Computational Biology: Genomes, Networks, Evolution

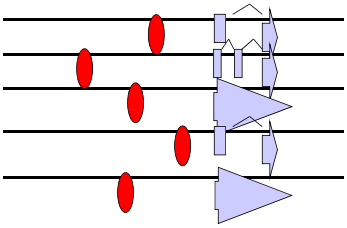
Motif Discovery

Lecture 9

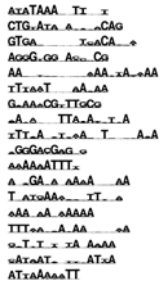
October 2, 2008

Regulatory Motifs

Find promoter motifs associated with co-regulated or functionally related genes



Transcription Factor Binding Sites

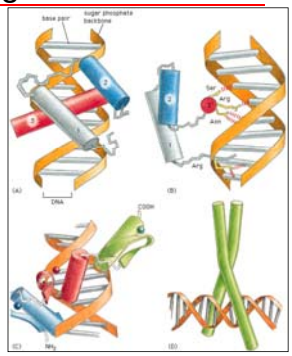


- Very Small
- Highly Variable
- ~Constant Size
- Often repeated
- Low-complexity-ish

Slide Credit: S. Batzoglou

Motifs Are Degenerate

- Protein-DNA interactions
  - Proteins read DNA by "feeling" the chemical properties of the bases
  - Without opening DNA (not by base complementarity)
- Sequence specificity
  - Topology of 3D contact dictates sequence specificity of binding
  - Some positions are fully constrained; other positions are degenerate
  - "Ambiguous / degenerate" positions are loosely contacted by the transcription factor



Other "Motifs"

- Splicing Signals
  - Splice junctions
  - Exonic Splicing Enhancers (ESE)
  - Exonic Splicing Surpressors (ESS)
- Protein Domains
  - Glycosylation sites
  - Kinase targets
  - Targetting signals
- Protein Epitopes
  - MHC binding specificities

Essential Tasks

- Modeling Motifs
  - How to computationally represent motifs
- Visualizing Motifs
  - Motif "Information"
- Predicting Motif Instances
  - Using the model to classify new sequences
- Learning Motif Structure
  - Finding new motifs, assessing their quality

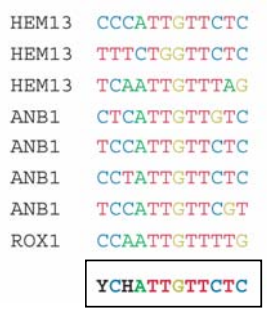
Modeling Motifs

Consensus Sequences

Useful for publication

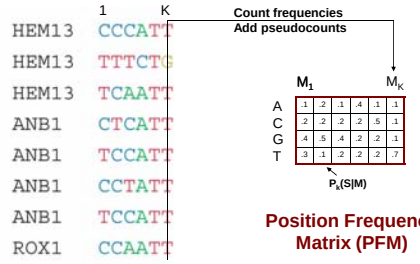
IUPAC symbols for degenerate sites

Not very amenable to computation



Nature Biotechnology 24, 423 - 425 (2006)

Probabilistic Model



Position Frequency Matrix (PFM)

Scoring A Sequence

To score a s… compare to a null model

Log likelihood ratio

$$Score = \log \frac{P(S \mid PFM)}{P(S \mid B)}$$

PFM

A	.1	.2	.1	.4	.1	.1
C	.2	.2	.2	.2	.5	.1
G	.4	.5	.4	.2	.2	.1
T	.3	.1	.2	.2	.2	.7

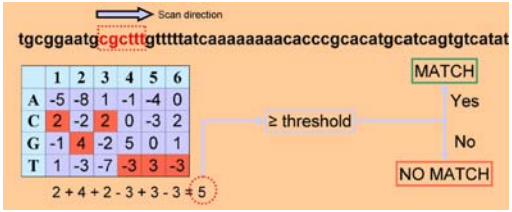
Background DNA (B)

A: 0.25
T: 0.25
G: 0.25
C: 0.25

Position Weight Matrix (PWM)

