

Lecture 5.0: Gene Regulation Bioinformatics

Wyeth W. Wasserman

University of British Columbia

www.cisreg.ca

Lecture 5.0: Overview

Part 1: Overview of transcription

Part 2: Prediction of transcription factor binding sites using binding profiles (“Discrimination”)

Part 3: Interrogation of sets of co-expressed genes to identify mediating transcription factors

Part 4: Detection of novel motifs (TFBS) over-represented in regulatory regions of co-expressed genes (“Discovery”)

Restrictions in Coverage

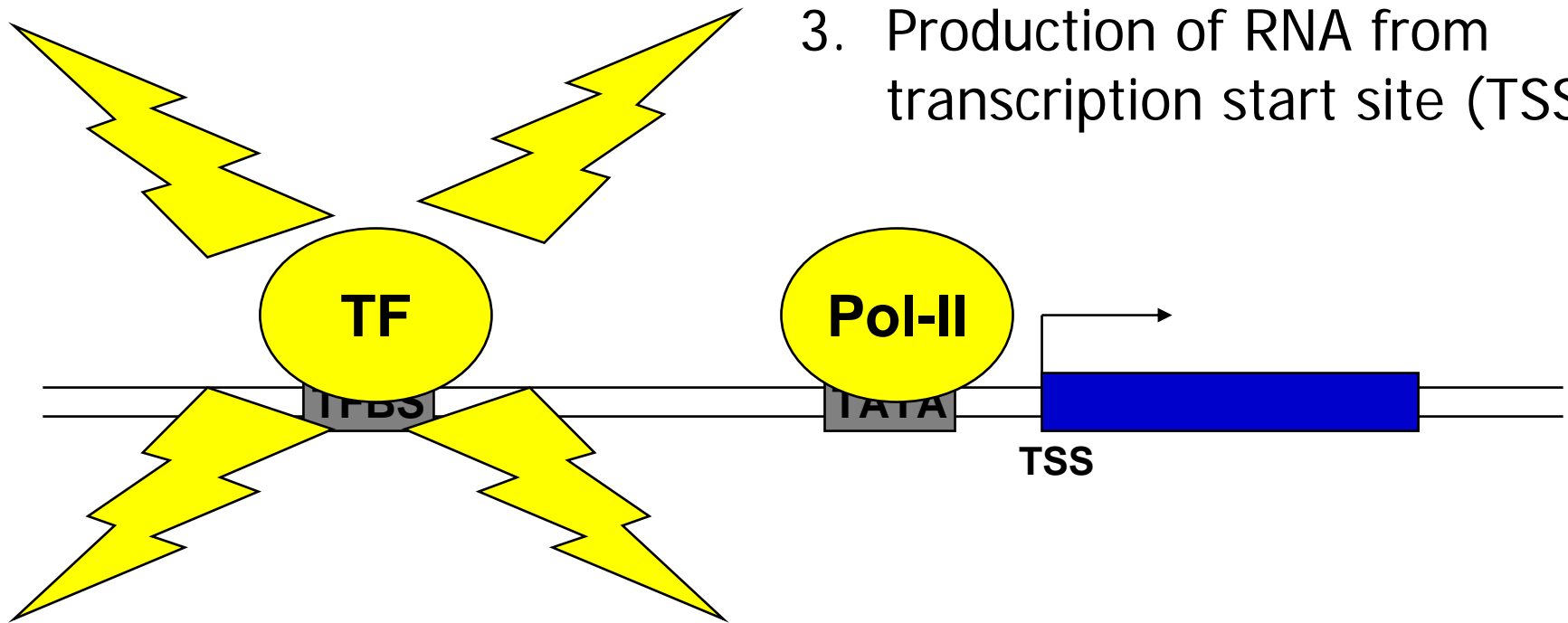
- Focus on Eukaryotic cells
 - Most principles apply to prokaryotes
- Polymerase II driven promoters
 - Generally protein coding genes
- All references are made to activating sequences
 - Information about repression is sparse

Part 1: Introduction to transcription in eukaryotic cells

Transcription Over-Simplified

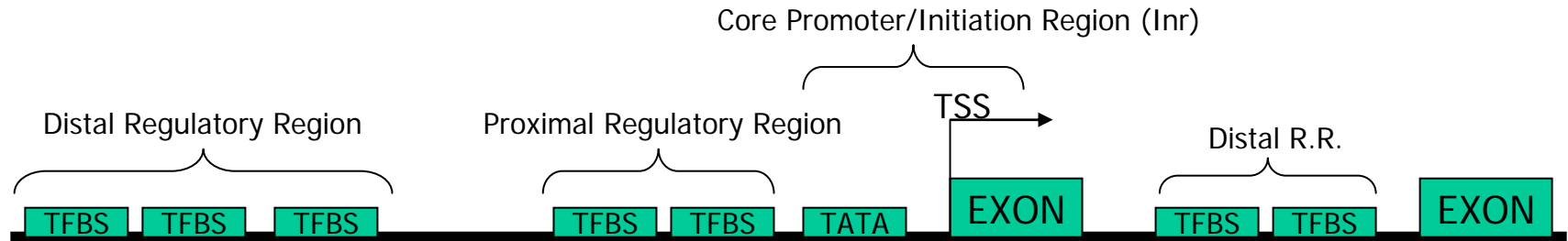
Three-step Process:

1. TF binds to TFBS (DNA)
2. TF catalyzes recruitment of polymerase II complex
3. Production of RNA from transcription start site (TSS)



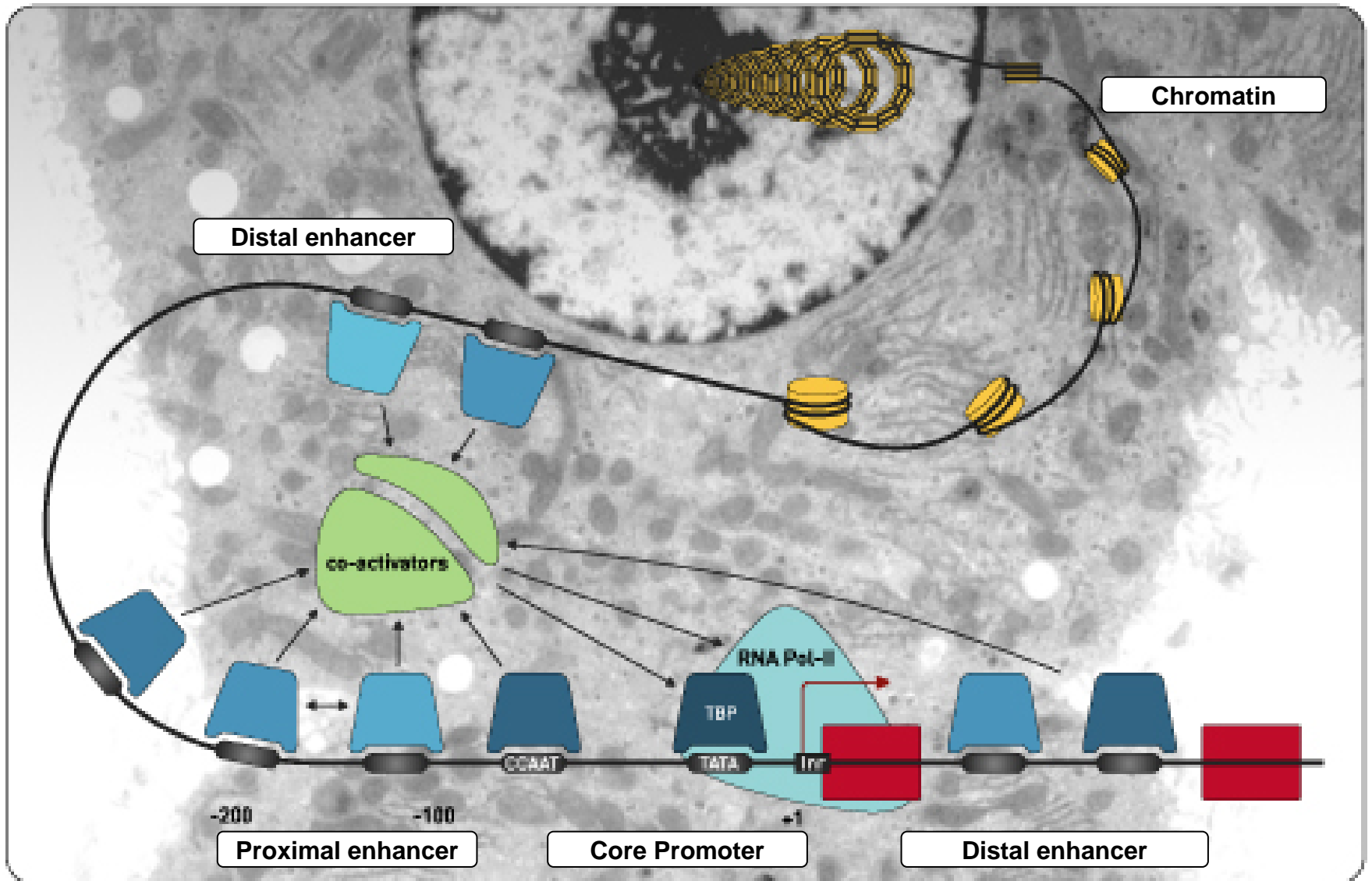
Anatomy of Transcriptional Regulation

WARNING: Terms vary widely in meaning between scientists

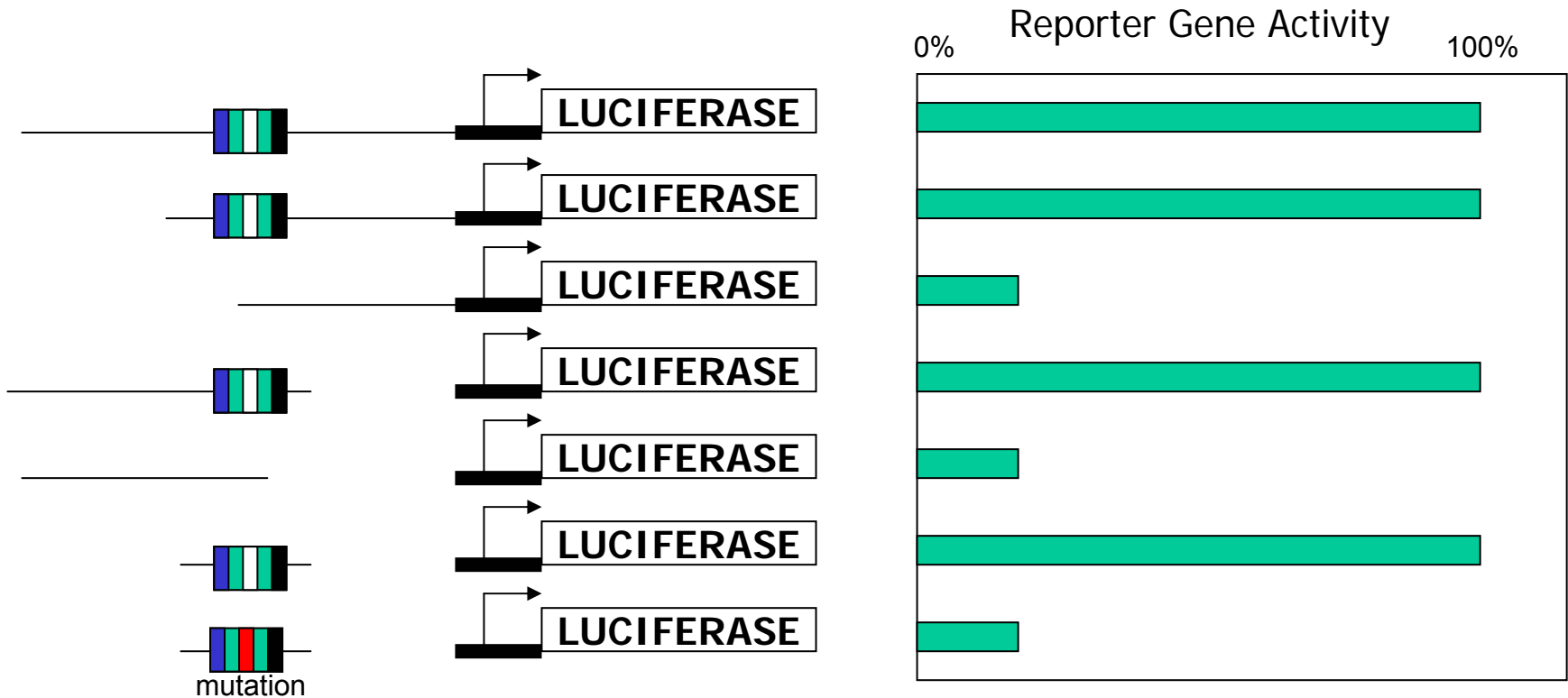


- Core Promoter – Sufficient to support the initiation of transcription; orientation dependent
 - TSS – transcription start site
 - Often a region rather than specific position
- TFBS – single transcription factor binding site
- Regulatory Regions
 - Proximal/Distal – vague reference to distance from TSS
 - May be positive (enhancing) or negative (repressing)
 - Orientation independent (generally)
 - Modules – Sets of TFBS within a region that function together

Complexity in Transcription



Lab Discovery of TF Binding Sites



Identify functional regulatory region within a sequence and delineate specific TFBS through mutagenesis (and in vitro binding studies)

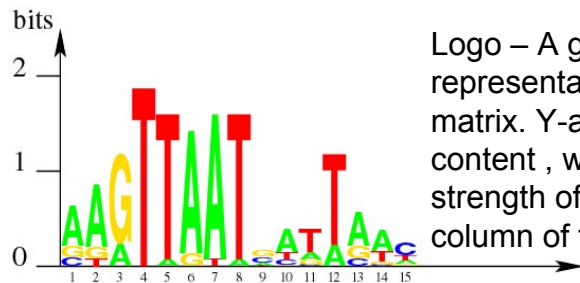
Part 2: Prediction of TF Binding Sites, Core Promoters and Regulatory Regions (Discrimination)

Teaching a computer to find TFBS...

