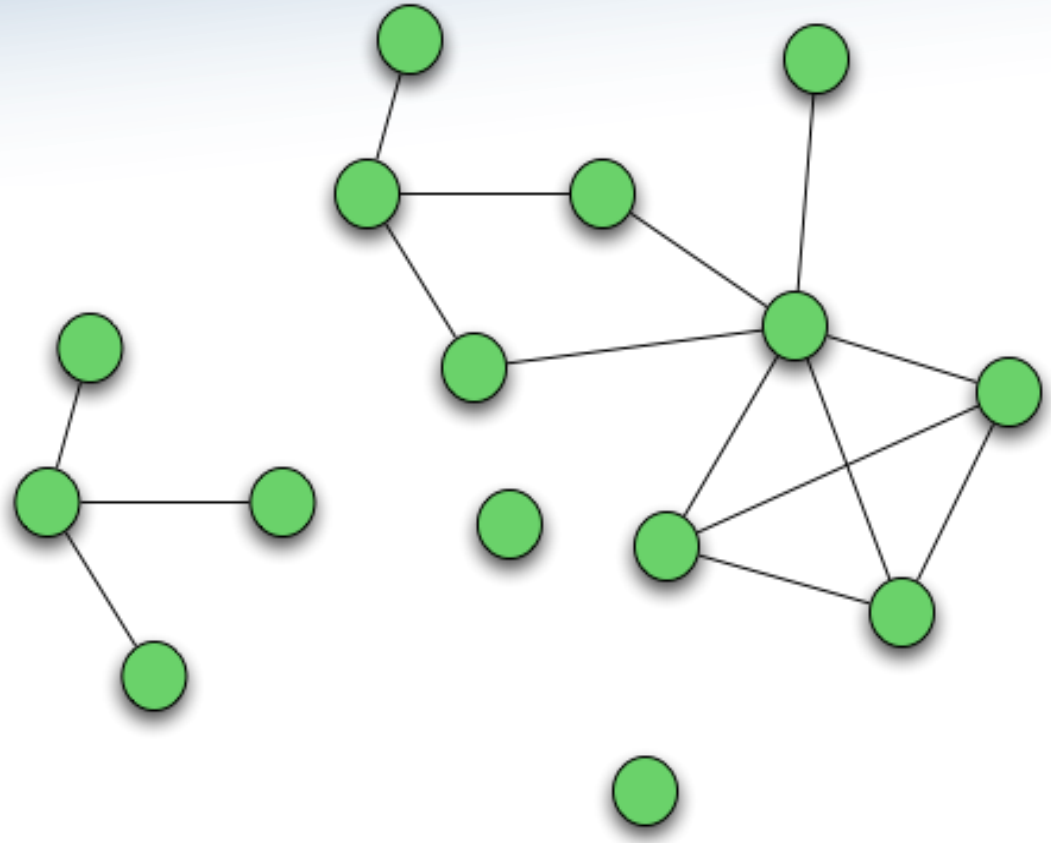
JASPAR motifs



MotifCut

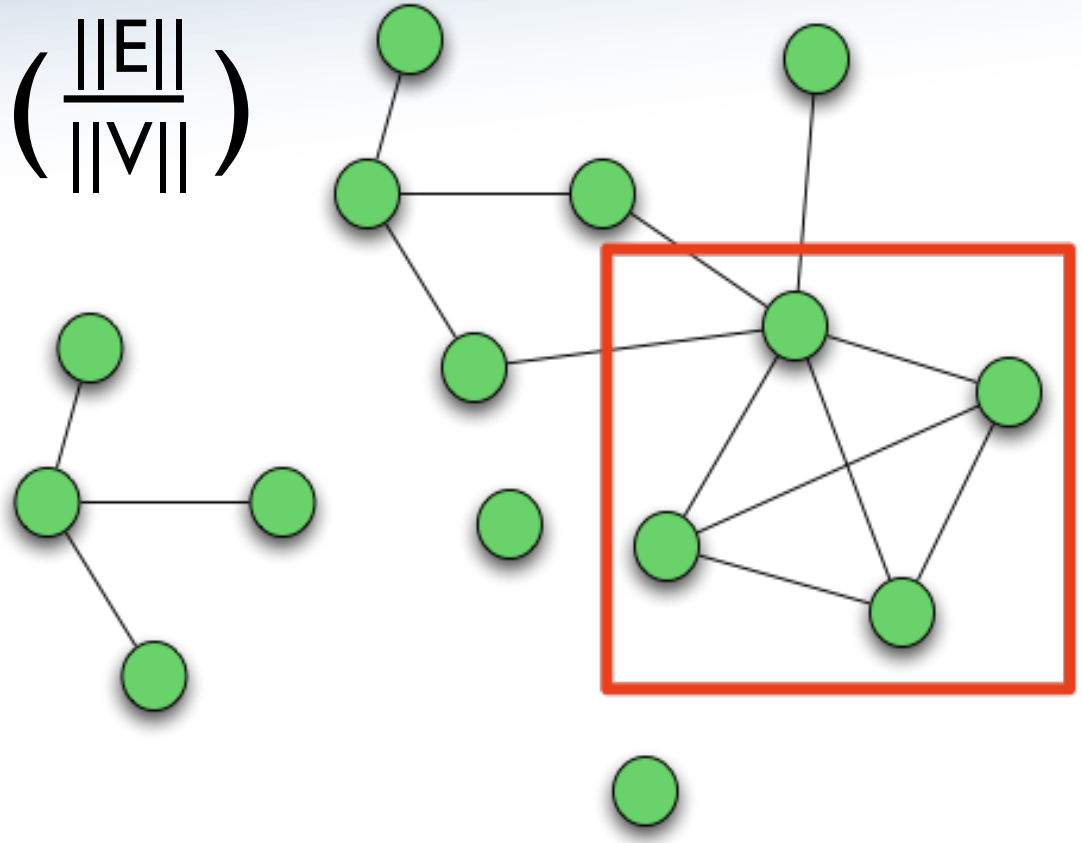


MotifCut



Maximum Density Subgraph

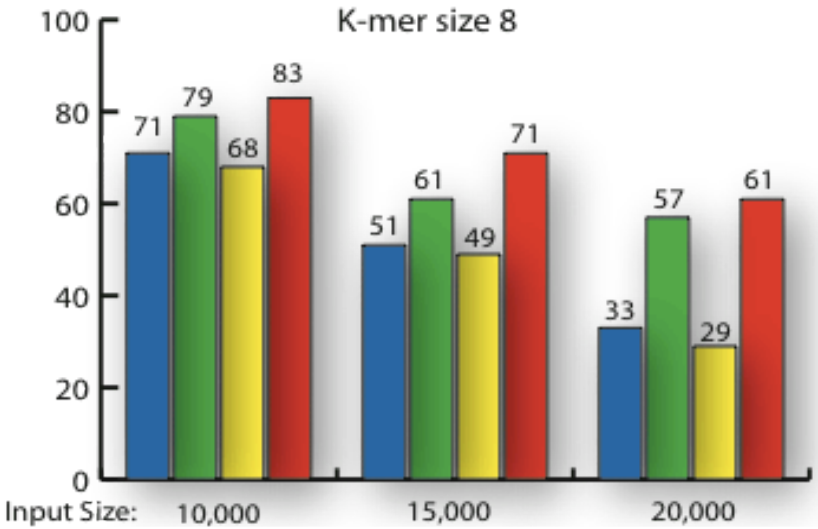
$$\max\left(\frac{||E||}{||V||}\right)$$



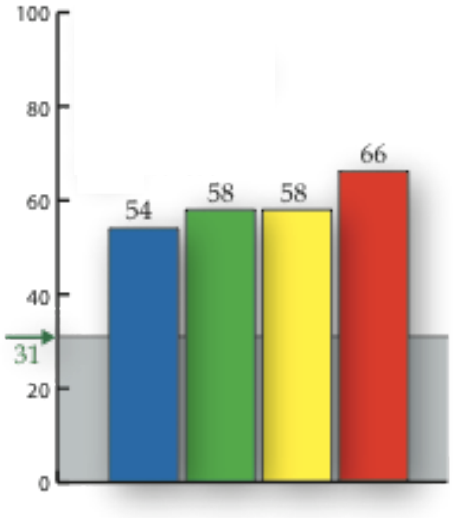
4 vertices
6 edges

MotifCut Performance

Synthetic data



Yeast data

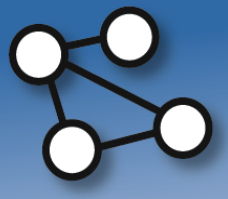


MotifCut AlignAce BioProspector MEME

MotifCut Performance

	AlignAce	BioProspector	MEME
MotifCut	0.14	0.10	0.12
MEME	0.20	0.31	
BioProspector	0.24		

