Bioinformatics Workshop 2009

Interpreting Gene Lists from -omics Studies

Deciphering regulatory networks by promoter sequence analysis

Elodie Portales-Casamar

University of British Columbia

www.cisreg.ca

Overview

Part 1: Overview of transcription

Lab 1: Promoters in Genome Browser (UCSC and PAZAR)

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

Lab 2: TFBS scan (ORCAtk)

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

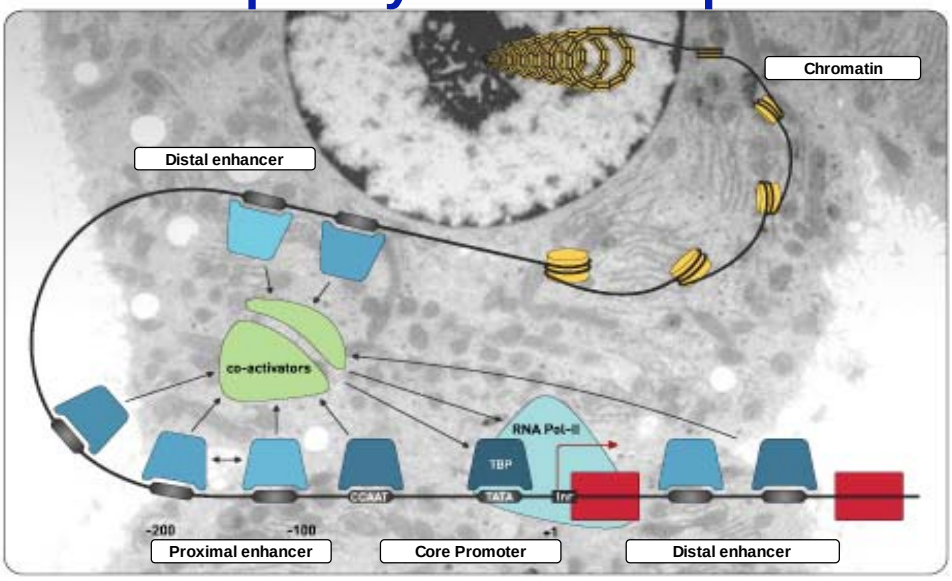
Lab 3: TFBS Over-Representation (oPOSSUM)

Restrictions in Coverage

- Focus on Eukaryotic cells and PolII Promoters
  - Principles apply to prokaryotes
  - Will provide suggestions for similar tools for other species as requested
- Many of the examples drawn from the Wasserman lab's work
  - there are equivalent tools

Part 1
Introduction to transcription in eukaryotic cells

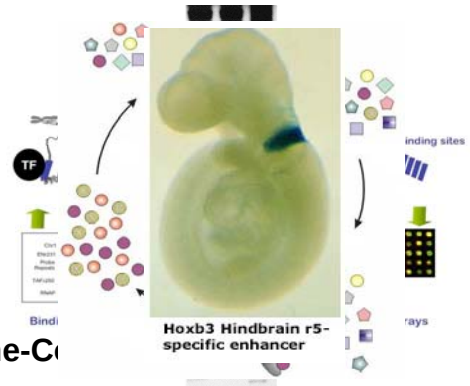
Complexity in Transcription



Studying gene expression at the bench

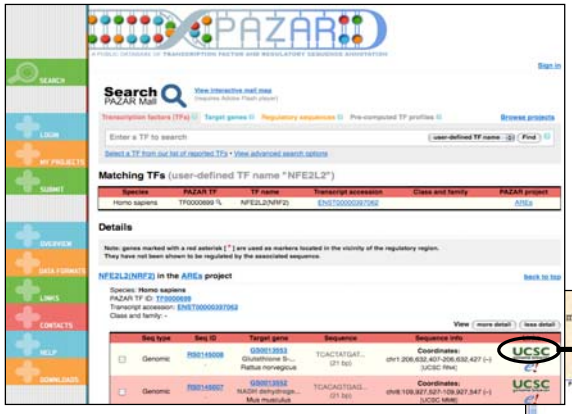
- EMSA
- DNase I footprinting
- ChIP- chip
- SELEX experiment
- Gene reporter assay

