BMI 203 Special Topics Lecture
Transcription Factor Binding Site Modeling

Lawrence Hon

Jain Lab

University of California, San Francisco

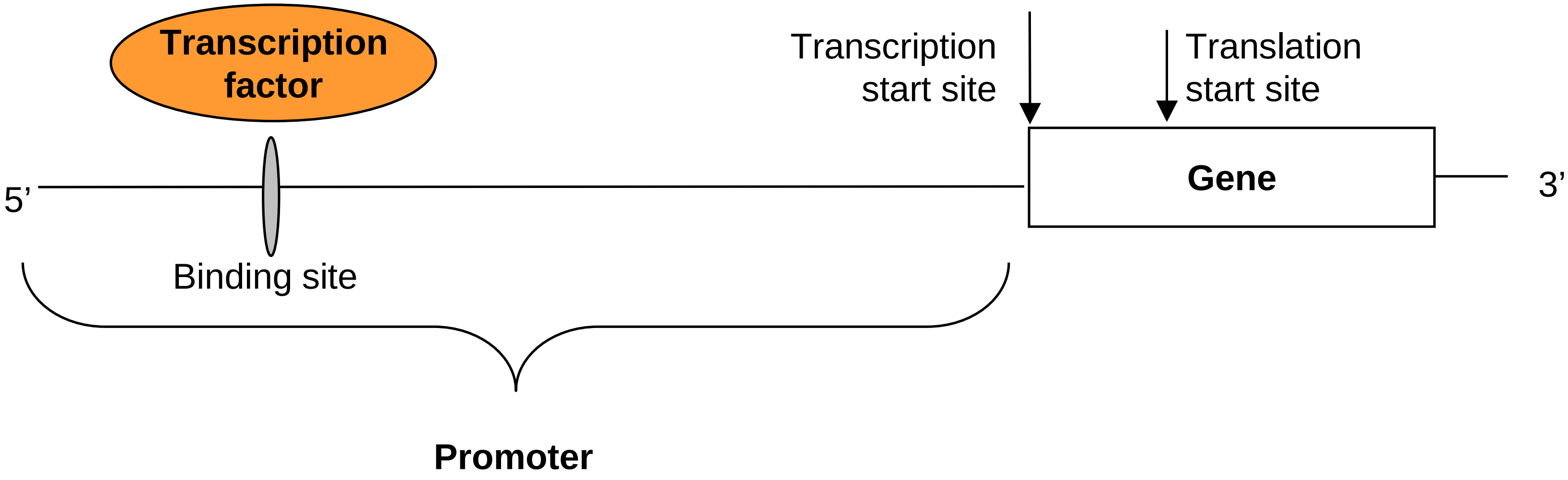
May 11, 2004

Outline

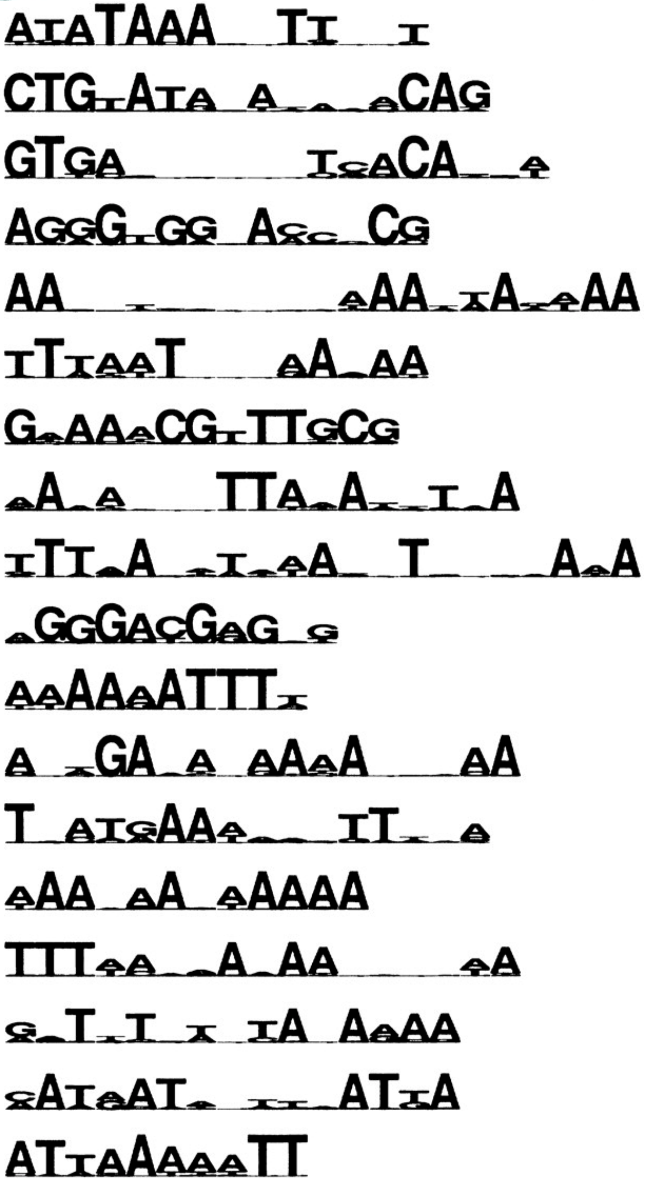
- Biology
- Basic Models
- Two Problems
  - Motif Finding
  - TF Binding Site Recognition

Biology

- Transcriptional regulation
  - – Mechanism to express genes as mRNA
  - – mRNA later becomes proteins



TF Binding Sites



- Tiny
- Highly Variable
- ~Constant Size
- Often repeated
- Low-complexity-ish

Motif vs. Binding site

- Binding site
  - – An individual short sequence which the TF binds onto
- Motif
  - – Represents all the possible sequences that a TF can bind onto
  - – E.g. PWM, other models

