Sequence Motif Analysis

Lecture in M.Sc. Biomedizin,
Module: "Proteinbiochemie und Bioinformatik"

Jonas Ibn-Salem

Andrade group
Johannes Gutenberg University Mainz
Institute of Molecular Biology

March 7, 2016

Morgane Thomas-Chollier (ENS, Paris) kindly shared some of her slides
Overview

1. **Sequence Motif Models**
   - Consensus sequence
   - Matrix representation

2. **De novo Motif Discovery**
   - Motif discovery problem
   - Word counting
   - Matrix-based

3. **Known Motifs and Pattern Matching**
   - Collections of Known Motifs
   - Scanning sequences for motif hits
Overview

1. Sequence Motif Models
   - Consensus sequence
   - Matrix representation

2. De novo Motif Discovery
   - Motif discovery problem
   - Word counting
   - Matrix-based

3. Known Motifs and Pattern Matching
   - Collections of Known Motifs
   - Scanning sequences for motif hits
Motif analysis of transcription factor binding sites

Question

How do transcription factors (TF) know where to bind DNA?

TFs recognize binding sites with specific DNA sequences. However, bound sequences are often slightly different.
Motif analysis of transcription factor binding sites

Question

How do transcription factors (TF) know where to bind DNA?

- TFs recognize binding sites with specific DNA sequences.
Motif analysis of transcription factor binding sites

Question

How do transcription factors (TF) know where to bind DNA?

- TFs recognize binding sites with specific DNA sequences.

- However, bound sequences are often slightly different.
Modelling the sequence specificity of TFs

**Question**

How can we describe the sequence specificity of a given TF?
Modelling the sequence specificity of TFs

Question
How can we describe the sequence specificity of a given TF?

Multiple alignment:

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>C</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>G</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
</tbody>
</table>
Modelling the sequence specificity of TFs

Question

How can we describe the sequence specificity of a given TF?

Sequence motifs are short, recurring patterns in DNA that are presumed to have a biological function.
Motif Representations

Question
How is a motif represented / described?

- String-based
  - Strict consensus
  - Degenerate consensus
- Regular expressions
- Matrix based
  - Position frequency matrix (PFM)
  - Position probability matrix (PPM)
  - Position specific scoring matrix (PSSM)
- Sequence Logos
- Hidden Markov Models (HMM)
Motif Representations

Question
How is a motif represented / described?

- String-based
  - Strict consensus
  - Degenerate consensus
- Regular expressions
- Matrix based
  - Position frequency matrix (PFM)
  - Position probability matrix (PPM)
  - Position specific scoring matrix (PSSM)
- Sequence Logos
- Hidden Markov Models (HMM)
Consensus sequence

- A consensus sequence is a motif description as string (=sequence).
- A **strict consensus** sequence is derived from the collection of binding sites by taking the predominant letter at each position (column) in the alignment.

- A **degenerate consensus** sequence is defined by selecting a degeneracy nucleotide symbol for each position.

---

Source: Morgane Thomas-Chollier (ENS, Paris)

Jonas Ibn-Salem (JGU Mainz/IMB)
Problem

TF binding sites (TFBS) are degenerated. A given TF is able to bind DNA on TFBSs with different sequences.
Matrix model of sequence motifs

Problem

TF binding sites (TFBS) are *degenerated*. A given TF is able to bind DNA on TFBSs with different sequences.

- A **Position Frequency Matrix (PFM)** is created by counting the occurrences of each base (rows) in each position (columns) of the alignment.
- This is a more quantitative description of the known binding sites.