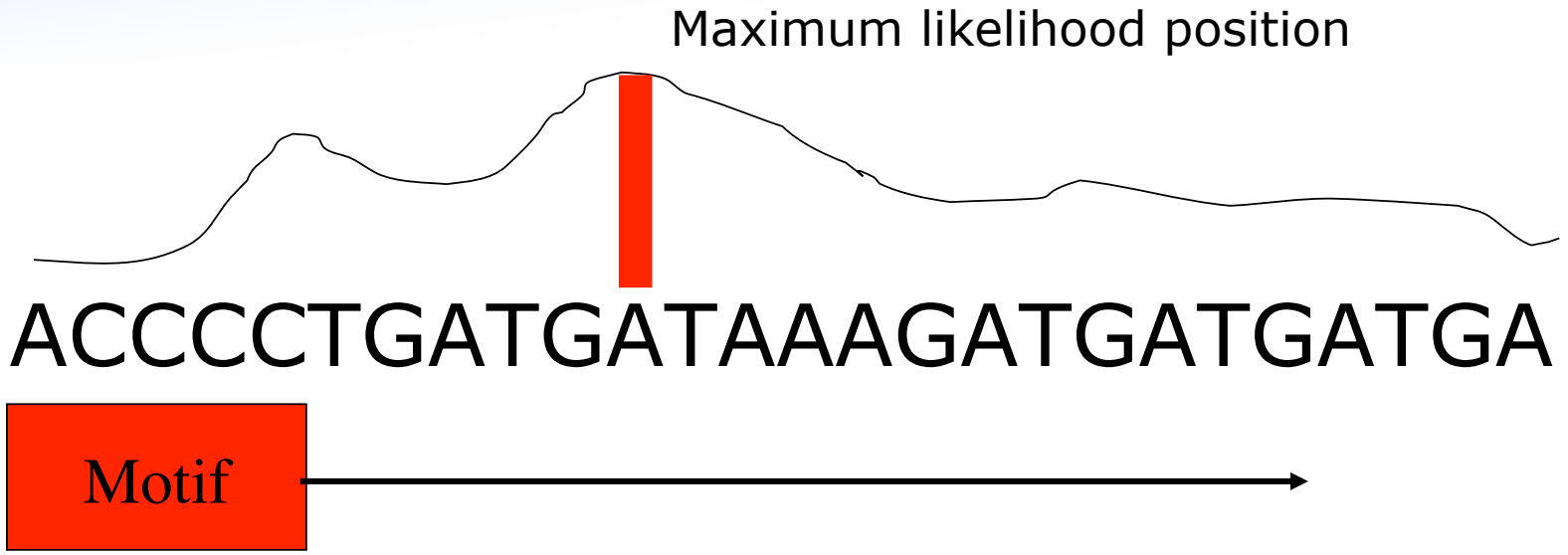
A log-odds measure of similarity of motifs found by different algorithms



