Position-specific scoring matrices (PSSM)

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Université d’Aix-Marseille, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univmed.fr/

FORMER ADDRESS (1999-2011)
Université Libre de Bruxelles, Belgique
Bioinformatique des Génomes et des Réseaux (BiGRe lab)
http://www.bigre.ulb.ac.be/
**Consensus representation**

- The TRANSFAC database contains 8 binding sites for the yeast transcription factor Pho4p
  - 5/8 contain the core of high-affinity binding sites (CACGTG)
  - 3/8 contain the core of medium-affinity binding sites (CACGTT)
- The IUPAC ambiguous nucleotide code allows to represent variable residues.
- 15 letters to represent any possible combination between the 4 nucleotides ($2^4 - 1 = 15$).
- This representation however gives a poor idea of the relative importance of residues.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>R06098</td>
<td>TCA\text{ACGTGGGA}</td>
</tr>
<tr>
<td>R06099</td>
<td>GG\text{ACGTGCAG}</td>
</tr>
<tr>
<td>R06100</td>
<td>TG\text{ACGTGGGT}</td>
</tr>
<tr>
<td>R06102</td>
<td>CA\text{GCACGTGGGG}</td>
</tr>
<tr>
<td>R06103</td>
<td>TT\text{CCACGTGCAG}</td>
</tr>
<tr>
<td>R06104</td>
<td>AC\text{GCACGTTGGT}</td>
</tr>
<tr>
<td>R06097</td>
<td>CA\text{GCACGTTTTC}</td>
</tr>
<tr>
<td>R06101</td>
<td>TA\text{CCACGTTTTC}</td>
</tr>
</tbody>
</table>

### IUPAC ambiguous nucleotide code

<table>
<thead>
<tr>
<th>Letter</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>R</td>
<td>puRine</td>
</tr>
<tr>
<td>Y</td>
<td>pYrimidime</td>
</tr>
<tr>
<td>W</td>
<td>Weak hydrogen bonding</td>
</tr>
<tr>
<td>S</td>
<td>Strong hydrogen bonding</td>
</tr>
<tr>
<td>M</td>
<td>aMino group at common position</td>
</tr>
<tr>
<td>K</td>
<td>Keto group at common position</td>
</tr>
<tr>
<td>H</td>
<td>not G</td>
</tr>
<tr>
<td>B</td>
<td>not A</td>
</tr>
<tr>
<td>V</td>
<td>not T</td>
</tr>
<tr>
<td>D</td>
<td>not C</td>
</tr>
<tr>
<td>N</td>
<td>aNy</td>
</tr>
</tbody>
</table>

TRANSFAC public version: [http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/search.cgi](http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/search.cgi)
From alignments to weights
### Residue count matrix

<table>
<thead>
<tr>
<th>Residue \ position</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>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>T</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
</tbody>
</table>

**Tom Schneider’s sequence logo**

(generated with Web Logo [http://weblogo.berkeley.edu/logo.cgi](http://weblogo.berkeley.edu/logo.cgi))

**Frequency matrix**

<table>
<thead>
<tr>
<th>Pos</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>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.13</td>
<td>0.38</td>
<td>0.25</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.13</td>
<td>0.25</td>
</tr>
<tr>
<td>C</td>
<td>0.25</td>
<td>0.25</td>
<td>0.38</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.25</td>
<td>0.00</td>
<td>0.25</td>
</tr>
<tr>
<td>G</td>
<td>0.13</td>
<td>0.25</td>
<td>0.38</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.63</td>
<td>0.50</td>
<td>0.63</td>
<td>0.25</td>
</tr>
<tr>
<td>T</td>
<td>0.50</td>
<td>0.13</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.38</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>Sum</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

\[
\hat{f}_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^{A} n_{i,j}}
\]

- \( A \) alphabet size (=4)
- \( n_{i,j} \) occurrences of residue i at position j
- \( p_i \) prior residue probability for residue i
- \( f_{i,j} \) relative frequency of residue i at position j

