(Aspekte der Thermodynamik in der Strukturbiologie)

# Einführung in die Bioinformatik

Wintersemester 2012/13
16:00-16:45 Hörsaal N100 B3

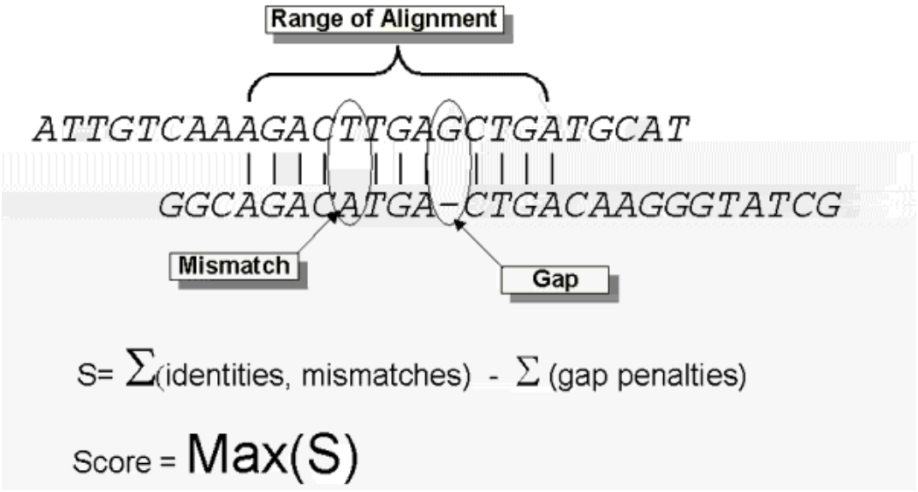
Peter Güntert

# Scoring of alignments

# Outline

- Substitution matrices
- Point Accepted Mutation (PAM)
- PAM substitution matrices
- log-odd substitution scores
- BLOSUM substitution and scoring matrices

# Calculation of an alignment score



## The alignment score is a sum of match, mismatch, gap creation, and gap extension scores

```
 Score = 18.1 bits (35),  Expect = 0.015, Method: Composition-based stats.
 Identities = 11/24 (45%), Positives = 12/24 (50%), Gaps = 2/24 (8%)

Query  12  VTALWGKVNVD--EVGGEALGRLL  33
           V  +WGKV  D    G E L RL
Sbjct  11  VLNVWGKVEADIPGHGQEVLIRLF  34

match      4  11 5   6    6 5 4 5    sum of matches: +60
              6 4                 4
mismatch  -1 1    0   -2 -2  -4  0    sum of mismatches: -13
          -2      0   -3   0
gap open              -11             sum of gap penalties: -12
gap extend             -1
                                      total raw score: 60 - 13 - 12 = 35
```

V matching V earns +4

T matching L earns -1

**These scores come from a "scoring matrix".**

## Where we're heading:

**to a PAM250 log odds scoring matrix that assigns scores and is forgiving of mismatches…**

**(such as +17 for W to W or -5 for W to T)**

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 2 | | | | | | | | | | | | | | | | | | |
| **R** | -2 | 6 | | | | | | | | | | | | | | | | | |
| **N** | 0 | 0 | 2 | | | | | | | | | | | | | | | | |
| **D** | 0 | -1 | 2 | 4 | | | | | | | | | | | | | | | |
| **C** | -2 | -4 | -4 | -5 | 12 | | | | | | | | | | | | | | |
| **Q** | 0 | 1 | 1 | 2 | -5 | 4 | | | | | | | | | | | | | |
| **E** | 0 | -1 | 1 | 3 | -5 | 2 | 4 | | | | | | | | | | | | |
| **G** | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | | | | | | | | | | | |
| **H** | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | | | | | | | | | | |
| **I** | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | | | | | | | | | |
| **L** | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | -2 | 6 | | | | | | | | |
| **K** | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | | | | | | | |
| **M** | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | | | | | | |
| **F** | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | | | | | |
| **P** | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | | | | |
| **S** | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | | | |
| **T** | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | | |
| **W** | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | |
| **Y** | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 |
| **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** |

…and to a whole series of scoring matrices such as PAM10 that are strict and do not tolerate mismatches (such as +13 for W to W or -19 for W to T)