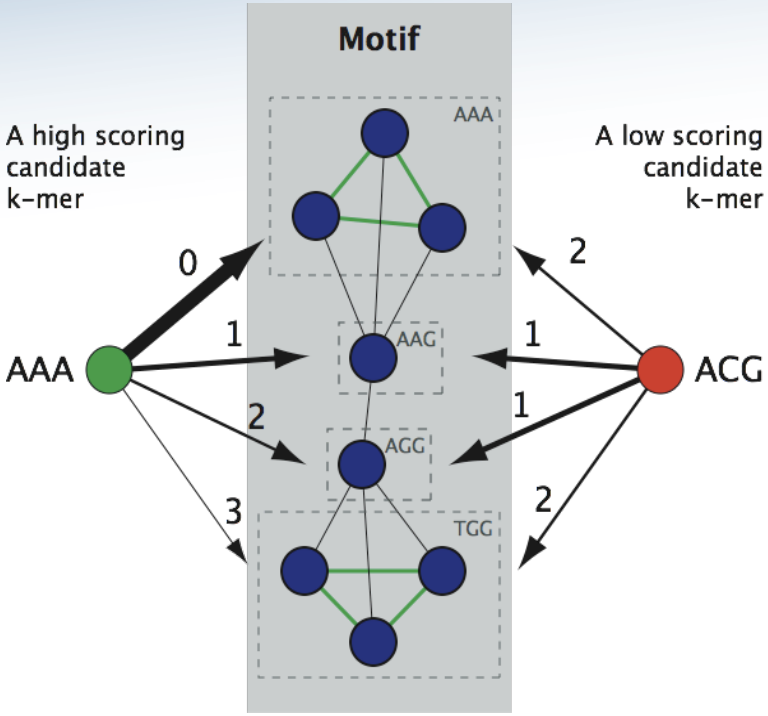
MotifCut

- Advantages:
  - Performance
  - Low correlation with present methods
  - Deterministic
  - Not alignment-based
  - Good for comparative genomics

