Position weight matrix

Position weight matrix

GAL4

Get regulated genes | Get sites | Get consensus | Get matrix | Get affinity data
Get genomewise distribution | Sort by copy No | Sort by function category

```
>YBR018C        AGCGCTCGGACAACTGTTGACC
>YBR018C        ATACTTCGGAGCACTGTTGAGCG
>YBR019C        CGGCGGCTTCTAATCCG
>YBR019C        TCGGAGGGCTGTCGCCCG
>YBR019C        CGGAGGAGAGTCTTCCG
>YBR019C        ATTGTTCGGAGCAGTGCGGCGCG
>YBR020W        CGCGCCGCACTGCTCCGAACAAT
>YBR020W        CGGAAGACTCTCCTCCG
>YBR020W        CGGGCGACAGCCCTCCGA
>YBR020W        CGGATTAGAAGCCGCCG
>YLR081W        TATCGGGGCGGATCACTCCGAAC
>YLR081W        CACCGGCGGTCTTTCGTCCGTGC
>YML051W        CGGCGCACTCTCGCCCG
>YOR120W        TCGGGGCAGACTATTCCGG
```

```
>YBR018C
CGGACAACTGTTGACCG
>YBR018C
CGGAGCACTGTTGAGCG
>YBR019C
CGGCGGCTTCTAATCCG
>YBR019C
CGGAGGGCTGTCGCCCG
>YBR019C
CGGAGGAGAGTCTTCCG
>YBR019C
CGGAGCAGTGCGGCGCG
>YLR081W
CGGGGCGGATCACTCCG
>YLR081W
CGGCGGTCTTTCGTCCG
>YML051W
CGGCGCACTCTCGCCCG
>YOR120W
CGGGGCAGACTATTCCG
```

Get regulated genes | Get sites | Get consensus | Get matrix | Get affinity data
Get genomewise distribution | Sort by copy No | Sort by function category

```
>GAL4    5        118      SCPD

A    0    0    0    4    1    1    7    0    5    1    0    2    0    2    0    0    0
T    0    0    0    0    1    1    1    0    5    2    7    2    1    6    0    0    0
G    0   10    9    4    5    3    2    3    0    3    1    1    4    1    1    0   10
C   10    0    1    2    3    5    0    7    0    4    2    5    5    1    9   10    0
```

Frequencies, weights

Copy-paste in WinEdit; save as .txt; open with Excel

Copy - paste special ‘transpose’

New order: A-C-G-T (instead of ATGC)

COUNTS

```
>GAL4    5        118      SCPD

A    0    0    0    4    1    1    7    0    5    1    0    2    0    2    0    0    0
T    0    0    0    0    1    1    1    0    5    2    7    2    1    6    0    0    0
G    0   10    9    4    5    3    2    3    0    3    1    1    4    1    1    0   10
C   10    0    1    2    3    5    0    7    0    4    2    5    5    1    9   10    0
```

FREQUENCIES

0.023	0.932	0.023	0.023
0.023	0.023	0.932	0.023
0.023	0.114	0.841	0.023
0.386	0.205	0.386	0.023
0.114	0.295	0.477	0.114
0.114	0.477	0.295	0.114
0.659	0.023	0.205	0.114
0.023	0.659	0.295	0.023
0.477	0.023	0.023	0.477
0.114	0.386	0.295	0.205
0.023	0.205	0.114	0.659
0.205	0.477	0.114	0.205
0.023	0.477	0.386	0.114
0.205	0.114	0.114	0.568
0.023	0.841	0.114	0.023
0.023	0.932	0.023	0.023
0.023	0.023	0.932	0.023

WEIGHTS

-2.398	1.316	-2.398	-2.398
-2.398	-2.398	1.316	-2.398
-2.398	-0.788	1.213	-2.398
0.435	-0.201	0.435	-2.398
-0.788	0.167	0.647	-0.788
-0.788	0.647	0.167	-0.788
0.969	-2.398	-0.201	-0.788
-2.398	0.969	0.167	-2.398
0.647	-2.398	-2.398	0.647
-0.788	0.435	0.167	-0.201
-2.398	-0.201	-0.788	0.969
-0.201	0.647	-0.788	-0.201
-2.398	0.647	0.435	-0.788
-0.201	-0.788	-0.788	0.821
-2.398	1.213	-0.788	-2.398
-2.398	1.316	-2.398	-2.398
-2.398	-2.398	1.316	-2.398

$$\ln \frac{(n_{i,j} + p_i) \,/\, (N+1)}{p_i} \;;$$

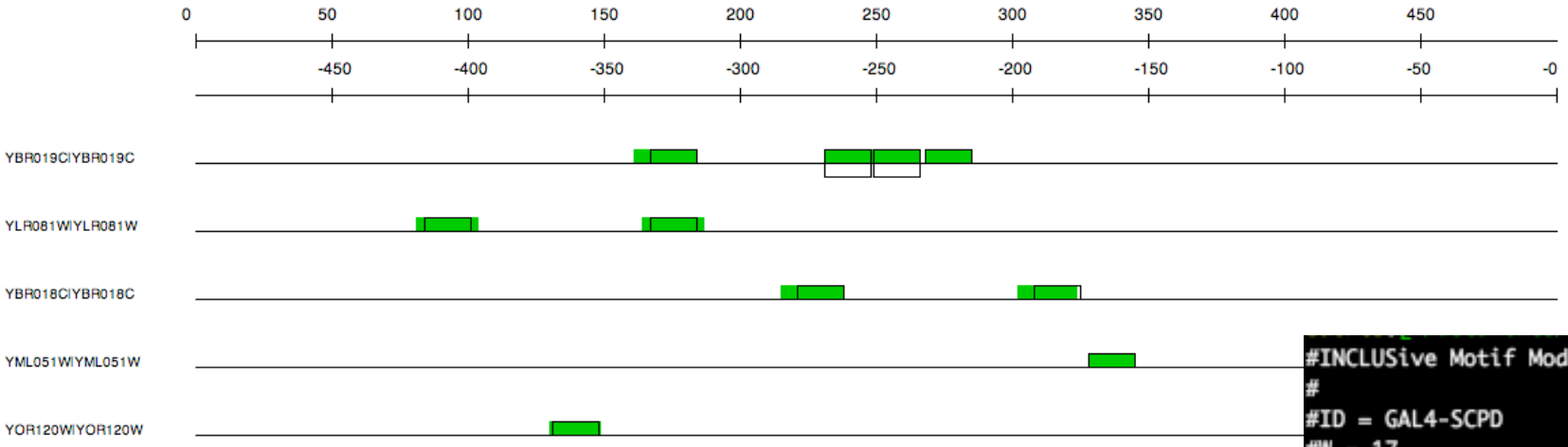
Score an instance with a PWM

			A	C	G	T	MAX	MIN
	1.316	C	-2.398	1.316	-2.398	-2.398	1.31567679	-2.398
	1.316	G	-2.398	-2.398	1.316	-2.398	1.31567679	-2.398
	1.213	G	-2.398	-0.788	1.213	-2.398	1.21302264	-2.398
	0.435	A	0.435	-0.201	0.435	-2.398	0.43531807	-2.398
	0.167	C	-0.788	0.167	0.647	-0.788	0.64662716	-0.788
	-0.788	A	-0.788	0.647	0.167	-0.788	0.64662716	-0.788
	0.969	A	0.969	-2.398	-0.201	-0.788	0.96940056	-2.398
	0.969	C	-2.398	0.969	0.167	-2.398	0.96940056	-2.398
	0.647	T	0.647	-2.398	-2.398	0.647	0.64662716	-2.398
	0.167	G	-0.788	0.435	0.167	-0.201	0.43531807	-0.788
	0.969	T	-2.398	-0.201	-0.788	0.969	0.96940056	-2.398
	-0.201	T	-0.201	0.647	-0.788	-0.201	0.64662716	-0.788
	0.435	G	-2.398	0.647	0.435	-0.788	0.64662716	-2.398
	-0.201	A	-0.201	-0.788	-0.788	0.821	0.82098055	-0.788
	1.213	C	-2.398	1.213	-0.788	-2.398	1.21302264	-2.398
	1.316	C	-2.398	1.316	-2.398	-2.398	1.31567679	-2.398
	1.316	G	-2.398	-2.398	1.316	-2.398	1.31567679	-2.398
SUM	11.259						15.5217066	-32.717
W-Wmin/Wmax-Wmin	0.91162331							

AGCGCT**CGGACAACTGTTGACCG**

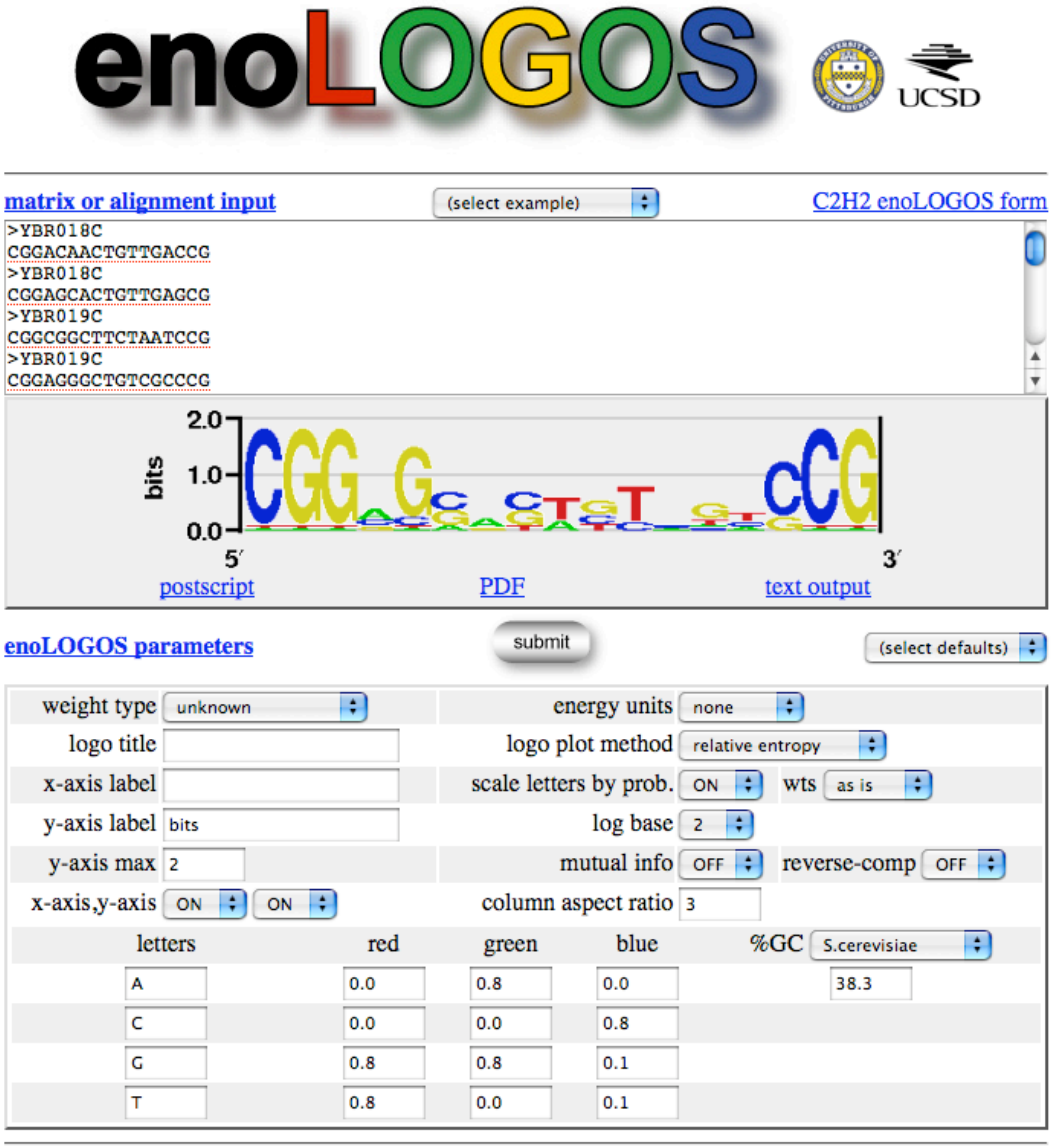
Toucan - find instances back?

- MotifLocator, threshold=0.9, bg0




```
#INCLUSive Motif Model
#
#ID = GAL4-SCPD
#W = 17
0       10      0       0
0       0       10      0
0       1       9       0
4       2       4       0
1       3       5       1
1       5       3       1
7       0       2       1
0       7       3       0
5       0       0       5
1       4       3       2
0       2       1       7
2       5       1       2
0       5       4       1
2       1       1       6
0       9       1       0
0       10      0       0
0       0       10      0
```

- http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi

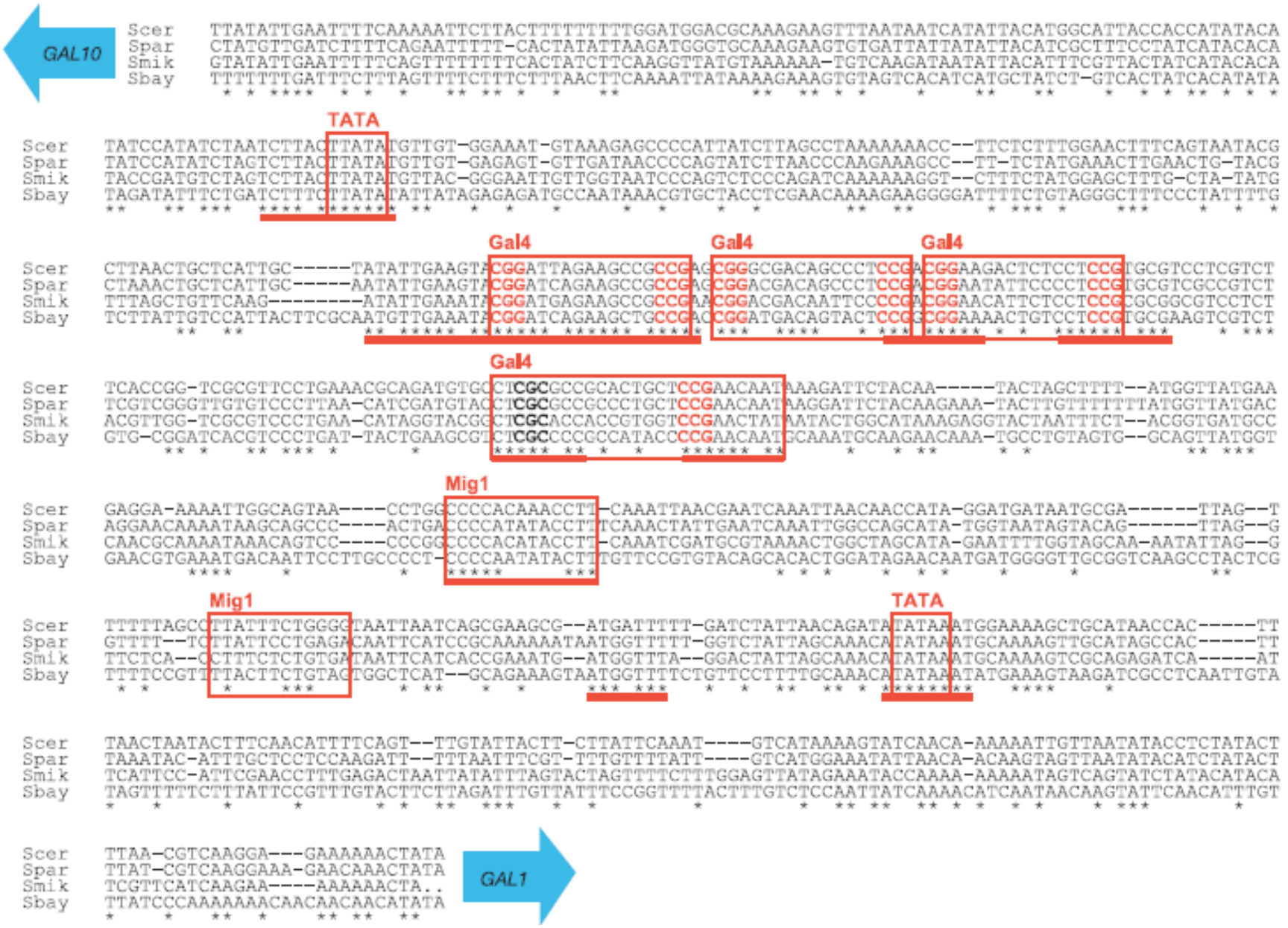


Phylogenetic footprinting

Use sequence conservation in scanning

- Basic assumption: functional sequences evolve more slowly than non-functional sequences, as they are subject to selection pressure.
- Basic approach:
  - Identify conserved regions by sequence alignment algorithms
  - Restrict motif finding in conserved regions.

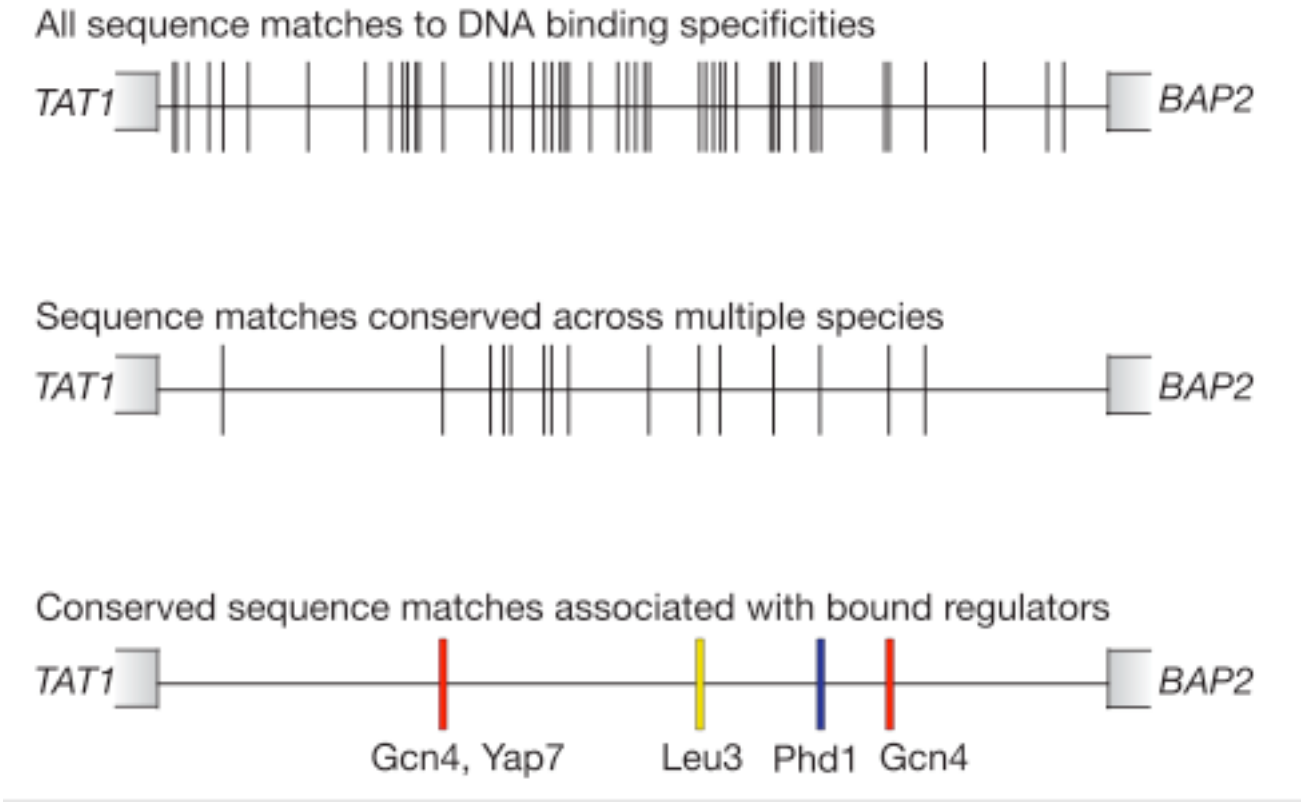
GAL4



Kellis et al Nature 2003

Use sequence conservation in scanning

- Harbison et al
- Criterium for conserved motif: the aligned sequence of at least 2 other *sensu stricto* yeast species have a motif match (>60% of motif maximum)



Use sequence conservation in scanning

- + add phylogenetic tree
  - Evolutionary model
  - Monkey

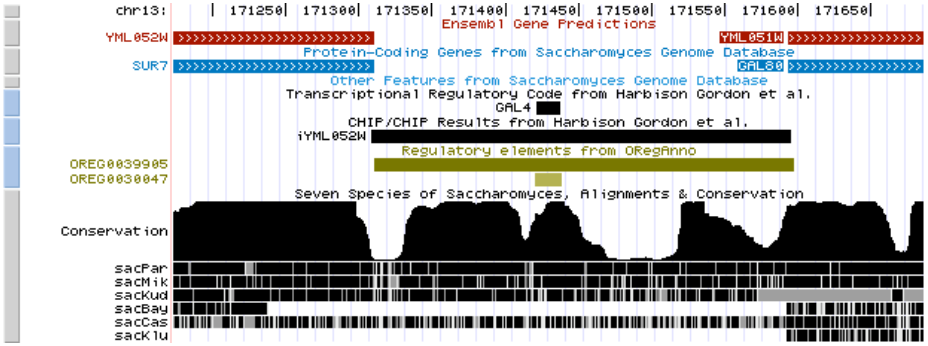
    (Moses et al. Genome Biology 2004)

Table 2

Performance of different scores in recognizing functional and nonfunctional sites

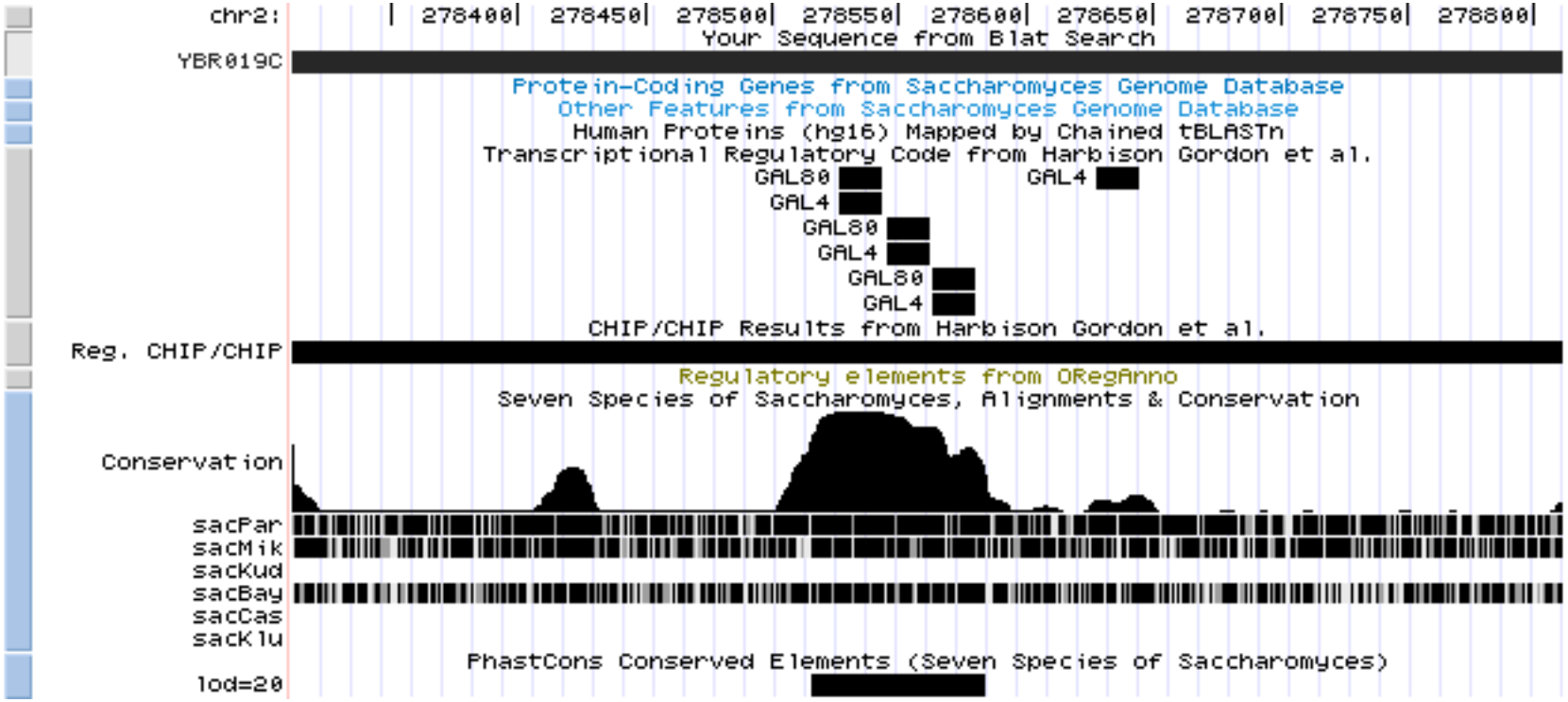
Percent of binding sites assigned a lower *p*-value	Gal4p	
	Positives (n = 10)	Negatives (n = 10)
Halpern-Bruno vs simple	90%	30%
Halpern-Bruno vs *S. cerevisiae* alone	100%	40%
Simple vs *S. cerevisiae* alone	100%	30%

- PhastCons scores



UCSC Tables

- Intersect Gal4 predicted binding sites with PhastCons conserved elements



Seven Species of Saccharomyces, Alignments & Conservation

Conservation score statistics

Capitalize coding exons based on SGD Genes show all bases

Place cursor over species for alignment detail. Click on 'B' to link to browser for aligned species, click on 'D' to get DNA for aligned species.