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 7 | | | | | | | | | | | | | | | | | | | |
| **R** | -10 | 9 | | | | | | | | | | | | | | | | | | |
| **N** | -7 | -9 | 9 | | | | | | | | | | | | | | | | | |
| **D** | -6 | -17 | -1 | 8 | | | | | | | | | | | | | | | | |
| **C** | -10 | -11 | -17 | -21 | 10 | | | | | | | | | | | | | | | |
| **Q** | -7 | -4 | -7 | -6 | -20 | 9 | | | | | | | | | | | | | | |
| **E** | -5 | -15 | -5 | 0 | -20 | -1 | 8 | | | | | | | | | | | | | |
| **G** | -4 | -13 | -6 | -6 | -13 | -10 | -7 | 7 | | | | | | | | | | | | |
| **H** | -11 | -4 | -2 | -7 | -10 | -2 | -9 | -13 | 10 | | | | | | | | | | | |
| **I** | -8 | -8 | -8 | -11 | -9 | -11 | -8 | -17 | -13 | 9 | | | | | | | | | | |
| **L** | -9 | -12 | -10 | -19 | -21 | -8 | -13 | -14 | -9 | -4 | 7 | | | | | | | | | |
| **K** | -10 | -2 | -4 | -8 | -20 | -6 | -7 | -10 | -10 | -9 | -11 | 7 | | | | | | | | |
| **M** | -8 | -7 | -15 | -17 | -20 | -7 | -10 | -12 | -17 | -3 | -2 | -4 | 12 | | | | | | | |
| **F** | -12 | -12 | -12 | -21 | -19 | -19 | -20 | -12 | -9 | -5 | -5 | -20 | -7 | 9 | | | | | | |
| **P** | -4 | -7 | -9 | -12 | -11 | -6 | -9 | -10 | -7 | -12 | -10 | -10 | -11 | -13 | 8 | | | | | |
| **S** | -3 | -6 | -2 | -7 | -6 | -8 | -7 | -4 | -9 | -10 | -12 | -7 | -8 | -9 | -4 | 7 | | | | |
| **T** | -3 | -10 | -5 | -8 | -11 | -9 | -9 | -10 | -11 | -5 | -10 | -6 | -7 | -12 | -7 | -2 | 8 | | | |
| **W** | -20 | -5 | -11 | -21 | -22 | -19 | -23 | -21 | -10 | -20 | -9 | -18 | -19 | -7 | -20 | -8 | -19 | 13 | | |
| **Y** | -11 | -14 | -7 | -17 | -7 | -18 | -11 | -20 | -6 | -9 | -10 | -12 | -17 | -1 | -20 | -10 | -9 | -8 | 10 | |
| **V** | -5 | -11 | -12 | -11 | -9 | -10 | -10 | -9 | -9 | -1 | -5 | -13 | -4 | -12 | -9 | -10 | -6 | -22 | -10 | 8 |

## Normalized frequencies of amino acids

| | | | |
|---|---|---|---|
| Gly | 8.9% | Arg | 4.1% |
| Ala | 8.7% | Asn | 4.0% |
| Leu | 8.5% | Phe | 4.0% |
| Lys | 8.1% | Gln | 3.8% |
| Ser | 7.0% | Ile | 3.7% |
| Val | 6.5% | His | 3.4% |
| Thr | 5.8% | Cys | 3.3% |
| Pro | 5.1% | Tyr | 3.0% |
| Glu | 5.0% | Met | 1.5% |
| Asp | 4.7% | Trp | 1.0% |

Frequencies sum to 1. Blue = 6 codons. Red = 1 codon.

## The relative mutability of amino acids

| | | | |
|---|---|---|---|
| Asn | 134 | His | 66 |
| Ser | 120 | Arg | 65 |
| Asp | 106 | Lys | 56 |
| Glu | 102 | Pro | 56 |
| Ala | 100 | Gly | 49 |
| Thr | 97 | Tyr | 41 |
| Ile | 96 | Phe | 41 |
| Met | 94 | Leu | 40 |
| Gln | 93 | Cys | 20 |
| Val | 74 | Trp | 18 |

**Value for Ala defined as 100.**

## Substitution Matrix

- A substitution matrix contains values proportional to the probability that amino acid *i* mutates into amino acid *j* for all pairs of amino acids.
- Substitution matrices are constructed by assembling a large and diverse sample of verified pairwise alignments (or multiple sequence alignments) of amino acids.
- Substitution matrices should reflect the true probabilities of mutations occurring through a period of evolution.
- Two major types of substitution matrices: PAM and BLOSUM.