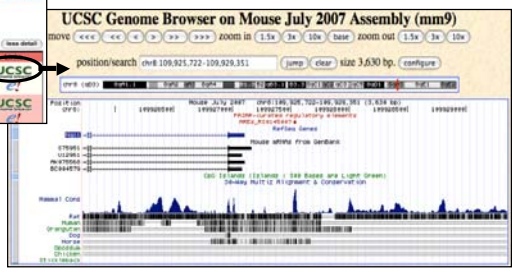
Expensive and Time-C



http://www.chiponchip.org/
http://www.hku.hk
http://dukehealth1.org
http://opbs.okstate.edu

PAZAR and UCSC





Part 2
Prediction of
TF Binding Sites

Teaching a computer to find TFBS…

TF Binding Profile

Aligned binding sites

TCACTATGATTCAGCAACAAA
TCACAGTGAGTCGGCAAAATT
TCATGCTGACTCAGCGGATCG
CAACCATGACACAGCATAAAA
CAGGCATGACATTGCATTTTT
TAATGGTGACAAAGCAACTTT
GGAGCATGACCCAGCAGAAGG
CTGGGATGACATAGCATTCAT
TCAGAATGACAAAGCAGAAAT
TCACCGTTACTCAGCACTTTG
AGGTGGTGATGTTGCATCACA
CCAGGATGACTTAGCAAAAAC
AGCCTGTGACTGGGCCGGGGC
AGACAATGACTAAGCAGAAAT
TCCCCGTGACTCAGCGCTTTG
TCAGCATGACTCAGCAGTCGC
CCTCCATGACAAAGCACTTTT
AGCGGGTGACCAAGCCCTCAA
TCAGGGTGACTCAGCAGCTTG
TCTGTGTGACTCAGCTTTGGA

Position Frequency Matrix (PFM)

A	10	0	0	20	0	6	5	16	0	0	15
C	1	0	0	0	17	2	10	0	0	20	2
G	9	0	19	0	1	1	1	2	20	0	2
T	0	20	1	0	2	11	4	2	0	0	1

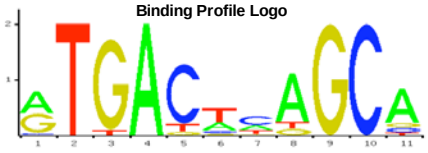
Position Specific Scoring Matrix (PSSM)

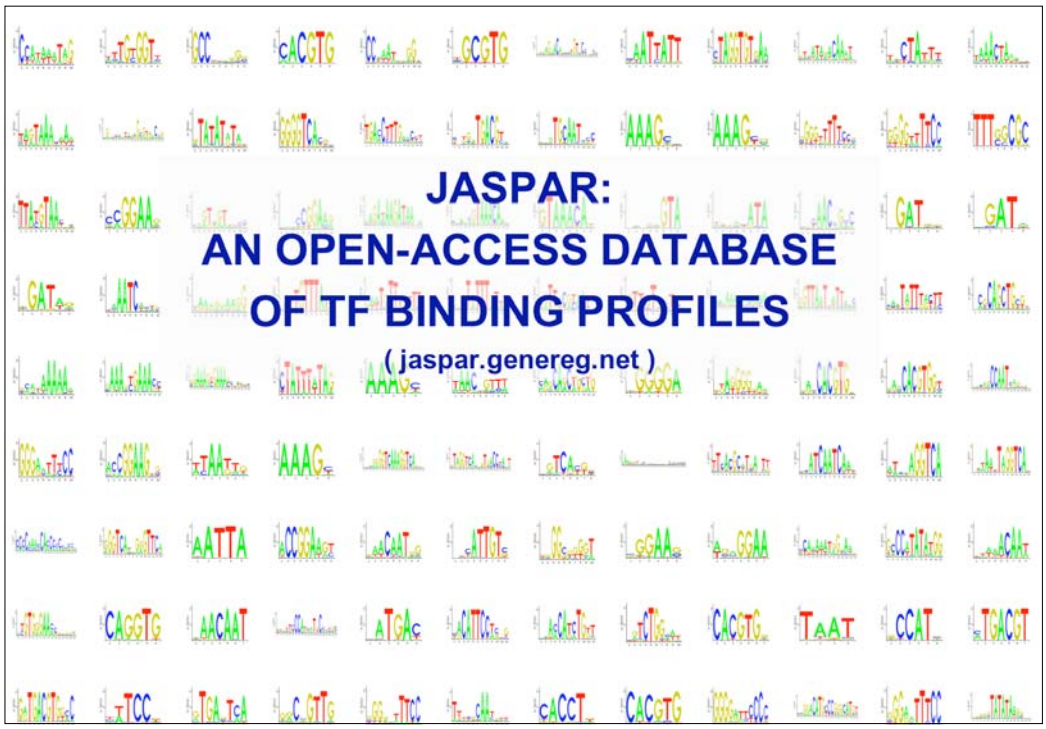
A	0.9	-2.5	-2.5	1.8	-2.5	0.2	0.0	1.5	-2.5	-2.5	1.4
C	-1.5	-2.5	-2.5	-2.5	1.6	-1.0	0.9	-2.5	-2.5	1.8	-1.0
G	0.7	-2.5	1.7	-2.5	-1.5	-1.5	-1.5	-1.0	1.8	-2.5	-1.0
T	-2.5	1.8	-1.5	-2.5	-1.0	1.0	-0.3	-1.0	-2.5	-2.5	-1.5