A	-1.3	-0.3	-1.3	0.6	-1.3	-1.3
C	-0.3	-0.3	0.3	-0.3	1	-1.3
G	0.6	1	0.6	-0.3	-0.3	-1.3
T	0.3	-1.3	-0.3	-0.3	-0.3	1.4

Scoring a Sequence

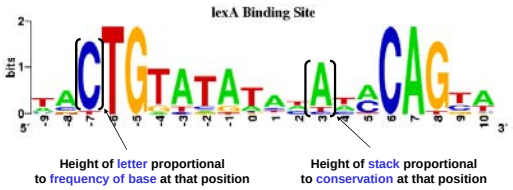


Common threshold = 60% of maximum score

MacIsaac & Fraenkel (2006) PLoS Comp Bio

Visualizing Motifs – Motif Logos

Represent both base frequency and conservation at each position

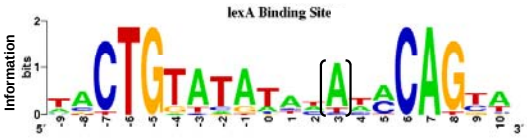


Height of letter proportional to frequency of base at that position

Height of stack proportional to conservation at that position

Motif Information

The height of a stack is often called the motif information at that position measured in bits



$$\text{Motif Position Information} = 2 - \sum_{b=\{A,T,G,C\}} -p_b \log p_b$$

Why is this a measure of information?

Uncertainty and probability

Uncertainty is related to our surprise at an event

"The sun will rise tomorrow" **Not surprising (p~1)**

"The sun will *not* rise tomorrow" ***Very*** **surprising (p<<1)**

Uncertainty is inversely related to probability of event

$$\text{Uncertainty} \propto \log \frac{1}{p_{event}}$$

Average Uncertainty

Two possible outcomes for sun rising

A "The sun will rise tomorrow" **$P(A)=p_1$**

B "The sun will *not* rise tomorrow" **$P(B)=p_2$**

What is our *average uncertainty* about the sun rising

Entropy

Entropy measures average uncertainty

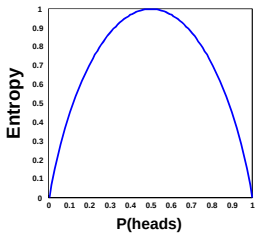
Entropy measures randomness

$$H(X) = -\sum_i p_i \log_2 p_i$$

If log is base 2, then the units are called bits

Entropy versus randomness

Entropy is maximum at maximum randomness

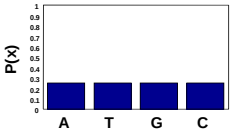


Example: Coin Toss

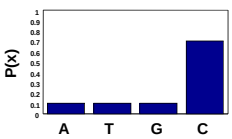
P(heads)=0.1 Not very random
H(X)=0.47 bits

P(heads)=0.5 Completely random
H(X)=1 bits

Entropy Examples



$$H(X) = -[0.25\log(0.25) + 0.25\log(0.25) + 0.25\log(0.25) + 0.25\log(0.25)] = 2 \text{ bits}$$



$$H(X) = -[0.1\log(0.1) + 0.1\log(0.1) + 0.1\log(0.1) + 0.75\log(0.75)] = 0.63 \text{ bits}$$

Information Content

Information is a *decrease in uncertainty*

Once I tell you the sun will rise, your uncertainty about the event decreases

Information = $H_{before}(X)$ - $H_{after}(X)$

Information is difference in entropy after receiving information

Motif Information

Motif Position Information = 2 - $\sum_{b=\{A,T,G,C\}} -p_b \log p_b$

$H_{background}(X)$ — **Prior uncertainty about nucleotide**

$H_{motif\_i}(X)$ — **Uncertainty after learning it is position i in a motif**

Uncertainty at this position has been reduced by 0.37 bits

Motif Logo

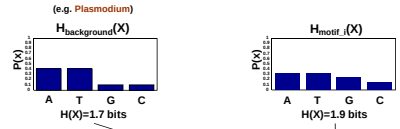
Conserved Residue
Reduction of uncertainty of 2 bits

Little Conservation
Minimal reduction of uncertainty

Background DNA Frequency

The definition of information assumes a uniform background DNA nucleotide frequency

What if the background frequency is not uniform?



