Sequence Analysis '17
lecture 20

MEME
Gibbs sampling
K-means
modular HMMs
Gene finding
Why are there Motifs?

Selective pressure for:

**structure** -- protein motifs
  - folding units
  - fibrous proteins
  - coiled coils
  - transmembrane helices

**function** -- protein motifs
  - active site
  - binding motifs
  - signal sequences

**expression** -- DNA motifs
  - transcription regulation
  - chromatin binding
Zinc finger motif

Example: selection for structure

two Cystines separated by 2 or 4 residues

two Histidines separated by 3 or 5 residues

\[ C-x(2,4)-C-x(3)-[LIVMFYW-C]-x(8)-H-x(3,5)-H \]

Loop must be length 12.
4th position in loop must be hydrophobic
What if you don’t know the pattern?

...and you also don’t know where it is....?
motif elucidation by maximizing expectation

1. Calculate the motif model given the locations of the motif.
2. Calculate the locations of the motif given the motif model.
3. Repeat until converged.

• T. L. Bailey & C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", ISMB, 2:28--36, 1994
Initial guess of motif location

From the motif locations, you make a profile model.

Motif Model: $L=4$

$P_1 = 2/3 \text{ T, } 1/3 \text{ G}$

Initial guesses underlined
Calculate the probability score for each position.

From the profile model and the sequence, get probability scores.

AGCTAGCTTTCTCGTGA

\[ P = P_1(A)P_2(G)P_3(C)P_4(T) = (0)(0.33)(0.67)(0) = 0. \]
Calculate the probability score for each position

Slide the model along the sequence to get the next score.

\[
P = P_1(G)P_2(C)P_3(T)P_4(A) = (.33)(.67)(.33)(0.) = 0.
\]
Calculate the probability score for each position

Slide the model along the sequence to get the next score.

\[ P = P_1(C)P_2(T)P_3(A)P_4(G) = (0.0)(0.0)(0.0)(0.67) = 0. \]
Calculate the probability score for each position

Slide the model along the sequence to get the next score.

\[ P = P_1(T)P_2(C)P_3(T)P_4(C) = (0.67)(0.67)(0.33)(0.33) = 0.05 \]
Calculate the probability score for each position

Slide the model along the sequence to get the next score.

\[ P = P_1(G)P_2(T)P_3(G)P_4(A) = (0.33)(0)(0)(0) = 0. \]
Calculate the probability score for each position

Do every sequence.
Calculate the probability score for each position

Do every sequence.
Re-Calculate the motif model

Probabilities are normalized to sum to one for each sequence, since we expect exactly one motif per sequence.

The new model is the profile built from the hits.
Recalculating the profile from the hits

$P_1(T) = \text{the probability of T in the first position} = \text{the sum of the scores for sequences with T in the first position, normalized.}$

$P_1(T) = \frac{1.0 + 0.5 + 0.3 + 0.6}{1.0 + 0.5 + 0.5 + 0.1 + 0.3 + 0.6} = 0.8$
Do it again: Re-calculate the probability scores using the refined model.

The new model is the profile built from the hits.
...and again, until converged.
EM converges on the conserved pattern if the initial guess was not too far off.

A summary of the exercise:

If the true motif was not one of the initial guesses, or some combination of the initial guesses, then EM would never find the true motif.
Pseudocounts, just in case

No A is observed in the first position, but if we set $P(A) = 0$, then we “rule out” a motif with A in the first position. Instead, $P_1(A) = \frac{\varepsilon}{\text{sum of weights}}$.

This is especially important in the initial guesses, so that the true motif is not missed.

$P_1(T) = \frac{\varepsilon}{1.0+0.5+0.5+0.1+0.3+0.6} = 0.8$

Pseudocounts may be decreased or removed ($\varepsilon=0$) in later stages.
Gibbs Sampling

Stochastic version of MEME.

Radius of convergence is wider than MEME. Doesn’t need to start with one correct guess.
GIBBS sampling

motif elucidation by maximizing expectation

1. Calculate the motif model given the locations of the motif.
2. Calculate the locations of the motif given the motif model.
3. Repeat until converged.

Expectation step

Start from random alignment. Select window size and position. Slide one sequence through window. Calculate scores.

