Part III. Systems Biology:  
7. Motif Finding

Lecture 20 – Mar 12, 2015  
CSE 427 Computational Biology  
Instructor: Su-In Lee  
TAs: Safiye Celik  
TTh 12:00-1:20 @ MGH 238

Outline (3/10 & 3/12)
- Clustering in gene expression data
  - K-means clustering
- Motif finding Background
- Motif representation
- Commonly used methods
  - Enumeration
  - Expectation-Maximization methods (MEME)
  - Greedy search method (CONSENSUS)

Challenge problem
- Find a motif in a sample of
  - 20 "random" sequences (e.g. 600 nt long)
  - each sequence containing an implanted pattern of length 15
  - each pattern appearing with 4 mismatches as (15,4)-motif.

Identifying motifs
- Genes are turned on or off by regulatory proteins (TFs).
- TFs bind to upstream regulatory regions of genes to either attract or block an RNA polymerase
- So, multiple genes that are regulated by the same TF will have the same motifs in their regulatory regions.
- How do we identify the genes that are regulated by the same TF?
Outline

- Clustering in gene expression data
  - K-means clustering
- Motif finding Background
  - Motif representation
- Commonly used methods
  - Enumeration
  - Expectation-Maximization methods (MEME)
  - Greedy search method (CONSENSUS)

Structural basis of interaction

- Key Feature:
  - Transcription factors are not 100% specific when binding DNA
- Not one sequence, but family of sequences, with varying affinities

Motif representation

- Structural discussion immediately raises difficulties
- Least expressive: \(\text{GACCG}\)
- Most expressive:
  - \(4^k\)-dimensional probability distribution
  - Independently assign probability for each of the possible \(k\)-mers

A specific \(n\)-tuple of nucleic acid that can be used to identify certain regions within DNA or proteins.
Outline (3/10 & 3/12)

- Clustering in gene expression data
  - K-means clustering
- Motif finding Background
- Motif representation
- Commonly used methods
  - Enumeration
  - Expectation-Maximization methods (MEME)
  - Greedy search method (CONSENSUS)

Finding regulatory motifs

- Say a transcription factor (TF) controls five different genes
- Each of the five genes will have binding sites for the TF in their promoter region

Finding regulatory motifs

- Given the upstream sequences of the genes that seem to be regulated by the same TFs,
- Find the TF-binding sites (motifs) in common

Identifying motifs: complications

- We do not know the motif sequence
- We do not know where it is located relative to the genes
- Motifs can differ slightly from one gene to another
- How to discern it from “random” motifs?
Problem statements:
- Given a set of promoters of \( n \) co-regulated genes, find a motif common to the promoters/
- Both the PWM and the motif sequences are unknown.

Enumeration (simplest method)
- Look at the frequency of all k-mers*

EM algorithm (MEME)
- Iteratively learn the most likely motif model

Gibbs sampling methods
- AlignAce, BioProspector

Generating k-mers
- Example (5-mers):
  at gacggaat gat ac cg tttg gct ac at t gtt a a ac
Example: MEME

1. MEME uses an initial EM heuristic to estimate the best starting-point PWM matrix:

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.26</td>
<td>0.24</td>
<td>0.18</td>
<td>0.26</td>
</tr>
<tr>
<td>A</td>
<td>0.24</td>
<td>0.26</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>T</td>
<td>0.25</td>
<td>0.23</td>
<td>0.30</td>
<td>0.25</td>
</tr>
<tr>
<td>C</td>
<td>0.25</td>
<td>0.27</td>
<td>0.24</td>
<td>0.25</td>
</tr>
</tbody>
</table>

2. MEME scores the match of all 6-mers to current matrix

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.29</td>
<td>0.24</td>
<td>0.17</td>
<td>0.27</td>
</tr>
<tr>
<td>A</td>
<td>0.22</td>
<td>0.26</td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td>T</td>
<td>0.24</td>
<td>0.23</td>
<td>0.33</td>
<td>0.23</td>
</tr>
<tr>
<td>C</td>
<td>0.24</td>
<td>0.27</td>
<td>0.23</td>
<td>0.28</td>
</tr>
</tbody>
</table>

