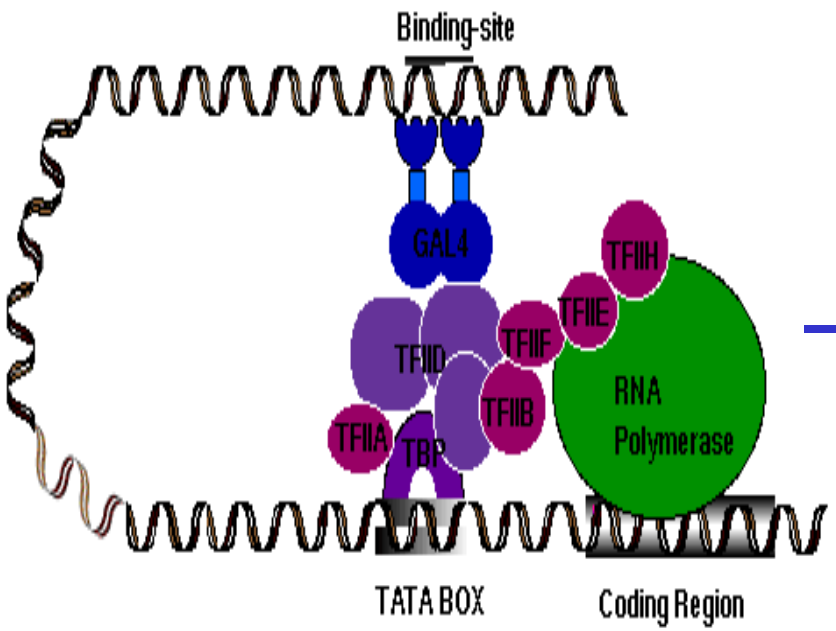
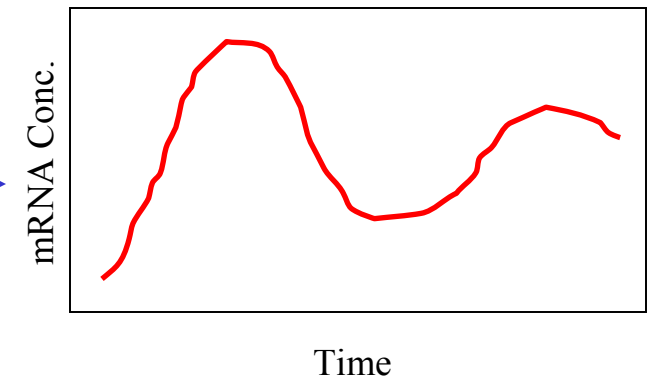
Motif Finding: Summary of Approaches

Lecture Outline

- Flashback: Gene regulation, the cis-region, and tying function to sequence
- Motivation
- Representation
  - simple motifs
  - weight matrices
- Problem: Finding motifs in sequences
- Approaches
  - enumerative (combinatorial)
  - statistical
- Comparison of approaches
- Higher Order Motifs and Approaches

Gene Regulation



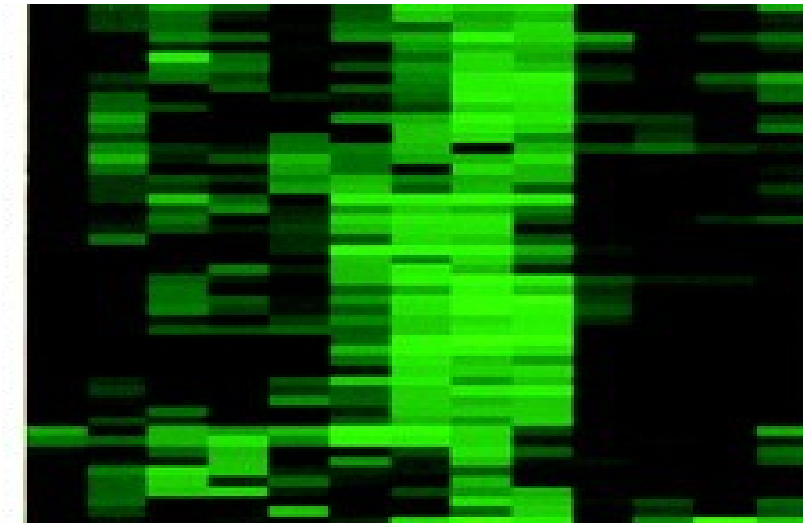


Sequence

Function

Motif Finding Motivation

Clustering genes based on their expressions groups co-expressed genes



Assuming co-expressed genes are co-regulated, we look in their promoter regions to find <u>conserved motifs,</u> confirming that the same TF binds to them