Source: Morgane Thomas-Chollier (ENS, Paris)
Jonas Ibn-Salem (JGU Mainz/IMB)

Sequence Motif Analysis
March 7, 2016 9 / 46
Conversion to Position Probability matrix (PPM)

Position Frequency Matrix (PFM):

<table>
<thead>
<tr>
<th>Residue</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>30</td>
<td>37</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
<td>0</td>
<td>35</td>
<td>37</td>
<td>41</td>
<td>9</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>19</td>
<td>32</td>
<td>38</td>
</tr>
<tr>
<td>Sum</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>44</td>
<td>44</td>
</tr>
</tbody>
</table>

\[ n_{i,j} \]

\[ \sum_{i=1}^{A} n_{i,j} \]

Position Probability Matrix (PPM):

<table>
<thead>
<tr>
<th>Residue</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.68</td>
<td>0.84</td>
<td>0.00</td>
<td>0.07</td>
<td>0.02</td>
<td>0.11</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>C</td>
<td>0.09</td>
<td>0.00</td>
<td>0.80</td>
<td>0.84</td>
<td>0.93</td>
<td>0.20</td>
<td>0.02</td>
<td>0.09</td>
</tr>
<tr>
<td>G</td>
<td>0.09</td>
<td>0.05</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.25</td>
<td>0.16</td>
<td>0.00</td>
</tr>
<tr>
<td>T</td>
<td>0.14</td>
<td>0.11</td>
<td>0.14</td>
<td>0.09</td>
<td>0.05</td>
<td>0.43</td>
<td>0.73</td>
<td>0.86</td>
</tr>
<tr>
<td>Sum</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ f_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^{A} n_{i,j}} \]

\( A \) alphabet size (=4)

\( n_{i,j} \) occurrences of residue \( i \) at position \( j \)

\( f_{i,j} \) relative frequency of residue \( i \) at position \( j \)

Reference: [Hertz and Stormo, 1999] Source: Morgane Thomas-Chollier (ENS, Paris)
Position specific scoring matrix (PSSM) or weight matrix

- To model the enrichment of an observed frequency over some background the PFM can be converted to a position specific scoring matrix (PSSM) or weight matrix $W$.

$$W_{k,j} = \log(M_{k,j}/b_k)$$

where $b_k$ is a background probability for each base (under a Bernoulli model).

- Usually a pseudo-count of $+1$ is added to the PFM before the transformation to avoid the logarithm of zero.
A **Sequence logo** is a graphical representation of a motif.

The height of each letter is proportional to the frequency of each residue at each position.

The total height of each column is proportional to the sequence conservation and indicate the amount of information contained in each position (information content measured in bits).

Allows easy identification of the most important positions in the motif.

\[
I_j = 2 - \left( -\sum_{i=1}^{A} f_{i,j} \log_2(f_{i,j}) \right)
\]

**Source:** Morgane Thomas-Chollier (ENS, Paris)

Jonas Ibn-Salem (JGU Mainz/IMB)
Overview

1. Sequence Motif Models
   Consensus sequence
   Matrix representation

2. De novo Motif Discovery
   Motif discovery problem
   Word counting
   Matrix-based

3. Known Motifs and Pattern Matching
   Collections of Known Motifs
   Scanning sequences for motif hits
De novo Motif Discovery

Motif discovery problem
If there is a common regulating factor, can we discover its motif by using a set of functionally related sequences?

The input sequences might be
- Upstream sequences from co-expressed genes
- Sequences of ChIP-seq peaks

Source: Morgane Thomas-Chollier (ENS, Paris)
Jonas Ibn-Salem (JGU Mainz/IMB)
Motif discovery using word counting

**Idea**

Since motifs corresponding to binding sites are generally repeated in the dataset, we can capture this signal of over-representation statistically.

- Binding sites are represented as "words" == "string" == "k-mer" (e.g. acgtga is a 6-mer).
- Signal is likely to be more frequent in the region of interest than in a random selection of regions.
- We can thus detect over-represented words.
Algorithm for motif discovery using word counting

Algorithm

1. Count occurrences of all k-mers in a set of related sequences.
2. Estimate the expected number of occurrences from a background model
   - Empirical based on observed k-mer frequencies
   - Theoretical background model (Markov Models)
3. Statistical evaluation of the deviation observed (P-value/E-value).
4. Select all words above a defined threshold.
Example of yeast under low nitrogen condition

- Expression of 7 yeast (Saccharomyces cerevisiae) genes under low nitrogen conditions (NIT)
- Count all possible 6-mers in upstream sequences.

Question:

- Does this mean that the word aaaaaa is enriched in upstream sequences of NIT genes?