Motif scanning

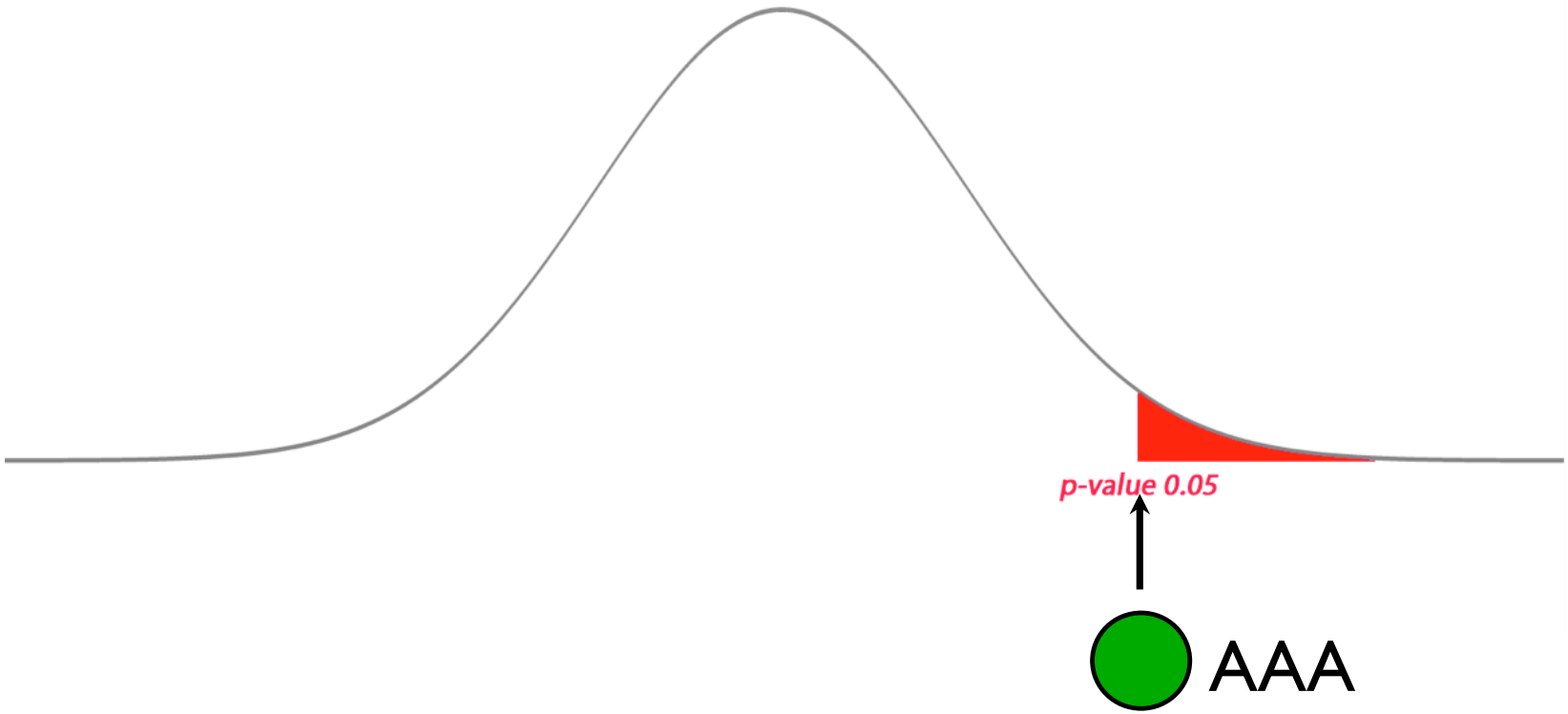


Motif Scanning with MotifScan

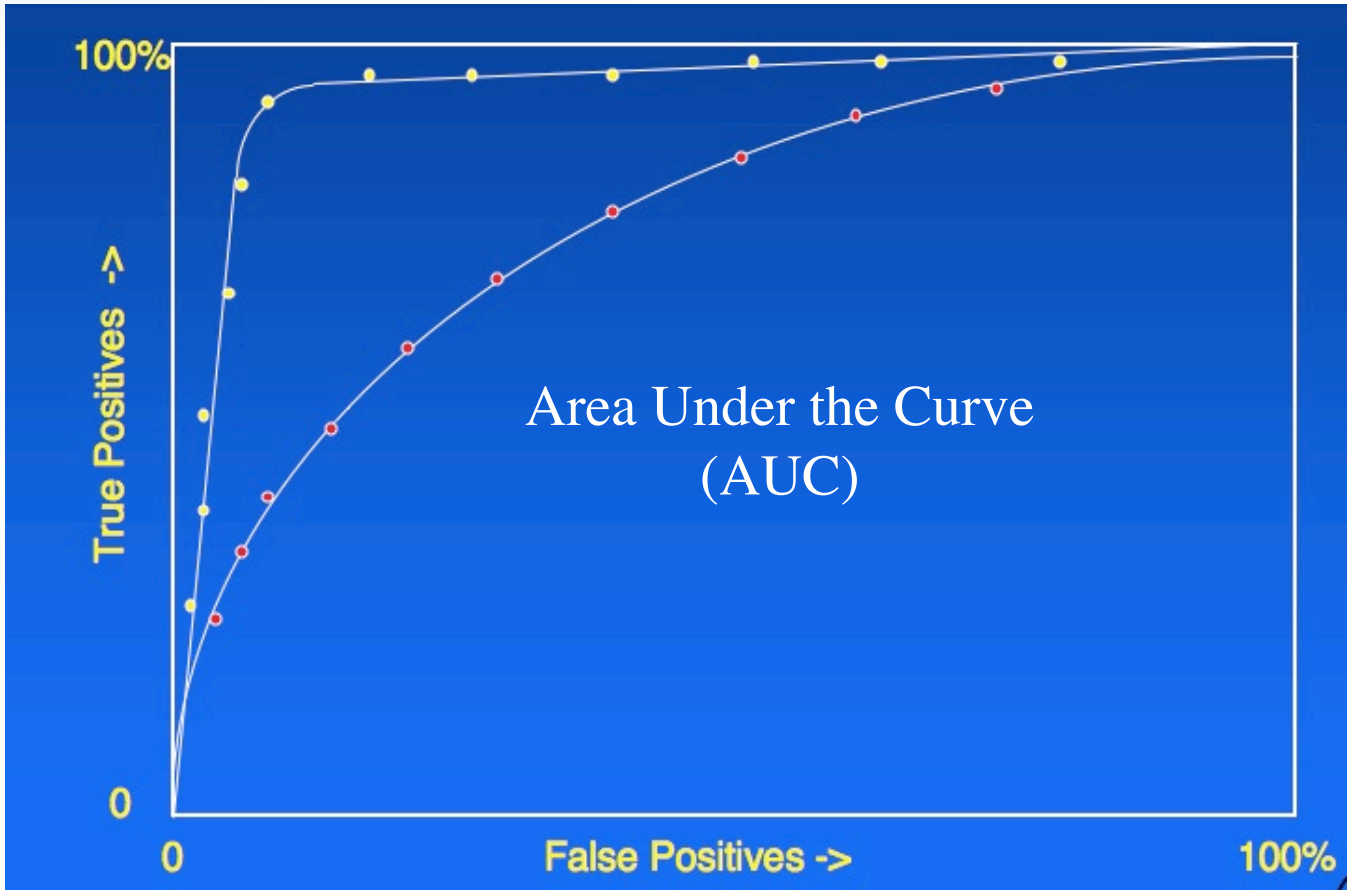


$$Score = \sum_{i=1}^{N} \Theta_{SS}^{d} \Theta_{NS(b1,b2)} \sum_{j=1}^{n_i} \Theta_{IK}^{j}$$