Representing Binding Sites for a TF

- A single site
 - AAGTTAATGA
- A set of sites represented as a consensus
 - VDRTWRWSHD (IUPAC degenerate DNA)
- A matrix describing a set of sites:

A	14	16	4	0	1	19	20	1	4	13	4	4	13	12	3
C	3	0	0	0	0	0	0	0	7	3	1	0	3	1	12
G	4	3	17	0	0	2	0	0	9	1	3	0	5	2	2
T	0	2	0	21	20	0	1	20	1	4	13	17	0	6	4



Logo – A graphical representation of frequency matrix. Y-axis is information content, which reflects the strength of the pattern in each column of the matrix

Set of binding sites

AAGTTAATGA
 CAGTTAATAA
 GAGTTAAACA
 CAGTTAATTA
 GAGTTAATAA
 CAGTTATTCA
 GAGTTAATAA
 CAGTTAATCA
 AGATTAAAGA
 AAGTTAACGA
 AGGTTAACGA
 ATGTTGATGA
 AAGTTAATGA
 AAGTTAACGA
 AAATTAATGA
 GAGTTAATGA
 AAGTTAATCA
 AAGTTGATGA
 AAATTAATGA
 ATGTTAATGA
 AAGTAAATGA
 AAGTTAATGA
 AAGTTAATGA
 AAATTAATGA
 AAGTTAATGA
 AAGTTAATGA
 AAGTTAATGA
 AAGTTAATGA
 AAGTTAATGA

Conversion of PFMs to Position Specific Scoring Matrices (PSSM)

Add the following features to the matrix profile:

1. Correct for nucleotide frequencies in genome
2. Weight for the confidence (depth) in the pattern
3. Convert to log-scale probability for easy arithmetic

<i>pfm</i>							<i>pssm</i>					
A	5	0	1	0	0	$\text{Log} \left(\frac{f(b,i) + s(n)}{p(b)} \right)$ \longrightarrow	A	1.6	-1.7	-0.2	-1.7	-1.7
C	0	2	2	4	0		C	-1.7	0.5	0.5	1.3	-1.7
G	0	3	1	0	4		G	-1.7	1.0	-0.2	-1.7	1.3
T	0	0	1	1	1		T	-1.7	-1.7	-0.2	-0.2	-0.2

TGCTG = 0.9



JASPAR: AN OPEN-ACCESS DATABASE OF TF BINDING PROFILES

(Transfac database is a commercial alternative)

The Good...

- Tronche (1997) tested 50 predicted HNF1 TFBS using an in vitro binding test and found that 96% of the predicted sites were bound!
- Hoffman and Fields (1998) found in detailed biochemical studies that the best weight matrices produce scores highly correlated with in vitro binding energy

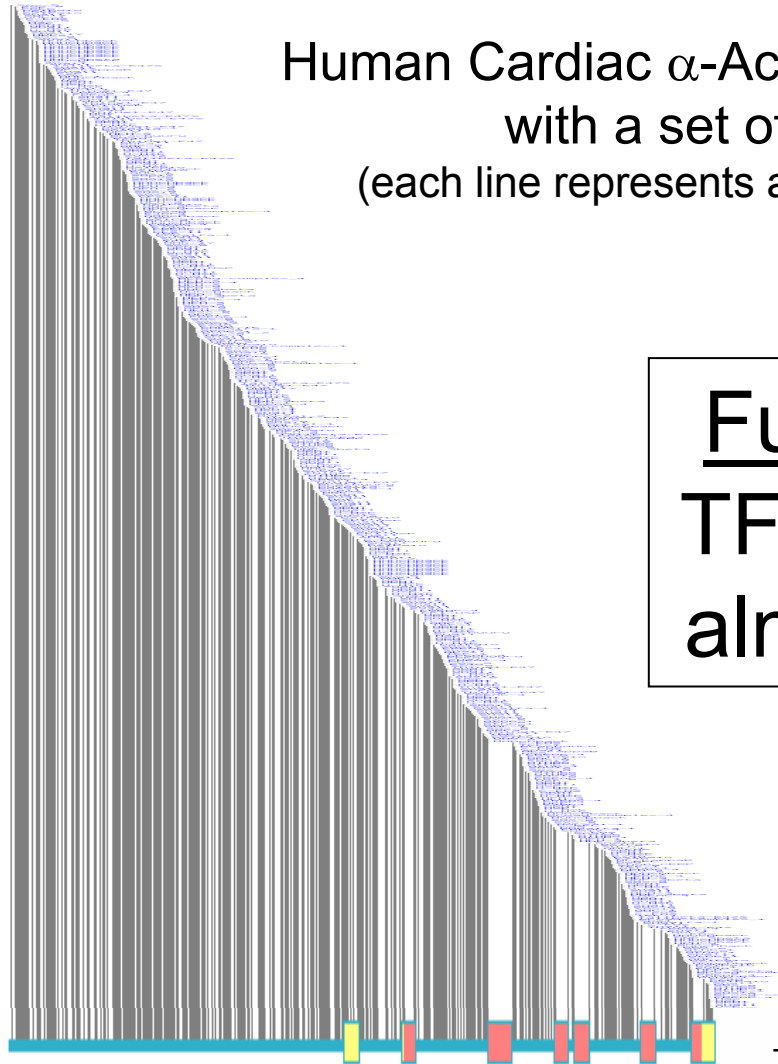
...the Bad...

- Fickett (1995) found that a profile for the myoD TF made predictions at a rate of 1 per ~500bp of human DNA sequence
 - This corresponds to an average of 20 sites / gene (assuming 10,000 bp as average gene size)

...and the Ugly!

Human Cardiac α -Actin gene analyzed
with a set of profiles
(each line represents a TFBS prediction)

Futility Conjecture:
TFBS predictions are
almost always wrong

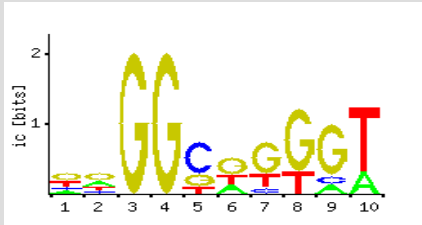


Red boxes are protein coding exons -
TFBS predictions excluded in this analysis

Detecting binding sites in a single sequence

Scanning a sequence against a PWM

Sp1



ACCCTCCCCAGGGGCGGGGGGCGGTGGCCAGGACGGTAGCTCC

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

Abs_score = 13.4 (sum of column scores)

Calculating the relative score

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

Max_score = 15.2 (sum of highest column scores)

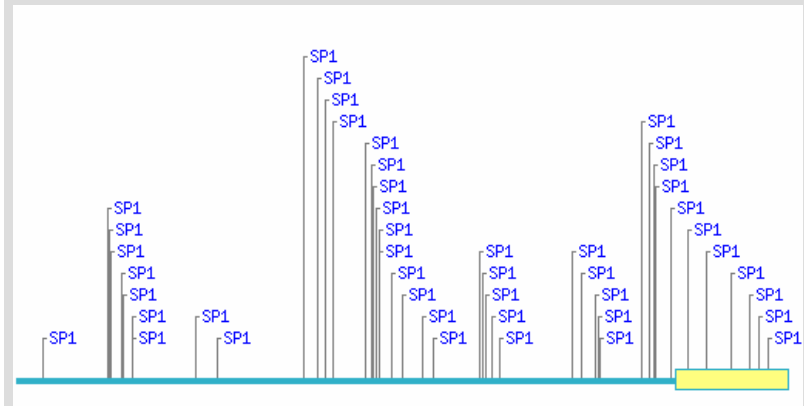
A	[-0.2284	0.4368	-1.5	-1.5	-1.5	0.4368	-1.5	-1.5	-0.2284	0.4368]
C	[-0.2284	-0.2284	-1.5	-1.5	1.5128	-1.5	-0.2284	-1.5	-0.2284	-1.5]
G	[1.2348	1.2348	2.1222	2.1222	0.4368	1.2348	1.5128	1.7457	1.7457	-1.5]
T	[0.4368	-0.2284	-1.5	-1.5	-0.2284	0.4368	0.4368	0.4368	-1.5	1.7457]

Min_score = -10.3 (sum of lowest column scores)

$$\text{Rel_score} = \frac{\text{Abs_score} - \text{Min_score}}{\text{Max_score} - \text{Min_score}} \cdot 100\%$$

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

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



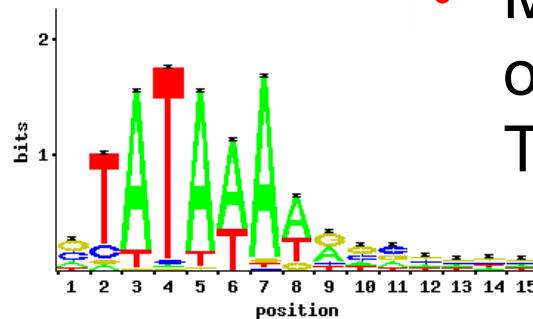
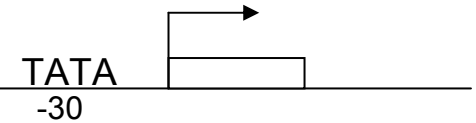
Ouch.

Observations

- PSSMs accurately reflect in vitro binding properties of DNA binding proteins
- Suitable binding sites occur at a rate far too frequent to reflect in vivo function
- Bioinformatics methods that use PSSMs for binding site studies must incorporate additional information to enhance specificity

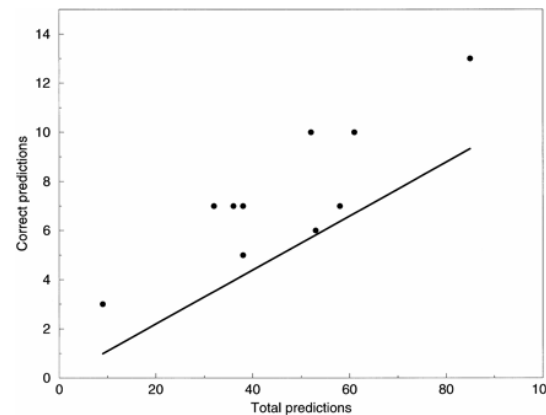
Core Promoter Prediction

- Amongst oldest topics in bioinformatics is core promoter detection



- Many methods based on PSSM detection of TATA motif
- Only ~60% of promoters have TATA motif

- Fickett & Hatzigeorgiou (1997) found that existing methods did as well as TATA box detection alone and *most* were slightly better than random guessing

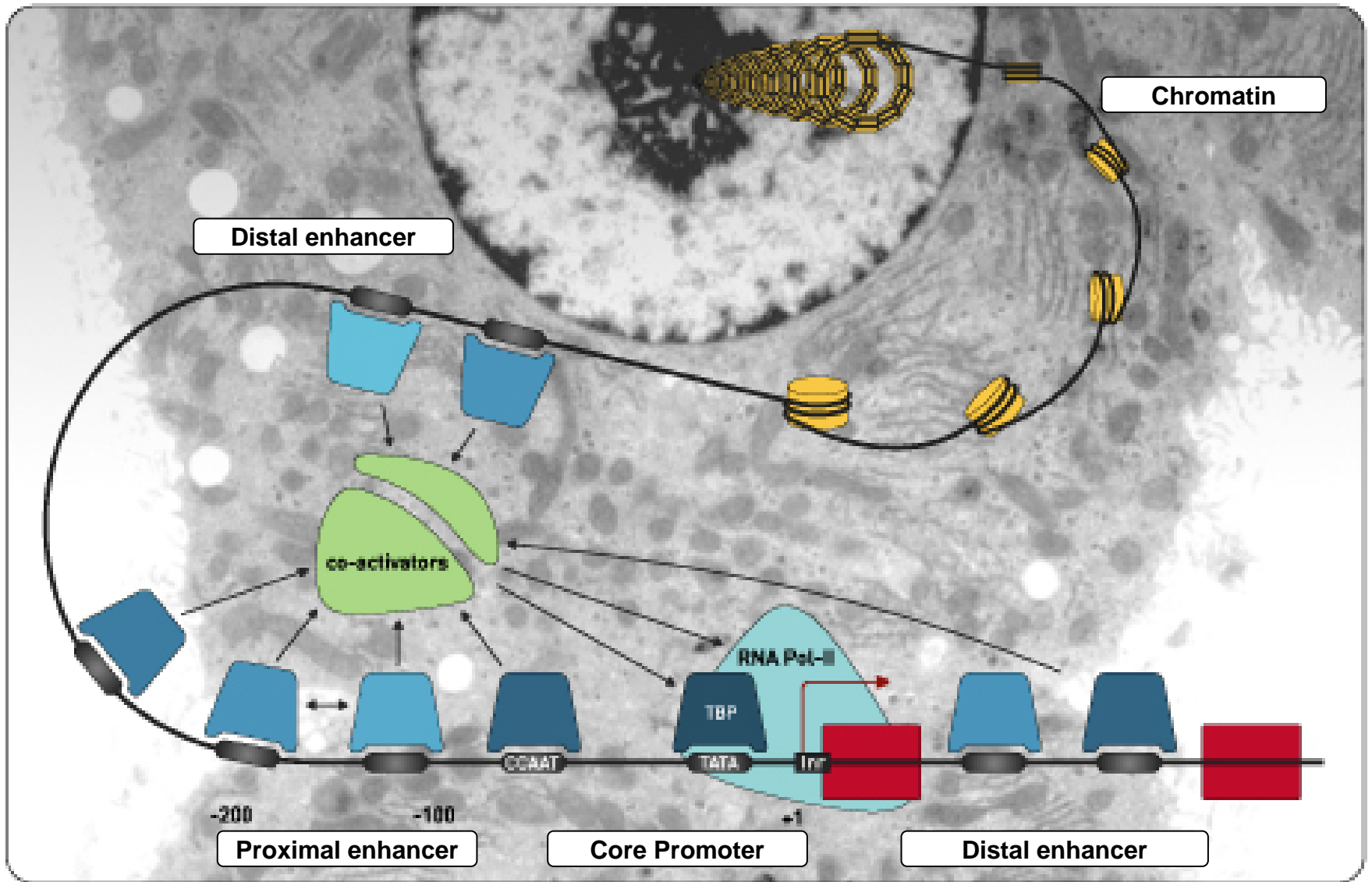


Line indicates random guessing

Changing the Question for Promoter Identification

- Recommendation from Fickett & Hatzigeorgiou to do two things to overcome the specificity problem for identification of promoters:
 - First, develop methods to predict regions containing promoters rather than predict specific transcription start sites
 - Second, find additional sources of information beyond TATA motif

