# CSE 549: Computational Biology

Substitution Matrices

Stony Brook University

# How should we score alignments

So far, we’ve looked at “arbitrary” schemes for scoring mutations. How can we assign scores in a more meaningful way?

Are these scores

|   | A | C | G | T |
|---|---|---|---|---|
| A | 5 | -5 | -3 | -5 |
| C | -5 | 5 | -5 | -3 |
| G | -3 | -5 | 5 | -5 |
| T | -5 | -3 | -5 | 5 |

better than these scores?

|   | A | C | G | T |
|---|---|---|---|---|
| A | 4 | -1 | -1 | -1 |
| C | -1 | 4 | -1 | -1 |
| G | -1 | -1 | 4 | -1 |
| T | -1 | -1 | -1 | 4 |

# How should we score alignments

So far, we've looked at "arbitrary" schemes for scoring mutations. How can we assign scores in a more meaningful way?

Are these scores

|  | A | C | G | T |
|---|---|---|---|---|
| A | 5 | -5 | -3 | -5 |
| [illegible] | [illegible] | 5 | [illegible] | [illegible] |
| G | -3 | -5 | 5 | -5 |
| T | -5 | -3 | -5 | 5 |

better than these scores?

|  | A | C | G | T |
|---|---|---|---|---|
| A | 4 | -1 | -1 | -1 |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| G | -1 | -1 | 4 | -1 |
| T | -1 | -1 | -1 | 4 |

**One option — "learn" the substitution / mutation rates from real data**

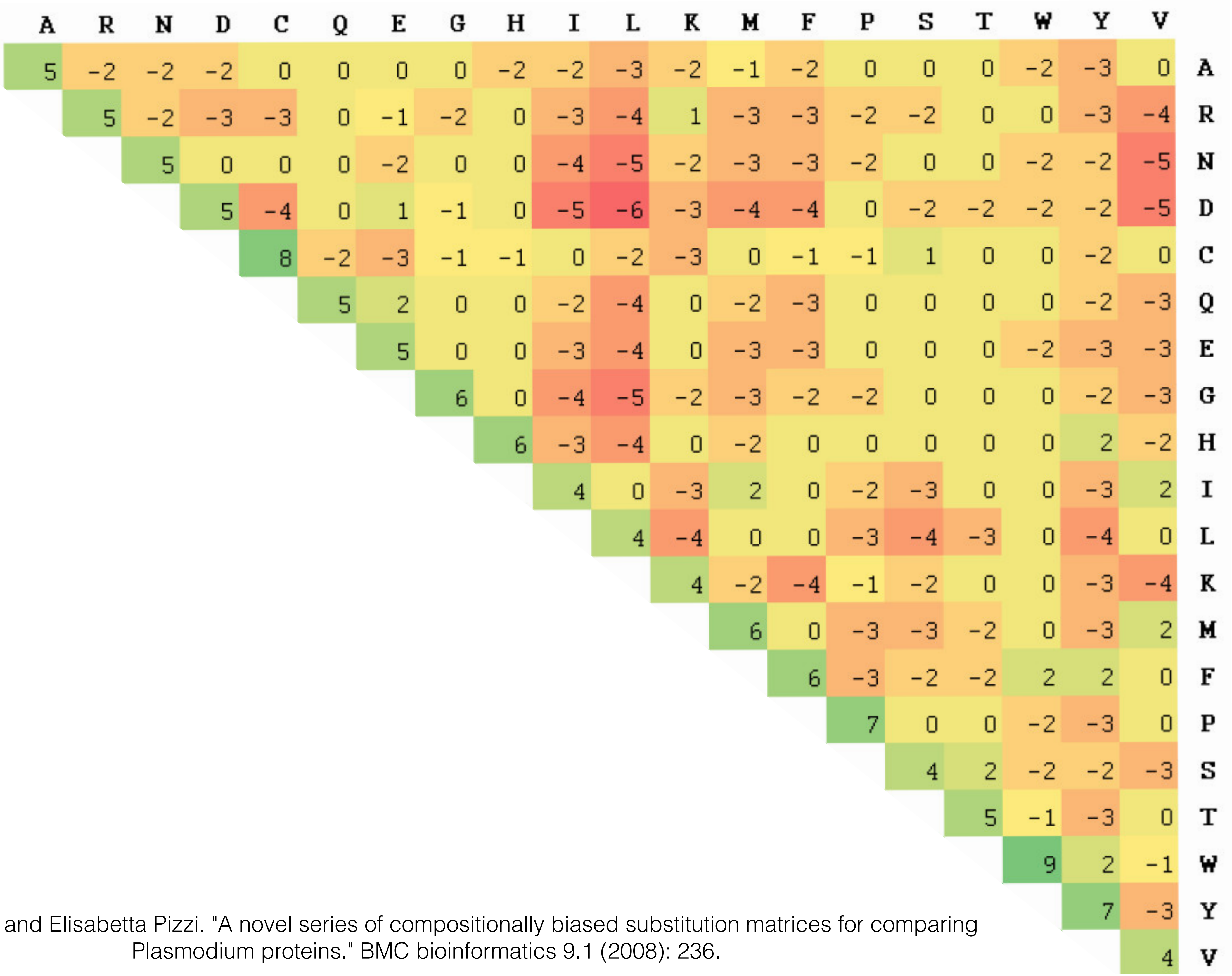
# How should we score alignments

**Main Idea:** **Assume** we can obtain (through a potentially burdensome process) a collection of high quality, high confidence sequence alignments.

We have a collection of sequences which, presumably, originated from the same ancestor — differences are mutations due to divergence.

**Learn** the frequency of different mutations from these alignments, and use the frequencies to derive our scoring function.

# BLOSUM62 matrix



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

Brick, Kevin, and Elisabetta Pizzi. "A novel series of compositionally biased substitution matrices for comparing Plasmodium proteins." BMC bioinformatics 9.1 (2008): 236.

# Probabilities to Scores

Assuming we have a reasonable process by which to compute frequencies, how can we use this to obtain a score?

# Probabilities to Scores

Assuming we have a reasonable process by which to compute frequencies, how can we use this to obtain a score?

Hypothesis we wish to test; two amino acids are correlated because they are homologous.