## Point Accepted Mutation (PAM)

Point accepted mutation (PAM), is a set of matrices used to score sequence alignments. The PAM matrices were introduced by Margaret Dayhoff in 1978 based on 1572 observed mutations in 71 families of closely related proteins.

Margaret O. Dayhoff (1925–1982)

## PAM units

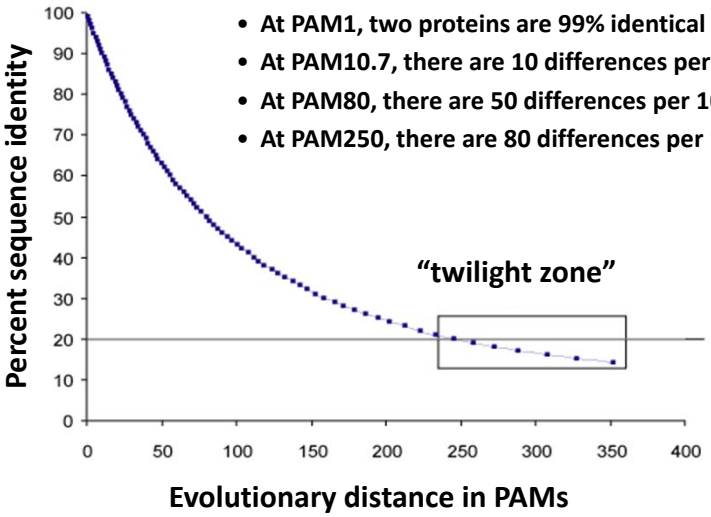
- PAM units measure the amount of evolutionary distance between two amino acid sequences.
- Two sequences $S_1$ and $S_2$ are said to be one PAM unit diverged if a series of accepted point mutations (and no insertions of deletions) has converted $S_1$ to $S_2$ with an average of one accepted point-mutation event per 100 amino acids.
- "Accepted" means a mutation that was incorporated into the protein and passed to its progeny. Therefore, either the mutation did not change the function of the protein or the change in the protein was beneficial to the organism.

## PAM units and sequence identity

Note that two sequences which are one PAM unit diverged do not necessarily differ in 1% of the positions, as often mistakenly thought, because a single position may undergo more than one mutation. The difference between the two notions grows as the number of units does:

| PAM | 0 | 30 | 80 | 110 | 200 | 250 |
|---|---|---|---|---|---|---|
| %identity | 100 | 75 | 50 | 40 | 25 | 20 |

## Relationship between PAM and sequence identity



## PAM substitution matrices

- PAM matrices are amino acid substitution matrices that encode the expected evolutionary change at the amino acid level.
- Each PAM matrix is designed to compare two sequences which are a specific number of PAM units apart.
- For example, the PAM120 score matrix is designed to compare between sequences that are 120 PAM units apart. The score it gives a pair of sequences is the (log of the) probabilities of such sequences evolving during 120 PAM units of evolution.

## PAM substitution matrices

- For a pair (Ai, Aj) of amino acids the (*i,j*) entry in the PAM *n* matrix reflects the frequency at which $A_i$ is expected to replace with $A_j$ in two sequences that are *n* PAM units diverged.
- These frequencies are estimated by gathering statistics on replaced amino acids.
- Collecting these statistics is difficult for distantly diverged sequences but easy for highly similar sequences, where only few insertions and deletions took place.

## PAM substitution matrices

- Therefore, in the first stage statistics were collected from aligned sequences that were believed to be approximately one PAM unit diverged and the PAM1 matrix could be computed based on this data, as follows:
- Let $M_{ij}$ denote the observed frequency (= estimated probability) of amino acid $A_i$ mutating into amino acid $A_j$ during one PAM unit of evolutionary change. $M$ is a 20 x 20 real matrix, with the values in each matrix column adding up to 1. There is a significant variance between the values in each column.

