# Sequence alignment

**Course code: MIM 5531**
**Course Title: Bioinformatics and Biostatistics**
**Course Teacher: Jennifer Michellin Kiruba N**

- A global sequence-alignment method aligns and compares two sequences along their entire length, and comes up with the best alignment that displays the maximum number of nucleotides or amino acids aligned.
- The algorithm that drives global alignment is the Needleman-Wunsch algorithm.
- A global alignment algorithm starts at the beginning of two sequences and adds gaps to each until the end of one is reached.
- Global alignment works the best when the sequences are similar in character and length.
- Because global alignment displays the best alignment between two sequences using the entire sequence, it may miss a small region of biological importance.
- Two of the available web servers for pairwise global alignment are EMBL-EBI EMBOSS (http://www.ebi.ac.uk/Tools/psa/), and NCBI specialized BLAST

# Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

# Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

input string

```
HEAGAWGHEEAHGEGAE
PAWHEAEHE
```

## Global alignment

```
HEAGAWGHEEAHGEGAE
--|-|||-|-||--|-||
--P-AW-H-EA--E-HE
```

## Local alignment

```
AWGHEEAH
||-|||||
AW-HEAEH
```

- Local sequence alignment is intended to find the most similar regions in two sequences being aligned. The algorithm that drives local alignment is the **SmithWaterman algorithm**.
- A local alignment algorithm finds the region of highest similarity between two sequences and builds the alignment outward from this region.
- If there are multiple regions of very high similarity, the same principle applies.
- Obviously, local alignment is useful for sequences that are not similar in character and length, yet are suspected to contain small regions of similarity, such as biologically important motifs.

TABLE 6.1 Online Pairwise Alignment Tools Using the Smith–Waterman Algorithm

| Online Tool | URL |
|---|---|
| PIR SSEARCH | http://pir.georgetown.edu/pirwww/search/pairwise.shtml[5] |
| NCBI specialized BLAST | bl2seq resource; look for the Align link on the NCBI BLAST home page under Specialized BLAST |
| SIM | http://web.expasy.org/sim/ |
| LALIGN* | http://www.ch.embnet.org/software/LALIGN_form.html |

**The LALIGN program is William Pearson's, and it implements the algorithm of X. Huang and W. Miller.*[6]

- Both these algorithms are examples of dynamic programming.
- An algorithm is a step-by-step procedure that utilizes a finite number of instructions for automated reasoning and the calculation of a function.
- Dynamic programming is a method for solving complex problems by breaking them down into simpler subproblems.
- In the case of sequence alignment, dynamic programming involves setting up a two-dimensional matrix in which one sequence is listed vertically and the other sequence is listed horizontally; then calculating the scores, one row at a time.
- A 100% perfect alignment will produce a diagonal straight line (with a negative slope) spanning from the top left to bottom right.
- If the alignment is not perfect, gaps are introduced in the matrix.
- For the sequence represented horizontally, gaps are introduced vertically,
- and for the sequence represented vertically, gaps are introduced horizontally,
- and the alignment is determined by a traceback step.

# Smith-Waterman Scoring

| | D | E | - | S |
|---|---|---|---|---|

| | | - | D | E | S | I | G | N |
|---|---|---|---|---|---|---|---|---|
| | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | I | 0 | 0 | 0 | 0 | 5 | 4 | 3 |
| D | D | 0 | 5 | 4 | 3 | 4 | 4 | 3 |
| E | E | 0 | 4 | 10 | 9 | 8 | 7 | 6 |
| A | A | 0 | 3 | 9 | 9 | 8 | 7 | 6 |
| S | S | 0 | 2 | 8 | 14 | 13 | 12 | 11 |