Slide first sequence through the motif window, calculate score.
Expectation step

AGCTAGCTTCTCGTG
TCTCGAGTGGCGCATG
TATTGCTCTCCGCAGC

score
aligned position
Example

AGCTAGCTTCTCGTGA
TCTCGAGTGGCGCATG
TATTGCTCTCCGCAGC

Select an aligned position at random from the score distribution.

score
aligned position

Do next sequence, and so on, cycling through the sequences many times.
Example

AGCTAGCTTTCTCGTGA
TCTCGAGTGGCGCATG
TATTGCTCTCCGCAGC

Select an aligned position at random from the score distribution.

Do next sequence, and so on, cycling through the sequences many times.
Convergence is when there are no more changes.

Exactly one segment is aligned to the motif region at each step.
Gibbs Sampling

Stochastic version of MEME.

1. Choose initial (or random) guesses of motif locations.

2. Calculate the motif model (w/ or w/o pseudocounts/noise) from the current motif positions.

3. Slide one sequence against the model. Calculate probability score at each position.

4. Randomly choose a motif position from the probability distribution.

5. Repeat until converged.

Radius of convergence is wider than MEME. Doesn’t need to start with a correct guess.
Example:

Transcription factor binding sites

We know that upstream of every gene that expresses in the presence of lysine, there must be a signal for a transcription factor (TF). But we don’t know which TF and we don’t know where it binds.
Transcription factor binding site

Example: selection for expression

This is what we want to find

NOTE: It’s palindromic in this case. Palindromy in TF footprints (binding sites) is due to the symmetry of the TFs, which are almost invariably dimeric.
What is Expectation/Maximization?

EM is any method that iterates between an “expectation” step and a “maximization” step. Starting with a statistical model and a set of data.

• **Expectation**
  Calculate the expected values for the parameters of the model, using the current model and the data.

• **Maximization**
  Replace the parameters of the model with their expected values.

MEME is an EM algorithm. Gibbs sampling is not.
Finding recurrent sequences by Kmeans clustering

Short, recurrent sequence patterns may exist in different protein because they are required to initiate folding.

Is is a recurrent structure?

Step 1: Compile a database of sequence profiles from MSAs

Step 2: Cluster protein sequence profiles

Each dot represents a segment of a profile from a MSA from a BLAST search

Each circle represents a cluster of profiles (same length)
K-means clustering

(1) Choose K.

(2) Randomly select K centers in the metric space.

(3) Get the distance from each center to each data point.

(4) Assign each data point to the nearest center.

(5) Calculate the new centers using the center-of-mass of the data points.

(6) Repeat from Step 3 until converged.

Final positions of the centers define K clusters of data points.
Example: $K=2$
Example: $K=2$
Example: $K=2$
Example: $K=2$
Example: $K=2$
Example: K=2
Example: $K=2$
Example: K=2

Final cluster centers

no change. Converged.
K-means clustering

(1) Requires knowledge of K, the number of classes
(2) Requires the objects to exist in a “metric space”.
(3) Stochastic. Depends on the starting points.
Distance/similarity metrics for profiles. We tried them all.

(1) Manhattan, or City-Block metric (a distance metric)

$$D(p, q) = \sum_{\text{positions}} \sum_{\text{amino acids}} \left| P(p_{ij}) - P(q_{ij}) \right|$$

(2) Entropy (a similarity metric)

NOTE: not symmetrical!

$$S(p, q) = \sum_{\text{positions}} \sum_{\text{amino acids}} p_{ij} \log(q_{ij})$$

(3) Correlation (a similarity metric)

$$S(p, q) = \frac{\sum_{\text{positions}} \sum_{\text{amino acids}} (p_{ij} - \langle p \rangle)(q_{ij} - \langle q \rangle)}{\sqrt{\sum_{\text{positions}} \sum_{\text{amino acids}} (p_{ij} - \langle p \rangle)^2 \sum_{\text{positions}} \sum_{\text{amino acids}} (q_{ij} - \langle q \rangle)^2}} = \frac{\sum_{\text{positions}} \sum_{\text{amino acids}} (p_{ij} - \langle p \rangle)(q_{ij} - \langle q \rangle)}{\sigma_p \sigma_q}$$

(4) Dpq (similarity metric)

$$D(p, q) = \sum_{\text{positions}} \sum_{\text{amino acids}} \text{LLR}(p_{ij})\text{LLR}(q_{ij})$$
Step 3: Refine using Supervised learning

Supervised learning finds predictive correlations between two spaces (sequence space and structure space)

Sequence profile

Remove all cluster members that do not conform with the paradigm

Training set

Search the database for the 400 nearest neighbors

Nearest neighbors

…as long as it is consistent with this structure.

