



The Centre for Molecular Medicine and Therapeutics



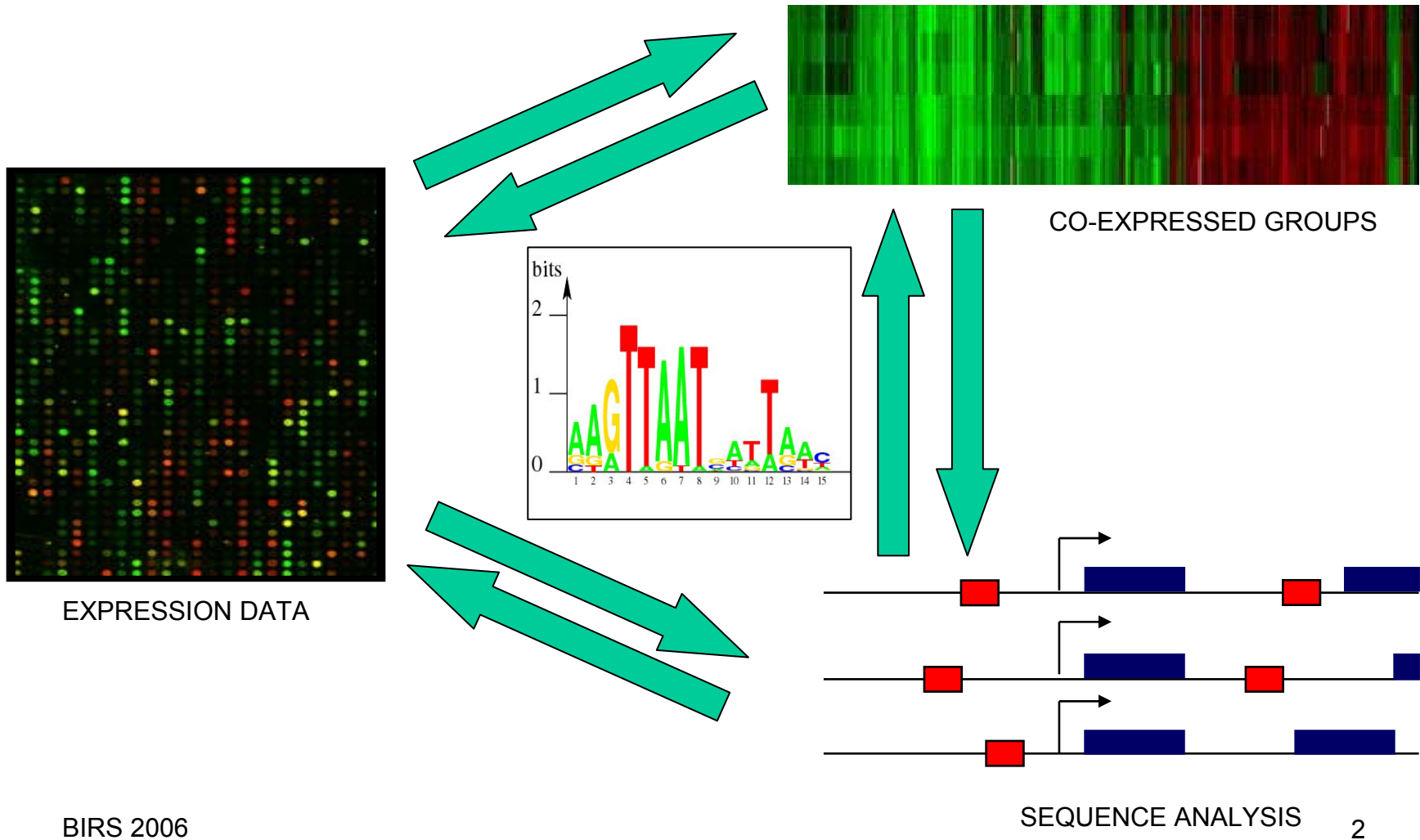
# Gene Regulation Bioinformatics

**Wyeth W. Wasserman**

University of British Columbia

[www.cisreg.ca](http://www.cisreg.ca)

# The Grand Challenge: Reliably Define Cis-Regulatory Mechanisms of Regulons



# REGULATORY PATHWAY INFERENCE from CO-EXPRESSED GENES

- What is the appeal?
  - Understand how perceived signals at surface result in downstream changes in cell phenotype
  - TFs occasionally serve as therapeutically relevant targets
    - PPAR $\gamma$ , Estrogen Receptor, Glucocorticoid Receptor
    -
  - Builds on data from powerful profiling technologies
    - Expression profiling; ChIP-chip

# Bioinformatics and Promoter Analysis

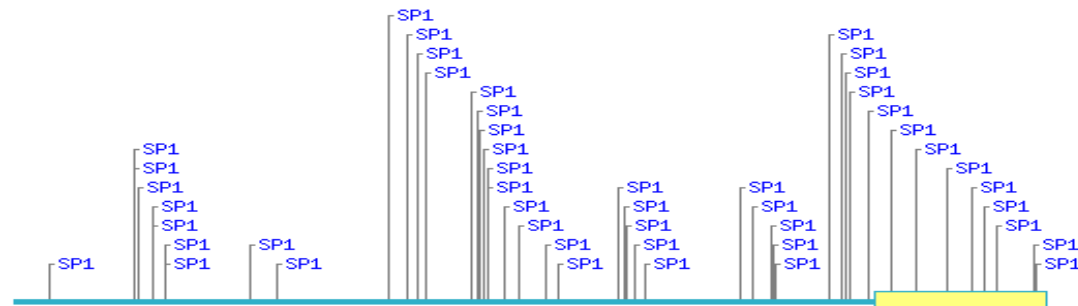
What can we do?

# Binding Profiles for a TF

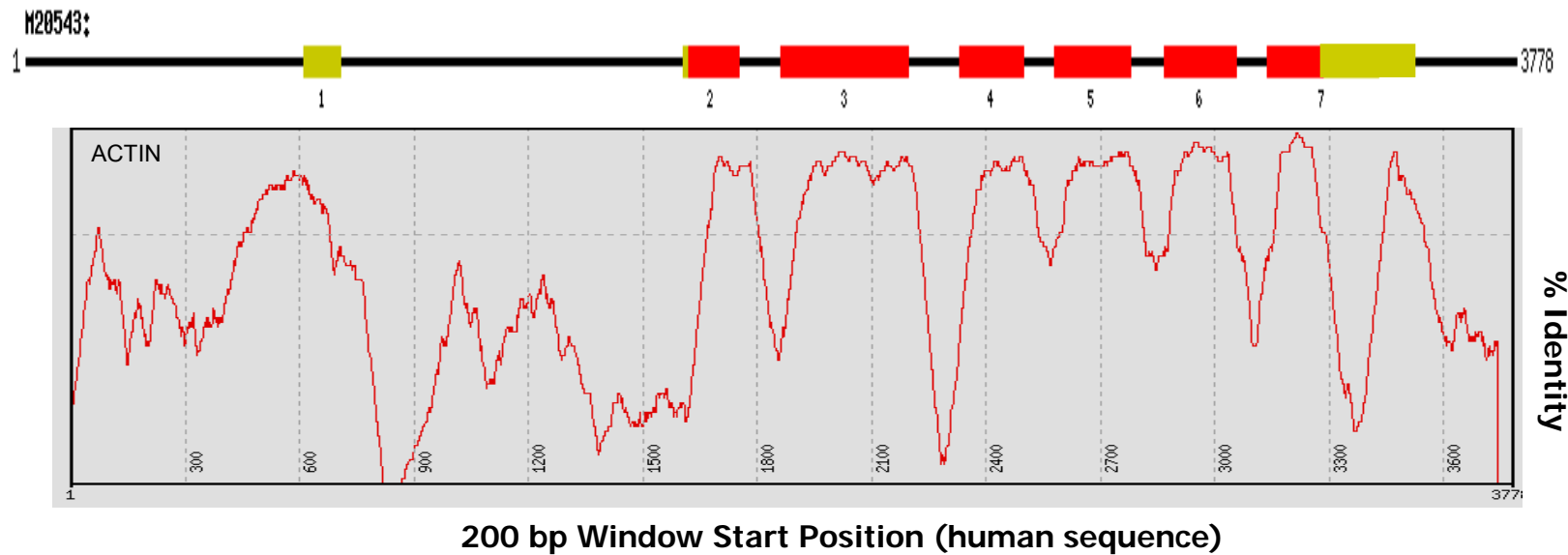
## Set of binding sites

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

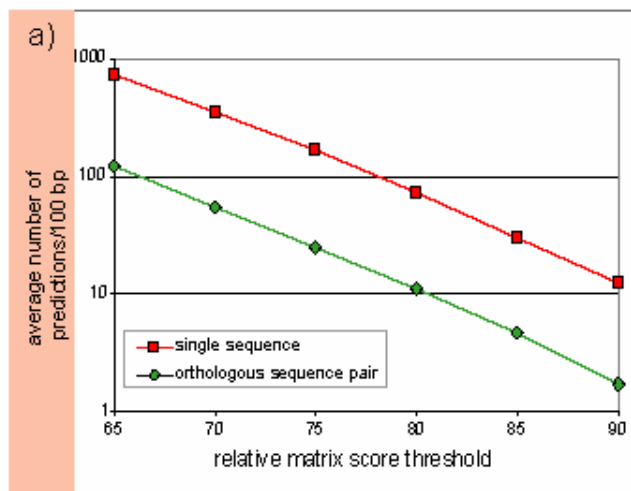
A	14	16	4	0	1	19	20	1	4	13	4	4	13	12	3
C	3	0	0	0	0	0	0	0	7	3	1	0	3	1	12
G	4	3	17	0	0	2	0	0	9	1	3	0	5	2	2
T	0	2	0	21	20	0	1	20	1	4	13	17	0	6	4