Alignment block 1 of 1 in window, 278312 - 278811, 500 bps

```
B D  S. cerevisiae  ttata---------ttgaattttcaaaaattcttacttttttttttggatggacgcaaagaagtttaat--
           sacPar   ctatg---------ttgatcttttcagaattttt-cactatattaagatgggtgcaaagaagtgtgat--
           sacMik   gtata---------ttgaattttcagtttttttttcactatcttcaaggttatgtaaaaaatgtcaa---
           sacBay   atacgtttatttttttgatttctttagttttctttctttaacttcaaaattataaaagaaagtgtagtca

     S. cerevisiae  aatcatattacatggcattaccaccatatacatatccatatctaatcttacttatatgttgtggaaa--t
           sacPar   tattatattacatcgctttcctatcatacacatatccatatctagtcttacttatatgttgtgagag--t
           sacMik   gataatattacatttcgttactatcatacacatacgatgtctagtcttacttatatgttacgggaat-t
           sacBay   catcatgctatct---gtcactatcacatatatagatatttctgatctttcttatatattatagagagat

     S. cerevisiae  gtaaagagccccattatcttagcctaaaaaaa--ccttctctttggaactttcagtaatacgcttaactg
           sacPar   gttgataaccccagtatcttaacccaagaaag---cctttctatgaaacttg-aactgtacgctaaactg
           sacMik   gttggtaatcccagtctcccagatcaaaaaag--gtctttctatggagcttt--gctatatgtttagctg
           sacBay   gccaataaacgtgctacctcgaacaaaagaaggggattttctgtagggctttccctattttgtcttattg

     S. cerevisiae  ctcatt-----gctatattgaagtacggattagaagccgccgagcgggcgacagccctccgacggaagac
           sacPar   ctcatt-----gcaatattgaagtacggatcagaagccgccgagcggacgacagccctccgacggaatat
           sacMik   ttca--------agatattgaaatacggatgagaagccgccgaacggacgacaattccccgacggaacat
           sacBay   tccattacttcgcaatgttgaaatacggatcagaagctgccgaccggatgacagtactccggcggaaaac

     S. cerevisiae  tctcctccgtgcgtcctcgtcttcaccggtcgcgttcctgaaacgcagatgtgcctcgcgccgcactgct
           sacPar   tcccctccgtgcgtcgccgtcttcgtcggttgtgtccctt-aacatcgatgtacctcgcgccgccctgct
           sacMik   tctcctccgcgcggcgtcctctacgttggtcgcgtccctg-aacataggtacggctcgcaccaccgtggt
           sacBay   tgtcctccgtgcgaagtcgtctgtgcggatcacgtccctg-attactgaagcgtctcgccccgccatacc

     S. cerevisiae  ccgaacaataaagattctacaa-----tactag--cttttatggttatg-aagaggaaaaattggcagta
           sacPar   ccgaacaataaggattctacaa-gaaatacttgtttttttatggttatgacaggaacaaaataagcagcc
           sacMik   ccgaactataatactggcataaagaggtactaa--tttctacggtgatgcccaacgcaaaataaacagtc
           sacBay   ccgaacaatgcaaatgcaagaa-caaatgcctg--tagtggcagttatggtgaacgtgaaat-gacaatt

     S. cerevisiae  acctggccccacaaacc-ttcaaattaacgaatcaaattaacaaccataggatgataat------gcg--
           sacPar   cactgaccccatataccttttcaaactattgaatcaaattggccagcatatggtaatagt------aca--
           sacMik   ccccggccccacatacc-ttcaaatcgatgcgtaaaactggctagcatagaattttggta-----gcaaa
           sacBay   ccttgcccctccccaatatactttgttccgtgtacagcacactgg-atagaacaatgatggggttgcg--

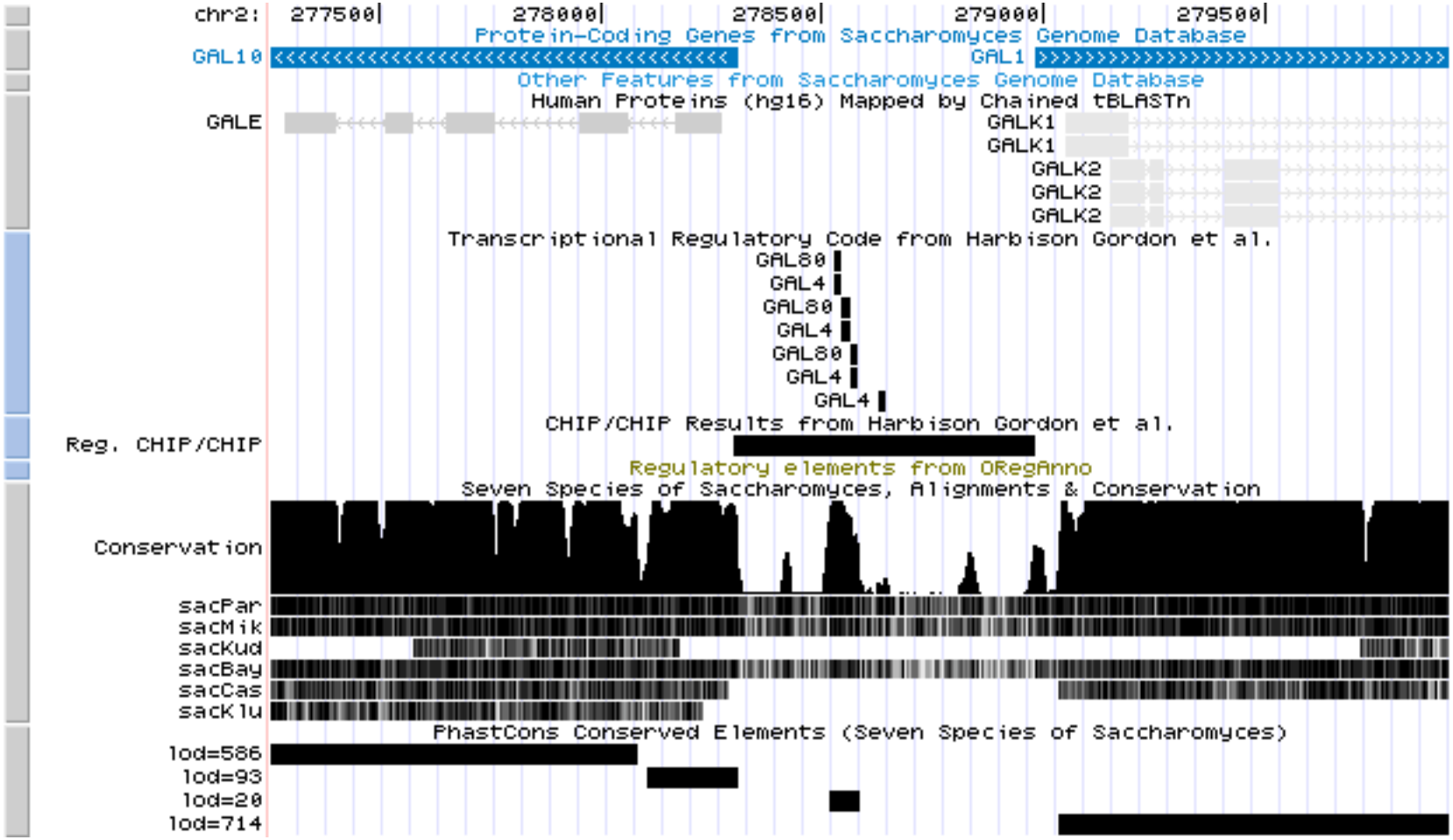
     S. cerevisiae  --attag--------ttttttagccttatttctggggtaattaatcagc--gaagcgat
           sacPar   --gttag--------ggttttc--cttattcctgagacaattcatccgcaaaaaataat
           sacMik   atattag--------gttctca--cctttctctgtgataattcatcacc--gaaatgat
           sacBay   --gtcaagcctactcgttttccgttttacttctgtagtggctcatgcag--aaagtaat
```

Words of caution

- Non-coding DNA alignment is more difficult
  - Incorrect alignment
  - Difficult to find the orthologous motif
  - Allow motif movement within N bp window
  - ‘preserved motifs’ instead of ‘conserved’
  - Motif is present in orthologous promoters, not necessarily aligned
- Motif conservation vs motif diversity
  - evolution of animal morphology by cis-evolution

Motif Clusters

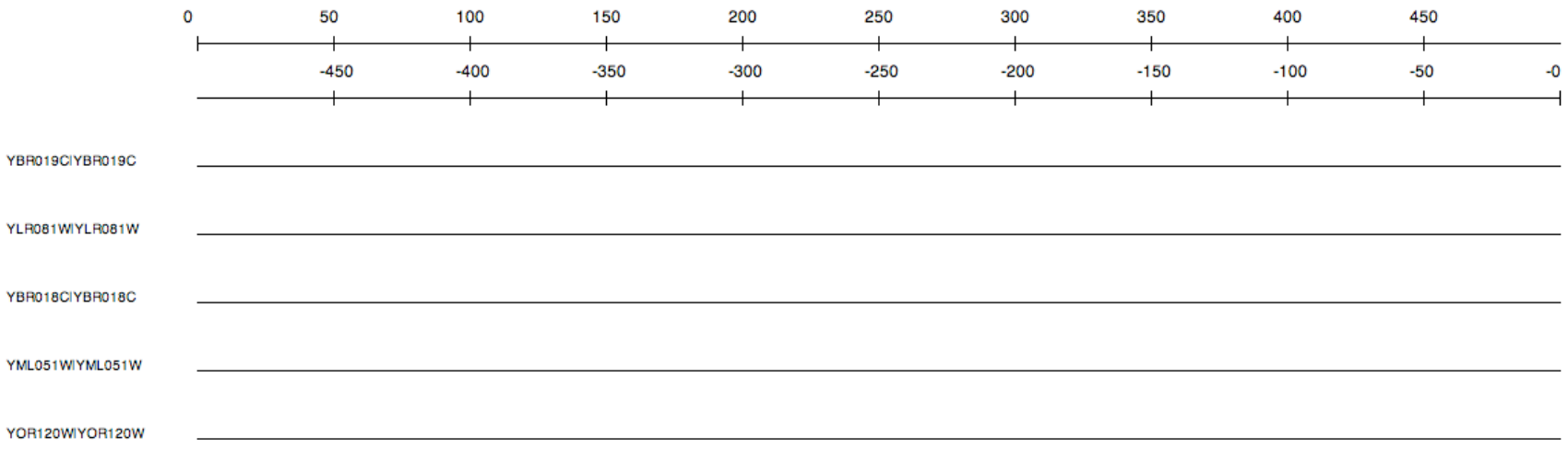
Motif clusters



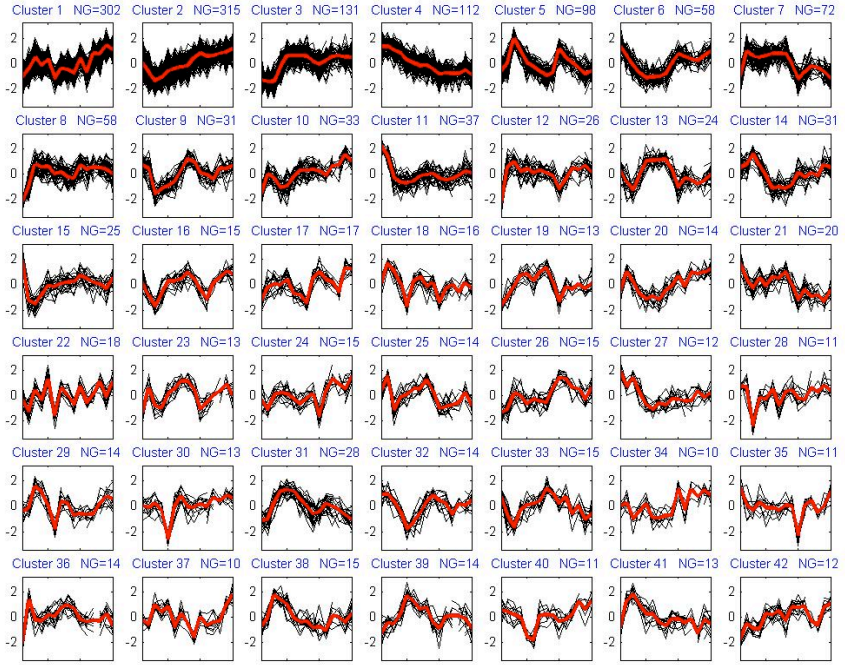
Motif discovery

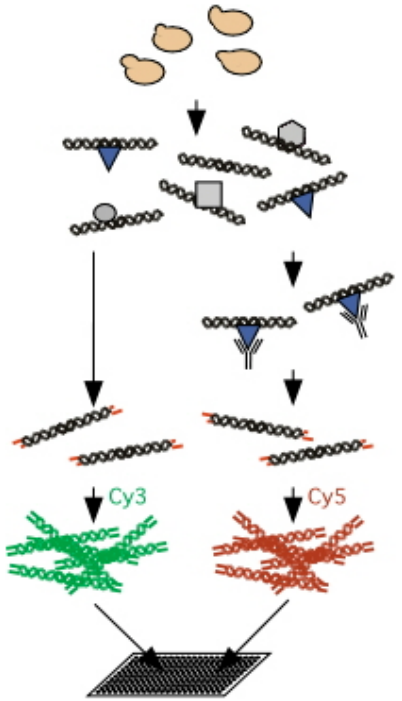
Motif discovery

- Given: set of co-regulated genes
- Question: what are the regulatory motifs?



Co-regulated genes: microarrays & ChIP-chip/seq





Crosslink protein to DNA in vivo with formaldehyde

Break open cells and shear DNA

Immunoprecipitate

Reverse-crosslinks, blunt DNA and ligate to unidirectional linkers LM-PCR

Hybridize to array

GATGGCTGCACCACGTGTATGC...ACGATGTCTCGC

CACATCGCATCACGTGACCAGT...GACATGGACGGC

GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA

TCTCGTTAGGACCATCACGTGA...ACAATGAGAGCG

CGCTAGCCCACGTGGATCTTGT...AGAATGGCCTAT

GTA...GATGGCTGCACCACGTGTATGC

CGA...CACATCGCATCACGTGACCAGT

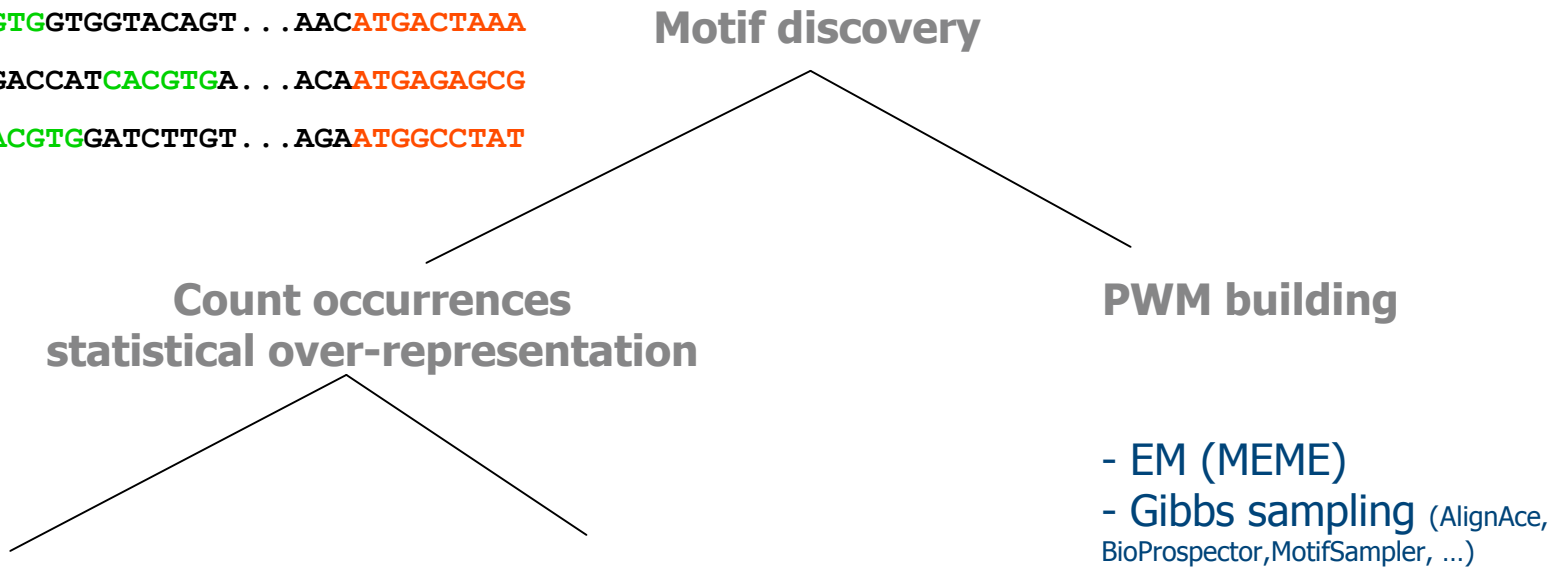
TTA...GCCTCGCACGTGGTGGTACAGT

GGC...TCTCGTTAGGACCATCACGTGA

CTG...CGCTAGCCCACGTGGATCTTGT

Motif discovery

```
GATGGCTGCACCACGTGTATGC...ACGATGTCTCGC
CACATCGCATCACGTGACCAGT...GACATGGACGGC
GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA
TCTCGTTAGGACCATCACGTGA...ACAATGAGAGCG
CGCTAGCCCACGTGGATCTTGT...AGAATGGCCTAT
```



- EM (MEME)
- Gibbs sampling (AlignAce, BioProspector,MotifSampler, …)

- all possible k-mers
- 4^k possibilities
- binomial analysis w/ background frequencies

- library of PWM's (TRANSFAC, JASPAR)
- binomial analysis w/ background frequencies

van Helden J, et al.. J Mol Biol. 1998 4;281(5):827-42.

Aerts S et al. Nucleic Acids Res. 2003 15;31(6):1753-64.

- *Lawrence CE, et al. Science. 1993 8;262(5131):208-14.*
- *Bailey T et al. Proceedings of ISMB, pp. 28-36, 1994.*

Statistical over-representation: Binomial probability

Example

On a ten question multiple choice test, with 4 options per question, the probability of getting 5 and only 5 answers correct if the answers are guessed can be calculated

$p = 0.25$

$q = 1 - p = 0.75$

$k = 5$

$n = 10$

$$\begin{aligned} P(pass) &= \binom{n}{k} p^k q^{n-k} \\ &= \binom{10}{5} (0.25)^5 (0.75)^5 \\ &\approx 5.8\% \end{aligned}$$

Therefore if somebody guesses 10 answers on a multiple choice test with 4 options, they have about a 5.8% chance of getting 5 and only 5 correct answers. If 5 or mo adding the probability of getting 5 (and only 5) answers correct, 6 (and only 6) answers correct, and so on up to 10 answers correct. This total probability is approximat

$$E(occ\{b\}) = F_e\{b\} \times 2 \times \sum_{i=1}^{S} (L_i - w + 1) = F_e\{b\} \times T$$

$$P(occ\{b\} = n) = \frac{T!}{(T-n)! \times n!} \times (F_e\{b\})^n \times (1 - F_e\{b\})^{(T-n)}$$

$$P(occ\{b\} \geq n) = \sum_{j=n}^{T} P(occ\{b\} = j)$$

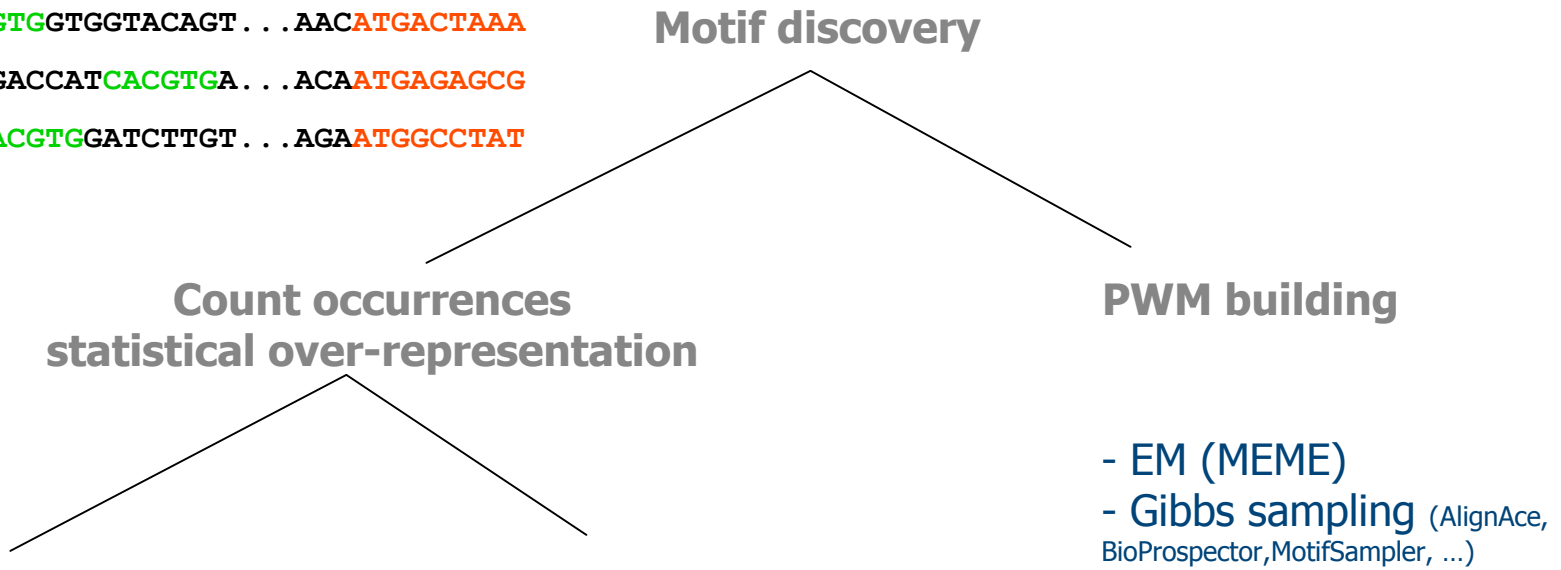
Oligo-analysis example

Family	Hexanucleotide analysis result: Sequence	Ms	Occ	Exp	*sig*	Sites previously characterized: Consensus	Bound factors
NIT	ATAAGA	6	20	6.0	**2.0**		Gln3p, Nillp,
	GATAAG	6	26	3.0	9.1	GATAAG	Gzf3p, Uga43p
	AGATAA	7	17	6.1	0.4		(Zn finger)
	CTGATA	6	11	3.1	0.1		
	CCGCGC	2	6	0.7	0.8	–	–
	CGGCAC	4	6	0.8	0.5	–	–
	ACATCT	4	11	2.9	0.4	–	–
MET	CACGTG	9	26	2.0	**7.0**		Cbflp-Met4p-
	TCACGT	9	19	2.9	**6.1**	TCACGTG	Met28p complex
	GTCACG	6	8	1.4	0.7		(bHLH-bLZ-bLZ)
	TGTGGC	7	10	2.4	0.5		
	CTGTGG	8	11	2.1	**1.6**	AAAACTGTGG	Met31p, Met32p
	ACTGTG	9	12	3.2	0.6		(Zn finger)
	AACTGT	10	17	5.5	0.9		
	ATATAT	19	82	42.3	0.8	–	–
	TATATA	11	80	43.9	0.2		
	GCTTCC	7	12	3.5	0.2	–	–
PHO	CGCACG	5	6	0.5	**1.5**		
	GCACGT	5	10	0.8	**4.4**		Pho4p
	CACGTG	5	12	0.9	**1.8**	GCACGTGGG	(bHLH)
	ACGTGG	5	8	0.7	**2.8**	(high affinity)	
	CGTGGG	3	5	0.5	0.5		
	CACGTT	5	7	1.2	0.3	GCACGTTTT	Pho4p
	ACGTTT	5	11	2.6	0.8	(medium affinity)	(bHLH)
	CTGCAC	4	8	1.0	**1.7**	–	–

van Helden J, Andre B, Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol. 1998 Sep 4;281(5):827-42.