Consensus Sequence

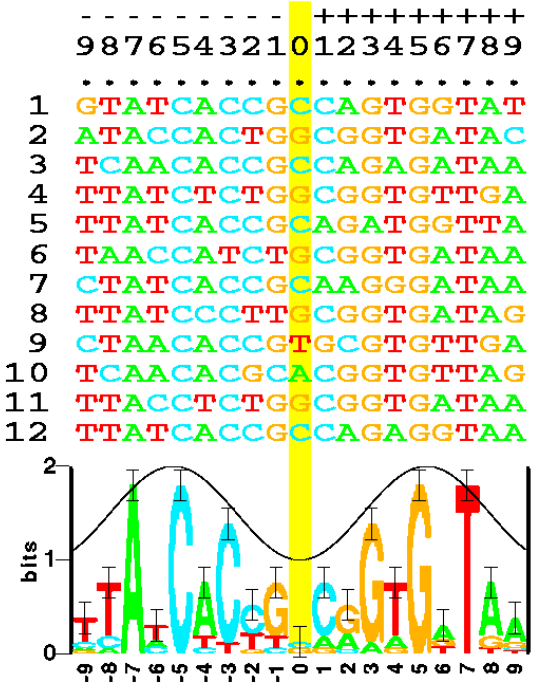
- Simplest model, intuitive to understand
- Represents the “average” sequence
- e.g. CACCCA
- Score of another sequence compared to consensus = number of matches
- Increase sophistication:
  - IUPAC codes: R=A or G, Y=C or T

Position Weight Matrix (PWM)

- For each position, state probability of each nucleotide
- Columns sum to 1
- Score of test sequence = sum of values corresponding to the correct letter for each position
  - CACCCA = .9+.8+.95+.8+.85+.8

	1	2	3	4	5	6
A	0.01	0.80	0.03	0.04	0.10	0.80
C	0.90	0.05	0.95	0.80	0.85	0.03
G	0.04	0.10	0.01	0.07	0.04	0.06
T	0.05	0.05	0.01	0.09	0.01	0.11

Representations compared



```
    ---------+++++++++
    9876543210123456789
    ...................
 1  GTATCACCGCCAGTGGTAT
 2  ATACCACTGGCGGTGATAC
 3  TCAACACCGCCAGAGATAA
 4  TTATCTCTGGCGGTGTTGA
 5  TTATCACCGCAGATGGTTA
 6  TAACCATCTGCGGTGATAA
 7  CTATCACCGCAAGGGATAA
 8  TTATCCCTTGCGGTGATAG
 9  CTAACACCGTGCGTGTTGA
10  TCAACACGCACGGTGTTAG
11  TTACCTCTGGCGGTGATAA
12  TTATCACCGCCAGAGGTAA
```

12 Lambda cI and cro binding sites

Fig. 1. Some aligned sequences and their sequence logo. At the top of the figure are listed the 12 DNA sequences from the P_L and P_R control regions in bacteriophage lambda. These are bound by both the cI and cro proteins [16]. Each even numbered sequence is the complement of the preceding odd numbered sequence. The sequence logo, described in detail in the text, is at the bottom of the figure. The cosine wave is positioned to indicate that a minor groove faces the center of each symmetrical protein. Data which support this assignment are given in reference [17].