3. Re-estimate the PWM based on the weighted contribution of all 6-mers.

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.40</td>
<td>0.20</td>
<td>0.15</td>
<td>0.42</td>
</tr>
<tr>
<td>A</td>
<td>0.30</td>
<td>0.30</td>
<td>0.20</td>
<td>0.24</td>
</tr>
<tr>
<td>T</td>
<td>0.15</td>
<td>0.30</td>
<td>0.16</td>
<td>0.15</td>
</tr>
<tr>
<td>C</td>
<td>0.15</td>
<td>0.20</td>
<td>0.16</td>
<td>0.15</td>
</tr>
</tbody>
</table>

4. MEME scores the match of all 6-mers to current matrix

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.85</td>
<td>0.05</td>
<td>0.10</td>
<td>0.80</td>
</tr>
<tr>
<td>A</td>
<td>0.05</td>
<td>0.60</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>T</td>
<td>0.05</td>
<td>0.30</td>
<td>0.70</td>
<td>0.05</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

5. Re-estimate the PWM based on the weighted contribution of all 6-mers.

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.85</td>
<td>0.05</td>
<td>0.10</td>
<td>0.80</td>
</tr>
<tr>
<td>A</td>
<td>0.05</td>
<td>0.60</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>T</td>
<td>0.05</td>
<td>0.30</td>
<td>0.70</td>
<td>0.05</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

6. MEME scores the match of all 6-mers to current matrix

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>0.85</td>
<td>0.05</td>
<td>0.10</td>
<td>0.80</td>
</tr>
<tr>
<td>A</td>
<td>0.05</td>
<td>0.60</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>T</td>
<td>0.05</td>
<td>0.30</td>
<td>0.70</td>
<td>0.05</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Iterations continue until convergence

Numbers do not change much between iterations

Final motif
Using EM algorithm

- MEME works by iteratively refining PWMs and identifying sites for each PWM
  1. Estimate motif model (PWM)
     - Start with a k-mer seed (random or specified)
     - Build a PWM by incorporating some of background frequencies
  2. Identify examples of the model
     - For every k-mer in the input sequences, identify its probability given the PWM model.
  3. Re-estimate the motif model
     - Calculate a new PWM, based on the weighted frequencies of all k-mers in the input sequences
  4. Iterate 2 & 3 until convergence.

Outline

- Clustering in gene expression data
  - K-means clustering
- Motif finding Background
  - Motif representation
- Commonly used methods
  - Enumeration
  - Expectation-Maximization methods (MEME)
  - Greedy search method (CONSENSUS)

CONSENSUS

- Hertz and Stormo, Bioinformatics 1999
- Popular algorithm for motif discovery, that uses a greedy approach
- **Motif model**: Position Weight Matrix (PWM)
- **Motif score**: information content

Information content

- PWM $W$:
  - $W_{\beta k} = $ frequency of base $\beta$ at position $k$
  - $q_\beta = $ frequency of base $\beta$ by chance
  
- Information content of $W$:
  \[
  \sum_k \sum_{\beta \in \{A,C,G,T\}} W_{\beta k} \log \frac{W_{\beta k}}{q_\beta}
  \]
Information content

- If $W_{\beta k}$ is always equal to $q_{\beta}$, i.e., if $W$ is similar to random sequence, information content of $W$ is 0.
- If $W$ is different from $q$, information content is high.
- Information content of $W$:
  \[
  \sum_{k} \sum_{\beta \in \{A,C,G,T\}} W_{\beta k} \log \frac{W_{\beta k}}{q_{\beta}}
  \]

CONSENSUS: Basic idea

- Find a set of subsequences, one in each input sequence.
- Set of subsequences define a PWM.
- Goal: This PWM should have high information content.
- High information content means that the motif "stands out".

CONSENSUS: Greedy heuristic

- Suppose we have built a partial set of subsequences $\{s_1, s_2, ..., s_i\}$ so far.
- Have to choose a subsequence $s_{i+1}$ from the input sequence $S_{i+1}$.
- Consider each subsequence $s$ of $S_{i+1}$.
- Compute the score (information content) of the PWM made from $\{s_1, s_2, ..., s_i, s\}$.
- Choose the $s$ that gives the PWM with highest score, and assign $s_{i+1} \leftarrow s$.