We want this profile to predict...
Results: I-sites library

diverging type-2 turn

Serine hairpin

Amino acids arranged from non-polar to polar

Backbone angles: $\psi$=green, $\phi$=red

Proline helix C-cap

alpha-alpha corner

Type-I hairpin

Frayed helix

glycine helix N-cap
How do you compare two predictive models?

**Accuracy** = percent of the predictions that are correct, of the ones that were made.  
OK. But accuracy depends on coverage

**Coverage** = number of possible predictions that were actually predicted.  
OK. But coverage depends on accuracy

**Confidence** = a score to sort the predictions. A more confident prediction should be a more accurate one.  
Does the confidence of a prediction match its accuracy?

<table>
<thead>
<tr>
<th>Selectivity = Accuracy</th>
<th>Sensitivity = Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T^+/(T^+ + F^+)$</td>
<td>$T^+/(T^+ + F^-)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>T</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>T+</td>
<td>F-</td>
</tr>
<tr>
<td>F</td>
<td>T-</td>
</tr>
</tbody>
</table>

+ | - |
Receiver Operator Characteristic (ROC)

• A way to describe the whole set of scores with a single number.
• Each score has a T or F.
• Sort the scores.
• Starting from the highest scoring, draw a vector up for a true, to the right for a false.
• Calculate ROC = the normalized area under this curve.
• If all of the true scores are greater that the greatest false score, then ROC = 1.0.
• $0.\leq\text{ROC}\leq1$. 

Better way to compare two predictive models
Sort the scores, for each score move up one if it is true, right one if it is false.

The area under the curve, divided by the total, is the ROC score. $0 \leq \text{ROC} \leq 1$. 
In class exercise: calculate ROC score

<table>
<thead>
<tr>
<th>Method A</th>
<th>Method B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.811 T</td>
<td>4 T</td>
</tr>
<tr>
<td>0.972 T</td>
<td>39 F</td>
</tr>
<tr>
<td>0.766 T</td>
<td>44 T</td>
</tr>
<tr>
<td>0.990 F</td>
<td>44 T</td>
</tr>
<tr>
<td>0.966 T</td>
<td>44 T</td>
</tr>
<tr>
<td>0.951 F</td>
<td>40 T</td>
</tr>
<tr>
<td>0.803 F</td>
<td>1 F</td>
</tr>
<tr>
<td>0.792 F</td>
<td>39 F</td>
</tr>
<tr>
<td>0.503 F</td>
<td>29 F</td>
</tr>
<tr>
<td>0.978 T</td>
<td>10 F</td>
</tr>
<tr>
<td>0.478 F</td>
<td>44 F</td>
</tr>
</tbody>
</table>

Which method is better?
Motifs can be words in a word-HMM

I-sites word HMM

= HMMSTR!

Hidden Markov Model for local protein STRucture

HMM of linked I-sites motifs. Each node is one amino acid.

Size of HMMSTR:
282 nodes
317 transitions

(Bystroff et al., JMB 2000)
Splicing
Gene finding
HMMs can be modular
Gene Prediction: Computational Challenge

Thanks to Pavel Pevzner (UCSD) and Chris Burge (MIT) for slides and images.
Two Approaches to Gene Prediction

- **Statistical**: coding segments (ORFs or exons) have typical sequences on either end and use different subwords than non-coding segments (introns or intergenic regions).

- **Similarity-based**: many human genes are similar to genes in mice, chicken, or even bacteria. Therefore, already known mouse, chicken, and bacterial genes may help to find human genes.
Gene Prediction and Motifs

- Upstream regions of genes often contain motifs that can be used for gene prediction.

- Reading frame should be “long enough”.

[Diagram showing transcription start site, TATA box, ribosomal binding site, and ATG and stop codon positions.]

TTCCAA
TATACT
Pribnow Box
Transcription start site

-35 -10 0 10 ATG

GGAGG
Ribosomal binding site

STOP
Promoter Structure in Prokaryotes (E.Coli)

- Pribnow Box (-10)
- Gilbert Box (-30)
- Ribosomal Binding Site (+10)

Transcription starts at offset 0.

