12 Sequence profiles

12.1 Sequence profiles and patterns

Protein secondary structures

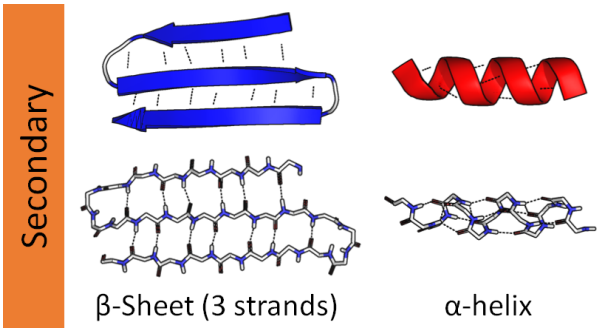

Figure 12.1: Protein secondary structures (source: Shafee, Wikimedia Commons)

Functional regions found in MSA

- `http://www.bioinformatics.org/strap/`
- `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0070843`

Applications of MSAs

- Position weight matrix
- Sequence profile
- HMM profile
- Motifs

12.2 Position weight matrix

A position weight matrix (PWM) is a two-dimensional array that contains position-specific scores. PWMs usually contain no gaps.

Creating a position probability matrix (PPM)

It requires an MSA without gaps.

Example of PPM

Make a PPM from the alignment below.

```
Seq1 AGT
Seq2 CAG
Seq3 AAT
Seq4 ATT
```

Position-specific frequencies

	1	2	3
A	3	2	0
G	0	1	1
C	1	0	0
T	0	1	3

PPM

	1	2	3
A	0.75	0.5	0
G	0	0.25	0.25
C	0.25	0	0
T	0	0.25	0.75

From PPM to PWM

Similar to pair-wise scores, log-odds scores can be used for profiles.

$$\text{PWM}_{ar} = \log \frac{\text{PPM}_{ar}}{q_a}$$

q_a : Background probability of a
r : Position in MSA

12.3 Sequence profiles

A protein sequence profile is a two-dimensional array that contains position-specific scores.

Profile values

A profile is based on position-specific weights and a score matrix.

Prof_{ra} : Position-specific score of a at position r

R_{ab} : Pair-wise score of a and b
r : Position in MSA
a, b : Nucleotide/amino acid element
M : All nucleotides/amino acids

W_{rb} : Weight value of b at position r

Profile with linear weights

$$\mathrm{Prof}_{ra} = \frac{1}{m_r} \sum_{b \in M} R_{ba} F_{rb}$$

$$W_{rb} = \frac{F_{rb}}{m_r}$$

F_{rb} : The number of occurrences of b at position r

m_r : The number of residues without gaps at position r

Example of profile with linear weights

Make a profile with linear weights.

Alignment

```
Seq1 AGC
Seq2 -AC
Seq3 AAT
```

Scoring matrix

	A	G	C	T
A	2	1	-3	-2
G	1	3	-2	-1
C	-3	-2	4	1
T	-2	-1	1	2

Scores can be calculated as follows.

$$\begin{aligned}
\text{A1}: &\quad 1/2 \times (2 \times 2 + 1 \times 0 + (-3) \times 0 + (-2) \times 0) = 1/2 \times 4 = 2 \\
\text{G1}: &\quad 1/2 \times (1 \times 2 + 3 \times 0 + (-2) \times 0 + (-1) \times 0) = 1/2 \times 2 = 1 \\
\text{C1}: &\quad 1/2 \times ((-3) \times 2 + (-2) \times 0 + 4 \times 0 + 1 \times 0) = 1/2 \times (-6) = -3 \\
\text{T1}: &\quad 1/2 \times ((-2) \times 2 + (-1) \times 0 + 1 \times 0 + 2 \times 0) = 1/2 \times (-4) = -2
\end{aligned}$$

$$\begin{aligned}
\text{A2}: &\quad 1/3 \times (2 \times 2 + 1 \times 1 + (-3) \times 0 + (-2) \times 0) = 1/3 \times 5 = 1.67 \\
\text{G2}: &\quad 1/3 \times (1 \times 2 + 3 \times 1 + (-2) \times 0 + (-1) \times 0) = 1/3 \times 5 = 1.67 \\
\text{C2}: &\quad 1/3 \times ((-3) \times 2 + (-2) \times 1 + 4 \times 0 + 1 \times 0) = 1/3 \times (-8) = -2.67 \\
\text{T2}: &\quad 1/3 \times ((-2) \times 2 + (-1) \times 1 + 1 \times 0 + 2 \times 0) = 1/3 \times (-5) = -1.67
\end{aligned}$$

$$\begin{aligned}
\text{A3}: &\quad 1/3 \times (2 \times 0 + 1 \times 0 + (-3) \times 2 + (-2) \times 1) = 1/3 \times (-8) = -2.67 \\
\text{G3}: &\quad 1/3 \times (1 \times 0 + 3 \times 0 + (-2) \times 2 + (-1) \times 1) = 1/3 \times (-5) = -1.67 \\
\text{C3}: &\quad 1/3 \times ((-3) \times 0 + (-2) \times 0 + 4 \times 2 + 1 \times 1) = 1/3 \times (9) = 3 \\
\text{T3}: &\quad 1/3 \times ((-2) \times 0 + (-1) \times 0 + 1 \times 2 + 2 \times 1) = 1/3 \times (4) = 1.33
\end{aligned}$$

Calculated profile with linear weights.