Motifs vs Transcription Factor Binding Sites

- Motifs:
  - statistical or computational entities
  - predicted
- Transcription Factor Binding Sites (or more generally cis-regulatory elements)
  - biological entities
  - Real
- The hope is that TFBS are conserved, or otherwise significant computationally, so motifs can be used to find them

Finding Motifs in a Set of Sequences

```
GTGGCTGCACCACGTGTATGC...ACGATGTCTC
ACATCGCATCACGTGACCAGT...GACATGGACG
CCTCGCACGTGGTGGTACAGT...AACATGACTA
CTCGTTAGGACCATCACGTGA...ACAATGAGAG
GCTAGCCCACGTGGATCTTGT...AGAATGGCCT
```

Finding Motifs in a Set of Sequences

```
GGCTGCACCACGTGTATGC...ACGATGTCTCGC
ATCGCATCACGTGACCAGT...GACATGGACGGC
TCGCACGTGGTGGTACAGT...AACATGACTAAA
CGTTAGGACCATCACGTGA...ACAATGAGAGCG
TAGCCCACGTGGATCTTGT...AGAATGGCCTAT
```

Finding Motifs in a Set of Sequences

```
TCTGCACCACGTGTATGC...ACGATGTCTCGC
ATCGCATCACGTGACCAGT...GACATGGACGGC
 GCCTCGCACGTGGTGGTACAGT...AACATGAC
GGACCATCACGTGA...ACAATGAGAGCG
 GCTAGCCCACGTGGATCTTGT...AGAATGGCC
```

Protein binding

Phylogenetic Footprinting

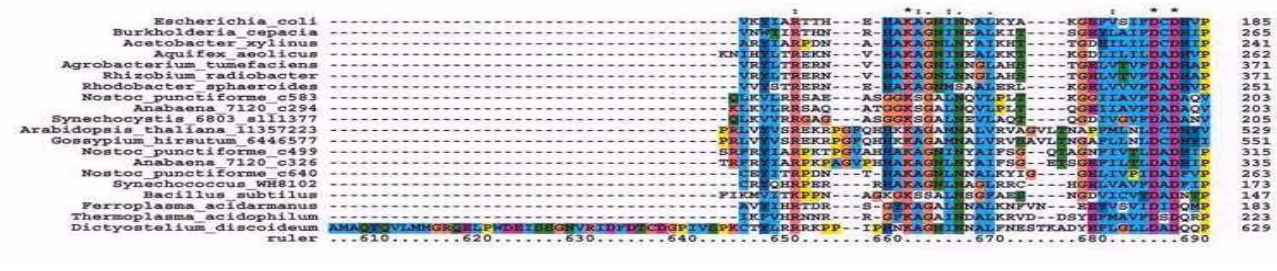
- Finding overrepresented short sequences in cis-regions
- Based on multiple alignment but short sequences don't have to be completely conserved
- Ex. FootPrinter (Blanchette and Tompa 2003)

Motif Finding Problem

Given n sequences, find a motif (or subsequence) present in many

This is essentially multiple alignment. The difference is that multiple alignment is global

- longer overlaps
- constant site sizes and gaps
- NP-complete!



Definition and Representation

- Motifs: Short sequences
- IUPAC notation
- <u>Regular Expressions</u>
  - consensus motif

    ACGGGTA
  - degenerate motif

    RCGGGTM

  {G|A}CGGGT{A|C}

Single-Letter Codes for Nucleotides	
Symbol	**Meaning**
G	G
A	A
T	T or U
C	C
U	U or T
R	G or A
Y	T, U or C
M	A or C
K	G, T or U
S	G or C
W	A, T or U
H	A, C, T or U
B	G, T, U or C
V	G, C or A
D	G, A, T or U
N	G, A, T, U or C

Position Specific Information

Seqs.