Corrected frequency matrix

<table>
<thead>
<tr>
<th>Pos</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>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.15</td>
<td>0.37</td>
<td>0.26</td>
<td>0.04</td>
<td>0.93</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.15</td>
<td>0.26</td>
</tr>
<tr>
<td>C</td>
<td>0.24</td>
<td>0.24</td>
<td>0.35</td>
<td>0.91</td>
<td>0.02</td>
<td>0.91</td>
<td>0.02</td>
<td>0.02</td>
<td>0.24</td>
<td>0.02</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.13</td>
<td>0.24</td>
<td>0.35</td>
<td>0.02</td>
<td>0.02</td>
<td>0.91</td>
<td>0.02</td>
<td>0.58</td>
<td>0.46</td>
<td>0.58</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.48</td>
<td>0.15</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.93</td>
<td>0.37</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

1st option: identically distributed pseudo-weight

\[
f'_{i,j} = \frac{n_{i,j} + k}{A}
\]

2nd option: pseudo-weights distributed according to residue priors

\[
f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}
\]

where:

- \( A \) = alphabet size (=4)
- \( n_{i,j} \) = occurrences of residue \( i \) at position \( j \)
- \( p_i \) = prior residue probability for residue \( i \)
- \( f_{i,j} \) = relative frequency of residue \( i \) at position \( j \)
- \( k \) = pseudo weight (arbitrary, 1 in this case)
- \( f'_{i,j} \) = corrected frequency of residue \( i \) at position \( j \)

# Weight matrix (Bernoulli model)

<table>
<thead>
<tr>
<th>Prior</th>
<th>Pos</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>
</tr>
</thead>
<tbody>
<tr>
<td>0.325</td>
<td>A</td>
<td>-0.79</td>
<td>0.13</td>
<td>-0.23</td>
<td>-2.2</td>
<td>1.05</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-0.79</td>
<td>-0.23</td>
</tr>
<tr>
<td>0.175</td>
<td>C</td>
<td>0.32</td>
<td>0.32</td>
<td>0.70</td>
<td>1.65</td>
<td>-2.2</td>
<td>1.65</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>0.32</td>
<td>-2.2</td>
</tr>
<tr>
<td>0.175</td>
<td>G</td>
<td>-0.29</td>
<td>0.32</td>
<td>0.70</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>1.65</td>
<td>-2.2</td>
<td>1.19</td>
<td>0.97</td>
<td>1.19</td>
<td>0.32</td>
</tr>
<tr>
<td>0.325</td>
<td>T</td>
<td>0.39</td>
<td>-0.79</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>-2.2</td>
<td>1.05</td>
<td>0.13</td>
<td>-0.23</td>
<td>-0.23</td>
<td>-0.23</td>
</tr>
<tr>
<td>1.000</td>
<td>Sum</td>
<td>-0.37</td>
<td>-0.02</td>
<td>-1.02</td>
<td>-4.94</td>
<td>-5.55</td>
<td>-4.94</td>
<td>-5.55</td>
<td>-3.08</td>
<td>-1.13</td>
<td>-2.03</td>
<td>0.19</td>
<td></td>
</tr>
</tbody>
</table>

The use of a weight matrix relies on Bernoulli assumption

If we assume, for the background model, an independent succession of nucleotides (Bernoulli model), the weight $W_S$ of a sequence segment $S$ is simply the sum of weights of the nucleotides at successive positions of the matrix ($W_{ij}$).

In this case, it is convenient to convert the PSSM into a weight matrix, which can then be used to assign a score to each position of a given sequence.

Properties of the weight function

- The weight is
  - positive when \( f_{i,j} > p_i \) (favourable positions for the binding of the transcription factor)
  - negative when \( f_{i,j} < p_i \) (unfavourable positions)

\[
W_{i,j} = \ln \left( \frac{f'_{i,j}}{p_i} \right)
\]

\[
f'_{i,j} = \frac{n_{i,j} + p_i k}{\sum_{i=1}^{A} n_{i,j} + k}
\]

\[
\sum_{i=1}^{A} f'_{i,j} = 1
\]
Information content
Shannon uncertainty

- Shannon uncertainty
  - \(H_s(j)\): uncertainty of a column of a PSSM
  - \(H_g\): uncertainty of the background (e.g. a genome)