<table>
<thead>
<tr>
<th>Sequence</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>aaaaaa</td>
<td>80</td>
</tr>
<tr>
<td>cttatc</td>
<td>26</td>
</tr>
<tr>
<td>tatata</td>
<td>22</td>
</tr>
<tr>
<td>ataaga</td>
<td>20</td>
</tr>
<tr>
<td>aagaaa</td>
<td>20</td>
</tr>
<tr>
<td>gaaaaaa</td>
<td>19</td>
</tr>
<tr>
<td>atatatat</td>
<td>19</td>
</tr>
<tr>
<td>agataaa</td>
<td>17</td>
</tr>
<tr>
<td>agaaaaa</td>
<td>17</td>
</tr>
<tr>
<td>aaagaaaa</td>
<td>16</td>
</tr>
<tr>
<td>aaaaacagataa</td>
<td>16</td>
</tr>
<tr>
<td>aaaaag</td>
<td>15</td>
</tr>
<tr>
<td>agaaga</td>
<td>14</td>
</tr>
<tr>
<td>tgataaa</td>
<td>14</td>
</tr>
<tr>
<td>atataata</td>
<td>14</td>
</tr>
</tbody>
</table>

Source: Morgane Thomas-Chollier (ENS, Paris)

Jonas Ibn-Salem (JGU Mainz/IMB)
Example of yeast under low nitrogen condition (part 2)

Problem

Upstream sequences have a compositional bias and are enriched for AT-rich sequences.

We need a background model to calculate the enrichment.

Source: Morgane Thomas-Chollier (ENS, Paris)
Jonas Ibn-Salem (JGU Mainz/IMB)
Sequence Motif Analysis
Empirical background based on observed k-mer frequencies

Idea

Count the occurrences of all k-mers in upstream sequences of all genes

<table>
<thead>
<tr>
<th>Sequence</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>aaaaaa</td>
<td>ttttttt</td>
</tr>
<tr>
<td>cttatc</td>
<td>gataag</td>
</tr>
<tr>
<td>tatata</td>
<td>tatata</td>
</tr>
<tr>
<td>ataaga</td>
<td>tctttat</td>
</tr>
<tr>
<td>aagaaa</td>
<td>tttctt</td>
</tr>
<tr>
<td>gaaaaa</td>
<td>tttttct</td>
</tr>
<tr>
<td>atatat</td>
<td>atatatat</td>
</tr>
<tr>
<td>agataa</td>
<td>ttatct</td>
</tr>
<tr>
<td>agaaaa</td>
<td>tttttct</td>
</tr>
<tr>
<td>aagaaa</td>
<td>tttcttt</td>
</tr>
<tr>
<td>aaaaac</td>
<td>tgttttt</td>
</tr>
<tr>
<td>aaaaag</td>
<td>ctttttt</td>
</tr>
<tr>
<td>agaaga</td>
<td>tcttct</td>
</tr>
<tr>
<td>tgataa</td>
<td>ttatca</td>
</tr>
<tr>
<td>atataa</td>
<td>ttatat</td>
</tr>
</tbody>
</table>

Source: Morgane Thomas-Chollier (ENS, Paris)
Jonas Ibn-Salem (JGU Mainz/IMB)
Estimation of background by Markov Models

Idea
Estimate the sequence composition using a statistical model.

Bernoulli model models only the frequency of each single base: \( p(A) \), \( p(C) \), \( p(G) \), \( p(T) \).

Simplest model: \( p(A) = p(C) = p(G) = p(T) = 0.25 \).

Example: \( p(ACGTGA) = 0.25^6 = 0.000244 \).

Frequencies in all upstream regions of yeast: \( p(A) = 0.3 \), \( p(C) = 0.2 \), \( p(G) = 0.2 \), \( p(T) = 0.3 \).

Example: \( p(ACGTGA) = (0.3)^2 \times (0.2)^2 \times (0.2)^2 \times (0.3)^2 \times (0.2) = 0.000216 \).

Markov model: The probability of each residue depends on the \( m \) preceding residues. The parameter \( m \) is the order of the Markov model.

