

Analysis of genomic regulatory sequences

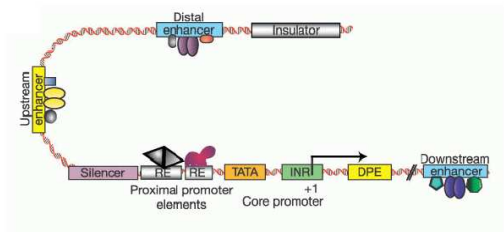
Andrew D Smith

Zhang Lab, CSHL

CSHL Systems Biology Meeting
Pre-meeting Workshop
March 29 2007

Assumed background

- Genes
- Promoters
- Transcription factors (TFs)
- Transcription factor binding sites
- Enhancers
- *cis*-Regulatory modules



(Levine & Tjian, 2003)

Overview

1. How to represent binding sites
2. How to use motifs to predict binding sites
3. Is my motif overrepresented in my sequences?
4. What if I suspect my sequences contain “novel” motifs?
5. When do we want to analyze regulatory sequences?
6. What about finding the functional sites?

Part I

How to represent binding sites

GATCATCATCATTGTGCAGCAGTC**GCCGTCCGCT**GAAAGAGAGAGAACATGACAACGA
ACAACGTACATGATGTGCCAGTC**GCCATCTTG**CACGTTTTTTAACACCGTGCCAAT
CCACGTGACGTAACCTGCATCACA**CCATCTTG**ACACGTGACCCAATATATGGACTT
AGTCTCGACAGCCTTCCCTTCGCG**GCCATTTTG**CAACCATGCACGAATTGAATTAAT
TGCGTATAACCCCATGATGCCCGA**GCCATCATG**GATGACCAACACACACCACACCAG

- Need some way to characterize the sites bound by a TF: Motifs
- Want to describe only important information

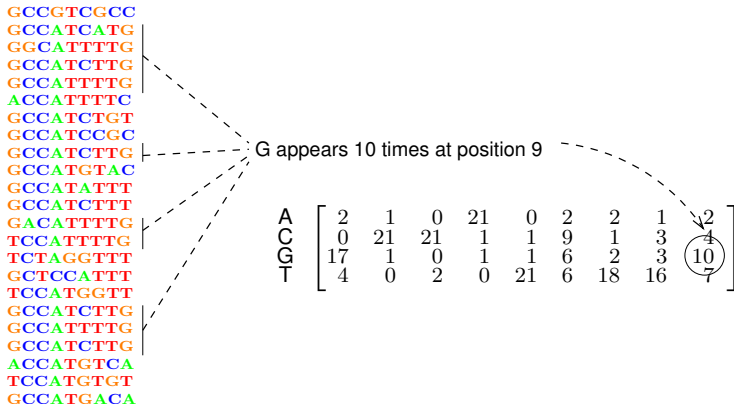
Consensus sequence:

Most frequent
nucleotide at each
position

GCCGTCGCC
GCCATCATG
GGCATTFTG
GCCATCTTG
GCCATTTTG
ACCATTTTC
GCCATCTGT
GCCATCCGC
GCCATCTTG
GCCATGTAC
GCCATATTT
GCCATCTTT
GACATTTTG
TCCATTTTG
TCTAGGTTT
GCTCCATTT
TCCATGGTT
GCCATCTTG
GCCATTTTG
GCCATCTTG
ACCATGTCA
TCCATGTGT
GCCATGACA
GCCATCTTG

Binding sites
(for YY1)

Matrix-based representation



- Matrix columns correspond to positions in sites
- Entries correspond to base counts at the site
- Assumptions: independent positions and no gaps

Sequence Logos

A	2	1	0	21	0	2	2	1	2
C	0	21	21	1	1	9	1	3	4
G	17	1	0	1	1	6	2	3	10
T	4	0	2	0	21	6	18	16	7



- Size of base is proportional to frequency in matrix
- Sometimes sizes are scaled to illustrate important ones

Known motifs

- TRANSFAC and JASPAR: databases of motifs
- TRANSFAC has free versions, JASPAR is free
- Hundreds of known motifs
- Essential resources for regulatory sequence analysis