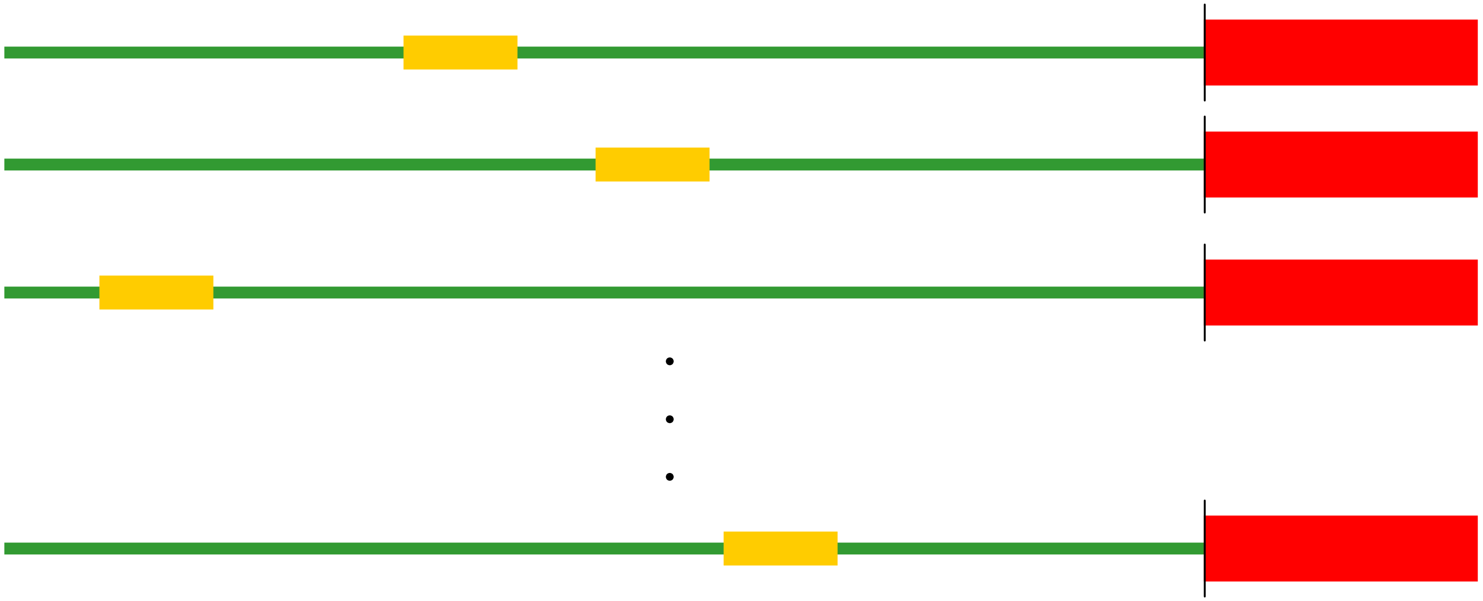
PWM

	9	8	7	6	…
A	1	1	12	3	
C	2	2	0	3	
G	1	0	0	0	
T	8	9	0	6	

Consensus

TTATCAC…

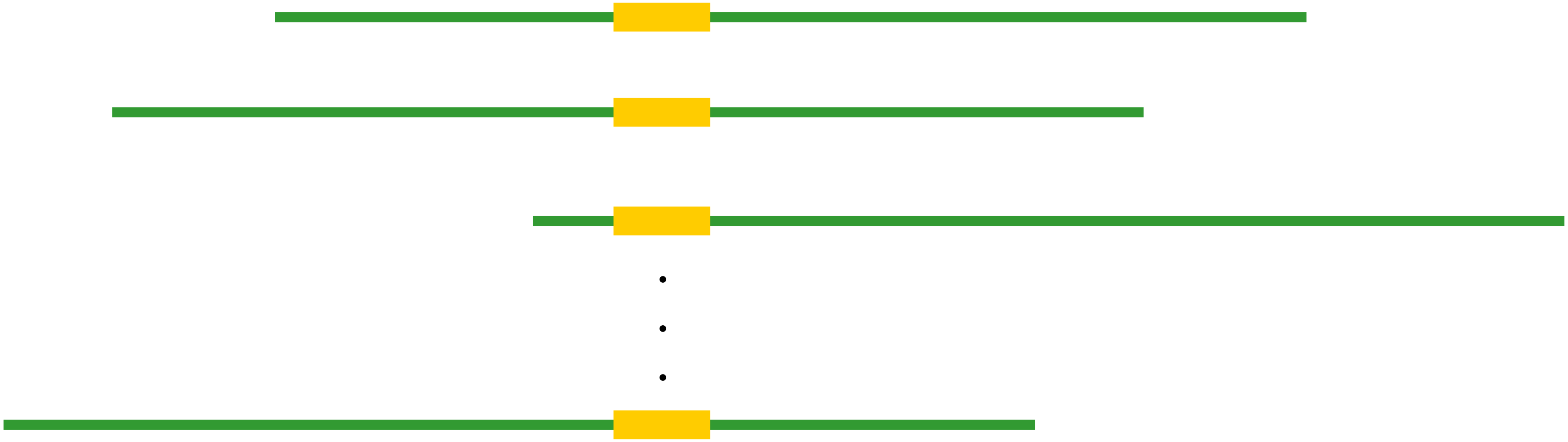
Problem 1: Motif Finding



Given a collection of genes with common expression,

Find the TF-binding motif in common

Essentially a Multiple Local Alignment



- Find "best" multiple local alignment
- Why can't we use standard multiple alignment algorithms?

Why?

- Experiments to determine TFBS are time consuming and expensive
- However, plenty of data
  - Microarrays, ChIP
  - Experiments determining that “gene X is responsive to transcription factor Y”
- Computational approaches to take advantage of this data are cheaper, will help understanding of transcriptional regulation

Scope of problem

- Rap1 binding site in yeast
  - – 6 bp core sequence CACCCA
  - – By chance, expect to see once very 4^6=4096 bases, or more if we allow mismatches
  - – Could be as many as one CACCCA type sequence in every gene, on average

State of the Art

- Most algorithms can handle finding correct motifs in yeast
  - ~1000 bases upstream
  - ~10 genes
- The goal: human
  - ~10,000 bases upstream
  - ? genes

Exhaustive Search

- For all *k*-length sequences (4^k)
  - Consider this as potential consensus motif
  - Compare against all k-mers in dataset
  - Motif is good if many close matches in dataset
- Advantage: finds “best” motif
- Disadvantage: slow – O(4^k)

Motif Finding algorithms

- Greedy search:
  - CONSENSUS
- Expectation Maximization:
  - MEME
- Gibbs Sampling:
  - AlignACE, BioProspector

Gibbs sampling

- Uses PWM as underlying model
- Stochastic algorithm
  - Multiple starting points
- Relatively fast
- Similar to EM, but easier to implement