### Algorithm

1. Start with a subsequence in one input sequence.
2. Build the set of subsequences incrementally, adding one subsequence at a time.
3. Until the entire set is built.

---

| $s_1$ | $s_2$ | $s_3$ | $:$ | $s_i$ |
Outline

- Background
  - Many classes of algorithms, differ in
    - Types of input data
    - Motif representation
  - Commonly used methods
    - Enumeration
    - Expectation-Maximization methods (MEME)
    - Gibbs sampling methods (AlignAce, BioProspector)
    - Greedy search method (CONSENSUS)

Cell = factory, proteins = machines

Gene expression

- Instruction for making the proteins
Gene expression

- Instruction for making the proteins
- Instruction for when and where to make them

“Coding” regions

“Regulatory” regions (regulons)

- What turns genes on (producing a protein) and off?
- When is a gene turned on or off?
- Where (in which cells) is a gene turned on?
- How many copies of the gene product are produced?

DNA regulation

Source: Richardson, University College London

Gene expression

- Instruction for making the proteins
- Instruction for when and where to make them

“Coding” regions

“Regulatory” regions (regulons)

- Regulatory regions contain “binding sites” (6-20 bp).
- “Binding sites” attract “transcription factors”.
- Bound transcription factors can initiate transcription.
- Proteins that inhibit transcription can also bind to their binding sites.

Structural basis of interaction
Structural basis of interaction

- **Key feature:**
  - Transcription factors are not 100% specific when binding DNA, because non-essential bases could mutate.
  - Not one sequence, but family of sequences, with varying affinities.

```
<table>
<thead>
<tr>
<th>Sequence</th>
<th>Affinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAGCC</td>
<td>0.54</td>
</tr>
<tr>
<td>CAGCC</td>
<td>0.48</td>
</tr>
<tr>
<td>CACTG</td>
<td>0.32</td>
</tr>
<tr>
<td>CAGCA</td>
<td>0.25</td>
</tr>
<tr>
<td>GGCCT</td>
<td>0.11</td>
</tr>
<tr>
<td>GGCTG</td>
<td>0.08</td>
</tr>
</tbody>
</table>
```

What is a motif?

- A subsequence (substring) that occurs in multiple sequences with a biological importance.
- Motifs can be totally constant or have variable elements.
- DNA motifs (regulatory elements)
  - Binding sites for proteins
  - Short sequences (5-25)
  - Up to 1000 bp (or farther) from gene
  - Inexactly repeating patterns

Motif finding

- Basic objective:
  - Find regions in the genome that transcription factors bind to.
- Motivations
  - Understanding which TFs regulate which genes
  - Major part of the gene regulation

Outline

- Background
  - Many classes of algorithms, differ in
    - Types of input data
    - Motif representation
- Commonly used methods
  - Enumeration
  - Expectation-Maximization methods (MEME)
  - Greedy search method (CONSENSUS)
Input data - single sequence

- Single sequence
  
  AGCATCAGCATCATCACTTACCACTTACCACTCACCATG
  
  AGCATCAGCATCATCACTTACCACTTACCACTCACCATG
  
  AGCATCAGCATCATCACTTACCACTTACCACTCACCATG

- Based on over-representation of short sequences

Random sample

Input data - single sequence

- Single sequence
  
  AGCATCAGCATCATCACTTACCACTTACCACTCACCATG
  
  AGCATCAGCATCATCACTTACCACTTACCACTCACCATG
  
  AGCATCAGCATCATCACTTACCACTTACCACTCACCATG

- Based on over-representation of short sequences

Random sample

Implanting motif AAAAAAGGGGGGG

Where is the implanted motif?
Implanting Motif $AAAAAAGGGGG$ with four mutations - (15,4)-motif

$aaaaaaaaggggg$  (15,4)-motif

Where is the motif ???