Example Markov model order 1:
\[
p(ACGTGA) = p(A) \times p(C|A) \times p(G|C) \times p(T|G) \times p(G|T) \times p(A|G)
\]

Example Markov model order 2:
\[
p(ACGTGA) = p(AC) \times p(G|AC) \times p(T|CG) \times p(G|GT) \times p(A|TG)
\]
Estimation of background by Markov Models

Idea
Estimate the sequence composition using a statistical model.

- **Bernoulli model** models only the frequency of each single base: \( p(A), p(C), p(G), p(T) \).
  - Simplest model: \( p(A) = p(C) = p(G) = p(T) = 0.25 \)
    - Example: \( p(ACGTGA) = 0.25^6 = 0.000244 \)
  - Frequencies in all upstream regions of yeast:
    - \( p(A) = 0.3, p(C) = 0.2, p(G) = 0.2, p(T) = 0.3 \)
    - Example: \( p(ACGTGA) = p(A)^2 \times p(C) \times p(G)^2 \times p(T) = 0.000216 \)
Estimation of background by Markov Models

Idea
Estimate the sequence composition using a statistical model.

- **Bernoulli model** models only the frequency of each single base: \( p(A), p(C), p(G), p(T) \).
  - Simplest model: \( p(A) = p(C) = p(G) = p(T) = 0.25 \)
    Example: \( p(ACGTGA) = 0.25^6 = 0.000244 \)
  - Frequencies in all upstream regions of yeast:
    \( p(A) = 0.3, p(C) = 0.2, p(G) = 0.2, p(T) = 0.3 \)
    Example: \( p(ACGTGA) = p(A)^2 \ast p(C) \ast p(G)^2 \ast p(T) = 0.000216 \)

- **Markov model**: The probability of each residue depends on the \( m \) preceding residues. The parameter \( m \) is the order of the Markov model.
  - Example Markov model order 1:
    \( p(ACGTGA) = p(A) \ast p(C|A) \ast p(G|C) \ast p(T|G) \ast p(G|T) \ast p(A|G) \)
  - Example Markov model order 2:
    \( p(ACGTGA) = p(AC) \ast p(G|AC) \ast p(T|CG) \ast p(G|GT) \ast p(A|TG) \)
Statistical evaluation of over representation

- Assume we found 18 occurrences of the word ACGTGA and estimated the expected number to be 2.95 using the background model.

Question:
- How big is the surprise to observe 18 occurrences when we expect 2.95?
Statistical evaluation of over representation

- At each position in the sequences, there is a probability $p$ that the word starting at this position is ACGTGA.
- We consider $n = 9000$ positions.
- From the background model we know the expected probability $p = \frac{2.95}{9000} = 0.00032$.
- What is the probability that $k = 18$ of these $n$ positions correspond to ACGTGA?

Binomial distribution to measure the significance:

$$P(X \geq k) = 1 - \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

In R:
```r
> binom.test(x=18, n=9000, p=0.00033)
```

$p$-value $= 3.03 \times 10^{-9}$

Multiple testing correction by multiplying the $p$-value by the number of tested words.
Statistical evaluation of over representation

- At each position in the sequences, there is a probability $p$ that the word starting at this position is ACGTGA.
- We consider $n = 9000$ positions.
- From the background model we know the expected probability $p = \frac{2.95}{9000} = 0.00032$.
- What is the probability that $k = 18$ of these $n$ positions correspond to ACGTGA?
- **Binomial distribution** to measure the significance:

\[
P(X \geq k) = 1 - \sum_{i=0}^{k} \binom{n}{i} p^i (1 - p)^{n-i}
\]

**In R:**

```r
> binom.test(x=18, n=9000, p=0.00033) ⇒ p-value = 3.029 \times 10^{-9}
```

- Multiple testing correction by multiplying the p-value by the number of tested words.
Assembling overlapping words