Motif discovery

```
GATGGCTGCACCACGTGTATGC...ACGATGTCTCGC
CACATCGCATCACGTGACCAGT...GACATGGACGGC
GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA
TCTCGTTAGGACCATCACGTGA...ACAATGAGAGCG
CGCTAGCCCACGTGGATCTTGT...AGAATGGCCTAT
```



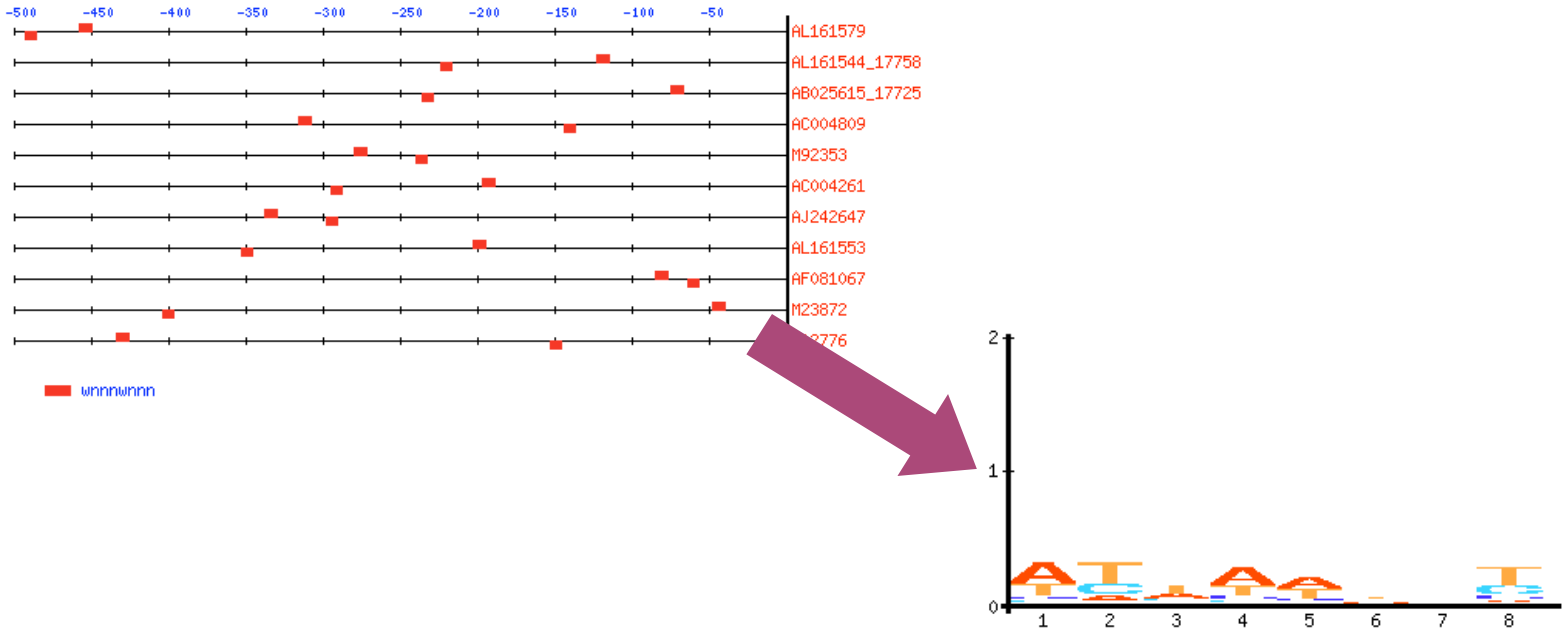
van Helden J, et al.. J Mol Biol. 1998 4;281(5):827-42.

Aerts S et al. Nucleic Acids Res. 2003 15;31(6):1753-64.

- *Lawrence CE, et al. Science. 1993 8;262(5131):208-14.*
- *Bailey T et al. Proceedings of ISMB, pp. 28-36, 1994.*

Gibbs sampling

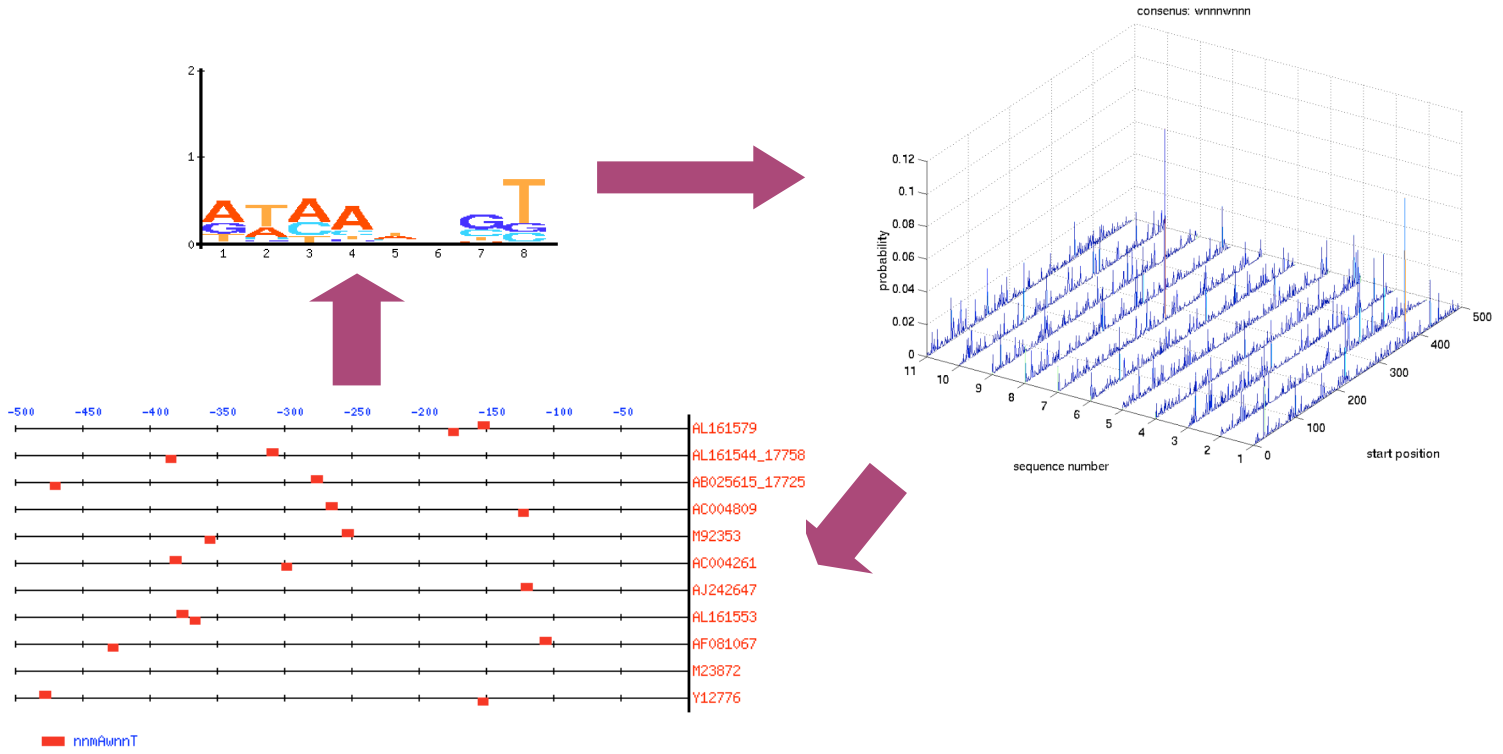
- Start from random set of motif positions
- Create initial motif model



Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol. 2002;9(2):447-64.

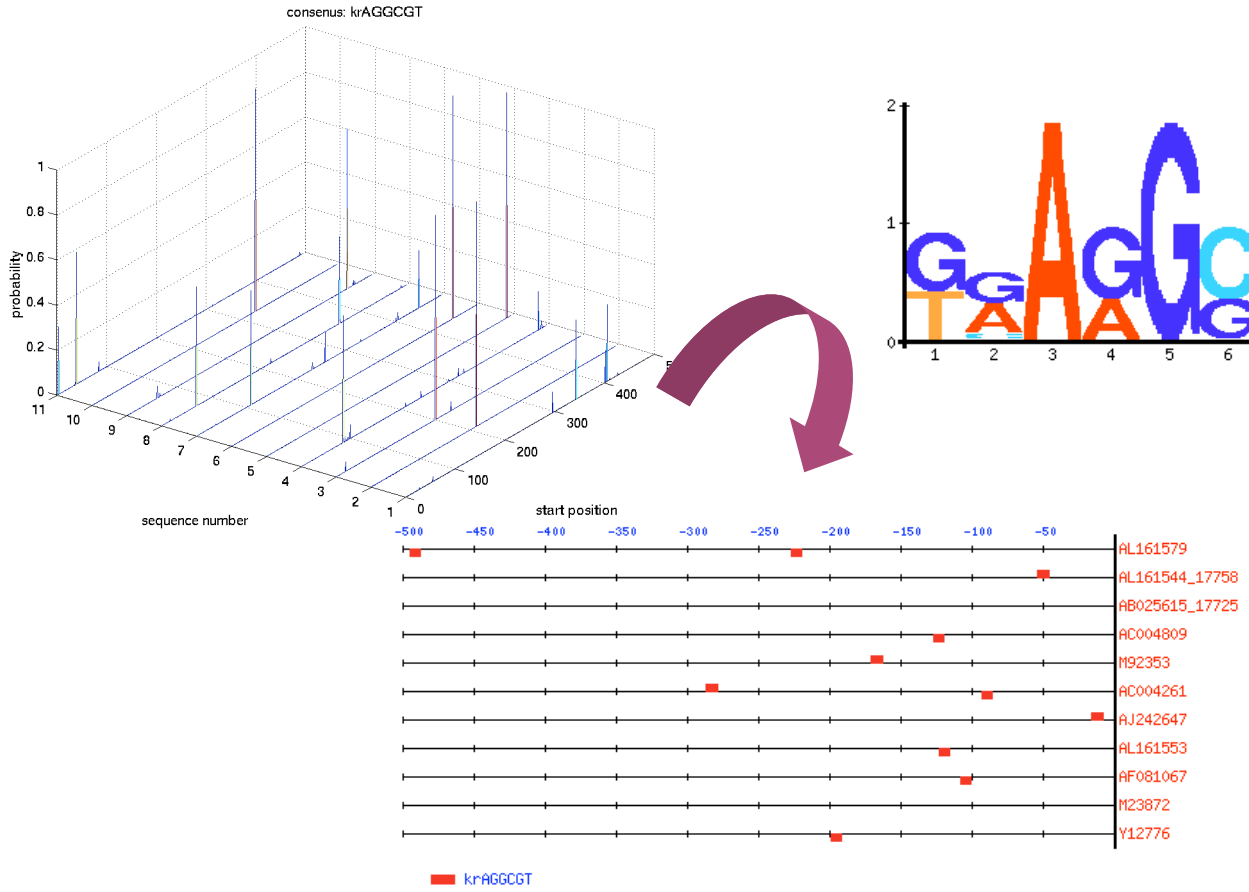
Gibbs sampling (ctd.)

1. Score sequences with current motif model and divide by score of the background model
2. Calculate distribution = probability of position being motif start
3. Sample new alignment position
4. Iterate for fixed number of steps



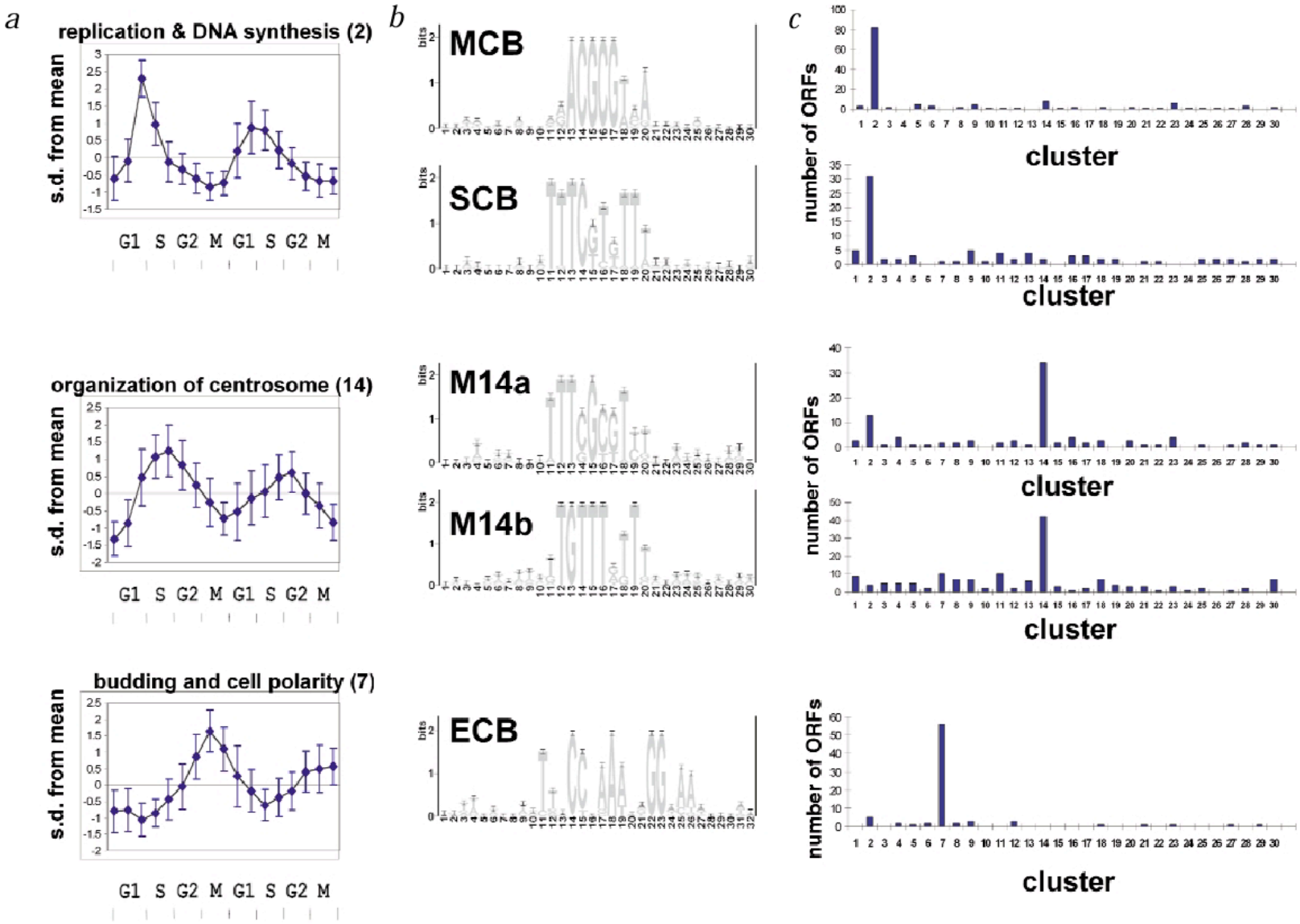
Gibbs sampling (ctd.)

Select best scoring positions from probability distribution to create motif and alingment



Variations: background models, gapped motifs, ChIP-chip data (MDScan)

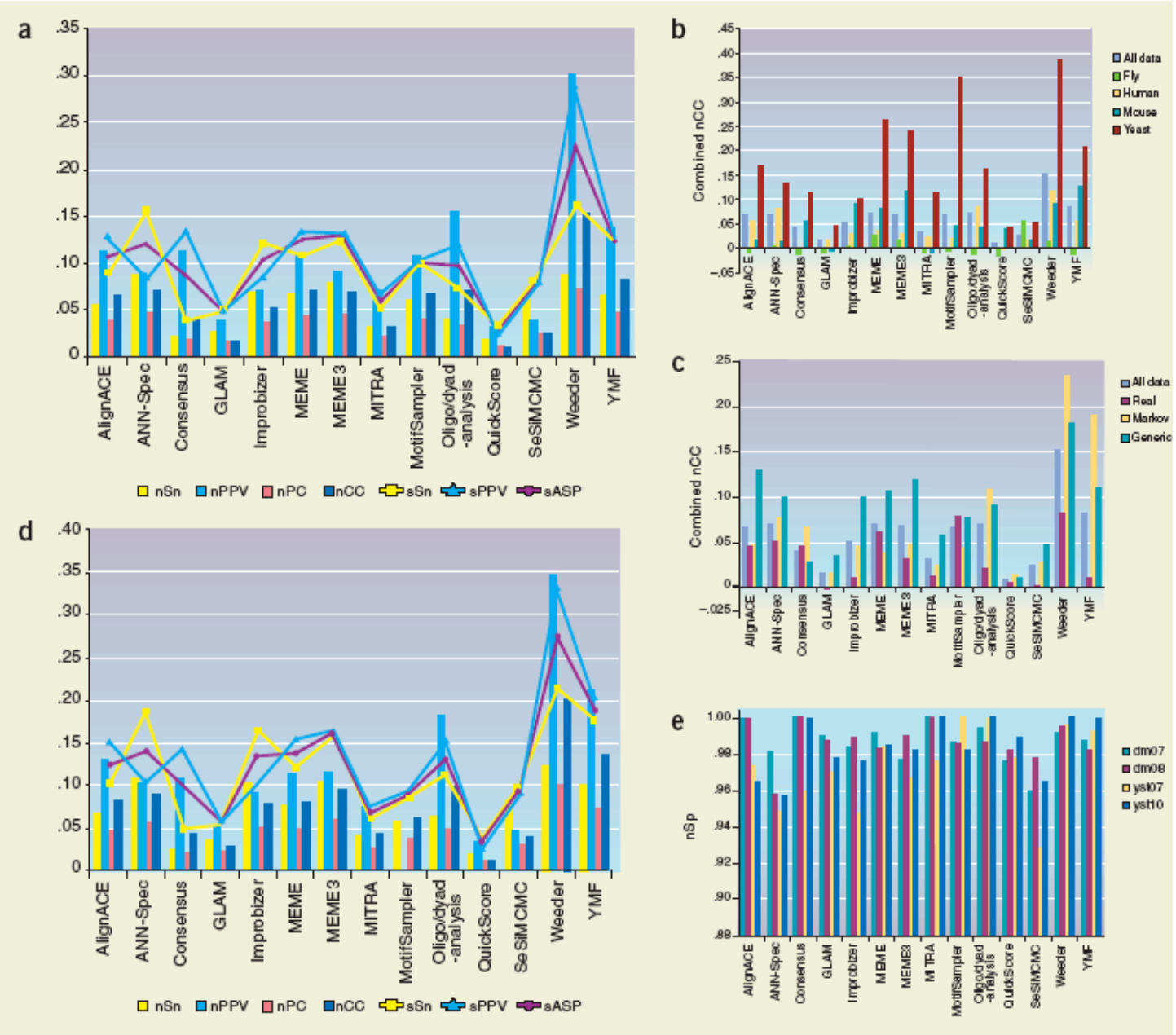
Gibbs sampling example



Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet. 1999 Jul;22(3):281-5.

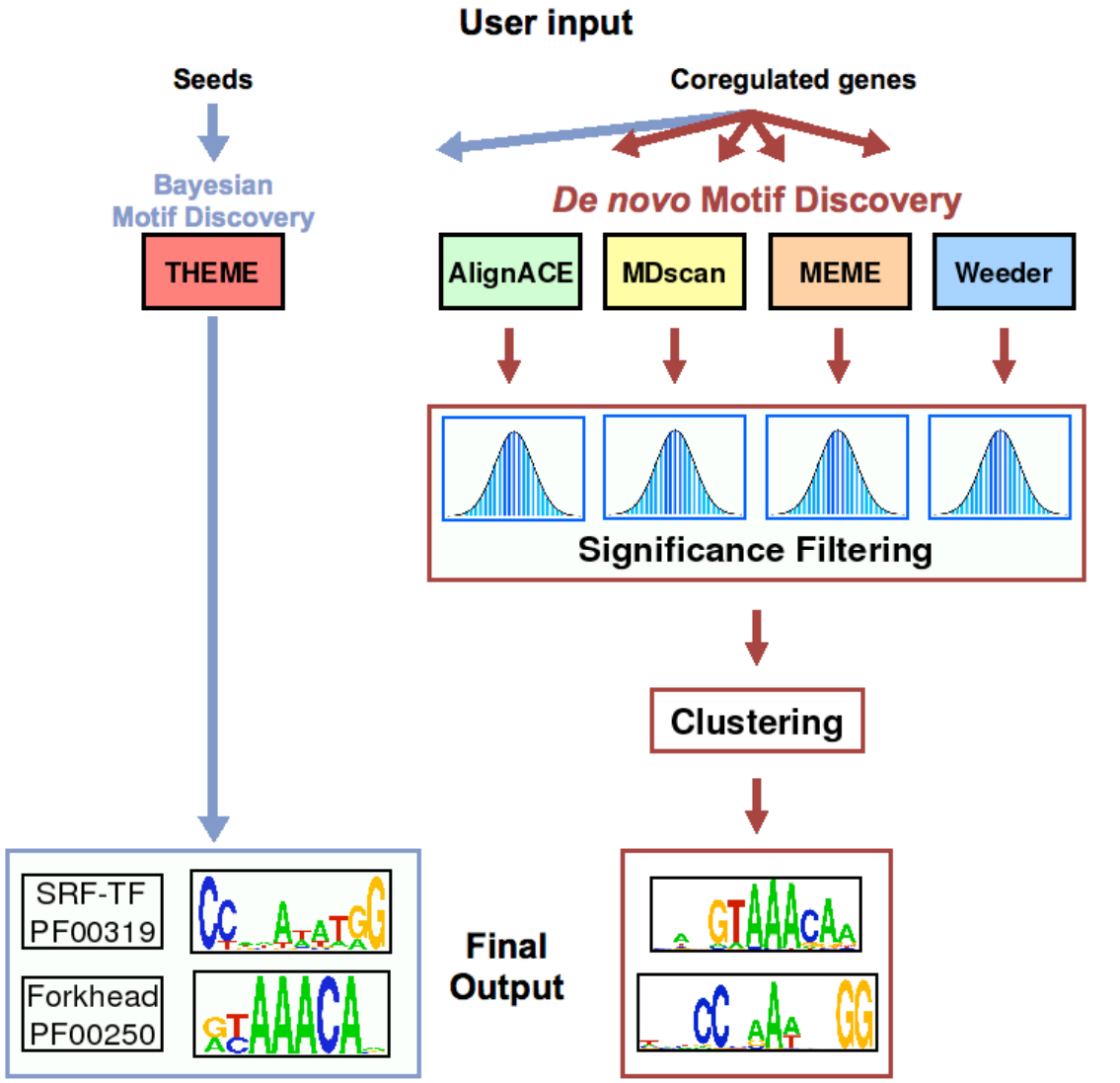
Assessing tools for motif discovery

Tompa et al. *Nature Biotechnology* (2005) 23:1



Combining multiple programs

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



Make PWM from known sites by motif discovery



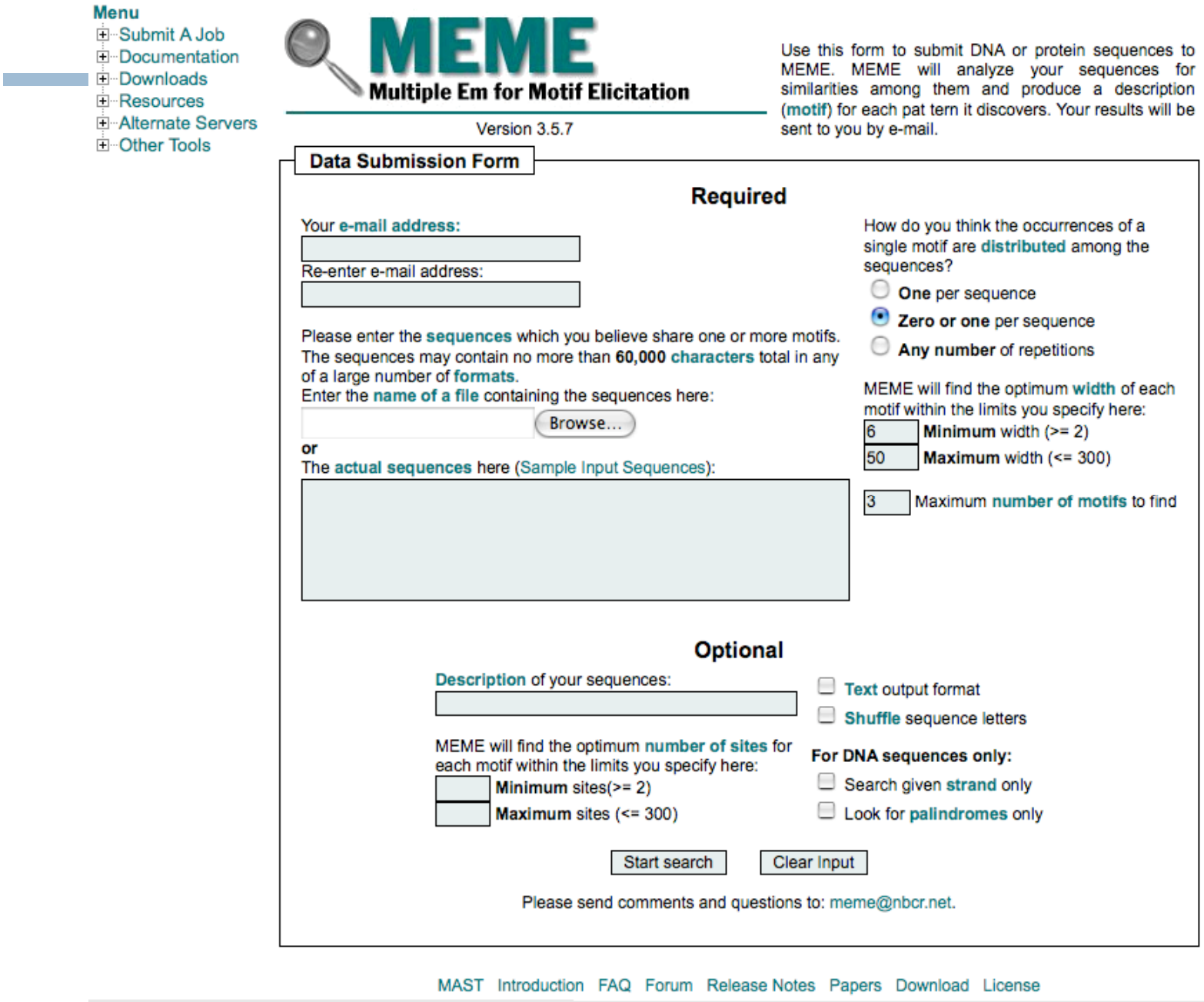
GCR1

Add NNNNN before and after

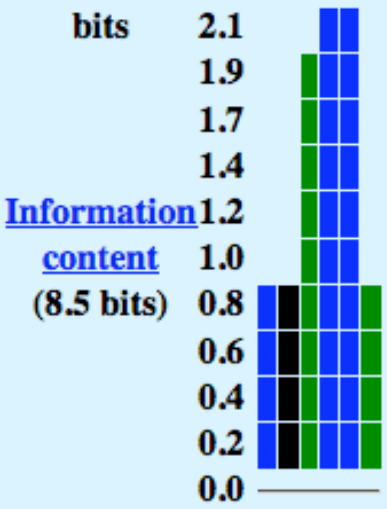
```
>YAL038W
NNNNNGGATTCCTNNNNN
>YAL038W
NNNNNCTTCCNNNNN
>YAL038W
NNNNNAGACATCGGGCTTCCACAATTTCGNNNNN
>YCR012W
NNNNNGACTTCCTGTCTTCCTATTGATTGCNNNNN
>YCR012W
NNNNNCTTCCNNNNN
>YCR012W
NNNNNCTTCCNNNNN
>YDR050C
NNNNNCATCCNNNNN
>YDR050C
NNNNNTCTGGCATCCAGTTTNNNNN
>YDR050C
NNNNNCTTCCNNNNN
>YDR050C
NNNNNAGCTTCCTCTATNNNNN
>YGR254W
NNNNNTCCACTAGGATAGNNNNN
>YHR174W
NNNNNCATCCNNNNN
>YOL086C
NNNNNACAATATGGACTTCCTCTTTTCTGGNNNNN
```

