Lecture 8: Patterns, Profiles, and Motifs

- Finding patterns in protein and DNA sequences
- Calculating profiles of DNA sequences
Definitions and Resources

§ **Motif**: A region of a protein or DNA sequence that may be functionally or structurally significant and/or conserved in other sequences
  • Motifs usually contain biologically important sequences

§ **Pattern**: Describes a motif using a qualitative consensus sequence (e.g., IUPAC or regular expression)

§ **Profile**: Describes a motif using quantitative information captured in a position specific scoring matrix (weight matrix)

§ PROSITE is a protein sequence pattern and profile database
  • http://www.expasy.ch/prosite
  • Contains >1100 entries describing >1600 patterns and profiles

§ DNA pattern and profile databases are more fragmented
  • JASPAR (http://jaspar.genereg.net/) and S. cerevisiae Promoter Database (SCPD) (http://rulai.cshl.edu/SCPD/)
Importance of Sequence Patterns in Proteins

• Conserved patterns in protein sequences usually have important biological functions

• Conserved sequence patterns may be indicative of e.g., a protein structural domain, enzyme active site, or a binding site for another protein or metal ion
Steps in the Development of a New PROSITE Pattern

(1) Construct a multiple sequence alignment of a protein family

(2) Use the alignment to identify conserved or biologically significant residues (e.g., residues in catalytic/active site, binding domain, structural features)

(3) Start by creating a core sequence pattern (approximately 4-5 contiguous amino acids in length)

(4) Expand the pattern to improve its sensitivity and specificity for detecting known protein family members

  • Sensitivity: Test the trial pattern against known positive sequences
  • Specificity: Test the trial pattern against known negative sequences

Finding Patterns in Multiple Sequence Alignments

Pattern of Cu/Zn Superoxide Dismutase

Example of a PROSITE pattern (PS00087; SOD_CU_ZN_1): [GA]-[IMFAT]-H-[LIVF]-H-{S}-x-[GP]-[SDG]-x-[STAGDE]

The two histidines (H) are copper ligands

Pattern Nomenclature for PROSITE database:
• Each position in pattern is separated with a hyphen
• x can match any residue
• [ ] are used indicate ambiguous positions in the pattern
e.g., [SDG] means the pattern can match S, D, or G at this position
• { } are used to indicate residues that are not allowed at this position
e.g., {S} means NOT S (not Serine)
• ( ) surround repeated residues, e.g., A(3) means AAA
• < or > indicate the pattern is restricted to the N- or C-terminus of the sequence

Many DNA patterns are binding sites for Transcription Factors

Gal4 binding sequence:
C-G-G-N(11)-C-C-G

TATA Box:
T-A-T-A-A-[AT](3)

Gene

Gal4

TATA box

GAL3  CGGTCCACTGTGTGCGG
GAL7  CGGAGCAGCTTGGAGGC
GAL80 CGGCCGACTCTCGCGCCG
GCY1  CGGGGGCAGACTATTCCG
GAL1  CGGATTAGAAGGCGCGG
          CGGGCGACAGCCCTCGG
          CGGAAAGACTCTCCTCCG
GAL10 CGGAGGAGGCTTGCTCGCGG
          CGGGCGCTTTCTAATCGG
GAL2  CGGAAGCTTCTCCTCCG
          CGGCGTGTCCTCGCGG
          CGGAGATATCTGCGCGG
          CGGGCGGATCAGTCCCG
          CGGATCAGTCCCGAAGG
PCL10 CGGAGTATATTGCAACCG
MTH1  CGGGGAATGGAGTCCG
# IUPAC DNA Codes for DNA Patterns

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>B</td>
<td>C or G or T</td>
<td>Not-A (B follows A in alphabet)</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>D</td>
<td>A or G or T</td>
<td>Not-C (D follows C in alphabet)</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>H</td>
<td>A or C or T</td>
<td>Not-G (H follows G in alphabet)</td>
</tr>
<tr>
<td>K</td>
<td>G or T</td>
<td>Keto</td>
</tr>
<tr>
<td>M</td>
<td>A or C</td>
<td>aMino</td>
</tr>
<tr>
<td>N</td>
<td>A or C or G or T</td>
<td>aNy</td>
</tr>
<tr>
<td>R</td>
<td>A or G</td>
<td>puRine</td>
</tr>
<tr>
<td>S</td>
<td>C or G</td>
<td>Strong interaction (3 H-bonds)</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>V</td>
<td>A or C or G</td>
<td>Not-T (or Not-U)</td>
</tr>
<tr>
<td>W</td>
<td>A or T</td>
<td>Weak interaction (2 H-bonds)</td>
</tr>
<tr>
<td>Y</td>
<td>T or C</td>
<td>pYrimidine</td>
</tr>
</tbody>
</table>
Pattern of TF binding site consensus sequences

