



Section 12.0

Transcription Factors, Binding Sites, and the Challenge of Finding Novel Problems in Bioinformatics ?

Wyeth Wasserman

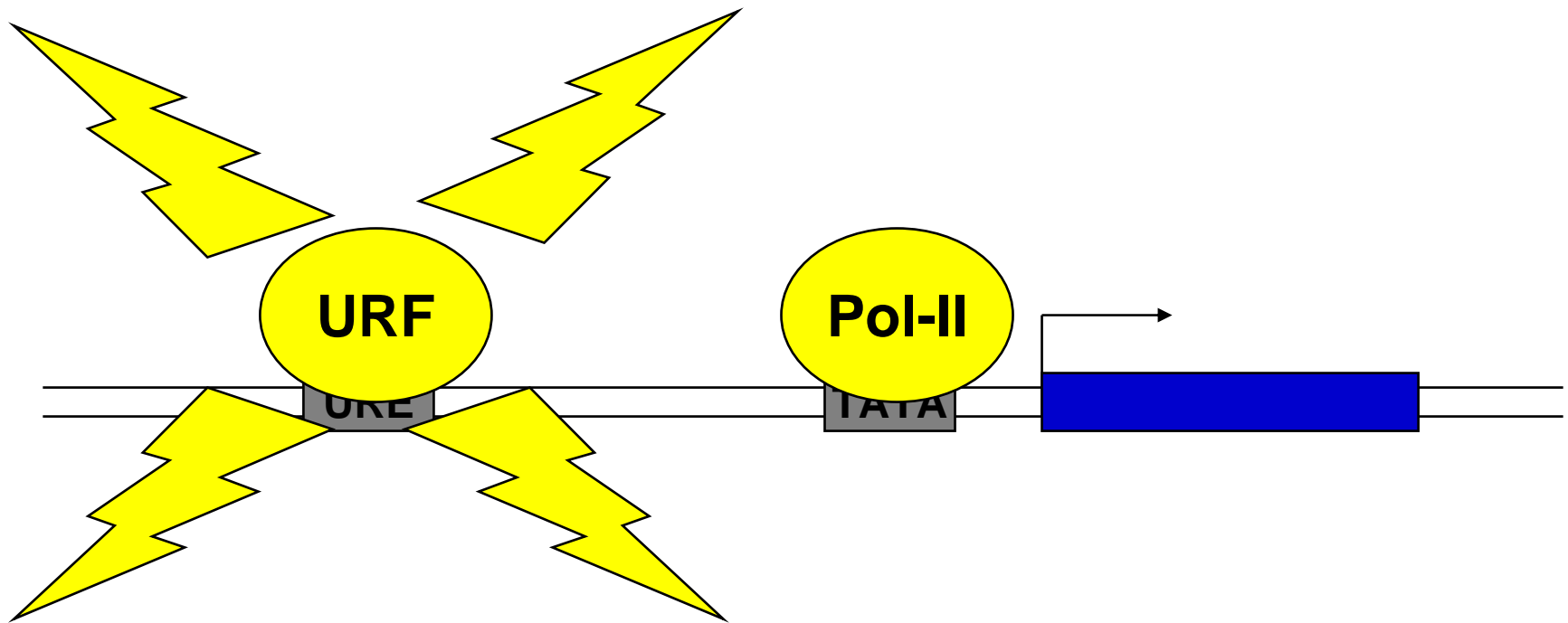


www.cisreg.ca

Overview

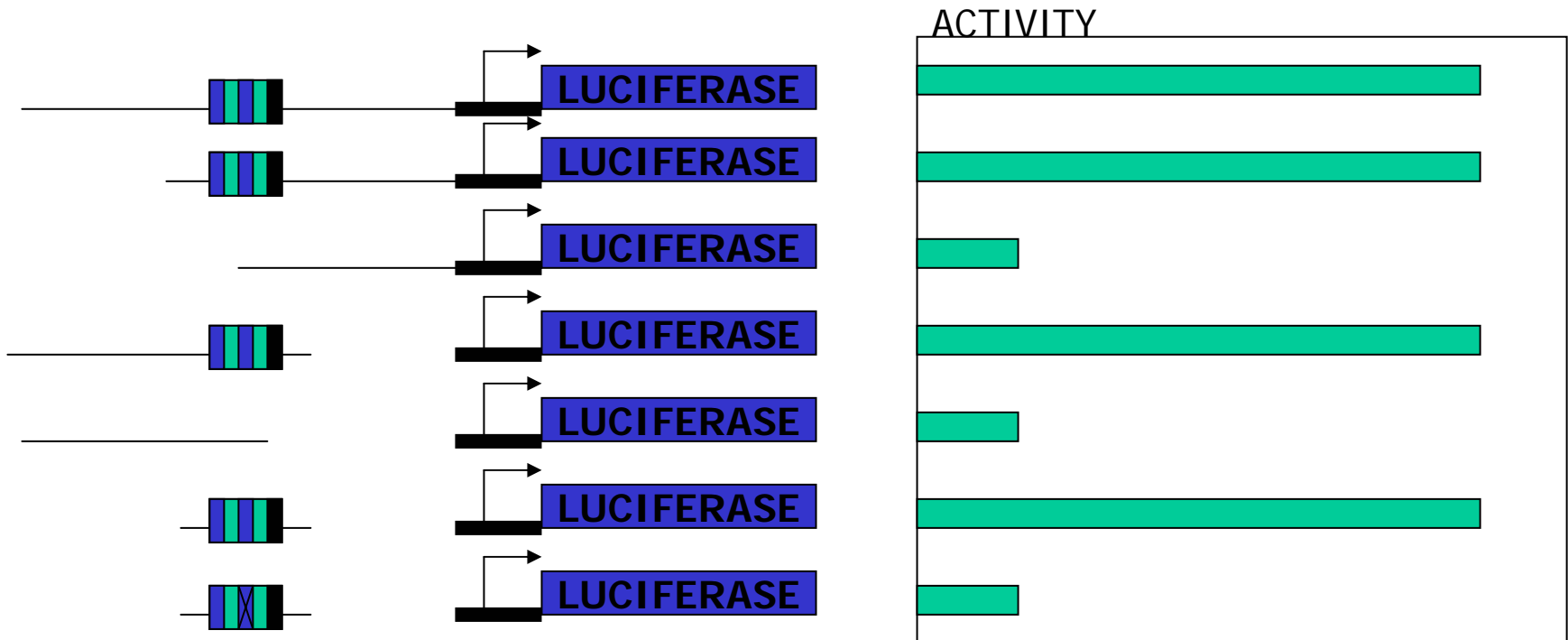
-
- TFBS Prediction with Motif Models
 - Improving Specificity of Predictions

Transcription Factor Binding Sites (over-simplified for pedagogical purposes)



**Teaching a computer
to find TFBS...**

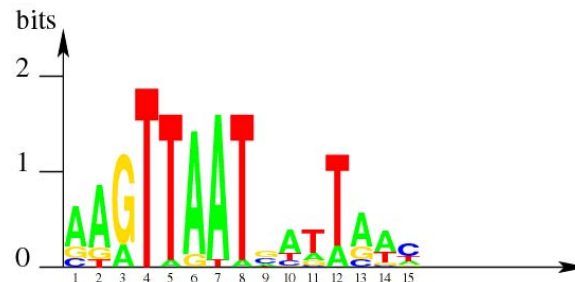
Laboratory Discovery of TFBS



Representing Binding Sites for a TF

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



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

PFMs to PWMs

Add the following features to the model:

1. Correcting for the base frequencies in DNA
2. Weighting for the confidence (depth) in the pattern
3. Convert to log-scale probability for easy arithmetic