A T G A T T C A G C A

Score = 13.6

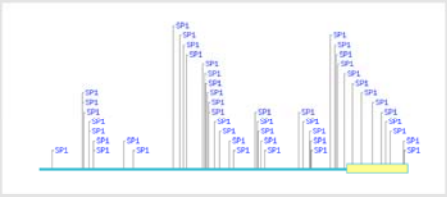
Binding Profile Logo





Analysis of TFBS with Phylogenetic Footprinting

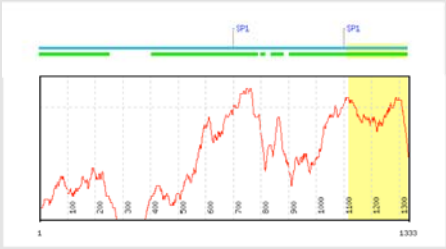
Scanning a single sequence



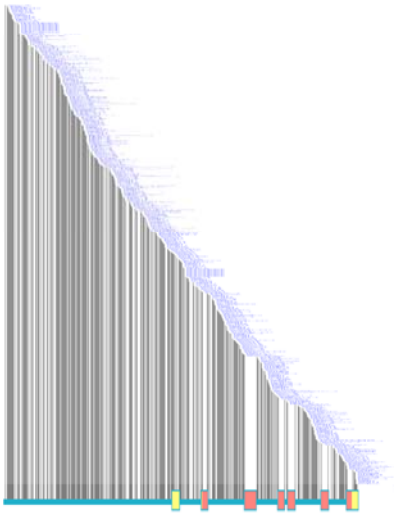
Low specificity of profiles:
- too many hits
- great majority not biologically significant

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

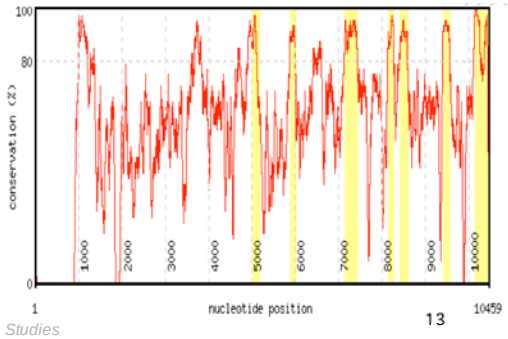
A dramatic improvement in the percentage of biologically significant detections



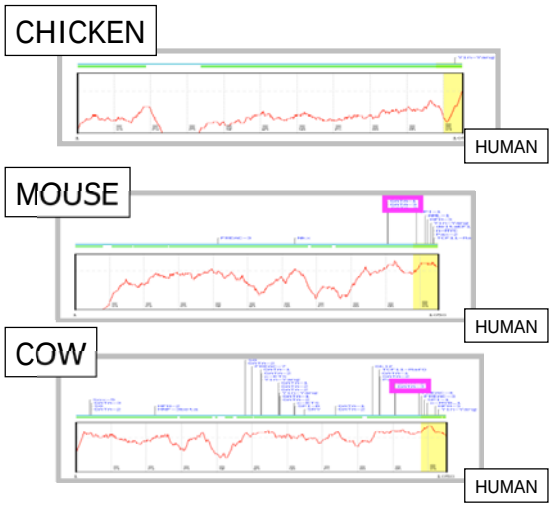
Phylogenetic Footprinting Dramatically Reduces Spurious Hits