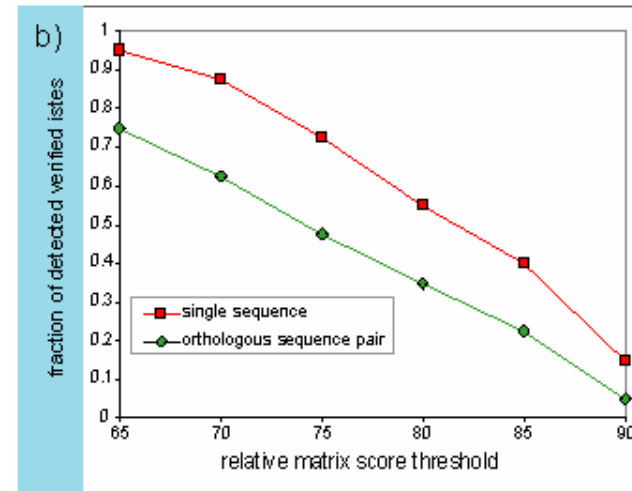
# Phylogenetic Footprinting



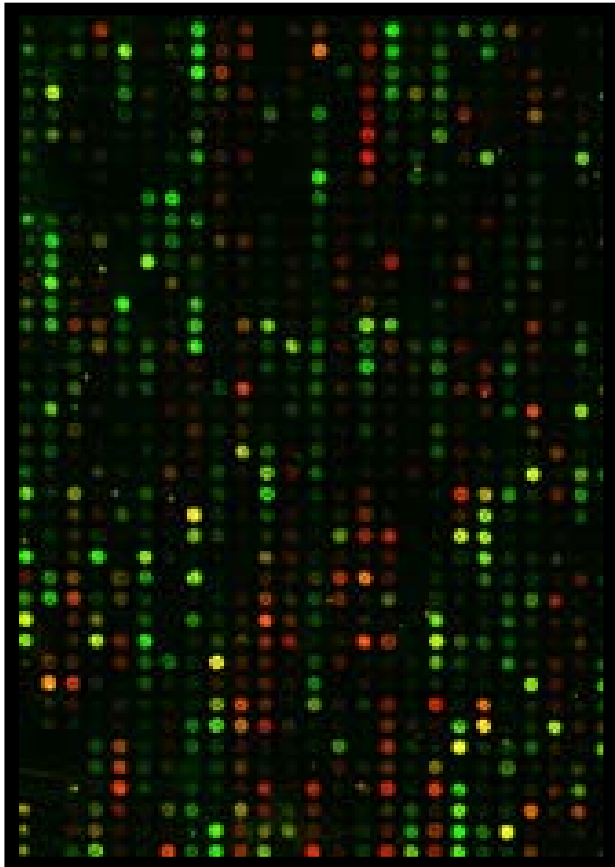
## SELECTIVITY



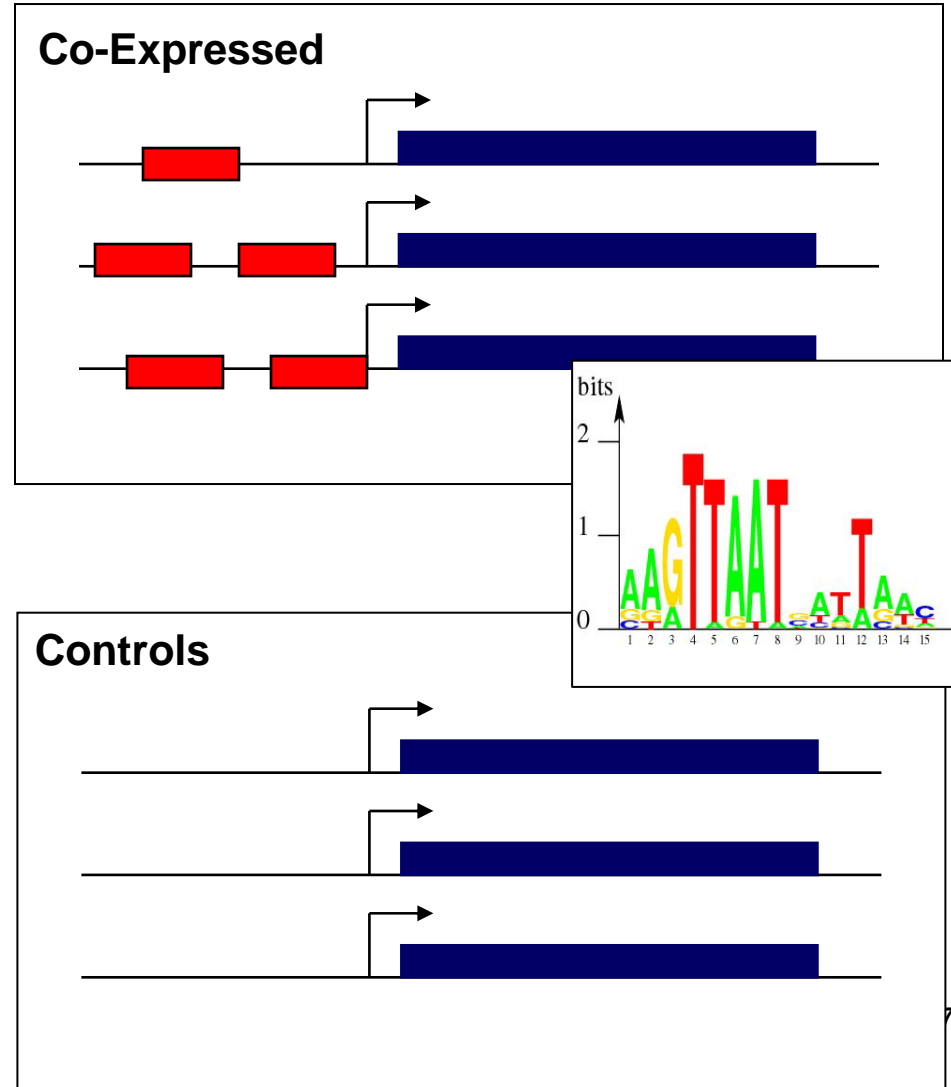
## SENSITIVITY



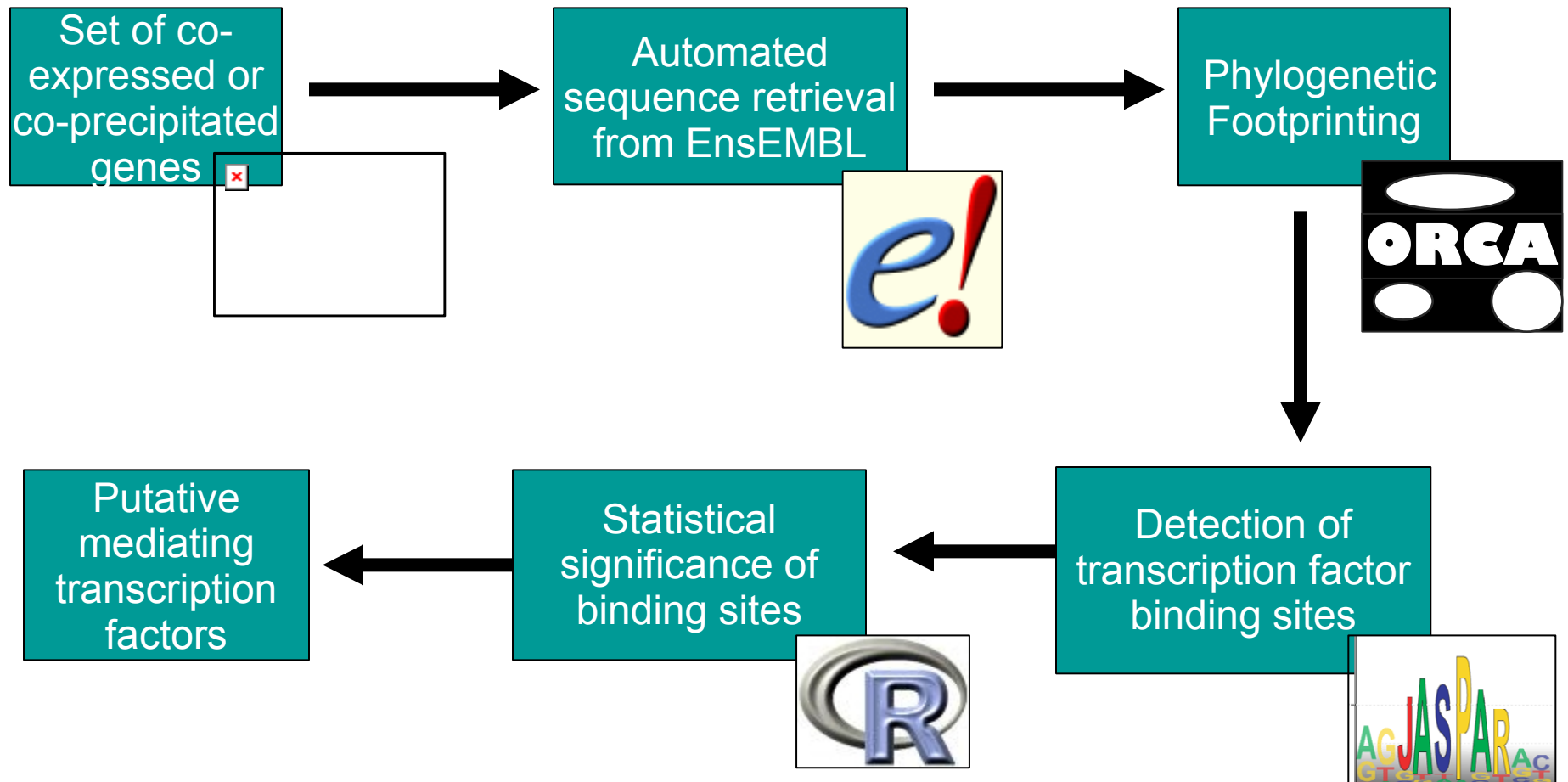
# Deciphering Regulation of Co-Expressed Genes



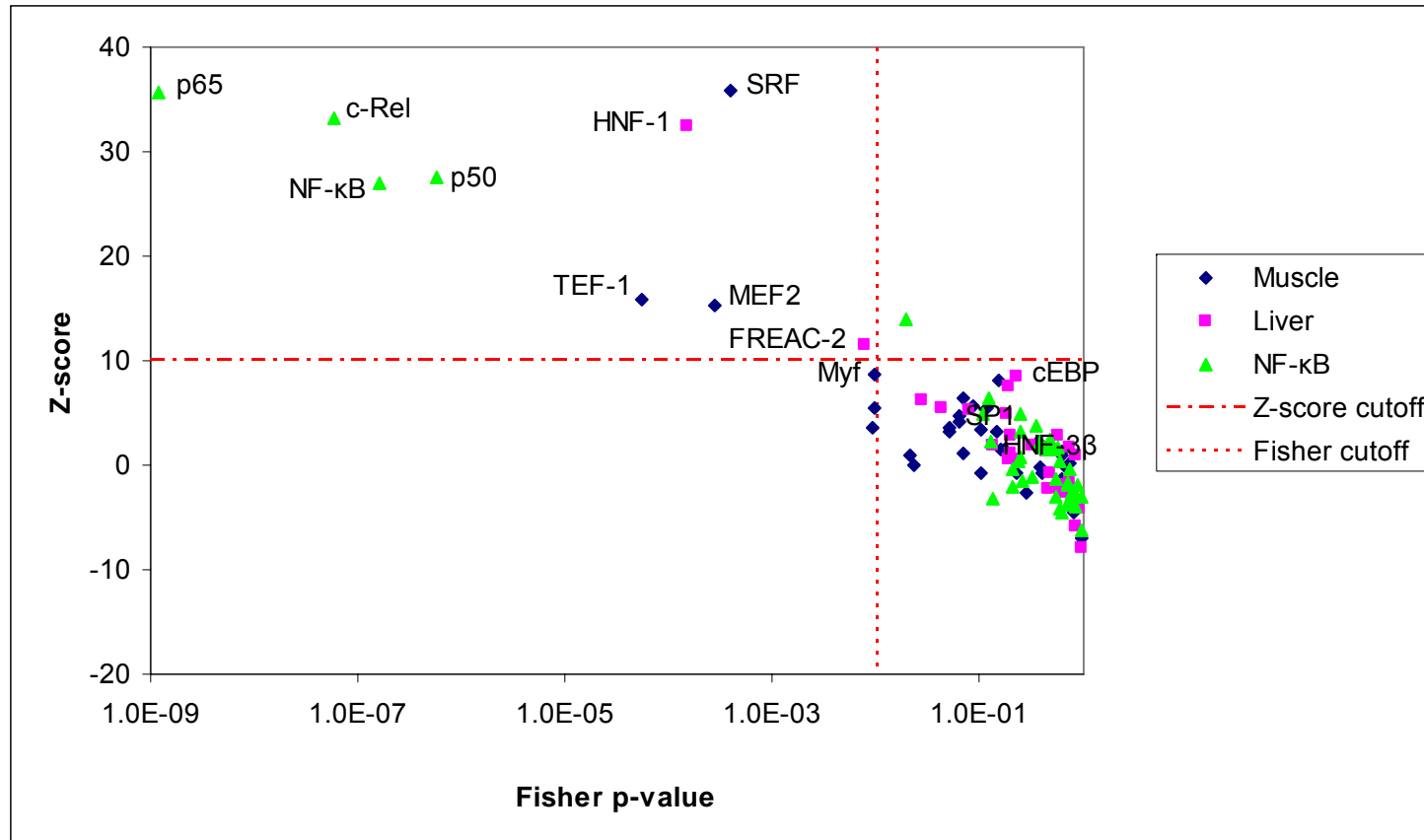
BIRS 2006



# oPOSSUM Procedure

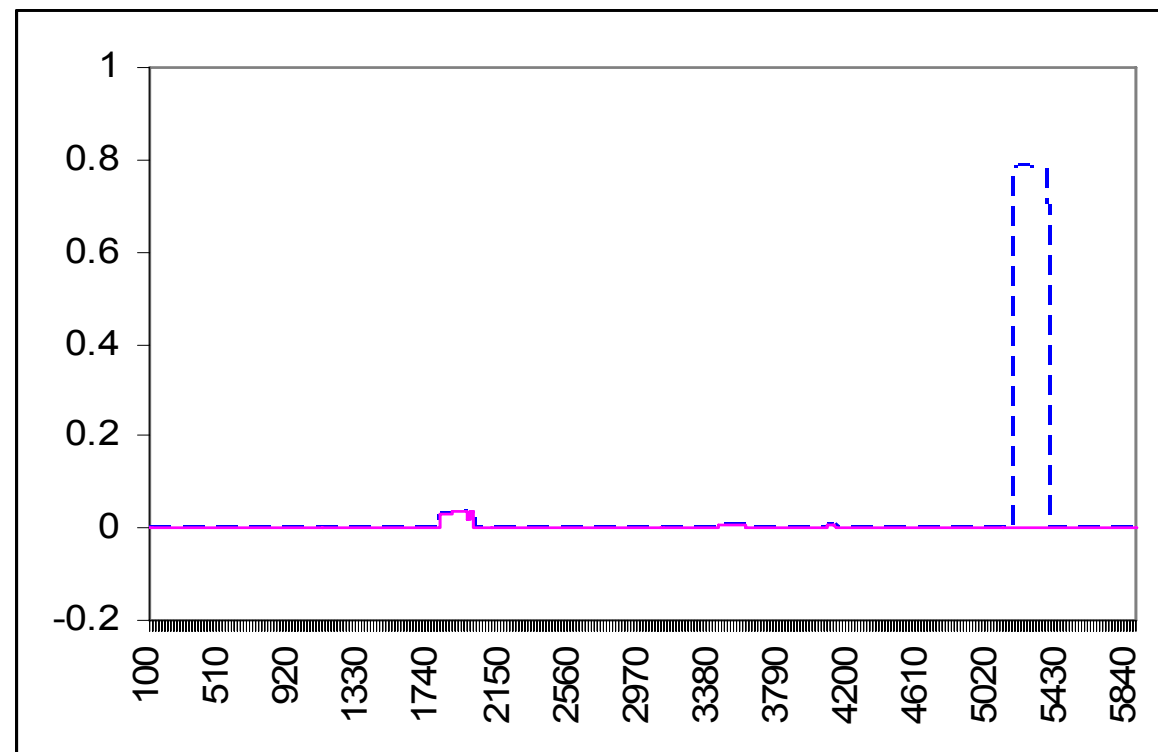
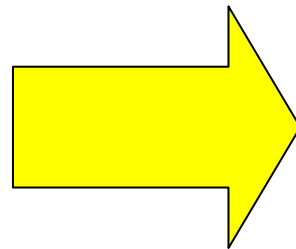
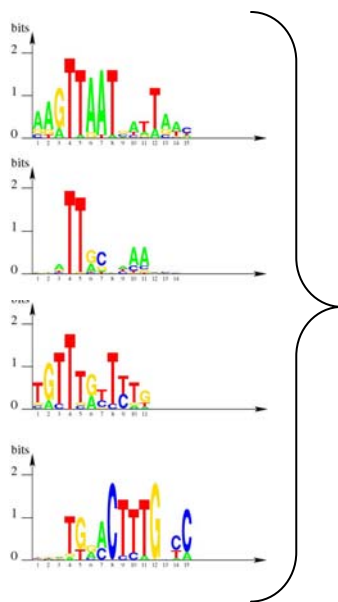