f matrix

A	5	0	1	0	0
C	0	2	2	4	0
G	0	3	1	0	4
T	0	0	1	1	1

$$\text{Log} \left(\frac{f(b,i) + s(n)}{p(b)} \right)$$

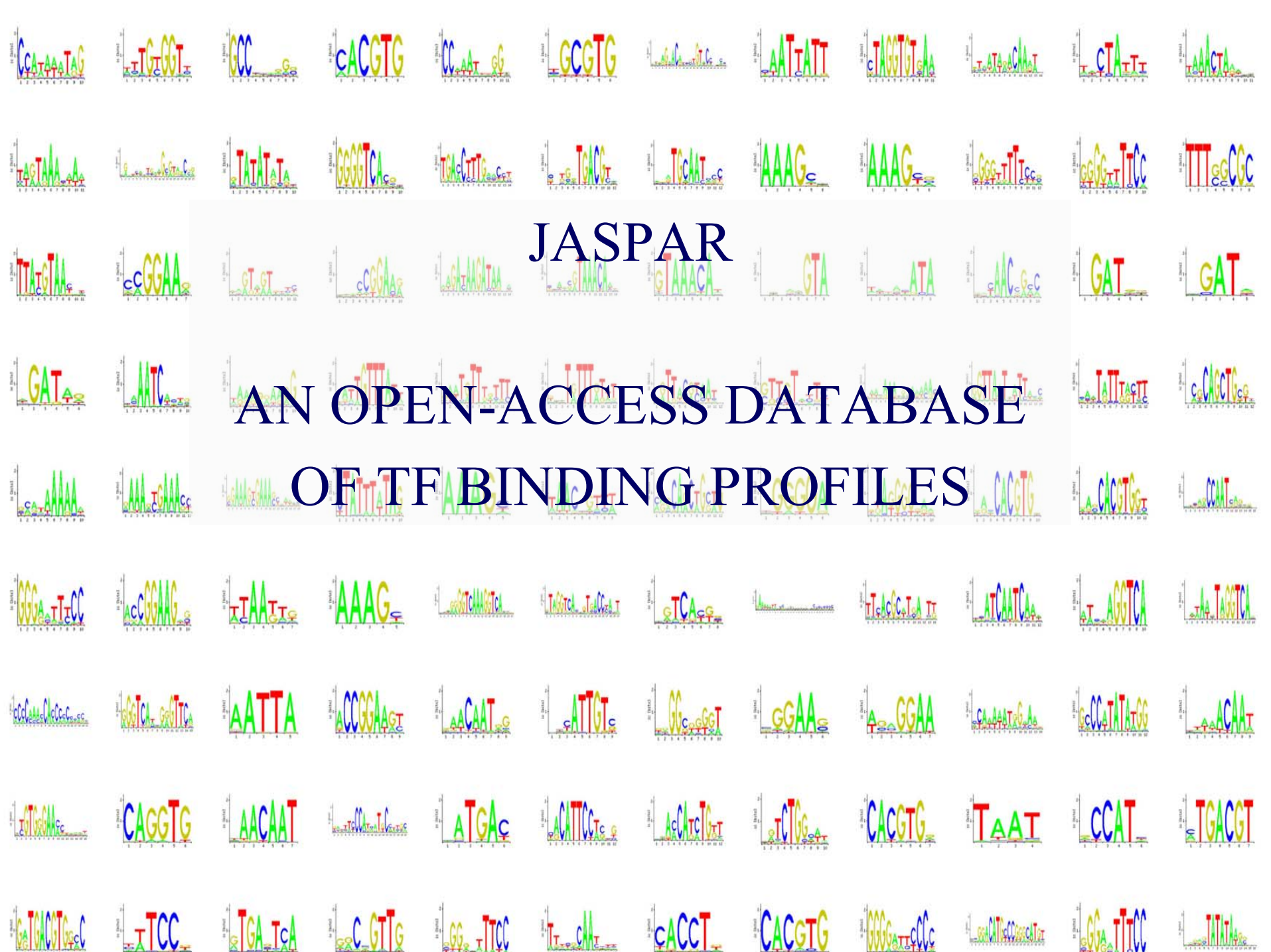
w matrix

A	1.6	-1.7	-0.2	-1.7	-1.7
C	-1.7	0.5	0.5	1.3	-1.7
G	-1.7	1.0	-0.2	-1.7	1.3
T	-1.7	-1.7	-0.2	-0.2	-0.2

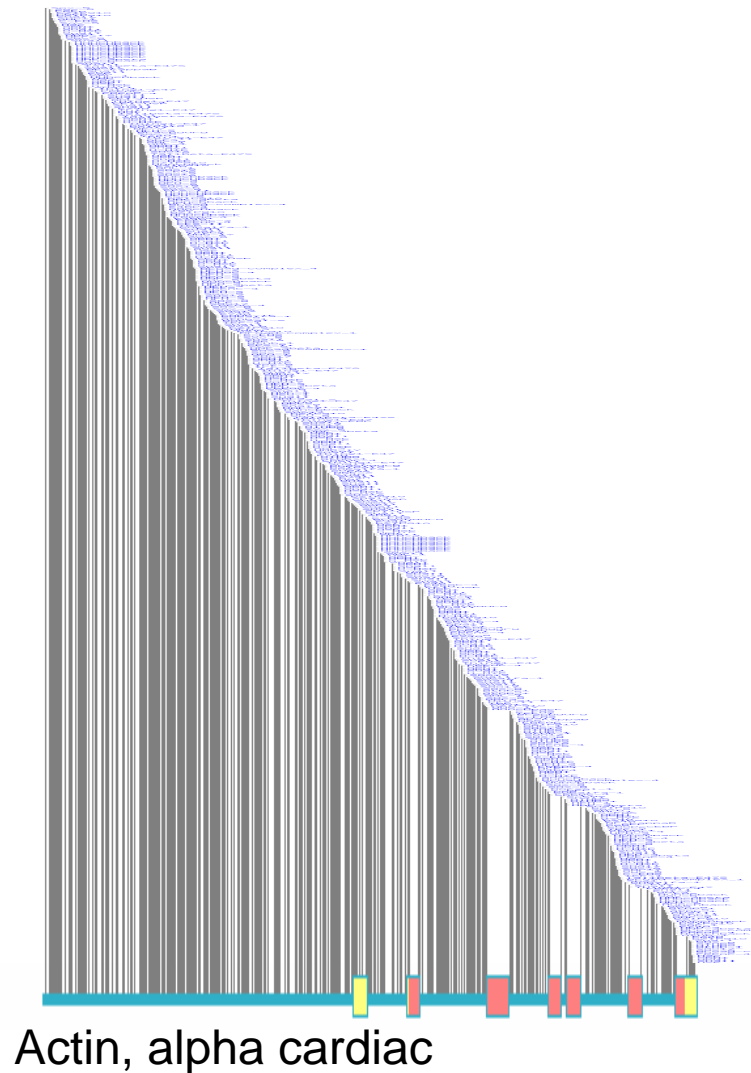
TGCTG = 0.9

Performance of Profiles

- 95% of predicted sites bound *in vitro* (Tronche 1997)
- MyoD binding sites predicted about once every 600 bp (Fickett 1995)
- The Futility Conjecture
 - Nearly 100% of predicted transcription factor binding sites have no function *in vivo*



PROBLEM: Too many spurious predictions



Terms



- **Specificity** – The portion of predictions that are correct
- **Sensitivity** – The portion of “positives” that are detected
- The detection of TFBS is limited by terrible specificity. **Why?**

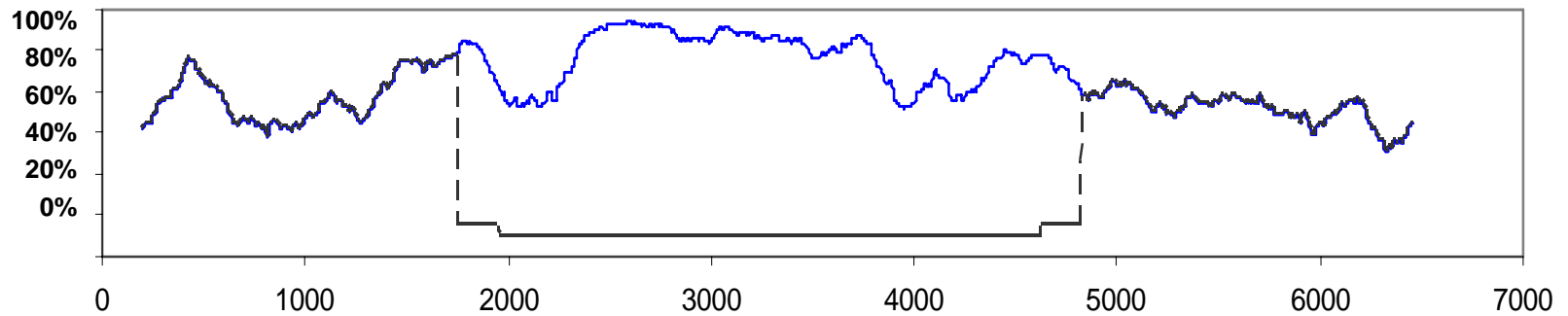
Method#1

Phylogenetic Footprinting

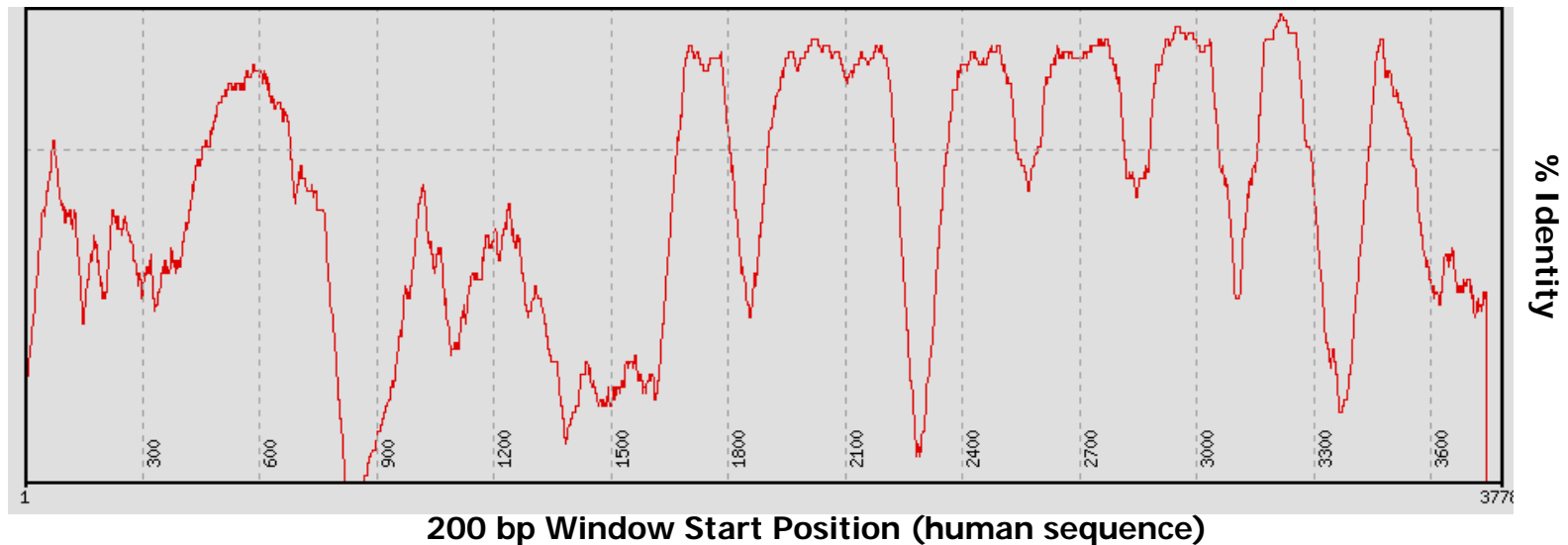
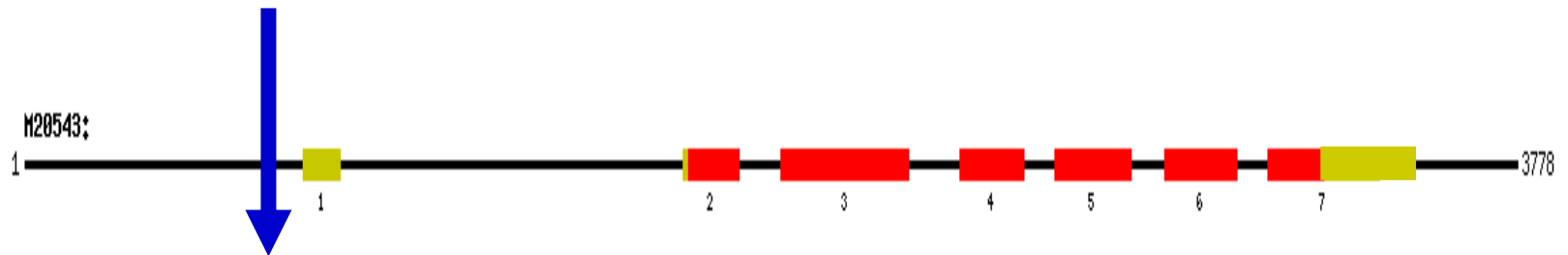
70,000,000 years of
evolution reveals most
regulatory regions