Summary

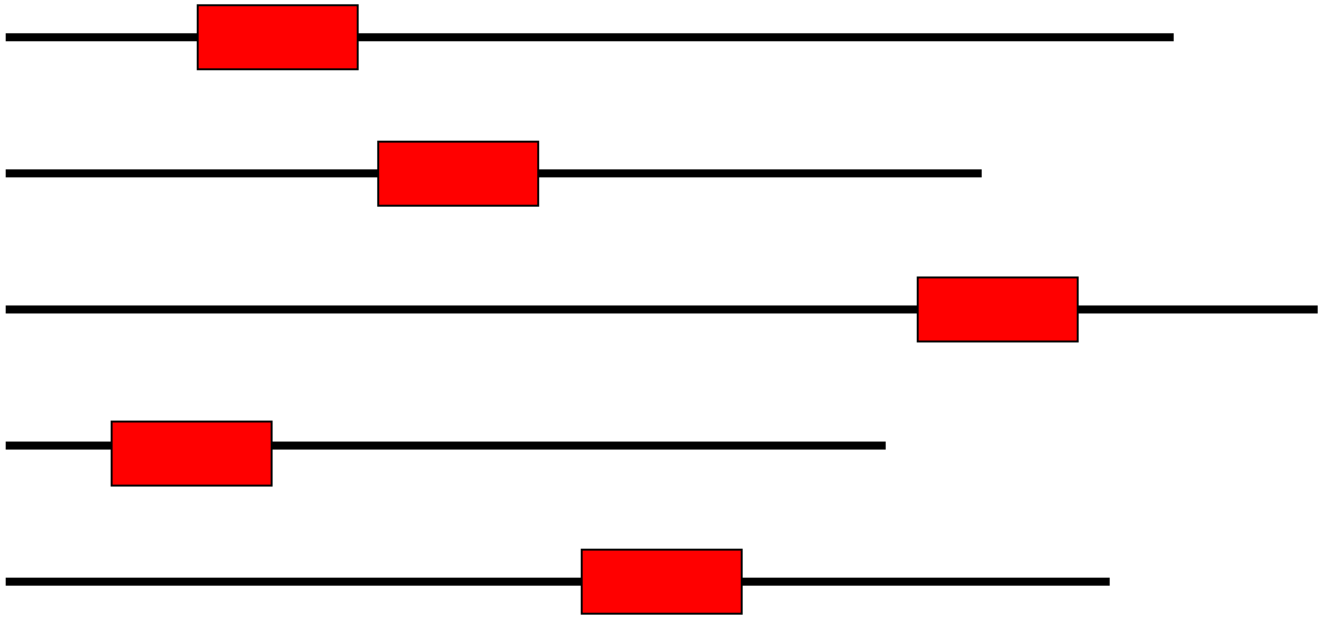
Algorithm (sketch):

1. Initialization:
   a. Select random locations in sequences $x^1, \ldots, x^N$
   b. Compute an initial PWM from these locations

2. Sampling Iterations:
   a. Remove one sequence x^i
   b. Recalculate PWM
   c. Pick a new location of site in x^i using highest scoring sequence according to PWM