# Empirical Selection of Parameters based on Reference Studies

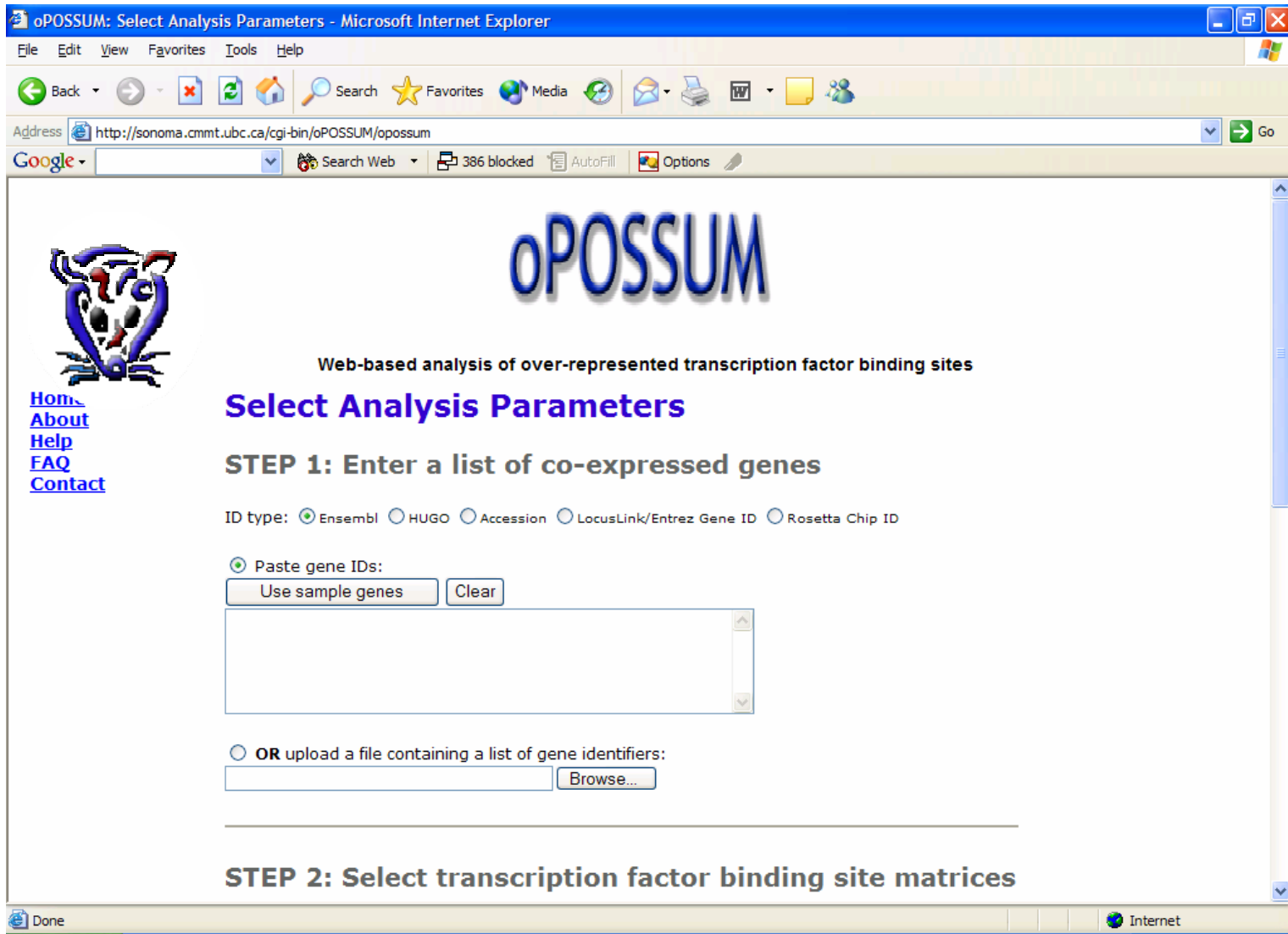


# CRM Models

Trained models take as input a set of TF binding profiles and return significant clusters of TFBS



# oPOSSUM Server



The screenshot shows a web browser window titled "oPOSSUM: Select Analysis Parameters - Microsoft Internet Explorer". The address bar shows the URL "http://sonoma.cmmt.ubc.ca/cgi-bin/oPOSSUM/opossum". The browser's toolbar includes navigation buttons (Back, Forward, Stop, Home), search, and other utility icons. The main content area features the oPOSSUM logo, a navigation menu with links for Home, About, Help, FAQ, and Contact, and a heading "Web-based analysis of over-represented transcription factor binding sites". The primary section is "Select Analysis Parameters", with "STEP 1: Enter a list of co-expressed genes" selected. Under this step, there are radio buttons for "ID type" (Ensembl, HUGO, Accession, LocusLink/Entrez Gene ID, Rosetta Chip ID) and "Paste gene IDs:" (selected). The "Paste gene IDs:" section includes a "Use sample genes" button, a "Clear" button, and a large text input field. Below this is an "OR upload a file containing a list of gene identifiers:" section with a "Browse..." button. "STEP 2: Select transcription factor binding site matrices" is partially visible at the bottom. The browser's status bar at the bottom shows "Done" and "Internet".


oPOSSUM: Select Analysis Parameters - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Refresh Print Mail News RSS Options

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

Google Search Web 386 blocked AutoFill Options



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

## 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...

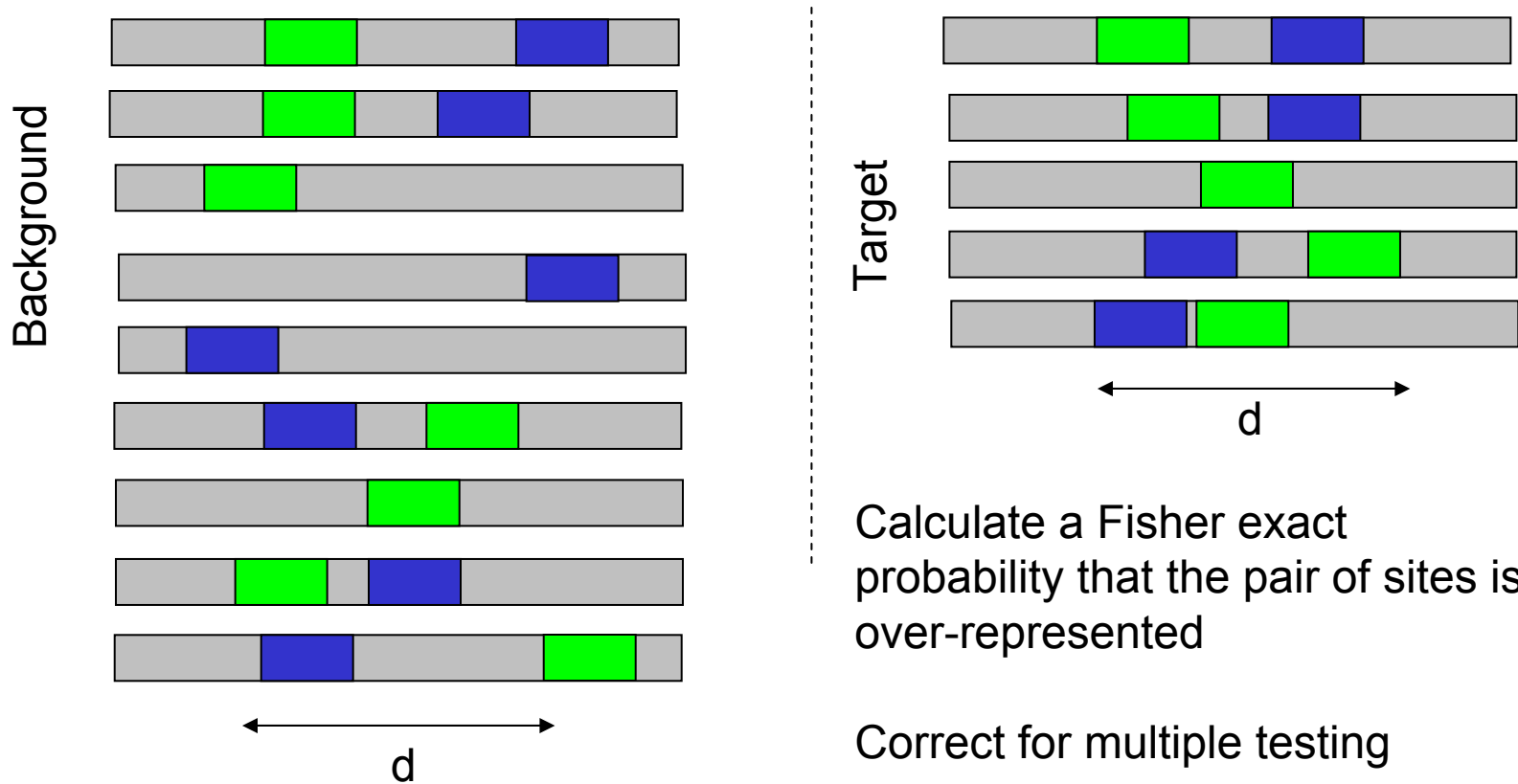
---

#### STEP 2: Select transcription factor binding site matrices

Done Internet

# WHAT CAN WE DO ?

# Identifying over-represented pairs of TFBSs in co-expressed genes



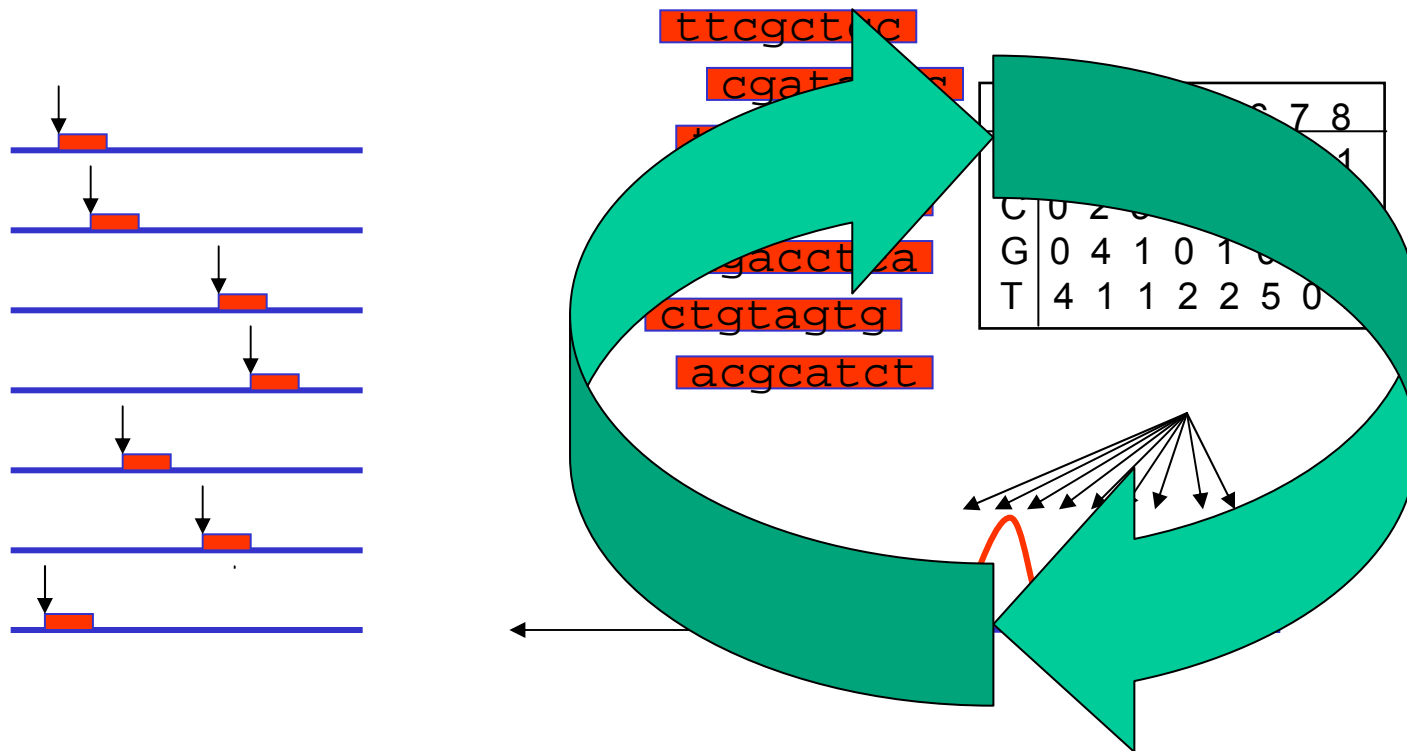
# Over-represented Pairs of Sites in Yeast Fermentation Clusters

cluster	motif1	motif2	Target		Background		p-value	Adjusted
			Hits	No hits	Hits	No hits		
4	CSRE	STRE	15	46	362	6311	8.33E-07	6.49E-04
4	CSRE	GCR1	43	18	2881	3792	1.62E-05	1.26E-02
7	STRE	ADR1P	67	262	835	5838	6.38E-05	4.97E-02
7	STRE	PHO2	70	259	881	5792	5.63E-05	4.39E-02
7	STRE	TBP	69	260	868	5805	6.36E-05	4.96E-02
7	STRE	UASPHR	55	274	628	6045	3.77E-05	2.94E-02
7	STRE	GCR1	68	261	813	5860	1.58E-05	1.23E-02
8	STRE	CAR1_r	25	150	372	6301	2.24E-05	1.75E-02
16	PAC	RRPE	188	293	1958	4715	6.54E-06	5.10E-03
16	RRPE	XBP1	424	57	5354	1319	5.11E-06	3.98E-03
16	RRPE	SCB	411	70	5121	1552	2.78E-06	2.17E-03
16	RRPE	PHO2	425	56	5388	1285	9.28E-06	7.24E-03
16	RRPE	ROX1	273	208	3056	3617	2.09E-06	1.63E-03
16	RRPE	TBP	425	56	5362	1311	3.74E-06	2.92E-03
16	RRPE	FKH1	404	77	5097	1576	4.72E-05	3.68E-02
17	LYS14	RRPE	31	23	1857	4816	5.47E-06	4.27E-03
18	PAC	RRPE	152	206	1958	4715	1.98E-07	1.55E-04
18	RAP1	RRPE	204	154	2901	3772	3.91E-07	3.05E-04
18	RRPE	XBP1	326	32	5354	1319	3.08E-08	2.40E-05
18	RRPE	SCB	309	49	5121	1552	6.59E-06	5.14E-03
18	RRPE	PHO2	325	33	5388	1285	2.38E-07	1.86E-04
18	RRPE	TBP	323	35	5362	1311	5.07E-07	3.96E-04
18	RRPE	UASPHR	256	102	4051	2622	2.02E-05	1.57E-02
18	RRPE	FKH1	312	46	5097	1576	4.20E-07	3.28E-04

# What can we do?

- Predict TFBS
- Predict CRMs
- Phylogenetic Footprinting
- Motif Over-Representation
- **Motif Discovery**