Phylogenetic Footprinting

FoxC2

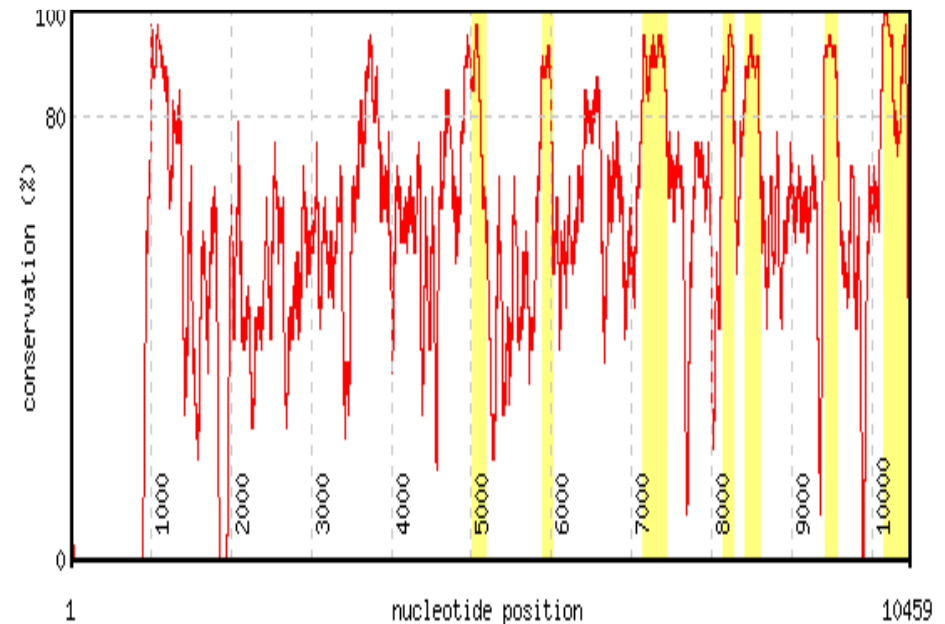
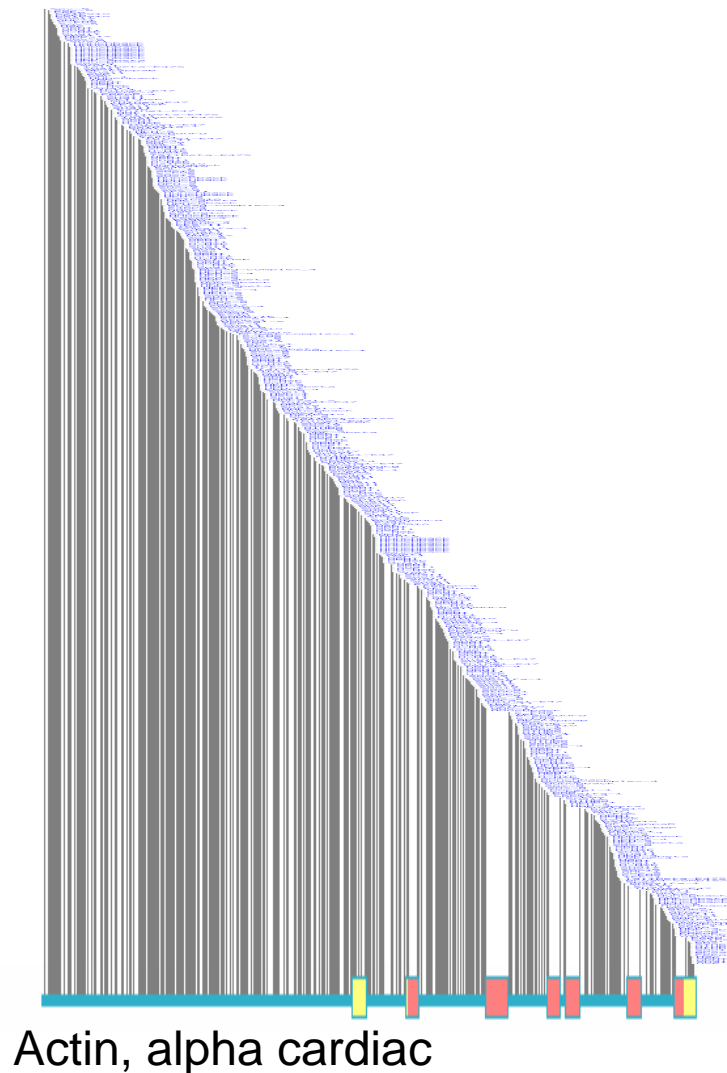


Phylogenetic Footprinting to Identify Functional Segments

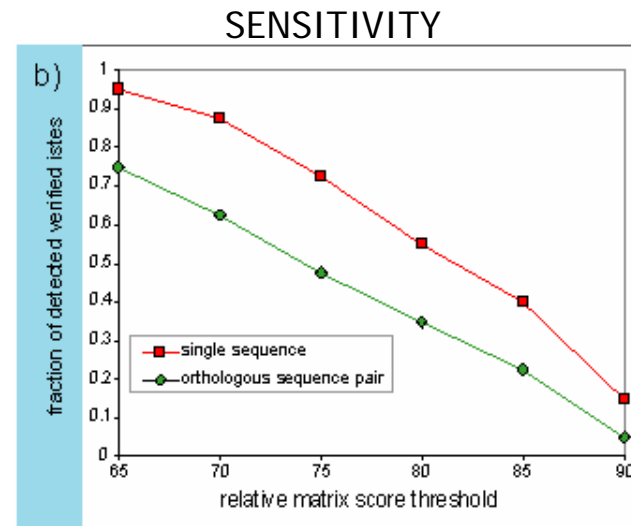
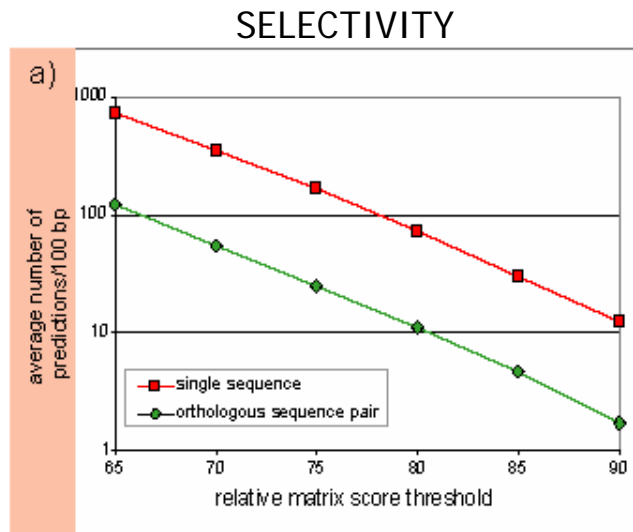


Actin gene compared between human and mouse with DPB.

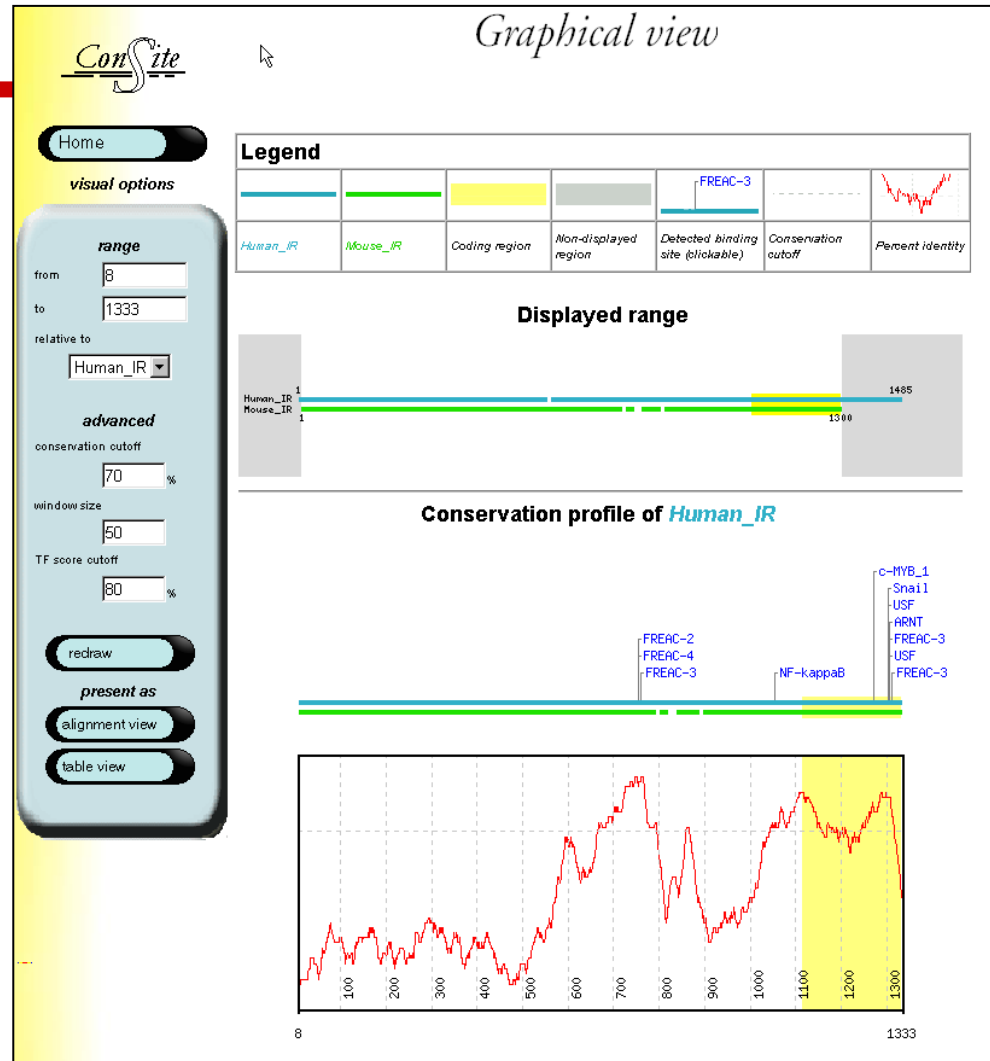
Phylogenetic Footprinting Dramatically Reduces Spurious Hits



Performance: Human vs. Mouse



- Testing set: 40 experimentally defined sites in 15 well studied genes (Replicated with 100+ site set)
- 75-90% of defined sites detected with conservation filter, while only 11-16% of total predictions retained



NEW: Ortholog Sequence Retrieval Service

Emerging Issues

- Multiple sequence comparisons
 - Incorporate phylogenetic trees
 - Visualization
- Analysis of closely related species
 - Phylogenetic shadowing
- Genome rearrangements
 - Inversion compatible alignment algorithm
- Higher order models of TFBS

OnLine Resources for Phylogenetic Footprinting



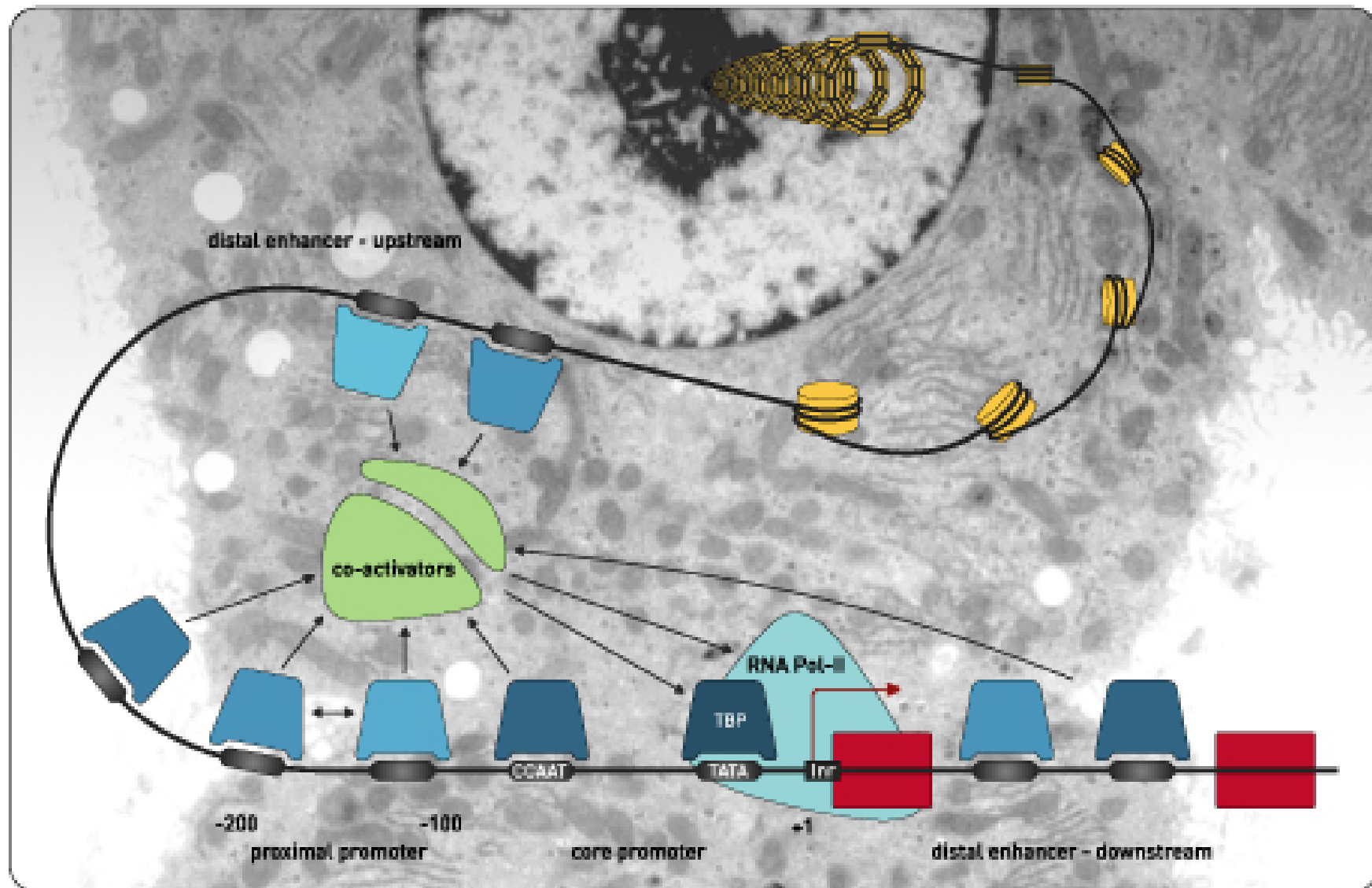
- Linked to TFBS
 - ConSite
 - rVISTA
- Alignments
 - Blastz
 - Lagan
 - Avid
 - ORCA
- Visualization
 - Sockeye
 - Vista Browser
 - PipMaker