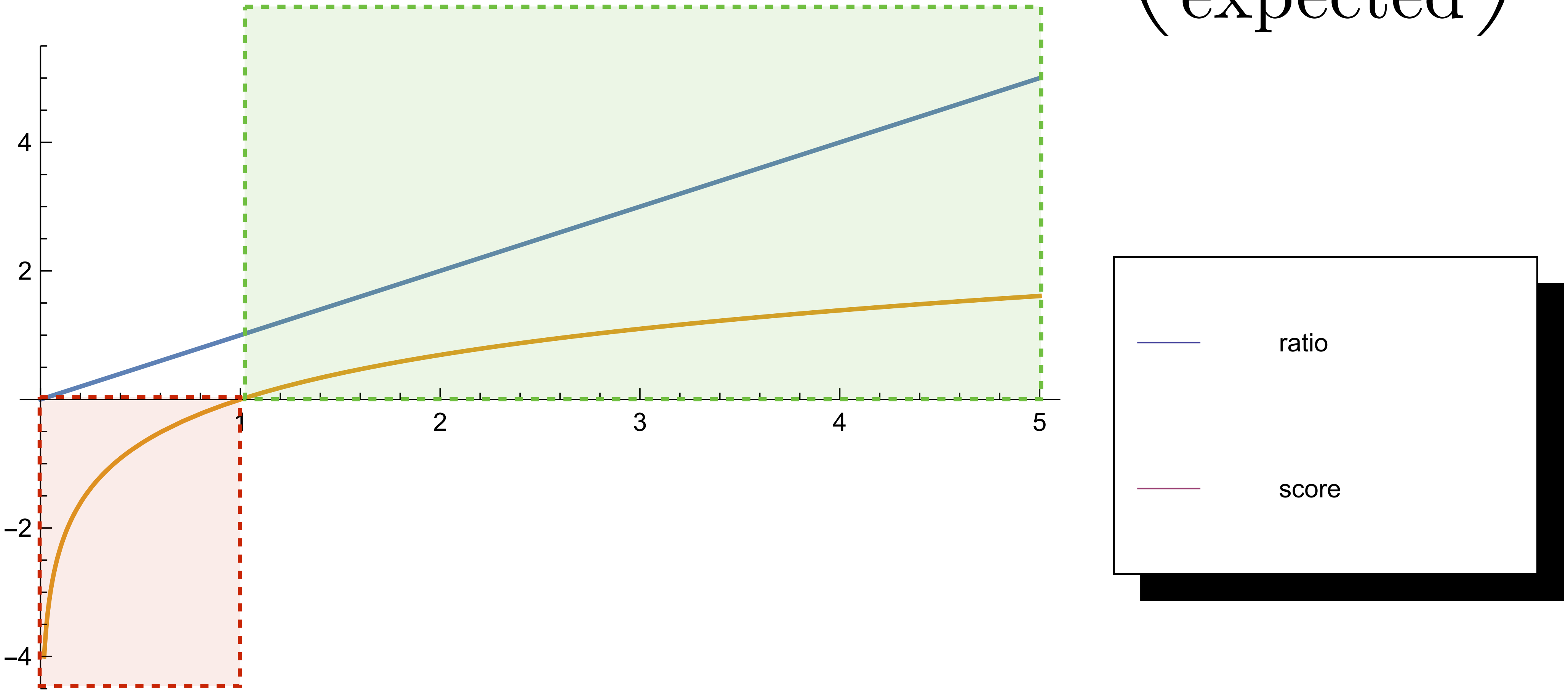
$$\text{score} = \text{log odds ratio} = s_{AB} \propto \log\left(\frac{\text{observed}}{\text{expected}}\right)$$

Null hypothesis; two amino acids occur independently (and are uncorrelated and unrelated).

Eddy, Sean R. "Where did the BLOSUM62 alignment score matrix come from?." Nature biotechnology 22.8 (2004): 1035-1036.

# Probabilities to Scores

$$\text{score} = \text{log odds ratio} = s_{AB} \propto \log\left(\frac{\text{observed}}{\text{expected}}\right)$$



Positive scores mean we find “conservative substitutions”

Negative scores mean we find “nonconservative substitutions”

Eddy, Sean R. "Where did the BLOSUM62 alignment score matrix come from?." Nature biotechnology 22.8 (2004): 1035-1036.

# BLOSUM matrix

Introduced by Henikoff & Henikoff in 1992

Start with the BLOCKS database (H&H ’91)

1. Look for conserved (gapless, <=62% identical) regions in alignments.

2. Count all pairs of amino acids in each column of the alignments.

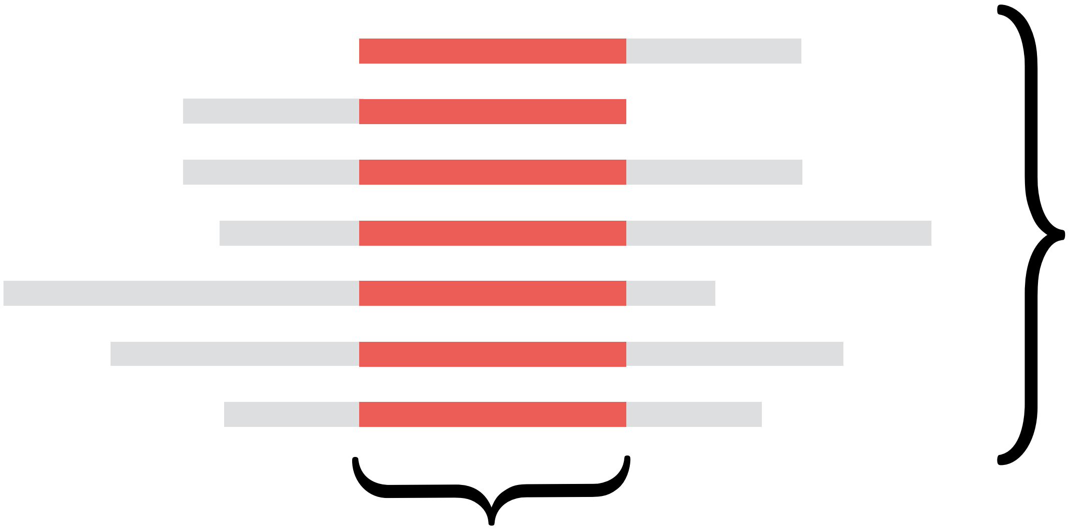
3. Use amino acid pair frequencies to derive “score” for a mutation/replacement

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# BLOSUM matrix

Start with the BLOCKS database (H&H ’91)

1. Look for conserved (gapless) regions in alignments.



collection of related proteins

conserved “block” within these proteins

sequences too similar are “clustered” & represented by either a single sequence, or a weighted combination of the cluster members