ACGGG

ATCGT

AAACC

TTAGC

ATGCC

Alignment Matrix (Profile)

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

Position (Frequency) Weight Matrix

Pos	A	C	G	T	Conse
1	0.8	0	0	0.2	A
2	0.2	0.2	0	0.6	T
3	0.4	0.2	0.4	0	A|G
4	0	0.4	0.6	0	G
5	0	0.6	0.2	0.2	C

1. Use PWM to Find the Motif in any Sequence

Frequency Weight Matrix

Pos	A	C	G	T	Conse
1	0.8	0	0	0.2	A
2	0.2	0.2	0	0.6	T
3	0.4	0.2	0.4	0	A|G
4	0	0.4	0.6	0	G
5	0	0.6	0.2	0.2	C

Given AAATC and the Weight Matrix of the data and for the background (i.e. prior), we want to calculate the joint probability

In general this is a lot of work, because of all possible ways a motif can depend on its sub-words.

E.g. TATTA=TAT.TA|TA.T.TA|T.A.T.T.A, etc.

2. Given Sequences Find Motifs

- Methods based on Position Weight Matrices (alignment)
  - Gibbs Sampling
  - Expectation Maximization
- Other Methods
  - HMMs
  - Bayesian methods
  - enumerative (combinatorial)

Simple Motif Finding

- Methods based on Position Weight Matrices (alignment)
  - Gibbs Sampling
  - Expectation Maximization
- Other Methods
  - HMMs
  - Bayesian methods
  - enumerative (combinatorial)

Popular Software:

- MEME (EM)

  http://meme.sdsc.edu/meme/website/intro.html
- AlignACE (Gibbs)

  http://atlas.med.harvard.edu/
- Cister (HMM)

  http://zlab.bu.edu/~mfrith/cister.shtml
- YMF (combinatorial)

  http://www.cs.washington.edu/homes/blanchem/software.html
- MITRA (combinatorial)
  http://www.cs.columbia.edu/compbio/mitra/
- NestedMICA
  http://www.sanger.ac.uk/Software/analysis/nmica

Overall Idea

- Enumerate motifs
- Score motifs based on their over-representation in all sequences
- The highest scoring ones, if occurring at <u>surprising</u> rates, are meaningful

Problems:

- How to enumerate?
- How to score motifs?
- What is surprise?

Using PWMs, main idea

- Capture the data in PWM
- Enumerate and score <u>all patterns</u>, w
  - suffix trees used to save space
- Update the PWM
- Scoring: <u>over-representation</u>

S=observed frequency/expected frequency

w in given sequences

w in genome

MEME

- Use Expectation-Maximization Algorithm to fit a two-component mixture model to the sequence data
- Component 1 is the motif
- Component 2 is the background

Algorithm:

- For each sequence s_i, (out of n)
- Start with a random PWM, P_i (i.e. alignment)
- Score every segment of s_i with P_i
- Update P_i=Sum all the scores with appropriate weights
- Perform EM until there is a convergence

The best 100 scoring motifs are kept overall

Gibbs Sampler

- Use a simple leave-one-out sampling strategy

Algorithm

- Given n sequences, s1, s2,...,sn
- Randomly initialize PWM (i.e. align)
- For each sequence s_i, take it out from the PWM