Method#2

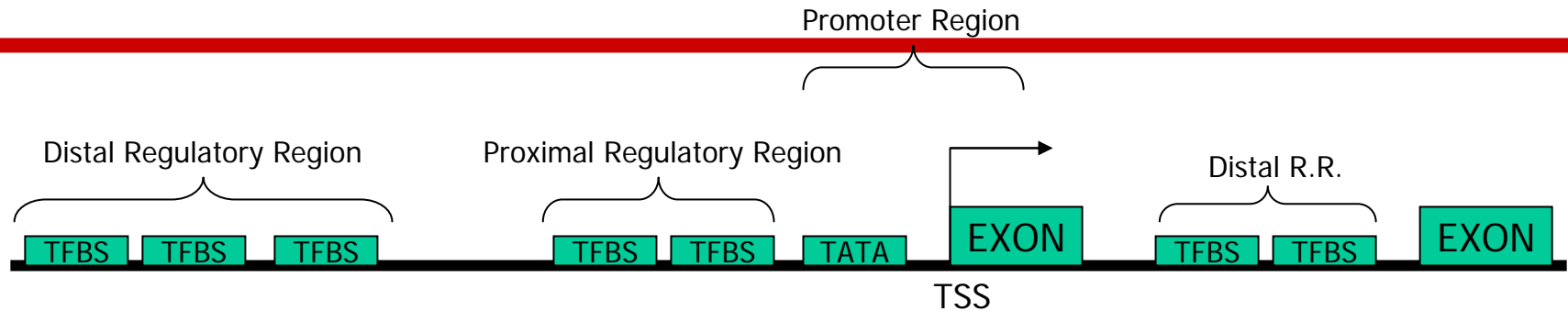
Discrimination of Regulatory Modules

TFs do NOT act in isolation

Layers of Complexity in Metazoan Transcription



Diverse and non-uniform use of terms: Partial glossary for tutorial

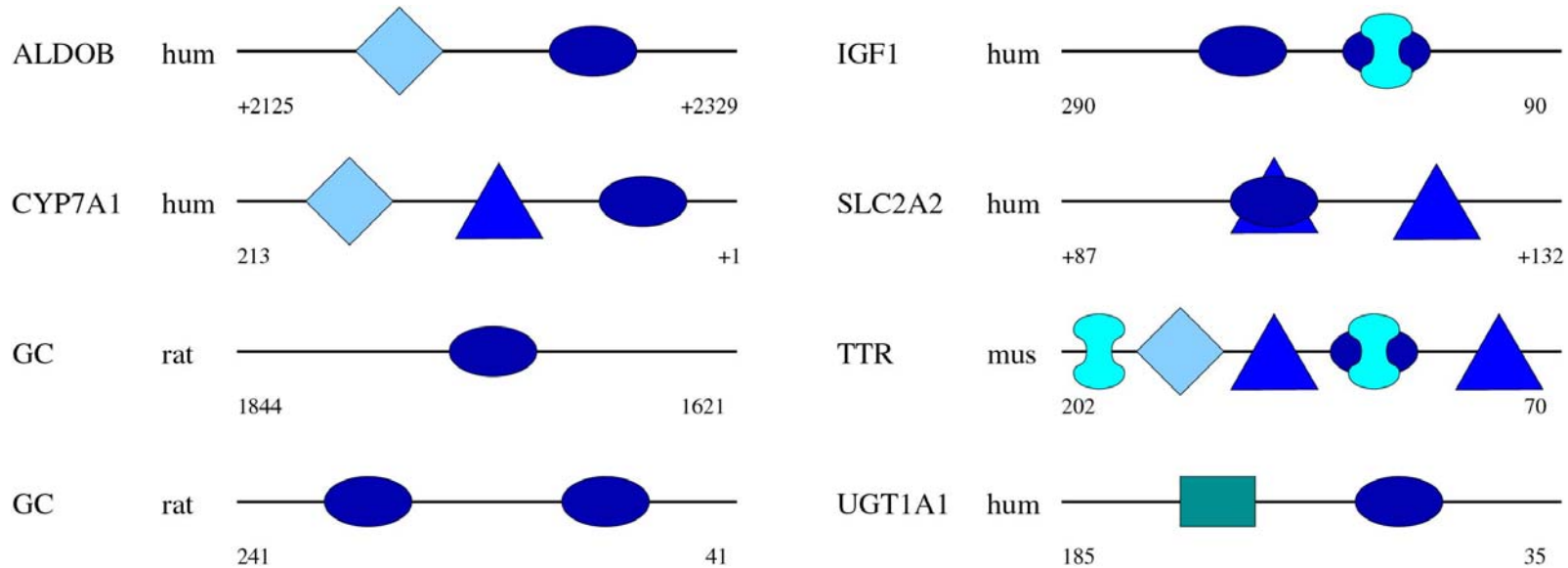


- Promoter – Sufficient to support the initiation of transcription; orientation dependent; includes TSS
- Regulatory Regions
 - Proximal – adjacent to promoter
 - Distal – some distance away from promoter (vague)
 - May be positive (enhancing) or negative (repressing)
- TSS – transcription start site
- TFBS – single transcription factor binding site
- Modules – Sets of TFBS that function together

Detecting Clusters of TF Binding Sites

- 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

Training for the detection of liver *cis*-regulatory modules (CRMs)



HNF1



HNF3



HNF4



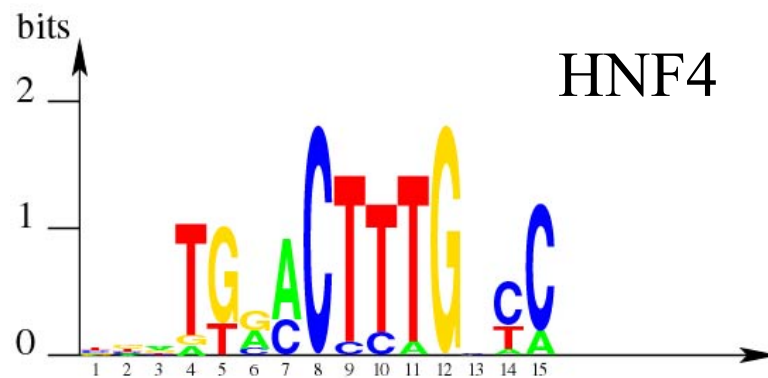
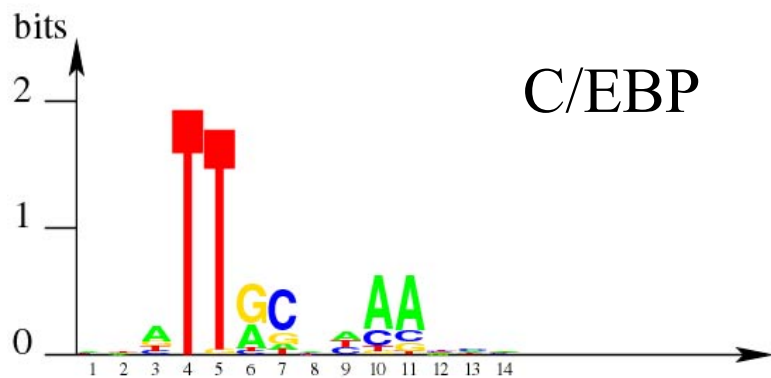
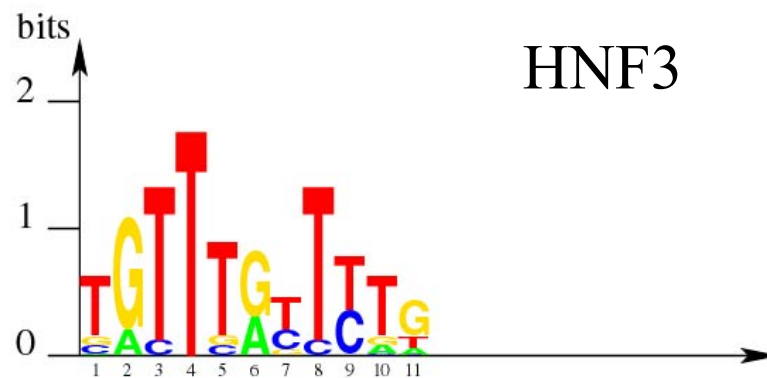
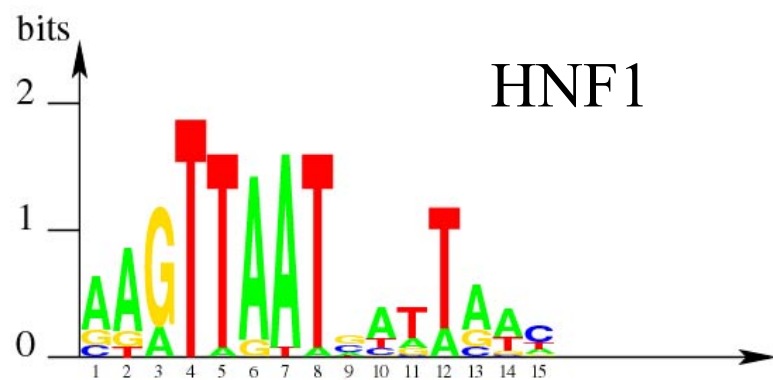
C/EBP



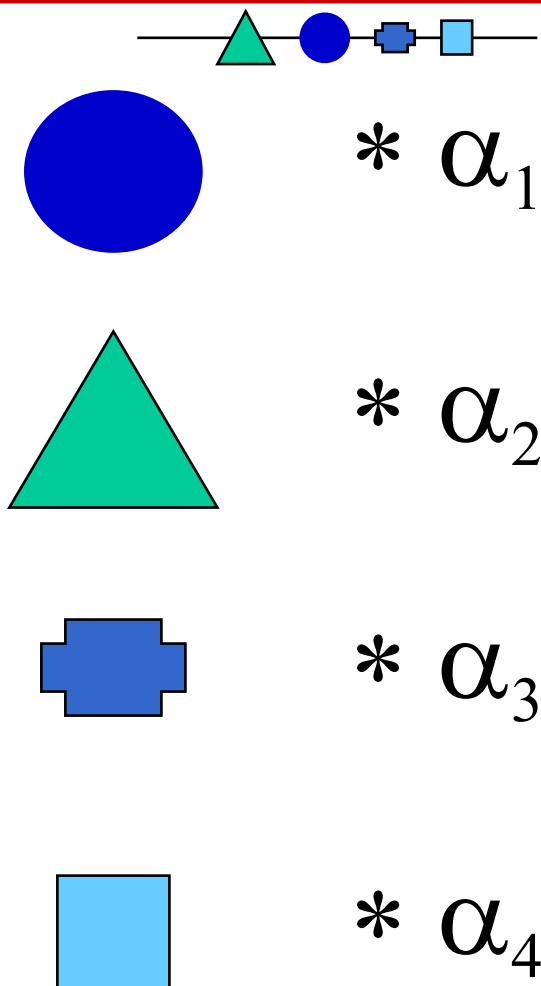
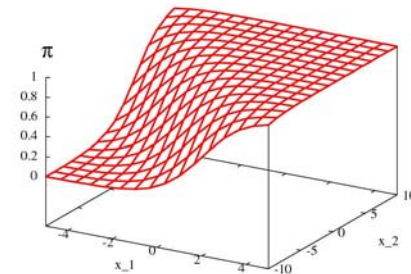
Sp1



Models for Liver TFs...



