Part V

11 Construction of scoring matrix

11.1 Scoring schemes for protein sequence alignment

Applying an appropriate scoring scheme is critical to create biologically accurate alignments and phylogenetic trees.

Different types of scoring schemes for proteins

- Use of identity
- Use of the genetic code
- Use of a classification of amino acids
- Scoring matrix

Use of identity

The score is calculated by counting identical amino acids. It is equivalent with a simple scoring scheme with match: 1, mismatch: 0, and gap penalty: 0.

Example of "use of identity"

Calculate the SP score by counting identical amino acids.

```
Seq1 F-NV
Seq2 FPN-
Seq3 FC-V
```

$S(\bar{s}^1, \bar{s}^2) = 2$
$S(\bar{s}^1, \bar{s}^3) = 2$
$S(\bar{s}^2, \bar{s}^3) = 1$

$$S(\mathcal{A}) = S(\bar{s}^1, \bar{s}^2) + S(\bar{s}^1, \bar{s}^3) + S(\bar{s}^2, \bar{s}^3) = 2 + 2 + 1 = 5$$

Score: 5

Use of the genetic code

The score is based on the distance between two amino acids at the codon level.

Example of "use of the genetic code"

```
Seq1 FFFF
Seq2 FCNG
```

```
Phe (UUU, UUC) & Phe (UUU, UUC): 3
Phe (UUU, UUC) & Cys (UGU, UGC): 2
Phe (UUU, UUC) & Asn (AAU, AAC): 1
Phe (UUU, UUC) & Glu (GAA, GAG): 0
```

Score: 6

Use of a classification of amino acids

The score is based on the physio-chemical properties. For example, AACH (amino acid class hierarchy) can be used as a scoring scheme.

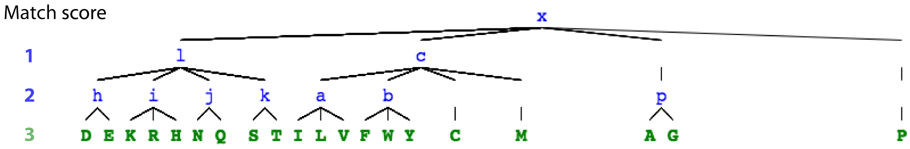


Figure 11.1: Example of amino acid class hierarchy (AACH)

Example of "Use of a classification of amino acids"

Calculate the score by using AACH.

```
Seq1 DDDP
Seq2 DEKD
```

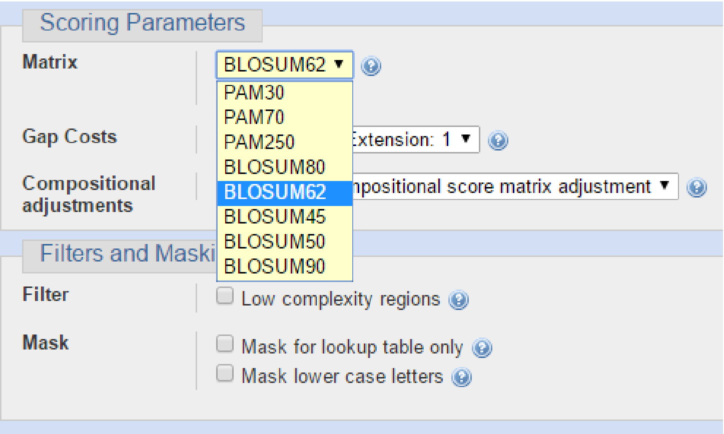
D & D: 3, D & E: 2, D & K: 1, P & D: 0
Score: 6

Scoring matrix

- DNA/RNA: 4 × 4
- Protein: 20 × 20

PAM and BLOSUM

BLAST parameters



Scoring Parameters

Matrix BLOSUM62 ▾

Gap Costs xtension: 1 ▾

Compositional adjustments npositional score matrix adjustment ▾

Filters and Maski

Filter □ Low complexity regions

Mask □ Mask for lookup table only
□ Mask lower case letters

Figure 11.2: BLAST score parameters (source: http://blast.ncbi.nlm.nih.gov)

Correspondence between PAM and BLOSUM

PAM 120	PAM 160	PAM 250
BLOSUM 80	BLOSUM 62	BLOSUM 45

Types of substitutions

There are several types of substitutions between two sequences from the common ancestor.

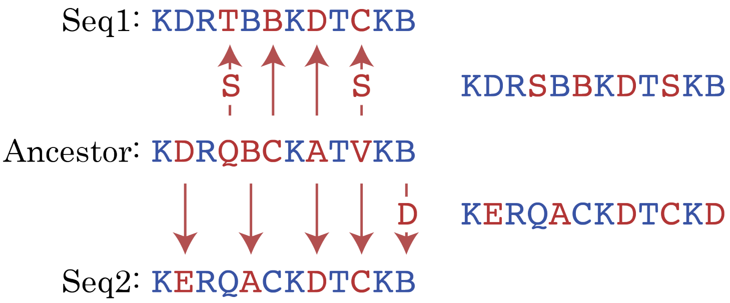


Figure 11.3: Different types of substitutions

Exercise 11.1

Calculate the score of the alignment by using different scoring schemes.

```
Seq1 K-RI
Seq2 KDCC
```

- Use the identity.