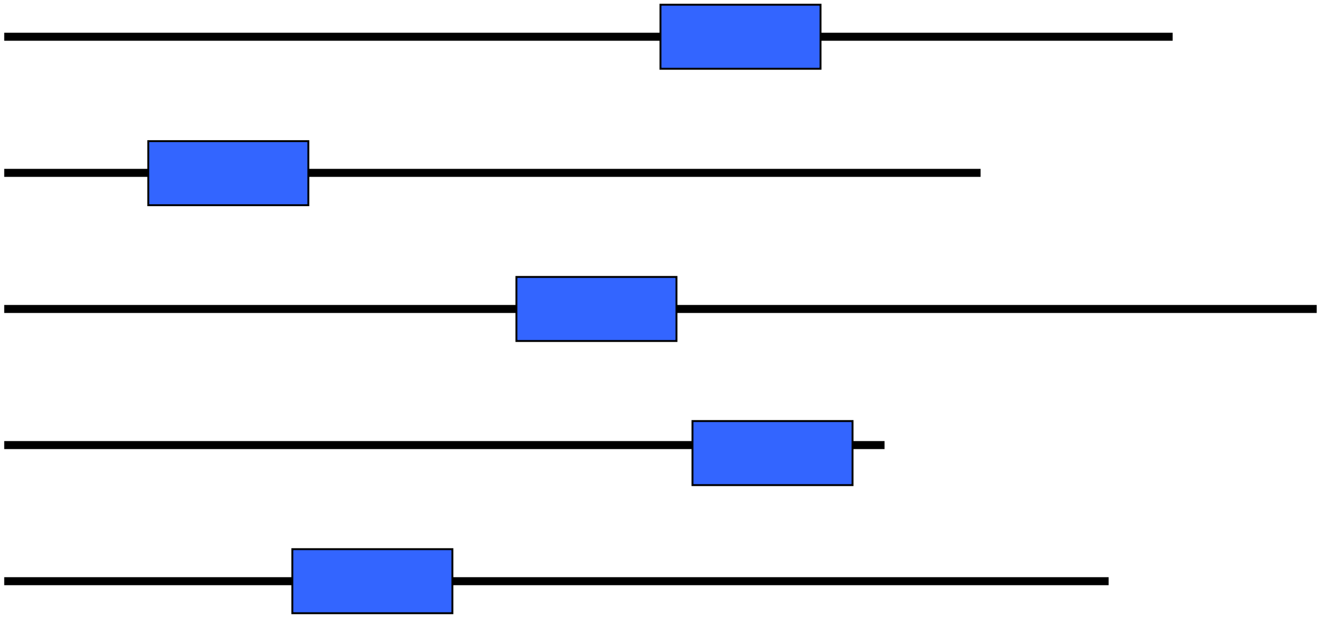
Data

- Binding site responsive to a TF is found in all 5 sequences



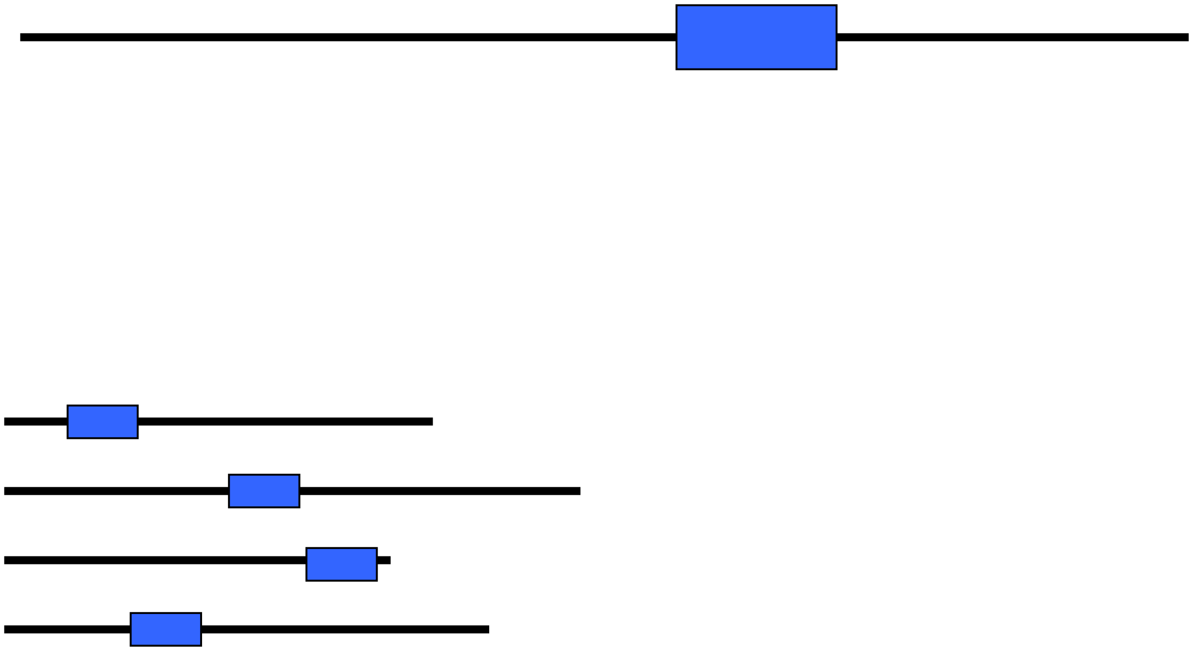
Step 1: Initialize

- Create random PWM



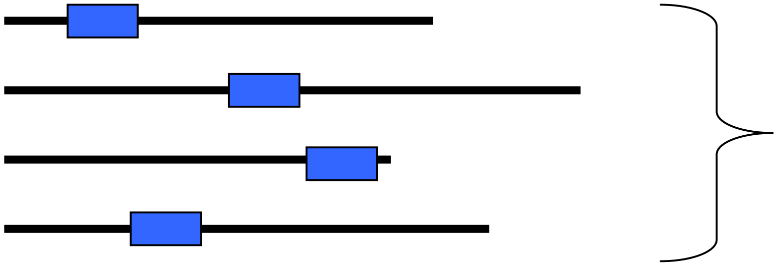
Step 2: Iterate

- Remove one sequence



Step 2: Iterate

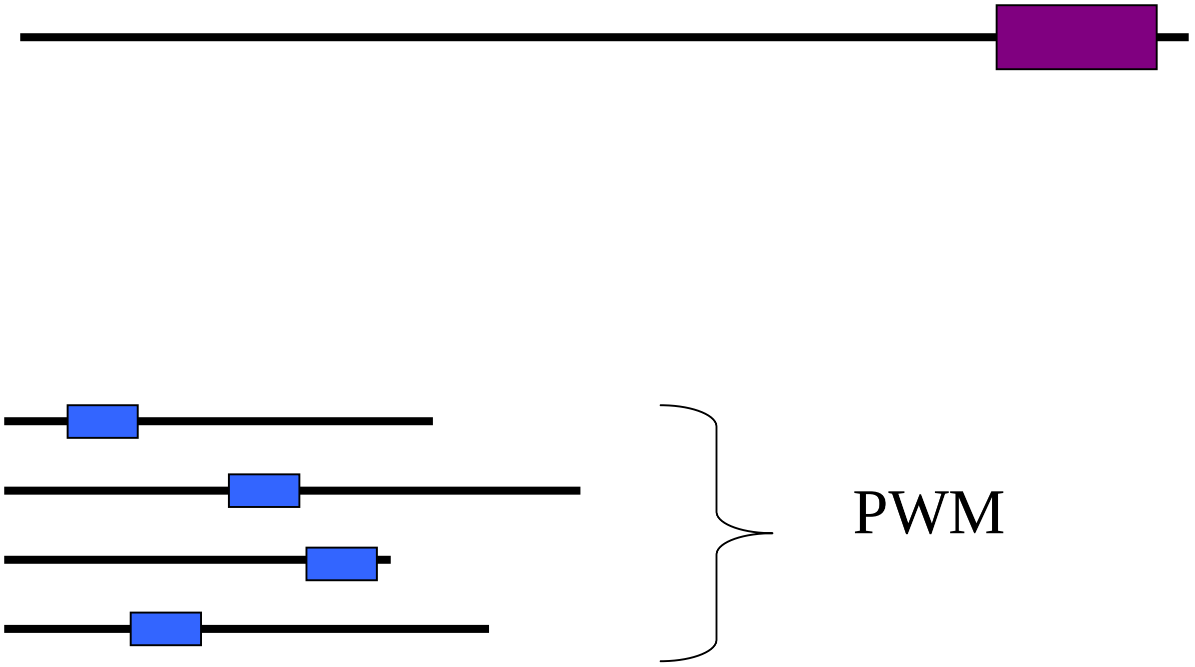
- Generate PWM from remaining sequences



Create PWM

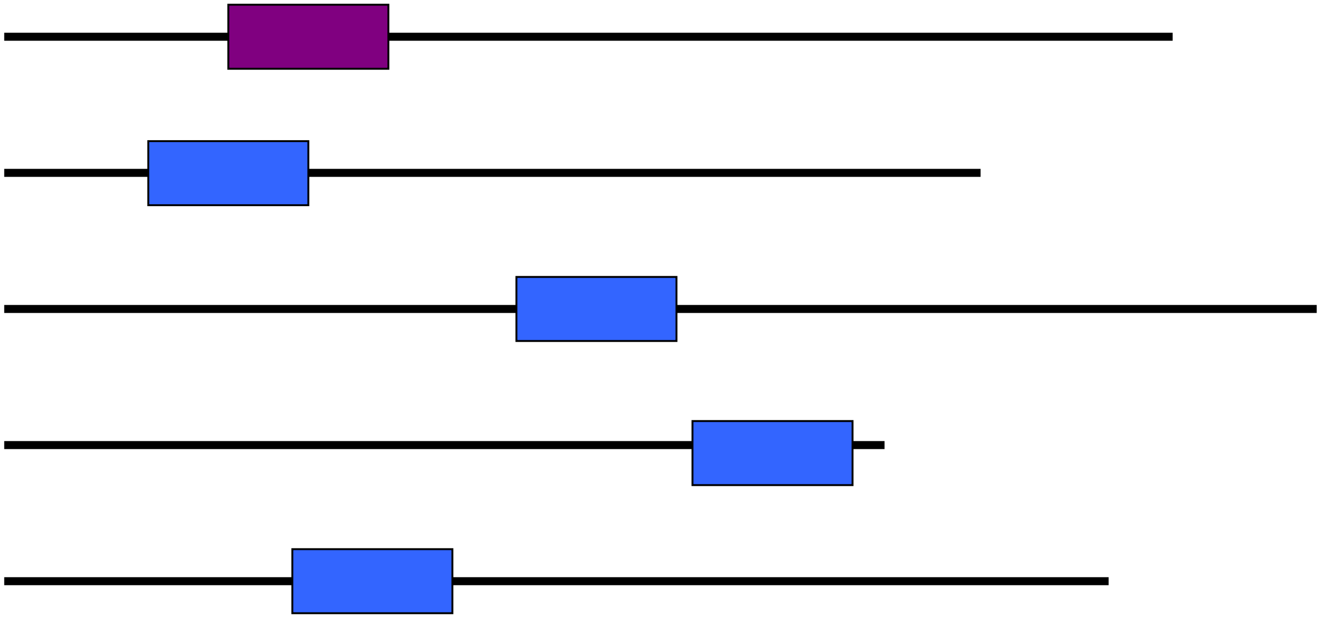
Step 2: Iterate

- Slide window across removed sequence to find best site that fits PWM



Step 2: Iterate

- Keep best site and merge this with remaining sites



Step 3

- Repeat step 2 until convergence
- Intuition:
  - You are more likely to see the real binding site than random sites
  - Once there’s one site in the motif, there’ll be a strong preference for other real sites to enter the motif (versus other random sequences)

Greedy Motif Finder

- In order to explore problem more carefully, we designed our own motif finder
- Code separated into search algorithm and scoring functions
  - New methods and functions can be plugged in easily

Scoring Function

- Don’t explicitly use a PWM or other model
- Instead:
  - Count number of pairwise nucleotide matches in a motif
  - Prefer sequences that are unique (i.e. not commonly found in genome as a whole)
- Or, prefer overrepresented but unique sequences

Search Algorithm

- Greedy Search (similar to CONSENSUS)