Match = +5

Mismatch = -1

Gap = -1

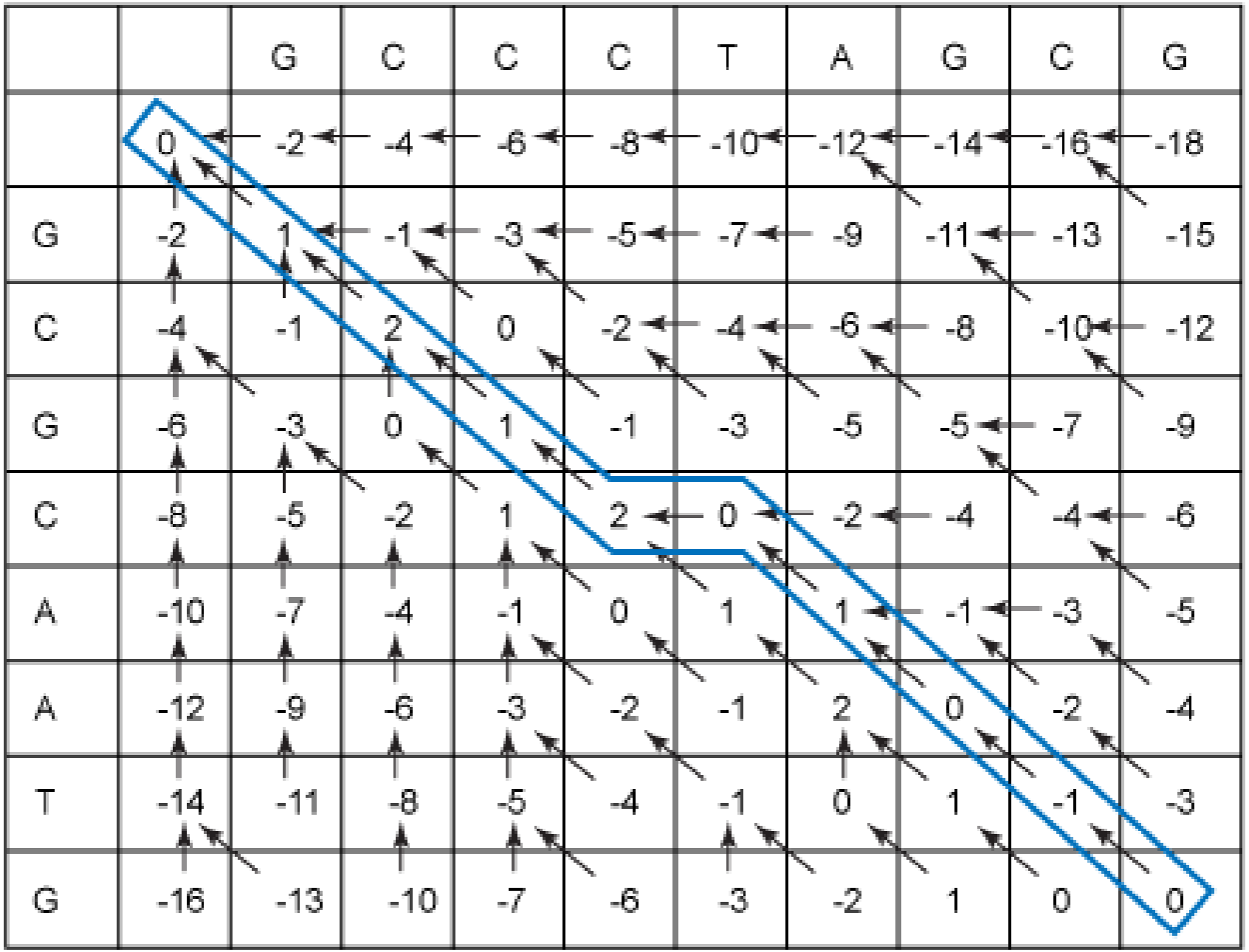
1: DESIGN

2: IDEAS

Aligned:

1: DE-S

2: DEAS


G C C C T A G C G
0 -2 -4 -6 -8 -10 -12 -14 -16 -18
G -2 1 -1 -3 -5 -7 -9 -11 -13 -15
C -4 -1 2 0 -2 -4 -6 -8 -10 -12
G -6 -3 0 1 -1 -3 -5 -5 -7 -9
C -8 -5 -2 1 2 0 -2 -4 -4 -6
A -10 -7 -4 -1 0 1 1 -1 -3 -5
A -12 -9 -6 -3 -2 -1 2 0 -2 -4
T -14 -11 -8 -5 -4 -1 0 1 -1 -3
G -16 -13 -10 -7 -6 -3 -2 1 0 0

# Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

# Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  ||||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

- In both global and local alignment, the final output is given an alignment score.
- Gaps have to be introduced to improve the alignment. The reason gaps are introduced is because one of the sequences may have gained or lost sequence characteristics (insertion-deletion) during evolution that did not happen with the other sequence.
- However, the number of gaps is kept to a minimum to keep the alignment meaningful; otherwise an artificially high alignment score can be obtained even when the two sequences are not related.
- The gap penalty value is subtracted from the gross alignment score to obtain the final alignment score
- The insertion of no more than 1 gap per 20 amino acid residues is ideal but that is not possible in most cases.
- For each gap opened, a gap-opening penalty value is assigned, and for each gap extended, a gap-extension penalty value is assigned.
- A gap-opening penalty is always much higher than a gap-extension penalty. Often, a default value of 210 for a gap-opening penalty and 21 for a gap-extension penalty are used.