Recall

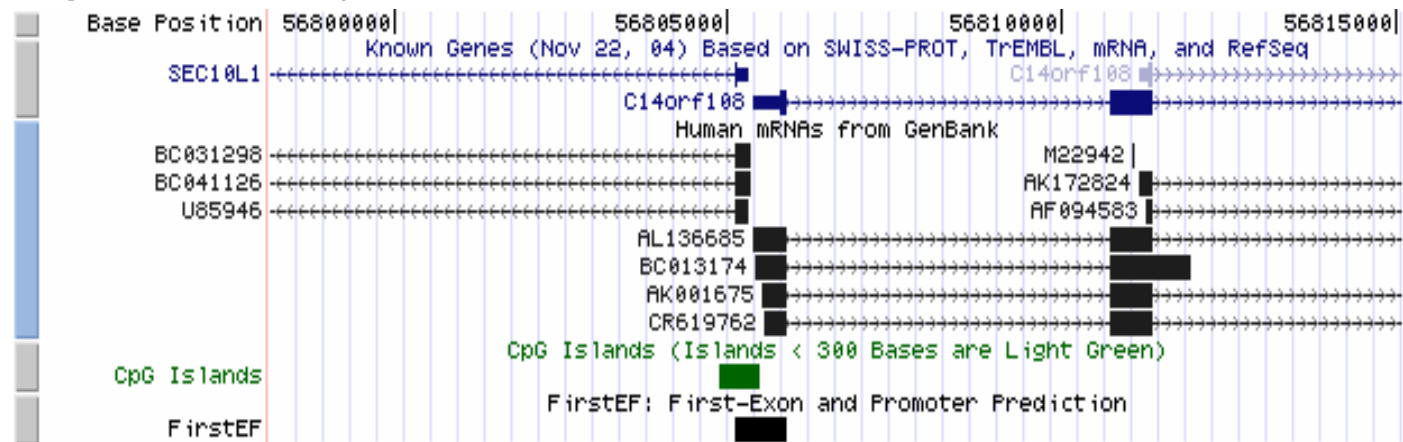


CpG Islands

- DNA methylation occurs in competition with histone acetylation
 - Acetylation promotes open chromatin structure that is permissive for TF binding to DNA
 - Methylation of DNA inhibits histone acetylation
 - Certain TFs promote histone acetylation by recruiting acetylases
- Methylation occurs on cytosines
 - Preferentially on cytosine adjacent to guanines (CG dinucleotides, generally referred to as CpG)
 - Methylated cytosines frequently undergo deamination to form thymidine (CpG → TpG)
- CpG Islands are regions of DNA where CG dinucleotides occur at a frequency consistent with C and G mononucleotide frequencies
 - Highlight of regions in which histones are acetylated – regions of active transcription

New Promoter Detection Programs

- Several second generation promoter detection methods (e.g. EpoNine) identify regions likely to contain transcription start sites based on nucleotide composition
 - Hannenhalli and Levy (2002) determined that the ratio $[CpG]/[C][G]$ is equally informative
- FirstEF combines composition analysis, TATA motifs and transcript data (cDNAs and ESTs) to predict regions likely to contain a TSS



chr14:56,798,150-56,815,078

Human May 2004

Notice bidirectional transcripts

Bidirectional Promoters (Aside)

- CpG islands reflect open chromatin
- Transcription initiation appears to occur more readily in such regions
- CpG islands are highly associated with transcript initiation in **BOTH** directions
 - Unclear if one direction is spurious or produces a functionally important transcript
 - » Probably a mixture of both

Promoter Recognition Summary

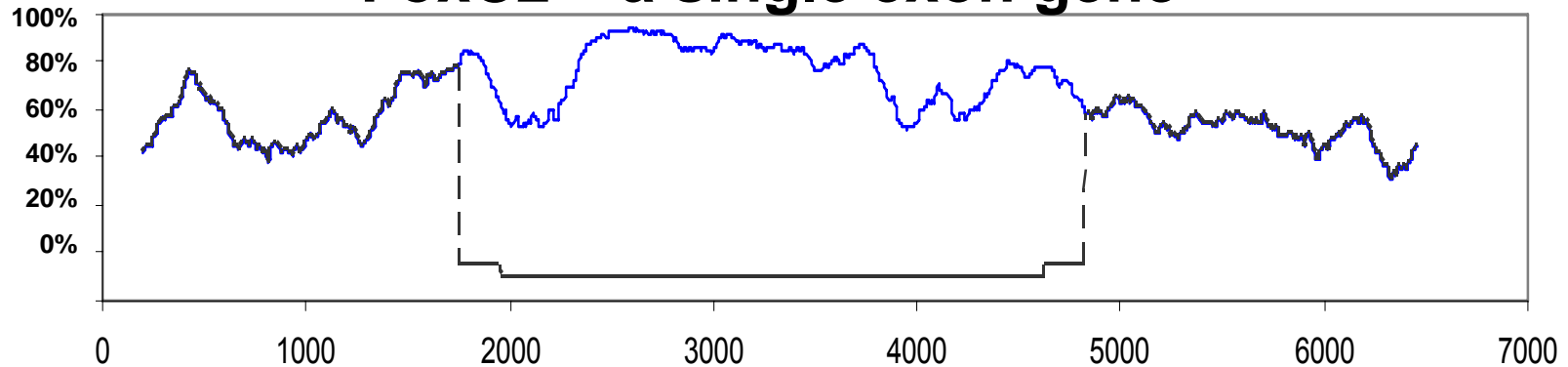
- TATA motif recognition is insufficient to specifically identify regions containing a transcription start site
- CpG island detection complements TATA motif detection in FirstEF
 - Biology insight dramatically improves pattern recognition
- Integration of independent information or properties can overcome specificity problems

Using Phylogenetic Footprinting to Improve TFBS Discrimination

70,000,000 years of evolution
can reveal regulatory regions

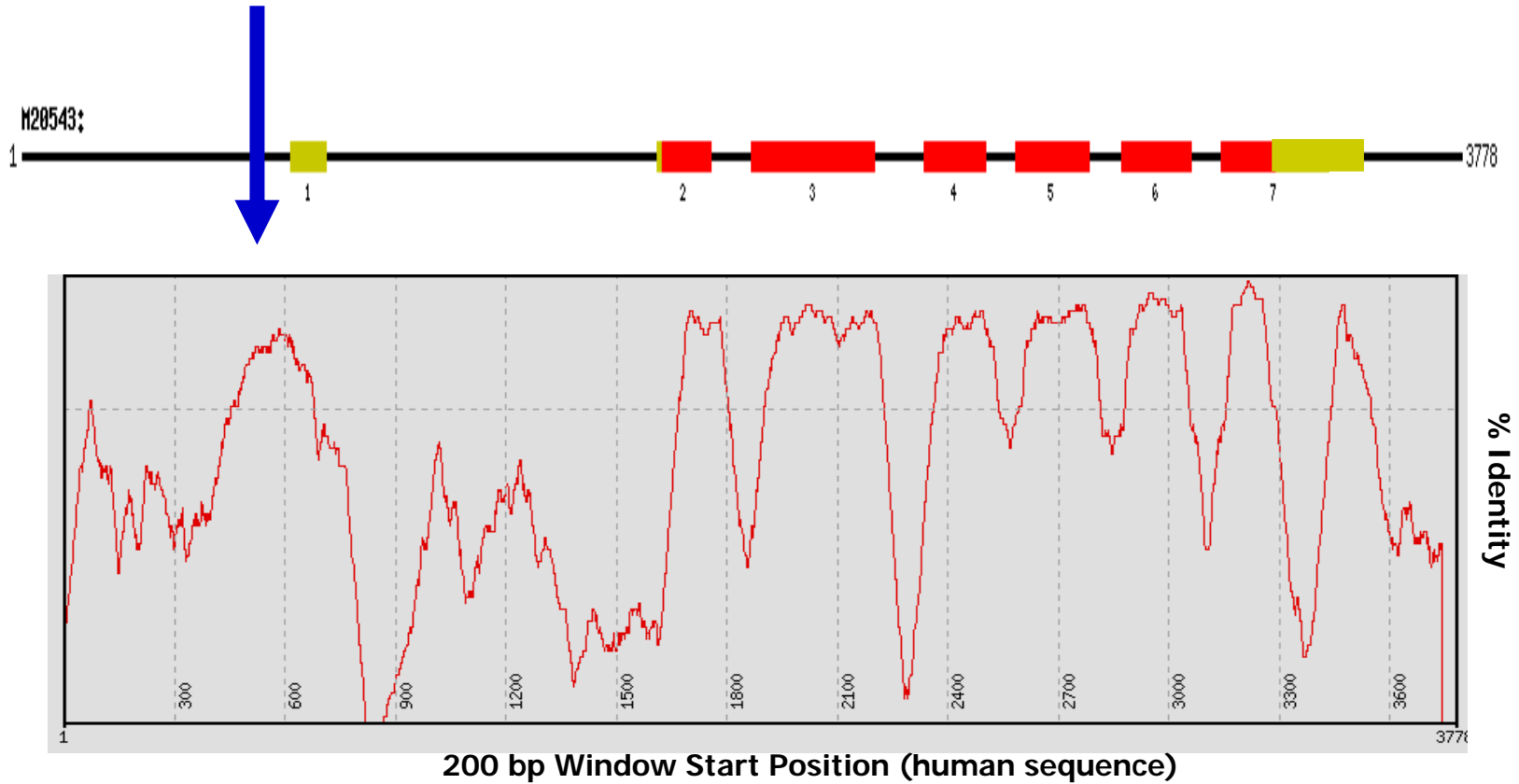
Phylogenetic Footprinting

FoxC2 – a single exon gene



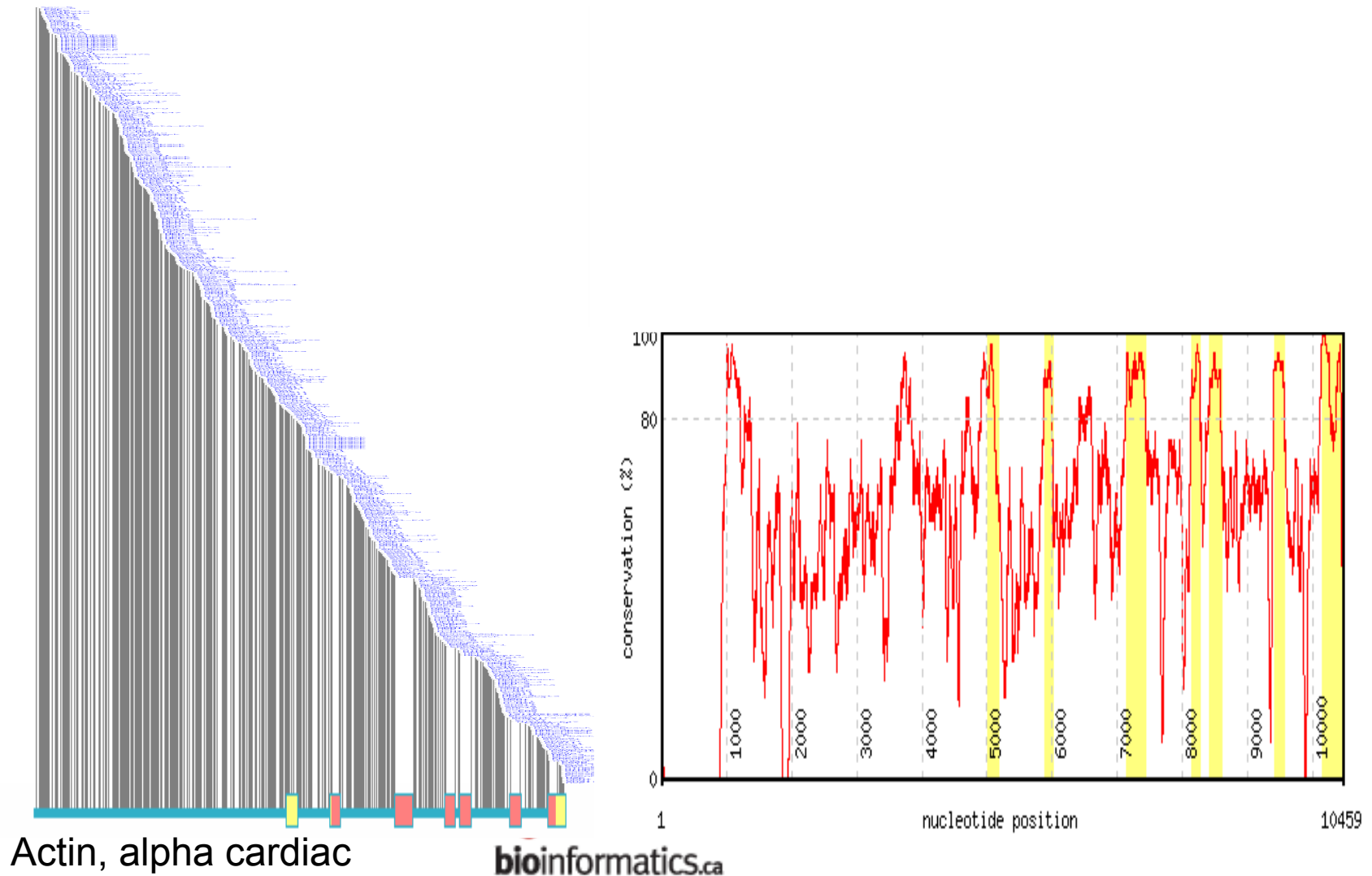
- Align orthologous gene sequences (e.g. LAGAN)
- For first window of 100 bp, of sequence#1, determine the % with identical match in sequence#2
 - Step across the first sequence, recording the percentage of identical nucleotides in each window
- Observe that single exon contains a region of high identity that corresponds to the ORF, with lower identity in the 5' and 3' UTRs
- Additional conserved region could be regulatory regions