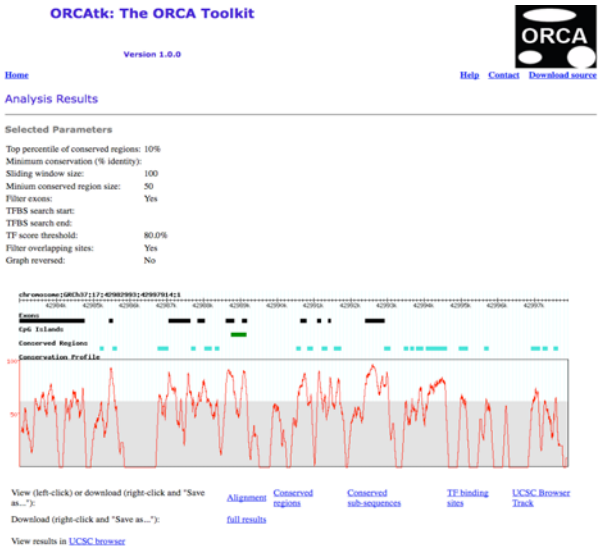
Actin, alpha cardiac



Choosing the ”right” species for pairwise comparison...



ORCAtk

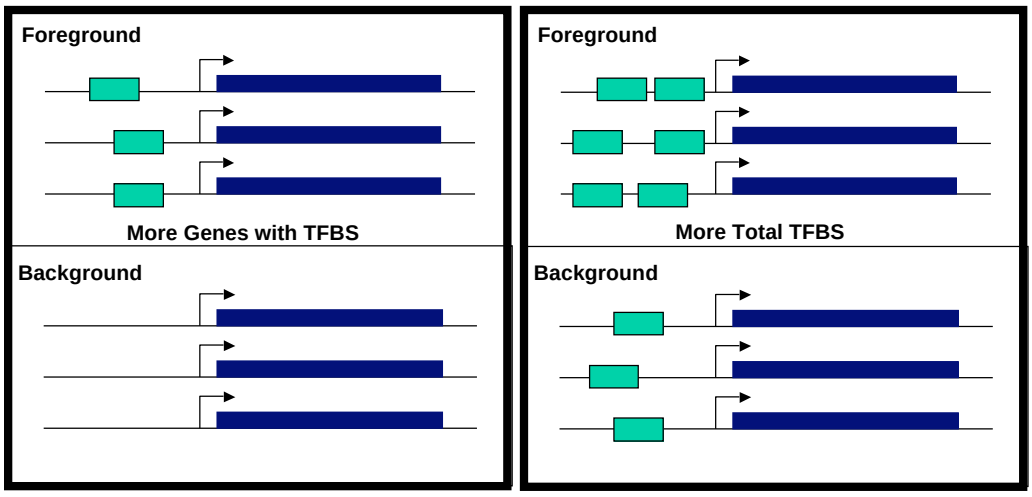


TFBS Discrimination Tools

- Phylogenetic Footprinting Servers
  - FOOTER http://biodev.hgen.pitt.edu/footer_php/Footerv2_0.php
  - CONSITE http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite/
  - rVISTA http://rvista.dcode.org/
  - ORCAtk http://burgundy.cmmt.ubc.ca/cgi-bin/OrcaTK/orcatk
- SNPs in TFBS Analysis
  - RAVEN http://burgundy.cmmt.ubc.ca/cgi-bin/RAVEN/a?rm=home
- Prokaryotes or Yeast
  - PRODORIC http://prodoric.tu-bs.de/
  - YEASTRACT http://www.yeastract.com/index.php
- Software Packages
  - TOUCAN http://homes.esat.kuleuven.be/~saerts/software/toucan.php
- Programming Tools
  - TFBS http://tfbs.genereg.net/
  - ORCAtk http://burgundy.cmmt.ubc.ca/cgi-bin/OrcaTK/orcatk