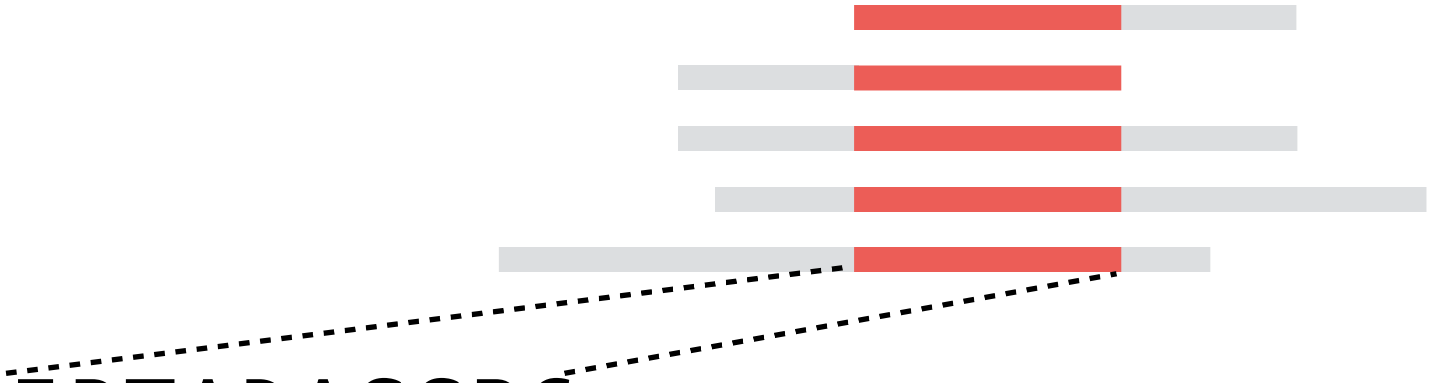
BLOSUM r: the matrix built from blocks with no more than r% of similarity – e.g., BLOSUM62 is the matrix built using sequences with no more than 62% similarity.*

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# BLOSUM matrix

Start with the BLOCKS database (H&H ’91)

2. Count all pairs of amino acids in each column of the alignments.



```
FPTADAGGRS
FVTADALGRS
FPTPDAGLRN
FVTAEAGIRQ
FPTAEAGGRS
```

$$c_{AB}^{(i)} = \begin{cases} \binom{c_A^{(i)}}{2} & \text{if } A = B \\ c_A^{(i)} \times c_B^{(i)} & \text{otherwise} \end{cases}$$

$$c_A^{(i)} = \text{num. of occurrences of } A \text{ in column } i$$

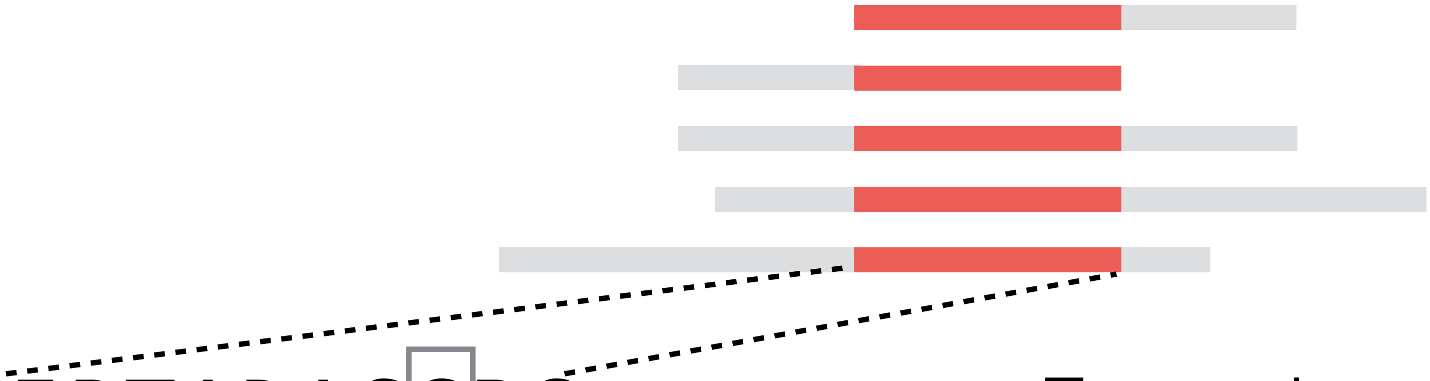
What is the intuition behind this expression?

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# BLOSUM matrix

Start with the BLOCKS database (H&H ’91)

2. Count all pairs of amino acids in each column of the alignments.



```
FPTADAGGRS
FVTADALGRS
FPTPDAGLRN
FVTAEAGLRQ
FPTAEAGGRS
```

Example:

$$c_{GG}^{(i)} = \binom{3}{2} = 3$$

$$c_{GL}^{(i)} = 3 \times 2$$

$$c_{LL}^{(i)} = \binom{2}{2} = 1$$

In this column, there are 3 ways to pair G with G, 6 potential ways to pair G with L and 1 potential way to pair L with L.

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# Computing Scores

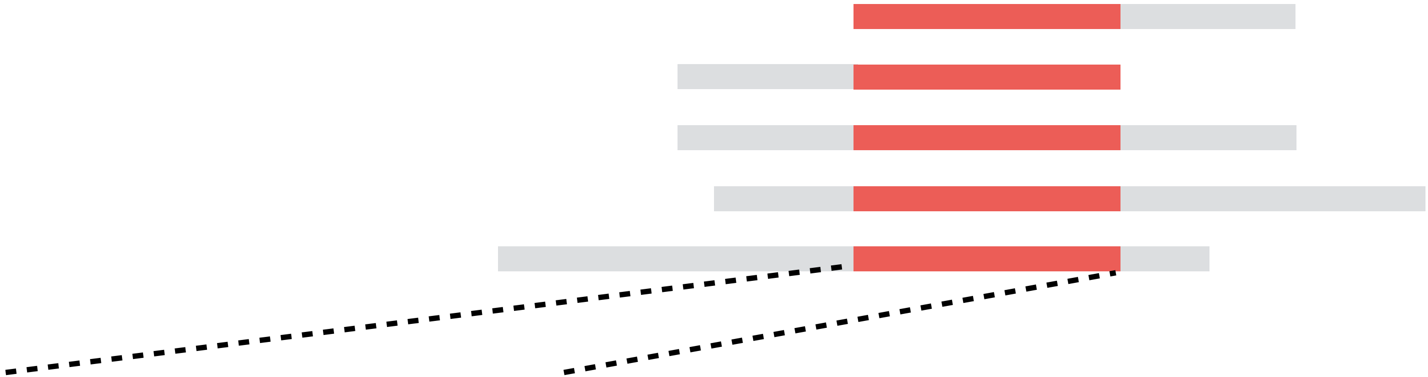
3. Use amino acid pair frequencies to derive "score" for a mutation/replacement