FASTA (replace \t by \r or \n)

```
>YAL038W
GGATTCCT
>YAL038W
CTTCC
>YAL038W
AGACATCGGGCTTCCACAATTTCG
>YCR012W
GACTTCCTGTCTTCCTATTGATTGC
>YCR012W
CTTCC
>YCR012W
CTTCC
>YDR050C
CATCC
>YDR050C
TCTGGCATCCAGTTT
>YDR050C
CTTCC
>YDR050C
AGCTTCCTCTAT
>YGR254W
TCCACTAGGATAG
>YHR174W
CATCC
>YOL086C
ACAATATGGACTTCCTCTTTTCTGG
```



http://meme.sdsc.edu/meme/meme.html



Multilevel consensus sequence: CATCCT / T

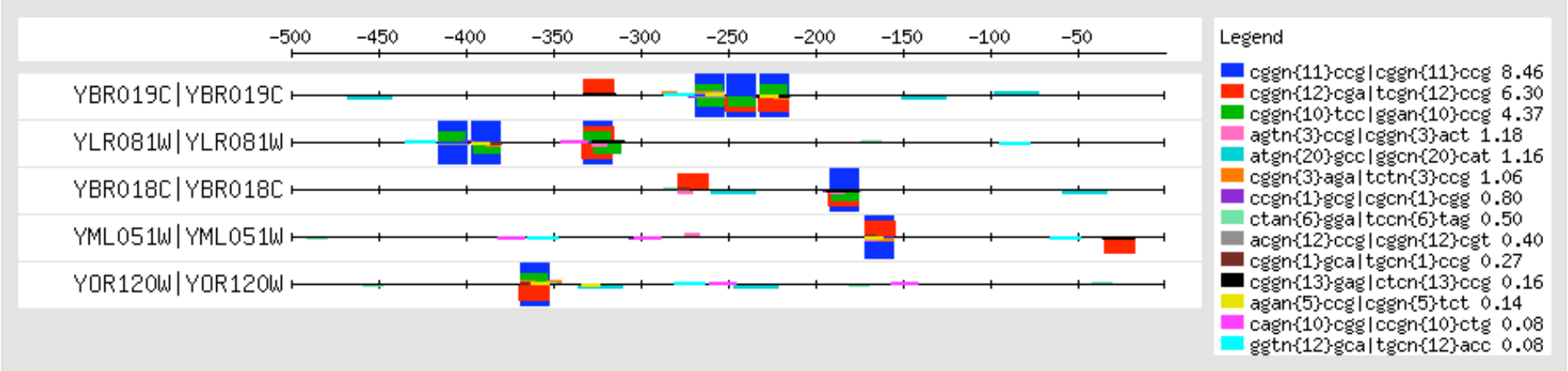
NAME	STRAND	START	P-VALUE	SITES
YOL086C	+	16	4.85e-04	ACAATATGGA CTTCCT CTTTTCTGGN
YCR012W	+	8	4.85e-04	NNNNNGA CTTCCT GTCTTCCTAT
YHR174W	+	6	1.55e-03	NNNNN CATCCN NNNN
YGR254W	-	12	1.55e-03	XXXXXC TATCCT AGTGGAXXXX
YDR050C	+	6	1.55e-03	NNNNN CATCCN NNNN
YAL038W	+	8	1.55e-03	NNNNNGG ATTCCT NNNNN

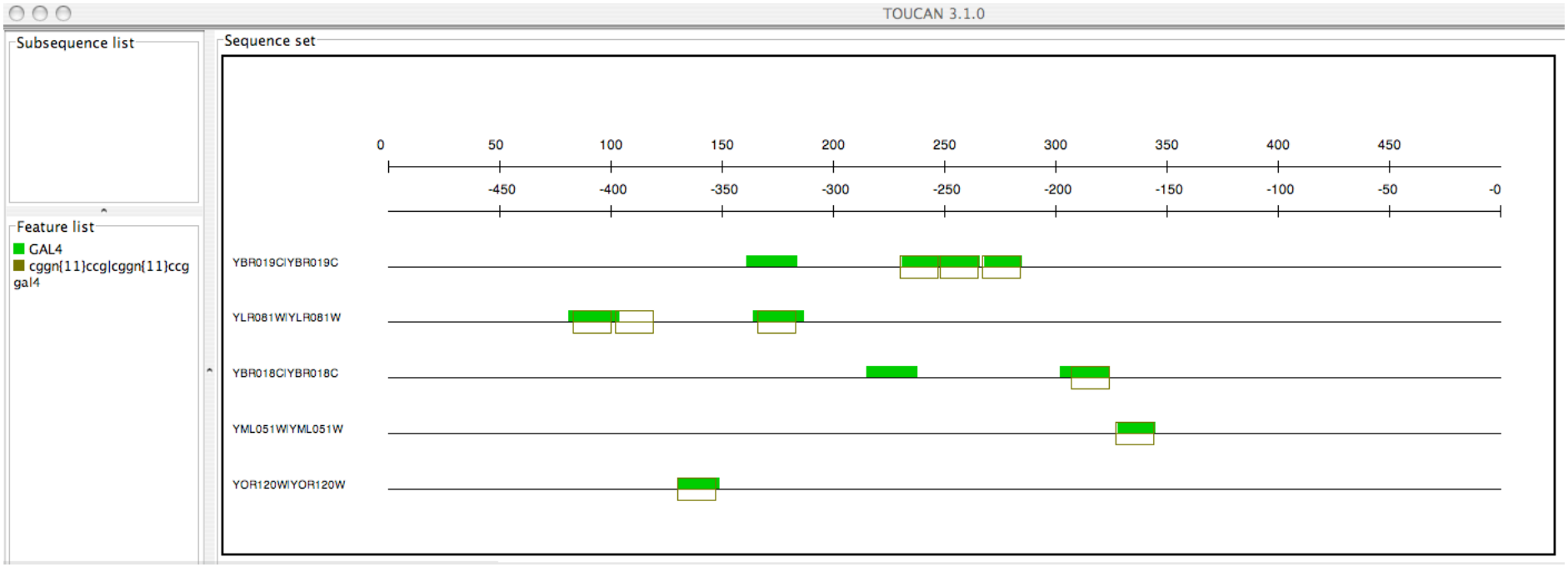
See also Position Specific Probability Matrix

Gal4

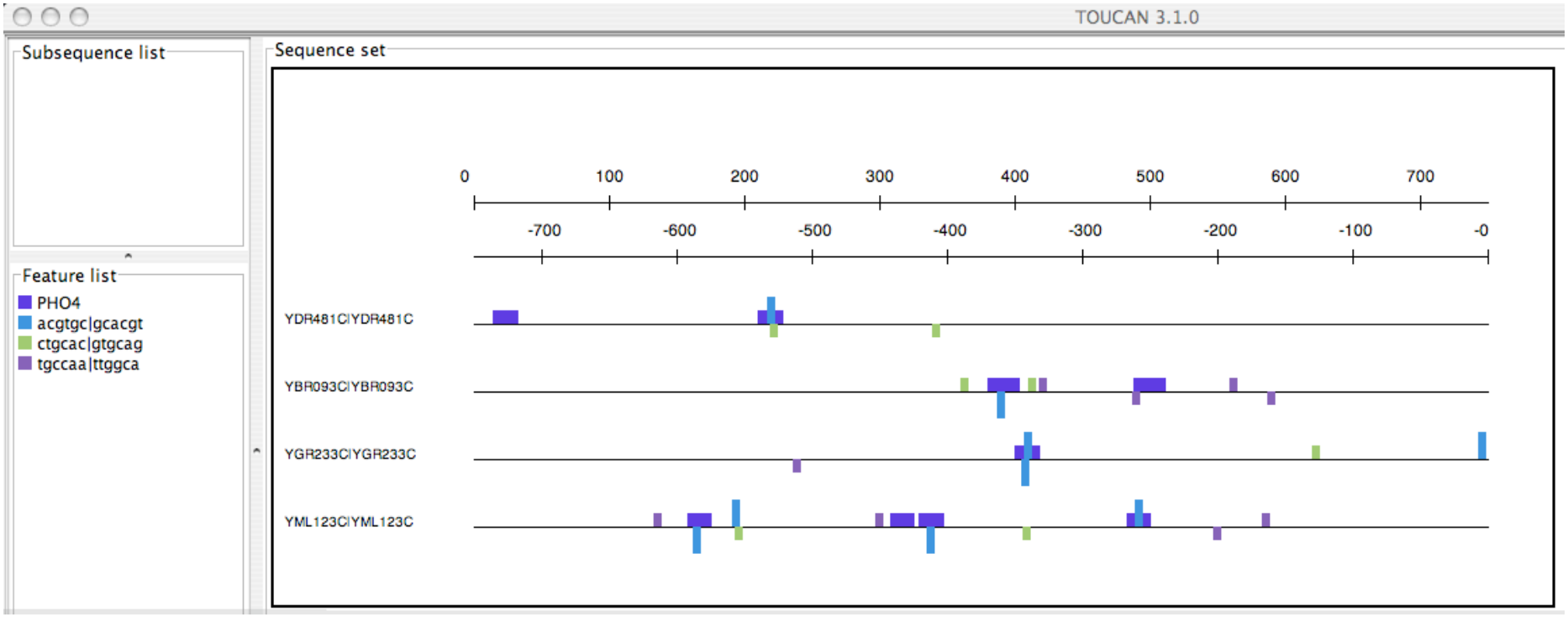
No over-represented hexamers found.

Dyad-analysis, Gal4



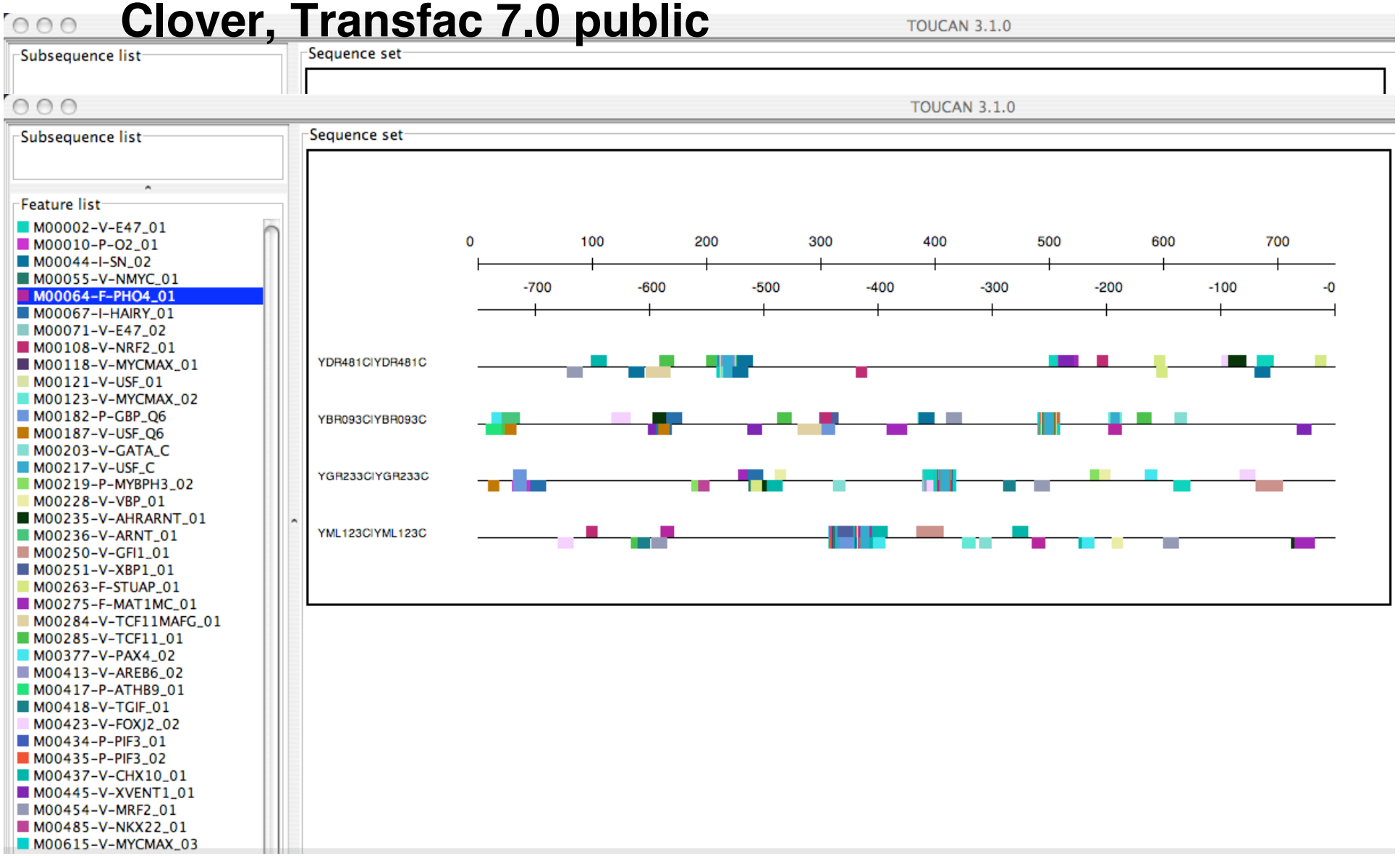


Pho4 - oligo-analysis

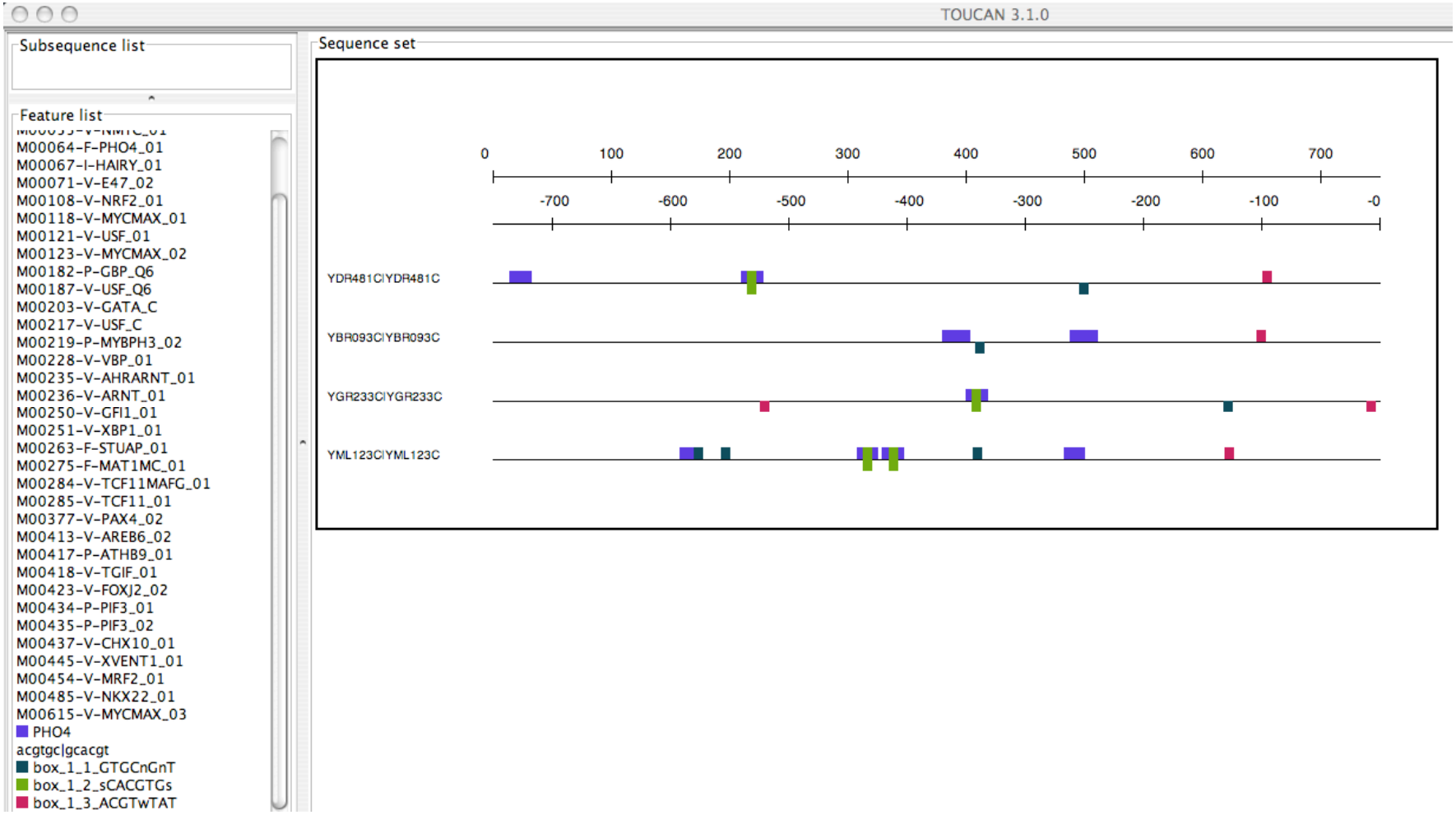


PHO4 - PWM over-representation

Clover, Transfac 7.0 public



PHO4 targets - MotifSampler



B

Alignment of conserved regions

```
YGL125W                                                  LEU3
S. cerevisiae    GAAAAAATAACAGCGACTTTTCTCCCGGTAGCGGGCCGTCGTTTAGTCATTCTATCCCTC
S. mikatae       AAAACATAACAGCGAATTTTCCTCCCGGTAGCGGGCCTTCGTTTAGTCATTCTCTCTCTT
S. bayanus       AAAAAATAACAGCGACTTTTCCCCCCGGTAGCGGGCCGTCGTTTAGTCATTCTCTCTCCC
S. kudriavzevii  GAAAAAAAACAACGGCGGCCTCCCCCGGTAGCGGGCCGTCGTTTAGTCATTCTCTCTCTC
                  ***** **** **   *** * *********************************

YOR108W                                  LEU3
S. cerevisiae    GCCATCATGGTCCGGTAACGGTCGTAGTGAATGACTCATATTTTTCCATCTCTTT
S. mikatae       GCCATCAAGGTCCGGTAACGGTCGTAGTGAATGACTCACATTTTCTTCGTTATTC
S. bayanus       ACCATTACGGTCCGGTAACGGACTTAGTGAATGATTCATCTTTTCTTCTTTTTTC
S. kudriavzevii  GTCGTTAAGGTCCGGTAACGGCCCTCAGCGAATGATTCATAATTTCATTTTTTTC
                 ***** * ***********  * ********** ***  ****    *** ***

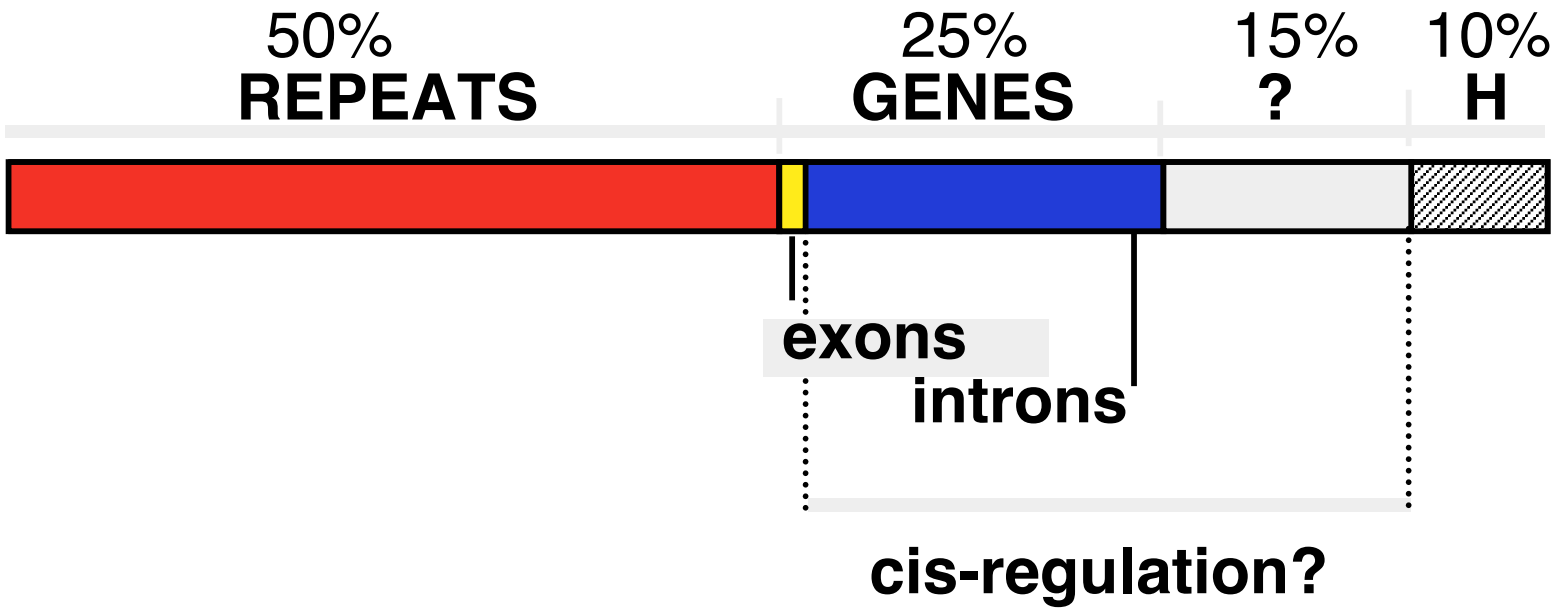
YMR108W                                          LEU3
S. cerevisiae    AACGCCTAGCCGCCGGAGCCTGCCGGTACCGGCTTGGCTTCAGTTGCTGATCTCGG
S. mikatae       CACAATGACACATACCTAACAGCCGGTACCGGCTTGAATGCCGCCGTTGGCTTCGG
S. bayanus       ATCTTCTAGTCACCGCAGTCTGCCGGTACCGGCTTGAATTCCGCCGTTGATCCTGG
S. kudriavzevii  CACATCTCTAGTCCGCGCTCTGCCGGTACCGGCTTAGACTAGCCACGAATCTCGGC
                  **  ***  * ****   ***********  ****  *** ** * **     **
```