Motif Position Information = 1.7 - $\sum_{b=\{A,T,G,C\}} -p_b \log p_b$ = -0.2 bits

Some motifs could have negative information!

A Different Measure

Relative entropy or Kullback-Leibler (KL) divergence

Divergence between a "true" distribution and another

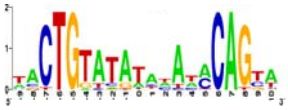
$$D_{KL}(P_{motif} \,||\, P_{background}) = \sum_{i=\{A,T,G,C\}} P_{motif}(i) \log \frac{P_{motif}(i)}{P_{background}(i)}$$

"True" Distribution **Other Distribution**

D_{KL} is larger the more different P_{motif} is from $P_{background}$

Comparing Both Methods

Information assuming uniform background DNA



KL Distance assuming 20% GC content (e.g. Plasmodium)

Online Logo Generation

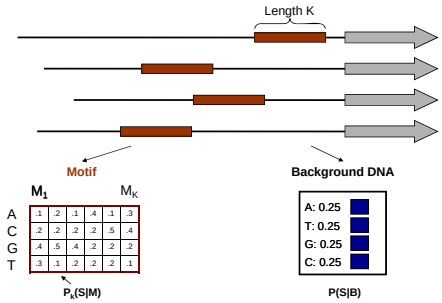
http://weblogo.berkeley.edu/

http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi

Finding New Motifs

Learning Motif Models

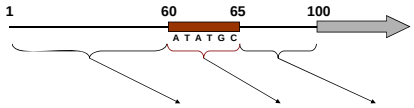
A Promoter Model



The same motif model in all promoters

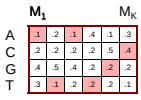
Probability of a Sequence

Given a sequence(s), motif *model* and motif *location*



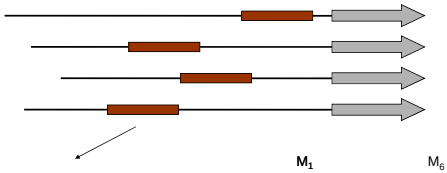
$$P(Seq \mid Mstart = 10, Model) = \prod_{i=1}^{59} P(S_i \mid B) \prod_{k=1}^{6} P_k(S_{k+63} \mid M) \prod_{i=66}^{100} P(S_i \mid B)$$

S_i = *nucleotide at position i in the sequence*



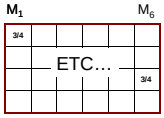
Parameterizing the Motif Model

Given multiple sequences and motif locations but no motif *model*



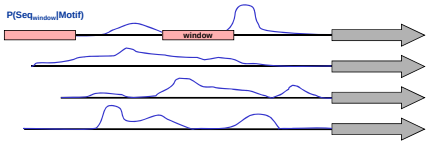
AATGCG
ATATGG
ATATCG
GATGCA

Count Frequencies
Add pseudocounts



Finding Known Motifs

Given multiple sequences and motif model but no motif *locations*



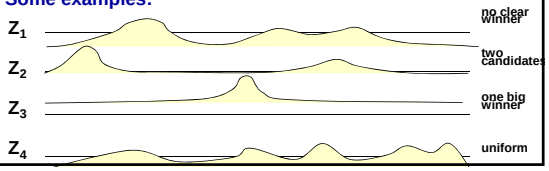
Calculate P(Seq$_{window}$|Motif) for every starting location