## PAM matrices: Point-accepted mutations

- PAM matrices are based on global alignments of closely related proteins (>85% amino acid identity).
- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. At an evolutionary interval of PAM1, one change has occurred over a length of 100 amino acids.
- Other PAM matrices are extrapolated from PAM1. For PAM250, 250 changes have occurred for two proteins over a length of 100 amino acids, i.e. PAM250 = $(PAM1)^{250}$

# Dayhoff's 34 protein superfamilies

| Protein | PAMs per 100 million years | Protein | PAMs per 100 million years |
|---|---|---|---|
| Ig kappa chain | 37 | thyroid stimulating hormone b | 7.4 |
| Kappa casein | 33 | parathyroid hormone | 7.3 |
| luteinizing hormone b | 30 | parvalbumin | 7.0 |
| lactalbumin | 27 | trypsin | 5.9 |
| complement component 3 | 27 | insulin | 4.4 |
| epidermal growth factor | 26 | calcitonin | 4.3 |
| proopiomelanocortin | 21 | arginine vasopressin | 3.6 |
| pancreatic ribonuclease | 21 | adenylate kinase 1 | 3.2 |
| haptoglobin alpha | 20 | triosephosphate isomerase 1 | 2.8 |
| serum albumin | 19 | vasoactive intestinal peptide | 2.6 |
| phospholipase A2, group IB | 19 | glyceraldehyde phosph. | |
| prolactin | 17 | dehydrogease | 2.2 |
| carbonic anhydrase C | 16 | cytochrome c | 2.2 |
| Hemoglobin a | 12 | collagen | 1.7 |
| Hemoglobin b | 12 | troponin C, skeletal muscle | 1.5 |
| apolipoprotein A-II | 10 | alpha crystallin B chain | 1.5 |
| lysozyme | 9.8 | glucagon | 1.2 |
| gastrin | 9.8 | glutamate dehydrogenase | 0.9 |
| myoglobin | 8.9 | histone H2B, member Q | 0.9 |
| nerve growth factor | 8.5 | ubiquitin | 0.0 |
| myelin basic protein | 7.4 | | |

# Dayhoff's 34 protein superfamilies

| Protein | PAMs per 100 million years | Protein | PAMs per 100 million years |
|---|---|---|---|
| Ig kappa chain | 37 | thyroid stimulating hormone b | 7.4 |
| **Kappa casein** | **33** | parathyroid hormone | 7.3 |
| | | parvalbumin | 7.0 |
| | | trypsin | 5.9 |
| | | insulin | 4.4 |
| | | calcitonin | 4.3 |
| phospholipase A2, group IB | 19 | | |
| prolactin | 17 | | |
| carbonic anhydrase C | 16 | | |
| Hemoglobin a | 12 | | |
| Hemoglobin b | 12 | | |
| apolipoprotein A-II | 10 | | |
| lysozyme | 9.8 | | |
| gastrin | 9.8 | | |
| myoglobin | 8.9 | | |
| nerve growth factor | 8.5 | | |
| myelin basic protein | 7.4 | **ubiquitin** | **0.0** |

**human versus mouse kappa casein**

```
 Score = 57.8 bits (138),  Expect = 3e-07
 Identities = 39/118 (33%), Positives = 61/118 (51%), Gaps = 2/118 (1%)

Query  1    MKSFLLVVNALALTLPFLAVEVQNQKPACHENDERPFYQKTAPYVPMYYVPNSYPYTGT  60
            M++F++V+N LALTLPFLA E+QN      E ++  + ++   Y P+  V N +  Y
Sbjct  2    MRNFIVVMNILALTLPFLAAEIQNPDSNCRGEKNDIVTDEQRVLYTPVRSVLN-FNQYEP  60

Query  61   NLYQRRPAI-AINNPYVPRTYYANPAVVRFHAQIPQRQYLPNSHPPTVVRLPNLHPSF  117
            N Y  RP++ A  +PY+         ++R  A I + Q +PN      V     +PSF
Sbjct  61   NYYHYRPSLPATASPYMYYPLVVRLLLLRSPAPISKWQSMPNFPQSAGVPYAIPNPSF  118
```

**human versus mouse ubiquitin**

```
 Score = 1316 bits (3407),  Expect = 0.0
 Identities = 681/685 (99%), Positives = 682/685 (99%), Gaps = 0/685 (0%)

Query  1    MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN  60
            MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN
Sbjct  1    MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN  60

Query  61   IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI  120
            IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI
Sbjct  61   IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI  120

Query  121  FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA  180
            FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA
Sbjct  121  FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA  180

Query  181  KIQDKEGIPSDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT  240
            KIQDKEGIP DQQRLIFAGKQLE GRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT
Sbjct  181  KIQDKEGIPPDQQRLIFAGKQLEGGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT  240

Query  241  ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR  300
            ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR
Sbjct  241  ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR  300
```

## Dayhoff's numbers of "accepted point mutations": what amino acid substitutions occur in proteins?

| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly |
|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | |
| R | 30 | | | | | | | |
| N | 109 | 17 | | | | | | |
| D | 154 | 0 | 532 | | | | | |
| C | 33 | 10 | 0 | 0 | | | | |
| Q | 93 | 120 | 50 | 76 | 0 | | | |
| E | 266 | 0 | 94 | 831 | 0 | 422 | | |
| G | 579 | 10 | 156 | 162 | 10 | 30 | 112 | |
| H | 21 | 103 | 226 | 43 | 10 | 243 | 23 | 10 |

(Some amino acids omitted for clarity.)