Alignment of profiles

| | | |

Higher eukaryotes

1-2 % of the human genome encodes proteins

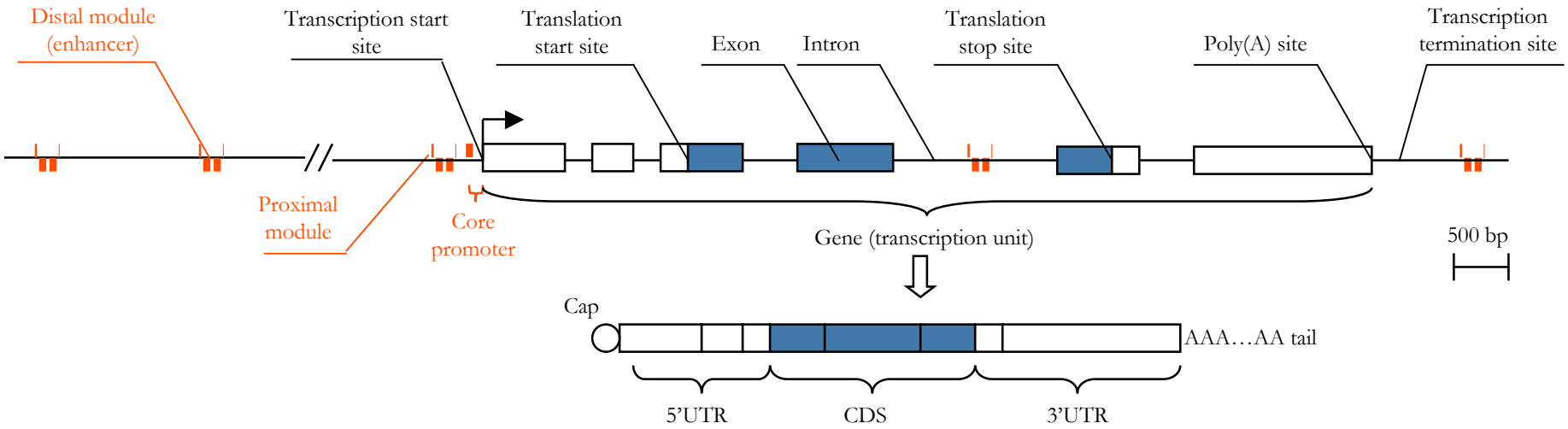


H = largely unsequenced heterochromatin

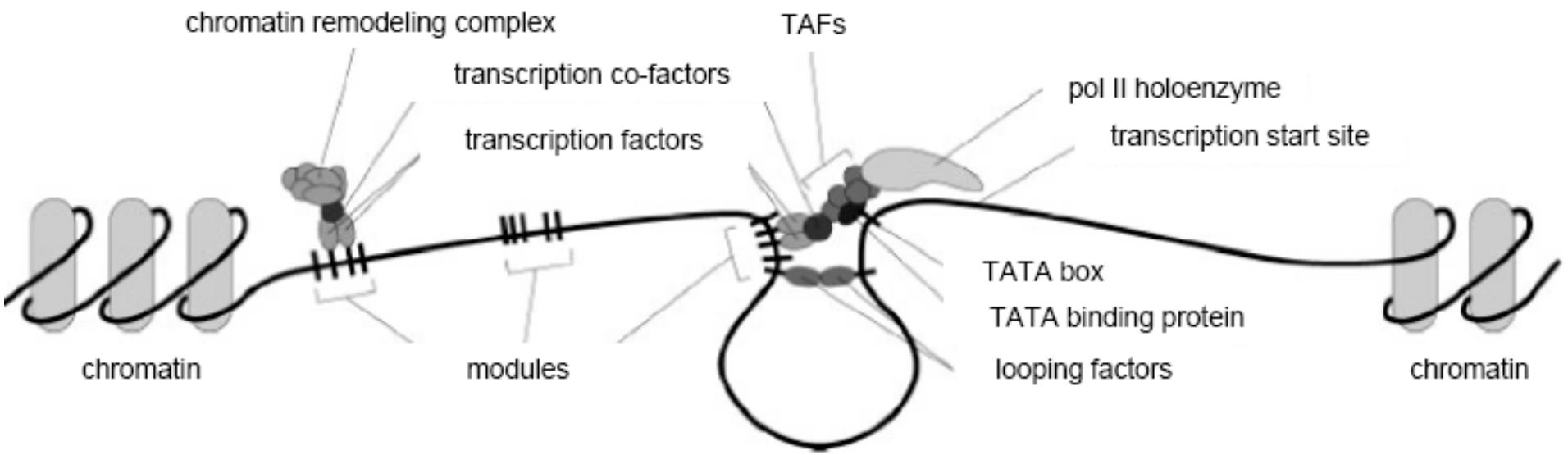
M. Markstein http://conferences.computer.org/bioinformatics/CSB2003/NOTES/Markstein.ppt

- Genes expressed at many different times and places during development
  - More signal integration
    - Simultaneous presence of multiple activators, each of which may be controlled by a separate signal
  - More combinatorial control
    - Many activators work in various combinations with other activators (and repressors) to regulate different sets of genes
  - Control by alternative sets of regulators
    - Same gene expressed in response to different sets of signals

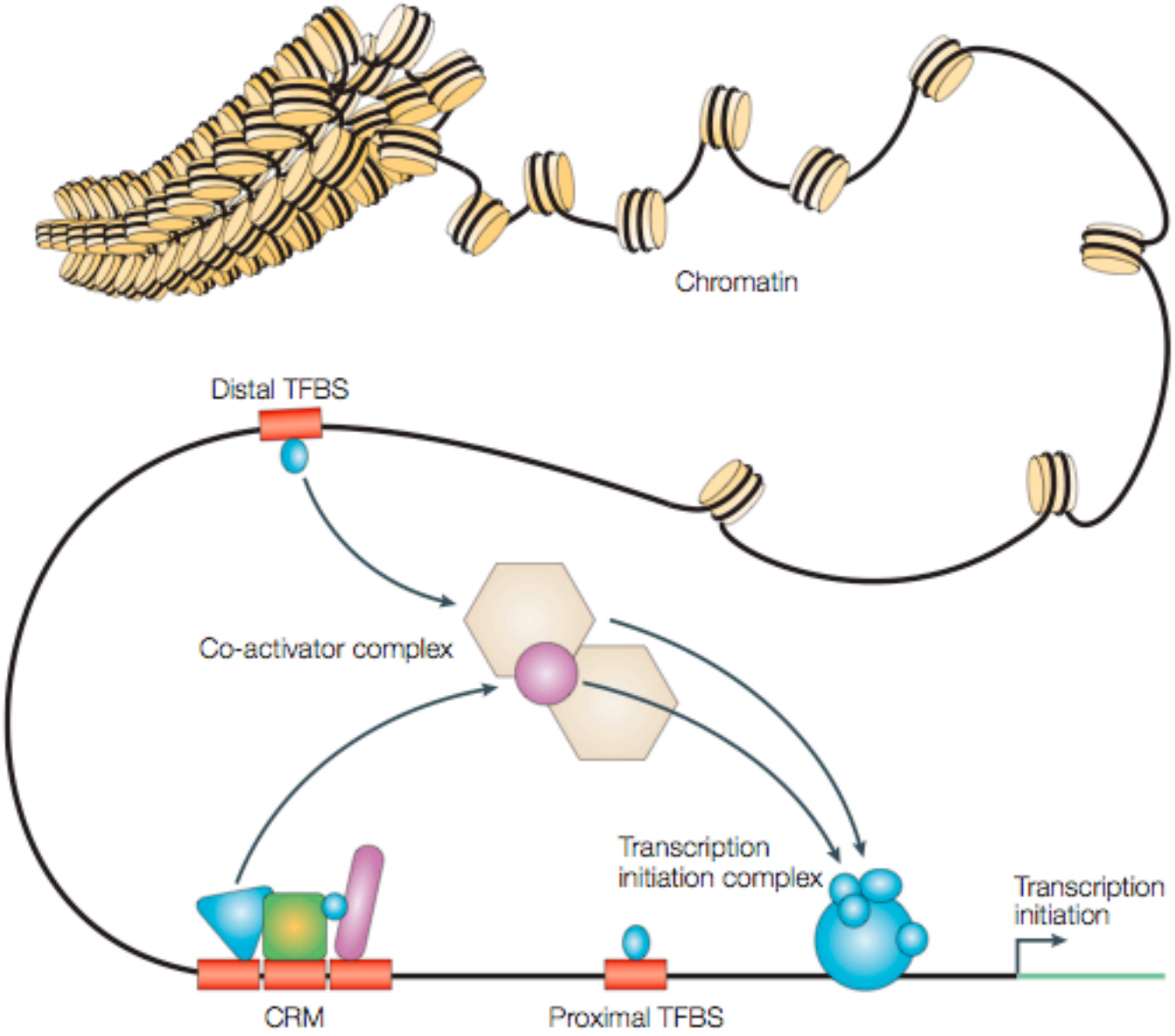
Cis-regulatory system



Cis-regulatory system



Wray et al. 2003


Chromatin
Distal TFBS
Co-activator complex
Transcription
initiation complex
Transcription
initiation
CRM
Proximal TFBS

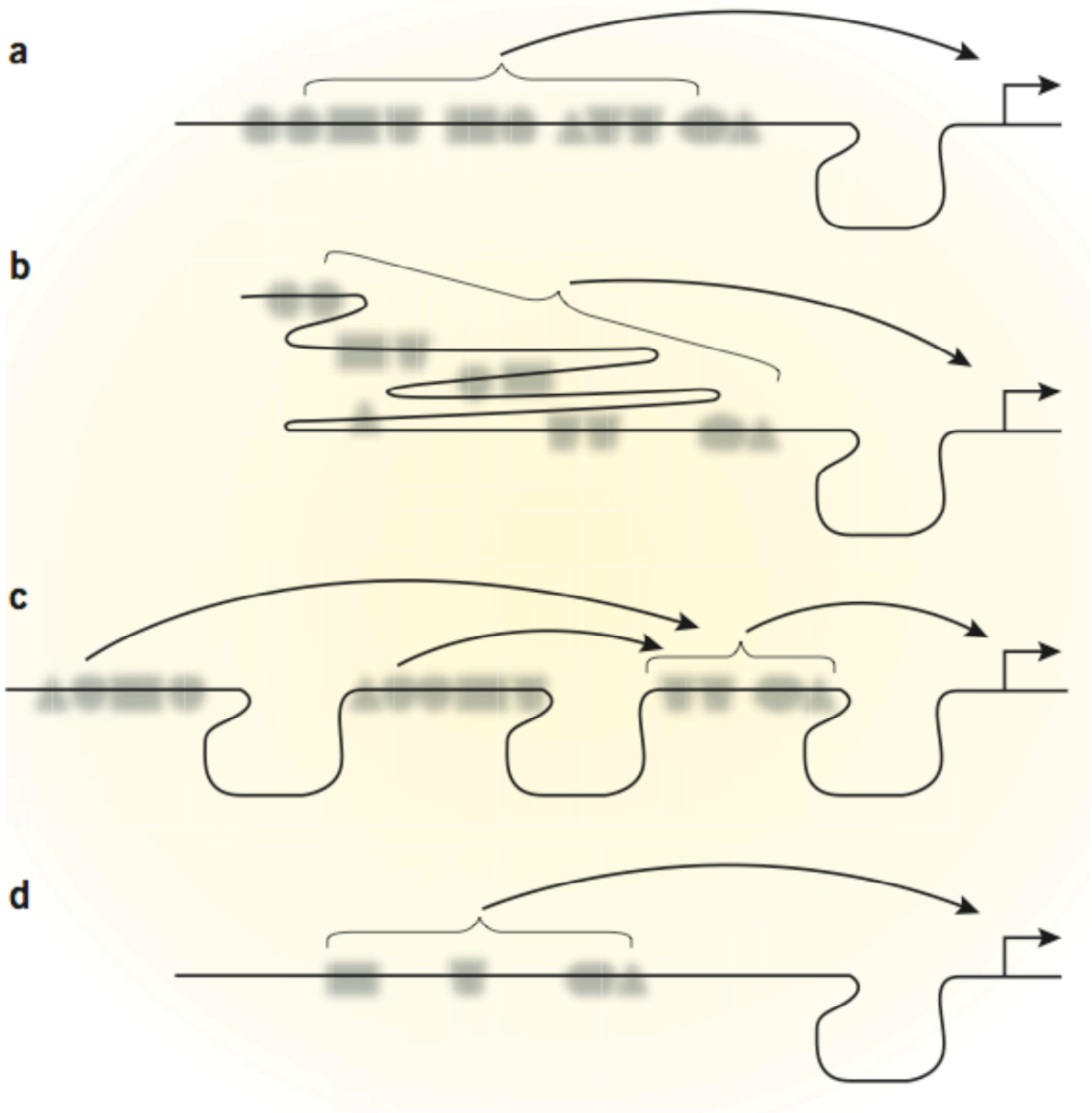
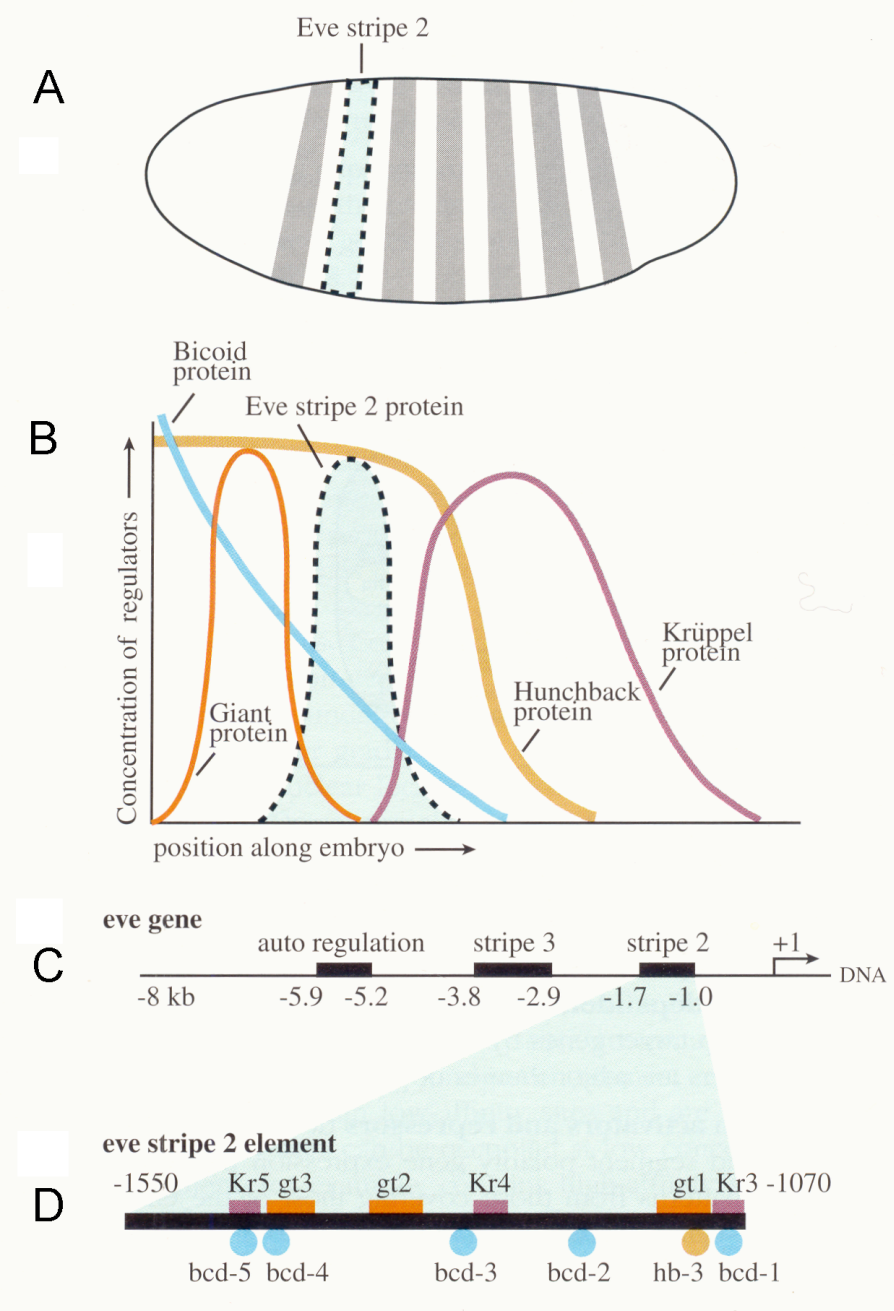


Figure 1 Different arrangements of *cis*-regulatory modules (CRMs). (**a**) 'Typical' CRM containing a compact arrangement of clustered binding sites. (**b**) Diffuse CRM in which binding sites are distributed over a large DNA segment. (**c**) Composite CRM. (**d**) Simple compact CRM without a dense cluster of binding sites.

M. Halfon, *Nat Genet* 2006

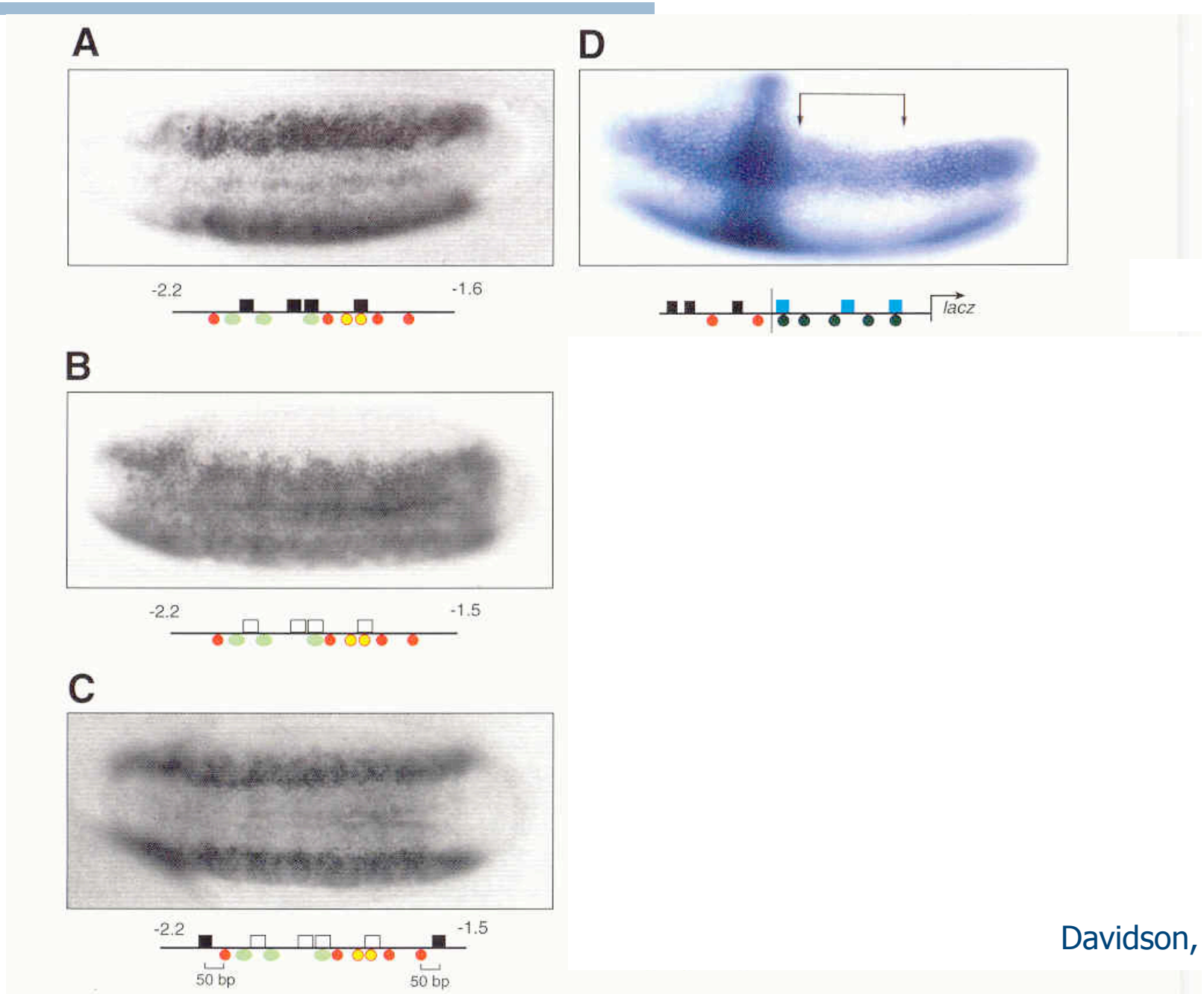
Module example: *eve*



- Features of CRMs:
  - 4-5 different factors per module
  - Length several 100's bp
  - TFs in competition-> TFBS may overlap
  - Multiple binding sites for same factor
  - Often conserved
  - Autonomous
  - Independent of location and orientation

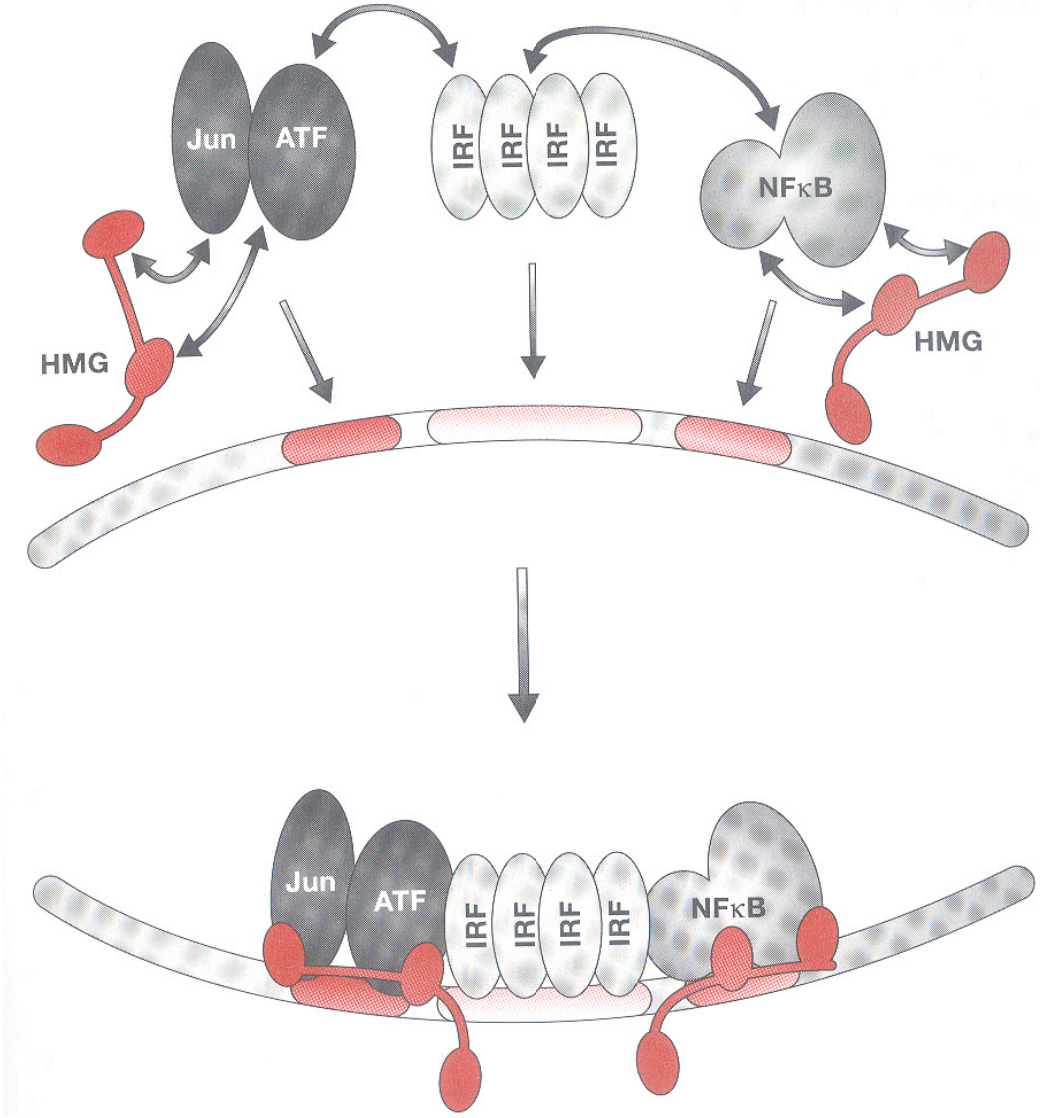
Caroll, 2001

Module example: *rho* and *eve*

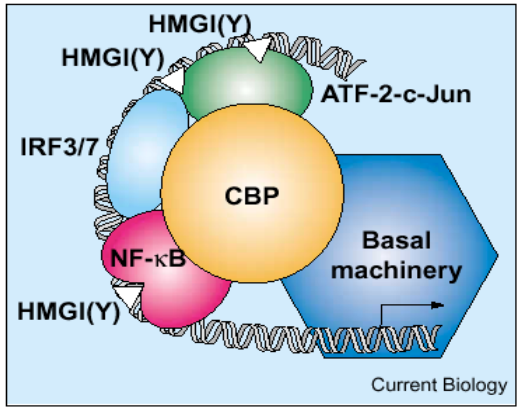


Davidson, 2001

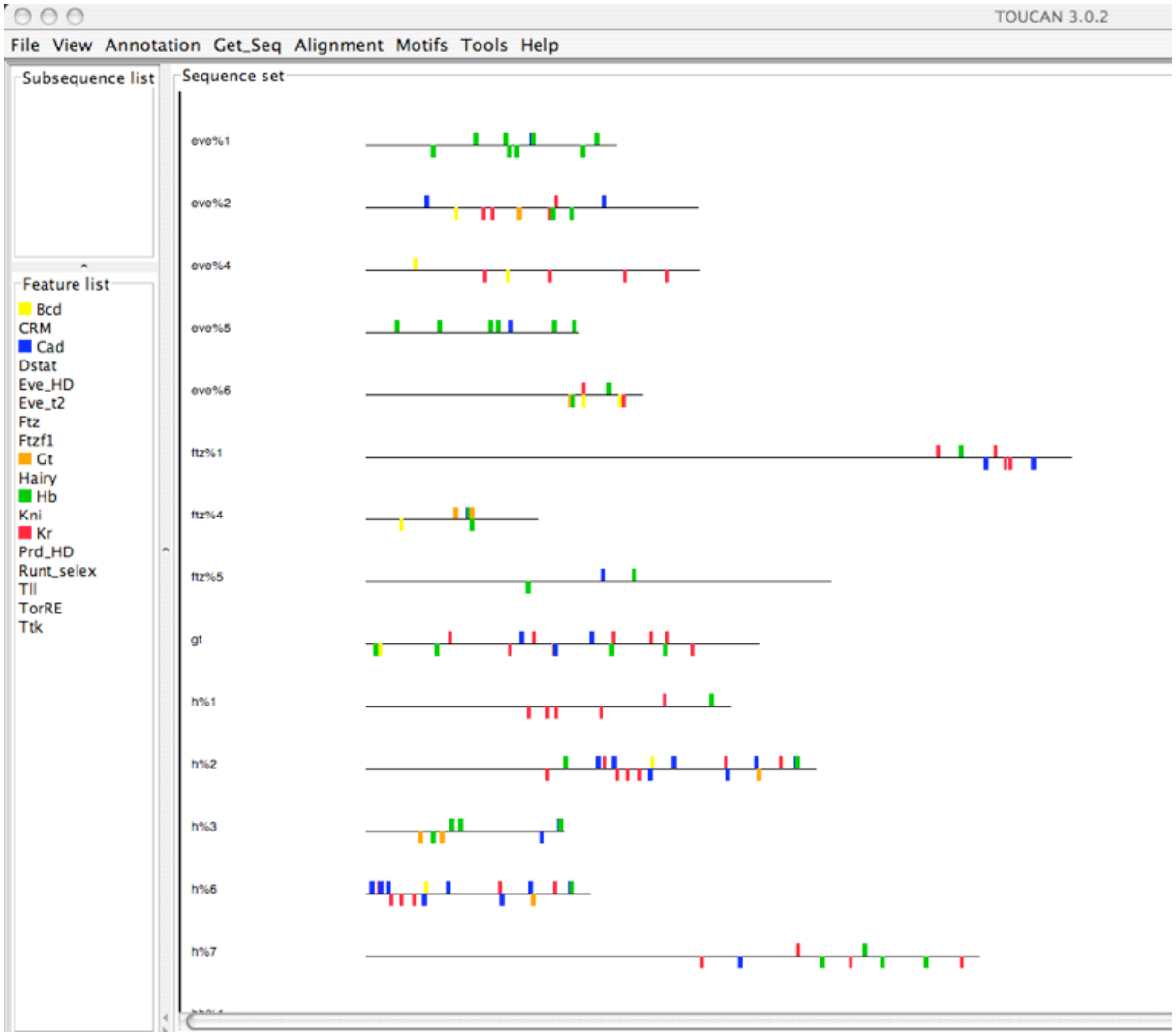
Module example: human interferon beta enhancer



“Enhanceosome” versus “billboard” model



Courey AJ (2001) Curr Biol. 11(7):R250-2



Example of billboard model: Functional conservation despite lack of seq conserv.