Phylogenetic Footprinting (cont)

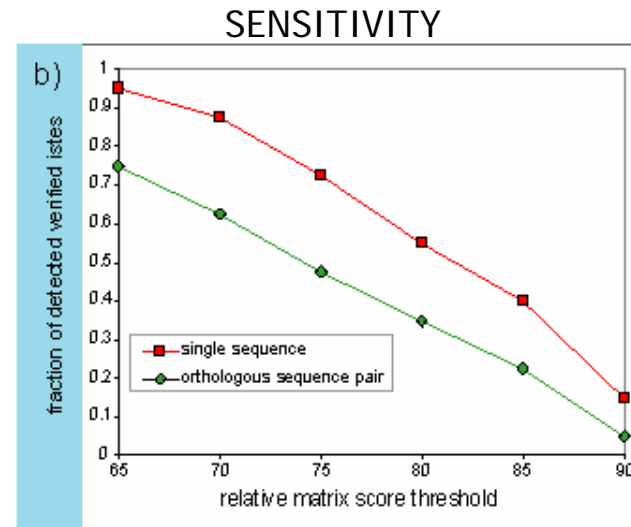
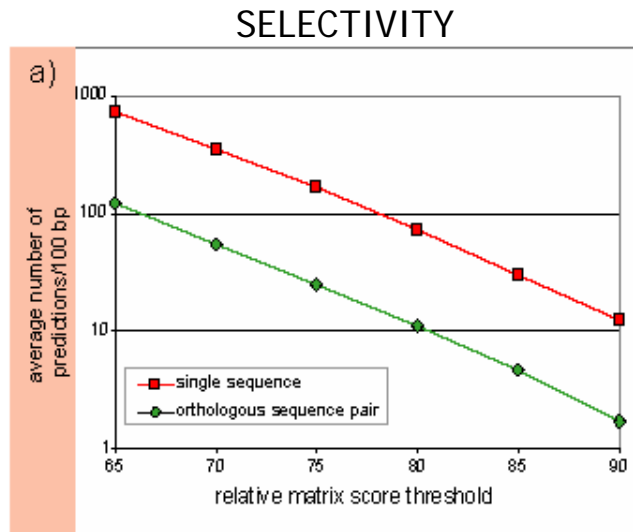


Actin gene compared between human and mouse

Phylogenetic Footprinting Dramatically Reduces Spurious Hits



TFBS Prediction with Human & Mouse Pairwise Phylogenetic Footprinting



- Testing set: 40 experimentally defined sites in 15 well studied genes (Replicated with 100+ site set)
- 75-80% of defined sites detected with conservation filter, while only 11-16% of total predictions retained

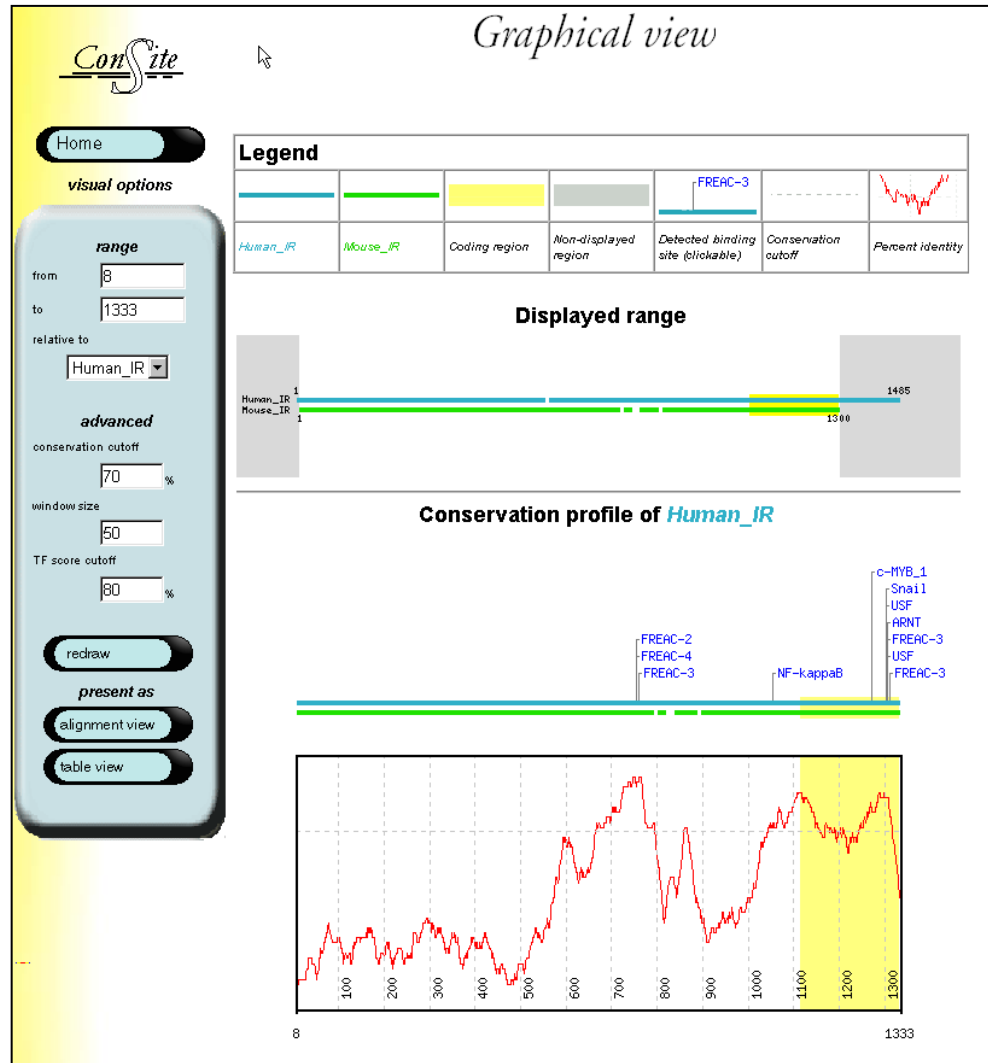
1 kbp insulin receptor promoter screened with footprinting



Multi-species Phylogenetic Footprinting

- In bioinformatics one never wishes to ignore useful information...
 - Pairwise comparisons do not take full advantage of the growing set of sequenced genomes
- New algorithms (e.g. Monkey) weight TFBS predictions based on retention over a branch of a species tree
 - Method is compute intensive, as each predicted TFBS is assessed against all other predictions
- Not clear what the relative benefits of multi-species methods will be...
 - Some suggestions that the best pairwise comparison gives similar results to a multi-species comparison

ConSite

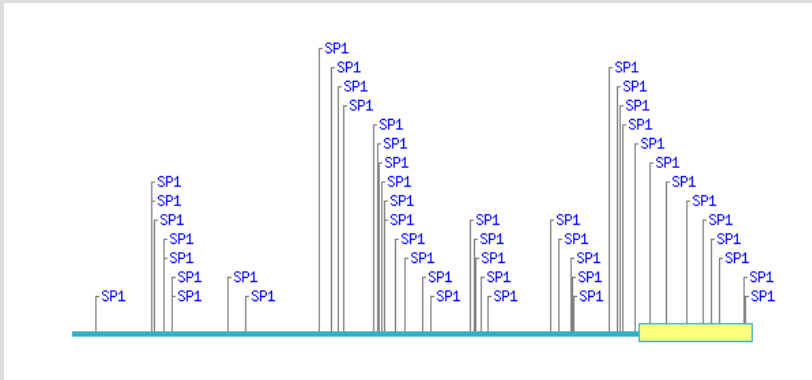


OnLine Resources for Phylogenetic Footprinting

- Linked to TFBS
 - ConSite
 - rVISTA
- Alignments
 - Blastz
 - Lagan/mLAGAN
 - Avid
 - ORCA
- Visualization
 - Sockeye
 - Vista Browser
 - PipMaker

Analysis of TFBS with Phylogenetic Footprinting

Scanning a single sequence

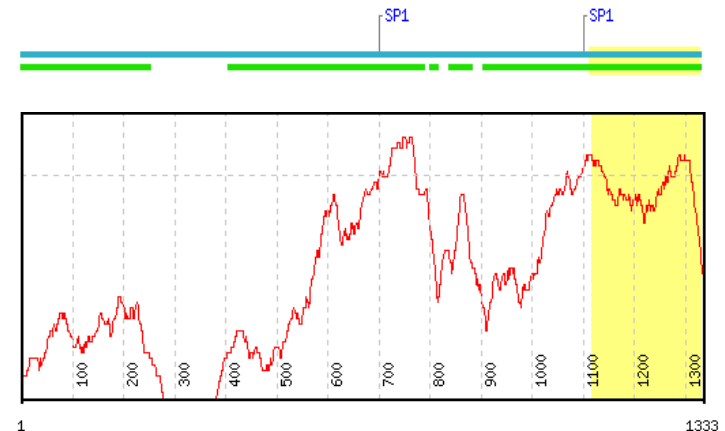


Low specificity of profiles:

- too many hits
- great majority not biologically significant

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

A dramatic improvement in the percentage of biologically significant detections



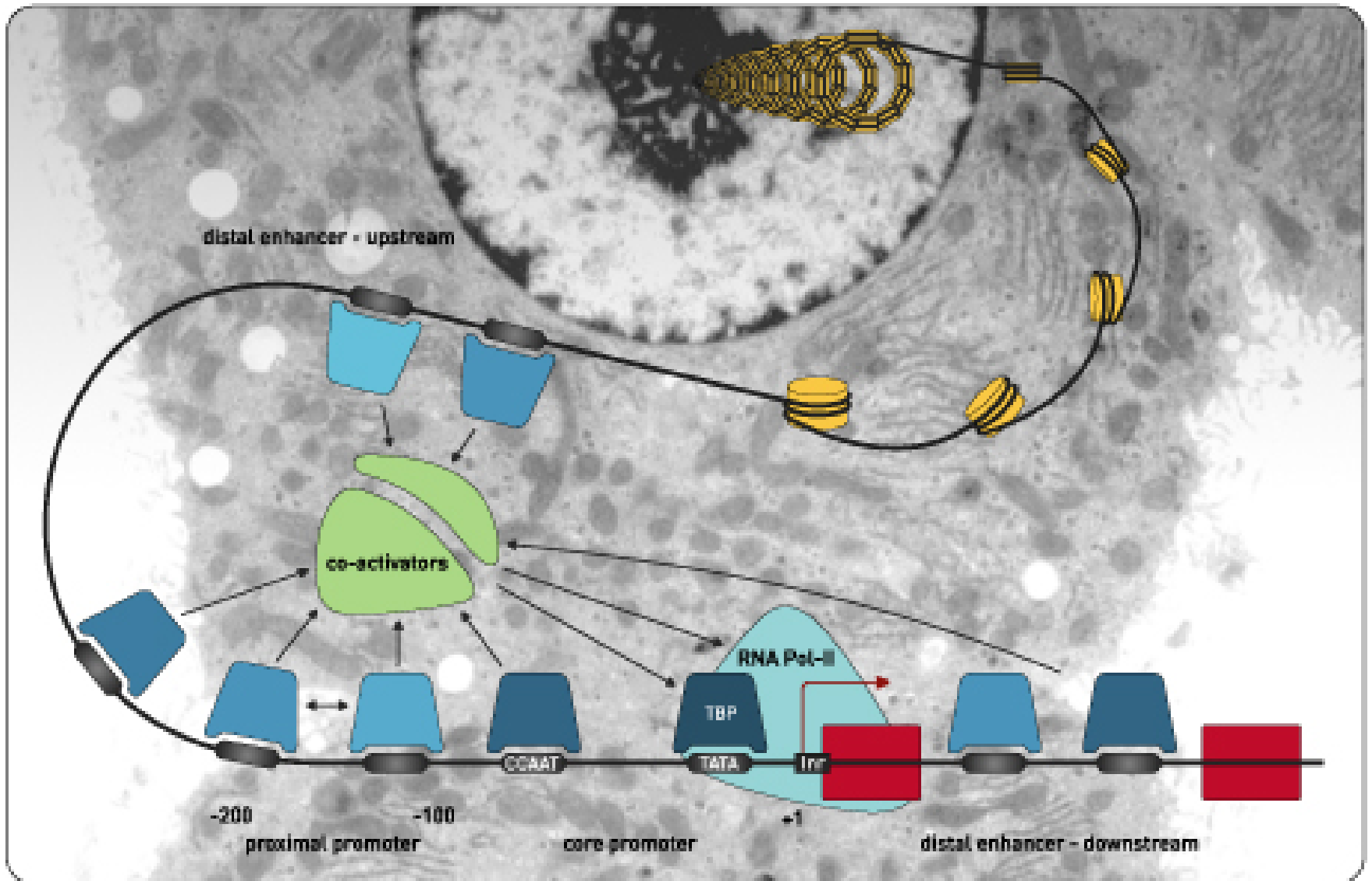
TFBS Phylogenetic Footprinting

- Binding site prediction coupled with
 - Assumes reasonable pairing of orthologs
- Available online resources support