- Use the genetic code.

K	Lys	AAA, AAG
D	Asp	GAU, GAC
R	Arg	CGU, CGC, CGA
I	Ile	AUU, AUC AUA
C	Cys	UGU, UGC

- Use AACH.

11.2 PAM accepted mutations

PAM is a popular scoring scheme for protein sequence alignments. It is based on substitution matrices created from experiment data.

Accepted point mutations

- Independent of positions and neighbor residues
- Independent from previous mutations at the same position
- Biological clock is assumed (the rate of mutations is constant)

PAM (point accepted mutation)

One PAM means one accepted point mutation per 100 residues.
Resources of constructing a PAM score

- 34 super-families
- 71 groups of homologous sequences (85% identity)

Preparations for constructing a PAM score

Counting the number of mutations is the fist step to make a PAM score. Several sub-steps are involved.

- Create a phylogenetic tree
- Estimate ancestor sequences
- Count all occurrences of mutations

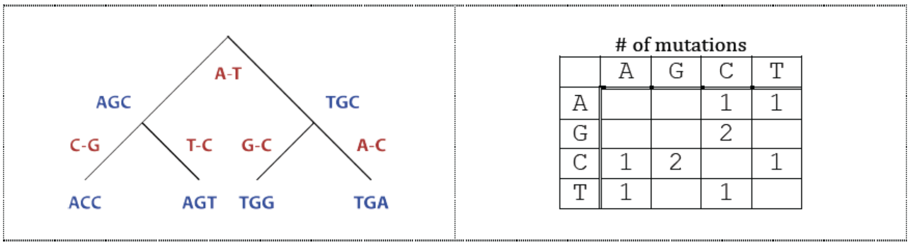
Frequencies of estimated mutations

Frequencies of estimated mutations are counted in internal nodes of the reconstructed tree.

$$
\begin{aligned}
f_{ab} &: \text{The number of mutations from } a \text{ to } b \text{ or from } b \text{ to } a \\
f_a &: \text{The total number of mutations in which a takes part} \\
f &: \text{Twice the total number of mutations}
\end{aligned}
$$

Example of frequency calculation

Calculate f_{CA}, f_C, and f from the phylogenetic tree and the table below.



# of mutations

	A	G	C	T
A			1	1
G			2	
C	1	2		1
T	1		1	

Figure 11.4: Phylogenetic tree and a table of the number of mutations

$$
\begin{aligned}
f_{CA} &= 1 \\
f_C &= 1 + 2 + 1 = 4 \\
f &= 10
\end{aligned}
$$

Background frequencies

The background probabilities are calculated from the data source.

p_a : The relative occurrence of a in the observed sequences

Example of background frequencies

Calculate p_G from the sequences below.

```
Seq1 ACC
Seq2 AGT
Seq3 TGG
Seq4 TGA
```

$$p_G = \frac{4}{12} \approx 0.333$$

11.3 PAM substitution matrix

PAM is based on a substitution matrix created from experimental data.

Relative mutability

The probabilities of amino acid mutations are calculated based on relative mutability.

$$m_a : \frac{1}{100p_a} \times \frac{f_a}{f}$$

Example of relative mutability calculation

- Frequencies of estimated mutations

f_A: 2	f_G: 2	f_C: 4	f_T: 2
f : 10			

- Background frequencies

p_A: 3/12	p_G: 4/12	p_C: 2/12	p_T: 3/12
$100p_A$: 23	$100p_G$: 33.33	$100p_C$: 16.67	$100p_T$: 25

- Relative mutability (1 PAM)

m_A: 0.008	m_G: 0.006	m_C: 0.024	m_T: 0.008

Mutation probability

Mutation probabilities are summarized in a matrix format called substitution matrix.

$$M_{ab} : m_a \times \frac{f_{ab}}{f_a} \quad M_{aa} : 1 - m_a$$

Example of substitution matrix

- Frequencies of estimated mutations

f_{AC}: 1	f_{AT}: 1	f_{GC}: 2	f_{CT}: 1
f_{CA}: 1	f_{TA}: 1	f_{CG}: 2	f_{TC}: 1
f_{A}: 2	f_{G}: 2	f_{C}: 4	f_{T}: 2

- Relative mutability (1 PAM)

m_A: 0.008	m_G: 0.006	m_C: 0.024	m_T: 0.008

- Mutation probabilities

m_{AC}: 0.004	m_{AT}: 0.004		
m_{GC}: 0.006			
m_{CA}: 0.006	m_{GC}: 0.012	m_{CT}: 0.006	
m_{TA}: 0.004	m_{TC}: 0.004		
m_{AA}: 0.992	m_{GC}: 0.994	m_{CC}: 0.976	m_{TT}: 0.992