<table>
<thead>
<tr>
<th>seq</th>
<th>identifier</th>
<th>exp_freq</th>
<th>occ</th>
<th>exp_occ</th>
<th>occ_P</th>
<th>occ_E</th>
</tr>
</thead>
<tbody>
<tr>
<td>cacgtg</td>
<td>cacgtg</td>
<td>0.0001569968432</td>
<td>17</td>
<td>1.47</td>
<td>5e-13</td>
<td>1.0e-09</td>
</tr>
<tr>
<td>acgtga</td>
<td>acgtga</td>
<td>0.0003355962588</td>
<td>18</td>
<td>3.15</td>
<td>7.3e-09</td>
<td>1.5e-05</td>
</tr>
<tr>
<td>ccacag</td>
<td>ccacag</td>
<td>0.0002365577659</td>
<td>14</td>
<td>2.22</td>
<td>1e-07</td>
<td>2.1e-04</td>
</tr>
<tr>
<td>gccaca</td>
<td>gccaca</td>
<td>0.0002897084237</td>
<td>15</td>
<td>2.72</td>
<td>2e-07</td>
<td>4.1e-04</td>
</tr>
<tr>
<td>actgtg</td>
<td>actgtg</td>
<td>0.0003762020409</td>
<td>16</td>
<td>3.53</td>
<td>1e-06</td>
<td>2.1e-03</td>
</tr>
<tr>
<td>cgtgca</td>
<td>cgtgca</td>
<td>0.0002325962261</td>
<td>11</td>
<td>2.18</td>
<td>1.8e-05</td>
<td>3.8e-02</td>
</tr>
<tr>
<td>aactgt</td>
<td>aactgt</td>
<td>0.0006168655788</td>
<td>17</td>
<td>5.78</td>
<td>0.00011</td>
<td>2.4e-01</td>
</tr>
<tr>
<td>agtcat</td>
<td>agtcat</td>
<td>0.0005306961969</td>
<td>15</td>
<td>4.73</td>
<td>0.00012</td>
<td>2.6e-01</td>
</tr>
<tr>
<td>ttagcta</td>
<td>ttagct</td>
<td>0.0004613751449</td>
<td>14</td>
<td>4.33</td>
<td>0.00017</td>
<td>3.5e-01</td>
</tr>
<tr>
<td>agccac</td>
<td>agccac</td>
<td>0.0002599968758</td>
<td>10</td>
<td>2.44</td>
<td>0.00023</td>
<td>4.7e-01</td>
</tr>
<tr>
<td>cgtgca</td>
<td>cgtgca</td>
<td>0.0001695417819</td>
<td>8</td>
<td>1.59</td>
<td>0.00025</td>
<td>5.2e-01</td>
</tr>
<tr>
<td>cgccga</td>
<td>cgccga</td>
<td>0.0001715224888</td>
<td>8</td>
<td>1.61</td>
<td>0.00027</td>
<td>5.6e-01</td>
</tr>
<tr>
<td>acgtgc</td>
<td>acgtgc</td>
<td>0.0002276443015</td>
<td>9</td>
<td>2.13</td>
<td>0.00038</td>
<td>7.9e-01</td>
</tr>
<tr>
<td>gactca</td>
<td>gactca</td>
<td>0.0002319359695</td>
<td>9</td>
<td>2.18</td>
<td>0.00043</td>
<td>9.0e-01</td>
</tr>
</tbody>
</table>

Word assembly to form longer motifs and matrices

```plaintext
;assembly # 1 seed: cacgtg
;   alignt rev_cpl
   gtcacg.... ....cgtgac
   .tcacgt.... ...acgtga.
   ..cagctg.. ..cagctg...
   ....cagtga. .tcagctg...
       ....cgtgac gtcacg....
   gtcacgac gtcacgac

;assembly # 2 seed: ccacag
;   alignt rev_cpl
   agccac.... ....tggtgc
   .gccacax.... .tggtgcx.
   ..ccacag.. .ctggtgcx..
   ....cagctg. .actggtg..
       ....cagtt aactgt....
   agccacagtt aactgtggct

;assembly # 3 seed: cgtgca
;   alignt rev_cpl
   gtcacg.... ....cgtgac
   .tcacgt.... ...acgtga.
   ..cagtgc.. ..cagctg...
   ....cagtc. .gctagct...
       ....cgtgca tgcacg....
   gtcacgac gtcacgac
```

Source: Morgane Thomas-Chollier (ENS, Paris)
Jonas Ibn-Salem (JGU Mainz/IMB)  Sequence Motif Analysis  March 7, 2016
Motif discovery using matrices

Idea:
Find a matrix model that optimally describes a motif that is enriched in the sequences.