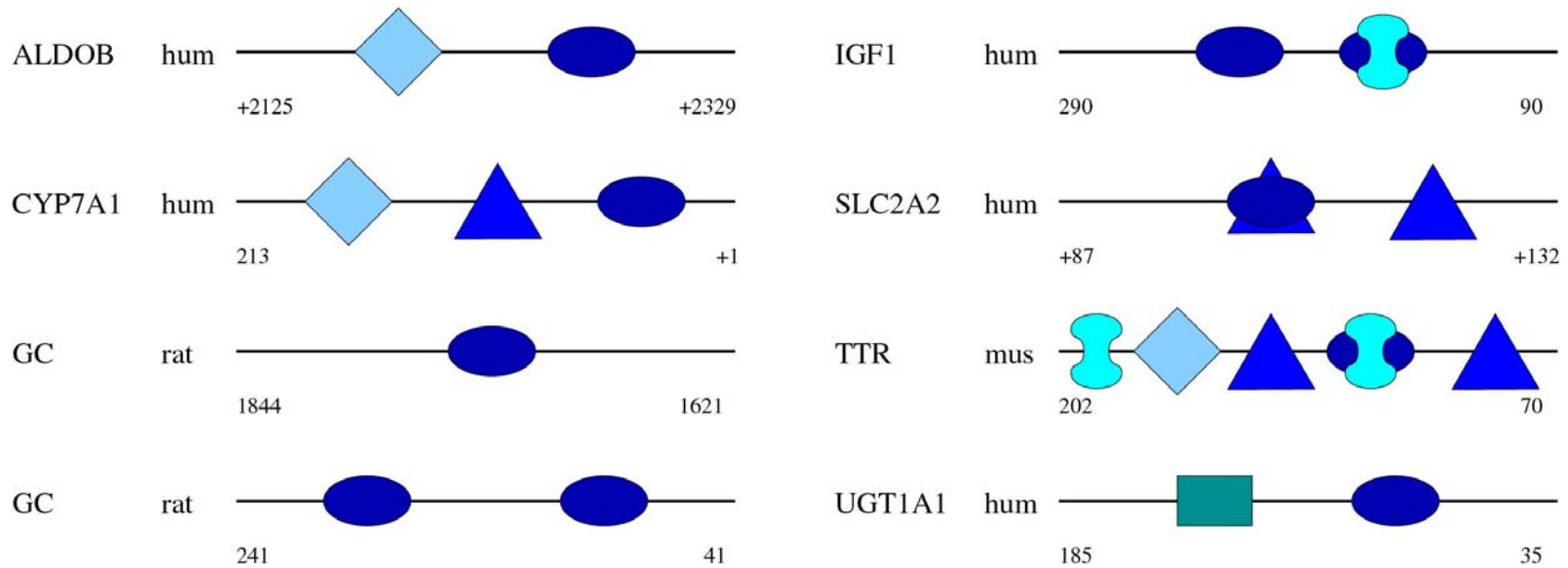
Discrimination of Regulatory Modules

TFs do NOT act in isolation
(THIS SECTION IS BRIEF DUE TO TIME CONSTRAINTS)

Recall (again)



Known *cis*-regulatory modules (CRMs) for specific expression in hepatocytes



HNF1



HNF3



HNF4



C/EBP



Sp1

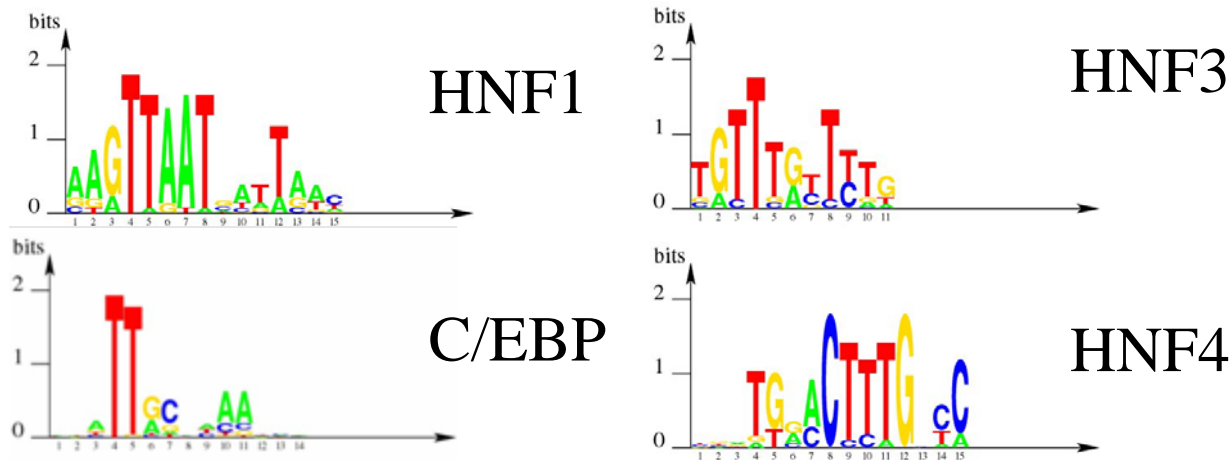


Detecting Clusters of TFBS

- GOAL: Given a set of profiles for TFs known (or hypothesized) to act together, teach computer to find clusters of TFBS
- Trained Methods
 - Sufficient examples of real clusters to establish weights on the relative importance of each TF
- Statistical Over-Representation of Combinations
 - Binding profiles available for a set of biologically motivated TFs
 - Usually confounded by the non-random properties of genomic sequences
 - Requires substantial effort to model local sequence properties in order to determine significance

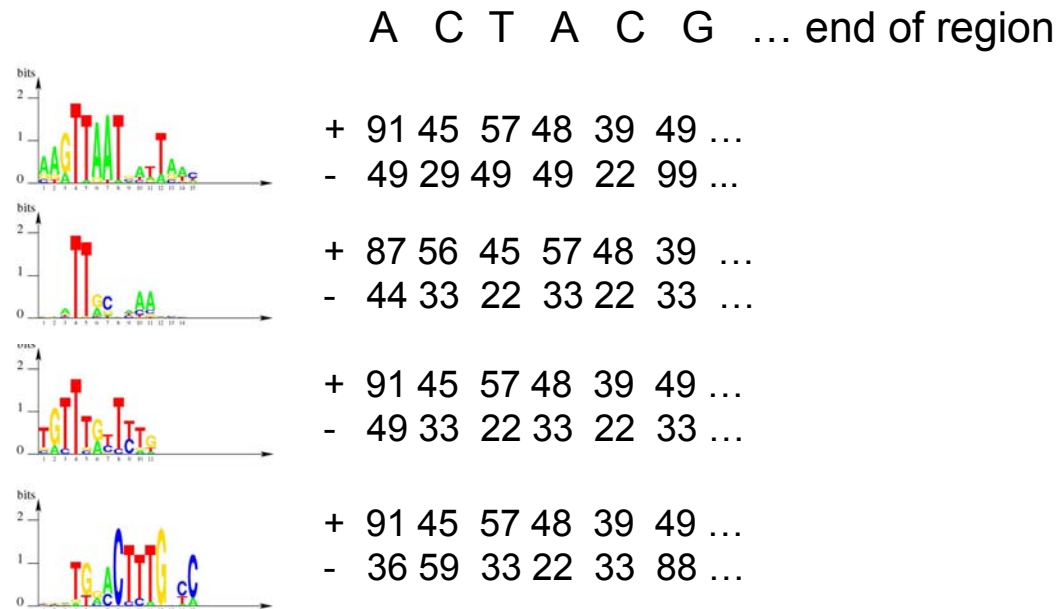
Building a trained model (1)

Step 1: Obtain a set of PSSMs for the mediating TFs



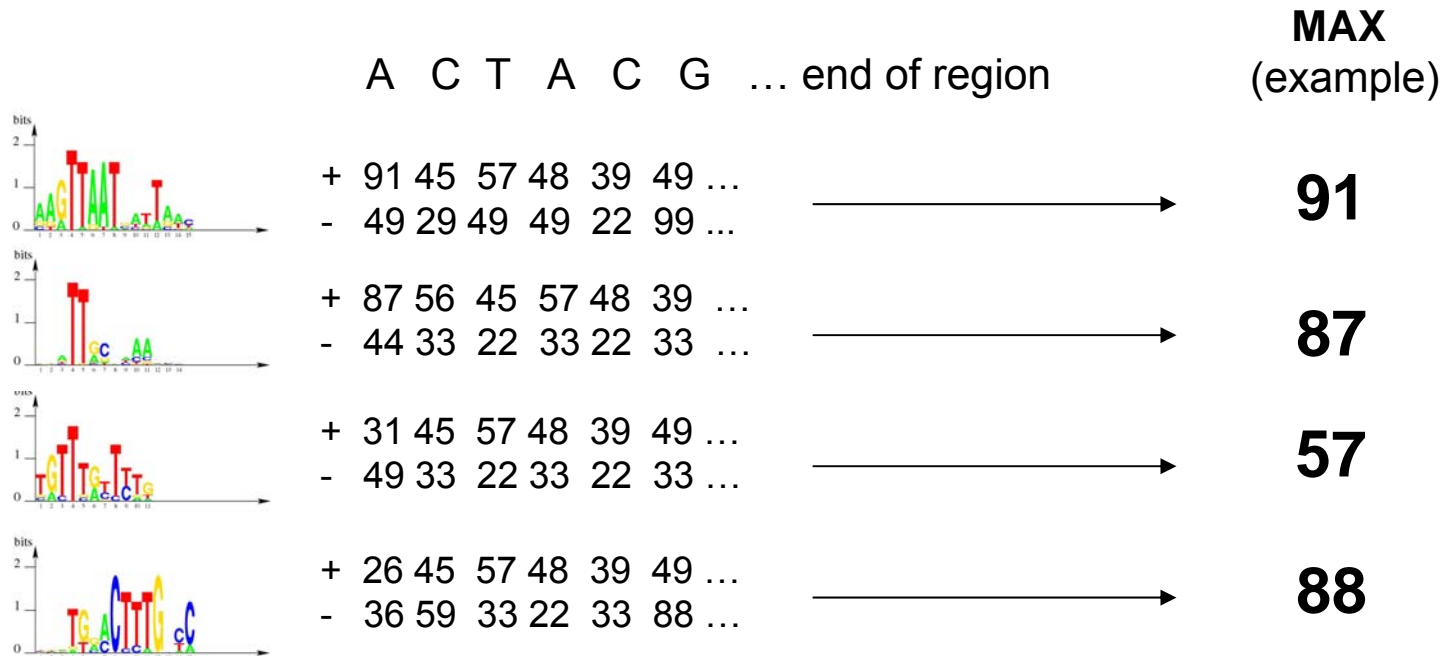
Building a trained model (2)

Step 2: Score all possible sites in each reference sequence with each profile (don't forget second strand)



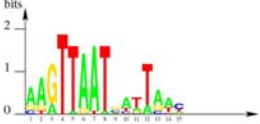
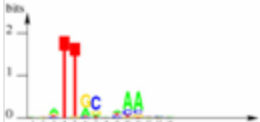
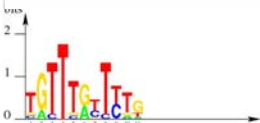
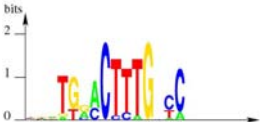
Building a trained model (3)

Step 3: Filter the scores (many possible approaches at this stage)



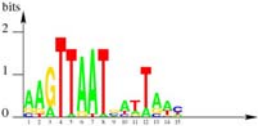
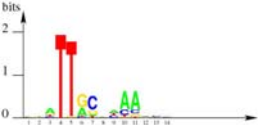
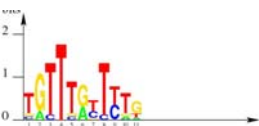

Building a trained model (4)

Step 4: Obtain scores for each sequence...

	HEPATOCYTE MODULES				NEGATIVE CONTROLS			
	MAX_{H1}	MAX_{H2}	...	MAX_{Hn}	MAX_{C1}	MAX_{C2}	MAX_{Cn}
	91	75	...	82	45	56	...	87
	87	34	...	56	33	44	...	28
	57	44	...	33	48	37	...	55
	88	44	...	27	22	33	...	44

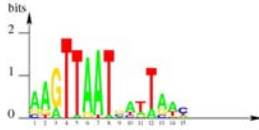
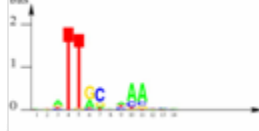
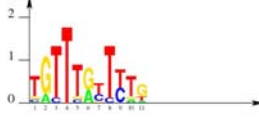
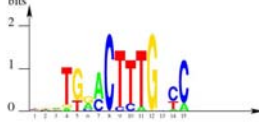
Building a trained model (5)

Step 5: Determine a weight to place upon the scores of each profile...