- Substitution matrix

	A	G	C	T
A	0.992		0.004	0.004
G		0.994	0.006	
C	0.006	0.012	0.976	0.006
T	0.004		0.004	0.992

Matrices for general evolutionary time

Markov chains can be used to generalize PAM with arbitrary values. For instance, the substitution value for 2 PAM (=2) for amino acids a to b can be calculated as:

$$M^2_{ab} = M_{ab}M_{bb} + M_{aa}M_{ab} + \sum_{c \notin \{a,b\}} M_{ac}M_{cb} = \sum_{c \in M} M_{ac}M_{cb}$$

Odds matrix

Substitution scores can be transformed to odds values. Odds values O_{ab} are equal to O_{ba}, and therefore an odds matrix is symmetrical.

$$O_{ab} = \frac{M_{ab}}{p_b}$$

when $a \neq b$:

$$O_{ab} = \frac{M_{ab}}{p_b} = m_a \times \frac{f_{ab}}{f_a} \times \frac{1}{p_b} = \frac{1}{100p_a} \times \frac{f_a}{f} \times \frac{f_{ab}}{f_a} \times \frac{1}{p_b} = \frac{f_{ab}}{100fp_ap_b}$$

Transformation of an odds matrix to a score matrix

Odds values can be further transformed to log-odds values.

$$R_{ab} = \log O_{ab} = \log \frac{M_{ab}}{p_b}$$

11.4 BLOSUM

BLOSUM is another popular method of constructing a scoring marix. It is more useful for diverse sequences than PAM.

Resources of constructing a BLOSUM score

- Scanned very conserved regions of protein families on the BLOCKS database
- Identified 2000 blocks
- A block contains multiple sequences that are highly conserved

Observed mutations

T : The total number of pairs from all blocks.
The number of amino acid pairs of a block with length w
and m sequences can be calculated as $1/2wm(m-1)$.
f_{ab} : The frequencies of an observed pair a and b.

Example of observed mutations

Block1	Block2
`AGCC`	`AGA`
`TAGC`	`TAC`
`AGCC`	

$$T = 1/2 \cdot 4 \cdot 3 \cdot 2 + 1/2 \cdot 3 \cdot 2 \cdot 1 = 12 + 3 = 15$$

f_{AA}: 1/15	f_{AC}: 1/15	f_{AG}: 3/15	f_{AT}: 3/15
f_{GG}: 1/15	f_{GC}: 2/15	f_{CC}: 4/15	

Example of observed mutations

- Frequencies of estimated mutations

f_A: 2	f_G: 2	f_C: 4	f_T: 2
f : 10			

Background frequencies

$$p_a = f_{aa} + \sum_{e \neq a} \frac{f_{ae}}{2}$$

$$\begin{aligned} e_{aa} &= p_a \cdot p_a \\ e_{ab} &= p_a \cdot p_b + p_b \cdot p_a = 2 \cdot p_a \cdot p_b \end{aligned}$$

Example of background frequencies

$$\begin{aligned} p_A &= \frac{1}{15} + \frac{1}{2} \times \left(\frac{3}{15} + \frac{1}{15} + \frac{3}{15} \right) = \frac{1}{15} + \frac{7}{30} = \frac{9}{30} \\ p_G &= \frac{1}{15} + \frac{1}{2} \times \left(\frac{3}{15} + \frac{2}{15} \right) = \frac{1}{15} + \frac{5}{30} = \frac{7}{30} \\ p_C &= \frac{4}{15} + \frac{1}{2} \times \left(\frac{2}{15} + \frac{1}{15} \right) = \frac{4}{15} + \frac{3}{30} = \frac{11}{30} \\ p_T &= \frac{0}{15} + \frac{1}{2} \times \left(\frac{3}{15} \right) = \frac{0}{15} + \frac{3}{30} = \frac{3}{30} \end{aligned}$$

$$\begin{aligned} e_{AA} &= p_A \cdot p_A = \frac{9}{30} \times \frac{9}{30} = \frac{81}{900} \\ e_{AG} &= p_A \cdot p_G = 2 \times \frac{9}{30} \times \frac{7}{30} = \frac{126}{900} \end{aligned}$$

Scoring matrix

BLOSUM scores are calculated as log ratios of observed and background probabilities.

$$R_{ab} = \log \frac{f_{ab}}{e_{ab}}$$

BLOSUM scores of different distances

One can categorize segments by an identify x to create BLOSUM x.

Example of BLOSUM x

1	`AGCC`
2	`TAGC`
3	`AGTC`
4	`AGTT`

100% identity

1	`AGCC`
2	`TAGC`
3	`AGTC`
4	`AGTT`

Number of mutations in the first column:
6 (3 ATs & 3 AAs)

75% identity

1,3	C AG C T
2	TAGC
4	AGTT

Number of mutations in the first column:
3 (2 ATs & 1 AA)

50% identity

1,3,4	CC AGTC TT
2	TAGC

Number of mutations in the first column:
1 (1 AT)