Logistic Regression Analysis



Optimize α vector to maximize the distance between output values for positive and negative training data.

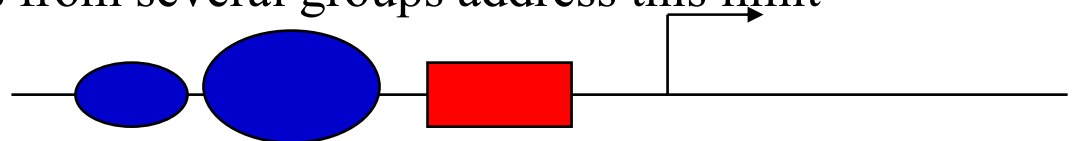
$$\Sigma \longleftarrow \text{“logit”}$$

Output value is:

$$p(x) = \frac{e^{\text{logit}}}{1 + e^{\text{logit}}}$$

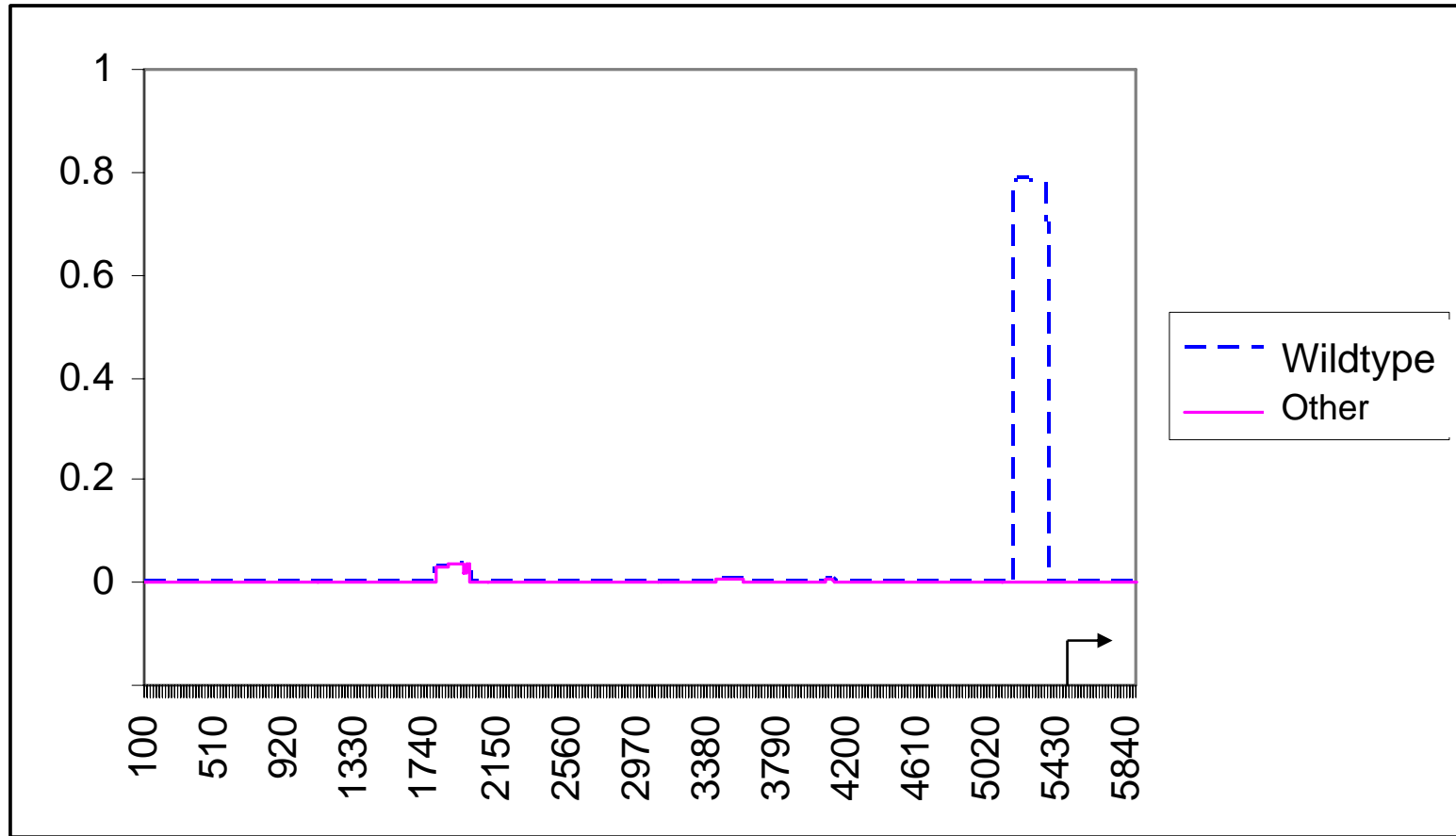
Performance of the Liver Model

- Performance
 - Sensitivity: 60% of known CRMs detected
 - Specificity: 1 prediction/35,000bp
- Limitations
 - Applies to genes expressed late in hepatocyte differentiation
 - Requires 10-15 genes in positive training set
 - This model doesn't account for multiple sites for the same TF
 - New methods from several groups address this limit



UGT1A1

Liver Module Model Score



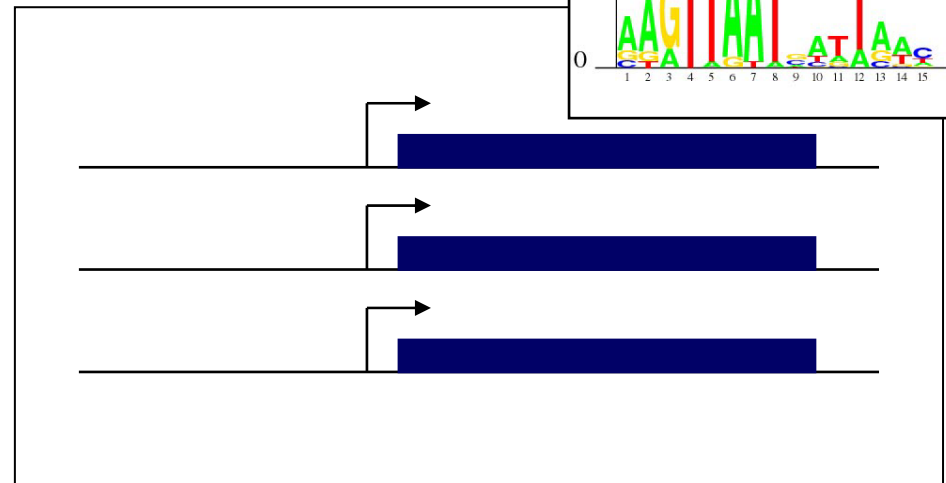
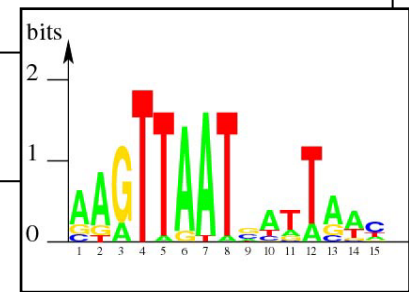
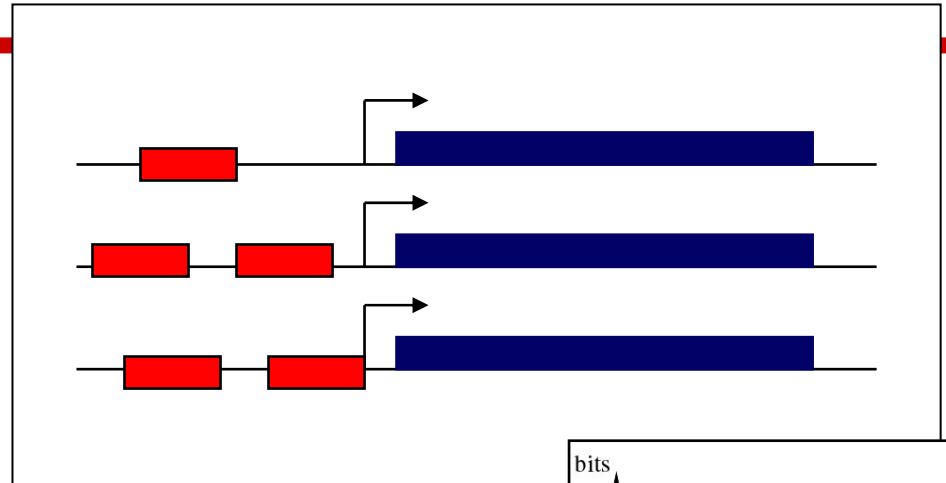
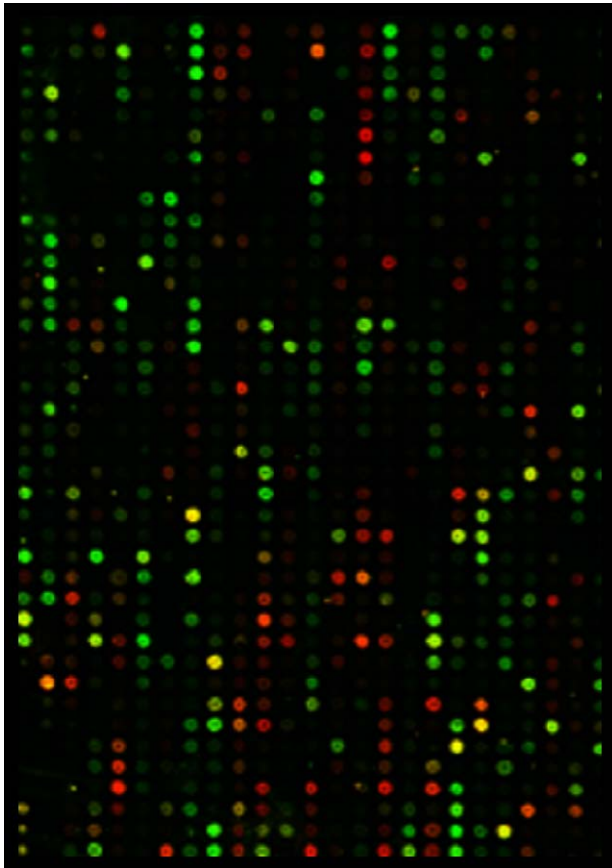
“Window” Position in Sequence