	HEPATOCYTE MODULES				NEGATIVE CONTROLS				WEIGHTS
	MAX_{H1}	MAX_{H2}	...	MAX_{Hn}	MAX_{C1}	MAX_{C2}	MAX_{Cn}	
	91	75	...	82	45	56	...	87	.1
	87	34	...	56	33	44	...	28	.2
	57	44	...	33	48	37	...	55	0
	88	44	...	27	22	33	...	44	.2

Building a trained model (6)

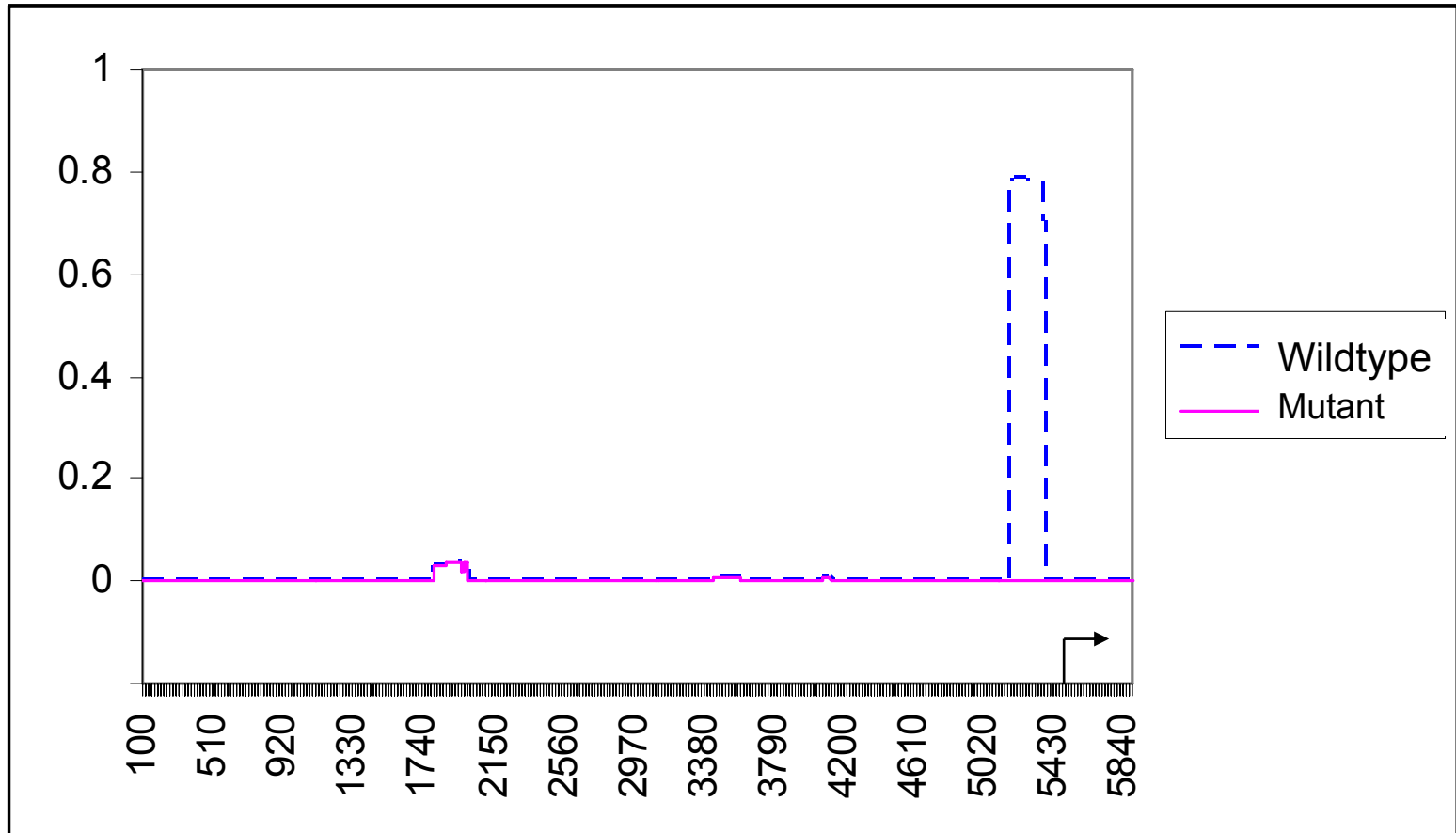
Step 6: Calculate score for test cases ...

	TEST CASE			
	MAX _{T1} * WEIGHT =			
	71	*	0.1	= 7
	88	*	0.2	= 17
	97	*	0	= 0
	87	*	0.2	= 17
				<hr/>
				41

FINAL SCORE FOR TEST SEQUENCE#1

UGT1A1

Liver Module Model Score/MaxScore



“Window” Position in Sequence

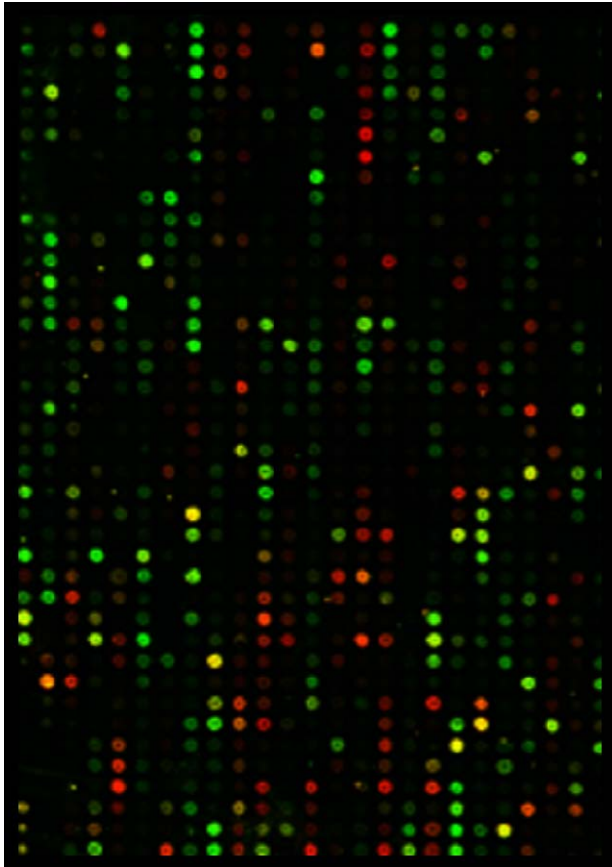
Final Points on CRM Detection

- Most procedures use advanced weighting procedures and do not limit to single maximum scoring TFBS
 - for instance HMMs and Logistic Regression Analysis
- Interpretation of score depends on tolerance for false predictions
 - Most publications assess the false positive rate of CRM prediction procedures at sensitivity of 66%
 - » Artifact of history
- Most trained methods generate false positives at a rate between 1/30000 bp – 1/60000
 - Untrained methods in best cases generate predictions at rates between 1/10000 bp – 1/18000

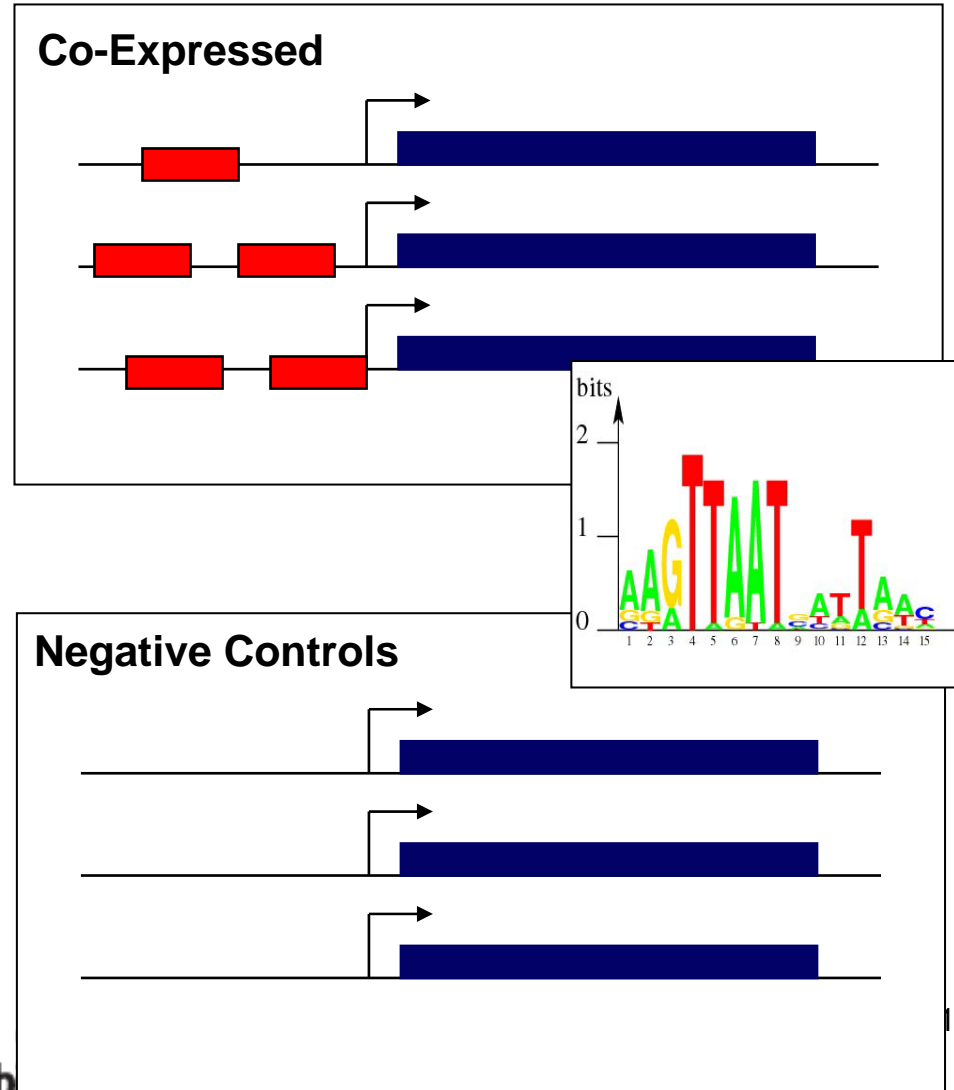
Part 3:

Inferring Regulating TFs for Sets of Co-Expressed Genes

Deciphering Regulation of Co-Expressed Genes



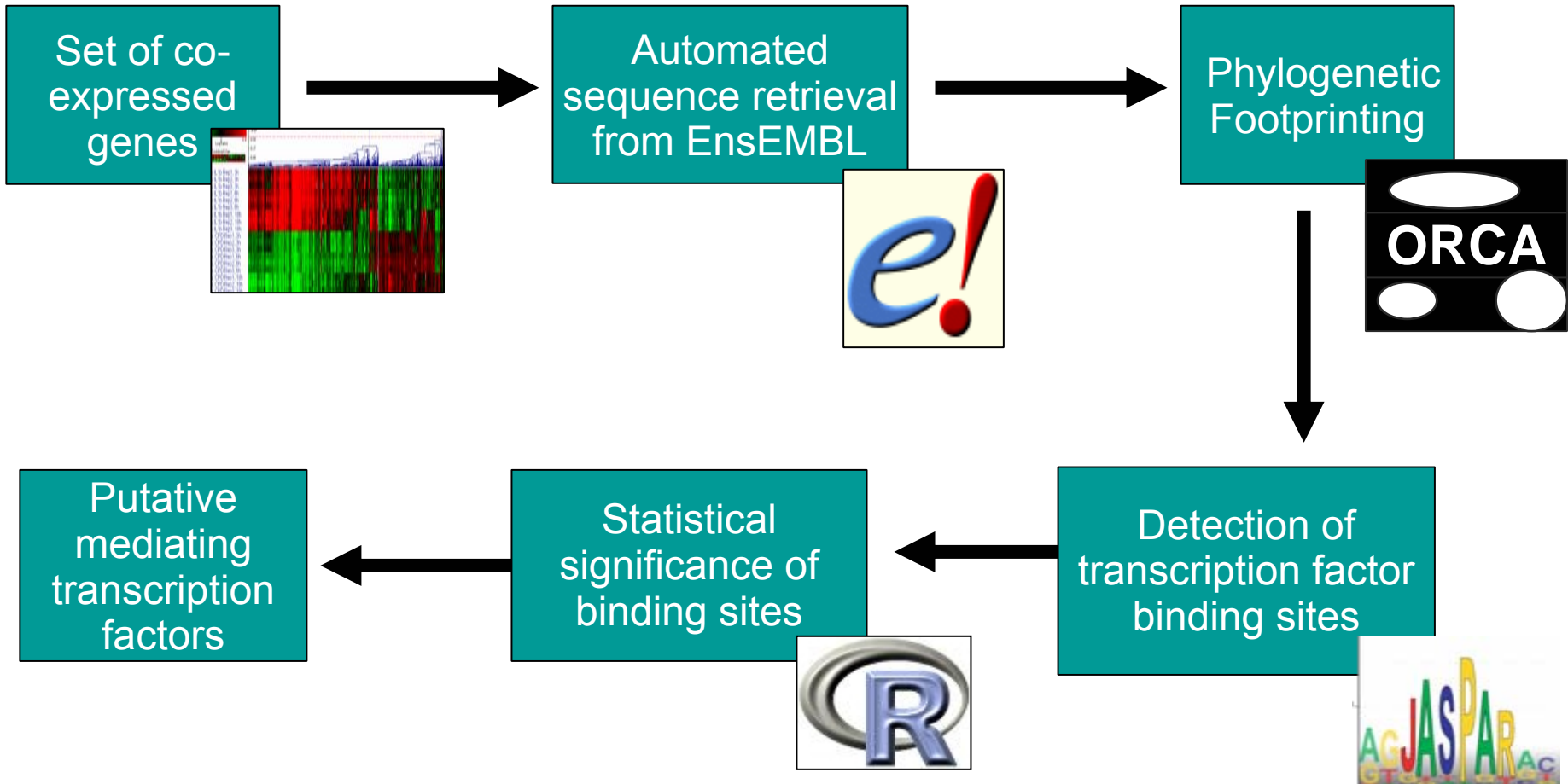
Lecture 5.0



TFBS Over-representation

- Akin to the GO studies yesterday, it would be convenient to identify if a set of co-expressed genes contains an over-abundance of binding sites for a known TF
- We will use phylogenetic footprinting to
- Can over-representation studies be successful?

oPOSSUM Procedure



Statistical Methods for Identifying Over-represented TFBS

- Z scores
 - Based on the **number of occurrences** of the TFBS relative to background
 - Normalized for sequence length
 - Simple binomial distribution model
- Fisher exact probability scores
 - Based on the **number of genes** containing the TFBS relative to background
 - Hypergeometric probability distribution