	A	G	C	T
1	2	1	-3	-2
2	1.67	1.67	-2.67	-1.67
3	-2.67	-1.67	3	1.33

Non-linear weights

Amino acids/nucleotides occurring many times are "favored".

$$\mathrm{W}_{rb} = \frac{\ln((1 - F_b)/(1 + m_r))}{\ln(1/(1 + m_r))}$$

Amino acids/nucleotides occurring many times are "punished".

$$\mathrm{W}_{rb} = \frac{1 + \ln(1 - F_b)}{1 + \ln m_r}$$

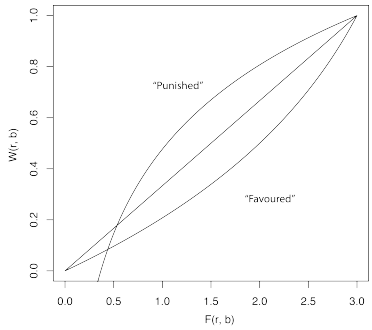

Figure 12.2: Two different weight functions)

Treating gaps

Position-specific gap penalties are usually added to profiles.

12.4 Profile search

A constructed profile can be used to find sequence patterns.

Profile score of a query sequence

The score of a query sequence can be calculated by adding all corresponding position-specific scores.

Example of profile score

Find the best score for q = AGCT.

Profile:

	A	G	C	T	Gap	Len
1	5	-5	-2	-1	10	10
2	-2	3	4	-7	10	10
3	1	2	1	-1	5	7
4	-3	3	-2	7	10	10

Score: $5 + 3 + 1 + 7 = 16$

Searching databases with a profile

A dynamic programming method can be used for a profile search.

$$H_{i,j} = \max \left\{ 0, \mathrm{Prof}_{jd_i} + \max \left\{ \begin{array}{l} H_{i-1,j-1} \\ \max_{2 \ll k \ll j-1} H_{i-1,j-k} - g_k^d \\ \max_{2 \ll l \ll i-1} H_{i-l,j-1} - g_l^d \end{array} \right. \right.$$

where g_k^d and g_l^P are database and profile gap penalties.

Example of database search with profile

d1 = ACT

Gap penalty: $5 + 2(l - 1)$

Profile:

	A	G	C	T	Gap	Len
1	5	-5	-2	-1	10	10
2	-2	3	4	-7	10	10
3	1	2	1	-1	5	7
4	-3	3	-2	7	10	10

DP table:

		1	2	3	4
	0	0	0	0	0
A	0	5	0	1	0
C	0	0	9	5	0
T	0	0	0	8	12

$H_{1,1}$: 5 $Prof_{1A}$: 5 **Diagonal: 5 + 0** Vertical: 5 + (0 - 10) Horizontal: 5 + (0 - 5)	$H_{1,2}$: 0 $Prof_{2A}$: -2 Diagonal: -2 + 0 Vertical: -2 + (0 - 10) Horizontal: -2 + (5 - 5)	$H_{1,3}$: 1 $Prof_{3A}$: 1 **Diagonal: 1 + 0** Vertical: 1 + (0 - 10) Horizontal: 1 + (5 - 7)	$H_{1,4}$: 0 $Prof_{4A}$: -3 Diagonal: -3 + 0 Vertical: -3 + (0 - 10) Horizontal: -3 + (1 - 5)
$H_{2,1}$: 0 $Prof_{1C}$: -2 Diagonal: -2 + 0 Vertical: -2 + (5 - 10) Horizontal: -2 + (0 - 5)	$H_{2,2}$: 9 $Prof_{2C}$: 4 **Diagonal: 4 + 5** Vertical: 4 + (0 - 10) Horizontal: 4 + (0 - 5)	$H_{2,3}$: 5 $Prof_{3C}$: 1 Diagonal: 1 + 0 Vertical: 1 + (1 - 10) **Horizontal: 1 + (9 - 5)**	$H_{2,4}$: 0 $Prof_{4C}$: -2 Diagonal: -2 + 1 Vertical: -2 + (0 - 10) Horizontal: -2 + (9 - 7)
$H_{3,3}$: 0 $Prof_{1T}$: -1 Diagonal: -1 + 0 Vertical: -1 + (0 - 10) Horizontal: -1 + (0 - 5)	$H_{3,2}$: 0 $Prof_{2T}$: -7 Diagonal: -7 + 0 Vertical: -7 + (9 - 10) Horizontal: -7 + (0 - 5)	$H_{3,3}$: 8 $Prof_{3T}$: -1 **Diagonal: -1 + 9** Vertical: -1 + (5 - 10) Horizontal: -1 + (0 - 5)	$H_{3,4}$: 12 $Prof_{4T}$: 7 **Diagonal: 7 + 5** Vertical: 7 + (0 - 10) Horizontal: 7 + (8 - 5)

Alignment:

```
profile: 1234
     d1: AC-T
```

12.5 PSI-BLAST

Position-specific iterated BLAST (PSI-BLAST) is an extension of BLAST. It is much more sensitive than BLAST. It can be used to find distantly related proteins.

Pseudo-code of linear progressive alignment (general progressive alignment)

Algorithm 12.1: Simplified procedure of PSI-BLAST

q: query sequence
t : threshold for significance

Q = BLAST(q, t);

do
  Q1 = Reduce(Q) ; `// Remove identical segments`
  M = MultipleAlignment(Q1);
  P = Profile(M);
  Q = ProfileSearch(P);
while convergence(Reduce(Q) = Q1) or maximum number of cycles;