Motif Position Distribution Z_{ij}

- the element Z_{ij} of the matrix Z represents the probability that the motif starts in position *j* in sequence *I*

$$Z = \begin{array}{c} \\ \text{seq1} \\ \text{seq2} \\ \text{seq3} \\ \text{seq4} \end{array} \begin{array}{cccc} 1 & 2 & 3 & 4 \\ 0.1 & 0.1 & 0.2 & 0.6 \\ 0.4 & 0.2 & 0.1 & 0.3 \\ 0.3 & 0.1 & 0.5 & 0.1 \\ 0.1 & 0.5 & 0.1 & 0.3 \end{array}$$

Some examples:



Calculating the Z Vector

$$P(Z_{ij}=1 \mid S,M) = \frac{P(S \mid Zij=1,M)P(Zij=1)}{P(S)}$$ **(Bayes' rule)**

$$P(Z_{ij}=1 \mid S,M) = \frac{P(S \mid Zij=1,M)\cancel{P(Zij=1)}}{\sum_{k=1}^{L-K+1} P(S \mid Zij=1,M)\cancel{P(Zij=1)}}$$

$$P(Z_{ij}=1 \mid S,M) = \frac{P(S \mid Zij=1,M)}{\sum_{k=1}^{L-K+1} P(S \mid Zij=1,M)}$$

Assume uniform priors (motif equally likely to start at any position)

Calculating the Z Vector - Example

$X_i =$ G C T G T A G

$$p = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{cccc} 0 & 1 & 2 & 3 \\ 0.25 & 0.1 & 0.5 & 0.2 \\ 0.25 & 0.4 & 0.2 & 0.1 \\ 0.25 & 0.3 & 0.1 & 0.6 \\ 0.25 & 0.2 & 0.2 & 0.1 \end{array}$$

$$Z_{i1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

$\vdots$

- then normalize so that $\sum_{j=1}^{L-W+1} Z_{ij} = 1$

Discovering Motifs

Given a set of co-regulated genes, we need to discover with only sequences

We have neither a motif model nor motif locations
Need to discover both

How can we approach this problem?

Expectation Maximization (EM)

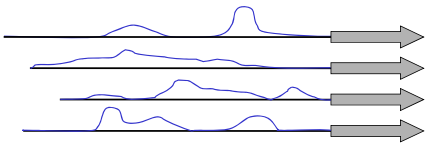
Remember the basic idea!

1.Use model to estimate distribution of missing data
2.Use estimate to update model
3.Repeat until convergence

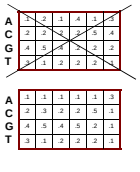
Model is the motif model

Missing data are the motif locations

EM for Motif Discovery



1. Start with random motif model
2. E Step: estimate probability of motif positions for each sequence
3. M Step: use estimate to update motif model
4. Iterate (to convergence)



ETC…

The M-Step Calculating the Motif Matrix

- M_{ck} is the probability of character c at position k
- With specific motif positions, we can estimate M_{ck}:

Counts of c at pos k In each motif position → $n_{c,k}$; Pseudocounts → $d_{c,k}$

$$M_{c,k} = \frac{n_{c,k} + d_{c,k}}{\sum_b n_{b,k} + d_{b,k}}$$

- But with probabilities of positions, Z_{ij}, we average:

$$n_{c,k} = \sum_{\text{sequences } S_i} \sum_{\{j|S_i=c\}} Z_{ij}$$

MEME

- MEME - implements EM for motif discovery in DNA and proteins
- MAST – search sequences for motifs given a model

http://meme.sdsc.edu/meme/

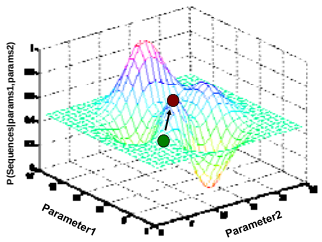
P(Seq|Model) Landscape

EM searches for parameters to increase P(seqs|parameters)

Useful to think of P(seqs|parameters) as a function of parameters

EM starts at an initial set of parameters

And then "climbs uphill" until it reaches a local maximum



Where EM starts can make a big difference

Search from Many Different Starts

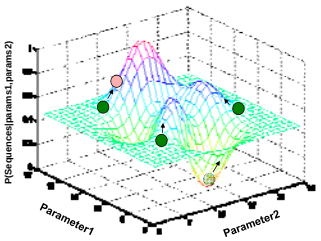
To minimize the effects of local maxima, you should search multiple times from different starting points