OPEN ACCESS Freely available online

PLoS GENETICS

Sepsid *even-skipped* Enhancers Are Functionally Conserved in *Drosophila* Despite Lack of Sequence Conservation

Emily E. Hare[1☯], Brant K. Peterson[1,2☯], Venky N. Iyer[1], Rudolf Meier[3], Michael B. Eisen[1,2,4,5*]

1 Department of Molecular and Cell Biology, University of California Berkeley, Berkeley, California, United States of America, **2** Center for Integrative Genomics, University of California Berkeley, Berkeley, California, United States of America, **3** Department of Biological Sciences, National University of Singapore, Singapore, **4** Genomics Division, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **5** California Institute for Quantitative Biosciences, Berkeley, California, United States of America

Abstract

The gene expression pattern specified by an animal regulatory sequence is generally viewed as arising from the particular arrangement of transcription factor binding sites it contains. However, we demonstrate here that regulatory sequences whose binding sites have been almost completely rearranged can still produce identical outputs. We sequenced the *even-skipped* locus from six species of scavenger flies (Sepsidae) that are highly diverged from the model species *Drosophila melanogaster*, but share its basic patterns of developmental gene expression. Although there is little sequence similarity between the sepsid *eve* enhancers and their well-characterized *D. melanogaster* counterparts, the sepsid and *Drosophila* enhancers drive nearly identical expression patterns in transgenic *D. melanogaster* embryos. We conclude that the molecular machinery that connects regulatory sequences to the transcription apparatus is more flexible than previously appreciated. In exploring this diverse collection of sequences to identify the shared features that account for their similar functions, we found a small number of short (20–30 bp) sequences nearly perfectly conserved among the species. These highly conserved sequences are strongly enriched for pairs of overlapping or adjacent binding sites. Together, these observations suggest that the local arrangement of binding sites relative to each other is more important than their overall arrangement into larger units of *cis*-regulatory function.

Citation: Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid *even-skipped* Enhancers Are Functionally Conserved in *Drosophila* Despite Lack of Sequence Conservation. PLoS Genet 4(6): e1000106. doi:10.1371/journal.pgen.1000106

Editor: Norbert Perrimon, Harvard Medical School, Howard Hughes Medical Institute, United States of America

Received February 28, 2008; **Accepted** May 22, 2008; **Published** June 27, 2008

This is an open-access article distributed under the terms of the Creative Commons Public Domain declaration which stipulates that, once placed in the public domain, this work may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose.

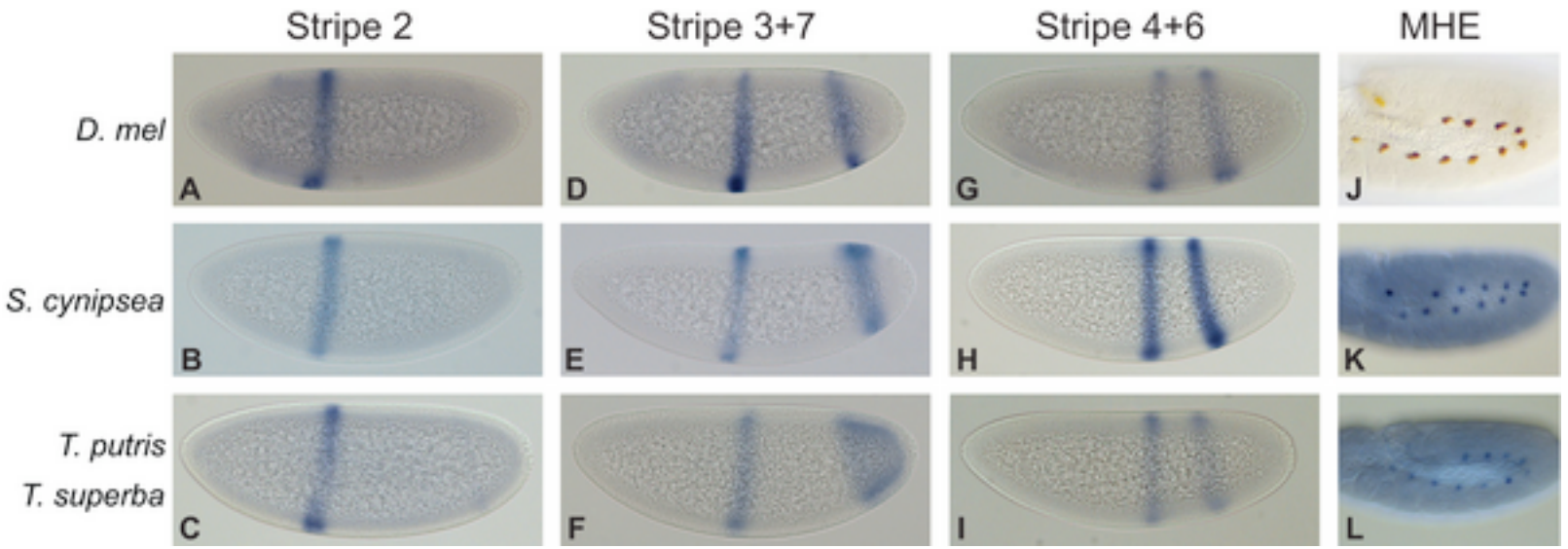
Funding: This work was supported by NIH grant HG002779 to MBE by the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and by NSF Grant No. 0334948 to RM. EEH was supported by an NSF Graduate Research Fellowship. Sequencing was conducted under a Community Sequencing Grant from the Department of Energy's Joint Genome Institute.

Competing Interests: The authors have declared that no competing interests exist.

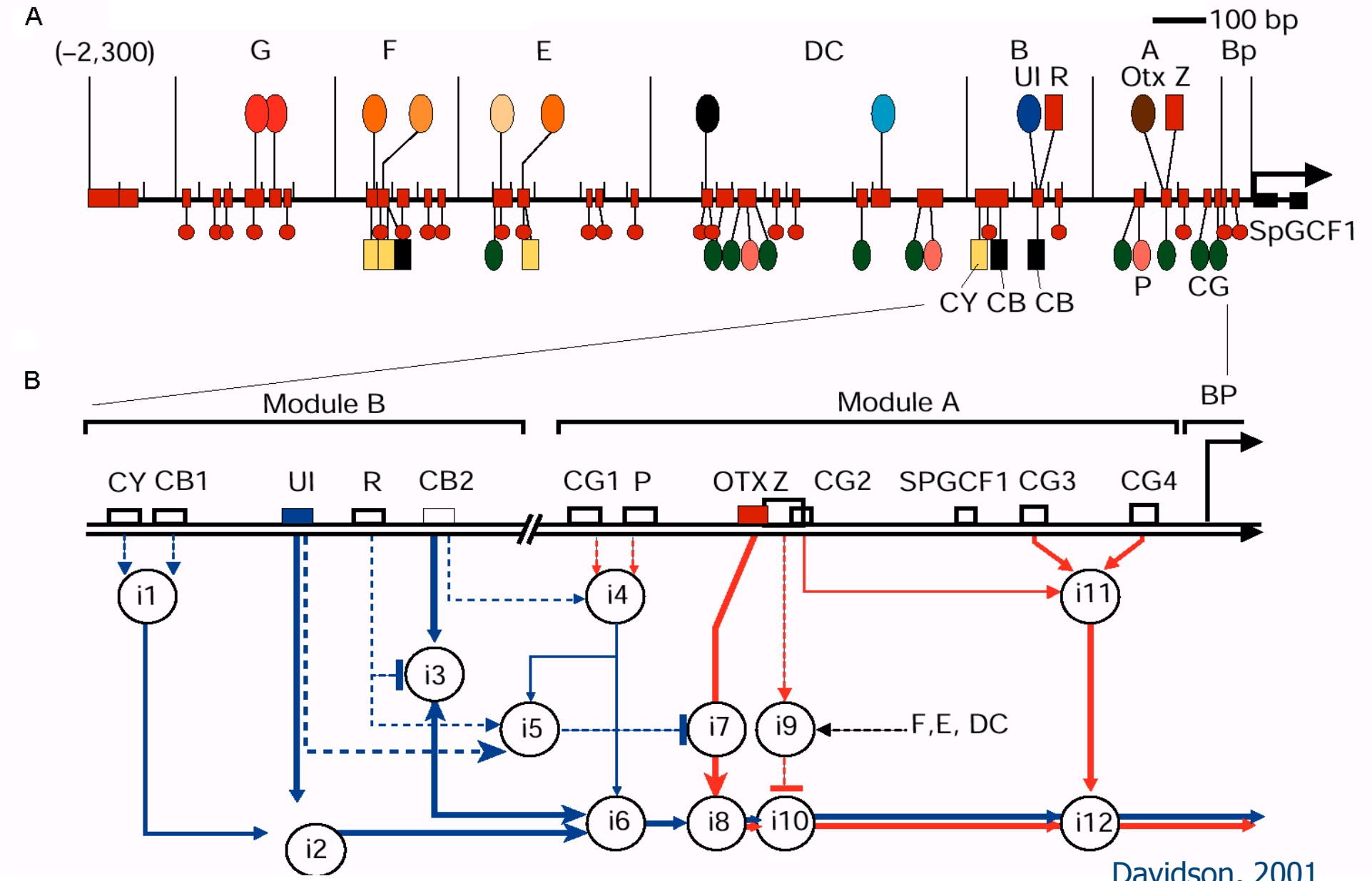
* E-mail: mbeisen@lbl.gov

☯ These authors contributed equally to this work


Stripe 2
0
400
800
1200
1600
2000
2400
2800
3200
3600
FOOTPRINTS
D. mel
D. ere
D. pse
D. vir
D. moj
D. gri
S. punctum
S. cynipsea
D. sp.
T. superba
T. minor
T. putris
*
BCD
ln(p) < -6
HB
ln(p) < -6
GT
ln(p) < -5.5
KR
ln(p) < -6
SLP1
ln(p) < -6

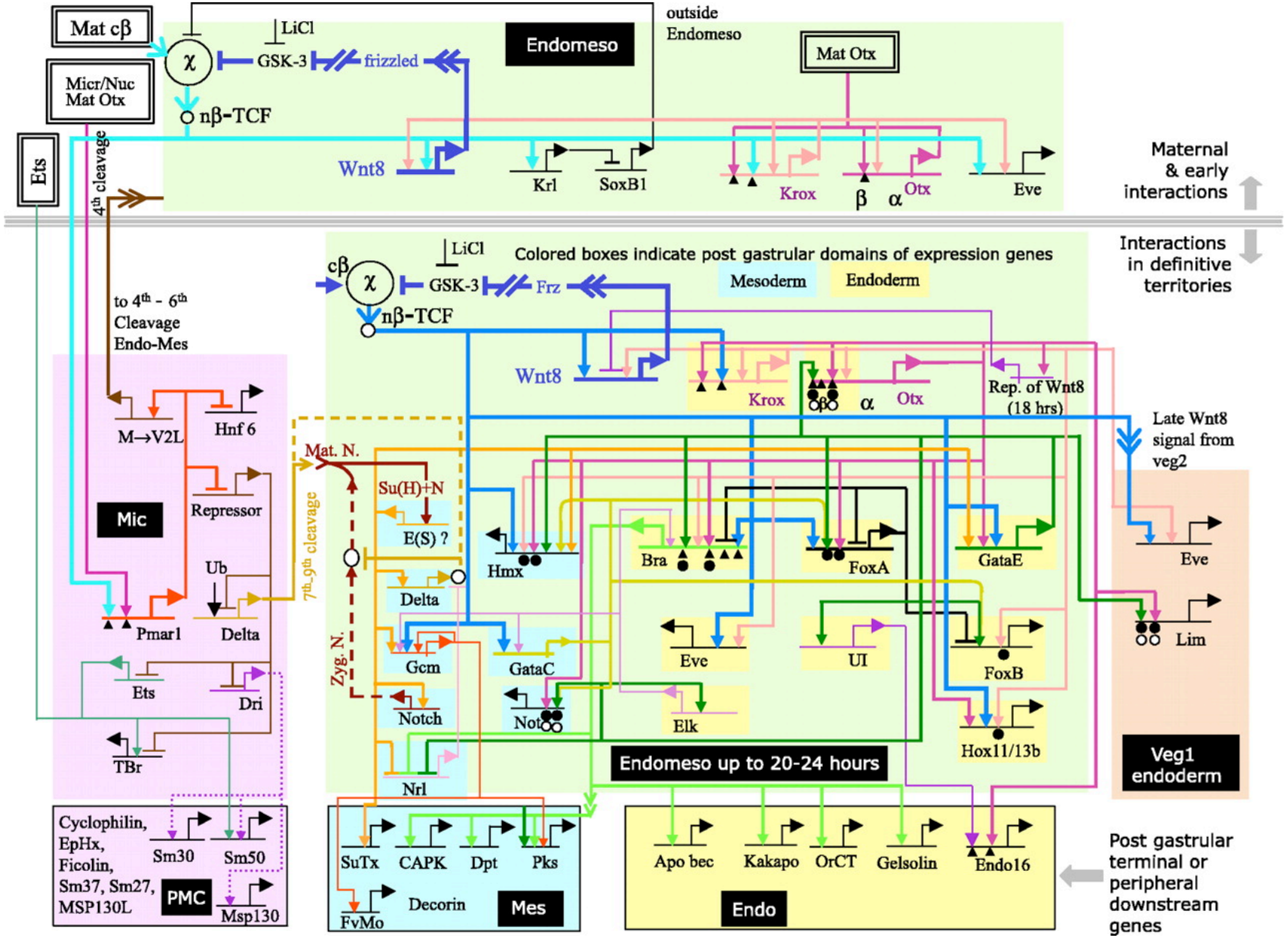

Stripe 2
Stripe 3+7
Stripe 4+6
MHE
D. mel
S. cynipsea
T. putris
T. superba
A
B
C
D
E
F
G
H
I
J
K
L

Example: Endo16 (*S.purpuratus*)



Davidson, 2001

Gene-regulatory network



Problems

- Problem: (very) large sequence space and “futility theorem”
  - 1000 fold excess of false positive predictions
- Solutions
  - Analyse only proximal promoters
    - Kellis et al
    - E2F case
  - Use phylogenetic footprinting
  - Use motif combinations
    - Scanning (cfr yeast)
    - discovery
  - Strong statistics with gene expression data (e.g., Lever)
    - Over-representation alone is not sufficient (signal/noise to low)
  - Use ChIP data

Prediction of *cis*-regulatory modules using input PWMs

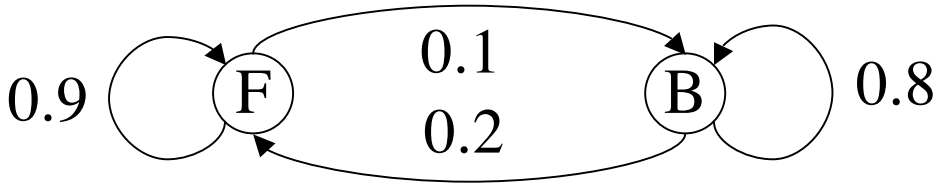
Motif Clusters

- Sliding window
  - E.g., sum log-odds scores
  - E.g., 500 bp
  - Stepsize 250 bp
  - minimally 3 sites
- Hidden Markov Model
  - E.g., cister, comet, cluster-buster, ahab, stubb, …

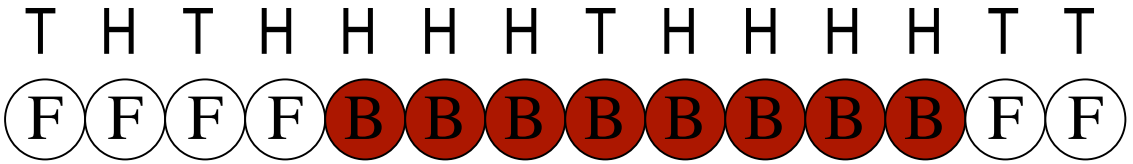
Hidden Markov Models (HMM)

Coin flip

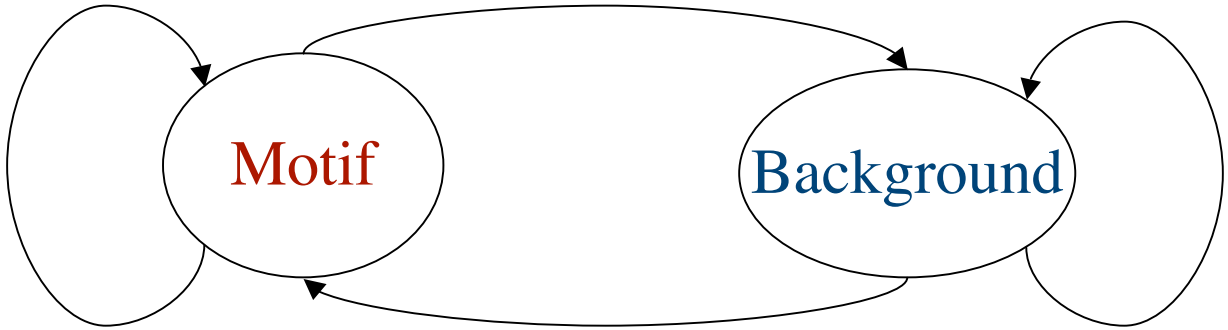
- Emission: P(H| F)=0.5 , P(H| B)=0.9
- Transition

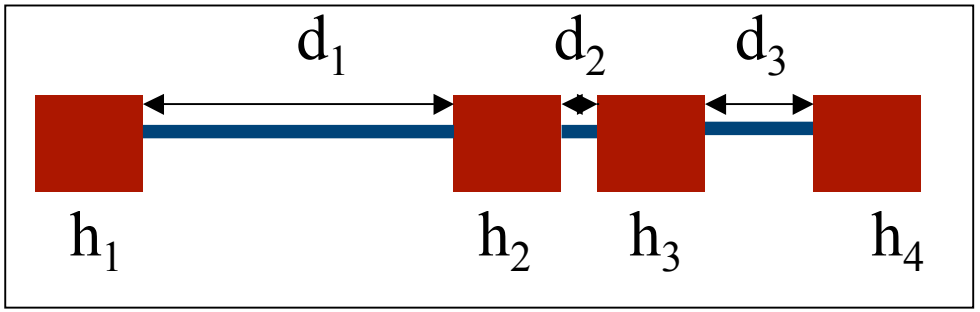


- Observe
- States



HMM for motif cluster

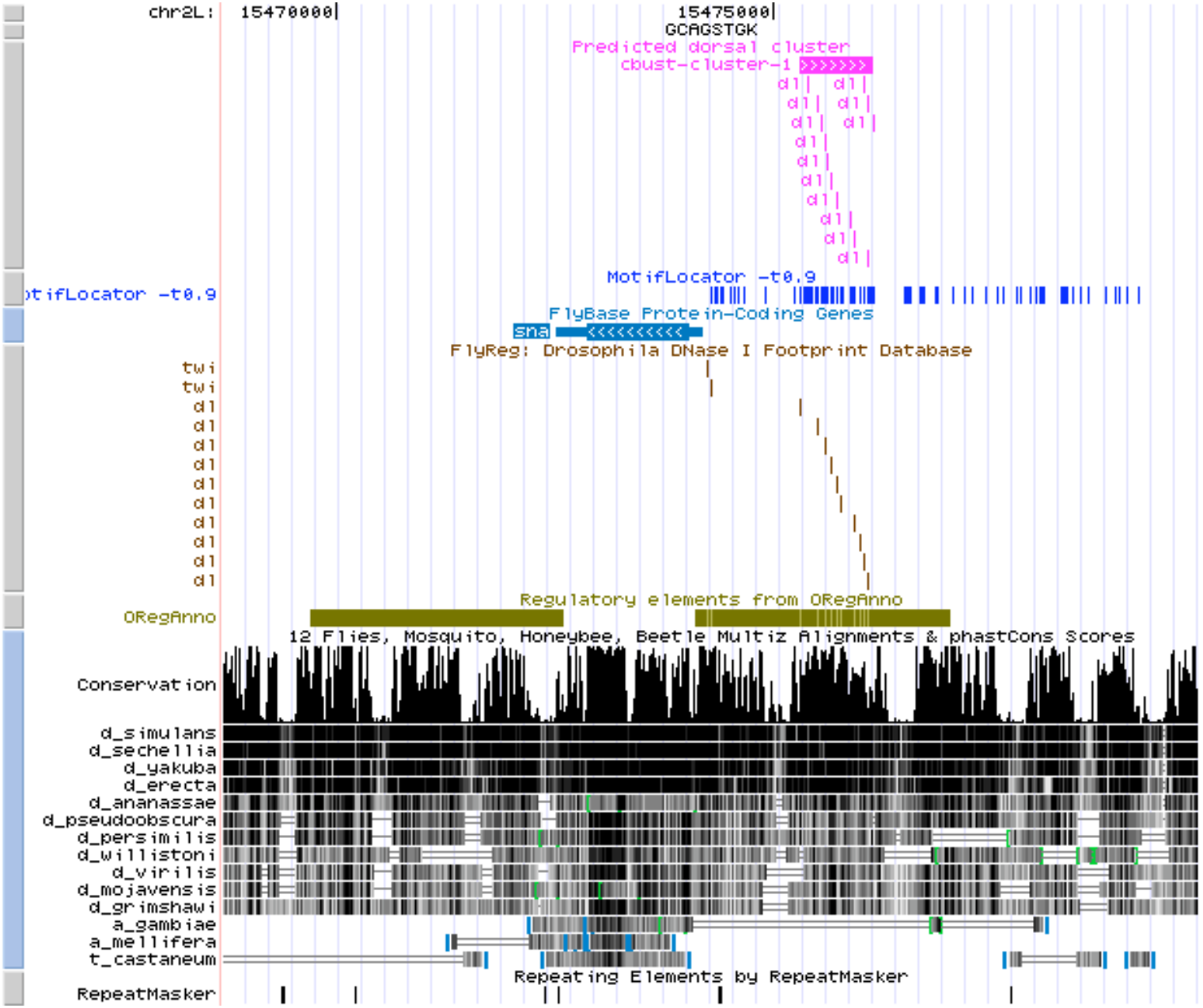




$$M = h_1 + h_2 + h_3 + h_4 + g(d_2 + d_3 + d_4)$$

The "match score" (M) for the cluster is the sum of the log-odds scores.