With four mutations - (15,4)-motif

Challenge problem

- Find a motif in a sample of
  - 20 "random" sequences (e.g. 600 nt long)
  - each sequence containing an implanted pattern of length 15
  - each pattern appearing with 4 mismatches as (15,4)-motif.
Input data

- Single sequence
  ```
  ...AGCATCAGCAGCA
  CATCATCAGCATACGACTCAGCATAGCCATGGGCTA
  CAGCA
  GATCGATCGAA
  CAGCA
  CG...
  ```

- Sequence + other data
  - Gene expression data
  - ChIP-seq
  - Others...

Identifying motifs

- Genes are turned on or off by regulatory proteins (TFs).
- TFs bind to upstream regulatory regions of genes to either attract or block an RNA polymerase
- So, multiple genes that are regulated by the same TF will have the same motifs in their regulatory regions.
- How do we identify the genes that are regulated by the same TF?

Sequence + gene expression data

- Say that a microarray experiment showed that when gene X is knocked out, 20 other genes are not expressed.
  - How can one gene have such drastic effects?
- Say that 5 different genes are co-expressed across many experiments in a gene expression data.
  - These genes are likely to share the same binding sites.

daf-19 binding sites in *C. elegans*

- Motifs and transcriptional start sites
  ```
  GTTGTGATGGGTGAC
  GTTTCCATGGAAAC
  GCTACCATGGCAAC
  GTTCCCATAGTAAC
  GTTTCCATGTGAAAC
  ```

source: Peter Swoboda
Input data

- Single sequence
  - ...GCATACGGCAAGCAGCATAGATCGATCGATCGATCGATCGACGGTACAGCA...
- Sequence + other data
  - Gene expression data
  - ChIP-seq
  - Others...
- Evolutionarily related set of sequences

Outline

- Background
- Many classes of algorithms, differ in
  - Types of input data
  - Motif representation
- Commonly used methods
  - Enumeration
  - Expectation-Maximization methods (MEME)
  - Gibbs sampling methods (AlignAce, BioProspector)
  - Greedy search method (CONSENSUS)

Outline

- Background
- Many classes of algorithms, differ in
  - Types of input data
  - Motif representation
- Commonly used methods
  - Enumeration
  - Expectation-Maximization methods (MEME)
  - Gibbs sampling methods (AlignAce, BioProspector)
  - Greedy search method (CONSENSUS)

Structural basis of interaction

- Key Feature:
  - Transcription factors are not 100% specific when binding DNA
- Not one sequence, but family of sequences, with varying affinities

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Affinity</th>
</tr>
</thead>
<tbody>
<tr>
<td>GACCG</td>
<td>0.54</td>
</tr>
<tr>
<td>GACCG</td>
<td>0.48</td>
</tr>
<tr>
<td>GACTG</td>
<td>0.32</td>
</tr>
<tr>
<td>GACCA</td>
<td>0.25</td>
</tr>
<tr>
<td>GCCCG</td>
<td>0.11</td>
</tr>
<tr>
<td>GCTCG</td>
<td>0.08</td>
</tr>
</tbody>
</table>
Motif representation

- Structural discussion immediately raises difficulties

- Least expressive: GACCG
  - Single sequence

- Most expressive:
  - 4^k-dimensional probability distribution
  - Independently assign probability for each of the possible k-mers

A specific n-tuple of nucleic acid that can be used to identify certain regions within DNA or proteins.

Motif logos

- The motif logo shows how well bases are conserved at each position.
- The higher the number of conserved bases, the higher the letters are.
- The height of the entire stack of the bases is the information measured in bits.

Motif logos

- Standard Solution:
  - Position-Specific Scoring Matrix (PSSM) or Position Weight Matrix (PWM)
  - Assuming independence of positions, assign a probability distribution for each position

### Motif logos

- The height of the entire stack of the bases is the information measured in bits.

\[
\text{height}_{\text{base}} = f_{a,i} \times \left[2 \left(1 - \left(H_i + c_{a,i}\right) \right)^{1/2}\right]
\]

- The approximation for the small-sample correction

\[
H_i = -\sum f_{a,i} \log_2 f_{a,i}
\]

- Uncertainty (Shannon entropy)

A specific n-tuple of nucleic acid that can be used to identify certain regions within DNA or proteins.
Oversimplicity of PSSMs