Part II

How to use motifs to predict binding sites

Motif as a probabilistic model

$$\begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{bmatrix} 2 & 1 & 0 & 21 & 0 & 2 & 2 & 1 & 2 \\ 0 & 21 & 21 & 1 & 1 & 9 & 1 & 3 & 4 \\ 17 & 1 & 0 & 1 & 1 & 6 & 2 & 3 & 10 \\ 4 & 0 & 2 & 0 & 21 & 6 & 18 & 16 & 7 \end{bmatrix}$$



$$\begin{array}{c} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{bmatrix} .09 & .04 & .00 & .91 & .00 & .09 & .09 & .04 & .09 \\ .00 & .91 & .91 & .04 & .04 & .39 & .04 & .13 & .17 \\ .74 & .04 & .00 & .04 & .04 & .26 & .09 & .13 & .43 \\ .17 & .00 & .09 & .00 & .91 & .26 & .78 & .70 & .30 \end{bmatrix}$$

- Normalize each column so it sums to 1
- Think of entries as probabilities

Probability from a motif

	T	C	C	A	G	G	T	T	C
A	.09	.04	.00	.91	.00	.09	.09	.04	.09
C	.00	.91	.91	.04	.04	.39	.04	.13	.17
G	.74	.04	.00	.04	.04	.26	.09	.13	.43
T	.17	.00	.09	.00	.91	.26	.78	.70	.30

$.17 \times .91 \times .91 \times .91 \times .04 \times .26 \times .78 \times .70 \times .17$

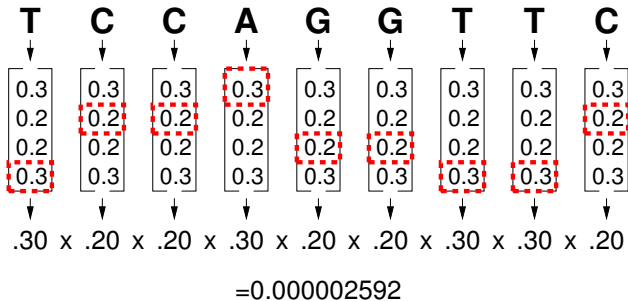
$=0.0001236653$

- Probability of sequence from matrix: multiply over positions
- Can do this because positions are independent
- $\Pr(\text{TCCAGGTTTC}) = 0.00012\dots$ but does that mean anything?

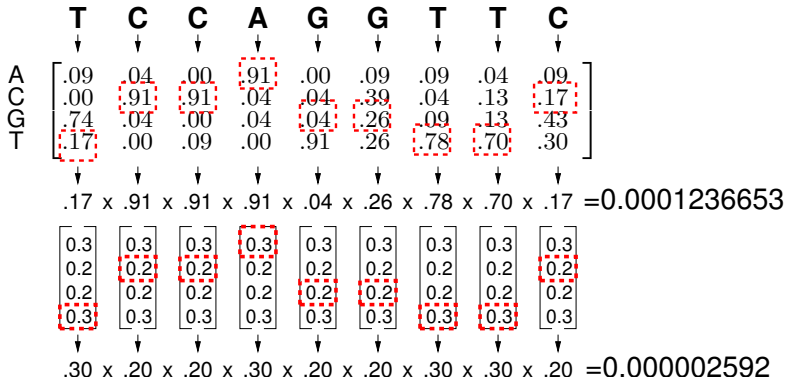
The alternative? Genomic frequencies

Use genomic base frequencies
for probability of any base at any position:

A	0.3
C	0.2
G	0.2
T	0.3



Was it more likely from motif or base composition?



- Ratio of the likelihoods: $0.0001236653/0.000002592 \approx 47.7$
- If we took the log of this:
 - positive score means more likely from motif
 - negative score means more likely from base composition

Making a scoring matrix

A	0	1	0	0	0.1	0
C	1	0	0.5	0	0	0
G	0	0	0.5	1	0	1
T	0	0	0	0	0.9	0

A	-4.96	1.69	-4.96	-4.96	-1.50	-4.96
C	2.28	-4.38	1.29	-4.38	-4.38	-4.38
G	-4.38	-4.38	1.29	2.28	-4.38	2.28
T	-4.96	-4.96	-4.96	-4.96	1.54	-4.96

A	0.3
C	0.2
G	0.2
T	0.3

$$\log \left(\frac{\text{matrix frequency}}{\text{background frequency}} \right) = \log \left(\frac{0.9}{0.3} \right) = 1.54$$

- Need **pseudocount** to prevent log of 0
- Take **sum** over positions instead of product