Making better predictions

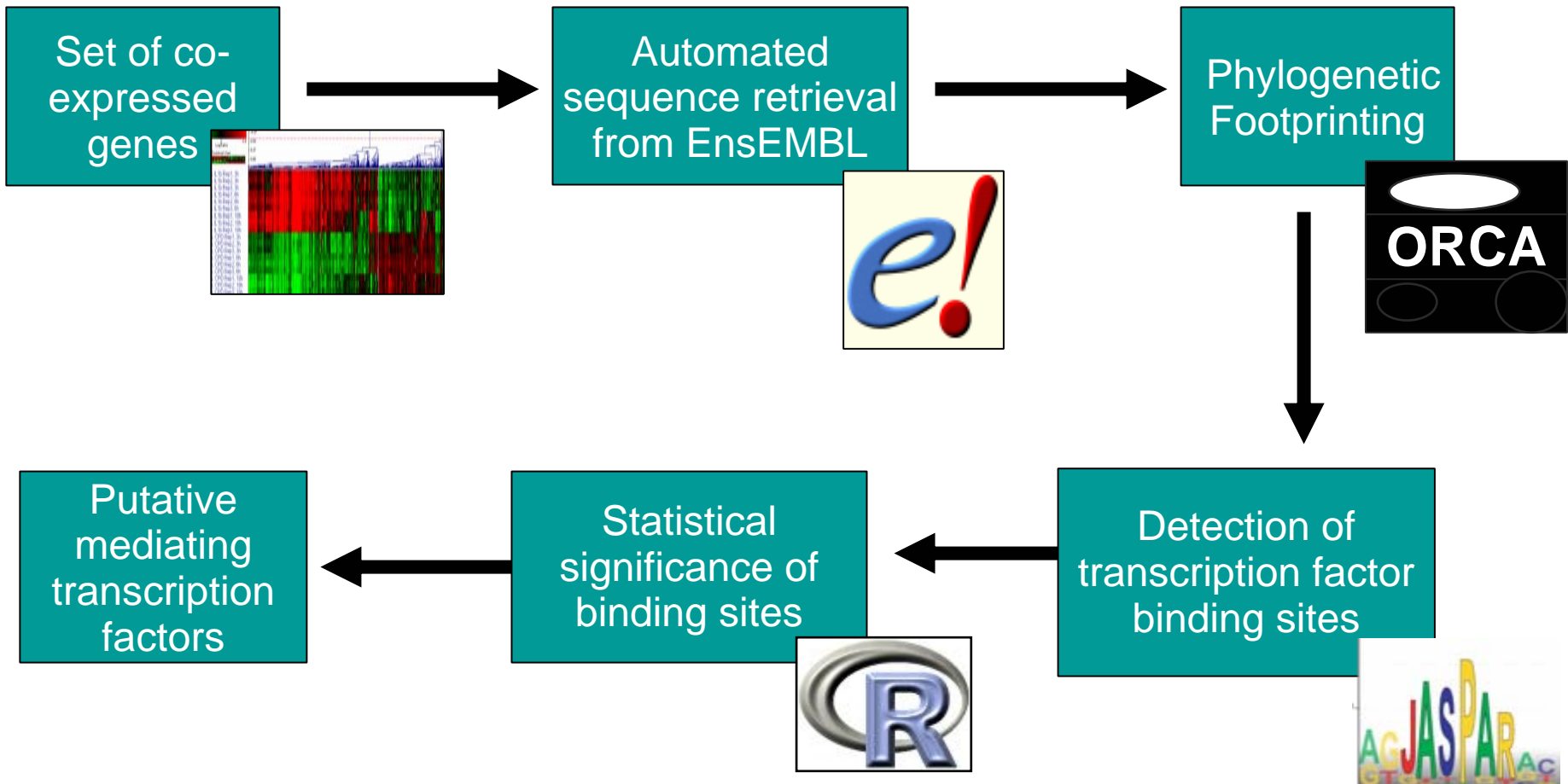
- Profiles make far too many false predictions to have predictive value in isolation
- Phylogenetic footprinting eliminates ~90% of false predictions
- Algorithms for detection of clusters of binding sites perform better, especially when possible to create train on known examples for the target context

Linking co-expressed genes to candidate transcription factors

Deciphering Regulation of Co-Expressed Genes



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%

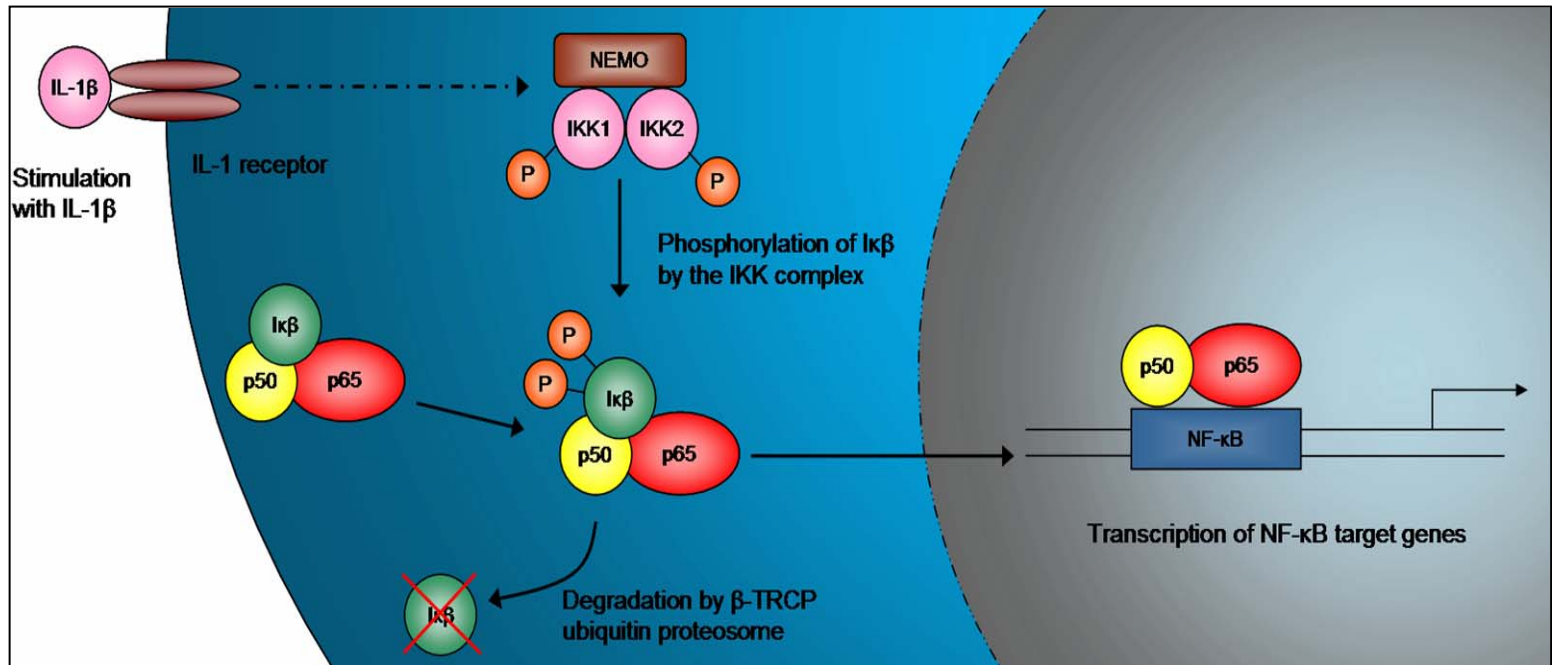
Validation using Reference Gene Sets

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.

Application to Microarray Data Sets

1. NF- κ B inhibition microarray study



Genes Significantly Down-regulated by the NF- κ B inhibitor (326 input; 179 analyzed)

		TF Class	Rank	Z-score	Fisher	No. Genes
p65		REL	1	36.57	5.66e-12	62
NF-kappaB		REL	2	32.58	5.82e-11	61
c-REL		REL	3	26.02	8.59e-08	63
Irf-2		TRP-CLUSTER	4	20.39	5.74e-04	6
SPI-B		ETS	5	16.59	1.23e-03	135
Irf-1		TRP-CLUSTER	6	15.4	9.55e-04	23
Sox-5		HMG	7	15.38	2.56e-02	126
p50		REL	8	14.72	2.23e-03	19
Nkx		HOMEO	9	13.66	2.29e-03	111
Bsap		PAIRED	10	13.2	9.92e-02	1
FREAC-4		FORKHEAD	11	12.05	1.66e-03	92
n-MYC		bHLH-ZIP	25	6.695	1.84e-03	102
ARNT		bHLH	26	6.695	1.84e-03	102
HNF-3beta		FORKHEAD	29	5.948	3.32e-03	47
SOX17		HMG	31	5.406	8.60e-03	79

oPOSSUM Server


oPOSSUM: Select Analysis Parameters - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Refresh Print Mail Stop 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-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

REVIEWING THE TOP POINTS

Orientation

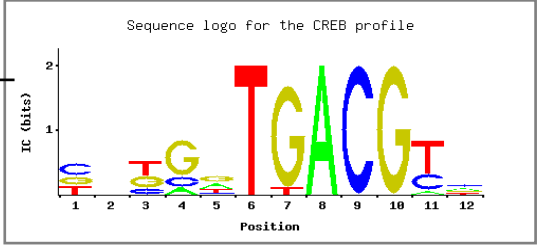
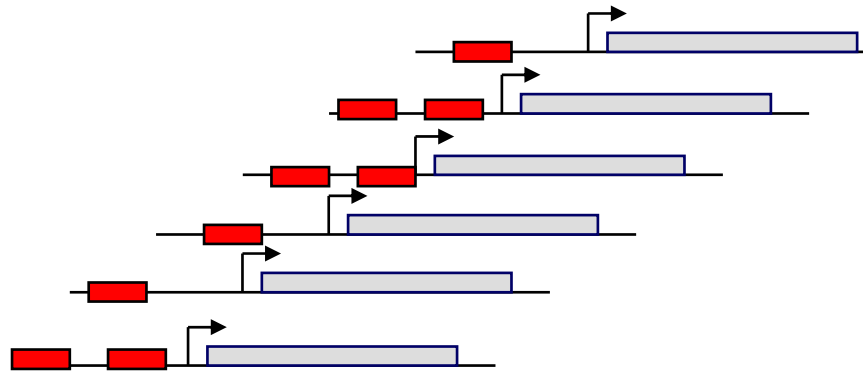
Regulatory regions problem space

Sets of binding sites

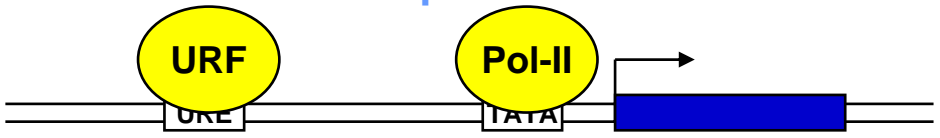
AATCACCA
 AATCACCA
 AATCACCA
 AATCACCA
 AATCTCCC
 AATCTCCG
 AATCACAC
 AATCATCA
 AATCTCAC
 AATCTCTG
 AGTCCCCA
 AATCCCGG
 AATCTGAG
 AATCCATA
 ATTCAGCC
 AATAACTT
 GATAACCT
 AATTAGAC
 GATTACAG
 GATTAGCG
 ATTCTTCC
 TATGAACA
 GATTAAAA
 AGACCCCA