Consensus sequences:

Pribnow box:

```
TTGACA
AACTGT
```

5' -36 -31 3' 5' -12 -7 3' 5' +1 +20 3'

mRNA:

```
```

AUG
Ribosomal Binding Site

1055 E. coli Ribosome binding sites listed in the Miller book
Central Dogma of biology

- **DNA**
  - CCTGAGCCAACTATTGATGAA

- **RNA**
  - CCUGAGCCAAACUUAUGAUUGAA

- **Protein**

- **Peptide**
Discovery of Split Genes

- In 1977, Phillip Sharp and Richard Roberts experimented with mRNA of *hexon*, a viral protein.
  - Map hexon mRNA in viral genome by hybridization to adenovirus DNA and electron microscopy
  - mRNA-DNA hybrids formed three curious loop structures instead of contiguous duplex segments
Expanded Central Dogma

- Exon 1
- Exon 2
- Exon 3
- Intron 1
- Intron 2

**Gene Structure**

- Replication
- Transcription
- Splicing
- Translation
- Folding

**Exon** = coding

**Intron** = non-coding
Splicing mechanism

**Spliceosome:** Def: A ribonucleoprotein complex, containing RNA and small nuclear ribonucleoproteins (snRNPs) that is assembled during the splicing of messenger RNA primary transcript to excise an intron.

(http://genes.mit.edu/chris/)
Splicing mechanism

Sans spliceosome

(1) pre-mRNA

(2) lariat loop forms

(3) exon 1 cleaved from lariat.
Splicing mechanism

(4) exon 1,2 positioned to ligate

(5) liberated lariat + ligated exons

gets degraded

Spliceosome (not shown) disassociates.

goes to ribosome
Splicing mechanism on the web

http://neuromuscular.wustl.edu/pathol/diagrams/splicefunct.html

google: wustl neuromuscular splicefunct

http://neuromuscular.wustl.edu/pathol/diagrams/splicemech.html

google: wustl neuromuscular splicemech

Much thanks to T. Wilson, UCSC!
Evolutionary models for introns

How to find splice points, using BLAST.

(1) Translate the DNA in all 6 frames.

(2) Search the database of protein sequences using the translations (blastx).

(3) Using the complete protein sequence, align it to the translation and find the regions of (near) perfect identity. These will abruptly end at the intron start site.

(4) Find the 5’-GT or 3’-AG signal at the point where the identity matches abruptly end.

(5) If your translation has an insertion with nearly perfect matches on either side, you have an alternative splicing.
In Class exercise: find the alternative spliced variants

Go to NCBI, search nucleotides for AKAP9 (you should get the sequence with accession number NM_005751.4, GI:197245395)

Select “BLAST sequence”

Select blastx (not tblastx)

Select the nr/nt database. Organism: homo sapiens

Submit.

While waiting, do exercise on the next page....
A sure sign of alternative splicing in blastx output:

Score = 160 bits (404), Expect = 8e-37
Identities = 85/116 (73%), Positives = 86/116 (74%)
Frame = +2

Query: 76820 RSHENGFMEDLDKTVRFQECDSRSNAPATLTENMAGAFSFIHSRVGSPWXXXXXX 76999
      +SHENGFMEDLDKTVRFQECDSRSNAPATLTENMA
Sbjct: 778 KSHENGFMEDLDKTVRFQECDSRSNAPATLTENMA------------------------ 814

Query: 77000 XXXXRHTGVFMLVAGGIVAGIFLIFEIAYKRHKDARRKQMQLAFAAVNVWRKNLQ 77167
       GVFMLVAGGIVAGIFLIFEIAYKRHKDARRKQMQLAFAAVNVWRKNLQ
Sbjct: 815 --------GVFMLVAGGIVAGIFLIFEIAYKRHKDARRKQMQLAFAAVNVWRKNLQ 863

Identical up to the insertion. Identical after the insertion. These must be the same gene.
HMM for pre-mRNA

5'UTR

AUG

exon

stop (UAA | UAG | UGA)

5' splice site (GU...)

3' splice site (...AG)

intron (0|1|2)

3'UTR

exon

intron

exon

intron

intron

XXXATG...XXXGUX...XXAGXX...XXGUX...XXAGXX...XXTAAXXX
Intron model for mammals

A genefinding HMM: Genescan

Internal exon model

Intron models

Initial exon model

Terminal exon model

Single exon model

Intergenic Regions

Mirrored models for reverse complement strand
Exon models, frame 1

Sequence enters the sub-model after first base in the codon, cycles through non-stop codons (Stop codons have zero transition probability), and transitions with small probability $p$ to final base, exiting in frame 1.
GENESCAN -- forward strand part