Scanning a sequence

<i>A</i>	-4.8	1.8	-3.0	-0.26	-7.1	-7.1
<i>C</i>	1.9	-2.1	1.6	0.37	-7.1	-3.4
<i>G</i>	-7.1	-3.9	-0.4	0.76	-7.1	1.9
<i>T</i>	-3.9	-2.5	-2.5	-2.49	2.0	-4.8

TGCTGTAAGCCAGCCTGTGGTGGCCTG**CAGCTG**CTGAACACTCTGTTGCTGTAAGC'

$$\text{match score} = 1.9 + 1.8 - 0.4 + 0.37 + 2.0 + 1.9 = 7.57$$

- Scan the sequence, find:
 - best scoring matches
 - matches that score above some cutoff
- Can measure significance of scores under various assumptions

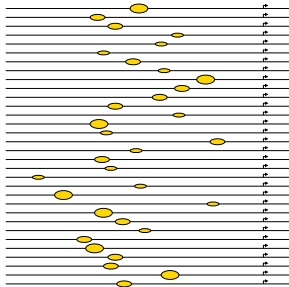
High false-positive rate when predicting sites this way!
(more on this later)

Part III

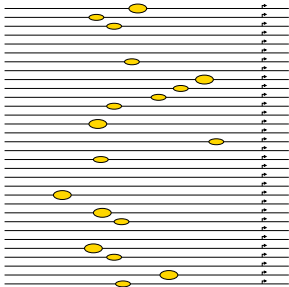
Is my motif overrepresented in my sequences?

Measuring enrichment in sequence sets

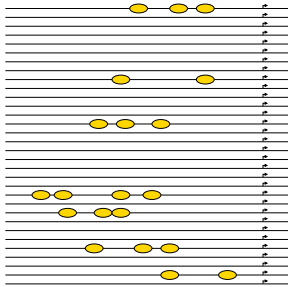
Exactly one occurrence
per sequence



At most one per sequence

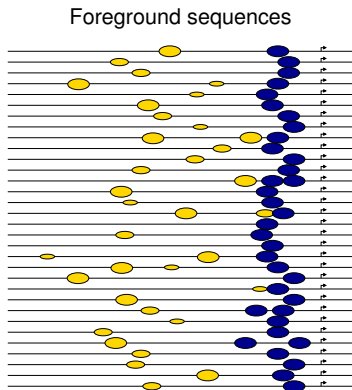


Any number per sequence



- Things we like to see:
 - more occurrences
 - stronger occurrences (higher scoring)
 - more sequences that contain an occurrence
- Different assumptions valid for different TFs

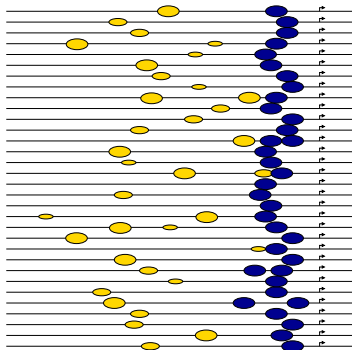
Using background sequences



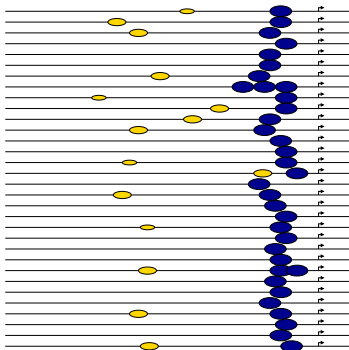
- Statistical models of “random” promoters don’t work
- Using a background can control many unknown variables
- Background should be selected carefully

Using background sequences

Foreground sequences



Background sequences



- Statistical models of “random” promoters don’t work
- Using a background can control many unknown variables
- Background should be selected carefully

Part IV

What if I suspect my sequences contain
“novel” motifs?