Specificity profiles for binding sites

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

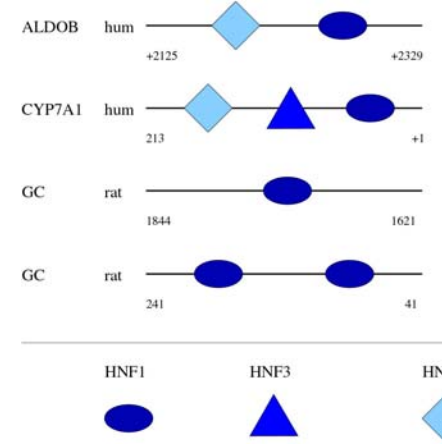


Transcription factors



Transcription factor binding sites
Regulatory nucleotide sequences

Clusters of binding sites

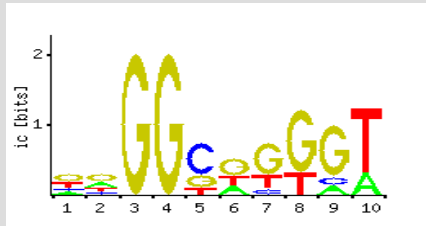


Analysis of regulatory regions with TFBS

Detecting binding sites in a single sequence

Scanning a sequence against a PWM

Sp1



ACCTTCCCCAGGGGCGGGGGGCGGTGGCCAGGACGGTAGCTCC

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)

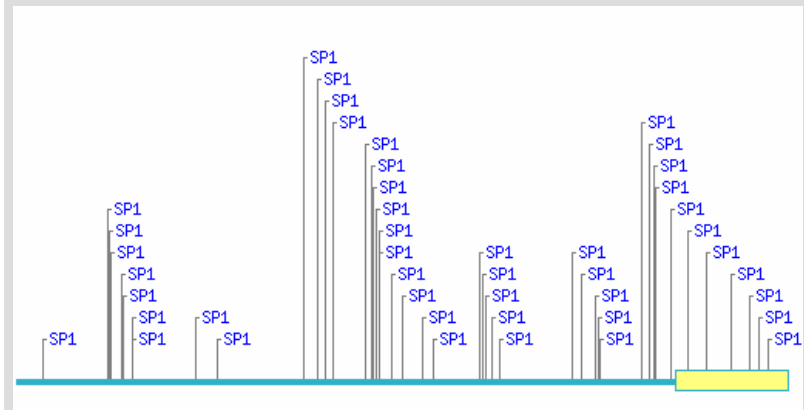
A	[-0.2284	0.4368	-1.5	-1.5	-1.5	0.4368	-1.5	-1.5	-0.2284	0.4368]
C	[-0.2284	-0.2284	-1.5	-1.5	1.5128	-1.5	-0.2284	-1.5	-0.2284	-1.5]
G	[1.2348	1.2348	2.1222	2.1222	0.4368	1.2348	1.5128	1.7457	1.7457	-1.5]
T	[0.4368	-0.2284	-1.5	-1.5	-0.2284	0.4368	0.4368	0.4368	-1.5	1.7457]

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%

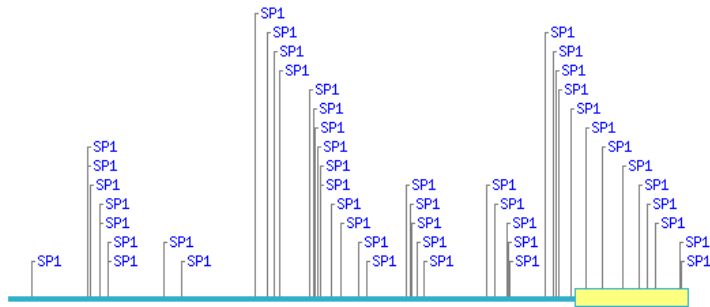


Ouch.

Analysis of regulatory regions with TFBS

Phylogenetic Footprints

Scanning a single sequence

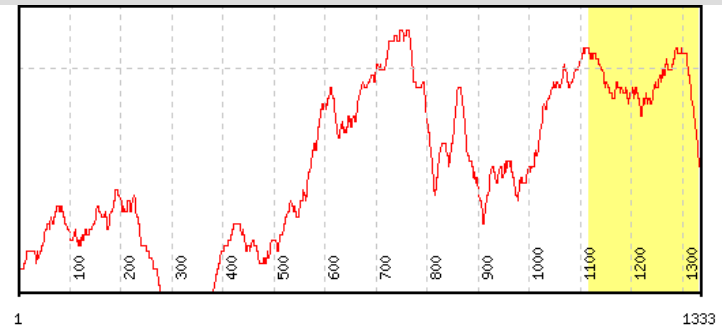


Low specificity of profiles:

- too many hits
- great majority not biologically significant

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

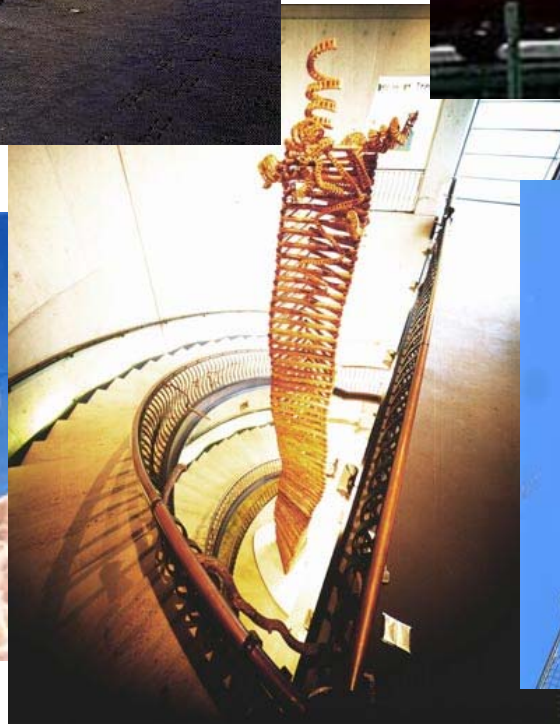
A dramatic improvement in the percentage of biologically significant detections



Congratulations on Your Completion of CBW Bioinformatics

How does one find new topics for
bioinformatics research?

DNA



The Study of the Absurd

Advances in Biology and
Bioinformatics are driven by the
investigation of the unusual

Deinococcus radiodurans

"strange berry that withstands radiation"

"World's Toughest Bacterium" – Guinness Book of World Records

- Survives DNA damaging conditions
- 4-10 copies of genome
 - Stacked with same sequences adjoining
- When damaged, single strand annealing brings copies together and homologous recombination reconstructs the full DNA sequence
- Bag full of protective enzymes
 - Protection against DNA damaging agents



<http://www.microbe.org/art/Deinococcus.jpg>

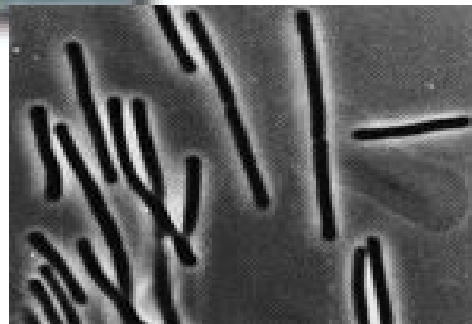
Thermus aquaticus

“Loves Hot Water”



<http://www.windowstowonderland.org>

- Thomas Brock sought organisms that could survive at high temperatures
- Identified *T.aquaticus* in geysers at Yellowstone Park
- Replicates at 100C
- Source of heat-stable enzymes for PCR and industrial processes



<http://webs.wichita.edu/mschneegurt/biol103/lecture05/21Taquaticus.jpg>

Nanoarchaeum equitans

(hyperthermophilic archaeal parasite)

- Recently discover Archaeal organism
- Missing genes for glutamate, histidine, tryptophan and initiator methionine transfer RNA
- Computational genome analysis revealed widely separated genes encoding tRNA halves
- RT-PCR demonstrated full-size tRNA



Cell of *Ignicoccus spec.* with four cells of *Nanoarchaeum equitans* attached. Electron micrograph by H. Huber et al . <http://www.genomenewsnetwork.org>

Ciliate Gene Reconstruction

(*Tetrahymena thermophila*)

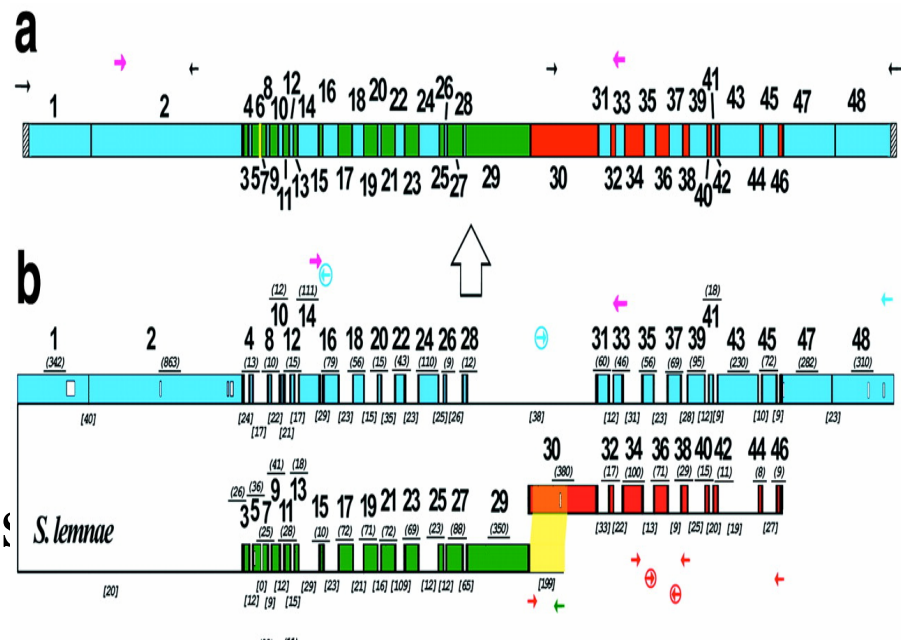
- Rearranges genome, excising extra DNA from somatic nucleus and placing the fragments into an auxiliary nucleus
- Sidenote: *Tetrahymena* was the original source for the discovery of catalytic RNA (Ribozymes)



<http://www.biology.wustl.edu/faculty/images/chalkercaption.jpg>

Building from pieces

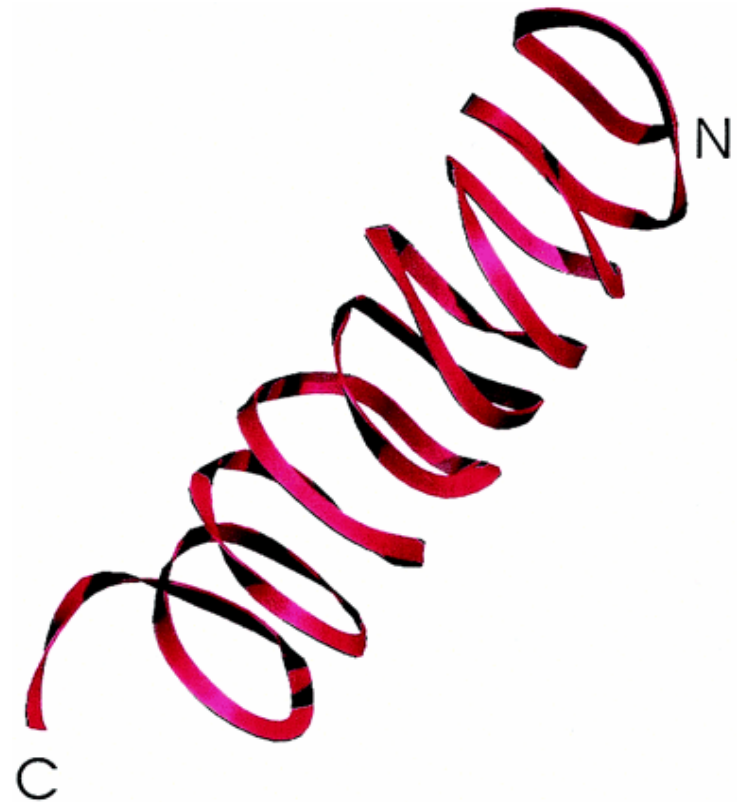
- *Stylonychia lemnae* pol-a gene is fragmented into 48 fragments
- Gene is reassembled from the pieces by complementary hybridization of edges of the fragments
 - Pol α rebuilt from 48 pieces



Pseudomonas syringae

(Knock-knock, can I come in?)