Part 3:
Inferring Regulating TFs for Sets of Co-Expressed Genes

Two Examples of TFBS Over-Representation

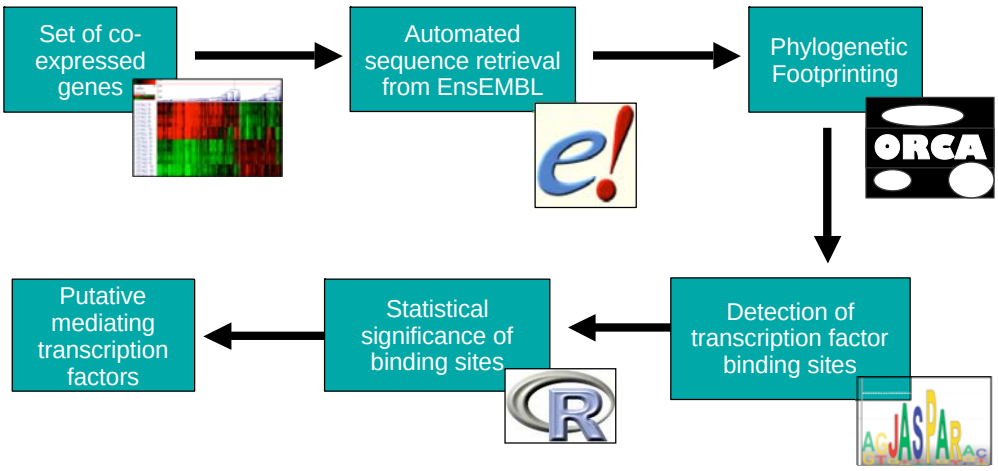


Statistical Methods for Identifying Over-represented TFBS

- Fisher exact probability scores
  - Based on the ***number of genes*** containing the TFBS relative to background
  - Hypergeometric probability distribution

- Binomial test (Z scores)
  - Based on the ***number of occurrences*** of the TFBS relative to background
  - Normalized for sequence length
  - Simple binomial distribution model

oPOSSUM Procedure

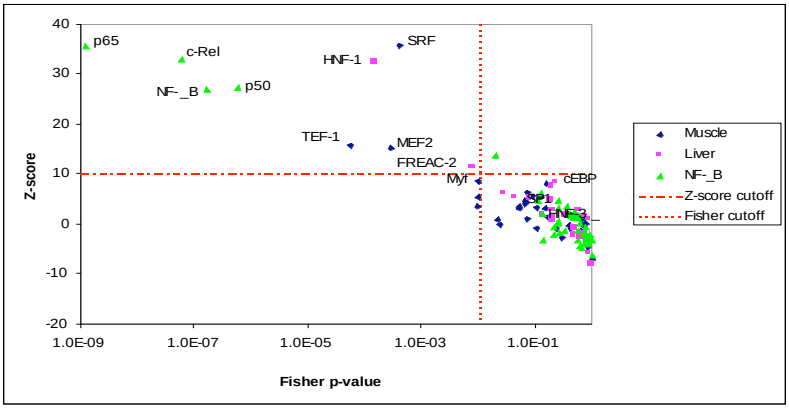


Validation using Reference Gene Sets

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

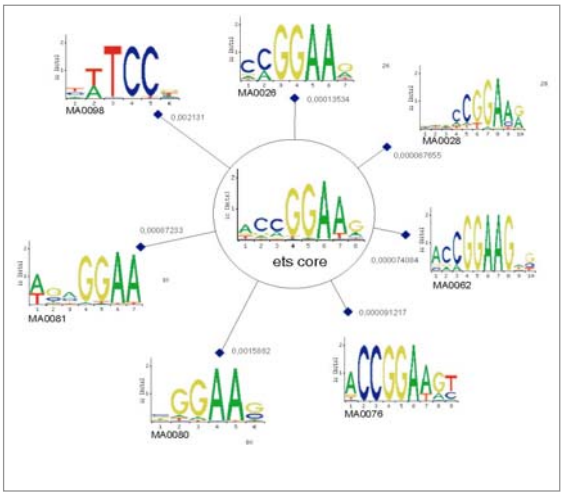
TFs with experimentally-verified sites in the reference sets.