Motif discovery by word counting

Table of words and their occurrences

AAAAA	521
AAAAC	534
AAAAG	243
AAAAT	847
AAACA	366
AAACC	504
GAGGT	622
GAGTA	718
GAGTC	???
GAGTG	
GAGTT	
TTTGA	
TTTTA	
TTTTC	
TTTTG	
TTTTT	

For each word of width k:
count number of occurrences
Apply statistics to counts

current word

GAGTC

AAGTCTACATGAGATCGATGGTTTCTTGGAGCTTCCACAAACTTAAACCATGAAACATCTATTATTGCTACTATTGTTA
 TAAATAAATTCATCTGATCAAAAGAAATTTAAAAACCAACSAACCCTAATGAGCTCTAAAGACAGCAGAGTCACACGCGA
 AGGAGCGCGCCCTTACCCCTCCGGCCTCAGCCCGCAGGGCTGCAACCCTTTCCGCACCTGGCTCCATCTCCCTGGCCCTC
 GGAGCGAGAAGGCGGCGGGGATCTGGCGCCGGCTTAGGGGGCAGACGGCCGACCCGGGAGCCTAGCGATCAGGGCCACC
 GCCACGCGCCGTGAGCCCGCCCAACATAGCCAGGAGTCGCTTCGGGTGATAAGCGTCCCGGTGGCGGAGGCCGCA
 AGAAGGGTGCCCTGTCTTGGAGTCCCTTTTTCAGCCACTCAGATGTGCTGCTGCGGTGTCTTTTGTGCTGGTGGCAGCC
 AGCCGTTCCAGCTTGACTTTCCCTTTAGCCTAGTGATTTGGGGCCCAAGGTTTATTTCTTTCCGCTAGCTTCGC
 TGTGTCTGGTGTCTTCTCTCTCAGCCTGTTTCTCATCTGGAAACATGAGGTGTCTGGCGAGGGCCGATAGCCGATG
 GGTGGGGTGGGAGGAAACCTTATCTGTGCCGATGGCCCTCGTTGTGAGTCTATTAAACTCTGGGAAACTGCTAT
 AAGACCCTGAGAAGCAAATCTTTAATTTTTTGTGAGACGGAGCACTCTGTCCAGGCTAGAGTGAATTAG
 GGTGCAATCTCGGCTCACTGGAACCTCCGCTCTCTGAGTCCATAGCGATTCTCTGCCTCAGCCTCCCGAGTAGCTGGTTA
 AGTAGAGACTGGATCACCATGTTGGCCAGGCTGTCTCGAACTCCTGACCCCAAGTATCCACCTGCCTCAGCCTCTT
 AAGTCTACATGAAAAGGATGGTTTCTTGGAGCTTCCACAAACTTAAAAATGGATTCAACATCTATTATTGCTACTATTGT
 TCTCCCGAGCAGGGCCCCAGCGGCACCATGTCTATGGATTCCGGAGTCCAGCTGGCCCTGCTCG

Word-based methods

- Very sophisticated word-based algorithms exist
- Words need not be exact
 - Counting approximate matches
 - Use wildcard characters
- Best used as initial “stage” in algorithms that eventually produce matrix-based motifs

Gibbs Sampling

Start with a given motif
and a set of occurrences

GCCATCTTT
GACATTTTG
TCCATTTTG
TCTAGGTTT
GCTCCATTT
TCCATGGTT
GCCATCTTG
GCCATTTTG
GCCATCTTG
ACCATGTCA
GCCATGACA
TCCATGTGT



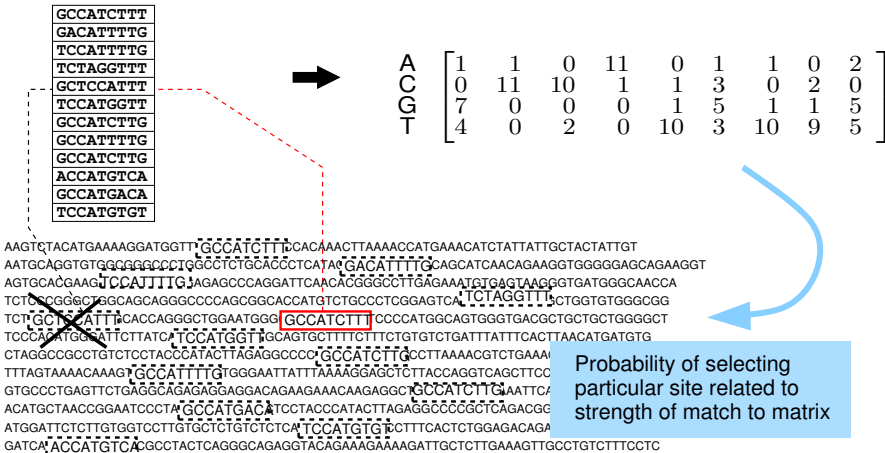
A	1	1	0	11	0	1	1	0	2
C	0	11	10	1	1	3	0	2	0
G	7	0	0	0	1	5	1	1	5
T	4	0	2	0	10	3	10	9	5