Total # of potential align. between A & B: $c_{AB} = \sum_i c_{AB}^{(i)}$

Total number of pairwise char. alignments: $T = \sum_{A \geq B} c_{AB}$

Normalized frequency of aligning A & B: $q_{AB} = \frac{c_{AB}}{T}$

# BLOSUM matrix



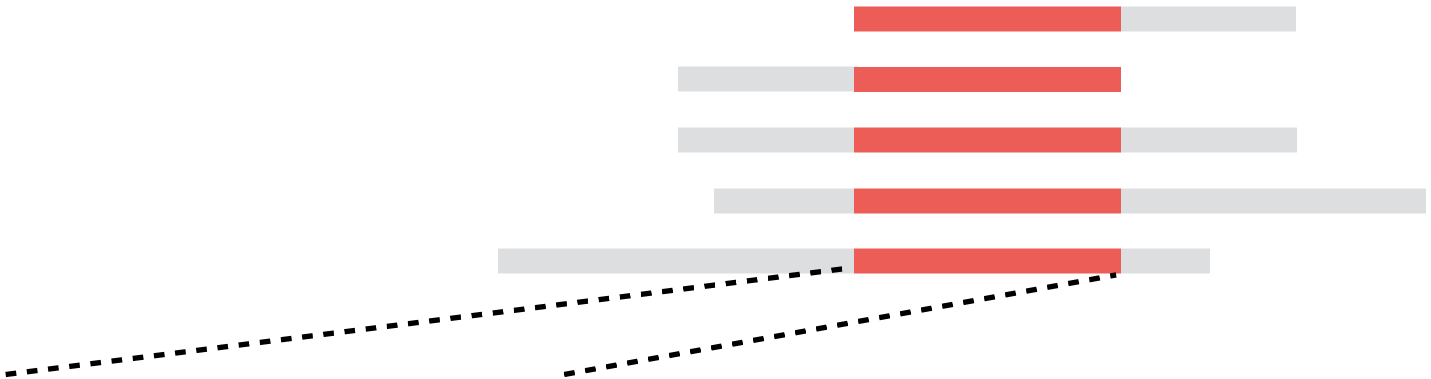
```
FPTADAGGRS
FVTADALGRS
FPTPDAGLRN
FVTAEAGLRQ
FPTAEAGGRS
```

In our example, we get

$$q_{GL} = \frac{0+0+0+0+0+0+4+6+0+0}{10\frac{(5)(4)}{2}} = \frac{10}{100}$$

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# BLOSUM matrix



FPTADAGGRS

FVTADALGRS

FPTPDAGLRN

FVTAEAGLRQ

FPTAEAGGRS

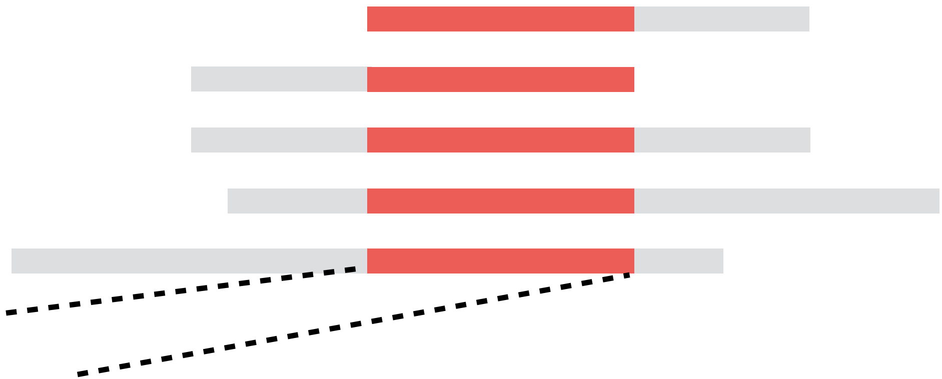
In our example, we get

$$q_{GL} = \frac{0+0+0+0+0+0+4+6+0+0}{\boxed{10\frac{(5)(4)}{2}}} = \frac{10}{100}$$

why does this denominator work?

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# BLOSUM matrix



```
FPTADAGGRS
FVTADALGRS
FPTPDAGLRN
FVTAEAGLRQ
FPTAEAGGRS
```

$c_{VP} = 2*3 = 6$

$c_{PP} = 3 \text{ choose } 2 = 3$

$c_{VV} = 2 \text{ choose } 2 = 1$

So $c_{VP} + c_{PP} + c_{VV} = 10 = 5 \text{ choose } 2$

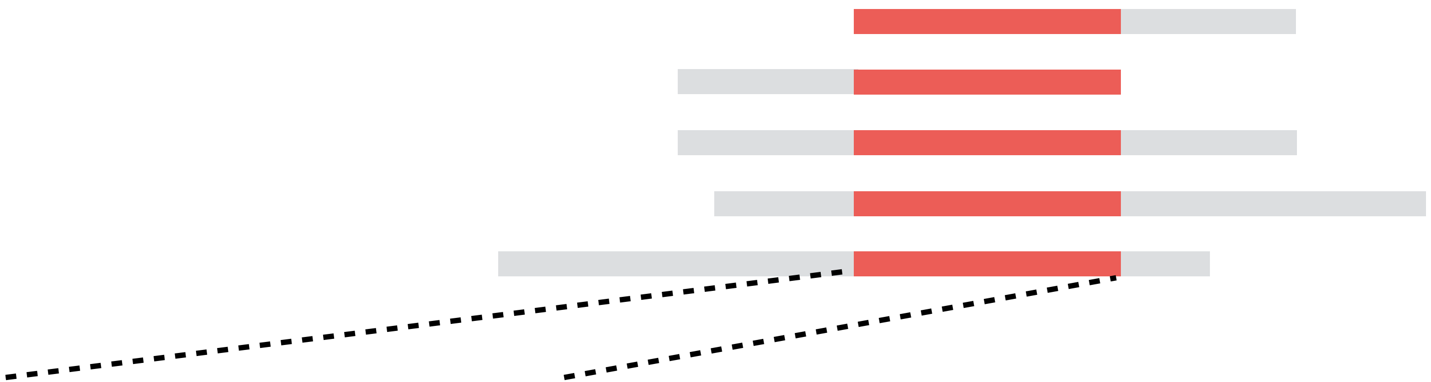
In our example, we get

$$q_{GL} = \frac{0+0+0+0+0+0+4+6+0+0}{\boxed{10\frac{(5)(4)}{2}}} = \frac{10}{100}$$

why does this denominator work?