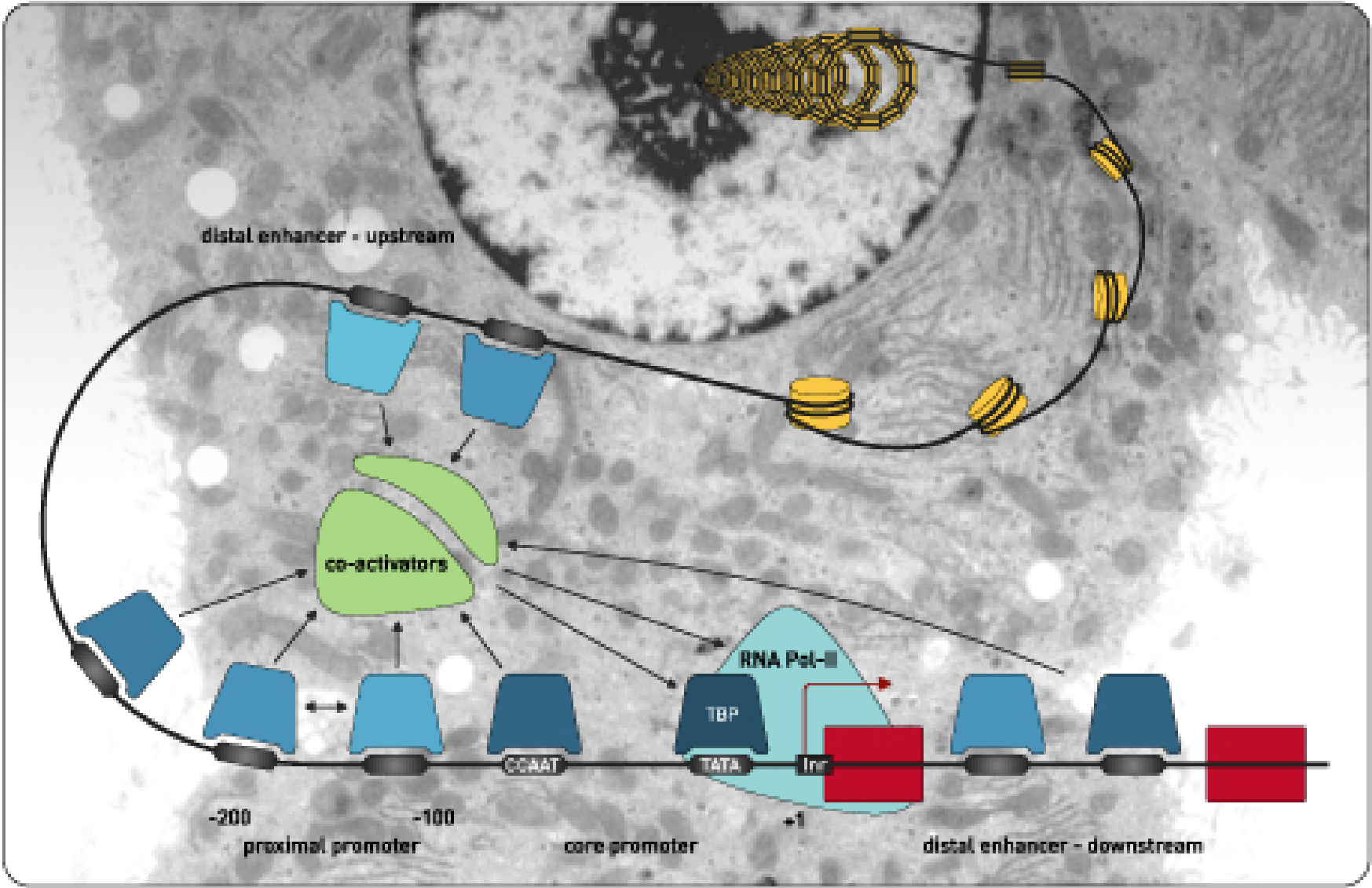
# Gibbs Sampling (grossly over-simplified)



**There are problems...**

Exploring limitations

# Combinatorial interactions between TFs

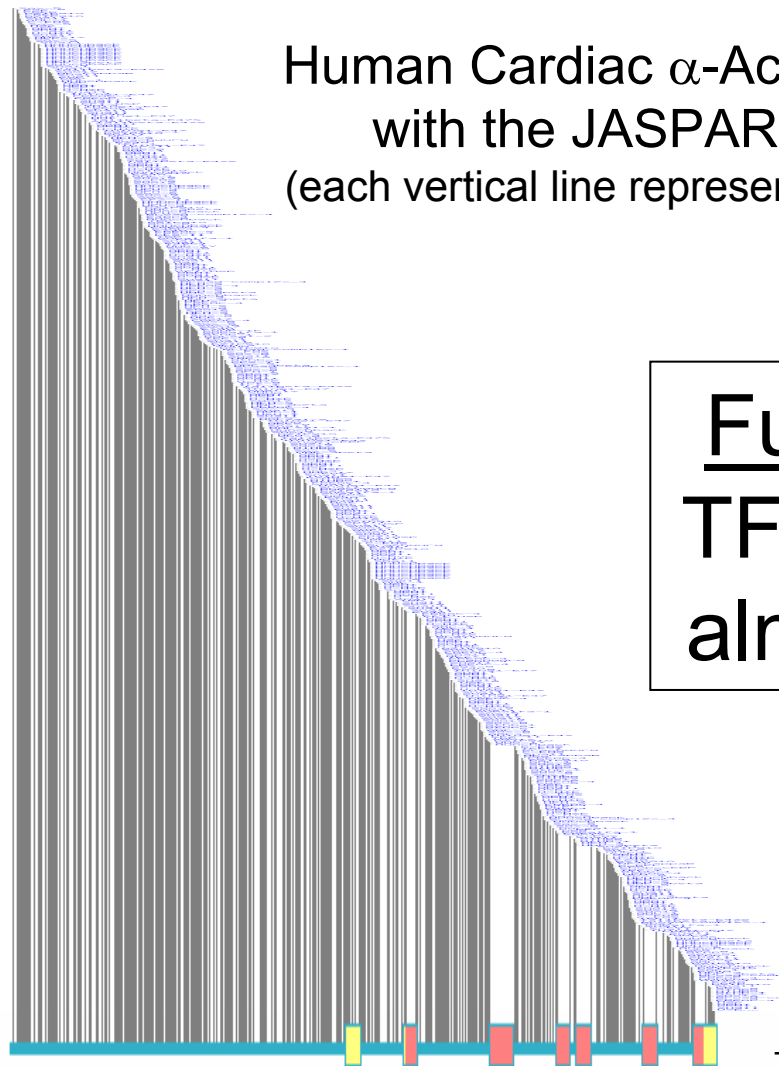


# Why can't we do better?

- Predict TFBS

# Futility Conjecture

Human Cardiac  $\alpha$ -Actin gene analyzed  
with the JASPAR set of profiles  
(each vertical line represents a TFBS prediction)



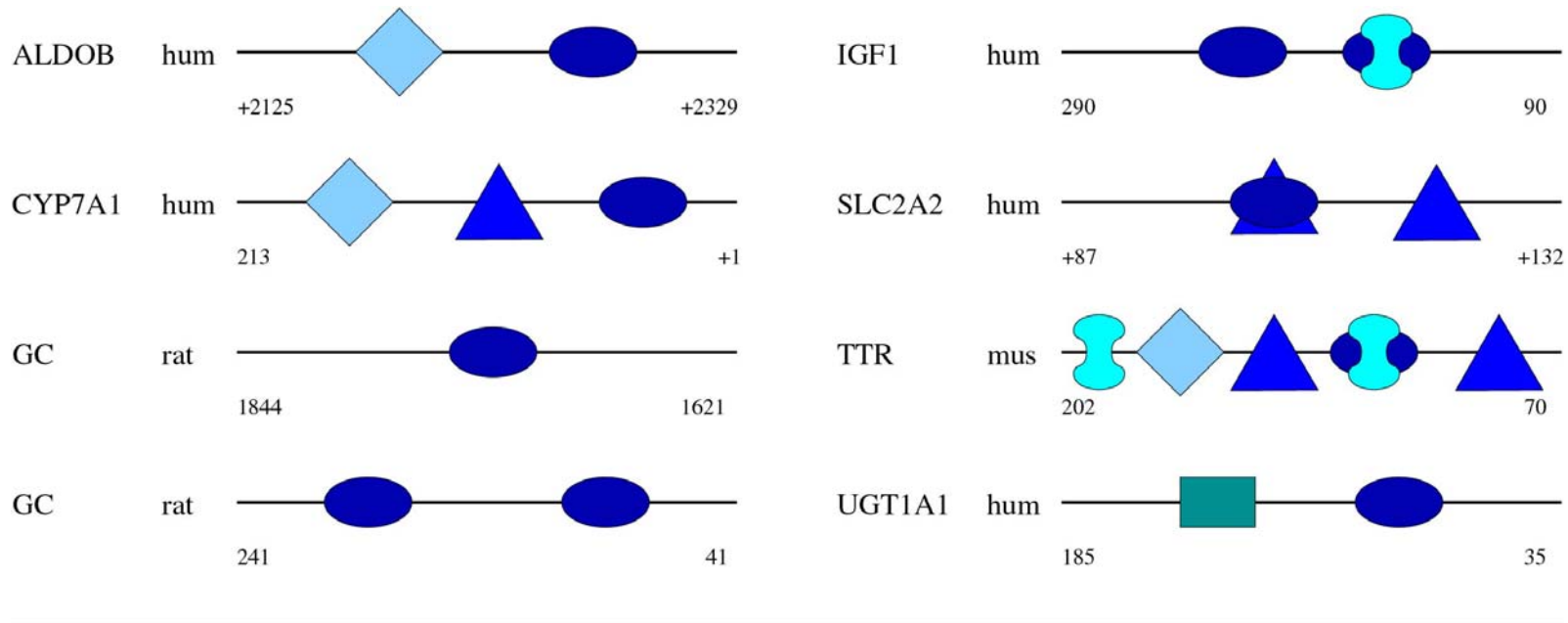
Futility Conjecture:  
TFBS predictions are  
almost always wrong