## Dayhoff's PAM1 mutation probability matrix

| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly | H His | I Ile |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 9867 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 | 6 |
| R | 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 10 | 3 |
| N | 4 | 1 | 9822 | 36 | 0 | 4 | 6 | 6 | 21 | 3 |
| D | 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 | 1 |
| C | 1 | 1 | 0 | 0 | 9973 | 0 | 0 | 0 | 1 | 1 |
| Q | 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 23 | 1 |
| E | 10 | 0 | 7 | 56 | 0 | 35 | 9865 | 4 | 2 | 3 |
| G | 21 | 1 | 12 | 11 | 1 | 3 | 7 | 9935 | 1 | 0 |
| H | 1 | 8 | 18 | 3 | 1 | 20 | 1 | 0 | 9912 | 0 |
| I | 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 | 9872 |

Each element of the matrix represents the probability (x $10^4$) that an original amino acid (top) will be replaced by another amino acid (side).

## Dayhoff's PAM0 mutation probability matrix: the rules for extremely slowly evolving proteins

| PAM0 | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly |
|---|---|---|---|---|---|---|---|---|
| A | 100% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| R | 0% | 100% | 0% | 0% | 0% | 0% | 0% | 0% |
| N | 0% | 0% | 100% | 0% | 0% | 0% | 0% | 0% |
| D | 0% | 0% | 0% | 100% | 0% | 0% | 0% | 0% |
| C | 0% | 0% | 0% | 0% | 100% | 0% | 0% | 0% |
| Q | 0% | 0% | 0% | 0% | 0% | 100% | 0% | 0% |
| E | 0% | 0% | 0% | 0% | 0% | 0% | 100% | 0% |
| G | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 100% |

= unit matrix

Top: original amino acid  Side: replacement amino acid

## Dayhoff's PAM2000 mutation probability matrix: the rules for very distantly related proteins

| PAM∞ | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly |
|---|---|---|---|---|---|---|---|---|
| A | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% |
| R | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% |
| N | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% |
| D | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% |
| C | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% |
| Q | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% |
| E | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% |
| G | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% |

Mutation probability = amino acid frequency

## PAM250 mutation probability matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 13 | 6 | 9 | 9 | 5 | 8 | 9 | 12 | 6 | 8 | 6 | 7 | 7 | 4 | 11 | 11 | 11 | 2 | 4 | 9 |
| R | 3 | 17 | 4 | 3 | 2 | 5 | 3 | 2 | 6 | 3 | 2 | 9 | 4 | 1 | 4 | 4 | 3 | 7 | 2 | 2 |
| N | 4 | 4 | 6 | 7 | 2 | 5 | 6 | 4 | 6 | 3 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | 2 | 3 | 3 |
| D | 5 | 4 | 8 | 11 | 1 | 7 | 10 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| C | 2 | 1 | 1 | 1 | 52 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 2 |
| Q | 3 | 5 | 5 | 6 | 1 | 10 | 7 | 3 | 7 | 2 | 3 | 5 | 3 | 1 | 4 | 3 | 3 | 1 | 2 | 3 |
| E | 5 | 4 | 7 | 11 | 1 | 9 | 12 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| G | 12 | 5 | 10 | 10 | 4 | 7 | 9 | 27 | 5 | 5 | 4 | 6 | 5 | 3 | 8 | 11 | 9 | 2 | 3 | 7 |
| H | 2 | 5 | 5 | 4 | 2 | 7 | 4 | 2 | 15 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 2 |
| I | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 10 | 6 | 2 | 6 | 5 | 2 | 3 | 4 | 1 | 3 | 9 |
| L | 6 | 4 | 4 | 3 | 2 | 6 | 4 | 3 | 5 | 15 | 34 | 4 | 20 | 13 | 5 | 4 | 6 | 6 | 7 | 13 |
| K | 6 | 18 | 10 | 8 | 2 | 10 | 8 | 5 | 8 | 5 | 4 | 24 | 9 | 2 | 6 | 8 | 8 | 4 | 3 | 5 |
| M | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 6 | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| F | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 5 | 6 | 1 | 4 | 32 | 1 | 2 | 2 | 4 | 20 | 3 |
| P | 7 | 5 | 5 | 4 | 3 | 5 | 4 | 5 | 5 | 3 | 3 | 4 | 3 | 2 | 20 | 6 | 5 | 1 | 2 | 4 |
| S | 9 | 6 | 8 | 7 | 7 | 6 | 7 | 9 | 6 | 5 | 4 | 7 | 5 | 3 | 9 | 10 | 9 | 4 | 4 | 6 |
| T | 8 | 5 | 6 | 6 | 4 | 5 | 5 | 6 | 4 | 6 | 4 | 6 | 5 | 3 | 6 | 8 | 11 | 2 | 3 | 6 |
| W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 55 | 1 | 0 |
| Y | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 15 | 1 | 2 | 2 | 3 | 31 | 2 |
| V | 7 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 15 | 10 | 4 | 10 | 5 | 5 | 5 | 7 | 2 | 4 | 17 |