- **Comprehensive approach**
  - Test all possible motif matrices that can build from the sequences.
  - Compute a score (e.g. information content, P-value) associated to each motif matrix.
  - Report the highest scoring matrix.

Problem: The number of possible matrices is too large to be computed in a reasonable time.

- **Heuristics**
  - Expectation-maximization (MEME)
  - Gibbs sampling (MotifSampler, info-gibbs,...)
MEME (Multiple EM for Motif Elicitation)

**Expectation-maximization (EM)**

- Instantiate a seed motif
- Iterate $N$ times to enhance matrix:
  - Maximization: select the $x$ highest scoring sites
  - Expectation: build a new matrix from the collected sites

---

**Multiple EM**

- Iterate over each $k$-mer fund in the input set
- Ranking / selection of the best matrices according to their score

Source: Morgane Thomas-Chollier (ENS, Paris)
Jonas Ibn-Salem (JGU Mainz/IMB)
Gibbs Sampling for stochastic optimisation

- Deterministic selection of the highest scoring site might result in local optimum.

![Diagram showing multiple peaks with one highlighted]

- Therefore, Gibbs sampling uses a random sampling step to better explore the "search space" of possible solutions.

Idea:

- The idea in Gibbs sampling is to choose a site stochastically according to the score distribution to find the global optimum.
Gibbs Sampling

- Step 0: Initialisation of the matrix and background model

random positions

N sequences

initial matrix
Gibbs Sampling

Step 1: Updating

the site of the selected sequence is **removed** from the matrix which is updated
Gibbs Sampling

- Step 2: Random sampling

\[ W_i = \log \frac{p_m(i)}{p_b(i)} \]

at each position i

the weight is computed

proba that position i corresponds to a site

proba that position i corresponds to a "non-site"
Gibbs Sampling

- Step 2: random sampling
  - selection of a new site with probability proportional to $W_i$
  - updating the matrix and the background model
Gibbs Sampling

- Step 1: updating (2nd iteration)

The site of the selected sequence is removed from the matrix which is updated.

*Further iterations ...*
Gibbs Sampling: Overview

0 Initialization
- Select a random set of sites in the sequence set
- Create a matrix with these sites

1 Sampling (Stochastic Expectation)
- Isolate one sequence from the set, and score each position (site) of this sequence.
- Select one “random” site, with a probability proportional to the score.

2 Predictive update (Maximization)
- Replace the old site with a new site, and update the matrix.

3 Iterate Step 1 and 2 for a fixed number of cycles.

Source: Morgane Thomas-Chollier (ENS, Paris)
De novo motif discovery with RSAT peak-motifs

- The RSAT tool suite implements a complete pipeline for ChIP-seq based motif analysis.
- For motif discovery it uses four complementary algorithms:
  - Global over-representation
  - oligo-analysis
  - dyad-analysis (spaced motifs)
  - Positional bias
  - position-analysis
  - local-words

source: [Thomas-Chollier et al., 2012]
Overview

1 Sequence Motif Models
   Consensus sequence
   Matrix representation

2 De novo Motif Discovery
   Motif discovery problem
   Word counting
   Matrix-based

3 Known Motifs and Pattern Matching
   Collections of Known Motifs
   Scanning sequences for motif hits
Collections of Known Motifs

Idea

After de novo motif discovery, we might want to know to which TF it belongs. We can compare a discovered motif to databases of known TF motifs.