  - Start with a pair of high scoring binding sites
  - Find other sites that look similar to current motif
  - Take the best site that maximizes score of augmented motif and add to motif
  - Repeat until motif size cutoff

Search Algorithm

- Obviously, resulting motif highly dependent on initial pair of sites chosen
- So, try out lots of different sequences pairs
- Highest scoring motif is the best motif

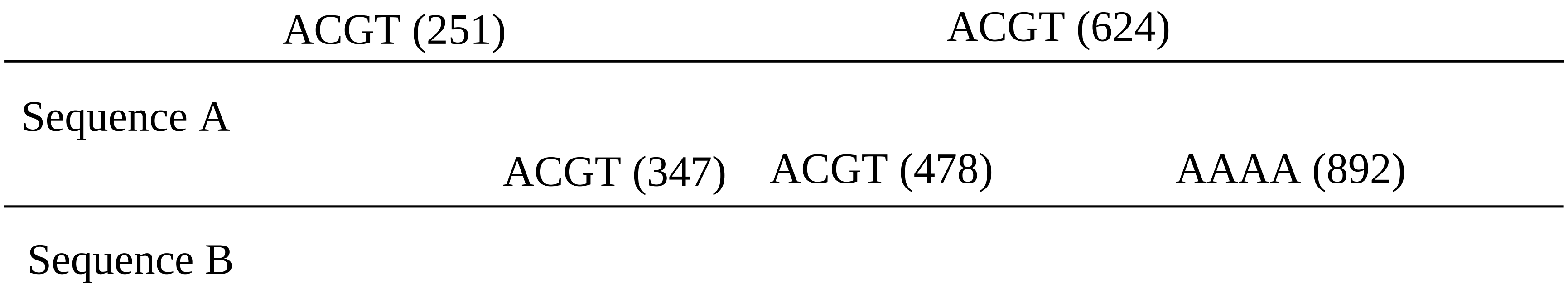
Straightforward method is Slow

- If there are *n* different potential binding sites (~10,000)
- Just to find initial sequences pairs, we need to do n^2 comparisons (~100 million)

Optimization

- Precalculate pairwise comparisons
  - – so we can quickly ask, “What other binding sites look similar to binding site x?”
  - – After precalculation, subsequent lookups are constant time
- Pairwise comparison uses indexing, so it takes O(n) instead of O(n^2) time
  - – Small decrease in sensitivity

Indexing Approach



ACGT (251)

ACGT (624)

Sequence A

ACGT (347)

ACGT (478)

AAAA (892)

Sequence B

Seq A Index	Seq B Index	ACGT Matches
AAAA	AAAA = 892	Seq A, 251 and Seq B, 347
…	…	Seq A, 251 and Seq B, 478
ACGT = 251, 624	ACGT = 347, 478	Seq A, 624 and Seq B, 347
ACTA	ACTA	Seq A, 624 and Seq B, 478
…	…	
TTTT	TTTT	

Current status

- Works in yeast
- Gunning for human
  - – scoring function
    - background models (uniqueness in genome)
  - – Other data
    - comparative genomics