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# BLOSUM matrix



```
FPTADAGGRS
FVTADALGRS
FPTPDAGLRN
FVTAEAGLRQ
FPTAEAGGRS
```

In our example, we get

$$q_{GL} = \frac{0+0+0+0+0+0+4+6+0+0}{\boxed{10\frac{(5)(4)}{2}}} = \frac{10}{100}$$

total column sum is always # rows choose 2

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# Computing Scores

3. Use amino acid pair frequencies to derive “score” for a mutation/replacement

Probability of occurrence of amino acid A in any {A,B} pair:

$$p_A = q_{AA} + \sum_{A \neq B} q_{AB}$$

Expected likelihood of each {A,B} pair, assuming independence:

$$e_{AB} = \begin{cases} (p_A)(p_B) = (p_A)^2 & \text{if } A = B \\ (p_A)(p_B) + (p_B)(p_A) = 2(p_A)(p_B) & \text{otherwise} \end{cases}$$

# Computing Scores

Recall the original idea (likelihood → scores)

$$\text{score} = \text{log odds ratio} = s_{AB} \propto \log\left(\frac{\text{observed}}{\text{expected}}\right)$$

score = log odds ratio = $s_{AB} = \text{Round}\left(\left(\frac{1}{\lambda}\right)\log_2\left(\frac{q_{AB}}{e_{AB}}\right)\right)$

Scaling factor used to produce scores that can be rounded to integers; set to 0.5 in H&H ‘92.

Henikoff, S.; Henikoff, J.G. (1992). "Amino Acid Substitution Matrices from Protein Blocks". PNAS 89 (22): 10915–10919.

# Scores are data-dependent

## distribution of amino acids across columns matters

```
GG
GA
WG
WA
NG
GA
GA
```

$p_G = 0.5$

$e_{GG} = 0.25$

$q_{GG} = 0.214$

$s_{GG} = \mathrm{Round}[(2)\log_2(0.214 / 0.25)]$
$= \mathrm{Round}[(2)(-0.22)] = 0$

```
GW
GA
GW
GA
GN
GA
GA
```

$p_G = 0.5$

$e_{GG} = 0.25$

$q_{GG} = 0.5$

$s_{GG} = \mathrm{Round}[(2)\log_2(0.5 / 0.25)]$
$= \mathrm{Round}[(2)(1)] = 2$

adopted from: http://www.bioinfo.rpi.edu/bystrc/courses/biol4540/lecture5.pdf

# Scores are data-dependent

{G,W} observed a lot

```
GG
GA
WG
AW
NG
GA
GA
```

$p_G = 0.5$ $p_W = 0.143$

$e_{GW} = 0.143$

$q_{GW} = 0.167$

$s_{GW} = \text{Round}[(2)\log_2(0.167 / 0.143)]$
$= \text{Round}[(2)(0.224)] = 0$

{G,W} observed rarely

```
GW
GA
GW
GA
GN
GA
AG
```

$p_G = 0.5$ $p_W = 0.143$

$e_{GW} = 0.143$

$q_{GW} = 0.048$

$s_{GW} = \text{Round}[(2)\log_2(0.048 / 0.143)]$
$= \text{Round}[(2)(-1.575)] = -3$

adopted from: http://www.bioinfo.rpi.edu/bystrc/courses/biol4540/lecture5.pdf

# Example

```
FPTADAGGRS
FVTADALGRS
FPTPDAGLRN
FVTAEAGLRQ
FPTAEAGGRS
```

↓

$$c_{AB} = \sum_i c_{AB}^{(i)}$$

→

## Matrix of $c_{AB}$ values

| | A | D | E | F | G | L | N | P | Q | R | S | T | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 16 | | | | | | | | | | | | |
| D | | 3 | | | | | | | | | | | |
| E | | 6 | 1 | | | | | | | | | | |
| F | | | | 10 | | | | | | | | | |
| G | | | | | 9 | | | | | | | | |
| L | | | | | 10 | 1 | | | | | | | |
| N | | | | | | | 0 | | | | | | |
| P | 4 | | | | | | | 3 | | | | | |
| Q | | | | | | | 1 | | 0 | | | | |
| R | | | | | | | | | | 10 | | | |
| S | | | | | | | 3 | | 3 | | 3 | | |
| T | | | | | | | | | | | | 10 | |
| V | | | | | | | | 6 | | | | | 1 |

# Example

$c_{AB}$

$$q_{AB} = \frac{c_{AB}}{T}$$

$$p_A = q_{AA} + \sum_{A \neq B} \frac{q_{AB}}{2}$$

## Matrix of $q_{AB}$ values

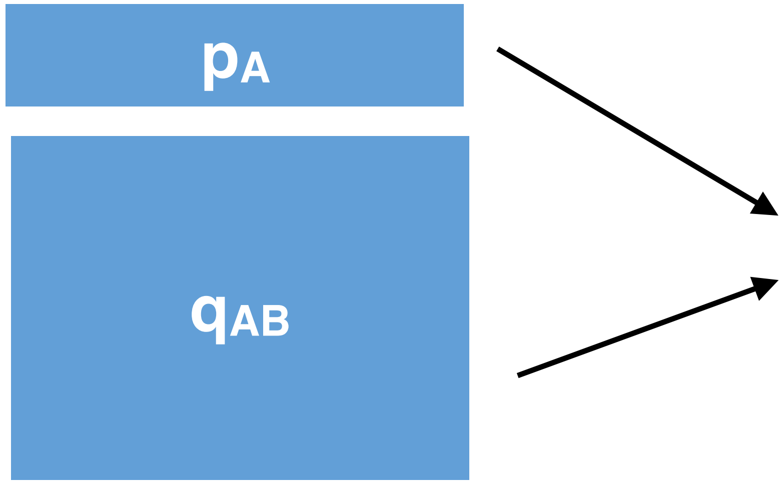
| | A | D | E | F | G | L | N | P | Q | R | S | T | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 0.16 | | | | | | | | | | | | |
| **D** | | 0.03 | | | | | | | | | | | |
| **E** | | 0.06 | 0.01 | | | | | | | | | | |
| **F** | | | | 0.1 | | | | | | | | | |
| **G** | | | | | 0.09 | | | | | | | | |
| **L** | | | | | 0.1 | 0.01 | | | | | | | |
| **N** | | | | | | | 0 | | | | | | |
| **P** | 0.04 | | | | | | | 0.03 | | | | | |
| **Q** | | | | | | | 0.01 | | 0 | | | | |
| **R** | | | | | | | | | | 0.1 | | | |
| **S** | | | | | | | 0.03 | | 0.03 | | 0.03 | | |
| **T** | | | | | | | | | | | | 0.1 | |
| **V** | | | | | | | | 0.06 | | | | | 0.01 |

| $P_A$ | $P_D$ | $P_E$ | $P_F$ | $P_G$ | $P_L$ | $P_N$ | $P_P$ | $P_Q$ | $P_R$ | $P_S$ | $P_T$ | $P_V$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.18 | 0.06 | 0.04 | 0.1 | 0.14 | 0.06 | 0.02 | 0.08 | 0.02 | 0.1 | 0.06 | 0.1 | 0.04 |