  - score each segment of s_i with the rest of the sequences

  - put the sequence back
- Important feature: convergence

YMF: Enumeration

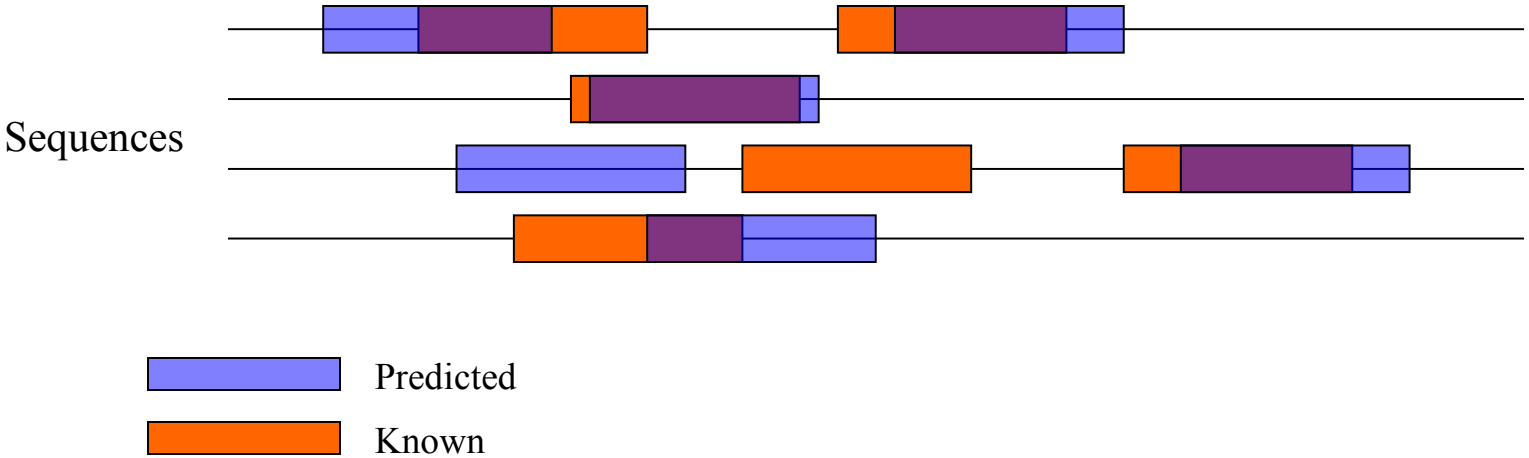
- Use a consensus model of motifs based on IUPAC alphabet
- Score motifs based on their significance of occurrence (vs. random)
- Clean up the found motifs to remove redundant motifs

Comparing the Methods

Tompa et al. (2005)

- Compared 13 different methods
- Used real sequences and searched for known binding sites (TRANSFAC)
  - 52 data sets + 4 negative controls
  - 4 organisms represented (fly, human, mouse and yeast)
- Scored methods based on confusion matrix statistics for the top motif observed

1. Assessing Method Performance



Score = Total overlap / Total span (Pevzner & Sze 2000)

Score = 1, if span = overlap
Score = 0, if overlap = 0

2. Comparing Binary Predictors

Comparing a Model to Reality:

Confusion matrix

Actual (reality)		
	TP	FN Type II error
	FP Type I error	TN

Model Predictions (Y,N)

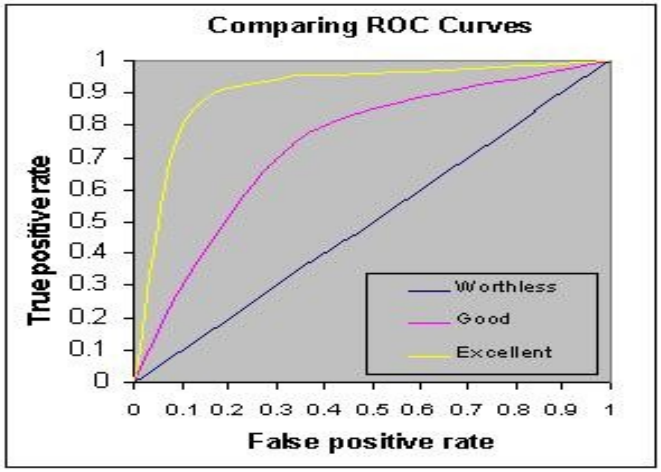
Measures of agreement:

- True Positives, True Negatives
- False Positives, False Negatives

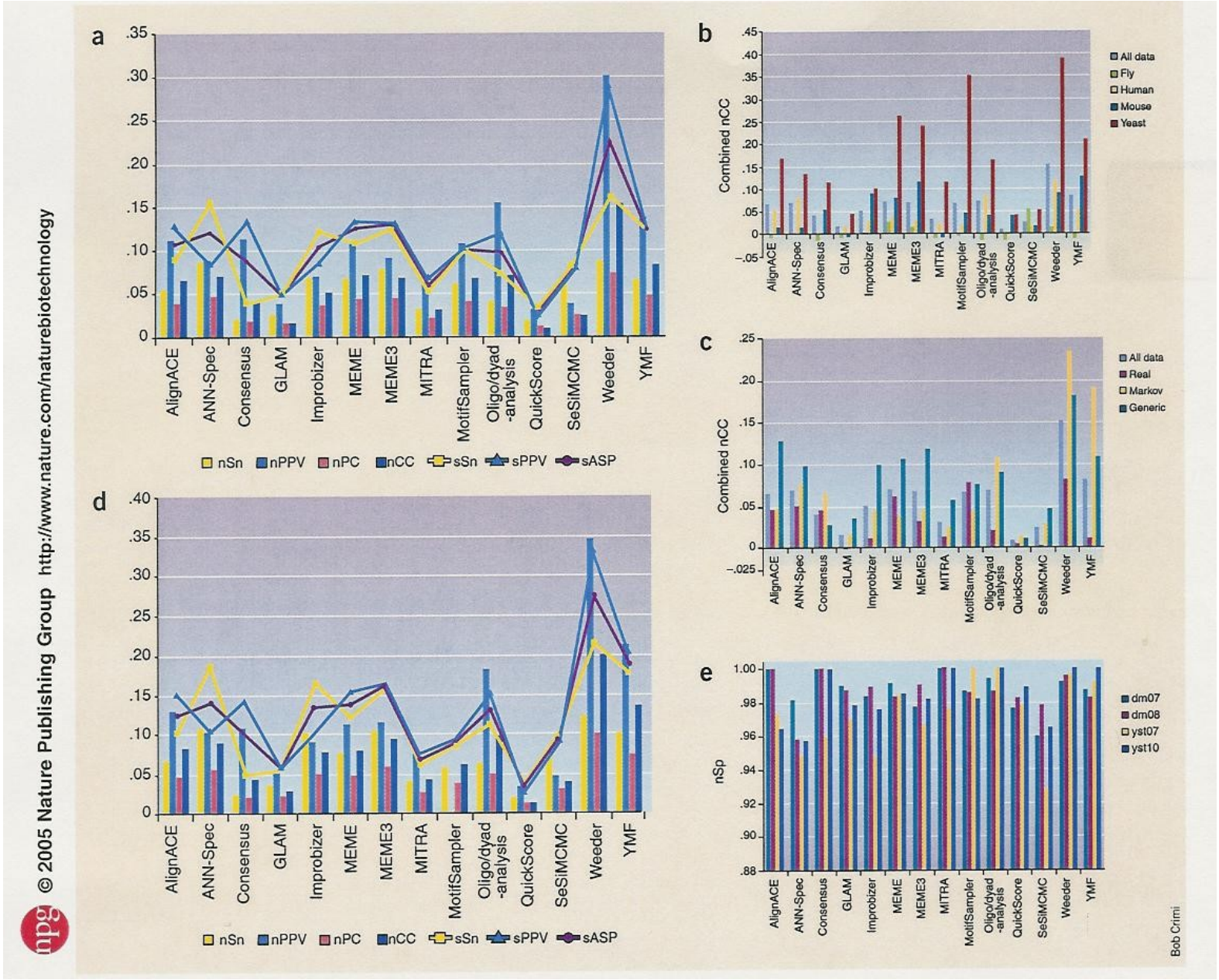
Measures of accuracy:

- Accuracy = (TP+TN)/(TP+TN+FP+FN)
- Sensitivity = TP/(TP+FN)
- PPV=TP/(TP+FP)
- Specificity = TN/(FP+TN)
- Correlation coefficient:

$$\frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}$$



ROC Curves: Tradeoff between Sens. And Specificity (sens vs. 1-spec.)



© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

Table 2 Number of data sets for which each tool predicted no motif[a]

Tool	Total (56)	Fly (8)	Mouse (12)	Human (26)	Yeast (10)
AlignACE	32	7	5	17	3
ANN-Spec	3	1	0	1	1
Consensus	37	4	3	26	4
GLAM	3	0	1	2	0
Improbizer	0	0	0	0	0
MEME	6	1	2	2	1
MEME3	14	0	5	8	1
QuickScore	20	2	4	14	0
SeSiMCMC	0	0	0	0	0
MITRA	11	7	3	0	1
MotifSampler	7	2	2	0	3
Oligo/dyad-analysis	23	1	5	13	4
Weeder	17	3	3	10	1
YMF	7	0	2	4	1

[a]The total number of data sets is given parenthetically in the column header.

Weeder's success is due to judicious choices regarding when to predict no motif in a data set: Weeder was run in a 'cautious mode,' where only the strongest motifs were reported. A few small exceptions to Weeder's domination are shown in **Figure 1b**, where SeSiMCMC did somewhat better on the fly data sets, and MEME3 and YMF somewhat better on the mouse data sets.