Problem 2: Binding Site Recognition

- Definition
  - – Given true binding sites
  - – identify other binding sites in test set
- Why?
  - – Computational method to identify new binding sites in genes not previously considered

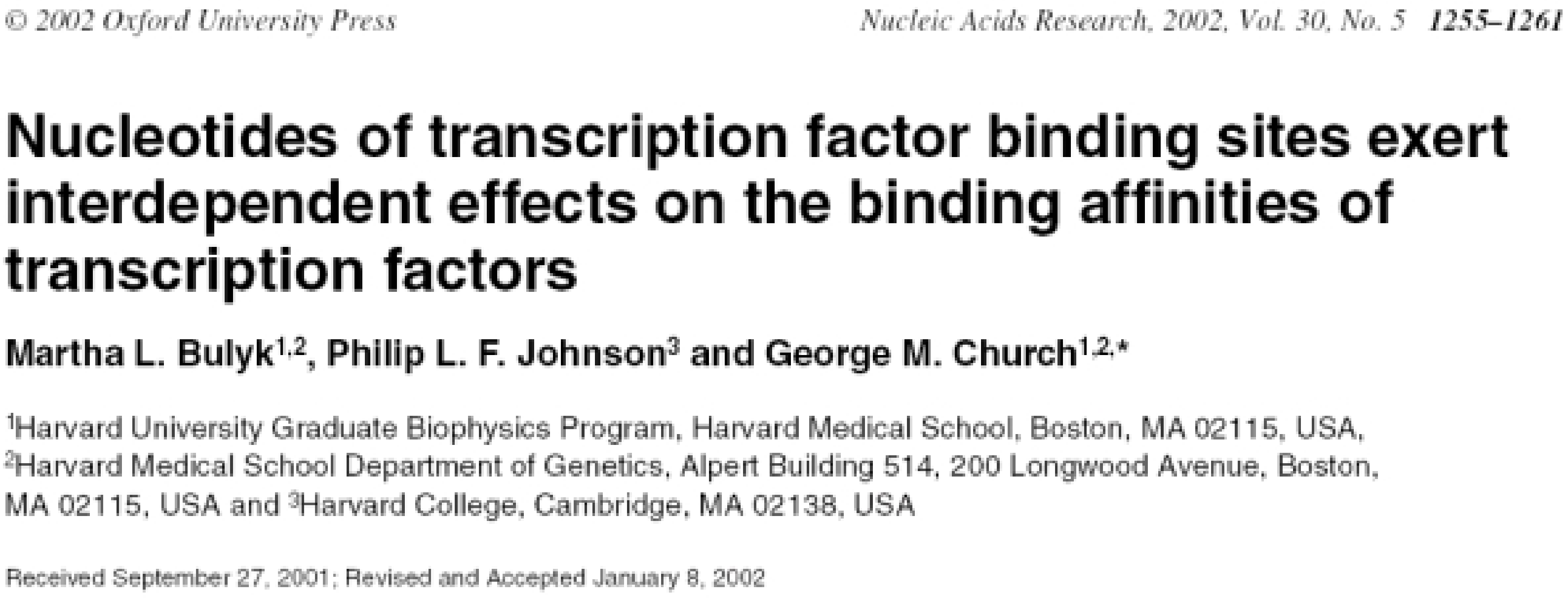
Standard approach

- Create a PWM from binding sites
- Run PWM across putative sites
- High scoring sequences are potential binding sites

Limitations of PWM

- Independence between positions
  - Choice of nucleotide in position x has no effect on that of position y
  - Can’t represent this: “If position 2 in the binding site is an A, then position 5 should be a G”
- Implicit background model
  - What if a repetitive sequence scores highly?
- Does it matter? Is a PWM good enough?

Dependence between positions in binding site



© 2002 Oxford University Press Nucleic Acids Research, 2002, Vol. 30, No. 5 1255–1261

Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors

Martha L. Bulyk[1,2], Philip L. F. Johnson[3] and George M. Church[1,2,*]

[1]Harvard University Graduate Biophysics Program, Harvard Medical School, Boston, MA 02115, USA,
[2]Harvard Medical School Department of Genetics, Alpert Building 514, 200 Longwood Avenue, Boston,
MA 02115, USA and [3]Harvard College, Cambridge, MA 02138, USA

Received September 27, 2001; Revised and Accepted January 8, 2002

- Show dependence between positions
  - Use microarray binding experiment
  - Enumerate central 3 bp of binding site of Zif268 zinc fingers
  - Analyze binding affinities

New motif model

- Use a neural network instead of a PWM
  - Three layer, fully connected
- Inputs
  - Binding site sequence
  - Other information?
- Output
  - Value between 0-1 showing categorization of binding site
    - Closer to 1: yes, is a binding site
    - Closer to 0: no, probably not

Binding site as Input

	C	A	T
A	0	1	0
C	1	0	0
G	0	0	0
T	0	0	1

= (0, 1, 0, 0,
1, 0, 0, 0,
0, 0, 0, 1)

For k-length sequence, 4k input nodes

Training Data

- Positive training data
  - – pre-curated binding sites
- Negative training data
  - – Random sequences drawn from the genome
  - – Actual genomic data negates need for a background model