MEME uses this idea

Start at many points

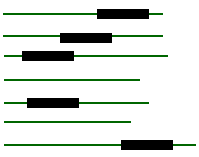
Run for one iteration

Choose starting point that got the "highest" and continue



The ZOOPS Model

- The approach as we've outlined it, assumes that each sequence has exactly one motif occurrence per sequence; this is the OOPS model
- The ZOOPS model assumes *zero or one* occurrences per sequence



E-step in the ZOOPS Model

- We need to consider another alternative: the *i*th sequence doesn't contain the motif
- We add another parameter (and its relative)

λ — **prior prob that any position in a sequence is the start of a motif**

$\gamma = (L - W + 1)\lambda$ — **prior prob of a sequence containing a motif**

E-step in the ZOOPS Model

$$P(Z_{ij}=1)=\frac{\Pr(S_i \mid Z_{ij}=1,M)\lambda}{\Pr(S_i \mid Q_i=0,M)(1-\gamma)+\sum_{k=1}^{L-W+1}\Pr(S_i \mid Z_{ik}=1,M)\lambda}$$

- here Q_i is a random variable that takes on 0 to indicate that the sequence doesn't contain a motif occurrence

$$Q_i=\sum_{j=1}^{L-W+1} Z_{i,j}$$

M-step in the ZOOPS Model

- update *p* same as before
- update λ, γ as follows

$$\lambda^{(t+1)}=\frac{\gamma^{(t+1)}}{(L-W+1)}=\frac{1}{n(L-W+1)}\sum_{sequences}^{n}\sum_{positions}^{m} Z_{i,j}^{(t)}$$

- **average of $Z_{i,j}^{(t)}$ across all sequences, positions**

The TCM Model

- The TCM (two-component mixture model) assumes *zero or more* motif occurrences per sequence

Likelihood in the TCM Model

- the TCM model treats each length *W* subsequence independently
- to determine the likelihood of such a subsequence:

$$\Pr(S_{ij} \mid Z_{ij}=1,M)=\prod_{k=j}^{j+W-1} M_{c_k,k-j+1}$$ **assuming a motif starts there**

$$\Pr(S_{ij} \mid Z_{ij}=0,p)=\prod_{k=j}^{j+W-1} P(c_k \mid B)$$ **assuming a motif doesn't start there**

E-step in the TCM Model

$$Z_{ij}=\frac{\Pr(S_{i,j} \mid Z_{ij}=1,M)\lambda}{\underbrace{\Pr(S_{i,j} \mid Z_{ij}=0,B)(1-\lambda)}_{\text{subsequence isn't a motif}}+\underbrace{\Pr(S_{i,j} \mid Z_{ij}=1,M)\lambda}_{\text{subsequence is a motif}}}$$

- M-step same as before

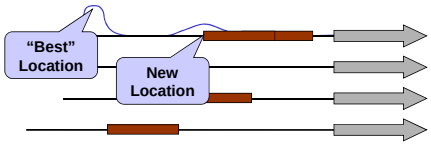
Gibbs Sampling

A stochastic version of EM that differs from deterministic EM in two key ways

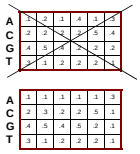
1. At each iteration, we only update the motif position of a single sequence

2. We may update a motif position to a "suboptimal" new position

Gibbs Sampling



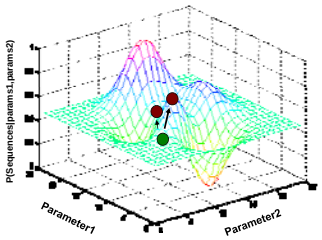
1. Start with random motif locations and calculate a motif model
2. Randomly select a sequence, remove its motif and recalculate tempory model
3. With temporary model, calculate probability of motif at each position on sequence
4. Select new position based on this distribution
5. Update model and Iterate