What is most striking about **Figure 1b** is the fact that so many tools perform much better on the yeast data sets than on other species. This suggests that computational biologists have been more successful at modeling binding sites in yeast than in metazoans. Little significance should be read into the slightly negative *nCC* values in **Figure 1b**: these are so close to zero that they should be interpreted simply as no correlation between the known and predicted binding sites.

Although the shapes of the curves are very similar in **Figure 1a** and **Figure 1d**, the scale is different. Nearly all tools performed better according to most of the seven measures when the data sets of type real were removed. For example, the correlation coefficient *nCC*, averaged over all tools, improved by 39% from **Figure 1a** to **Figure 1d**. This seems to say more about the experimental design than about the tools themselves: it is likely that the data sets of type real contain functional motifs other than the single TRANSFAC binding site on which they were scored, and that tools that discovered other functional motifs were unduly penalized. The tool most affected by this is YMF, whose seven measures each improved between 45% and 67% when the real data sets were removed. Interestingly, there is one tool that did not improve by this removal: MotifSampler's performance was somewhat better on the data sets of type real than on the others. This aspect of MotifSampler can also be seen in **Figure 1c** for the measure *nCC*.

We have not discovered any simple feature, such as type of motif search, that determines the accuracy of these tools. Nor should we expect such a simple conclusion: the tools are based on algorithms and motif models that are varied and complex, and predicting their performance on complex data is beyond our current analytical ability.

Table 3 shows some very interesting complementary behaviors among certain pairs of tools. For example, MotifSampler's predictions complement well the predictions of MEME, oligo/dyad-analysis, ANN-Spec and YMF, improving their individual *nCC* scores by 64–92%. It is also informative to see that MEME's predictions improve the individual *nCC* score of MEME3 by 53%. This gives some idea of the improvement possible by allowing a given tool to predict two motifs rather than just one.

Exploiting comparative sequence analysis, using tools not covered in this assessment, provides a powerful adjunct to these methods. As an example, a recent tool called PhyME that combines intraspecies overrepresentation and interspecies conservation reported success[23] in predicting the binding sites for one of the most difficult human data

Table 3 Correlation coefficient (*nCC*) for all pairs of tools[a]

	Quick score	GLAM	SeSi MCMC	MITRA	Consen	Improb	Align ACE	Motif sampler	MEME3	MEME	Oligo/ dyad	ANN-Spec	YMF	Weeder
QuickScore	0.009	0.020	0.042	0.030	0.025	0.052	0.068	0.072	0.072	0.074	0.038	0.064	0.061	0.084
GLAM	0.031	0.016	0.060	0.037	0.039	0.068	0.066	0.084	0.088	0.086	0.052	0.082	0.090	0.113
SeSiMCMC	0.049	0.059	0.024	0.068	0.042	0.083	0.071	0.091	0.081	0.088	0.058	0.103	0.104	0.092
MITRA	0.042	0.041	0.072	0.031	0.054	0.082	0.084	0.097	0.106	0.105	0.070	0.101	0.103	0.131
Consensus	0.067	0.060	0.075	0.053	0.042	0.077	0.079	0.109	0.084	0.077	0.074	0.082	0.081	0.098
Improbizer	0.065	0.069	0.083	0.077	0.056	0.052	0.089	0.117	0.096	0.098	0.083	0.112	0.091	0.117
AlignACE	0.088	0.084	0.089	0.090	0.085	0.111	0.068	0.097	0.102	0.091	0.088	0.091	0.115	0.119
MotifSampler	0.071	0.092	0.107	0.097	0.077	0.103	0.099	0.068	0.112	0.119	0.103	0.127	0.130	0.134
MEME3	0.089	0.094	0.092	0.102	0.074	0.102	0.093	0.124	0.069	0.106	0.094	0.129	0.126	0.114
MEME	0.091	0.090	0.100	0.102	0.077	0.091	0.095	0.120	0.100	0.073	0.104	0.123	0.121	0.121
Oligo/dyad	0.073	0.088	0.111	0.088	0.082	0.082	0.099	0.136	0.119	0.112	0.071	0.106	0.107	0.130
ANN-Spec	0.085	0.091	0.111	0.094	0.090	0.100	0.085	0.122	0.114	0.110	0.089	0.074	0.118	0.117
YMF	0.094	0.095	0.112	0.101	0.093	0.100	0.114	0.146	0.121	0.129	0.092	0.131	0.084	0.137
Weeder	0.164	0.169	0.162	0.167	0.157	0.171	0.166	0.186	0.168	0.164	0.173	0.167	0.167	0.156

[a]The primary tool is listed in the row header and the secondary tool in the column header. The score shown for the same tool on both axes (that is, along the main diagonal) is the individual *nCC* score from **Figure 1**. Numerical values are categorized by color, ranging from dark blue (poorer predictions) to red (better predictions).