- PSSM might be a too simple representation
- Assumes independence between positions
- ~25% of TRANSFAC motifs have been shown to violate this assumption
  - Two Examples: ADR1 and YAP6

Outline

- Background
  - Many classes of algorithms, differ in
    - Types of input data
    - Motif representation
  - Commonly used methods
    - Enumeration
    - Expectation-Maximization methods (MEME)
    - Gibbs sampling methods (AlignAce, BioProspector)
    - Greedy search method (CONSENSUS)

Finding regulatory motifs

- Say a transcription factor (TF) controls five different genes
- Each of the five genes will have binding sites for the TF in their promoter region

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
<th>Gene 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binding sites for TF</td>
<td>ADR1</td>
<td>YAP6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Finding regulatory motifs

- Given the upstream sequences of the genes that seem to be regulated by the same TFs,
- Find the TF-binding sites (motifs) in common

Identifying motifs: complications

- We do not know the motif sequence
- We do not know where it is located relative to the genes
- Motifs can differ slightly from one gene to another
- How to discern it from “random” motifs?

Common methods

- Problem statements:
  - Given a set of promoters of $n$ co-regulated genes, find a motif common to the promoters/
  - Both the PWM and the motif sequences are unknown.
- Enumeration (simplest method)
  - Look at the frequency of all k-mers
- EM algorithm (MEME)
  - Iteratively learn the most likely motif model
- Gibbs sampling methods
  - AlignAce, BioProspector

Generating k-mers

- Example (5-mers):
  atgacgggat act gat acgcgt attt gcct gg ccgt acacattagata aacg
Motif finding using EM algorithm

- MEME (Multiple EM for Motif Elucidation)
    http://meme.sdsc.edu/meme/intro.html
- Expectation-Maximisation
  - In each iteration, it learns the PWM model and identifies examples of the matrix (sites in the input sequences)
  - Identify binding locations for all PWMs
    - Optimize recognition preferences

Example: MEME

- Find a 6-mer motif in 4 sequences
  - S1: GGCTATTGCAATATGACGAGATGAGGCCCAGACC
  - S2: GGATGACAAATTATATAAAGGACGAGATGAC
  - S3: CTAGCTCGTAGCTCGTTGAGATGCGCTCCCCGCTC
  - S4: GATGACGGAGTATTAAAGACTCGATGAGTTATACGA

1. MEME uses an initial EM heuristic to estimate the best starting-point PWM matrix:

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.26</td>
<td>0.24</td>
<td>0.18</td>
<td>0.26</td>
</tr>
<tr>
<td>S2</td>
<td>0.24</td>
<td>0.26</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>S3</td>
<td>0.25</td>
<td>0.23</td>
<td>0.30</td>
<td>0.25</td>
</tr>
<tr>
<td>S4</td>
<td>0.25</td>
<td>0.27</td>
<td>0.24</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Motif finding using EM algorithm

- MEME works by iteratively refining PWMs and identifying sites for each PWM
  - 1. Estimate motif model (PWM)
    - Start with a k-mer seed (random or specified)
    - Build a PWM by incorporating some of background frequencies
  - 2. Identify examples of the model
    - For every k-mer in the input sequences, identify its probability given the PWM model.
  - 3. Re-estimate the motif model
    - Calculate a new PWM, based on the weighted frequencies of all k-mers in the input sequences
  - 4. Iteratively refine the PWMs and identify sites until convergence.

2. MEME scores the match of all 6-mers to current matrix

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
<th>A</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCTATTGC ATATGACGAGGTGAGGCCCAGACC</td>
<td>0.29</td>
<td>0.24</td>
<td>0.17</td>
<td>0.27</td>
<td>0.24</td>
<td>0.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATGAC AAATAGGCCGTGATAAGATGAC</td>
<td>0.22</td>
<td>0.26</td>
<td>0.27</td>
<td>0.22</td>
<td>0.28</td>
<td>0.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTAGCT CGTGC AGTGAGGTGCTCGTCCCAGCTC</td>
<td>0.24</td>
<td>0.23</td>
<td>0.33</td>
<td>0.23</td>
<td>0.24</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GATGAC CTGTTAAGACTCGATGAGTTACGA</td>
<td>0.24</td>
<td>0.27</td>
<td>0.23</td>
<td>0.28</td>
<td>0.24</td>
<td>0.24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. MEME scores the match of all 6-mers to current matrix

```
G 0.40 0.20 0.15 0.42 0.24 0.30
A 0.30 0.30 0.20 0.24 0.46 0.18
T 0.15 0.30 0.45 0.16 0.15 0.28
C 0.15 0.20 0.20 0.16 0.15 0.24
```