AAGTCTACATGAAAAGGATGGTTGCCATCTTTCCACAAACTTAAAACCATGAAACATCTATTATTGCTACTATTGT
AATGCAGGTGTGGCGGGCCCTGGCCTCTGCACCCCTCATAGACATTTTGCAGCATCAACAGAAGGTGGGGGAGCAGAAGGT
AGTGACGAAAGTCCAJTTTIGAGAGCCAGGATCAACACGGGCCTTGAGAAATGTGAGTAAGGGTATGGGCAACCA
TCTCCCGGGCTGGCAGCAGGGCCCCAGCGGCACCATGTCTGCCCTCGGAGTCACTCTAGGTTTCTGGTGTGGGCGG
TCTGCTCCATTTCCACCAGGGCTGGAATGGGGCCGCCCGCTCCCCATGGCAGTGGGTGACGCTGCTGCTGGGGCT
TCCCACATGGGATTCATTATCTCCATGGTTGCAGTGCTTTTCTTTCTGTGTCTGATTATTTCACTTAACATGATGTG
CTAGGCGCCGTGTCTCCTACCCTACTTAGAGGCCCCGCCATCTTGCCTTAAAACGCTGAAAGGCCGTTCTGCGCA
TTTAGTAAAACAAGTGCCATTTTGTGGGAATTATTTAAAAGGAGCTTTACCAGGTCAGCTTCCTTCGGTGTGCGG
GTGCCCTGAGTCTGAGGCAGAGAGGAGGACAGAAGAAACAAGAGGCTGCCATCTTGAATTTCAGTATCCCAGTTG
ACATGCTAACCGGAATCCCTAGCCATGACATCCTACCATACTTAGAGGCCCGCTCAGACGGTCTTAAAACGCTC
ATGGATTCTCTGTGGTCTGTGCTCTGTCTCTCACTCCATGTGTCTTTCACTCTGGAGACAGAGCTCTGGGAG
GATCAACCATGTCAAGCCTACTCAGGCGAGAGGTACAGAAAGAAAAGATTGCTCTTGAAAGTTGCTGTCTTTCTCT

Gibbs Sampling

Iterate these steps:

1) Sample a new occurrence from one sequence



Gibbs Sampling

Iterate these steps:

- 1) Sample a new occurrence from one sequence
- 2) Update the matrix based on new occurrence

GCCATCTTT
GACATTTTG
TCCATTTTG
TCTAGGTTT
GCTCCATTT
TCCATGGTT
GCCATCTTG
GCCATTTTG
GCCATCTTG
ACCATGTCA
GCCATGACA
TCCATGTGT

A
C
G
T

1	1	0	11 ¹²	0	1 ⁰	1	0	2
0	11	10 ¹¹	1 ⁰	1 ⁰	3 ⁴	0	2	0
7	0	0	0	1	5	4	1	1
4	0	2 ¹	0	10 ¹¹	3	10	9	5

Usually the changes will move matrix toward stronger motif

AAGTCTACATGAAAAGGATGGTTGCCATCTTTCCACAACCTTAAAACCATGAAACATCTATTATTGCTAC
 AATGCAGGTGTGGCGGCCCTGGCCCTCTGCACCCTCATACGACATTTTGCCAGCATCAACAGAAGGTGC
 AGTGACACGAAGTCCAJJJJJAGAGCCCAGGATTCACACGGGCCCTTGAGAAATGTGAGTAAGGGTGTAGGGCAACCA
 TCTCCCGGGCTGGCAGCAGGGCCCCAGCGGCACCATGTCTGCCCTCGGAGTCACTCTAGGTTTCTGGTGTGGGCGG
 TCTGCTCCATTTCCACCAGGCTGGAATGGGGCCATCTTTCCCCATGGCAGTGGGTGACGCTGCTGCTGGGGCT
 TCCCACATGGGATTCATTATCTCCATGGTTGCAGTGTCTTTCTTCTGTGTCTGATTATTTCACTTAACATGATGTG
 CTAGGCCGCTGTCTCCTACCCTAFACTTAGAGGCCCCGCCATCTTGCCTTAAAACGCTGAAAGGCCGCTTCTGCCA
 TTTAGTAAAACAAGTGGCCATTTTGGGAATTATTTAAAAGGAGCTTTACCAGGTCAGCTTCCTCGGTGTTGCCGG
 GTGCCCTGAGTCTGAGGCAGAGAGGAGGACAGAAGAAACAAGAGGCTGCCATCTTTGAATTCAGTATCCCAGTTG
 ACATGCTAACCGGAATCCCTAGCCATGACATCCTACCCATACTTAGAGGCCCCGCTCAGACGCTCCTAAAACGCTC
 ATGGATTCTCTGTGGTCTGTGCTCTGTCTCTCACTCCATGTGTCTTTCACTCTGGAGACAGAGCTCTGGGAG
 GATCAACCATGTCACGCCTACTCAGGCGAGAGGTACAGAAAGAAAAGATTGCTCTTGAAAGTTGCTGTCTTCTCTC