- Special cases of uncertainty (for a 4 letter alphabet)
  - \(\text{min}(H)=0\)
    - No uncertainty at all: the nucleotide is completely specified (e.g. \(p=\{1,0,0,0\}\))
  - \(H=1\)
    - Uncertainty between two letters (e.g. \(p=\{0.5,0,0,0.5\}\))
  - \(\text{max}(H)=2\) (Complete uncertainty)
    - One bit of information is required to specify the choice between each alternative (e.g. \(p=\{0.25,0.25,0.25,0.25\}\)).
    - Two bits are required to specify a letter in a 4-letter alphabet.

- \(R_{seq}\)
  - Schneider (1986) defines an information content based on Shannon’s uncertainty.

- \(R^*_{seq}\)
  - For skewed genomes (i.e. unequal residue probabilities), Schneider recommends an alternative formula for the information content.
  - This is the formula that is nowadays used.

\[
H_s(j) = -\sum_{i=1}^{A} f_{i,j} \log_2(f_{i,j})
\]

\[
H_g = -\sum_{i=1}^{A} p_i \log_2(p_i)
\]

\[
R_{seq}(j) = H_g - H_s(j) \quad R_{seq} = \sum_{j=1}^{w} R_{seq}(j)
\]

\[
R^*_{seq}(j) = \sum_{i=1}^{A} f_{i,j} \log_2 \left( \frac{f_{i,j}}{p_i} \right) \quad R^*_{seq} = \sum_{j=1}^{w} R^*_{seq}(j)
\]

### Information content of a PSSM

<table>
<thead>
<tr>
<th>Prior</th>
<th>Pos.</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>
</tr>
</thead>
<tbody>
<tr>
<td>0.325</td>
<td>A</td>
<td>-0.12</td>
<td>0.05</td>
<td>-0.06</td>
<td>-0.08</td>
<td>0.97</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.12</td>
<td>-0.06</td>
</tr>
<tr>
<td>0.175</td>
<td>C</td>
<td>0.08</td>
<td>0.08</td>
<td>0.25</td>
<td>1.50</td>
<td>-0.04</td>
<td>1.50</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.04</td>
<td>0.08</td>
<td>-0.04</td>
<td>0.08</td>
</tr>
<tr>
<td>0.175</td>
<td>G</td>
<td>-0.04</td>
<td>0.08</td>
<td>0.25</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.04</td>
<td>1.50</td>
<td>-0.04</td>
<td>0.68</td>
<td>0.45</td>
<td>0.68</td>
<td>0.08</td>
</tr>
<tr>
<td>0.325</td>
<td>T</td>
<td>0.19</td>
<td>-0.12</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.08</td>
<td>-0.08</td>
<td>0.97</td>
<td>0.05</td>
<td>-0.06</td>
<td>-0.06</td>
<td>-0.06</td>
</tr>
<tr>
<td>1.000</td>
<td>Sum</td>
<td>0.11</td>
<td>0.09</td>
<td>0.36</td>
<td>1.29</td>
<td>0.80</td>
<td>1.29</td>
<td>1.29</td>
<td>0.80</td>
<td>0.61</td>
<td>0.39</td>
<td>0.47</td>
<td>0.04</td>
</tr>
</tbody>
</table>

\[ f'_{i,j} = \frac{n_{i,j} + p_i k}{A} \sum_{i=1}^{A} n_{i,j} + k \]

\[ I_{i,j} = f'_{i,j} \ln\left( \frac{f'_{i,j}}{p_i} \right) \]

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

\[ I_{\text{matrix}} = \sum_{j=1}^{w} \sum_{i=1}^{A} I_{i,j} \]

- \( A \): alphabet size (=4)
- \( n_{i,j} \): occurrences of residue i at position j
- \( w \): matrix width (=12)
- \( p_i \): prior residue probability for residue i
- \( f_{i,j} \): relative frequency of residue i at position j
- \( k \): pseudo weight (arbitrary, 1 in this case)
- \( f'_{i,j} \): corrected frequency of residue i at position j
- \( W_{i,j} \): weight of residue i at position j
- \( I_{i,j} \): information of residue i at position j