# Example

## Matrix of $e_{AB}$ values

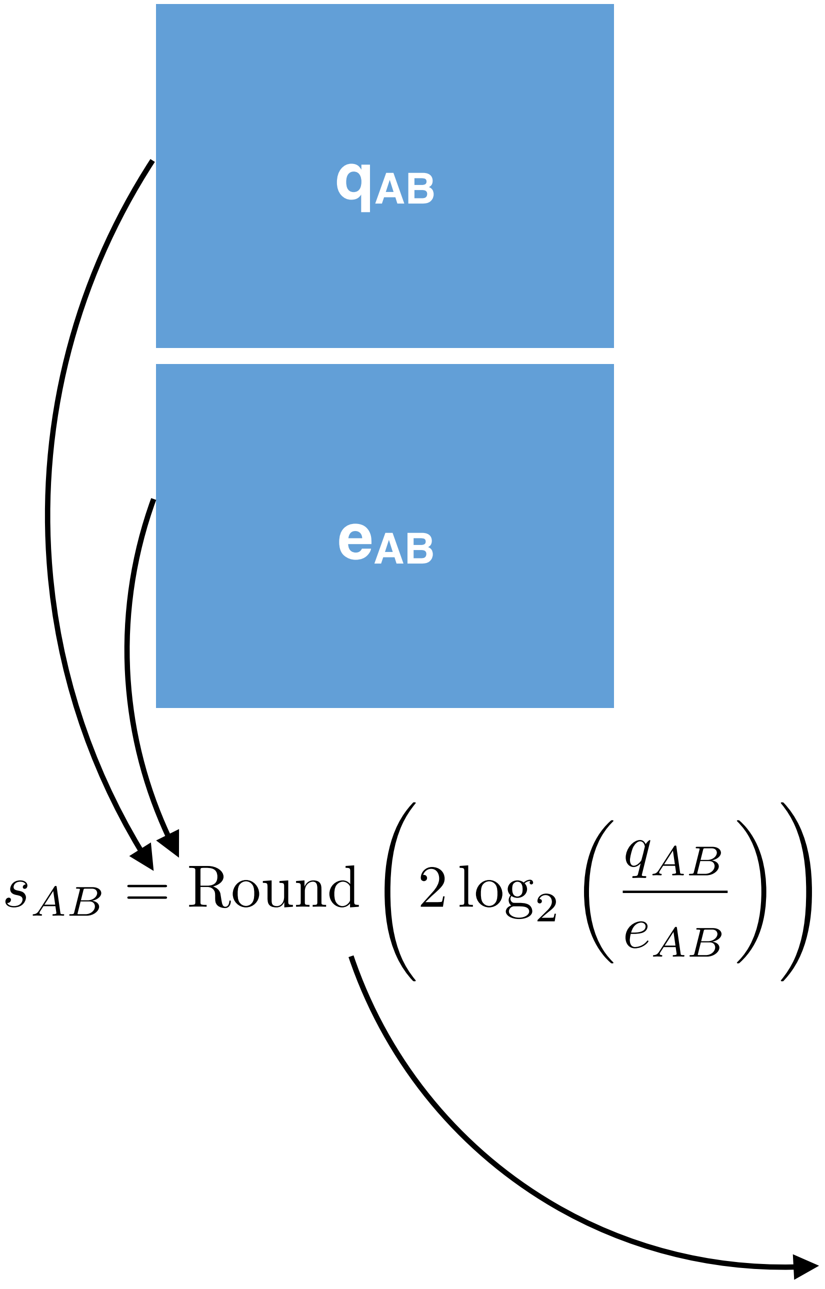
| | A | D | E | F | G | L | N | P | Q | R | S | T | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 0.0324 | | | | | | | | | | | | |
| **D** | 0.0216 | 0.0036 | | | | | | | | | | | |
| **E** | 0.0144 | 0.0048 | 0.0016 | | | | | | | | | | |
| **F** | 0.0360 | 0.0120 | 0.0080 | 0.0100 | | | | | | | | | |
| **G** | 0.0504 | 0.0168 | 0.0112 | 0.0280 | 0.0196 | | | | | | | | |
| **L** | 0.0216 | 0.0072 | 0.0048 | 0.0120 | 0.0168 | 0.0036 | | | | | | | |
| **N** | 0.0072 | 0.0024 | 0.0016 | 0.0040 | 0.0056 | 0.0024 | 0.0004 | | | | | | |
| **P** | 0.0288 | 0.0096 | 0.0064 | 0.0160 | 0.0224 | 0.0096 | 0.0032 | 0.0064 | | | | | |
| **Q** | 0.0072 | 0.0024 | 0.0016 | 0.0040 | 0.0056 | 0.0024 | 0.0008 | 0.0032 | 0.0004 | | | | |
| **R** | 0.0360 | 0.0120 | 0.0080 | 0.0200 | 0.0280 | 0.0120 | 0.0040 | 0.0160 | 0.0040 | 0.0100 | | | |
| **S** | 0.0216 | 0.0072 | 0.0048 | 0.0120 | 0.0168 | 0.0072 | 0.0024 | 0.0096 | 0.0024 | 0.0120 | 0.0036 | | |
| **T** | 0.0360 | 0.0120 | 0.0080 | 0.0200 | 0.0280 | 0.0120 | 0.0040 | 0.0160 | 0.0040 | 0.0200 | 0.0120 | 0.0100 | |
| **V** | 0.0144 | 0.0048 | 0.0032 | 0.0080 | 0.0112 | 0.0048 | 0.0016 | 0.0064 | 0.0016 | 0.0080 | 0.0048 | 0.0080 | 0.0016 |



$$e_{AB} = \begin{cases} (p_A)\,(p_B) = (p_A)^2 & \text{if } A = B \\ (p_A)\,(p_B) + (p_B)\,(p_A) = 2\,(p_A)\,(p_B) & \text{otherwise} \end{cases}$$