Probabilities in %

## Dayhoff's approach to assigning scores for any two aligned amino acid residues: log-odd scores

Dayhoff et al. defined the score $S_{ij}$ of two aligned residues *i*, *j* as 10 times the (base 10) logarithm of how likely it is to observe these two residues (based on the empirical observation of how often they are aligned in nature) divided by the background probability of finding these amino acids by chance. This provides a score for each pair of residues.

$$S_{ij} = 10 \times \log_{10}\left(\frac{M_{ij}}{p_i}\right)$$

## PAM250 log odds scoring matrix

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 2 | | | | | | | | | | | | | | | | | | |
| **R** | -2 | 6 | | | | | | | | | | | | | | | | | |
| **N** | 0 | 0 | 2 | | | | | | | | | | | | | | | | |
| **D** | 0 | -1 | 2 | 4 | | | | | | | | | | | | | | | |
| **C** | -2 | -4 | -4 | -5 | 12 | | | | | | | | | | | | | | |
| **Q** | 0 | 1 | 1 | 2 | -5 | 4 | | | | | | | | | | | | | |
| **E** | 0 | -1 | 1 | 3 | -5 | 2 | 4 | | | | | | | | | | | | |
| **G** | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | | | | | | | | | | | |
| **H** | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | | | | | | | | | | |
| **I** | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | | | | | | | | | |
| **L** | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | -2 | 6 | | | | | | | | |
| **K** | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | | | | | | | |
| **M** | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | | | | | | |
| **F** | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | | | | | |
| **P** | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | | | | |
| **S** | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | | | |
| **T** | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | | |
| **W** | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | |
| **Y** | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 |
| **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** |

## Why do we go from a mutation probability matrix to a log odds matrix?

- We want a scoring matrix so that when we do a pairwise alignment (or a BLAST search) we know what score to assign to two aligned amino acid residues.
- Logarithms are easier to use for a scoring system. They allow us to sum the scores of aligned residues (rather than multiplying the probabilities for independent mutations).

## How do we go from a mutation probability matrix to a log odds matrix?

The cells in a log odds matrix consist of an "odds ratio":

$$\frac{\text{the probability that an alignment is authentic}}{\text{the probability that the alignment was random}}$$

The score *S* for an alignment of residues *i*, *j* is given by:

$$S_{ij} = 10 \times \log_{10}\left(\frac{M_{ij}}{p_i}\right)$$

Example: Tryptophan, $S$(Trp,Trp) = 10 $\log_{10}$ (0.55/0.010) = 17.4

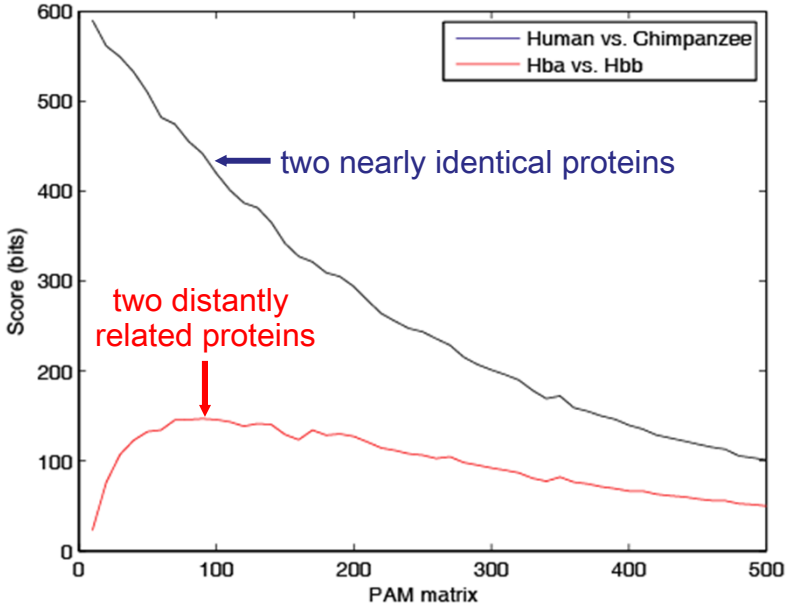
## PAM250 log odds scoring matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 2 | | | | | | | | | | | | | | | | | | | |
| **R** | -2 | 6 | | | | | | | | | | | | | | | | | | |
| **N** | 0 | 0 | 2 | | | | | | | | | | | | | | | | | |
| **D** | 0 | -1 | 2 | 4 | | | | | | | | | | | | | | | | |
| **C** | -2 | -4 | -4 | -5 | 12 | | | | | | | | | | | | | | | |
| **Q** | 0 | 1 | 1 | 2 | -5 | 4 | | | | | | | | | | | | | | |
| **E** | 0 | -1 | 1 | 3 | -5 | 2 | 4 | | | | | | | | | | | | | |
| **G** | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | | | | | | | | | | | | |
| **H** | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | | | | | | | | | | | |
| **I** | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | | | | | | | | | | |
| **L** | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | -2 | 6 | | | | | | | | | |
| **K** | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | | | | | | | | |
| **M** | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | | | | | | | |
| **F** | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | | | | | | |
| **P** | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | | | | | |
| **S** | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | | | | |
| **T** | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | | | |
| **W** | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | | |
| **Y** | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | |
| **V** | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |

## What do the numbers mean in a log odds matrix?

- $S_{ij} = 10 \times \log_{10}\left(\frac{M_{ij}}{p_i}\right) \quad \Rightarrow \quad \frac{M_{ij}}{p_i} = 10^{S_{ij}/10}$
- A score of +2 indicates that the amino acid replacement occurs $10^{+2/10} = 1.6$ times more often than expected by chance.
- A score of 0 is neutral.
- A score of –10 indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is $10^{-10/10} = 0.1$ times as frequent as the chance alignment of these amino acids.

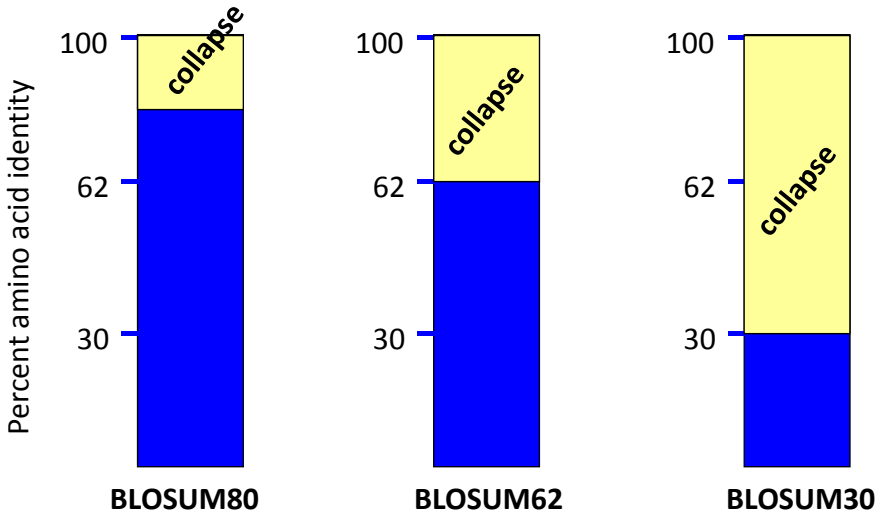
## Alignment scores using a series of PAM matrices



## Alternative to PAM: BLOSUM scoring matrices

- BLOSUM = Blocks Substitution Matrix
- Introduced by S. und J. G. Henikoff (1992)
- Based on the BLOCKS database consisting of over 500 groups of local multiple alignments (blocks) of distantly related proteins.
- $S_{ij} = 2 \times \log_2 \left(\frac{M_{ij}}{p_i}\right)$
- BLOSUM*n* matrices: Sequences with identity > *n*% are weighted (grouped) as one sequence. → BLOSUM*n* matrix is useful for scoring proteins with less than *n*% identity.