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

# Why can't we do better?

- Predict TFBS
- Predict CRMs

# Cis-regulatory modules (CRMs) for specific expression in hepatocytes



HNF1



HNF3



HNF4



C/EBP



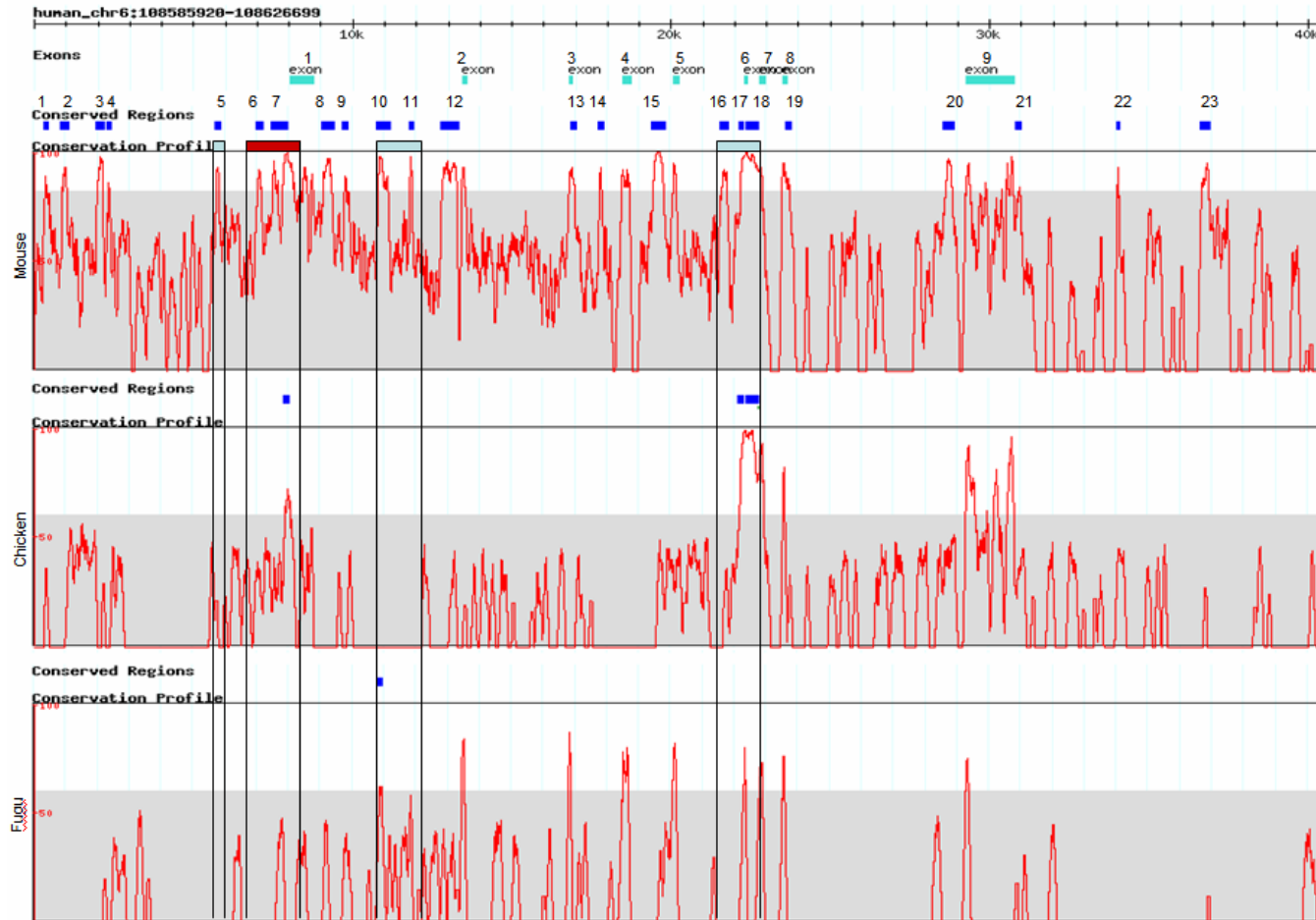
Sp1



# Why can't we do better?

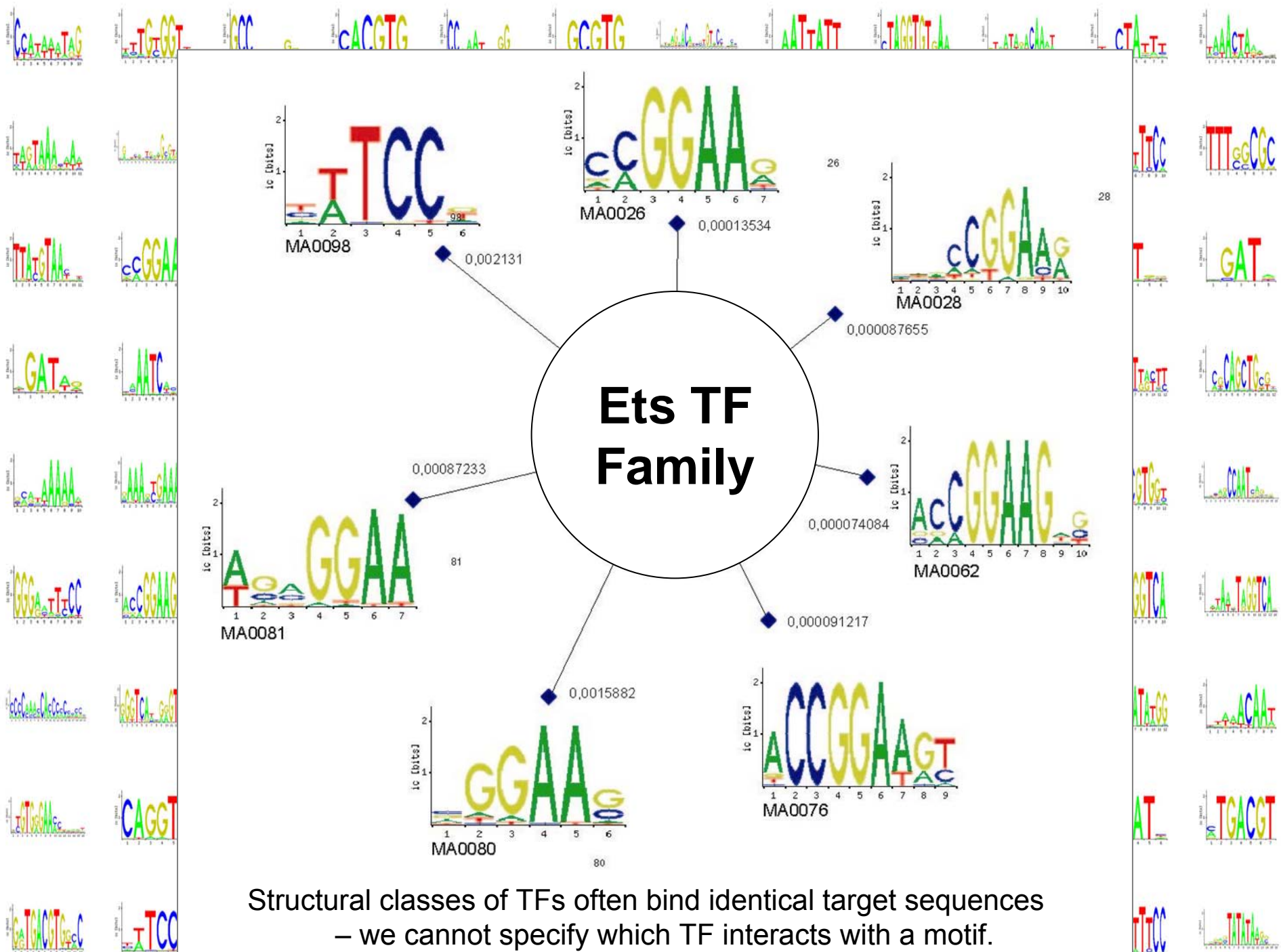
- Predict TFBS
- Predict CRMs
- Phylogenetic Footprinting

# Regulatory Resolution Varies Widely Between Genes



# Why can't we do better?

- Predict TFBS
- Predict CRMs
- Phylogenetic Footprinting
- Motif Over-Representation



Structural classes of TFs often bind identical target sequences  
 – we cannot specify which TF interacts with a motif.

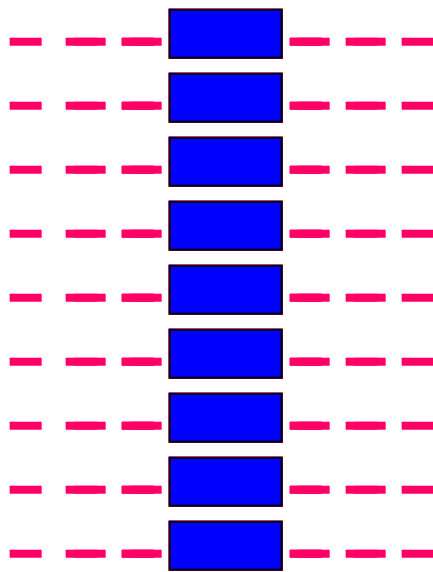
# Challenges for Motif Over-Representation

- Methods fail when noise (genes not co-regulated) exceeds 20-50%
- Most expression profiling experiments are not sufficiently resolved to identify such co-regulated clusters
  - Works well for studies linked to a primary TF response, but fail over long time periods or complex (multi-pathway) responses

# Why can't we do better?

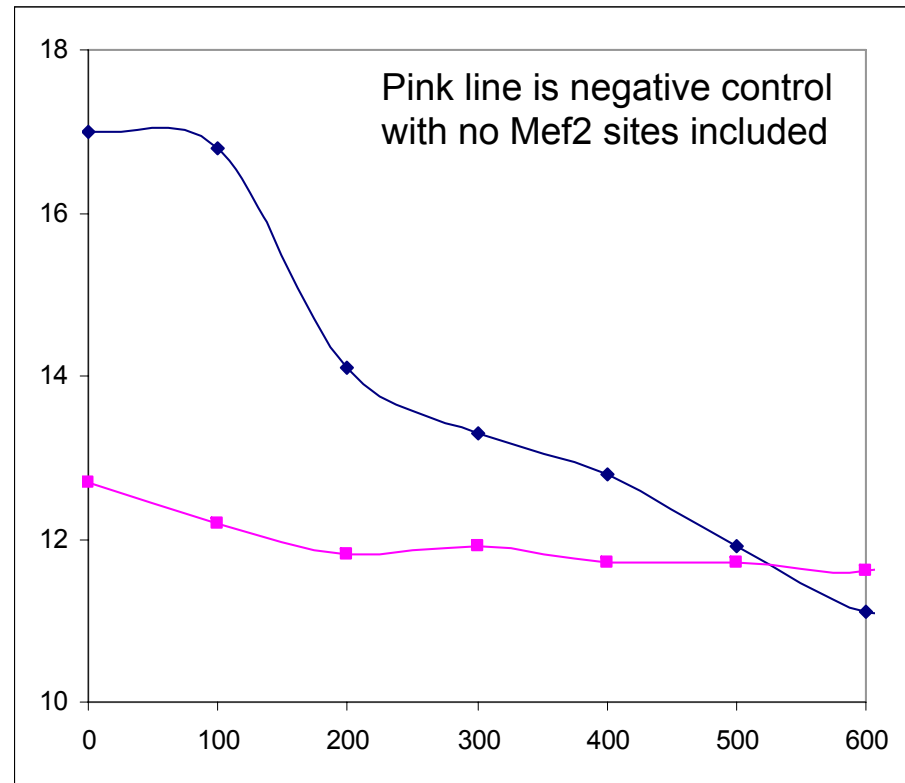
- Predict TFBS
- Predict CRMs
- Phylogenetic Footprinting
- Motif Over-Representation
- Motif Discovery

# Applied Pattern Discovery is Acutely Sensitive to Noise



True Mef2 Binding Sites

PATTERN SIMILARITY  
VS. TRUE MEF2 PROFILE



SEQUENCE LENGTH

# The Signal-to-Noise Battle

- Background models
- Phylogenetic footprinting
- Motif combinations
- Familial Binding Profiles
- Concurrent motif discovery and expression clustering

# Where are we going now?

Snippets of Active Projects

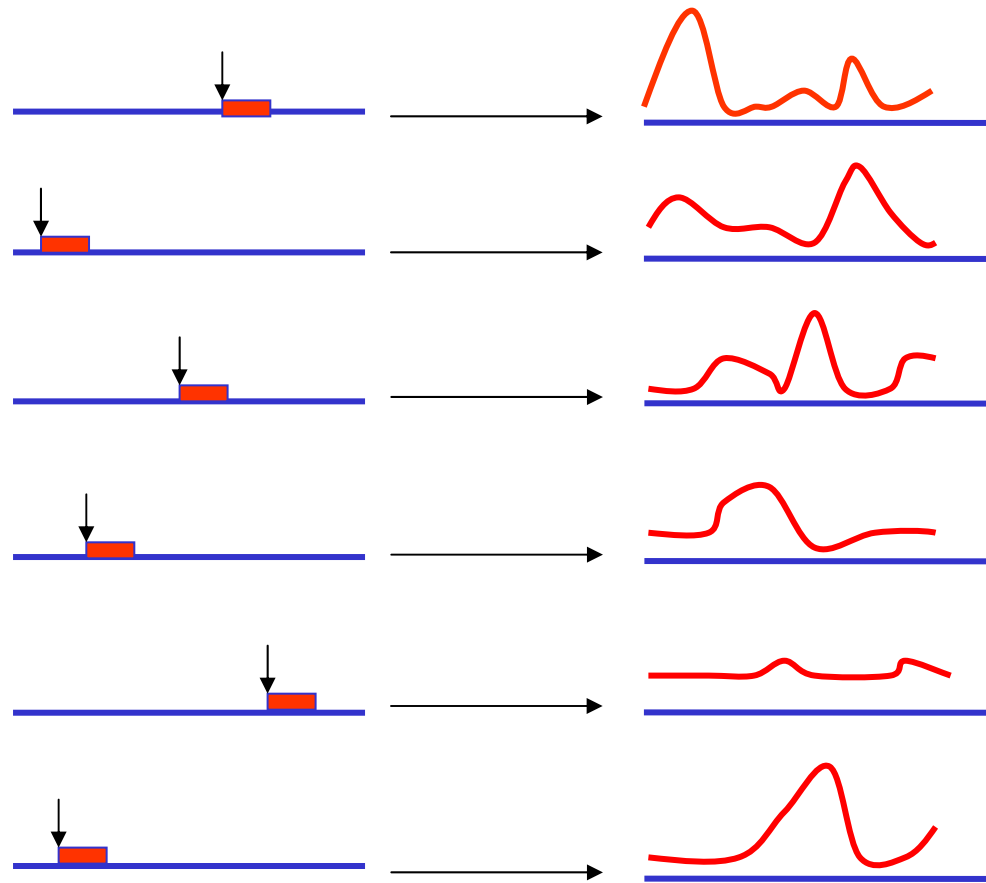
# An impending transition in promoter analysis...

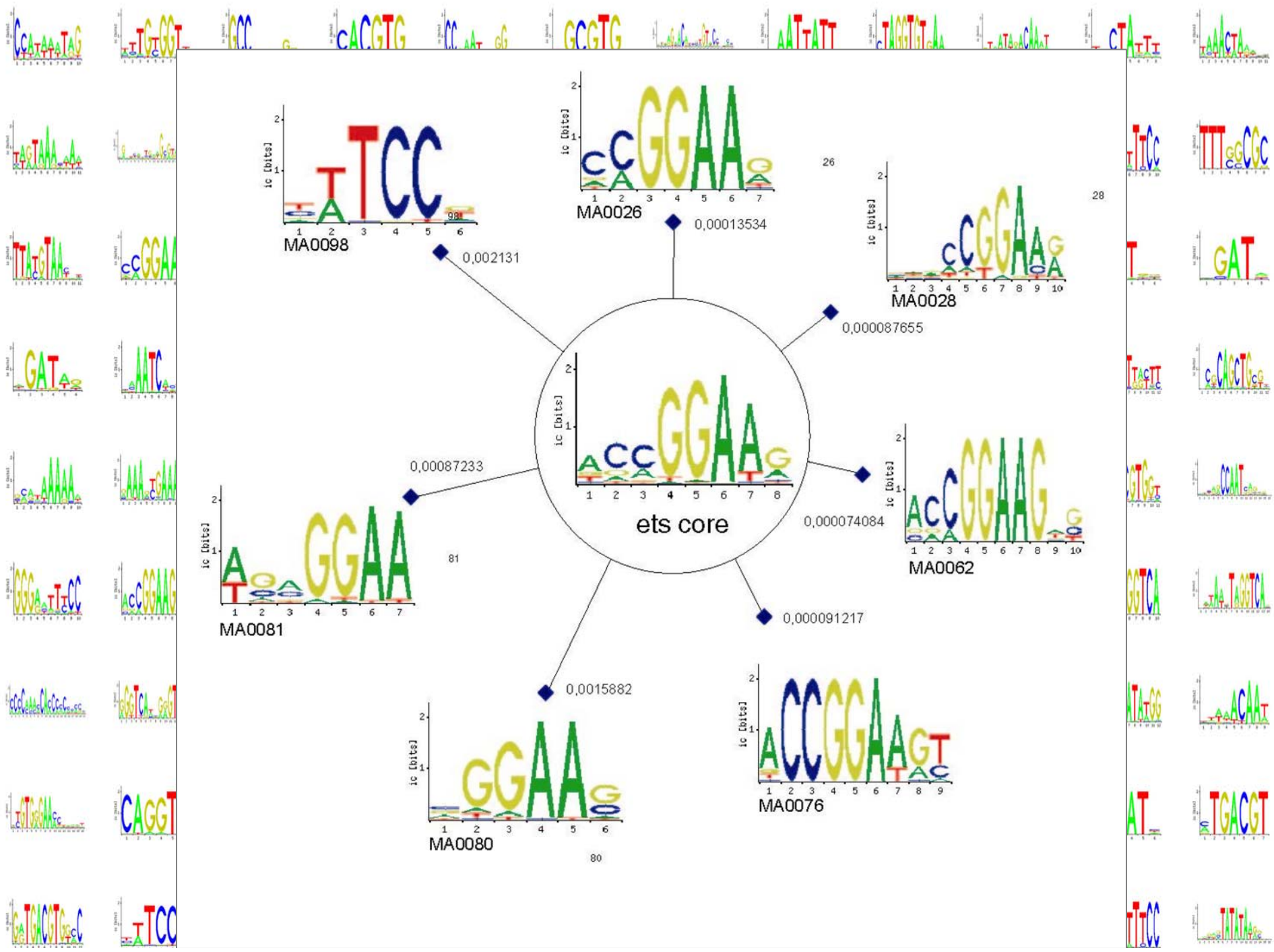
- Transitions in promoter analysis algorithms separated by periods of slow progress
  - Focus on same tired reference collections using progressively more convoluted algorithms
- Advances can be triggered from new data producing technologies, but more commonly from adopting principles well-known to laboratory researchers
  - CpG islands; CRMs; phylogenetic footprinting
- **The next transition: Incorporating data from laboratory studies**