Information content $I_{ij}$ of a cell of the matrix

- For a given cell of the matrix
  - $I_{ij}$ is positive when $f'_{ij} > p_i$
    (i.e. when residue $i$ is more frequent at position $j$ than expected by chance)
  - $I_{ij}$ is negative when $f'_{ij} < p_i$
  - $I_{ij}$ tends towards 0 when $f'_{ij} \to 0$
    - because $\lim_{x \to 0} (x \ln(x)) = 0$
Information content of a column of the matrix

- For a given column $i$ of the matrix
  - The information of the column ($I_j$) is the sum of information of its cells.
  - $I_j$ is always positive
  - $I_j$ is 0 when the frequency of all residues equal their prior probability ($f_{ij}=p_i$)
  - $I_j$ is maximal when
    - the residue $i_m$ with the lowest prior probability has a frequency of 1 (all other residues have a frequency of 0)
    - and the pseudo-weight is null ($k=0$).

$$I_j = \sum_{i=1}^{A} I_{i,j} = \sum_{i=1}^{A} f_{i,j}' \ln \left( \frac{f_{i,j}'}{p_i} \right)$$

$$i_m = \arg \min_i (p_i) \quad k = 0$$

$$\max(I_j) = 1 \cdot \ln \left( \frac{1}{p_i} \right) = -\ln(p_i)$$
Schneider & Stephens (1990) propose a graphical representation based on his previous entropy (H) for representing the importance of each residue at each position of an alignment. He provides a new formula for $R_{seq}$

- $H_s(j)$ uncertainty of column $j$
- $R_{seq}(j)$ “information content” of column $j$ (beware, this definition differs from Hertz’ information content)
- $e(n)$ correction for small samples (pseudo-weight)

**Remarks**

- This information content does not include any correction for the prior residue probabilities ($p_i$)
- This information content is expressed in bits.

**Boundaries**

- $\min(R_{seq})=0$ equiprobable residues
- $\max(R_{seq})=2$ perfect conservation of 1 residue with a pseudo-weight of 0,

**Sequence logos can be generated**

- from aligned sequences on the Weblogo server [http://weblogo.berkeley.edu/logo.cgi](http://weblogo.berkeley.edu/logo.cgi)
- From matrices or sequences on enologos [http://www.benoslab.pitt.edu/cgi-bin/enologos/enologos.cgi](http://www.benoslab.pitt.edu/cgi-bin/enologos/enologos.cgi)

$$H_s(j) = - \sum_{i=1}^{A} f_{ij} \log_2(f_{ij})$$

$$R_{seq}(j) = 2 - H_s(j) + e(n)$$

$$h_{ij} = f_{ij} R_{seq}(j)$$
Information content of the matrix

- The total information content represents the capability of the matrix to make the distinction between a binding site (represented by the matrix) and the background model.
- The information content also allows to estimate an upper limit for the expected frequency of the binding sites in random sequences.
- The pattern discovery program consensus (developed by Jerry Hertz) optimises the information content in order to detect over-represented motifs.
- Note that this is not the case of all pattern discovery programs: the gibbs sampler algorithm optimizes a log-likelihood.

\[
I_{\text{matrix}} = \sum_{j=1}^{w} \sum_{i=1}^{A} I_{i,j}
\]

\[
P(\text{site}) \leq e^{-I_{\text{matrix}}}
\]

The upper bound of $I_j$ increases when $p_i$ decreases

- $I_j \rightarrow \infty$ when $p_i \rightarrow 0$

The information content, as defined by Gerald Hertz, has thus no upper bound.
References - PSSM information content

- Seminal articles by Tom Schneider
  - Tom Schneider’s publications online

- Seminal article by Gerald Hertz

- Software tools to draw sequence logos
  - Weblogo
    • [http://weblogo.berkeley.edu/logo.cgi](http://weblogo.berkeley.edu/logo.cgi)
  - Enologos
    • [http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi](http://biodev.hgen.pitt.edu/cgi-bin/enologos/enologos.cgi)