- Getting past plant cell walls/membranes is a goal for some microbes
- Placing a protein on the surface of the membrane that catalyzes ice formation, results in a hole at which the bacteria can gain access to a good meal...
 - Ice nucleation protein
- Protein analysis reveals a beautiful helical structure

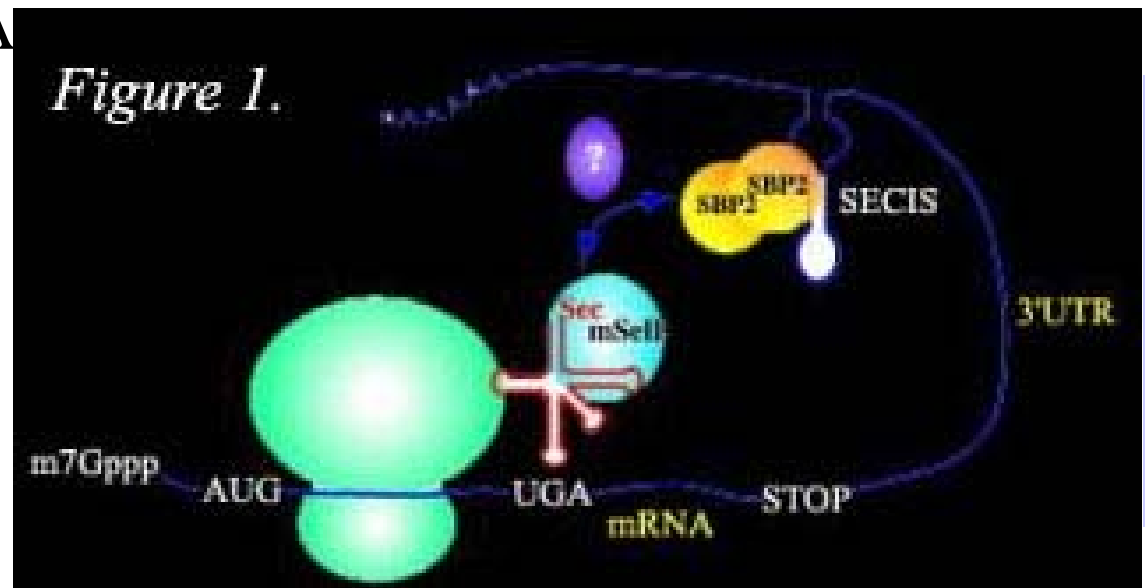


Unusual Transcription?

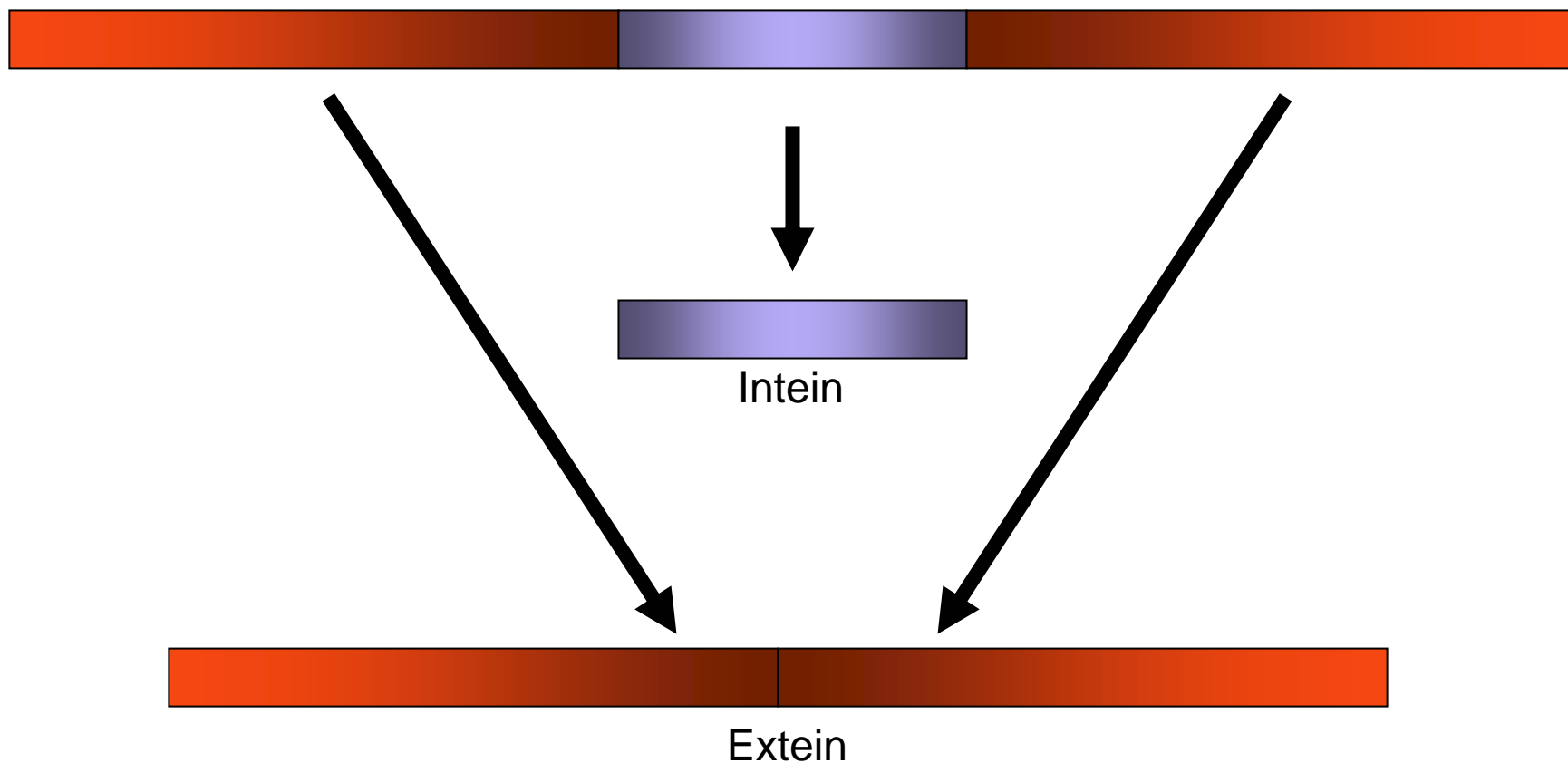
- Missing a tRNA
- Generated from the fusion of two distinct transcripts

Selenocysteine Insertion vs Translation Termination

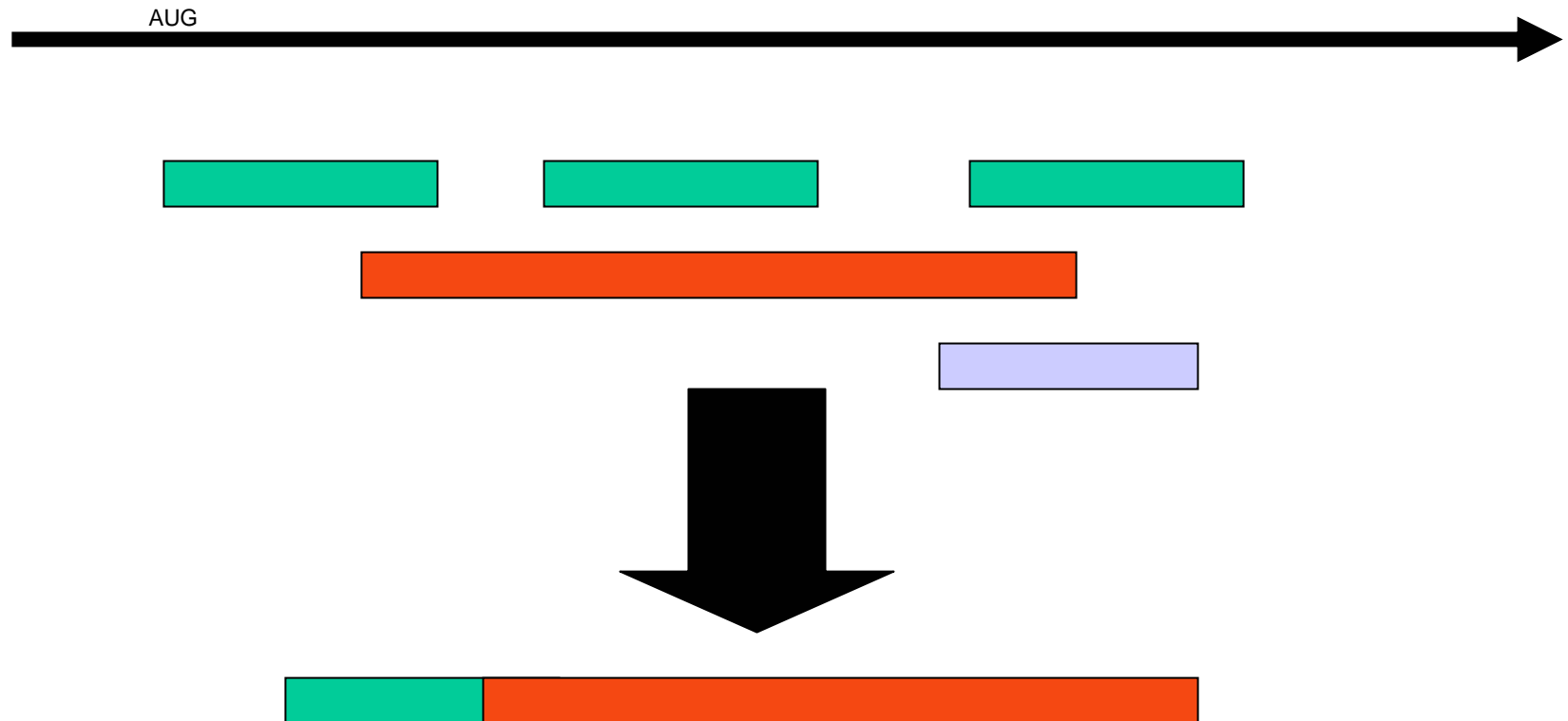
- Selenocysteine is an alternative amino acid that is inserted by a tRNA interacting with the codon UGA STOP codon!



Intein Protein Splicing



Translation Frameshifting



Thoughts

- New problems in bioinformatics are driven by unique datasets
- Incremental improvements in existing methods are valued
- Keep thinking about biological observations – how could computational approaches be based on the concepts?

Sources for the Weird and Unusual

- <http://www.genomenewsnetwork.org/>