ETC…

Gibbs Sampling and Climbing

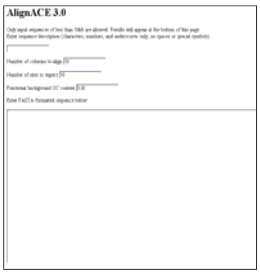
Because gibbs sampling does not always choose the best new location it can move to another place not directly uphill



***In theory,* Gibbs Sampling less likely to get stuck a local maxima**

AlignACE

- **Implements Gibbs sampling for motif discovery**
  - **Several enhancements**
- **ScanAce – look for motifs in a sequence given a model**
- **CompareAce – calculate "similarity" between two motifs (i.e. for clustering motifs)**

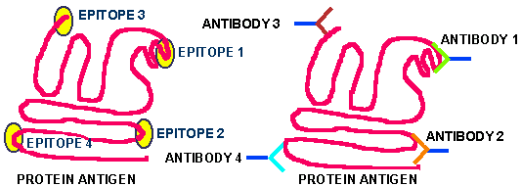


http://atlas.med.harvard.edu/cgi-bin/alignace.pl

Antigen Epitope Prediction

Antigens and Epitopes

- Antigens are molecules that induce immune system to produce antibodies
- Antibodies recognize parts of molecules called epitopes



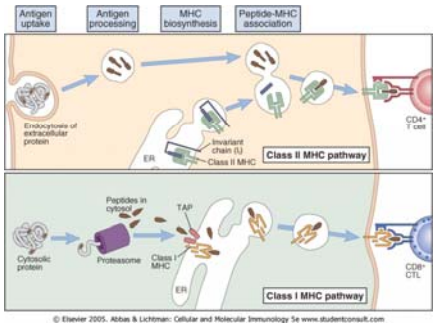
Genome to "Immunome"

Pathogen genome sequences provide define all proteins that could illicit an immune response

- Looking for a needle…
  - Only a small number of epitopes are typically antigenic
- …in a very big haystack
  - *Vaccinia* virus (258 ORFs): 175,716 potential epitopes (8-, 9-, and 10-mers)
  - *M. tuberculosis* (~4K genes): 433,206 potential epitopes
  - *A. nidulans* (~9K genes): 1,579,000 potential epitopes

Can computational approaches predict all antigenic epitopes from a genome?

Antigen Processing & Presentation

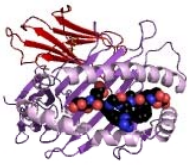


Modeling MHC Epitopes

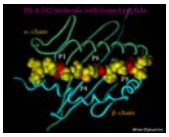
- Have a set of peptides that have been associate with a particular MHC allele
- Want to discover motif within the peptide bound by MHC allele
- Use motif to predict other potential epitopes

Motifs Bound by MHCs

- MHC 1
  - Closed ends of grove
  - Peptides 8-10 AAs in length
  - *Motif is the peptide*



- MHC 2
  - Grove has open ends
  - Peptides have broad length distribution: 10-30 AAs
  - ***Need to find binding motif within peptides***



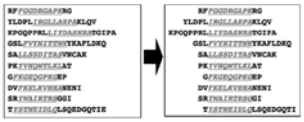
MHC 2 Motif Discovery

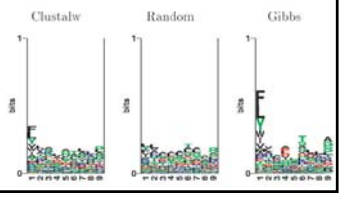
Use Gibbs Sampling!

462 peptides known to bind to MHC II HLA-DR4(B1*0401)

9-30 residues in length

Goal: identify a common length 9 binding motif





Nielsen et al (2004) Bioinf

Vaccinia Epitope Prediction

Mutaftsi et al (2006) Nat. Biotech.

- Predict MHC1 binding peptides
- Using 4 matrices for H-2 Kb and Db
- Top 1% predictions experimentally validated

49 validated epitopes accounting for 95% of immune response