A raw alignment score can be calculated based on the following simple formula:

$$S = \Sigma_i + \Sigma_m - G_t, \tag{6.4}$$

where $S$ = raw score, $\Sigma_i$ = total score for identities, $\Sigma_m$ = total score for mismatches, and $G_t$ = total gap penalty.

relevant for sequence alignment. Affine gap penalty is calculated as follows:

$$G_t = G_o + G_e \times L_n, \tag{6.3}$$

where $G_t$ = total gap penalty, $G_o$ = gap-opening penalty, $G_e$ = gap-extension penalty, and $L_n$ = length of the extension gaps. For any given block of gaps, $L_n$ = # of total gaps − 1, because the first gap is the opening, the rest in the block are extensions.

# Scoring matrices

Unit II

**Histone H1** (residues 120-180)

| Buried residue | Conservative substitutions | Nonconservative substitutions |
|---|---|---|
| Val | Ala, Phe | Gly, Asp, Thr |
| Ile | Phe, Cys | Gly, Glu, Thr |
| Leu | Phe, Val | Gly, Glu, Thr |
| Met | Phe, Val | Gly, Glu, Thr |
| Phe | Leu, Met | Gly, Glu, Thr |
| Cys | Val, Met | Gly, Asp, Ser |
| Trp | Phe, Leu | Gly, Glu, Thr |

- For both nucleic acids and proteins, the alignment score is calculated using a scoring matrix.
- A scoring matrix is a set of values representing the likelihood of one residue being substituted by another during sequence divergence through evolution.
- This is why the scoring matrix is also known as the substitution matrix.

- A scoring matrix for comparing DNA sequences can be simple because there are only four nucleotides and the mutation frequencies are assumed to be equal (the Jukes and Cantor assumption).
- A high positive score (e.g. 5) is assigned for a match and a low negative score (e.g. 24) for a mismatch, thus creating a simple model.
- However, the frequency of transition mutations (purine replaced by purine or pyrimidine replaced by pyrimidine) is higher than transversion mutations (purine replaced by pyrimidine or vice versa): Kimura and others

# Scoring Matrices for DNA Sequences

- Transition: A ←→ G  C ←→ T
- Transversion: a purine (A or G) is replaced by a pyrimadine (C or T) or vice versa

Identity Matrix

| | A | T | C | G |
|---|---|---|---|---|
| A | 1 | 0 | 0 | 0 |
| T | 0 | 1 | 0 | 0 |
| C | 0 | 0 | 1 | 0 |
| G | 0 | 0 | 0 | 1 |

BLAST Matrix

| | A | T | C | G |
|---|---|---|---|---|
| A | 5 | –4 | –4 | –4 |
| T | –4 | 5 | –4 | –4 |
| C | –4 | –4 | 5 | –4 |
| G | –4 | –4 | –4 | 5 |

Transition Transversion Matrix

| | A | T | C | G |
|---|---|---|---|---|
| A | 1 | –5 | –5 | –1 |
| T | –5 | 1 | –1 | –5 |
| C | –5 | –1 | 1 | –5 |
| G | –1 | –5 | –5 | 1 |

**FIGURE 2.4** *Scoring matrices for aligning DNA sequences.*

- Scoring matrices for amino-acid substitutions are more complex, reflecting the similarity of physicochemical properties, as well as the likelihood of one amino acid being substituted by another at a particular position in homologous proteins.
- The scoring matrices for proteins are 20 X 20 matrices.
- Two well-known types of scoring matrices for proteins are PAM and BLOSUM.

# PAM Matrices

- PAM (point accepted mutation—that is, accepted point mutation—also called percent accepted mutation) matrices were first developed by Margaret Dayhoff and colleagues in 1978 and hence are also known as Dayhoff PAM matrices.
- A PAM represents a substitution of one amino acid by another that has been fixed by natural selection because either it does not alter the protein function or it is beneficial to the organism.
- In a PAM1 matrix, which is the original PAM matrix generated, a PAM unit is an evolutionary time over which 1% of the amino acids in a sequence are expected to undergo accepted mutations, resulting in 1% sequence divergence.
- The PAM1 matrix was built by aligning closely related protein sequences (71 protein families) that had at least 85% sequence identity

- Subsequently, in order to deal with protein sequences that are more diverged and distantly related, other PAM matrices, such as PAM100 and PAM250, were generated.
- These later PAM matrices were generated by multiplying the PAM1 matrix by itself hundreds of times.
- For example, the PAM250 matrix can be obtained by multiplying the PAM1 matrix by itself 250 times over.
- The values in the matrix are log odds scores