MotifScan p-values



Receiver-Operator Characteristic Curves



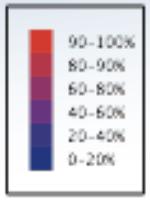
MotifScan Results

Yeast motifs

MotifScan > PSSM+ 5%

Motif Name	# k-mers	Area under ROC curve: MotifScan	Area under ROC curve: PSSM
ADR1	29	100%	49%
CAD1	22	78%	67%
CIN5	135	100%	84%
FKH1	154	100%	90%
GCN4	177	100%	82%
GLN3	79	98%	79%
HAP4	37	90%	82%
MSN2	32	72%	60%
MSN4	37	100%	73%
PHD1	116	100%	62%
RAP1	112	89%	76%
RCS1	59	90%	75%
RDS1	10	100%	88%
RFX1	11	93%	84%
ROX1	28	99%	82%
SKN7	125	91%	22%
SOK2	184	100%	65%
SPT2	13	65%	56%
SPT23	53	100%	83%
SUT1	42	31%	22%
SWI4	128	100%	90%
SWI6	179	100%	81%
TEC1	96	100%	92%
UME6	88	95%	88%
YAP7	91	90%	62%

26 motifs



JASPAR motifs

MotifScan > PSSM+ 10%

Motif Name	# k-mers	Area under ROC curve: MotifScan	Area under ROC curve: PSSM
MA0001	97	86%	58%
MA0002	29	88%	67%
MA0005	90	54%	33%
MA0006	24	100%	80%
MA0008	25	100%	78%
MA0011	12	100%	26%
MA0014	12	22%	7%
MA0015	80	82%	66%
MA0020	21	100%	86%
MA0031	20	100%	65%
MA0034	25	29%	12%
MA0037	63	100%	65%
MA0038	53	66%	37%
MA0040	18	72%	60%
MA0041	47	77%	61%
MA0044	13	50%	5%
MA0054	70	100%	55%
MA0056	20	100%	79%
MA0057	16	36%	22%
MA0063	17	100%	48%
MA0067	31	95%	25%
MA0070	18	67%	53%
MA0077	76	93%	63%
MA0080	57	100%	65%
MA0081	49	100%	70%
MA0084	28	73%	39%
MA0086	40	100%	83%
MA0087	23	100%	78%
MA0089	34	100%	87%
MA0092	29	67%	35%
MA0095	17	100%	86%
MA0098	40	100%	82%
MA0103	41	100%	69%
MA0105	18	74%	60%

34 motifs

PSSM > MotifScan + 5%

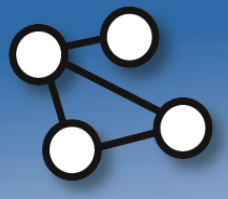
Motif Name	# k-mers	Area under ROC curve: MotifScan	Area under ROC curve: PSSM
ABF1	29	79%	85%

1 motif

PSSM > MotifScan + 10%

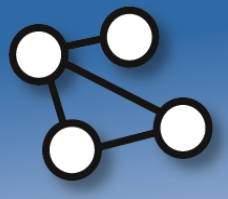
Motif Name	# k-mers	Area under ROC curve: MotifScan	Area under ROC curve: PSSM
MA0007	24	48%	59%
MA0018	16	18%	31%
MA0024	10	79%	92%
MA0045	14	4%	15%

4 motifs



Conclusion

- **MotifScan** uses a **graph-based model** of transcription factor binding sites, which retains **all the known motif instances**.
- This model works **significantly better than a PSSM.**



Conclusions

- Our **graph-based methods** perform better than the current methods.
- They make **fewer assumptions** about the distribution of k-mers in the motif.
- They deal naturally with **k-mer clustering.**
- They represent positional correlations implicitly