# Informed Motif Discovery

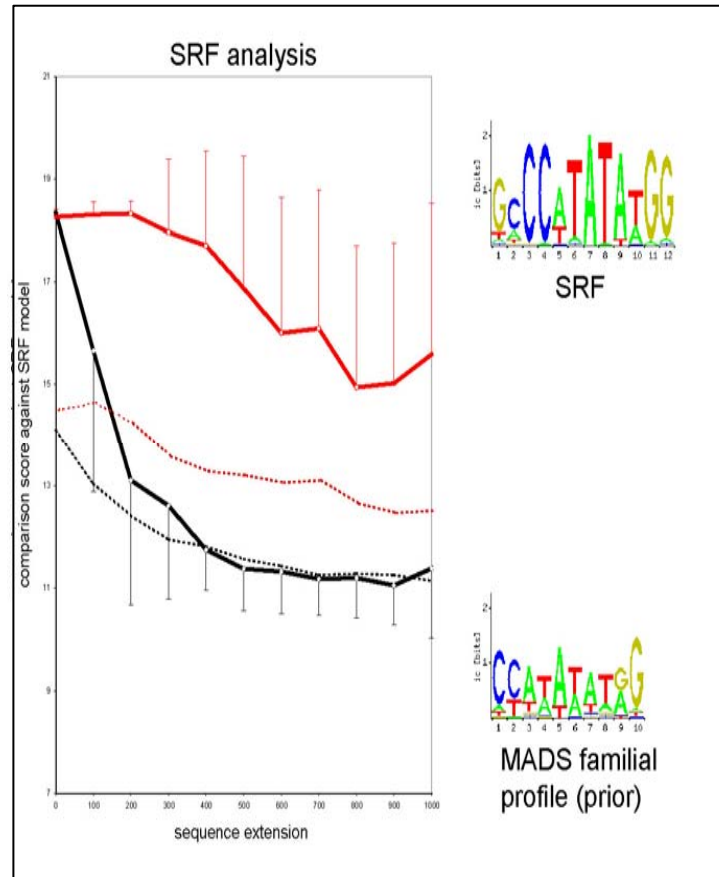
Enhance the Signal  
or  
Reduce the Noise

# Informed Initial Choice



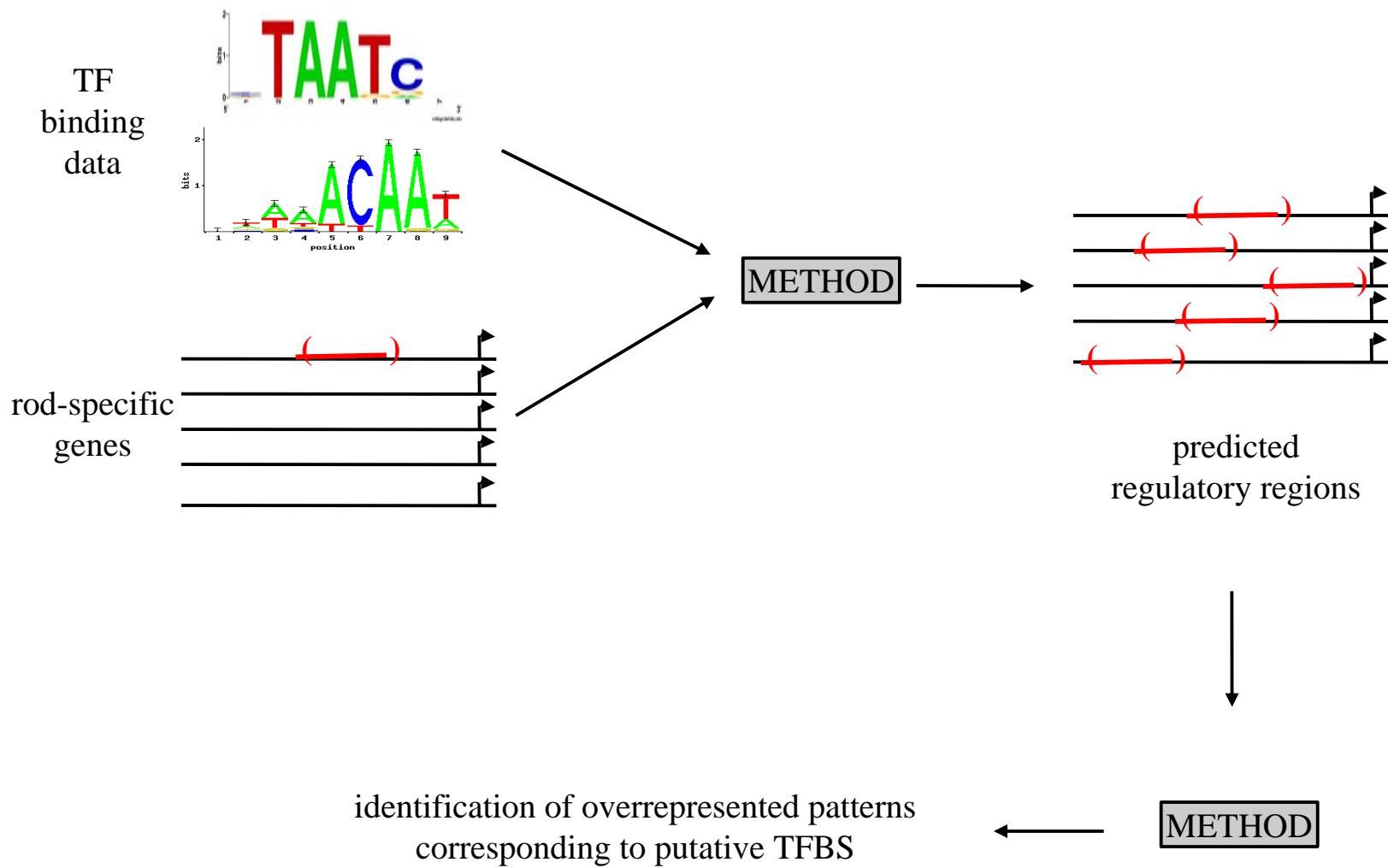


# FBPs enhance sensitivity of pattern detection

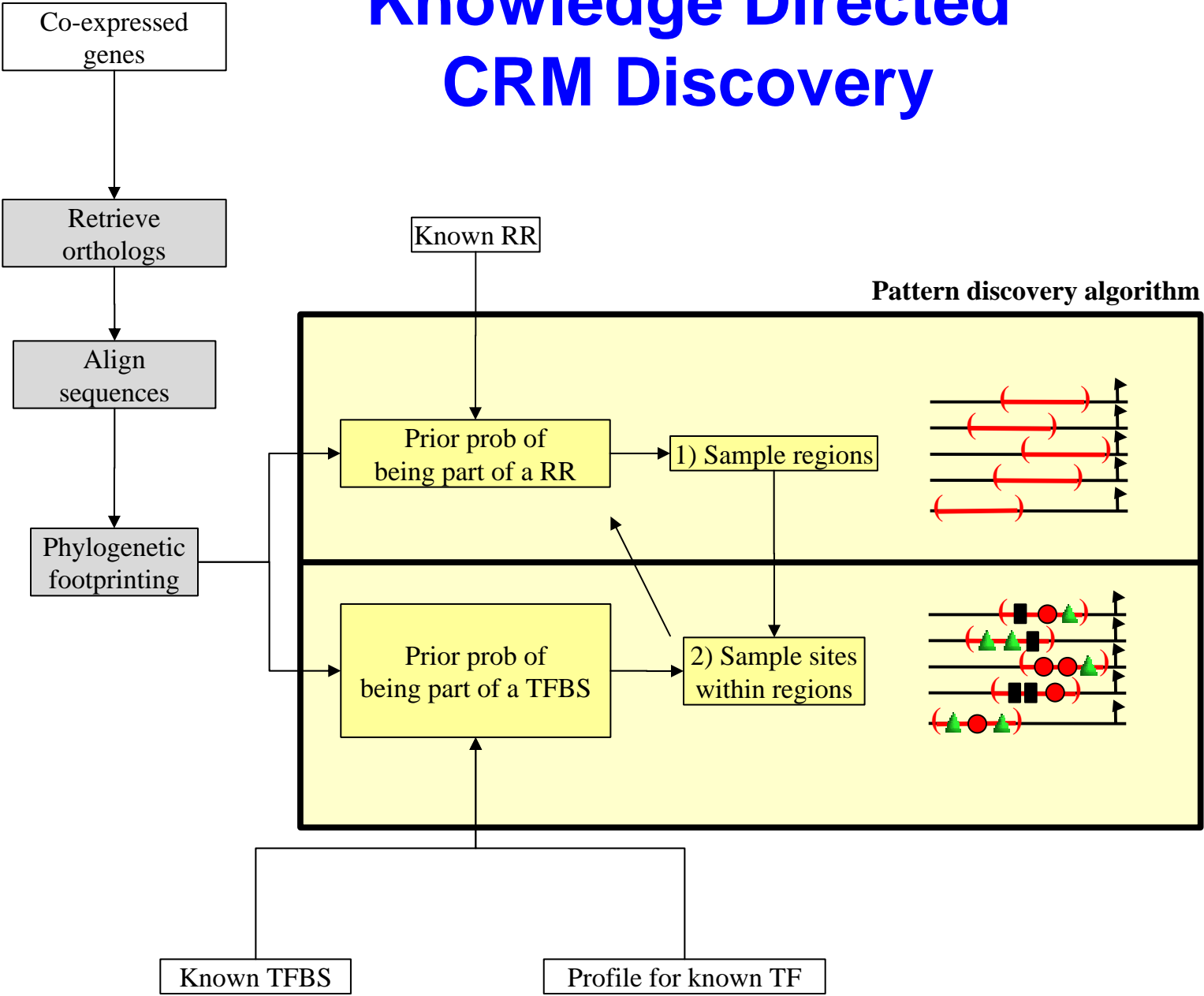


# A new direction?

- Laboratory (WET) data indicating the locations of regulatory regions and/or specific TFBS can constrain the motif discovery process to improve the success rate
- Extension – We should be able to determine how much WET data is required for successful prediction

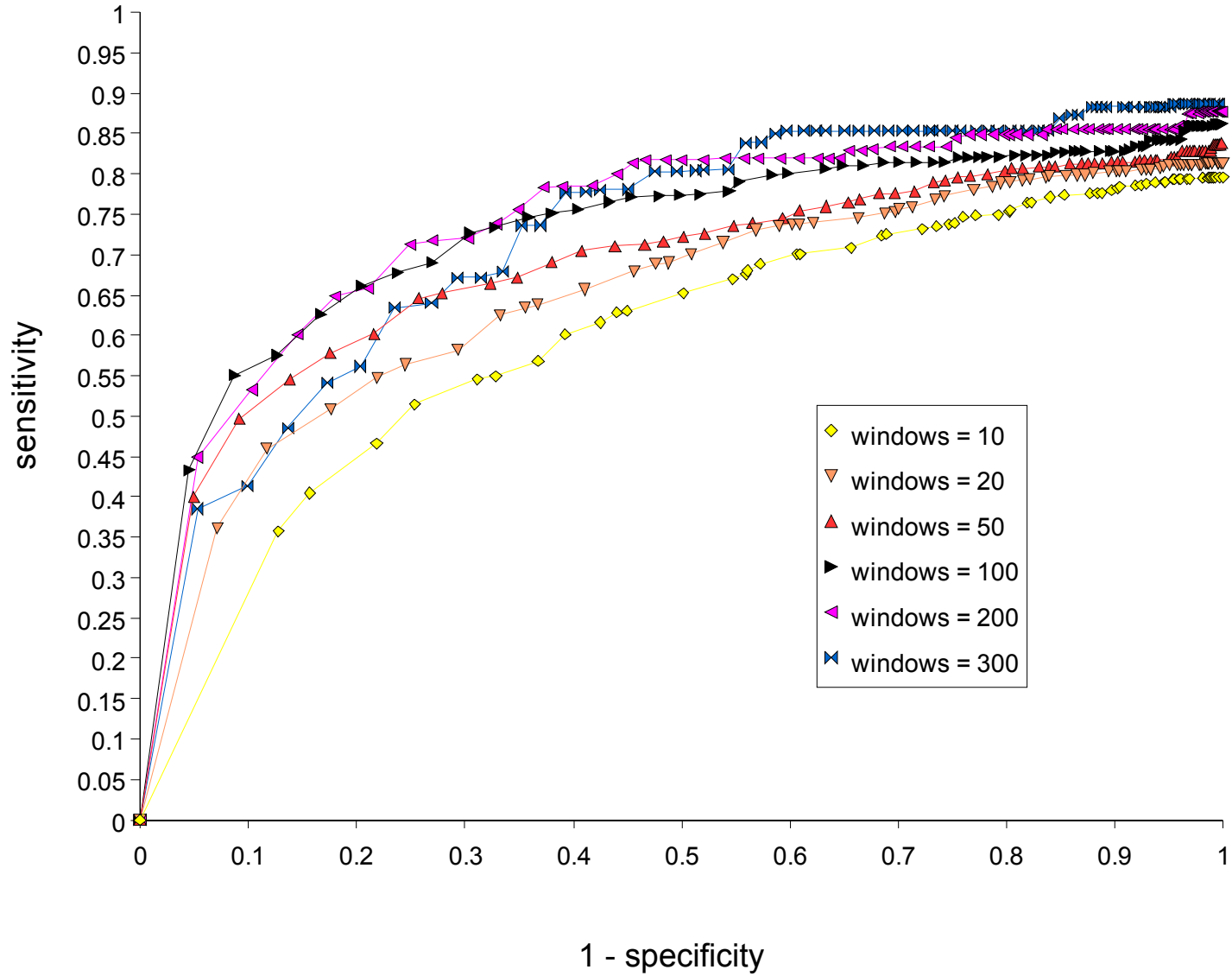


# Knowledge Directed CRM Discovery



CRMs, TFBS and profiles

# ROC curve (exons excluded)



# Software Just Finished

- Test all forms of prior knowledge
  - CRM Length
  - Locations of Known CRMs
  - Location of Known TFBS
  - PSSMs for Contributing TFs
  - Etc
- A limitation - Where to get organized prior data?