Many database of known TF motifs exist:

- Public (free accessible) and commercial databases
- General, organism, or experimental specific databases
- Often incomplete, redundant, and of heterogeneous quality

Source: Morgane Thomas-Chollier (ENS, Paris)

Jonas Ibn-Salem (JGU Mainz/IMB)
- Commercial (BioBase™)
- Eukaryotic TF matrix models from binding sites with evidence and publications
- Pattern matching tools
JASPAR

- Public (http://jaspar.genereg.net/)
- Contain matrix models and access to sites from which they were constructed
- Tools for pattern matching and matrix randomization

Reference: [Mathelier et al., 2015]

Jonas Ibn-Salem (JGU Mainz/IMB)
Pattern Matching Problem:

Given a known TF recognition motif and a sequence, we want to know all positions where the motif matches the sequence.
Motif discovery vs. Pattern Matching

Motif Discovery Problem:
Given a set of related sequences, identify common sequence motifs in from these sequences

Input: Set of sequences
Output: Motif model

Example
- What is the sequence motif of the TF in a ChIP-seq experiment?
- Are there other co-factors involved?
- Which TF bind promoters of some co-expressed genes?

Pattern Matching Problem:
Given a known TF recognition motif and a sequence, we want to know all positions where the motif matches the sequence.

Input: Motif, Sequence
Output: Positions of motif hits

Example
- Where does a TF bind exactly in a ChIP-seq peak region?
- Does a TF bind direct or indirect at a ChIP-seq peak region?
Searching with Known Motif: String based

Problem:

Given a motif as consensus sequence, find all matches in an input sequence.

- Compare each position along the sequence with the motif (scanning) and report matches.
Searching with Known Motif: Matrix based

Problem:
Given a matrix motif model, find all significant hits (motif instances) in a sequence.

- The similarity of a given sequence segment $S$ with the motif can be computed as weight score $W_S$

$$W_S = \log \frac{P(S|M)}{P(S|B)}$$

- $P(S|M)$ is the probability that the sequence segment $S$ occurs according to the motif model $M$. This can be simply computed from the matrix.
- $P(S|B)$ is the probability of $S$ under a background model $B$. This can be a first order Markov model.
- The higher the score $W_S$ the more similar is the sequence $S$ to the matrix model $M$. 

Probability of a sequence under the matrix model $P(S|M)$

Given a matrix and sequence we can calculate

$$P(S|M) = \prod_{j=1}^{w} M_{r_j,j}$$

where $M$ is the frequency matrix of width $w$, $S = \{r_1, r_2, ..., r_2\}$ is the sequence segment and $r_j$ the residue found at position $j$ in the sequence $S$.

Note, that usually a corrected matrix is used by adding pseudo-counts (e.g. $+1$) to avoid zeros in the calculation.
The probability under the background model can be calculated as

\[ P(S|B) = \prod_{j=1}^{w} p_{r_j} \]

where \( p_{r_j} \) is the probability of observing the residue \( r_j \) according to the background model.

A background model \((B)\) should estimate the probability of a sequence motif in non-biding sites.

Various possible background models. E.g.
- Bernoulli model with residue-specific probabilities \((p_r)\)
- Markov models to take dinucleotide frequencies into account (e.g. GC-bias)
Assigning a score to each segment and threshold usage

- The sequence is scanned with the matrix and a score is assigned to each position.
- Resulting hits depend highly on the threshold on the weight score:
  - Stringent threshold $\Rightarrow$ high confidence in predicted sites, but many sites missed.
  - Loose threshold $\Rightarrow$ real sites hidden in many false positives.
- Some programs compute theoretical p-values from the weight score.
- A threshold on a p-value can be easier interpreted: For $p = 10^{-4}$ we expect 1 false prediction every 10,000 base pairs (on one strand).
Summary

- Transcription factors (TFs) recognize specific DNA sequence motifs.
- Sequence motifs can be represented as consensus sequence (string) or more quantitatively as matrix model to capture the specificity of each residue at each position.
- A sequence logo plot is a graphical representation of a motif matrix.
- An enriched sequence motif can be detected in a set of functionally related sequences using de novo motif discovery approaches.
- Known TF recognition motifs can be retrieved as matrix model from online databases.
- Pattern matching aims at finding putative TF binding sites (for which the binding motif is known) in DNA sequences.