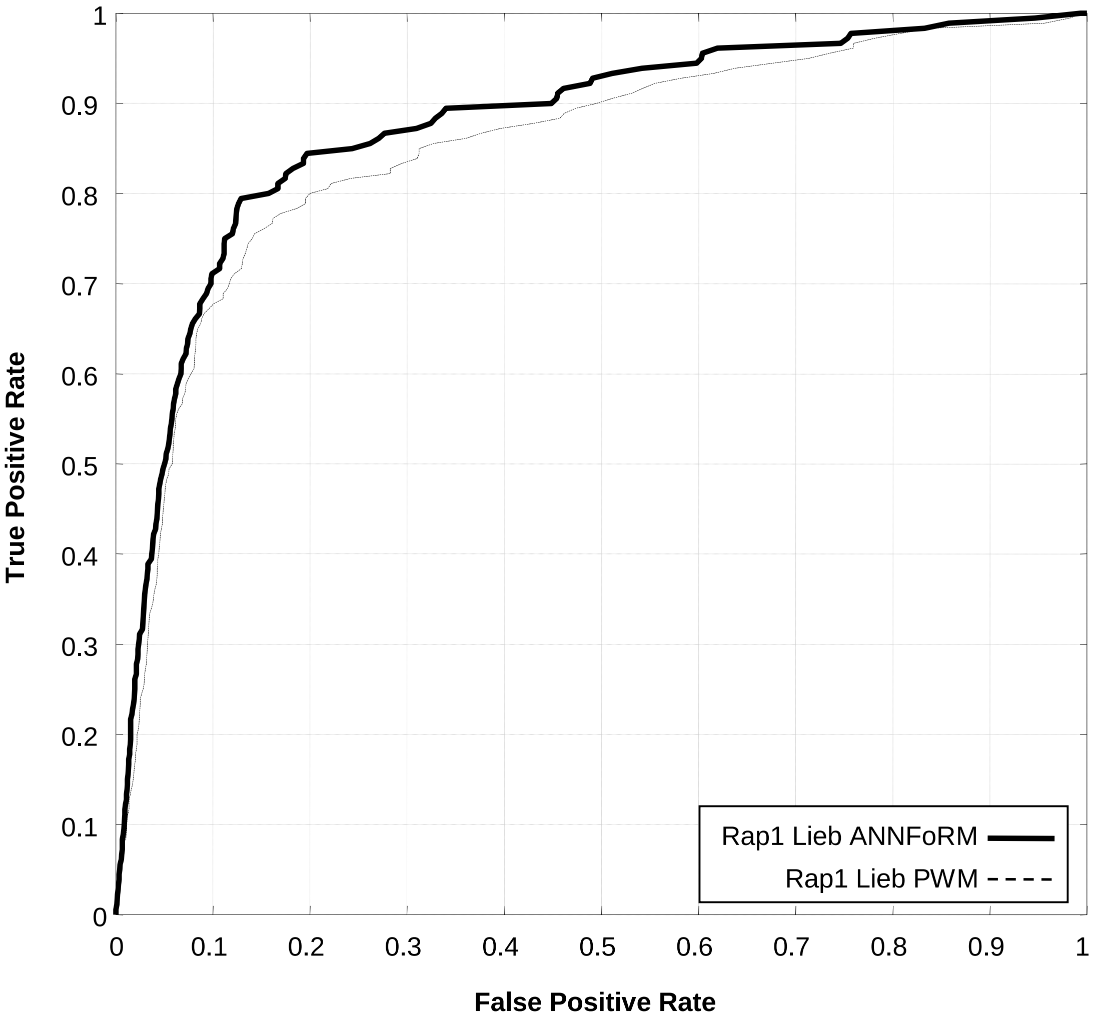
Challenge

- How do we train network with small number of positives but a large number of negatives?
- Solution: Sample negatives, don’t use all of them
    - Some negative sequences may provide more value for the training of the network (i.e. result in large errors)
    - High value sequences should be exposed to the ANN more often

Results

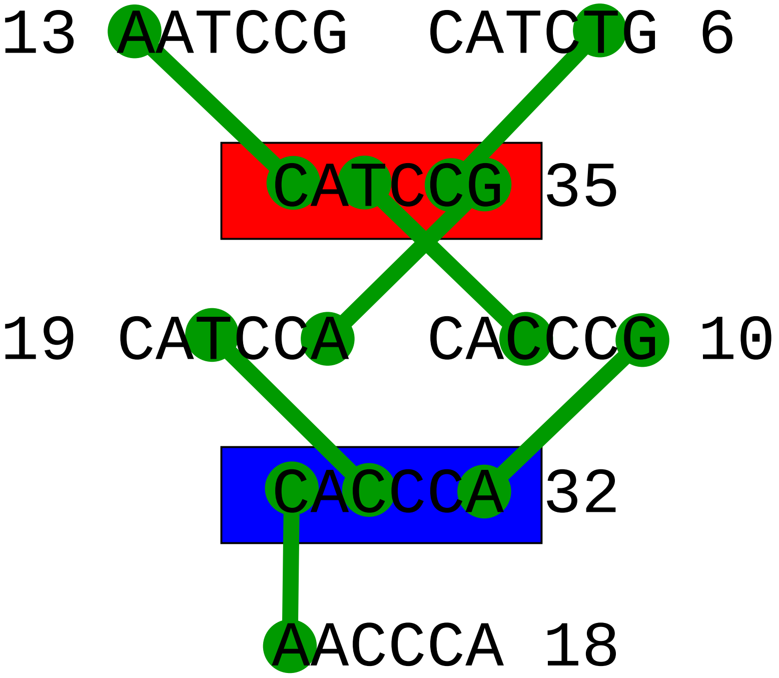
- Resultant neural net robust to choice of parameters
- For small datasets, equivalent performance compared to PWM
- For larger datasets, neural network does better

Results



ROC Plot Comparing ANNFoRM and PWM

Results



13 AATCCG

CATCTG 6

CATCCG 35

19 CATCCA

CACCCG 10

CACCCA 32

AACCCA 18

Position 11 12**3**45**6**78901	Rank of ANN output	Rank of PWM output
CATCCGTACAT	1	2
CACCCATACAT	2	3
CATCCATACAT	3	1
CACCCGTACAT	4	4

Summary

- Transcription regulation
- Consensus sequences, PWMs
- Motif Finding
  - Gibbs Sampling
  - My greedy search algorithm
- Motif Recognition