# PAZAR

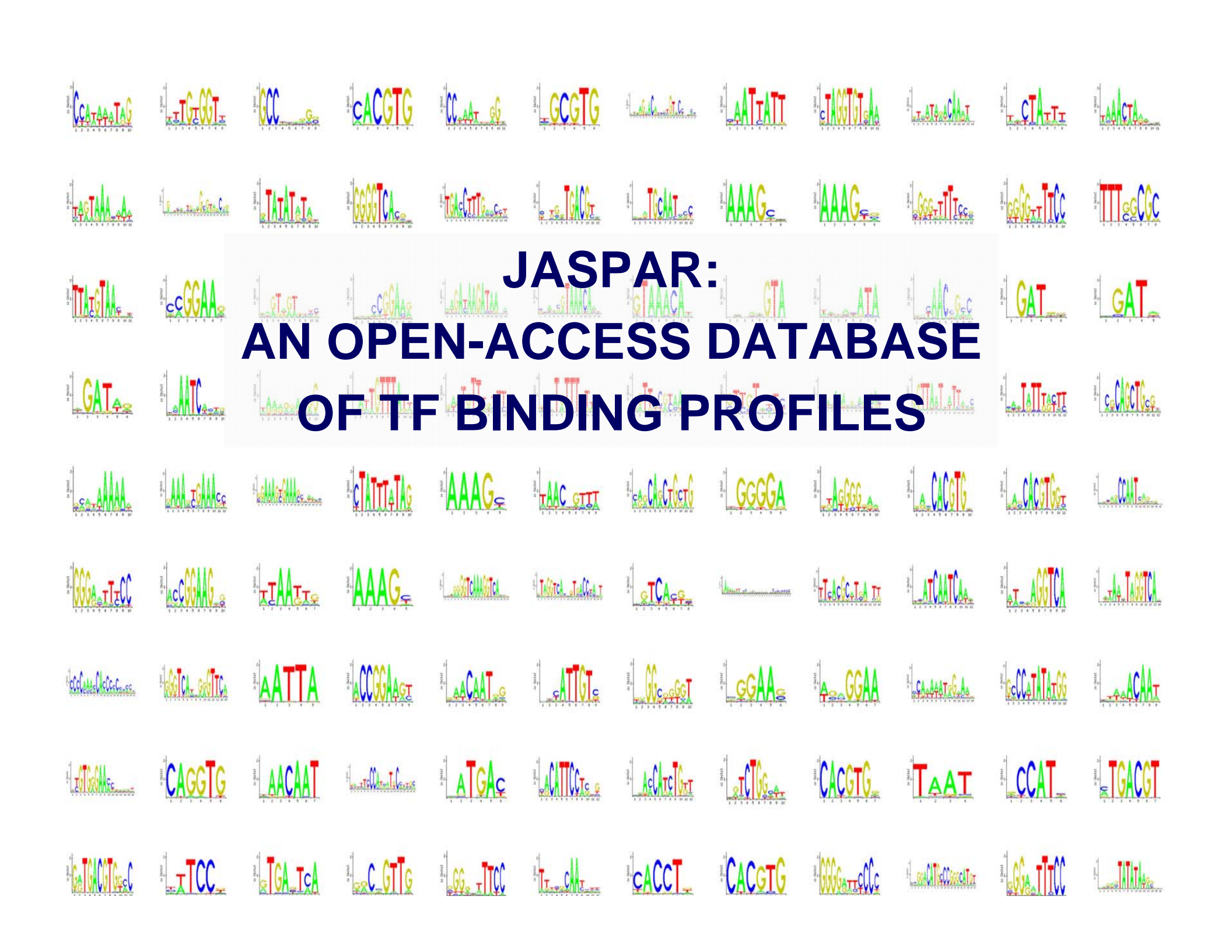
Open-access regulatory sequence  
repository – an information mall

Stefan Kirov  
Elodie Portales-Casamar  
Jonathan Lim  
Jay Snoddy

# PAZAR



Grand Bazaar, Istanbul



**JASPAR:  
AN OPEN-ACCESS DATABASE  
OF TF BINDING PROFILES**

## EXISTING RESOURCES

**Problem:** Many databases with little or no shared data

### commercial

TRANSFAC	eukaryotic transcription factors and their binding profiles <a href="http://www.gene-regulation.de/">http://www.gene-regulation.de/</a>
transcription factors dd	transcription factors of humans and other organisms <a href="http://www.proteinlounge.com/trans_home.asp">http://www.proteinlounge.com/trans_home.asp</a>
TRRD	Transcription Regulatory Regions Db <a href="http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/">http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/</a>

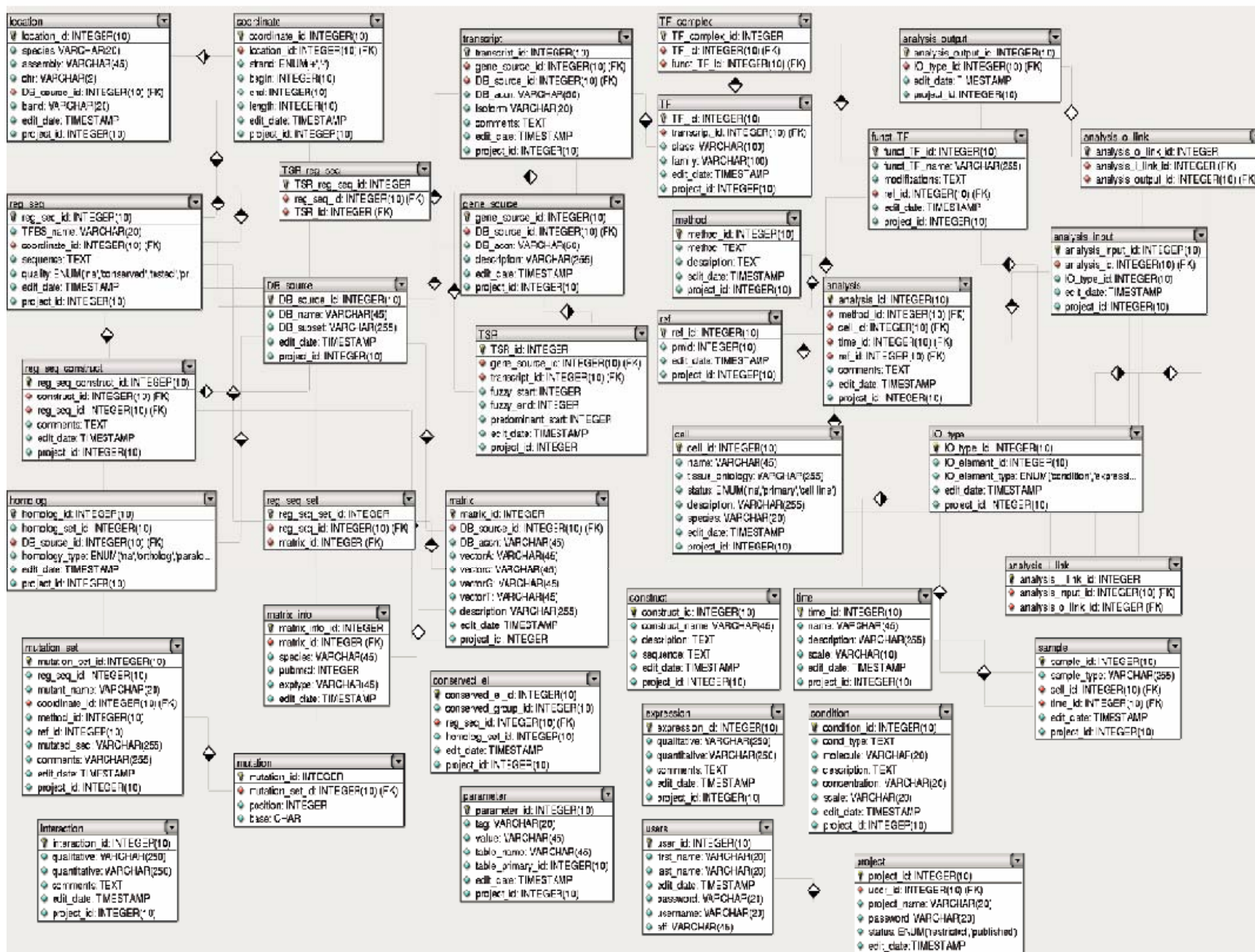
### transcription factors

JASPAR	high-quality transcription factor binding profile db <a href="http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl">http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl</a>
RARTF	RIKEN Arabidopsis Transcription Factor db <a href="http://rarge.gsc.riken.jp/rartf/">http://rarge.gsc.riken.jp/rartf/</a>
ooTFD	object-oriented Transcription Factors Db <a href="http://www.ifti.org/oofd/">http://www.ifti.org/oofd/</a>
TFdb	RIKEN Mouse Transcription Factor Db <a href="http://genome.gsc.riken.jp/TFdb/">http://genome.gsc.riken.jp/TFdb/</a>
RiceTFDB	rice genes involved in transcriptional control <a href="http://ricetfdb.bio.uni-potsdam.de/">http://ricetfdb.bio.uni-potsdam.de/</a>
AGRIS	AtcisDB (Arabidopsis thaliana cis-regulatory db) and AtTFDB (Arabidopsis thaliana transcription factor db). <a href="http://arabidopsis.med.ohio-state.edu/">http://arabidopsis.med.ohio-state.edu/</a>

### cis-regulatory sequences

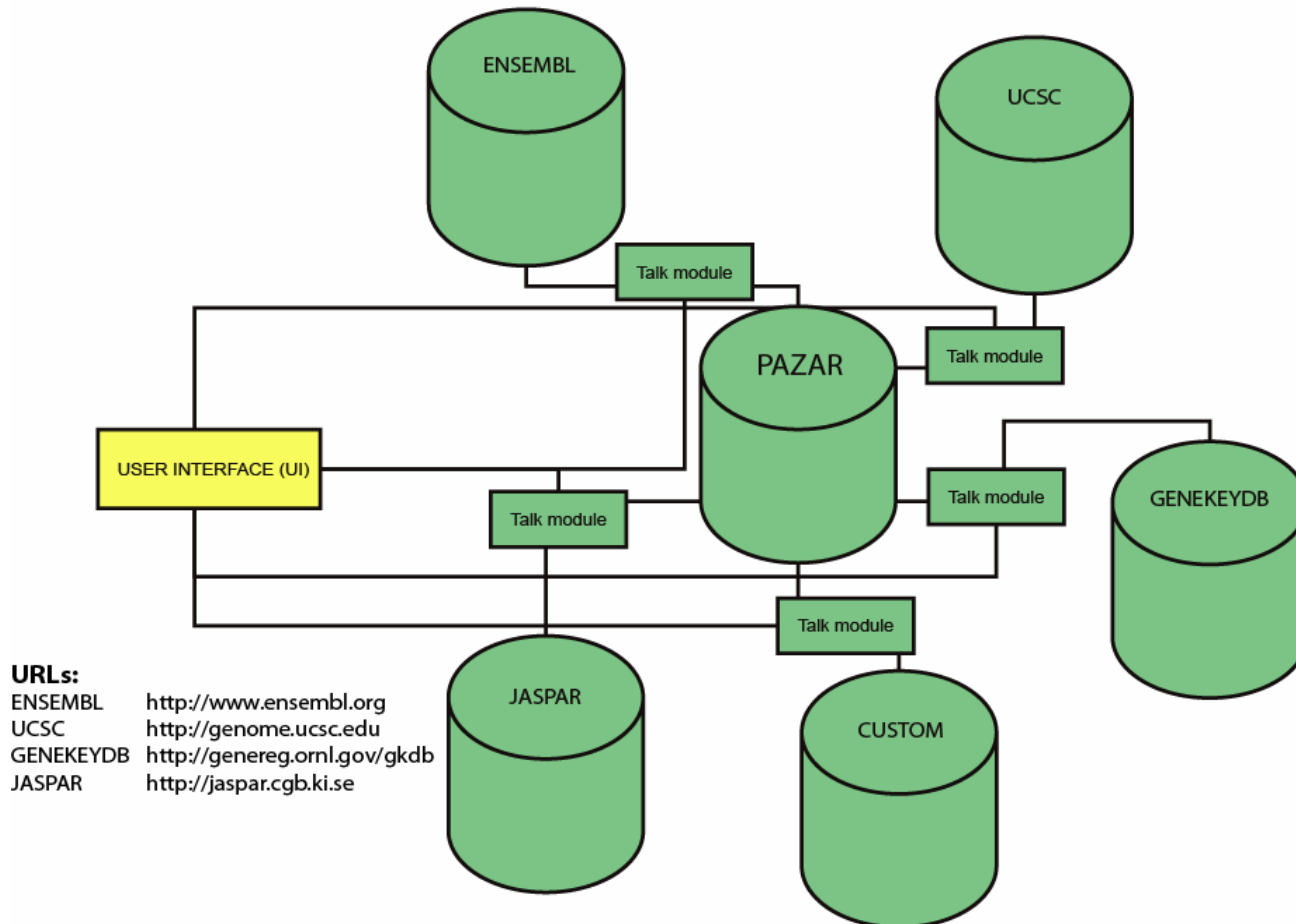
MPD	Mammalian Promoter Db (human, mouse and rat) <a href="http://rulai.cshl.edu/CSHLmpd2">http://rulai.cshl.edu/CSHLmpd2</a>
MPromDb	Mammalian Promoter Db with experimentally supported annotations <a href="http://bioinformatics.med.ohio-state.edu/MPromDb/">http://bioinformatics.med.ohio-state.edu/MPromDb/</a>
OMGProm	Orthologous Mammalian Gene Promoters <a href="http://bioinformatics.med.ohio-state.edu/OMGProm/">http://bioinformatics.med.ohio-state.edu/OMGProm/</a>
DoOP	Orthologous clusters of promoters. <a href="http://doop.abc.hu/">http://doop.abc.hu/</a>
EPD	Eukaryotic Promoter Db <a href="http://www.epd.isb-sib.ch/">http://www.epd.isb-sib.ch/</a>
SCPD	S. cerevisiae Promoter Db <a href="http://rulai.cshl.edu/SCPD/">http://rulai.cshl.edu/SCPD/</a>
CEPDB	C. elegans Promoter Db <a href="http://rulai.cshl.edu/cgi-bin/CEPDB/home.cgi">http://rulai.cshl.edu/cgi-bin/CEPDB/home.cgi</a>
PLACE	Plant Cis-acting Regulatory DNA Elements <a href="http://www.dna.affrc.go.jp/PLACE/">http://www.dna.affrc.go.jp/PLACE/</a>
Plant CARE	Cis-Acting regulatory element. <a href="http://intra.psb.ugent.be:8080/PlantCARE/">http://intra.psb.ugent.be:8080/PlantCARE/</a>
PlantProm DB	Plant Promoter Sequences <a href="http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom">http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom</a>
OPD	Osteo-Promoter Db (promoters of genes in the osteogenic pathway) <a href="http://www.opd.tau.ac.il/">http://www.opd.tau.ac.il/</a>
HemoPDB	Hematopoiesis Promoter Db <a href="http://bioinformatics.med.ohio-state.edu/HemoPDB/">http://bioinformatics.med.ohio-state.edu/HemoPDB/</a>
LSPD	The Liver Specific Gene Promoter Database <a href="http://cgsigma.cshl.org/LSPD">http://cgsigma.cshl.org/LSPD</a>
MTIR	Muscle-specific regulation of transcription <a href="http://www.cbil.upenn.edu/MTIR/HomePage.html">http://www.cbil.upenn.edu/MTIR/HomePage.html</a>
the Globin Gene Server	experimental data on the regulation of the globin gene cluster <a href="http://globin.cse.psu.edu/">http://globin.cse.psu.edu/</a>
Oreganno	open regulatory annotation <a href="http://oreganno.org">http://oreganno.org</a>