The oPOSSUM Database

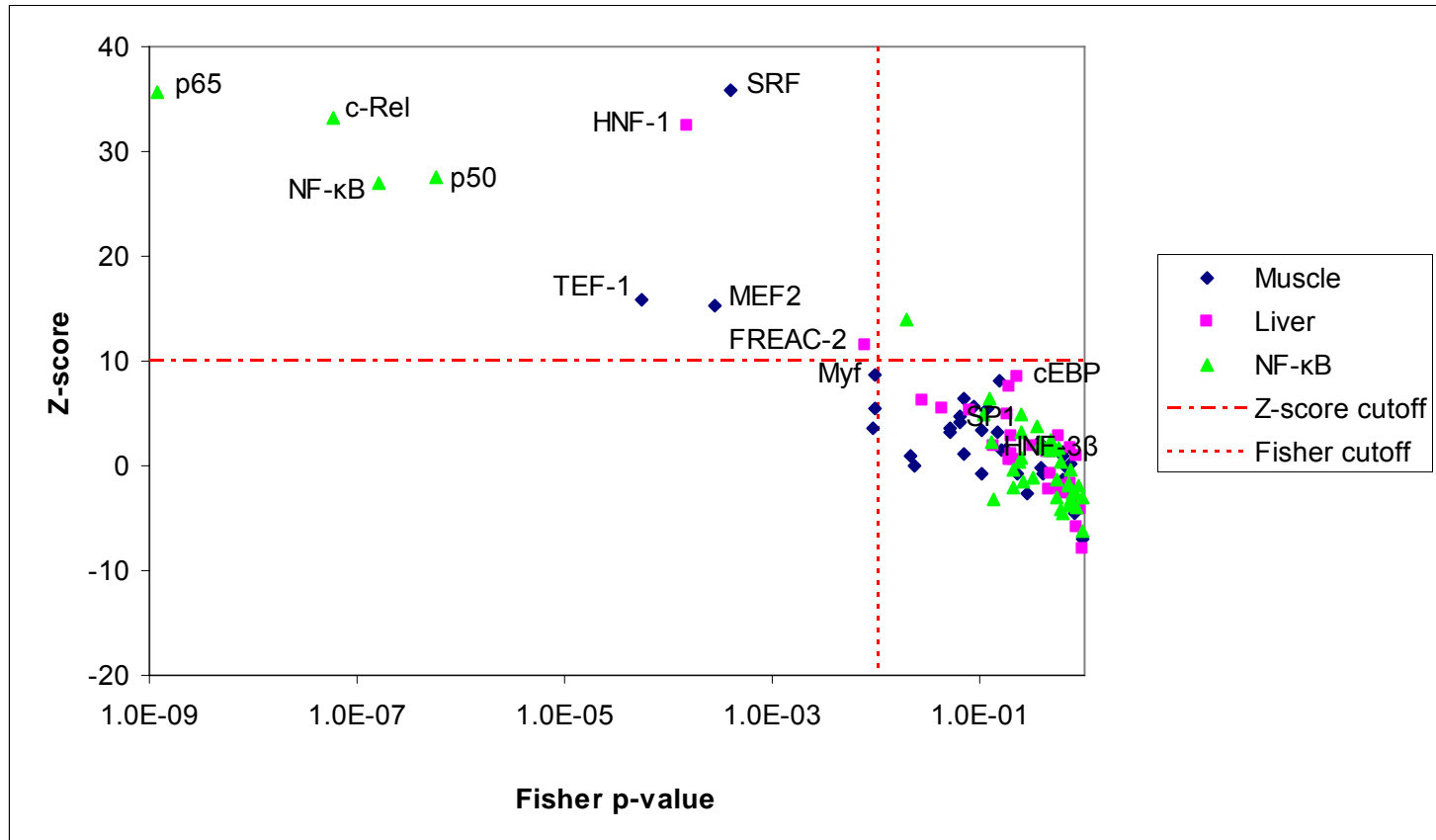
- Orthologous genes: 8468
- Promoter pairs: 6911
- Promoters with TFBS: 6758
- Total # of TFBS predictions: 1638293
- Overall failure rate: 20.2%

Validation using Reference Gene Sets

A. Muscle-specific (23 input; 16 analyzed)				B. Liver-specific (20 input; 12 analyzed)			
	Rank	Z-score	Fisher		Rank	Z-score	Fisher
SRF	1	21.41	1.18e-02	HNF-1	1	38.21	8.83e-08
MEF2	2	18.12	8.05e-04	HLF	2	11.00	9.50e-03
c-MYB_1	3	14.41	1.25e-03	Sox-5	3	9.822	1.22e-01
Myf	4	13.54	3.83e-03	FREAC-4	4	7.101	1.60e-01
TEF-1	5	11.22	2.87e-03	HNF-3beta	5	4.494	4.66e-02
deltaEF1	6	10.88	1.09e-02	SOX17	6	4.229	4.20e-01
S8	7	5.874	2.93e-01	Yin-Yang	7	4.070	1.16e-01
Irf-1	8	5.245	2.63e-01	S8	8	3.821	1.61e-02
Thing1-E47	9	4.485	4.97e-02	Irf-1	9	3.477	1.69e-01
HNF-1	10	3.353	2.93e-01	COUP-TF	10	3.286	2.97e-01

 TFs with experimentally-verified sites in the reference sets.

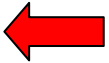

Empirical Selection of Parameters based on Reference Studies



C-Myc SAGE Data

- c-Myc transcription factor dimerizes with the Max protein
- Key regulator of cell proliferation, differentiation and apoptosis
- Menssen and Hermeking identified 216 different SAGE tags corresponding to unique mRNAs that were induced after adenoviral expression of c-Myc in HUVEC cells
- They then went on to confirm the induction of 53 genes using microarray analysis and RT-PCR


Induced Genes after Ectopic Expression of c-Myc (SAGE) (53 input; 36 analyzed)

	TF Class	Rank	Z-score	Fisher	No. Genes
Myc-Max 	bHLH-ZIP	1	21.68	5.35e-03	7
Staf	ZN-FINGER, C2H2	2	20.17	1.70e-02	2
Max 	bHLH-ZIP	3	18.32	2.16e-02	12
SAP-1	ETS	4	13.23	1.61e-04	13
USF	bHLH-ZIP	5	11.90	1.84e-01	16
SP1	ZN-FINGER, C2H2	6	11.68	4.40e-02	12
n-MYC 	bHLH-ZIP	7	11.11	1.55e-01	20
ARNT	bHLH	8	11.11	1.55e-01	20
Elk-1	ETS	9	10.92	3.88e-03	19
Ahr-ARNT	bHLH	10	10.17	1.11e-01	25

C-Fos Microarray Experiment

- In a study examining the role of transcriptional repression in oncogenesis, Ordway *et al.* compared the gene expression profiles of fibroblasts transformed by c-fos to the parental 208F rat fibroblast cell line
- We mapped the list of 252 induced Affymetrix Rat Genome U34A GeneChip sequences to 136 human orthologs

Induced Genes after Ectopic Expression of c-Fos (Affymetrix) (136 input; 86 analyzed)

	TF Class	Rank	Z-score	Fisher	No. Genes
c-FOS 	bZIP	1	17.53	2.60e-05	45
RREB-1	ZN-FINGER, C2H2	2	8.899	1.41e-01	1
PPARgamma-RXRalpha	NUCLEAR RECEPTOR	3	3.991	2.98e-01	1
CREB	bZIP	4	3.626	1.25e-01	10
E2F	Unknown	5	2.965	7.67e-02	15


oPOSSUM Server

oPOSSUM: Select Analysis Parameters - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://sonoma.cmmt.ubc.ca/cgi-bin/oPOSSUM/opossum> Go

Google Search Web 386 blocked AutoFill Options



[Home](#)
[About](#)
[Help](#)
[FAQ](#)
[Contact](#)

oPOSSUM

Web-based analysis of over-represented transcription factor binding sites

Select Analysis Parameters

STEP 1: Enter a list of co-expressed genes

ID type: Ensembl HUGO Accession LocusLink/Entrez Gene ID Rosetta Chip ID

Paste gene IDs:

Use sample genes Clear

OR upload a file containing a list of gene identifiers:
Browse...

STEP 2: Select transcription factor binding site matrices

Done Internet


oPOSSUM: Select Analysis Parameters - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Refresh Mail Print TV Options

Address <http://sonoma.cmmt.ubc.ca/cgi-bin/oPOSSUM/opossum> Go

Google Search Web 386 blocked AutoFill Options



oPOSSUM

Web-based analysis of over-represented transcription factor binding sites

Select Analysis Parameters

STEP 1: Enter a list of co

ID type: Ensembl HUGO Accession

Paste gene IDs:

Use sample genes Clear

OR upload a file containing a list of gene identifiers:

Browse...

STEP 2: Select transcription factor binding site matrices

Done Internet

INPUT A LIST OF
CO-EXPRESSED GENES

oPOSSUM: Select Analysis Parameters - Microsoft Internet Explorer

File Edit View Favorites Tools Help Google

Back Forward Stop Refresh Home Search Favorites Media

Address <http://www.cisreg.ca/cgi-bin/oPOSSUM/opossum> Go

STEP 2: Select transcription factor binding site matrices

All profiles with a minimum

OR select by taxonomic s

plant vertebrate inse

OR select specific profiles

- AGL3
- AML-1
- ARNT
- Agamous
- Ahr-ARNT
- Androgen
- Athb-1
- Brachyury

SELECT YOUR TFBS PROFILES

STEP 3: Select parameters

Level of conservation:

Done Internet

oPOSSUM: Select Analysis Parameters - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Google

Back Forward Stop Refresh Home Search Favorites Media

Address <http://www.cisreg.ca/cgi-bin/oPOSSUM/opussum> Go

Ahr-ARNT
Androgen
Athb-1
Brachyury

STEP 3: Select parameters

Level of conservation:
1 (top 10.0% of conserved regions)

Matrix match threshold:
80.0 %

Amount of upstream / downstream sequence:
5000 / 5000

Statistical measure for over-representation:
Both

Press the **Submit** button to perform the analysis or **Reset** to reset the analysis parameters to their default values. Depending on server load, the analysis may take anywhere from a few seconds to a minute or more to perform. Please be patient.

SELECT:

1. CONSERVATION
2. PSSM MATCH THRESHOLD
3. PROMOTER REGION
4. STATISTICAL MEASURE

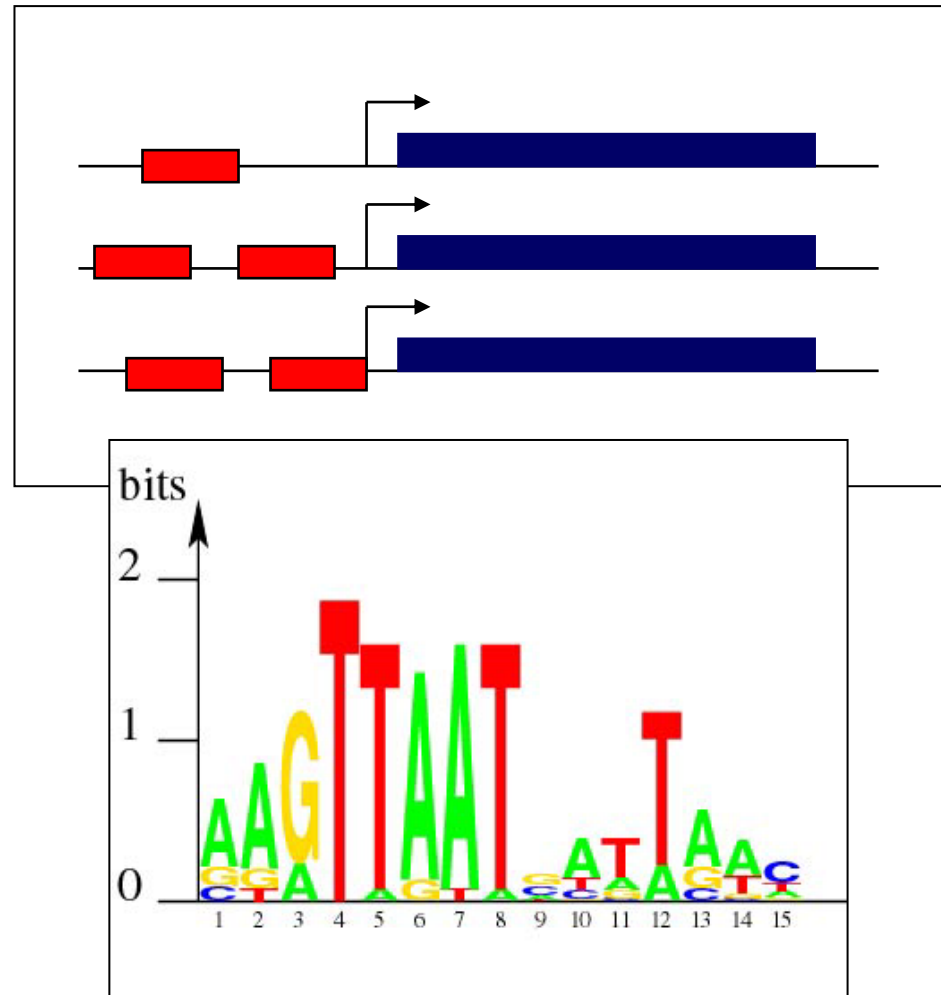
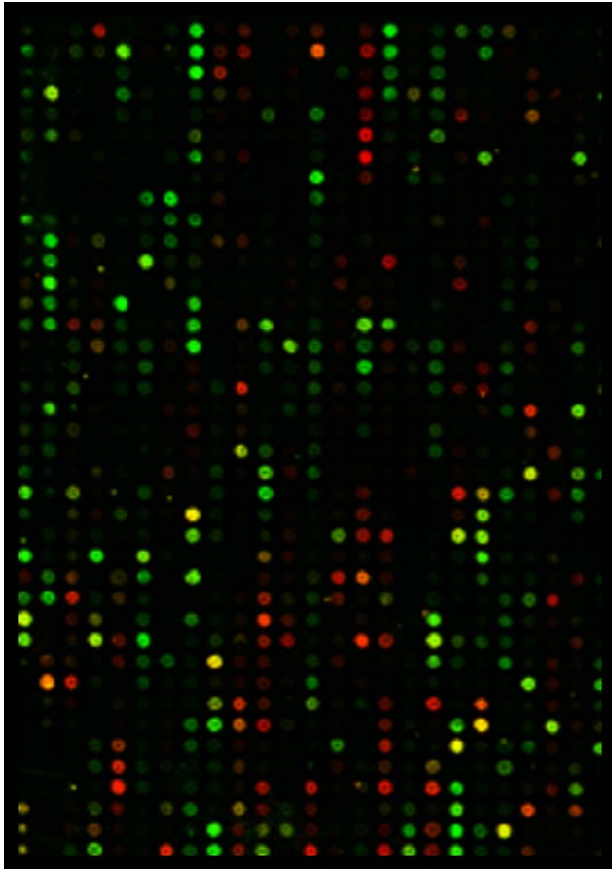
TFBS Over-Representation Summary