© 2005 Nature Publishing Group http://www.nature.com/naturebiotechnology

Multi-site Motif

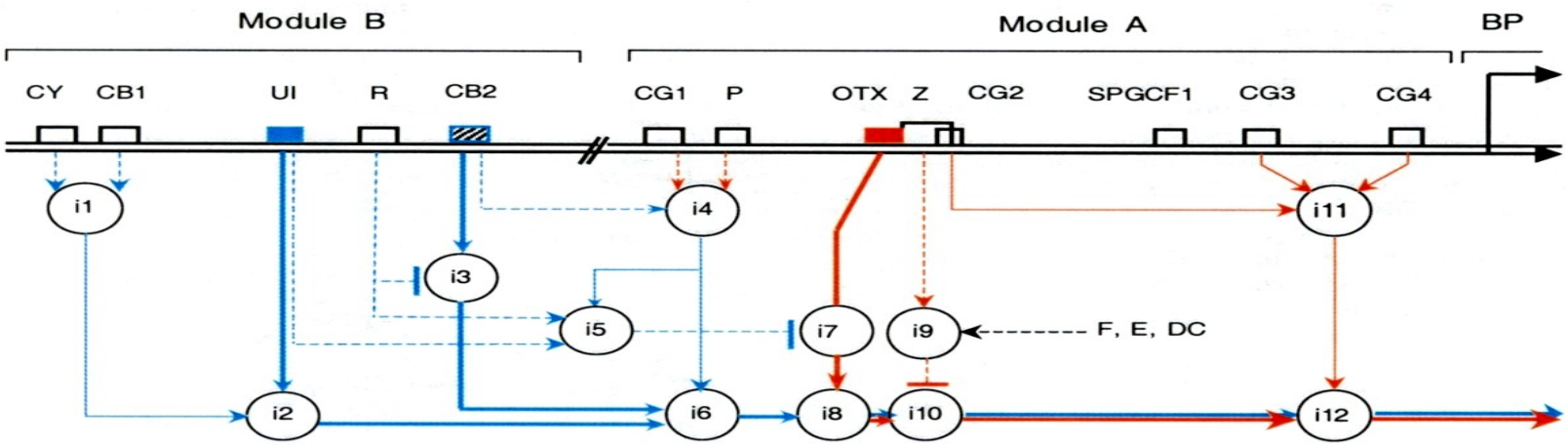
- Two-site: Dimer, dyad
- Gapped Motif
- In general, a motif is an ordered set of binding sites

Table 3 • Dimer alignment for MCM1 binding site

```
.ACC.....AGGA.
.ACC......GGAA
..CCTA...AGGA.
.ACCT...AAGG..
..CCT.....GGAA
..CCTA....GGAA
TACC....AAGG..
.ACCT.....GGA.
.ACCT....AGGA.
TACC......GGA.
TACC.....AGGA.
.ACCT.....GGAA
TACC...    GGAA
```

Higher Order Motifs

- Nature of course is more complicated...