- Internal exon 0-frame: $E^0$
- Exon 1-frame: $E^1$
- Exon 2-frame: $E^2$

- Donor splice site
- Acceptor splice site
- Non-emitting state

- Initial exon 0-frame: $E_{\text{init}}^0$
- Terminal exon: $E_{\text{term}}$

- Intron (short): $I_{\text{short}}^0$
- Intron (long): $I_{\text{short}}^2$

- Non-emitting state

- Transition probabilities:
  - $I_{\text{short}}^0 \rightarrow I_{\text{short}}^{1}$: 0.9
  - $I_{\text{short}}^1 \rightarrow I_{\text{short}}^{2}$: 0.9
  - $I_{\text{short}}^2 \rightarrow I_{\text{short}}^0$: 0.9
  - $I_{\text{short}}^0 \rightarrow I_{\text{short}}^1$: 0.1
  - $I_{\text{short}}^1 \rightarrow I_{\text{short}}^2$: 0.1
  - $I_{\text{short}}^2 \rightarrow I_{\text{short}}^0$: 0.1

- Additional states:
  - $DSS^0$
  - $ASS^0$
  - $DSS^1$
  - $ASS^1$
  - $DSS^2$

- GENESCAN -- forward strand part
TwinScan

- Aligns two sequences and marks each base as gap ( - ), mismatch ( : ), match ( | ) , resulting in a new alphabet of 12 letters: \( \Sigma \{ A-, A:, A |, C-, C:, C |, G-, G:, G |, T-, T:, T| \} \).

- Run Viterbi algorithm using emissions \( e_k(b) \) where \( b \in \{ A-, A:, A|, \ldots, T| \} \).
TwinScan

Exon (E) state emits matches (Nl), Intron state (I) emits mismatches (N:) and gaps (N-). Gamma values (forward backward values) predict introns, exons. May be combined with motif models for intron, codon models.
Splicing facts

**Exons** average 145 nucleotides in length
- Contain regulatory elements:
  - ESEs: Exonic splicing enhancers
  - ESSs: Exonic splicing silencers

**Introns** average more than 10x longer than exons
- Contain regulatory elements (bind regulatory complexes)
  - ISEs: Intronic splicing enhancers
  - ISSs: Intronic splicing silencers

**Splice sites**
- 5' splice site
  - Sequence: AGguragu (r = purine)
  - U1 snRNP: Binds to 5' splice site
- 3' splice site
  - Sequence: yyyyyyy nagG (y = pyrimidine)
- Branch site
  - Sequence: ynyuraray (r = purine)
  - U2 snRNP: Binds to branch site via RNA:RNA interactions between snRNA and pre-mRNA
Alternative splicing

**Definition:** Joining of different 5' and 3' splice sites

- ~80% of alternative splicing results in changes in the encoded protein
- Up to 59% of human genes express more than one mRNA by alternative splicing

**Functional effects:** Generates several forms of mRNA from single gene

- Allows functionally diverse protein isoforms to be expressed according to different regulatory programs

**Structural effects:**

- Insert or remove amino acids
- Shift reading frame
- Introduce termination codon

**Gene expression effects**

- Removes or inserts regulatory elements controlling translation, mRNA stability, or localization

**Regulation**

- Splicing pathways modulated according to:
  - Cell type
  - Developmental stage
  - Gender
  - External stimuli
Review

1. What information is used to find genes in prokaryotes?
2. How were introns discovered?
3. What is the better way to find genes: by homology, or by statistical models?
4. Where is the ribosome binding site relative to the transcription start site?
5. Where is the Pribnow Box? Gilbert box?
6. How can HMMs be linked together to make a larger model?
7. What happens to the intron once it is removed?
8. What is alternative splicing?
9. What do DNA alignments tell you about introns/exons?
10. What is a splicing enhancer? silencer?
Review

Why are there motifs in proteins?
What does MEME do?
How is Gibbs sampling different from MEME?
What kind of data can be clustered using K-means?
What is supervised learning? How is it different from machine learning?
What kind of data goes into a ROC score?
How to find motifs, signatures, footprints

MEME -- *deterministic* EM algorithm for motif finding, needs initial guess.

Gibbs sampling -- *stochastic* EM algorithm for motif finding, doesn’t need initial guess.

K-means -- unsupervised learning of recurrent patterns, requires a metric space (distance or similarity).

Supervised learning -- EM. Expectation in one "space", maximization in the other.

Method for comparing prediction methods

ROC

Aligning (or not) low complexity sequences

masking

null models for repeats

word HMMs