Expectation Maximization

- Instead of sampling sites with particular probability:
 - All possible sites contribute to the matrix
 - Contribution of each site related to probability (score)
- Like deterministic version of Gibbs: no random choices

Expectation Maximization

- Instead of sampling sites with particular probability:
 - All possible sites contribute to the matrix
 - Contribution of each site related to probability (score)
- Like deterministic version of Gibbs: no random choices

Variants of EM or Gibbs

- Gibbs Motif Sampler (Lawrence et al., 1993)
- MEME (Bailey & Elkan, 1995)
- AlignACE (Hughes et al., 2000)
- MDscan (Liu et al., 2002)

Good starting points are **critical** for Gibbs and EM

De novo motif discovery

Current status

- Field starting to mature: many great algorithms exist!
- Probably none will be “perfect” for your application
- Try several algorithms
- Understand what they do, and post-process results

De novo motif discovery

Current status

- Field starting to mature: many great algorithms exist!
- Probably none will be “perfect” for your application
- Try several algorithms
- Understand what they do, and post-process results

How to improve

- Combine best aspects of different algorithms
- No single algorithm universally appropriate
- Incorporate more biological knowledge

De novo motif discovery

Current status

- Field starting to mature: many great algorithms exist!
- Probably none will be “perfect” for your application
- Try several algorithms
- Understand what they do, and post-process results

How to improve

- Combine best aspects of different algorithms
- No single algorithm universally appropriate
- Incorporate more biological knowledge

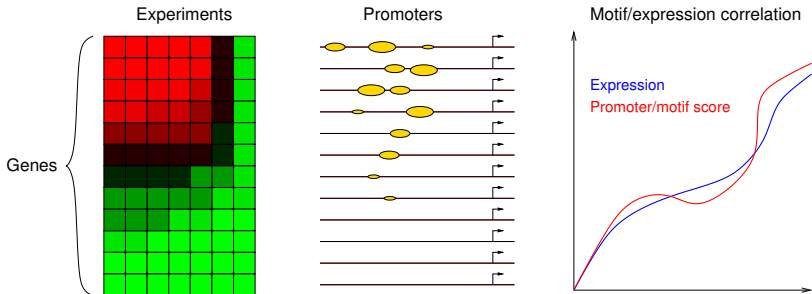
DME: Discriminating Motif Enumerator

- Enumerative search strategy, matrix-based motifs
- Smith et al. (2005)

Part V

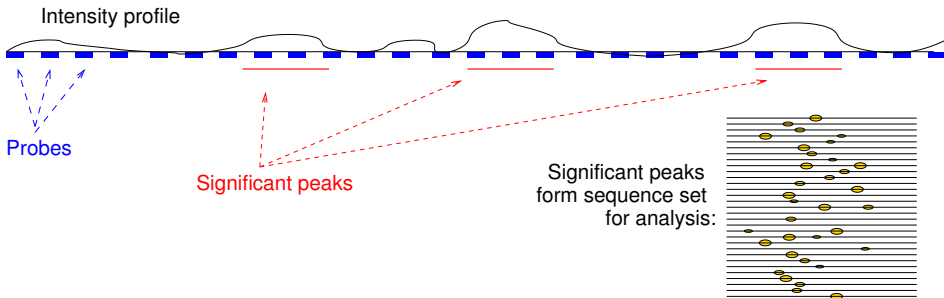
When do we want to analyze regulatory sequences?