- New generation of tools to help interrogate the meaning of observed clusters of co-expressed genes
- Still in development, so procedures have the potential to improve
 - For example, seek over-represented clusters of TFBS
- Generally best performance has been in studies directly linked to a transcription factor
 - Highly dependent on the experimental design – cannot overcome noisy data from poor design

Part 4:

de novo Discovery of TF Binding Sites

De novo Pattern Discovery



de novo Pattern Discovery

- String-based
 - e.g. YMF (Sinha & Tompa)
 - Generalization: Identify over-represented oligomers in comparison of “+” and “-” (or complete) promoter collections
 - Used often for yeast promoter analysis
- Profile-based
 - e.g. AnnSpec (Workman & Stormo) or MEME (Bailey & Elkin)
 - Generalization: Identify strong patterns in “+” promoter collection vs. background model of expected sequence characteristics

String-based methods(1)

How likely are X words in a set of sequences, given background sequence characteristics?

```
CCCCCCGGAATGAAATCTGATTGACATTTTCC >EP71002 (+) Ce[IV] msp-56 B; range -100 to -75
TTCAAATTTTAAACGCCGGAATAATCTCCTATT >EP63009 (+) Ce Cuticle Col-12; range -100 to -75
TCGCTGTAACCGGAATATTTAGTCAGTTTTTG >EP63010 (+) Ce Cuticle Col-13; range -100 to -75
TATCGTCATTCTCCGCCTCTTTTCTT >EP11013 (+) Ce vitellogenin 2; range -100 to -75
GCTTATCAATGCGCCCGGAATAAAACGCTATA >EP11014 (+) Ce vitellogenin 5; range -100 to -75
CATTGACTTTATCGAATAAATCTGTT >EP11015 (-) Ce vitellogenin 4; range -100 to -75
ATCTATTTACAATGATAAACTTCAA >EP11016 (+) Ce vitellogenin 6; range -100 to -75
ATGGTCTCTACCGGAAAGCTACTTTCAGAATT >EP11017 (+) Ce calmodulin cal-2; range -100 to -75
TTTCAAATCCGGAATTTCCACCCGGAATTACT >EP63007 (-) Ce cAMP-dep. PKR P1+; range -100 to -75
TTTCCTTCTTCCGGAATCCACTTTTCTTCC >EP63008 (+) Ce cAMP-dep. PKR P2; range -100 to -75
ACTGAACCTGTCTTCAAATTTCAACACCGGAA >EP17012 (+) Ce hsp 16K-1 A; range -100 to -75
TCAATGCCGGAATTCTGAATGTGAGTCGCCCT >EP55011 (-) Ce hsp 16K-1 B; range
```

String-based methods(2)

Find all words of length n in the yeast promoters (e.g. $n=7$)

```
GTCTTATCTTCAAAGTTGTCTGTCCAAGATTTGGACTTGAAGG
ACAAGCGTGTCTTCTCAGAGTTGACTTCAACGTCCCATTGGAC
GGTAAGAAGATCACTTCTAACCAAAGAATTGTTGCTGCTTTGC
CAACCATCAAGTACGTTTTGGAACACCACCCAAGATACGTTGT
CTTGTCTCACTTGGGTAGACCAAACGGTCAAAGAAACGAAAA
ATACTCTTTGGCTCCAGTTGCTAAGGAATTGCAATCATTGTTG
GGTAAGGATGTCACCTTCTTGAACGACTGTGTCGGTCCAGAA
GTTGAAGCCGCTGTCAAGGCTTCTGCCCCAGGTTCCGTTATTT
TGTTGGAAACTGCGTTACCACATCGAAGAAGAAGGTTCCAGA
AAGGTCGATGGTCAAAAAGGTCAAGGCTCAAGGAAGATGTTCA
AAAGTTCAGACACGAATTGAGCTCTTTGGCTGATGTTTACATC
ACGATGCCTTCGGTACCGCTCACAGAGCTCACTCTTCTATGGT
CGGTTTCGACTTGCCAACGTGCTGCCGGTTTCTTGTTGGAAAA
GGAATTGAAGTACTTCGGTAAGGCTTTGGAGAACCCAACCAG
ACCATTCTTGCCATCTTAGGTGGTGCCAAGGTTGCTGACAAG
ATTCAATTGATTGACAACCTTGTGGACAAGGTCGACTCTATCAT
CATTGGTGGTGGTATGGCTTTCCTTCAAGAAGGTTTTGGAAA
ACACTGAAATCGGTGACTCCATCTTCGACAAGGCTGGTGCTG
AAATCGTTCCAAAGTTGATGGAAAAGGCCAAGGCCAAGGGTG
TCGAAGTCGTCTTGACGTCGACTTCATCATTGCTGATGCTTTC
TCTGCTGATGCCAACACCAAGACTGTCACTGACAAGGAAGGT
ATTCCAGCTGGCTGGCAAGGTTGGACAATGGTCCAGAATCT
AGAAAGTGTGTTGCTGCTACTGTTGCAAAGGCTAAGACCATTGT
CTGGAACGGTCCACCAGGTGTTTTCGAATTCGAAAAGTTCGCT
GCTGGTACTAAGGCTTTGTTAGACGAAGTTGTCAAGAGCTCTG
CTGCTGGTAACACCGTCATCATTGGTGGTGGTGACTGCCA
```

Make a lookup table:

AAACCTTT	456
TTTTTTTT	57788
GATAGGCA	589

Etc...

String-based methods(3)

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

X_w : Instances of a word w within our set of X genes

$E[X_w]$: Average number of instances of w based on number of genes in our set

$\text{Var}[X_w]$: Variance – how much deviation from the average is expected for w

Limitations of String-based Methods

- Longer word lengths not possible
- While many methods use degeneracy codes, TFBS are not words – we lose quantitation for variable positions
 - Imagine position with 7 A's and 1 T, at which we would represent W or throw out the instance with T

Probabilistic Methods for Pattern Discovery

- What is a probabilistic method?
- The Gibbs sampler algorithm

Probabilistic Methods

Overview:

Find a local alignment of width x of sites that **maximizes information content** (or related measure) in reasonable time

Usually by Gibbs sampling or EM methods

Motivation:

TFBS are not words

Efficiency – can handle longer patterns than string-based methods

Can be intentionally influenced to reflect prior knowledge

What does probabilistic mean?

- Based on probability
- Functionally, it means we're going to guess our way to a good pattern (TFBS)
 - We're going to try to make a good guess
- Two different flavours of the approach
 - Expectation Maximization in which we try to make the best guess each time
 - Gibbs Sampling in which we make our guesses based on the strength of our conviction

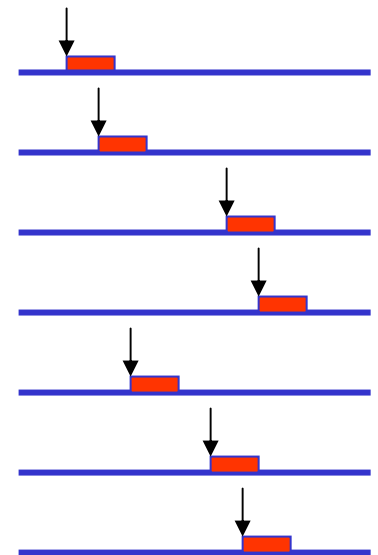
Gibbs Sampling

Two data structures used:

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

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

```
tgacttcc
tgatctct
agacctca
tgacctct
```



Iterations in Gibbs Sampling

A

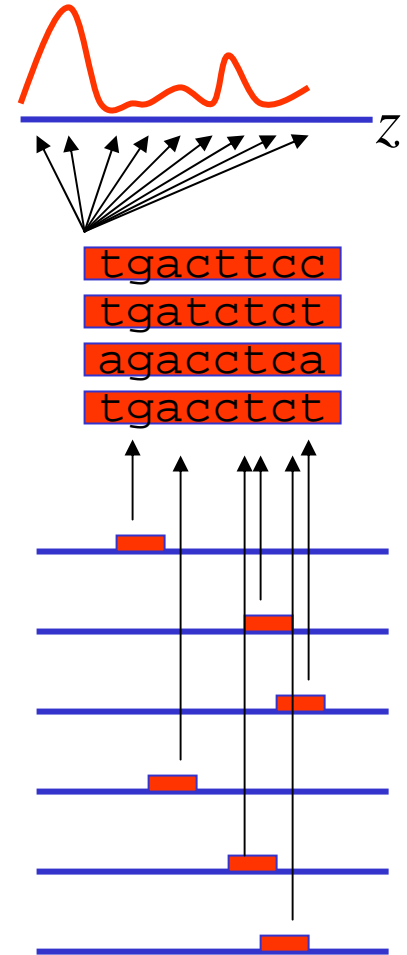
Remove one sequence z from the set. Update the current pattern according to

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

Pseudocount for symbol j
Sum of all pseudocounts in column

B

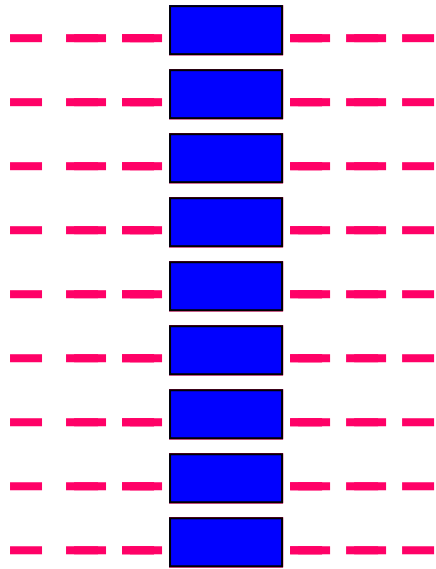
'Score' the current pattern against each possible occurrence a_k in z . Draw a new a_k with probabilities based on respective score divided by the background model



Pattern Discovery

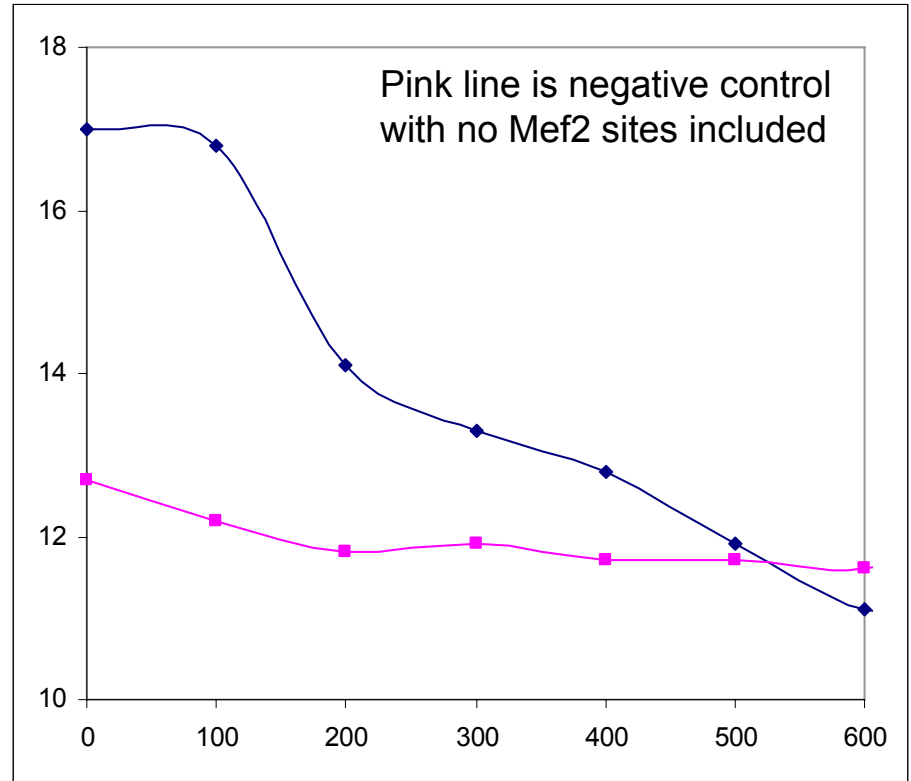
- Gibbs sampling is guaranteed to return an optimal pattern if repeated sufficiently often
 - Procedure is fast, so running many 1000s of times is feasible
- Unfortunately, we have a problem...what if our pattern of interest is not strong relative to other possible patterns...

Applied Pattern Discovery is Acutely Sensitive to Noise



True Mef2 Binding Sites

PATTERN SIMILARITY
VS. TRUE MEF2 PROFILE



SEQUENCE LENGTH

Four Approaches to Improve Sensitivity

- Better background models
 - Higher-order properties of DNA
- Phylogenetic Footprinting
 - Human:Mouse comparison eliminates ~75% of sequence
- Regulatory Modules
 - Architectural rules
- Limit the types of binding profiles allowed
 - TFBS patterns are NOT random

Pattern Discovery Summary

- Pattern discovery methods can recover over-represented patterns in the promoters of co-expressed genes
- Methods are acutely sensitive to noise, indicating that the signal we seek is weak
 - TFs tolerate great variability between binding sites
- As for pattern discrimination, supplementary information/approaches are required to overcome the noise
- Except in yeast, not quite ready for real world problems

REFLECTIONS

- Part 2
 - Futility Theorem – Essentially predictions of individual TFBS have no relationship to an *in vivo* function
 - Successful bioinformatics methods for site discrimination incorporate additional information (clusters, conservation)
- Part 3
 - TFBS over-representation is a power new means to identify TFs likely to contribute to observed patterns of co-expression
- Part 4
 - Pattern discovery methods are severely restricted by the Signal-to-Noise problem
 - Observed patterns must be carefully considered
 - Successful methods for pattern discovery will have to incorporate additional information (conservation, structural constraints on TFs)

THE END

- Questions before the break?
- Lab exercises address Sections 2 and 3