- Alignment of transcription factor binding sites

  Motif
  
  CCAAAATTAGGAAA
  CCTATTAAGAAAA
  CCAAATTAGGAAAA
  CCAAATTCGGATA
  CCCATTTCGAAAA
  CCTATTTAGTATA
  CCAAATTAGGAAAA
  TCTATTTTGGAAA
  CCAATTTTCAAAA
  CCAAATTGGCAAA

Consensus: YCHAWTWNNSAWA or CCHAWTTNGNAWA
Patterns: Summary

§ Advantages:
- Relatively straightforward to identify => exact pattern matching is fast
- Patterns are relatively intuitive to read and understand
- Databases with large numbers of protein and DNA sequence patterns are available (e.g., PROSITE, etc)

§ Disadvantages:
- Patterns are a qualitative description of motif => lose information about relative frequency of each residue or nucleotide at an ambiguous position, e.g. [GAC] versus 0.6 G, 0.28 A, and 0.12 C
- Can be difficult to write complex motifs using regular expression notation
- Cannot represent subtle sequence motifs
DNA Sequence Profiles

§ A profile is a position-specific scoring matrix that gives a quantitative description of a sequence motif

§ For DNA sequences, the profile scoring matrix has $N$ columns and $4+$ rows, $N$ being the length of the profile (# of sequence positions)

§ The first 4 rows of each column specify the score (log odds ratio) for finding, at that position in the target sequence, each of the 4 nucleotides (A, C, G, T)

§ The rows after the first 4 rows contain penalties for insertions/deletions at that position in the target sequence

§ $M_{kj} =$ score for the $j^{th}$ nucleotide (or gap) at the $k^{th}$ position in the sequence
DNA Profile: Position Specific Scoring Matrix

Sequence profile position $k$

\[ M_{kj} = \log \left( \frac{p_{kj}}{p_j} \right) \]

- $p_{kj} =$ probability of nucleotide $j$ at position $k$ in the profile
- $p_j =$ “background” probability of nucleotide $j$ in genome sequence
Simple Method for Calculating DNA Sequence Profiles

Adapted from Hertz and Stormo, *Bioinformatics* 15:563-577

§ Recall: \( M_{kj} = \log_e \left( \frac{p_{kj}}{p_j} \right) \)

§ As the number of aligned sequences grow (for large \( Z \)): \( p_{kj} = \frac{C_{kj}}{Z} \)
  - \( C_{kj} \) = Number of \( j^{th} \) type nucleotide at position \( k \)
  - \( Z \) = Total number of aligned sequences

§ For small numbers of aligned sequences, better to use the following method of calculating \( p_{kj} \):

\[
p_{kj} = \frac{C_{kj} + p_j}{Z + 1}
\]

- Where \( p_j \) = background probability of that nucleotide type in the genome (based on GC content of genome)
Example of calculating a DNA sequence profile (PSSM)

Alignment of Transcription factor consensus binding sequence:

```
CCAAATTAGGAAA
CCTATTAAGAAAA
CCAAATTAGGAAA
CCAAATTCCGATA
CCCATTTCGAAAA
CCTATTAGTATA
CCAAATTAGGAAA
CCAAATTGGCAAA
TCTATTTTGGAAA
CCAATTTTCAAAA
```

Alignment Matrix:

<table>
<thead>
<tr>
<th>Position k</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>C:</td>
<td>9</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T:</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Computing the DNA Sequence Profile (PSSM)

**Alignment Matrix:**

<table>
<thead>
<tr>
<th>Position k =</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>C:</td>
<td>9</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T:</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Recall: \( M_{kj} = \log_e \left( \frac{p_{kj}}{p_j} \right) = \log_e \left( \frac{(C_{kj} + p_j) / (Z + 1)}{p_j} \right) \)

Profile matrix values for k = 1 (assume \( p_j = 0.25 \) for all nucleotides):

\[
M_{1A} = \log_e \left( \frac{(C_{1A} + p_A) / (Z + 1)}{p_A} \right) = \log_e \left( \frac{(0 + 0.25) / (10 + 1)}{0.25} \right) = -2.4
\]

\[
M_{1C} = \log_e \left( \frac{(C_{1C} + p_C) / (Z + 1)}{p_C} \right) = \log_e \left( \frac{(9 + 0.25) / (10 + 1)}{0.25} \right) = 1.2
\]

\[
M_{1G} = \log_e \left( \frac{(C_{1G} + p_G) / (Z + 1)}{p_G} \right) = \log_e \left( \frac{(0 + 0.25) / (10 + 1)}{0.25} \right) = -2.4
\]

\[
M_{1T} = \log_e \left( \frac{(C_{1T} + p_T) / (Z + 1)}{p_T} \right) = \log_e \left( \frac{(1 + 0.25) / (10 + 1)}{0.25} \right) = -0.8
\]
Computing the DNA Sequence Profile (PSSM)