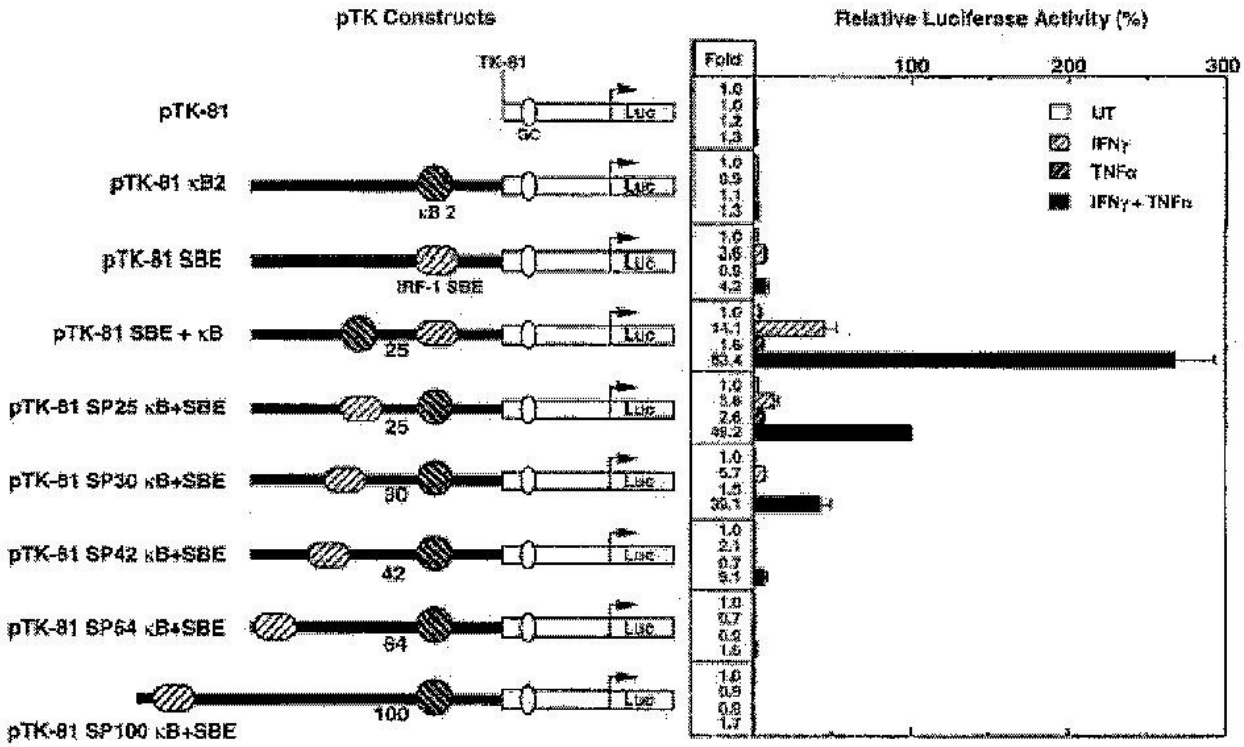
- <u>Combinatorial motifs:</u> combinations of binding sites to which an interacting group of TFs binds
- More realistic, but difficult to look for
- Sinha, 2002

What is Nature Like?

Now that we are talking about realistic motifs, what is it that we know about them from biology?

- Combinatorial motifs are sets of simple motifs separated by a stretch of DNA
- Changing the order of the simple motifs within it doesn't kill transcription, but changes it
- Changing the distance between the simple motifs usually kills transcription
- The distances between motifs are usually small (<20bp)
- The distance restriction is sometimes strict, and other times not
- Randomly distributed simple motifs do not activate transcription

Dependence of Simple Motif Pairs on Distance and Order Between Them



Ohmori et al., 1997

Finding Higher Order Motifs

Sinha (2002) reviews methods for finding higher order motifs, and groups the approaches based on their general relationship to simple motif finders

- find simple motifs and discover patterns made of these
- start with simple motifs and build higher order ones
- find higher order motifs from scratch (e.g. Marsan and Sagot, 2000)

Models of Higher Order Motifs

- The set model $\{M_1, M_2, \ldots, M_k\}$
- Tuples with distance constraints
  (M_1, M_2, d_{12})
- Hidden Markov Model
- Boolean Combinations

Usually two step approaches:

- Enumerate the motif models
- Determine significance (Monte Carlo experiments)

Tricky Business

- All these models have a lot of parameters (e.g. distances between motifs)
- They depend on the initial choice of parameters and/or an initial set of simple motifs
- Using these tools is more of an art than science so far

Conclusions

- PWMs do well for simple motifs
- Combinatorial methods are probably doing better
- Should use all available tools to determine strong simple motifs
- Higher order motifs:
  - positive: knowing your biochemistry helps
  - negative: nobody knows the biochemistry fully!

References:

- Saurabh Sinha, Ph.D. thesis, U of Washington, 2002
- Sinha and Tompa, *Performance Comparison of Algorithms for Finding Transcription Factor Binding Sites,* BIBE 2003
- Marsan and Sagot, *Algorithms for Extracting Structured Motifs Using a Suffix Tree*, JCB, v.7, 2000, 345-362
- Ohmori et al., Journal of Biological Chemistry, 1997
- Tompa et al (2005), Nat. Biotech