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.

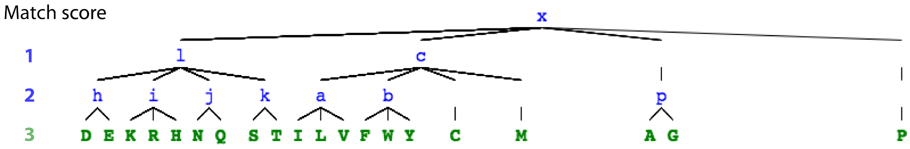


Figure 11.1: Example of amino acid class hierarchy (AACH)

Example of "Use of a classification of amino acids"

Calculate the score by using AACH.

```
Seq1 DDDP
Seq2 DEKD
```

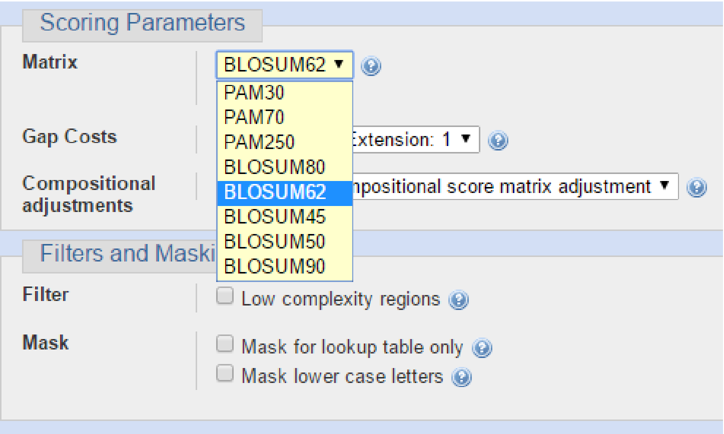
D & D: 3, D & E: 2, D & K: 1, P & D: 0
Score: 6

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:)

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.

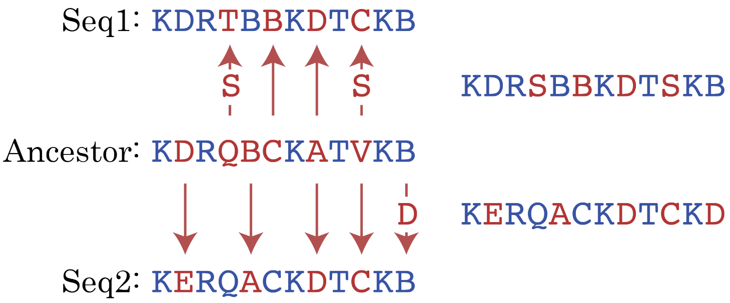


Figure 11.3: Different types of substitutions

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.