# Example

## Matrix of scores



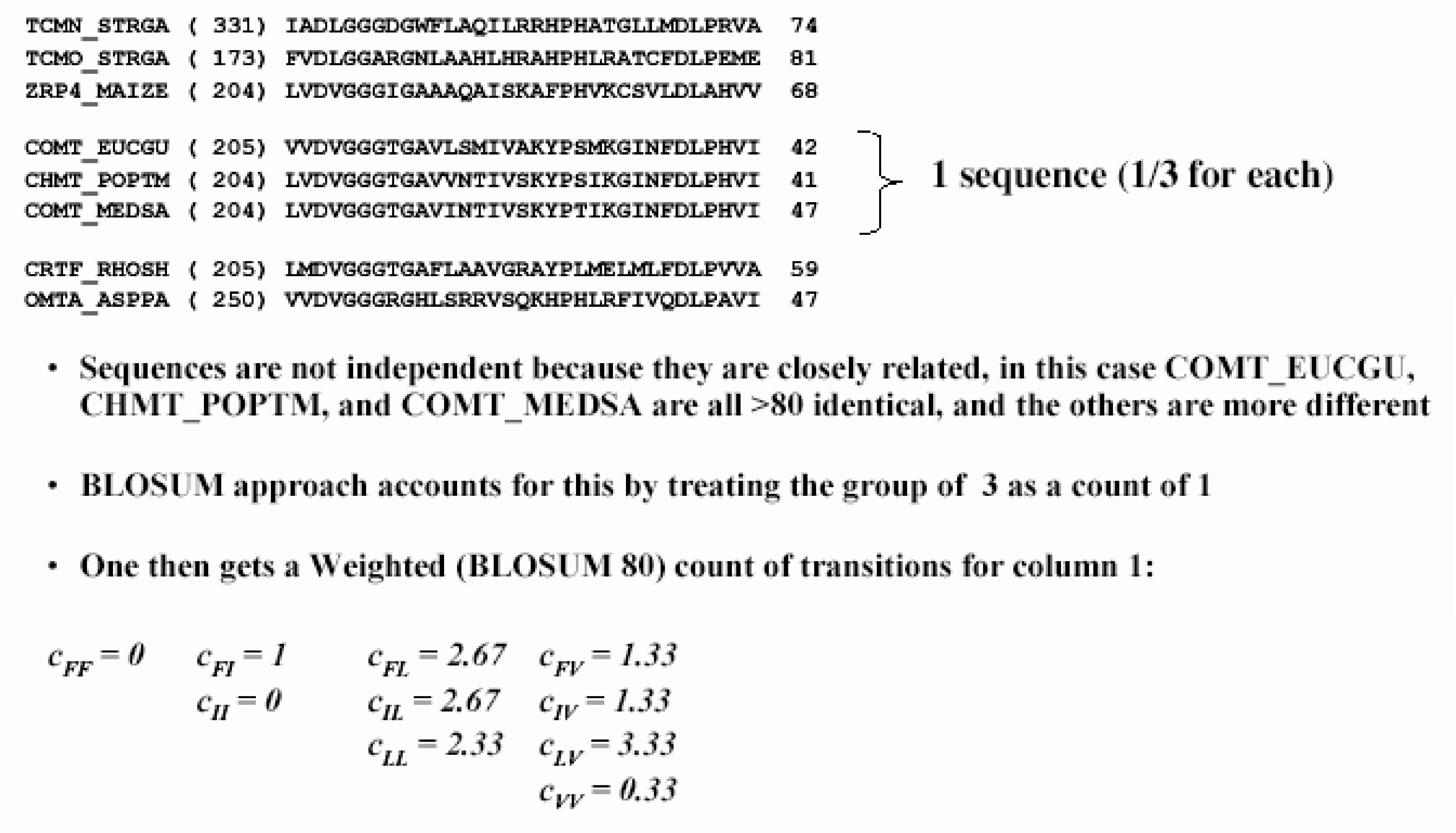
$$s_{AB} = \mathrm{Round}\left(2\log_2\left(\frac{q_{AB}}{e_{AB}}\right)\right)$$

| | A | D | E | F | G | L | N | P | Q | R | S | T | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 5 | | | | | | | | | | | | |
| **D** | | 6 | | | | | | | | | | | |
| **E** | | 7 | 5 | | | | | | | | | | |
| **F** | | | | 7 | | | | | | | | | |
| **G** | | | | | 4 | | | | | | | | |
| **L** | | | | | 5 | 3 | | | | | | | |
| **N** | | | | | | | | | | | | | |
| **P** | 1 | | | | | | | 4 | | | | | |
| **Q** | | | | | | | 7 | | | | | | |
| **R** | | | | | | | | | | 7 | | | |
| **S** | | | | | | | 7 | | 7 | | 6 | | |
| **T** | | | | | | | | | | | | 7 | |
| **V** | | | | | | | | 6 | | | | | 5 |

Blank cells are “missing data” (i.e. no observed values); wouldn’t happen with sufficient training data.

# Dealing with sequence redundancy

E.g., for BLOSUM-80, group sequences that are >80% similar



```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68

COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41   } 1 sequence (1/3 for each)
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47

CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47
```

- **Sequences are not independent because they are closely related, in this case COMT_EUCGU, CHMT_POPTM, and COMT_MEDSA are all >80 identical, and the others are more different**

- **BLOSUM approach accounts for this by treating the group of 3 as a count of 1**

- **One then gets a Weighted (BLOSUM 80) count of transitions for column 1:**

$c_{FF} = 0$ $c_{FI} = 1$ $c_{FL} = 2.67$ $c_{FV} = 1.33$

$c_{II} = 0$ $c_{IL} = 2.67$ $c_{IV} = 1.33$

$c_{LL} = 2.33$ $c_{LV} = 3.33$

$c_{VV} = 0.33$

(slide from Michael Gribskov)

# Point Accepted Mutation Matrix

Introduced by Margaret Dayhoff in 1978

Observed mutation probabilities between amino acids over 71 families of closely related proteins (85% sequence identity within a family)

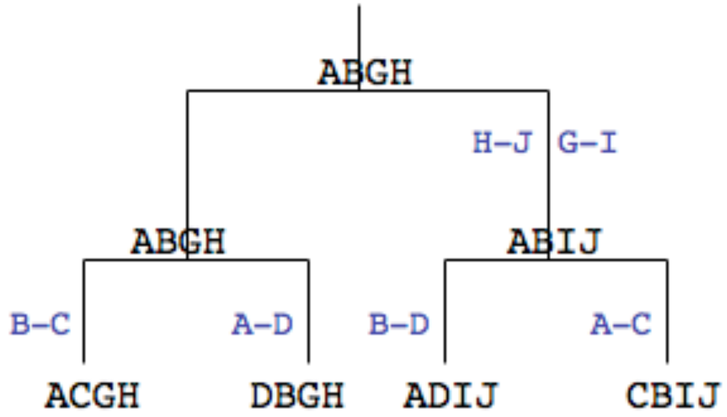
Based on a Markov mutation model; The PAM is a “unit of evolutionary mutation”. 1 PAM is the unit for which 1 mutation to occurs per 100 amino acids (this varies e.g. by species). The PAM$_1$ matrix express the log odds ratio of the likelihood of a point accepted mutation from one amino acid to another to the likelihood that these amino acids were aligned by chance.