**PAM250 matrix**

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | A | 2 | | | | | | | | | | | | | | | | | | |
| Arg | R | -2 | 6 | | | | | | | | | | | | | | | | | |
| Asn | N | 0 | 0 | 2 | | | | | | | | | | | | | | | | |
| Asp | D | 0 | -1 | 2 | 4 | | | | | | | | | | | | | | | |
| Cys | C | -2 | -4 | -4 | -5 | 12 | | | | | | | | | | | | | | |
| Gln | Q | 0 | 1 | 1 | 2 | -5 | 4 | | | | | | | | | | | | | |
| Glu | E | 0 | -1 | 1 | 3 | -5 | 2 | 4 | | | | | | | | | | | | |
| Gly | G | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | | | | | | | | | | | |
| His | H | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | | | | | | | | | | |
| Ile | I | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | | | | | | | | | |
| Leu | L | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | 2 | 6 | | | | | | | | |
| Lys | K | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | | | | | | | |
| Met | M | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | | | | | | |
| Phe | F | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | | | | | |
| Pro | P | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | | | | |
| Ser | S | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | | | |
| Thr | T | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | | |
| Trp | W | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | |
| Tyr | Y | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 |
| Val | V | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |
| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |

FIGURE 6.9 A PAM250 substitution matrix made by writing the amino acids in alphabetical order.

- Jones et al. updated the PAM matrix by taking into account 2621 families of sequences (16,000 homologous protein sequences) from the Swiss-Prot database.
- The sequences were clustered at 85% identity level as was done in the original PAM matrix, and the raw mutation frequency matrix was processed in a similar way as in the PAM matrix.
- This updated PAM matrix is called the PET91 matrix (1991).
- Thus, PET91 takes into account the substitutions that were poorly represented in the original Dayhoff matrix.
- The overall character of PAM and PET91 matrices is similar.

# BLOSUM

- BLOSUM (blocks substitution matrices) scoring matrices were proposed by Steven Henikoff and Jorja Henikoff in 1992.
- BLOSUM represents an alternative set of scoring matrices, which are widely used in sequence alignment algorithms.
- Like PAM, BLOSUM matrices are also log-odds matrices.
- BLOSUM matrices were developed based on multiple alignment of 500 groups of related protein sequences, which yielded.
- 2000 blocks of conserved amino-acid patterns.
- Blocks are ungapped multiple sequence alignments corresponding to the most conserved regions of the proteins involved

- In each multiple alignment, the sequences showing similar % identity were clustered into groups and averaged. Using these groups, the substitution frequencies for all pairs of amino acids were calculated and the matrix was developed.
- Therefore, the blocks of ungapped multiple sequence alignments, which are the cornerstone of BLOSUM matrices, reveal the evolutionary relationship between proteins.
- BLOCKS database was developed to host these multiple sequence alignments that reveal the blocks.
- By 1996, there were B3000 blocks reported, based on 770 protein families.
- Different BLOSUM matrices differ in the % sequence identity used in clustering

**BLOSUM62 matrix**

| | | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | A | 4 | | | | | | | | | | | | | | | | | | | |
| Arg | R | -1 | 5 | | | | | | | | | | | | | | | | | | |
| Asn | N | -2 | 0 | 6 | | | | | | | | | | | | | | | | | |
| Asp | D | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | | |
| Cys | C | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | | |
| Gln | Q | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | | |
| Glu | E | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | | |
| Gly | G | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | | |
| His | H | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | | |
| Ile | I | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | | |
| Leu | L | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | | |
| Lys | K | -1 | 2 | 0 | -1 | -3 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | | |
| Met | M | -1 | -1 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | | |
| Phe | F | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | | |
| Pro | P | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | | |
| Ser | S | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | | |
| Thr | T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | | |
| Trp | W | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | | |
| Tyr | Y | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | |
| Val | V | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

FIGURE 6.10 BLOSUM62 substitution matrix made by writing the amino acids in alphabetical order.

- BLOSUM62 is useful for aligning and scoring proteins that show less than 62% identity.
- Henikoff and Henikoff tested the performance of hierarchical multiple alignment of three serine proteases using BLOSUM45, BLOSUM62, BLOSUM80,PAM120, PAM160, and PAM250 matrices.
- All BLOSUM matrices performed better than PAM matrices;

To summarize, PAM and BLOSUM matrices can be compared as follows:

1. PAM matrices are constructed based on an evolutionary model—that is, from the estimation of mutation rates through constructing phylogenetic trees and inferring the ancestral sequence—but BLOSUM matrices are constructed based on direct observation of ungapped multiple alignment-driven sequence relationships.
2. Thus, PAM matrices are often used for reconstructing phylogenetic trees, whereas BLOSUM matrices are suitable for local sequence alignments.
3. PAM matrix construction involves global alignment of the full-length sequences consisting of both conserved and diverged regions, but BLOSUM matrix construction involves local sequence alignment of conserved sequence blocks.
4. Additionally, when Henikoff and Henikoff compared the two equivalent matrices PAM160 and BLOSUM62, they found that BLOSUM62 is less tolerant to hydrophilic-amino-acid substitution, but more tolerant to hydrophobic-amino-acid substitution than PAM160. Also, for rare amino acids, such as cysteine and tryptophan, BLOSUM62 is typically more tolerant to mismatches than PAM160.