## BLOSUM Matrices



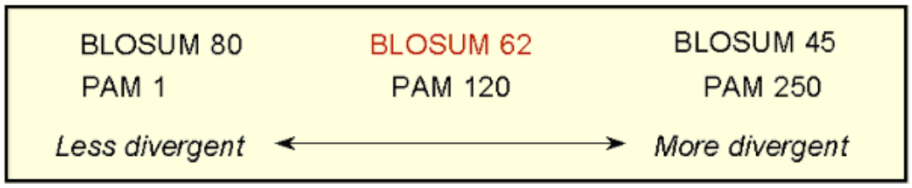
# BLOSUM Matrices

- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
- The BLOCKS database contains thousands of groups of multiple sequence alignments.
- BLOSUM performs better than PAM especially for weakly scoring alignments.
- BLOSUM62 is the default matrix in BLAST 2.0 at NCBI.
- Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships.
- A search for distant relatives may be more sensitive with a different matrix.

# Blosum62 scoring matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 4 | | | | | | | | | | | | | | | | | | | |
| **R** | -1 | 5 | | | | | | | | | | | | | | | | | | |
| **N** | -2 | 0 | 6 | | | | | | | | | | | | | | | | | |
| **D** | -2 | -2 | 1 | 6 | | | | | | | | | | | | | | | | |
| **C** | 0 | -3 | -3 | -3 | 9 | | | | | | | | | | | | | | | |
| **Q** | -1 | 1 | 0 | 0 | -3 | 5 | | | | | | | | | | | | | | |
| **E** | -1 | 0 | 0 | 2 | -4 | 2 | 5 | | | | | | | | | | | | | |
| **G** | 0 | -2 | 0 | -1 | -3 | -2 | -2 | 6 | | | | | | | | | | | | |
| **H** | -2 | 0 | 1 | -1 | -3 | 0 | 0 | -2 | 8 | | | | | | | | | | | |
| **I** | -1 | -3 | -3 | -3 | -1 | -3 | -3 | -4 | -3 | 4 | | | | | | | | | | |
| **L** | -1 | -2 | -3 | -4 | -1 | -2 | -3 | -4 | -3 | 2 | 4 | | | | | | | | | |
| **K** | -1 | 2 | 0 | -1 | -1 | 1 | 1 | -2 | -1 | -3 | -2 | 5 | | | | | | | | |
| **M** | -1 | -2 | -2 | -3 | -1 | 0 | -2 | -3 | -2 | 1 | 2 | -1 | 5 | | | | | | | |
| **F** | -2 | -3 | -3 | -3 | -2 | -3 | -3 | -3 | -1 | 0 | 0 | -3 | 0 | 6 | | | | | | |
| **P** | -1 | -2 | -2 | -1 | -3 | -1 | -1 | -2 | -2 | -3 | -3 | -1 | -2 | -4 | 7 | | | | | |
| **S** | 1 | -1 | 1 | 0 | -1 | 0 | 0 | 0 | -1 | -2 | -2 | 0 | -1 | -2 | -1 | 4 | | | | |
| **T** | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 1 | 5 | | | |
| **W** | -3 | -3 | -4 | -4 | -2 | -2 | -3 | -2 | -2 | -3 | -2 | -3 | -1 | 1 | -4 | -3 | -2 | 11 | | |
| **Y** | -2 | -2 | -2 | -3 | -2 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | -1 | 3 | -3 | -2 | -2 | 2 | 7 | |
| **V** | 0 | -3 | -3 | -3 | -1 | -2 | -2 | -3 | -3 | 3 | 1 | -2 | 1 | -1 | -2 | -2 | 0 | -3 | -1 | 4 |

## Choice of scoring matrix should be adapted to expected sequence divergence




More conserved

Rat *versus* mouse globin

Less conserved

Rat *versus* Bacterial globin

## Unterlagen zur Vorlesung

http://www.bpc.uni-frankfurt.de/guentert/wiki/index.php/Teaching