Homotypic clusters

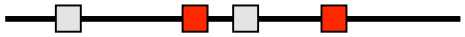


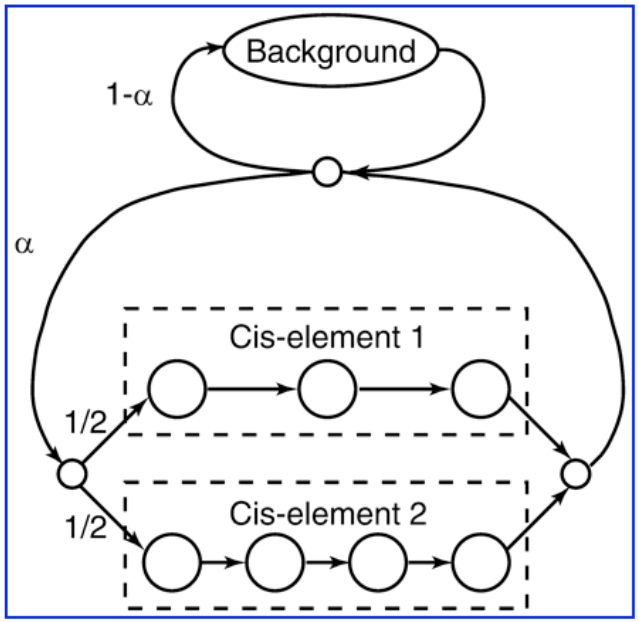
HMM for module scanning

DNA sequence segment

$$\text{score}\left(\text{segment}\right) = \ln\left[\frac{\text{Prob}\left(\text{segment} \mid \text{cluster model}\right)}{\text{Prob}\left(\text{segment} \mid \text{random model}\right)}\right]$$

Cluster model:



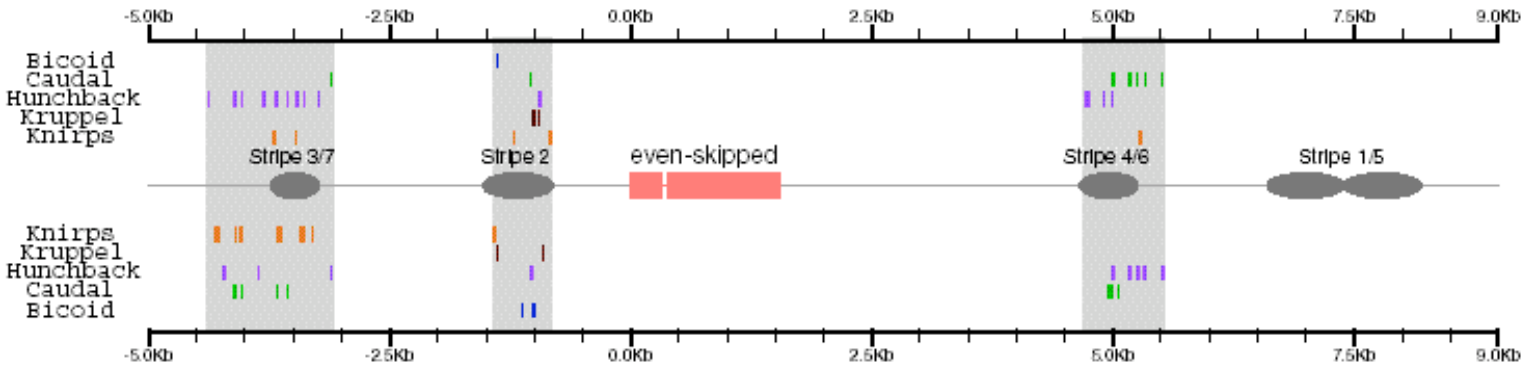


HMM: Cister, COMET, Ahab, Stubb, ClusterBuster, MCAST

Combined statistical significance: MSCAN

Example: sliding window approach in *Drosophila*

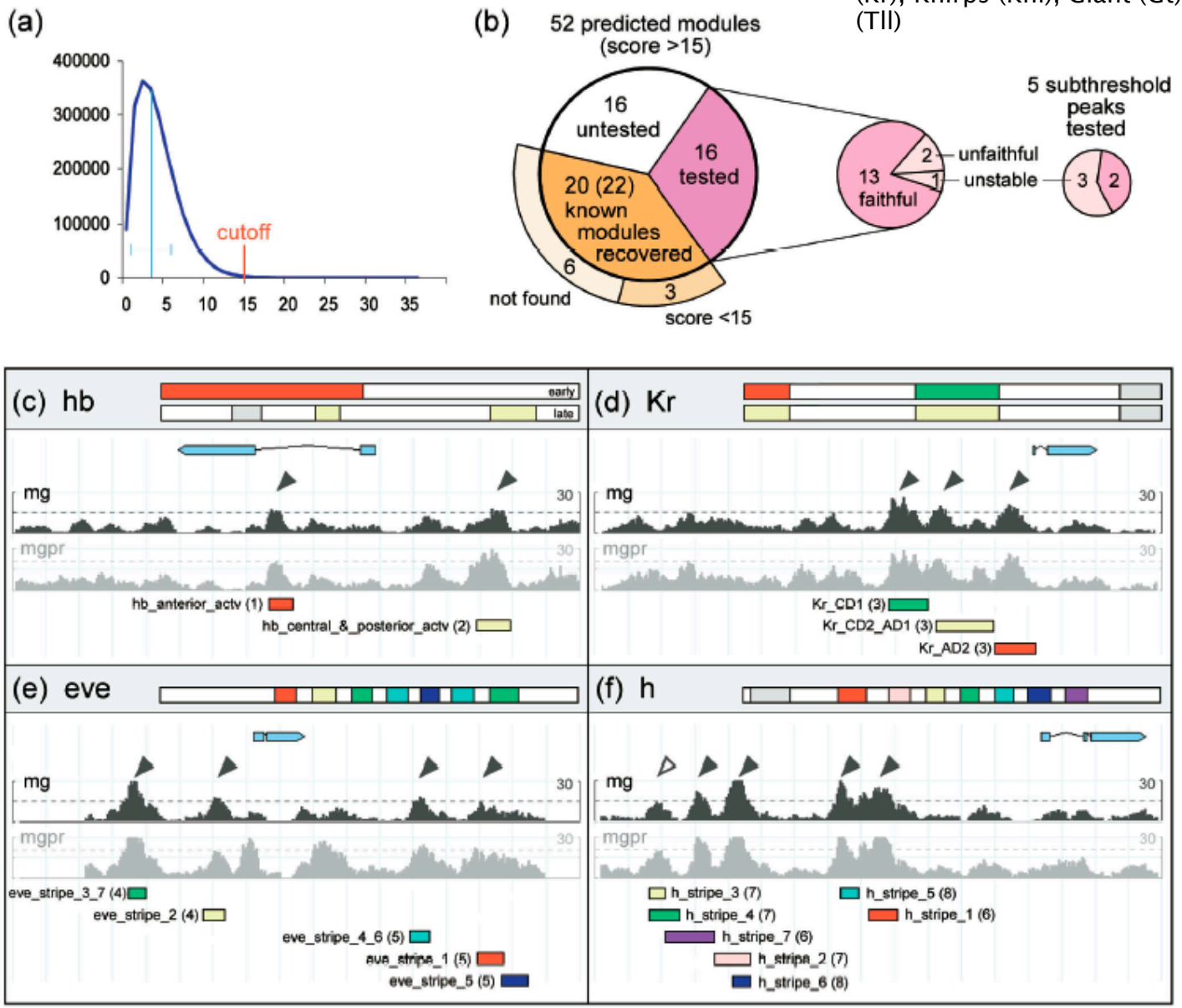
Cis-regulatory modules often involve clusters of binding sites for one or more transcription factors (e.g., drosophila EVE gene).



Clusters of 13 or more pattern matches in a window of 700 bp

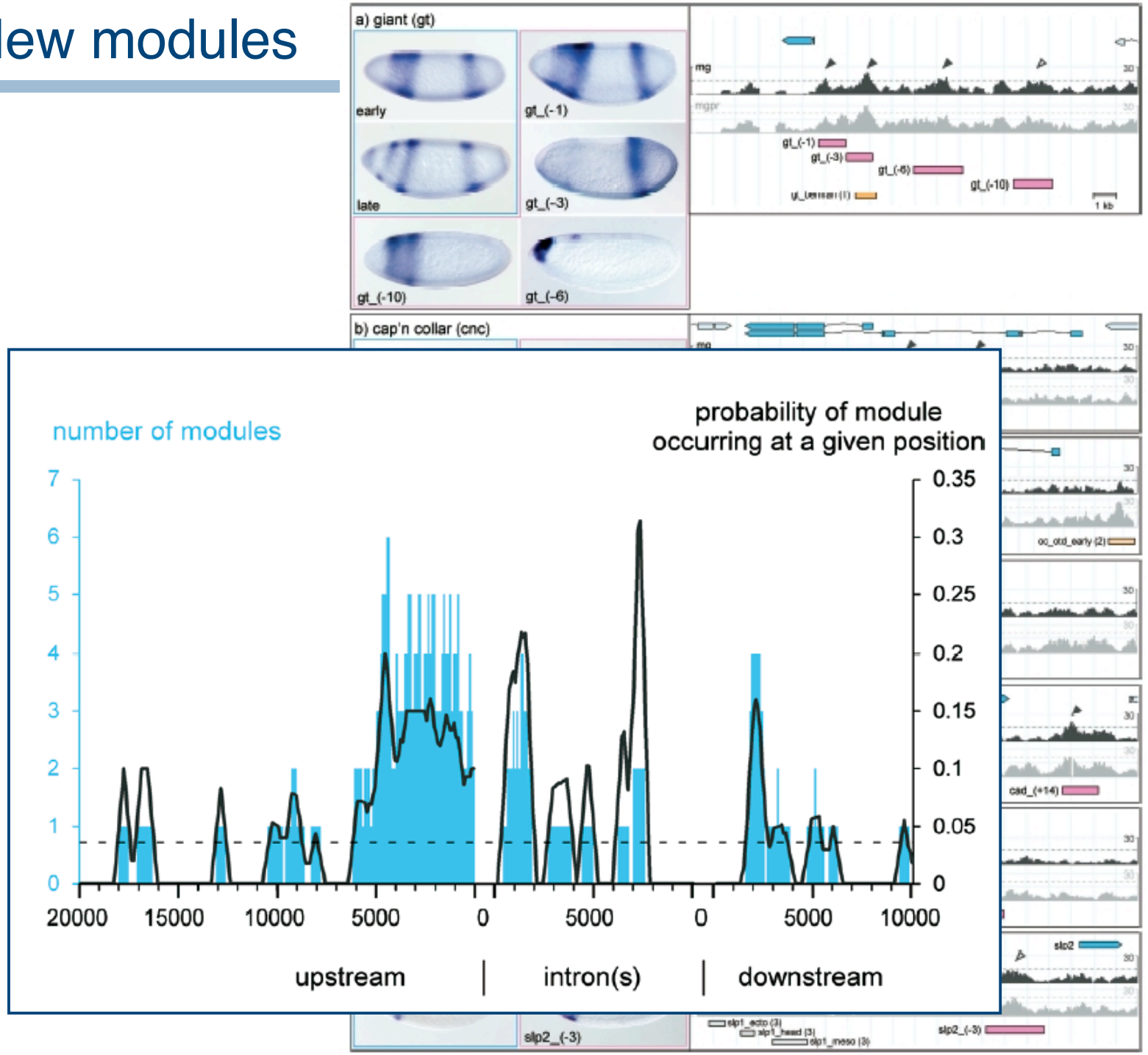
Recovery of known modules

Bicoid (Bcd), Hunchback (Hb), Caudal (Cad), the Torso-response element (TorRE), and Stat92E (D-Stat); Kruppel (Kr), Knirps (Kni), Giant (Gt), and Tailless (Tll)



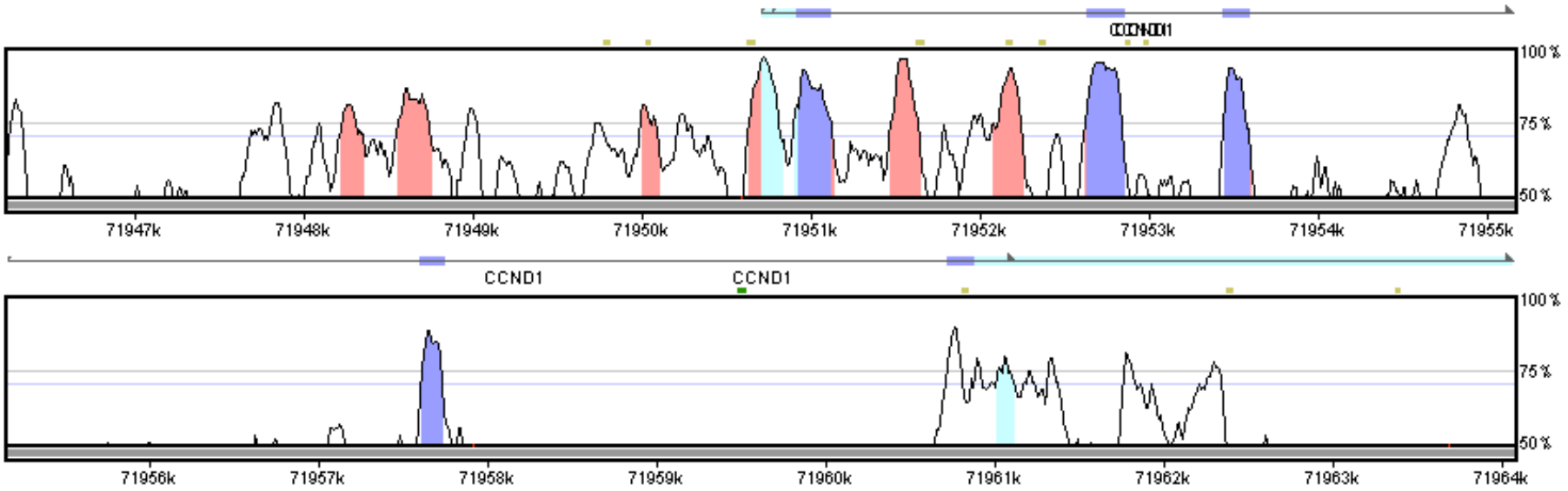
Schroeder et al. 2004, PLOS Biology

New modules

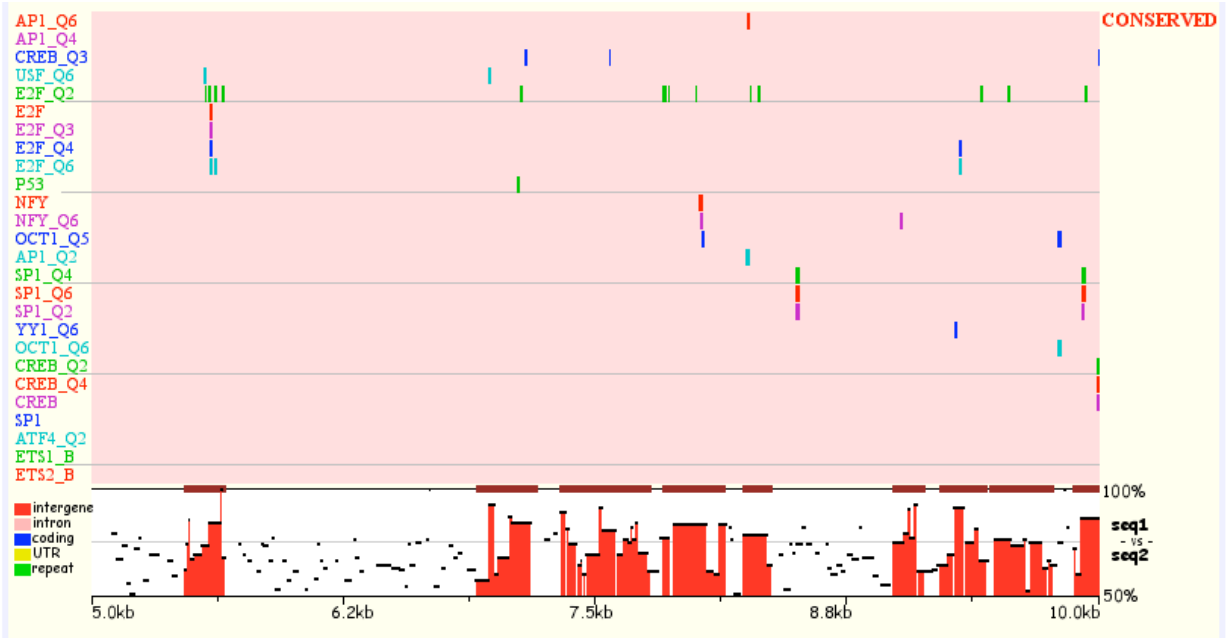

a) giant (gt)
early
gt_(-1)
late
gt_(-3)
gt_(-10)
gt_(-6)
b) cap'n collar (cnc)
slp2_(-3)
number of modules
probability of module
occurring at a given position
7
6
5
4
3
2
1
0
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
20000
15000
10000
5000
0
5000
0
5000
10000
upstream
intron(s)
downstream

Motif and module discovery in conserved sequences

Conserved non-coding sequences (CNS)



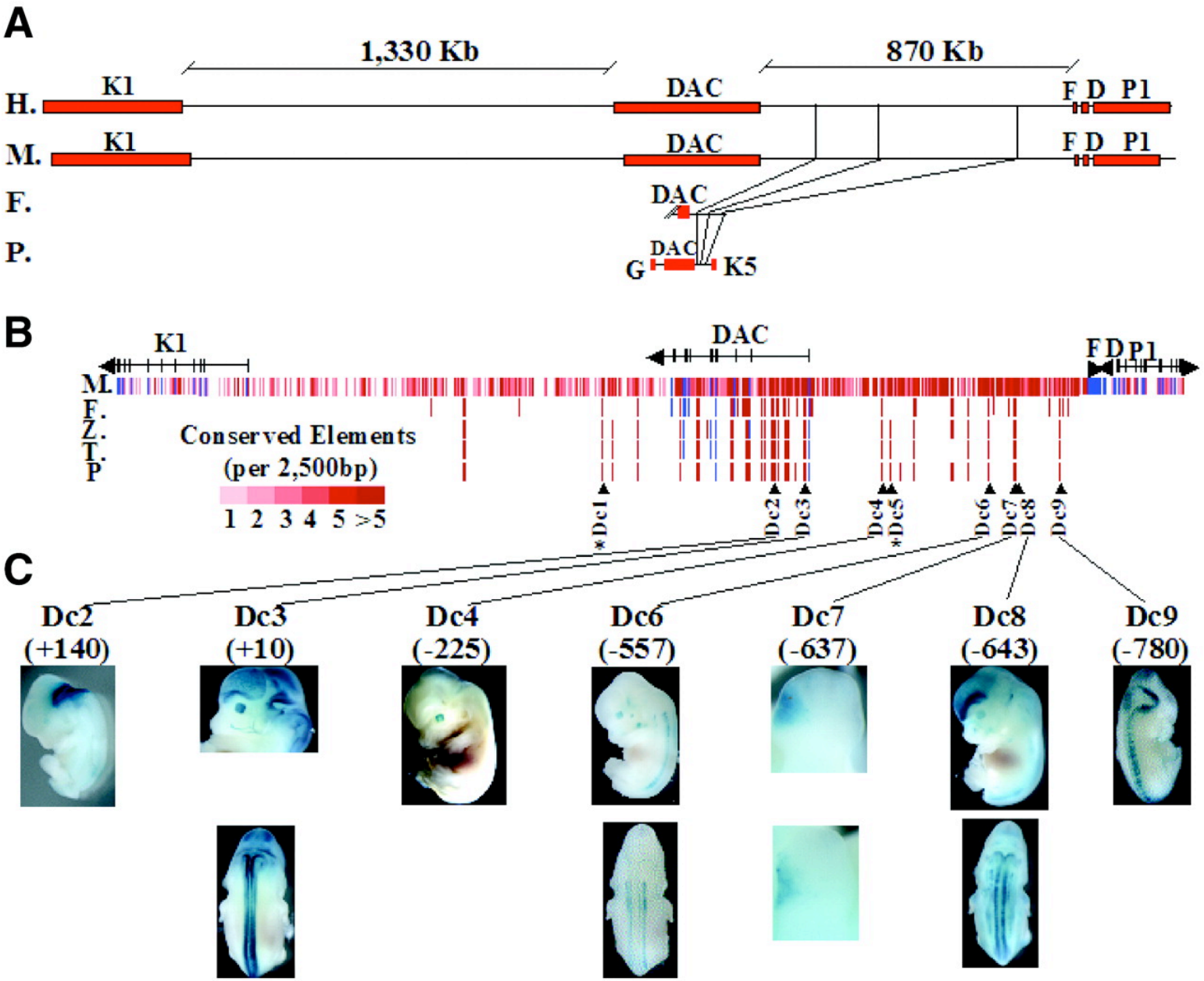
AVID, LAGAN, BLASTZ



TraFac, ConSite, rVISTA

Conserved non-coding sequences

Nobrega et al. (2003) *Science* 302 (5644)



Ultraconserved sequences

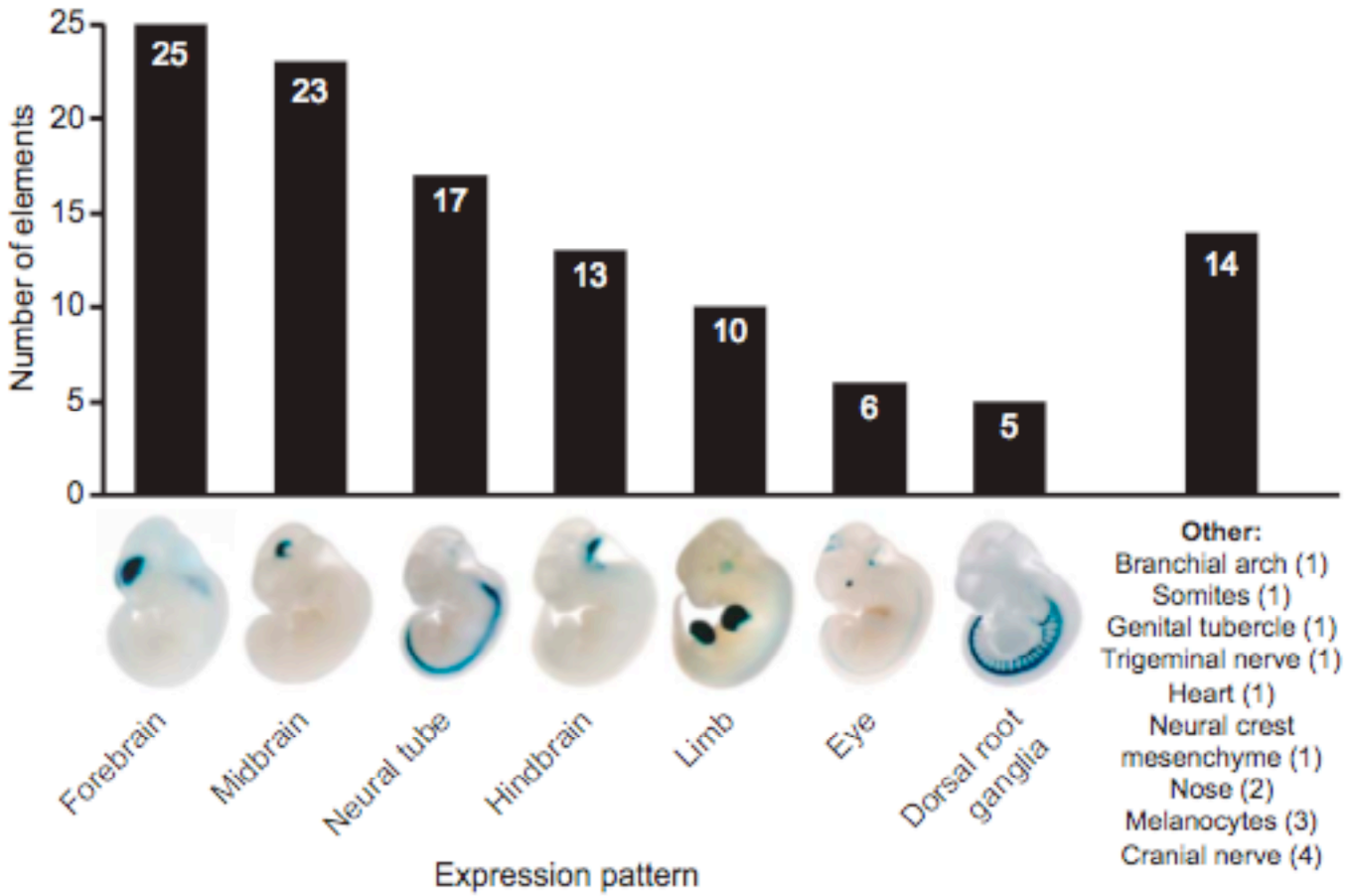
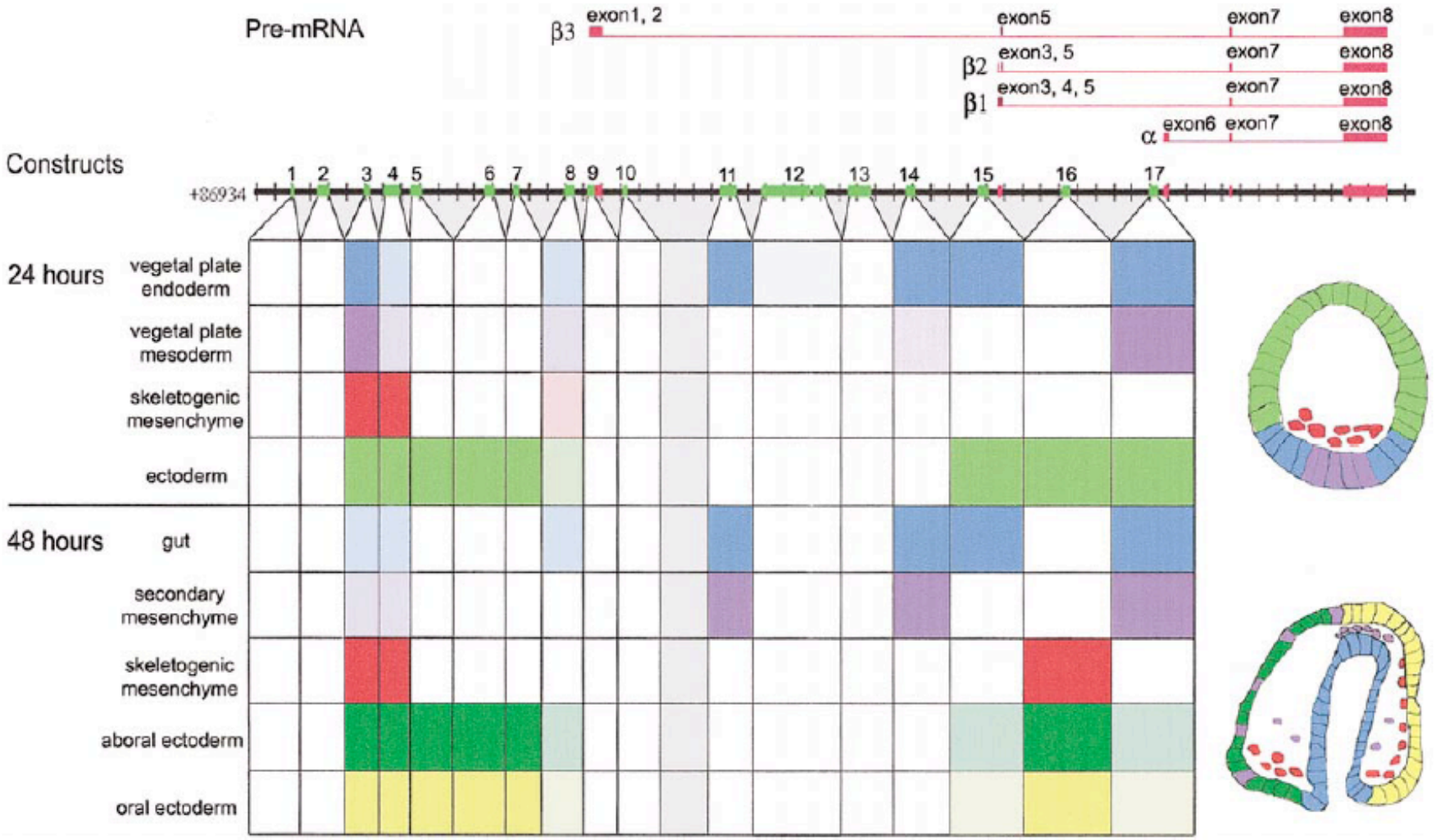


Figure 3 | Grouping of positive expression patterns captured in the transgenic mouse enhancer assay. The total number of elements displaying a given anatomical pattern is depicted by the height of the bars in the chart. A representative transgenic embryo is provided for each expression pattern. Elements with reproducible staining in more than one structure are included in each respective category.