Understand regulation of gene expression



- Expression \Rightarrow co-regulated genes \Rightarrow similar regulatory elements
- Major goal: predict expression from sequence
- Sequence analysis can identify:
 - TFs functioning in the conditions
 - Interactions between TFs in the conditions
 - Binding sites of those TFs
 - Direct vs indirect targets

ChIP-chip data

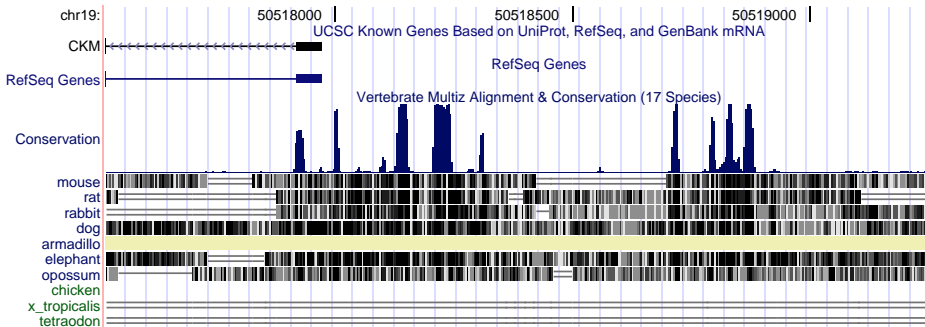


- Similar to expression case, but different...
- Try to predict binding (or intensity) using sequence
- What we can get:
 - Characterize motif if not already known
 - Identify precise binding sites (inside peak regions)
 - Infer co-factors (more on this later)
 - Quality control

Part VI

What about finding the functional sites?

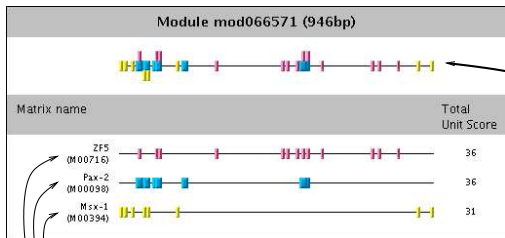
Using cross-species conservation



- Easiest: use phastCons profile or “conserved elements”
- Whole-genome alignments are good (especially close to genes)
- Rapid progress in this area
- Expect interesting results at the meeting!

Distal regulatory regions

PReMod (Blanchette et al, 2006)

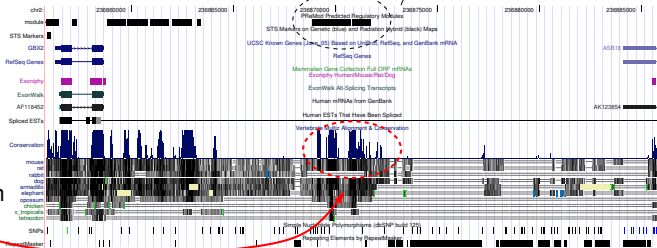


Occurrences tightly clustered

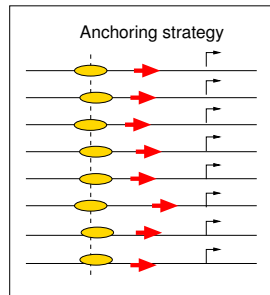
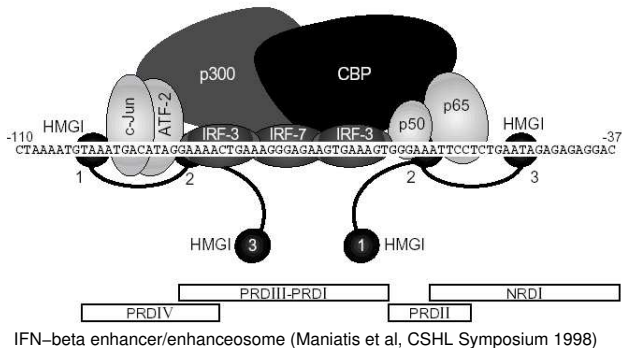
Far from gene

Strong occurrences of known motifs

Highly conserved region



cis-Regulatory modules



- Sets of interacting sites
- Two things to look for
 - Interaction can better predict expression
 - Relative positional preference of sites

Acknowledgments

Workshop organizers:
Scott Tenenbaum & Tristan Fiedler

Meeting organizers:
Julia Bailey-Serres, Nir Friedman and Bing Ren