Empirical Selection of Parameters based on Reference Studies

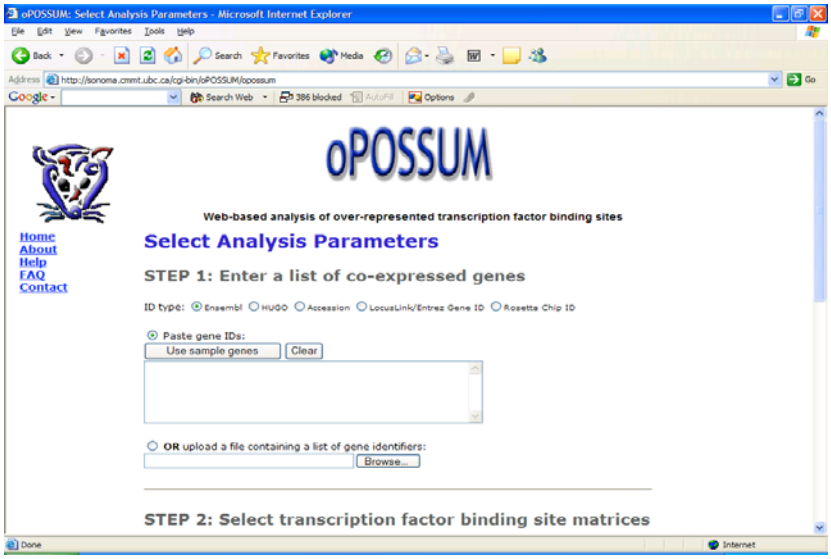


Structurally-related TFs with Indistinguishable TFBS

- Most structurally related TFs bind to highly similar patterns
  - Zn-finger is a big exception



oPOSSUM Server



TFBS Over-representation Analysis Tools

- oPOSSUM: http://www.cisreg.ca/oPOSSUM
- TFM-Explorer: http://bioinfo.lifl.fr/TFME/form
- Asap: http://asap.binf.ku.dk/Asap/Home.html

REFLECTIONS

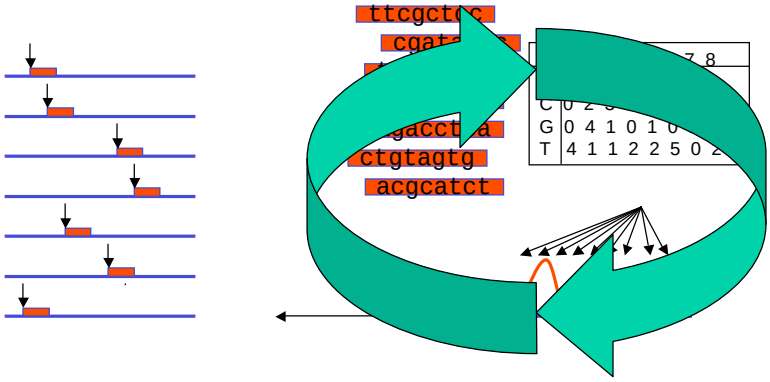
- Part 2
  - Futility Theorem – Essentially predictions of individual TFBS have no relationship to an *in vivo* function
  - Successful bioinformatics methods for site discrimination incorporate additional information (clusters, conservation)
- Part 3
  - TFBS over-representation is a powerful new means to identify TFs likely to contribute to observed patterns of co-expression
  - Generally best performance has been with data directly linked to a transcription factor
  - Statistical significance is extremely sensitive to gene set size
  - TFs in the same structural family tend to have similar binding preferences

The end

More tomorrow in the lab…

Part 4:
de novo Discovery
of TF Binding Sites
(Gibbs sampling method)

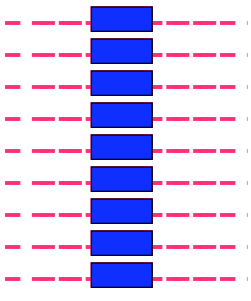
Gibbs Sampling
(grossly over-simplified)



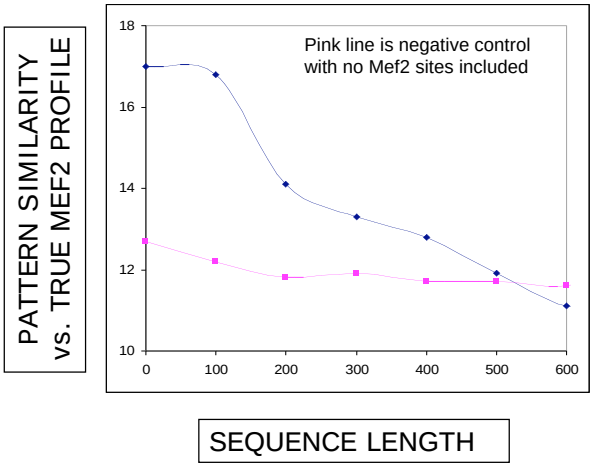
Pattern Discovery

- 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 the mediating TFBS are not strongly over-represented relative to other patterns…

Applied Pattern Discovery is Acutely Sensitive to Noise



True Mef2 Binding Sites



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 over-come the noise