## A complex database schema to allow flexibility

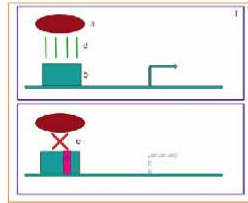


## PAZAR can be linked to external data resources (ensembl, genekeydb) using a “talk” module

PAZAR is confined to the description of regulatory sequence features. There is often need for other information, such as gene identifiers, genomic DNA sequence, etc. The API talk module grants access to external resources. It is easily extensible to support other databases, including new “malls”, while providing standard accessor methods.



## XML exchange format



Relationship Between Entities and Tables	
Entity	Table(s) involved
a Transcription Factor (TF)	TF, transcript, gene_source
b Transcription Factor Binding Site (TFBS)	reg_seq, TSR, gene_source
c TFBS Mutation	mutation, mutation_set
d TF - TFBS Interaction (induction of expression)	interaction / expression
e No TF - TFBS Interaction (no induction of expression)	interaction / expression
f Experiment Description	method, ref, cell, time, condition
g Project Description	project, users

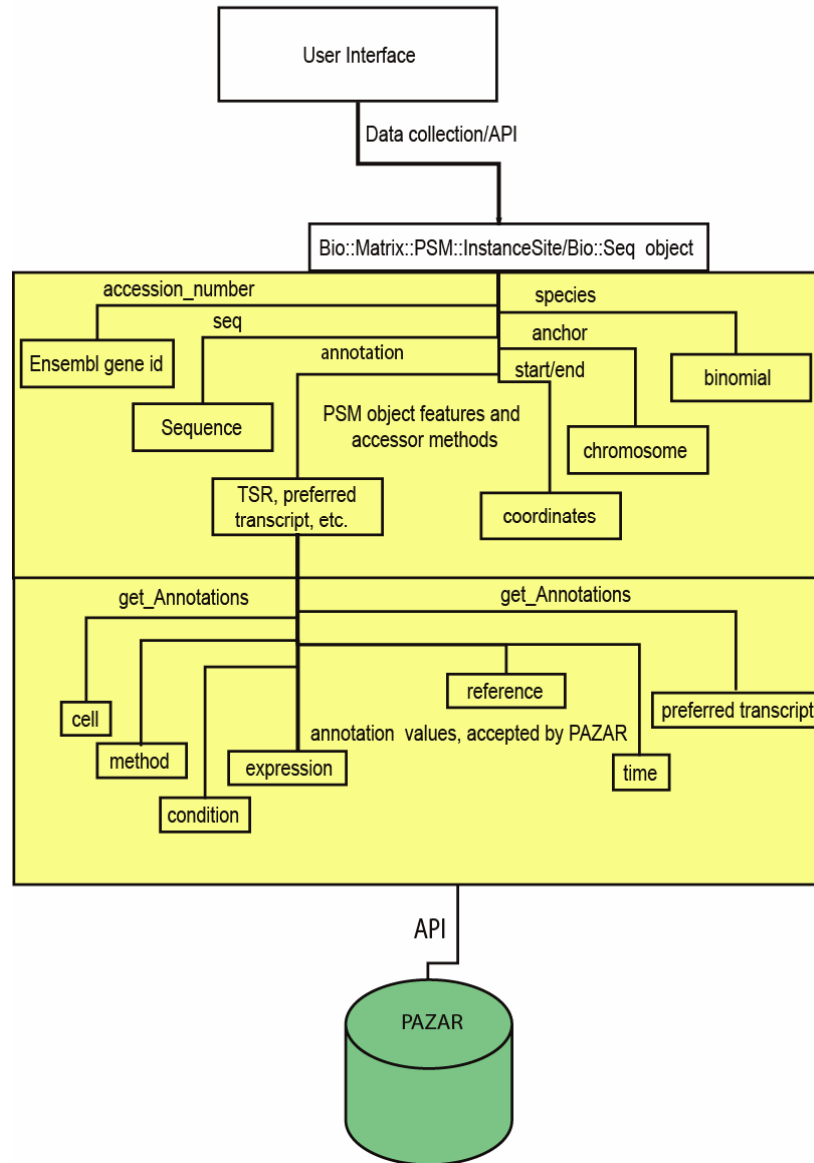
```

<pazar>
  <project name="example_project" pazar_id="project_0" status="restricted">
    <user affiliation="affiliation" first_name="first_name" last_name="last_name"
pazar_id="user_0" username="username"/>
  </project>
  <data>
    <gene_source description="PDE6B" pazar_id="gene_0">
      <db_accession db_accn="ENSG00000133256" db_name="ensembl" />
      <TSR fuzzy_start="609373" fuzzy_end="609373" pazar_id="TSR_0">
        <transcript pazar_id="transcript_0">
          <db_accession db_accn="ENST00000255622" db_name="ensembl" />
        </transcript>
        <reg_seq TFBS_name="NRE" quality="tested" pazar_id="reg_seq_0"
sequence="ATTGTAGGAGTGAGTCAGCTGACCCGC">
          <coordinate begin="609283" end="609310" length="28" strand="+">
            <location assembly="NCBI 35" band="4p16.3" species="human">
              <db_accession db_name="ensembl" />
            </location>
          </coordinate>
          <mutation_set pazar_id="mutation_set_0">
            <coordinate begin="609294" end="609299" length="6" strand="+">
              <location assembly="NCBI 35" band="4p16.3" species="human">
                <db_accession db_name="ensembl" />
              </location>
            </coordinate>
            <mutation base="g" position="1" pazar_id="mutation_0"/>
            ...
            <mutation base="a" position="6" pazar_id="mutation_5"/>
          </mutation_set>
        </reg_seq>
      </TSR>
    </gene_source>
    <gene_source description="NRL" pazar_id="gene_1">
      <db_accession db_accn="ENSG00000129535" db_name="ensembl" />
      <transcript pazar_id="transcript_1">
        <db_accession db_accn="ENST_00000250471" db_name="ensembl" />
        <tf class="bZIP" family="MAF" pazar_id="tf_0"/>
      </transcript>
    </gene_source>
    <funct_tf pazar_id="funct_tf_0" tf_ids="tf_0"/>
    <interaction qualitative="yes" pazar_id="interaction_0"/>
    <interaction qualitative="no" pazar_id="interaction_1"/>
  </data>
  <analysis >
    <method method="EMSA"/>
    <ref pmid="11438531"/>
    <cell name="Y79" species="human" status="cell_line"/>
    <input_output >
      <input inputs="funct_tf_0"/>
      <output outputs="interaction_0"/>
    </input_output>
    <input_output>
      <input inputs="mutation_set_0"/>
      <output outputs="interaction_1"/>
    </input_output>
  </analysis>
</pazar>

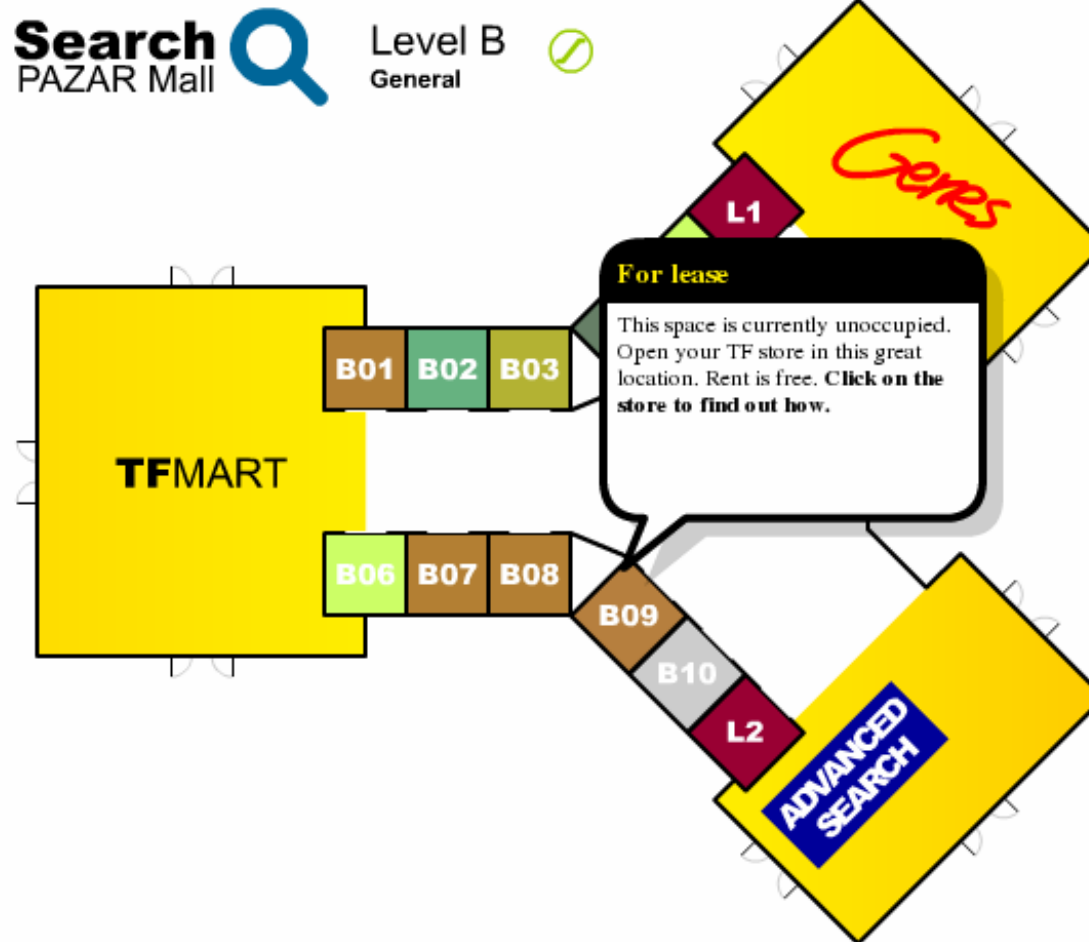
```

## API data structure

The API is based on existing Bioperl data structures and methods. Using Bioperl allows the PAZAR project to use standardized procedures.



# Retrieval/Browsing Interface



# Status

- PAZAR – Database Implemented
- API/Perl Modules – Available
- Streamlined Submission Interface – Available
- COHO - In Progress
- Release impending
- Open-Access/Open-Software: see [www.pazar.info](http://www.pazar.info) for details

# Putting It All Together



Control stick

Pan

Zoom

Reset

Viewing data for

Select other TF

Buffer

50

Refresh

Out of range

MA0052

Position

510



50 bp

75 bp

456

451

564

569

1000

0

0

82

Test

MEF2

CREB

E2F

GATA-2

GATA-3

109 bp

541

current cycle

restart

play

GATA-3

MEF2

CREB E2F

CREB MEF2

E2F

# Final Thoughts

- The grand challenge remains for the analysis of co-regulated human genes
- Significant progress in the past five years suggests that we will be able to decipher regulatory mechanisms for targeted experiments
- Numerous attractive problems remain available for bioinformatics students

# Thanks!

## THE AMAZING PEOPLE WHO DID THE WORK!

- Elodie Portales-Casamar
- David Martin
- David Arenillas
- Jochen Brumm
- Alice Chou
- Debra Fulton
- Miroslav Hatas
- Shannan Ho Sui
- Andrew Kwon
- Jonathan Lim
- Dora Pak
- Raf Podowski
- Diane Wu
- Dimas Yusuf



- James Mortimer
- Brian Kennedy



- Jay Snoddy
- Stefan Kirov (BMS)



- CIHR
- IBM
- MSFHR
- MerckFrosst
- GenomeBC
- GenomeCanada
- CFI