5. Re-estimate the PWM based on the weighted contribution of all 6-mers.

```
GATGACGGTATGACGAGATGAGGCCCAGACC
GGATGACAATTATATAAAGGACCGTGATAAGAGATTAC
CTAGCTCGTAGCTCGTTGAGATGCGCTCCCCGCTC
GATGACGGTATGACGAGATGAGTATACGA
```

6. MEME scores the match of all 6-mers to current matrix

```
G 0.85 0.05 0.10 0.80 0.20 0.35
A 0.05 0.60 0.10 0.05 0.60 0.10
T 0.05 0.30 0.70 0.05 0.20 0.10
C 0.05 0.05 0.10 0.10 0.10 0.35
```

Using EM algorithm

- MEME works by iteratively refining PWMs and identifying sites for each PWM
  1. Estimate motif model (PWM)
     - Start with a k-mer seed (random or specified)
     - Build a PWM by incorporating some of background frequencies
  2. Identify examples of the model
     - For every k-mer in the input sequences, identify its probability given the PWM model.
  3. Re-estimate the motif model
     - Calculate a new PWM, based on the weighted frequencies of all k-mers in the input sequences
  4. Iterate 2 & 3 until convergence.

Outline

- Background
- Many classes of algorithms, differ in
  - Types of input data
  - Motif representation
- Commonly used methods
  - Enumeration
  - Expectation-Maximization methods (MEME)
  - Greedy search method (CONSENSUS)
Outline

- Background
  - Many classes of algorithms, differ in
    - Types of input data
    - Motif representation
  - Commonly used methods
    - Enumeration
    - Expectation-Maximization methods (MEME)
    - Greedy search method (CONSENSUS)
    - Gibbs sampling methods (AlignAce, BioProspector)

CONSENSUS

- Hertz and Stormo, Bioinformatics 1999
- Popular algorithm for motif discovery, that uses a greedy approach
- Motif model: Position Weight Matrix (PWM)
- Motif score: information content

Information content

- PWM $W$:
  - $W_{\beta k}$ = frequency of base $\beta$ at position $k$
  - $q_{\beta}$ = frequency of base $\beta$ by chance
  - Information content of $W$:

$$\sum_k \sum_{\beta \in \{A,C,G,T\}} W_{\beta k} \log \frac{W_{\beta k}}{q_{\beta}}$$

- If $W_{\beta k}$ is always equal to $q_{\beta}$, i.e., if $W$ is similar to random sequence, information content of $W$ is 0.
- If $W$ is different from $q$, information content is high.

- Information content of $W$:

$$\sum_k \sum_{\beta \in \{A,C,G,T\}} W_{\beta k} \log \frac{W_{\beta k}}{q_{\beta}}$$
CONSENSUS: Basic idea

- Find a set of subsequences, one in each input sequence

Set of subsequences define a PWM.

Goal: This PWM should have high information content.

High information content means that the motif “stands out”.

CONSENSUS: Greedy heuristic

- Suppose we have built a partial set of subsequences \( \{s_1, s_2, \ldots, s_i\} \) so far.
- Have to choose a subsequence \( s_{i+1} \) from the input sequence \( S_{i+1} \)
- Consider each subsequence \( s \) of \( S_{i+1} \)
- Compute the score (information content) of the PWM made from \( \{s_1, s_2, \ldots, s_i, s\} \)
- Choose the \( s \) that gives the PWM with highest score, and assign \( s_{i+1} \leftarrow s \)

CONSENSUS: Basic idea

- Start with a subsequence in one input sequence
- Build the set of subsequences incrementally, adding one subsequence at a time
- Until the entire set is built