Alignment Matrix:

<table>
<thead>
<tr>
<th>Position k =</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>10</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>C:</td>
<td>9</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G:</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T:</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

DNA Profile Matrix (PSSM):

\[
M_{kj} = \log_e \left( \frac{p_{kj}}{p_j} \right) = \log_e \left( \frac{(C_{kj} + p_j) / (Z + 1)}{p_j} \right)
\]
Scoring a Test Sequence using the DNA profile (PSSM)

**Test Sequence (potential binding site):**

```
CCTATTTAGGATA
```

**DNA sequence profile (PSSM) for Transcription Factor binding site:**

<table>
<thead>
<tr>
<th>Position k</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>A:</td>
<td>-2.4</td>
<td>-2.4</td>
<td>0.8</td>
<td>1.3</td>
<td>0.6</td>
<td>-2.4</td>
<td>-0.8</td>
<td>0.6</td>
<td>-2.4</td>
<td>0.2</td>
<td>1.3</td>
<td>1.1</td>
<td>1.3</td>
</tr>
<tr>
<td>C:</td>
<td>1.2</td>
<td>1.3</td>
<td>-0.8</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-0.2</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
</tr>
<tr>
<td>G:</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-0.8</td>
<td>1.2</td>
<td>0.6</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
</tr>
<tr>
<td>T:</td>
<td>-0.8</td>
<td>-2.4</td>
<td>0.2</td>
<td>-2.4</td>
<td>0.6</td>
<td>1.3</td>
<td>1.2</td>
<td>-0.2</td>
<td>-2.4</td>
<td>-0.8</td>
<td>-2.4</td>
<td>-0.2</td>
<td>-2.4</td>
</tr>
</tbody>
</table>

**Test seq:**

```
C    C    T    A    T    T    T    T    A    G    G    A    T    A
```

**Total Score for test sequence:**

Score = 1.2 + 1.3 + 0.2 + 1.3 + 0.6 + 1.3 + 1.2 + 0.6 + 1.2 + 0.6 + 1.3 + -0.2 + 1.3

**Score = 11.9**

• Does the Test Sequence match the DNA sequence profile?
Simple Test for a Match to the DNA sequence profile

Score of Test Sequence (**CCTATTTAGGATA**) : **11.9**

Maximum possible score (**CCAATTTAGGAAA**):

<table>
<thead>
<tr>
<th>Position k =</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>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>-2.4</td>
<td>-2.4</td>
<td><strong>0.8</strong></td>
<td><strong>1.3</strong></td>
<td>0.6</td>
<td>-2.4</td>
<td>-0.8</td>
<td><strong>0.6</strong></td>
<td>-2.4</td>
<td>0.2</td>
<td><strong>1.3</strong></td>
<td><strong>1.1</strong></td>
<td><strong>1.3</strong></td>
</tr>
<tr>
<td><strong>C</strong></td>
<td><strong>1.2</strong></td>
<td><strong>1.3</strong></td>
<td>-0.8</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-0.2</td>
<td>-0.8</td>
<td>-0.8</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
<td>-0.8</td>
<td><strong>1.2</strong></td>
<td><strong>0.6</strong></td>
<td>-2.4</td>
<td>-2.4</td>
<td>-2.4</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>-0.8</td>
<td>-2.4</td>
<td>0.2</td>
<td>-2.4</td>
<td><strong>0.6</strong></td>
<td><strong>1.3</strong></td>
<td><strong>1.2</strong></td>
<td>-0.2</td>
<td>-2.4</td>
<td>-0.8</td>
<td>-2.4</td>
<td>-0.2</td>
<td>-2.4</td>
</tr>
</tbody>
</table>

Max Score: **C C A A T T T T A G G A A A A**

Total Score for Best Matching Sequence:

Max Score = 1.2 + 1.3 + 0.8 + 1.3 + 0.6 + 1.3 + 1.2 + 0.6 + 1.2 + 0.6 + 1.3 + 1.1 + 1.3

Max Score = **13.8**
Simple Test for a Match to the DNA sequence profile

Score of Test Sequence (CCTATTTAGGATA): 11.9

Maximum possible score (CCAATTTAGGAAA): 13.8

• Example threshold: if the score of the test sequence is >60% of the Maximum Score, we will designate it a match

Score Threshold for Match = 60% x Max Score = 0.6 x 13.8 = 8.28

For Match:
Score of test sequence > Score threshold
11.9 > 8.28

Hence, test sequence (CCTATTTAGGATA) matches the DNA sequence profile

Test sequence is a potential binding site of Transcription Factor