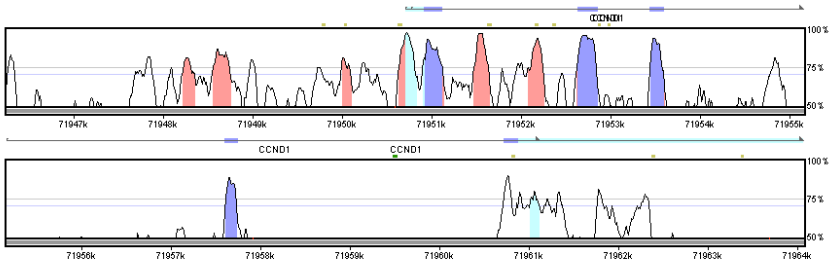
Conserved non-coding sequences

Strongylocentrotus purpuratus - Lytechinus variegatus: otx gene Yuh et al. (2002) *Dev Biol* 246:1

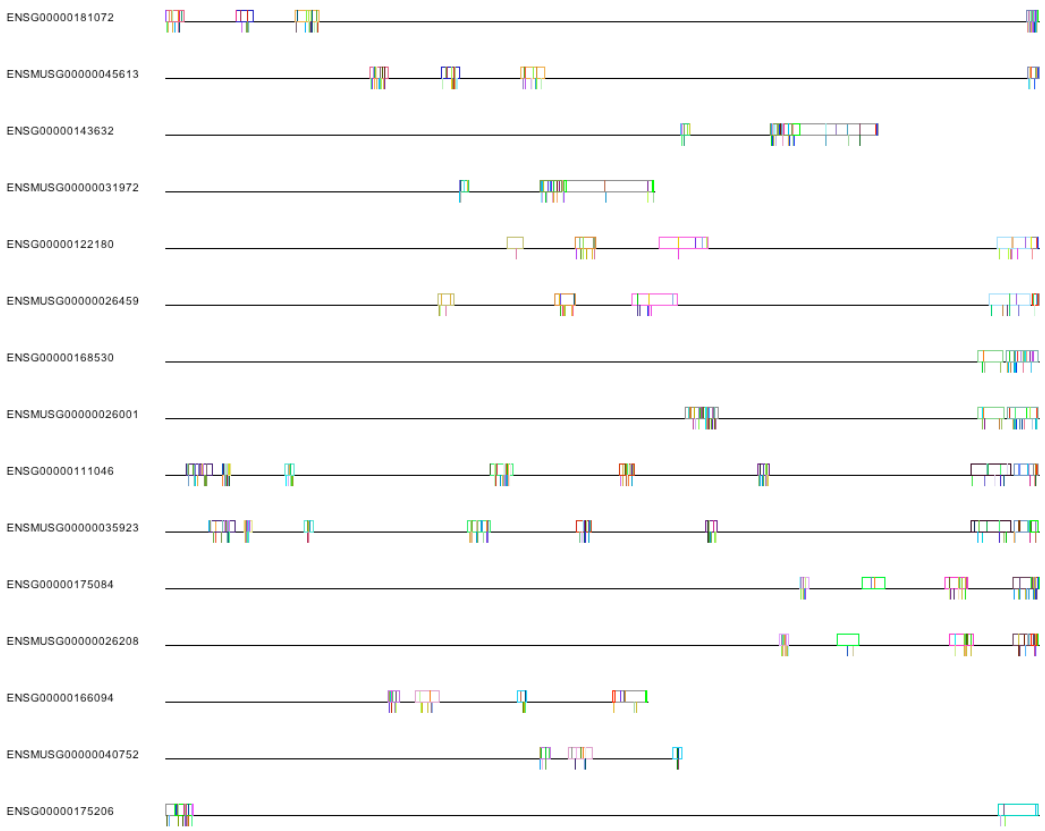


Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH. Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. Dev Biol. 2002 Jun 1;246(1):148-61.

Conserved non-coding sequences (CNS)



AVID, LAGAN, BLASTZ



Nobrega et al. (2003) *Science* 302 (5644)

Woolfe et al. (2005) *PLoS Biology* 3(1)

Yuh et al. (2002) *Dev Biol* 246:1

$$E(occ\{b\}) = F_e\{b\} \times 2 \times \sum_{i=1}^{S}(L_i - w + 1) = F_e\{b\} \times T$$

$$P(occ\{b\} = n) = \frac{T!}{(T-n)!\times n!} \times (F_e\{b\})^n \times (1 - F_e\{b\})^{(T-n)}$$

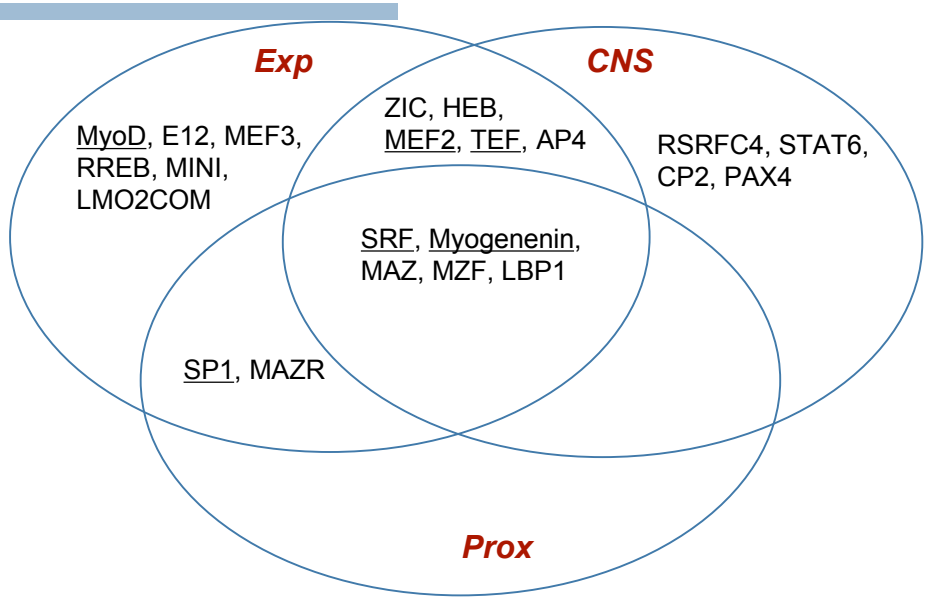
$$P(occ\{b\} \geq n) = \sum_{j=n}^{T} P(occ\{b\} = j)$$

Statistical over-representation

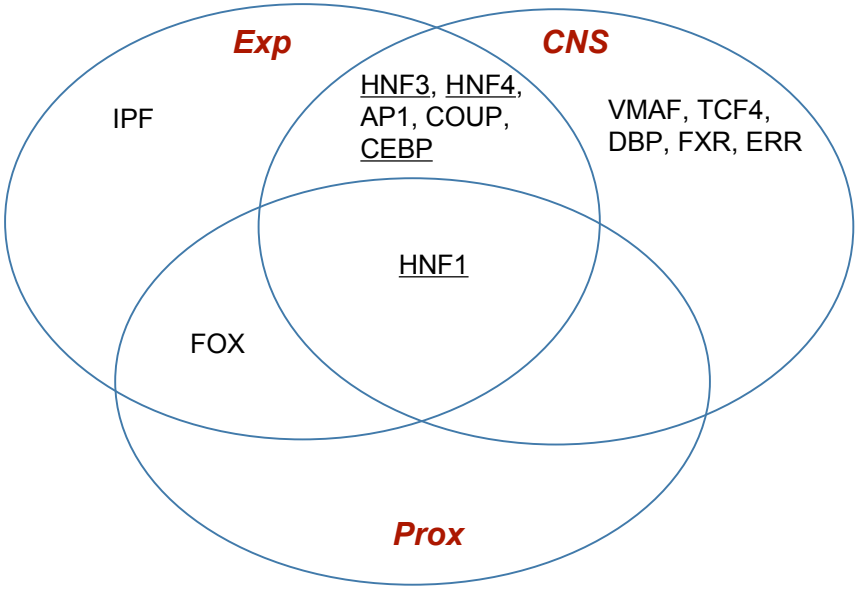
Aerts et al. (2003) *Nucleic Acids. Res.* 15;31(6):1753-64

Statistical over-representation

Wasserman & Fickett (1998)



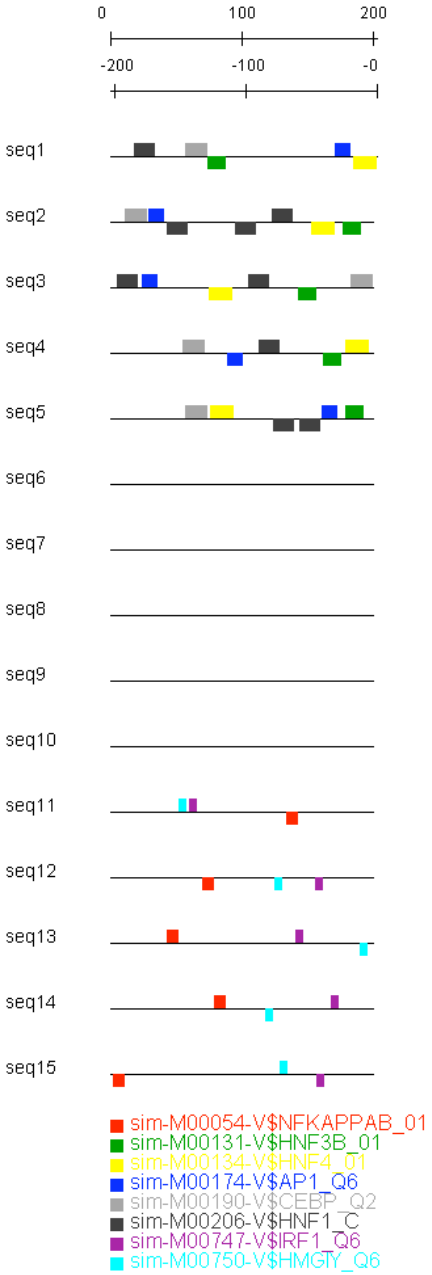
Krivan & Wasserman (2001)



Aerts et al. (2003) *Nucleic Acids. Res.* 15;31(6):1753-64

Module discovery

Module score function



$\mathcal{M}=(\Theta_1, \Theta_2,\ldots, \Theta_n)$ Module model

$$\text{PWM}=\Theta=\begin{bmatrix} 0{,}4 & 0{,}1 & \ldots & 0{,}2 \\ 0{,}2 & 0{,}1 & \cdots & 0{,}6 \\ 0{,}3 & 0{,}7 & \ldots & 0{,}1 \\ 0{,}1 & 0{,}1 & \cdots & 0{,}1 \end{bmatrix}$$

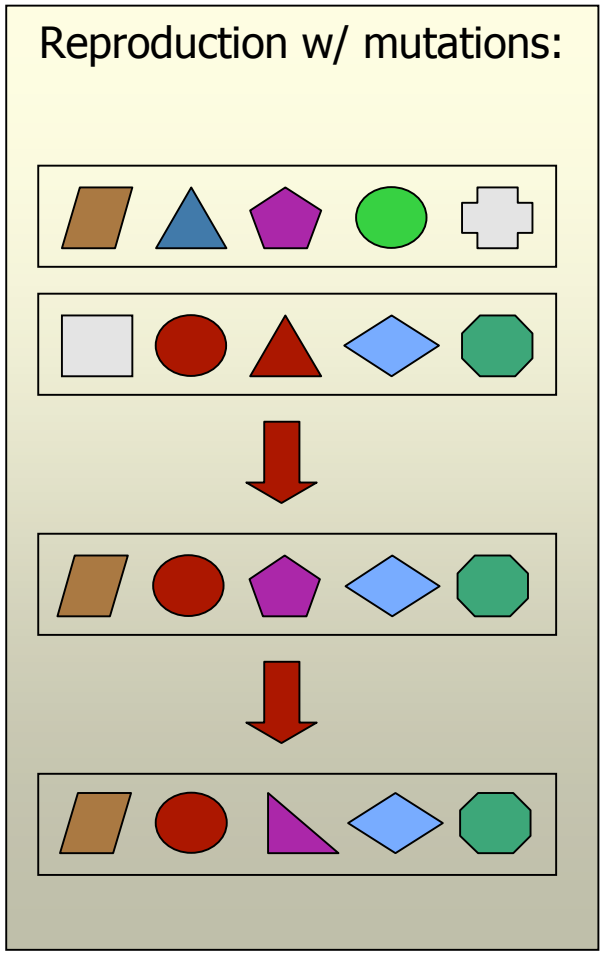
$$\mathcal{S}_{\mathcal{M}}(\mathbf{set}) = \sum_{i=1}^{n} \mathcal{S}_{\mathcal{M}}(seq_i)$$

$$\mathcal{S}_{\mathcal{M}}(seq) = \max_{T} p(\mathbf{m}) \times \sum_{\mathbf{x}\in\mathbf{m}} W(\mathbf{x})$$

T = all valid modules m's
- within window of length maxLength
- overlapAllowedYesNo

n=5 ⇒ $400^5=10^{13}$ combinations ⇒ Efficient search algorithm

Genetic algorithm



Population: generate *p* random modules

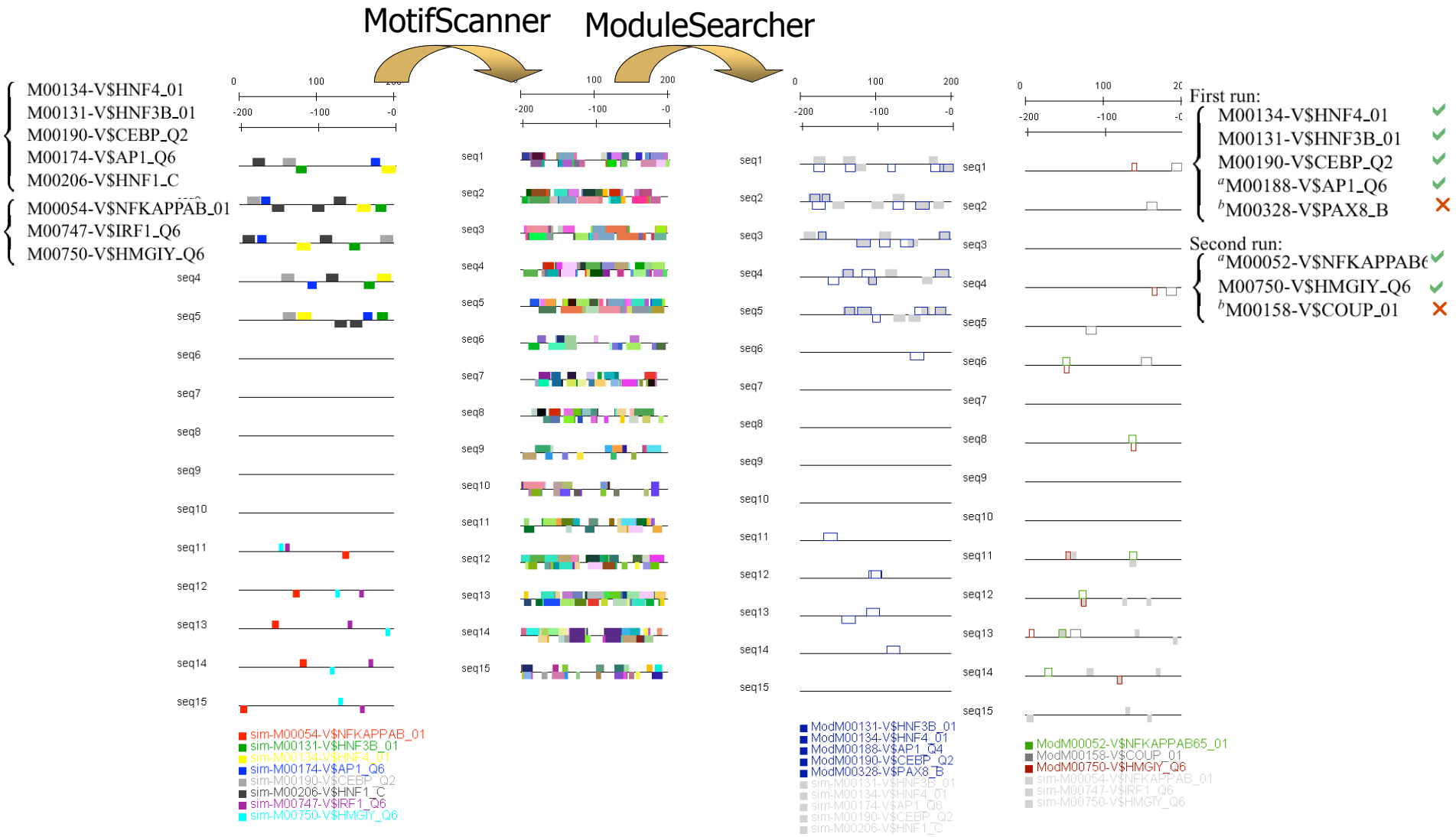
Evaluation: fitness value for each module
s best scoring modules are survivors

g generations

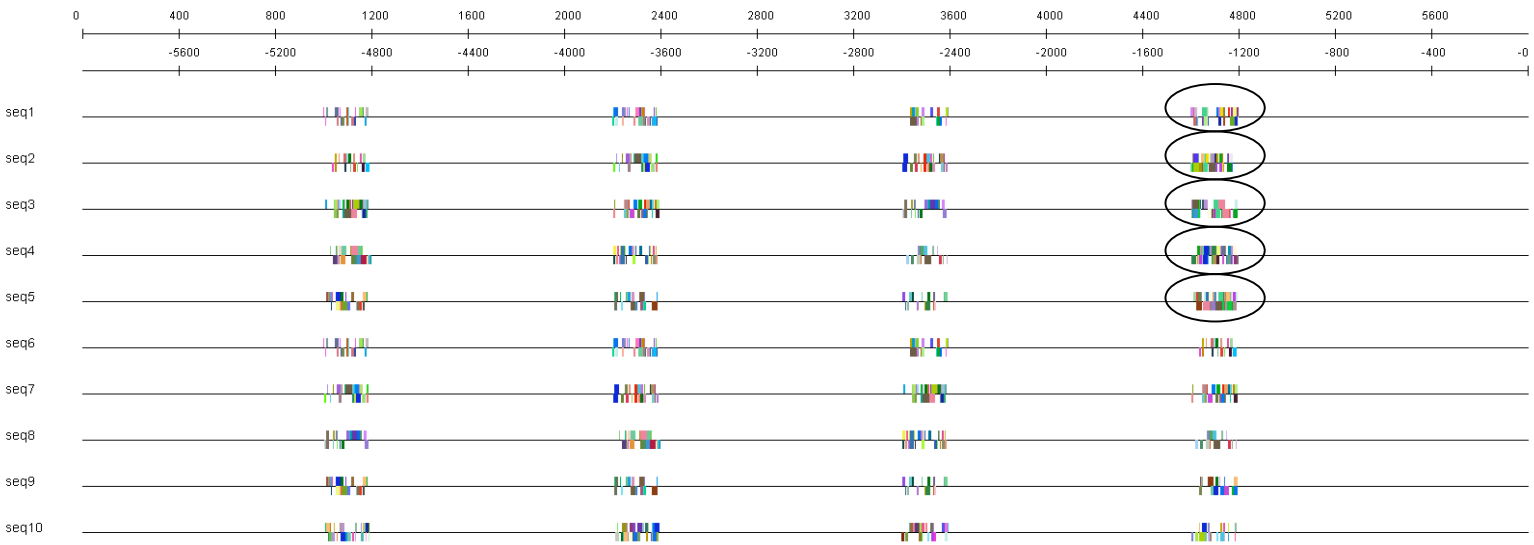
Reproduction:
survivors are paired to create *p* new modules by crossover

Mutation: replace a factor with another factor with mutation probability ρ

Test on simulated data

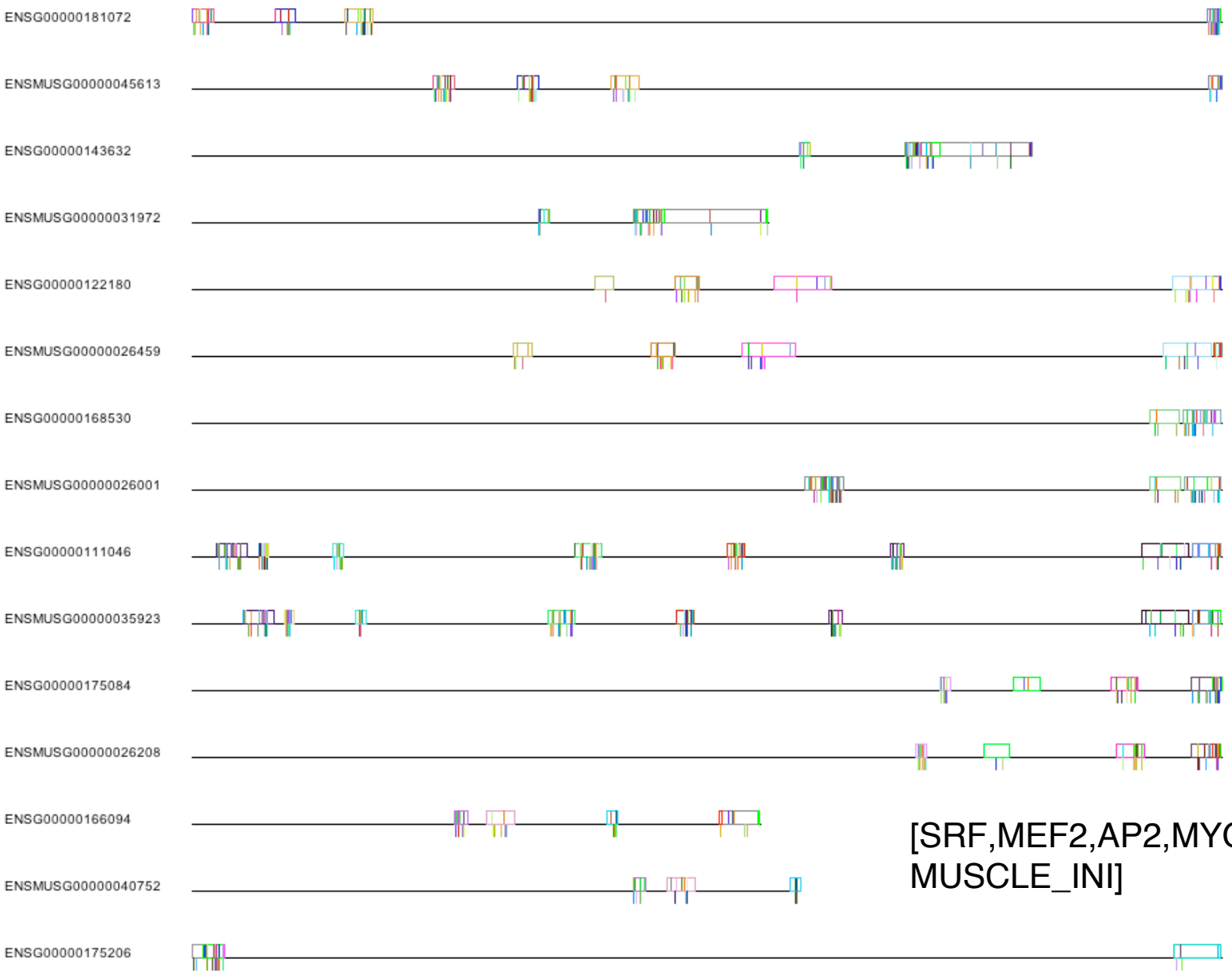


Test on simulated data



- M00134-V$HNF4_01
- M00131-V$HNF3B_01
- M00190-V$CEBP_Q2
- M00188-V$AP1_Q6
- M00206-V$HNF1_Cc

ModuleSearcher + phylogenetic footprinting



[SRF,MEF2,AP2,MYOGENIN, MUSCLE_INI]

Exercise / competition motif and CRM discovery

- Simulated set
  - 5 PWMs
  - 10 seq with 1 instance of each PWM
  - 10 seq generated from 0th order background model
  - What and where are the motifs, and which are the 10 sequences?
    - Upload GFF to **http://med.kuleuven.be/cme-mg/lng/student/** !!
    - Put method, parameters, explanation in "comment" field
    - Scored by calculating the nucleotide-level correlation with the answer vectors

End