**PAM matrix slides below courtesy of Didier Gonze**
(http://homepages.ulb.ac.be/~dgonze/TEACHING/pam_blosum.pdf)

Picture from: http://en.wikipedia.org/wiki/Margaret_Oakley_Dayhoff#mediaviewer/File:Margaret_Oakley_Dayhoff_cropped.jpg

# PAM scoring matrices

The substitution score is expected to depend on the rate of divergence between sequences.



accepted mutations

The **PAM matrices** derived by Dayhoff (1978):

- are based on evolutionary distances.
- have been obtained from carefully aligned closely related protein sequences (71 gapless alignments of sequences having at least 85% similarity).

**M. Dayhoff**

**Reference:** Dayhoff *et al.* (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–352. National Biomedical Research Foundation, Silver Spring, MD, 1978.

# PAM scoring matrices

**PAM = Percent (or Point) Accepted Mutation**

The PAM matrices are **series of scoring matrices**, each reflecting a certain level of divergence:

PAM = unit of evolution (1 PAM = 1 mutation/100 amino acid)

- PAM1 proteins with an evolutionary distance of 1% mutation/position
- PAM50 idem for 50% mutations/position
- PAM250 250% mutations/position (a position could mutate several times)

**Reference:** Dayhoff *et al.* (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, 345–352. National Biomedical Research Foundation, Silver Spring, MD, 1978.

# Derivation of the PAM matrices

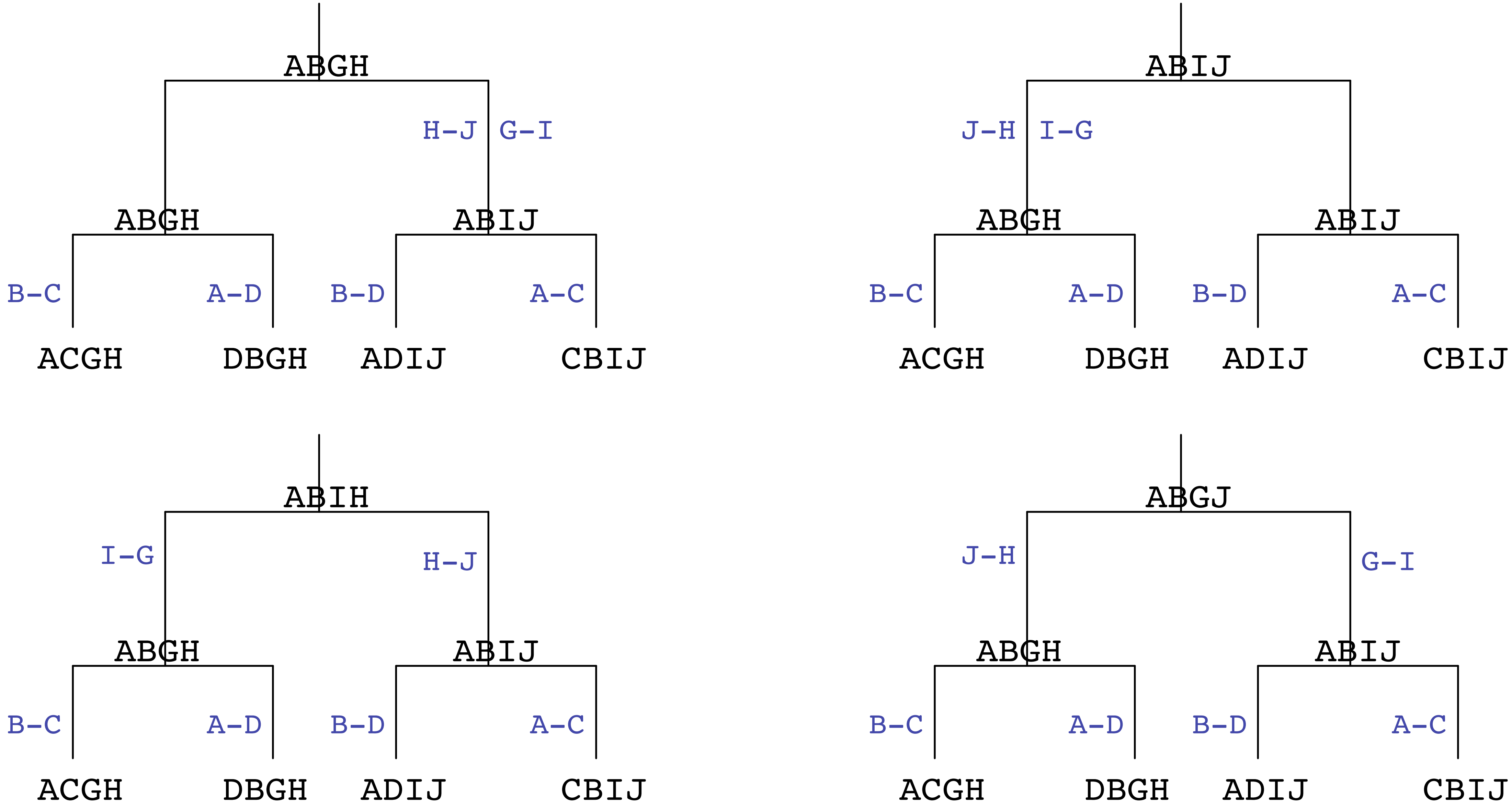
To illustrate how the PAM substitution matrices have been derived, we will consider the following artificial ungapped aligned sequences:

```
A C G H
D B G H
A D I J
C B I J
```

**Reference:** Borodovsky & Ekisheva (2007) Problems and Solutions in Biological sequence analysis. *Cambridge Univ Press*.

# Derivation of the PAM matrices

## **Phylogenetic trees** (maximum parsimony)



Here are represented the four more parsimonious (minimum of substitutions) phylogenetic trees for the alignment given above.

# Derivation of the PAM matrices

**Matrix of accepted point mutation counts (A)**

| | A | B | C | D | G | H | I | J |
|---|---|---|---|---|---|---|---|---|
| **A** | | 0 | 4 | 4 | 0 | 0 | 0 | 0 |
| **B** | 0 | | 4 | 4 | 0 | 0 | 0 | 0 |
| **C** | 4 | 4 | | 0 | 0 | 0 | 0 | 0 |
| **D** | 4 | 4 | 0 | | 0 | 0 | 0 | 0 |
| **G** | 0 | 0 | 0 | 0 | | 0 | 4 | 0 |
| **H** | 0 | 0 | 0 | 0 | 0 | | 0 | 4 |
| **I** | 0 | 0 | 0 | 0 | 4 | 0 | | 0 |
| **J** | 0 | 0 | 0 | 0 | 0 | 4 | 0 | |

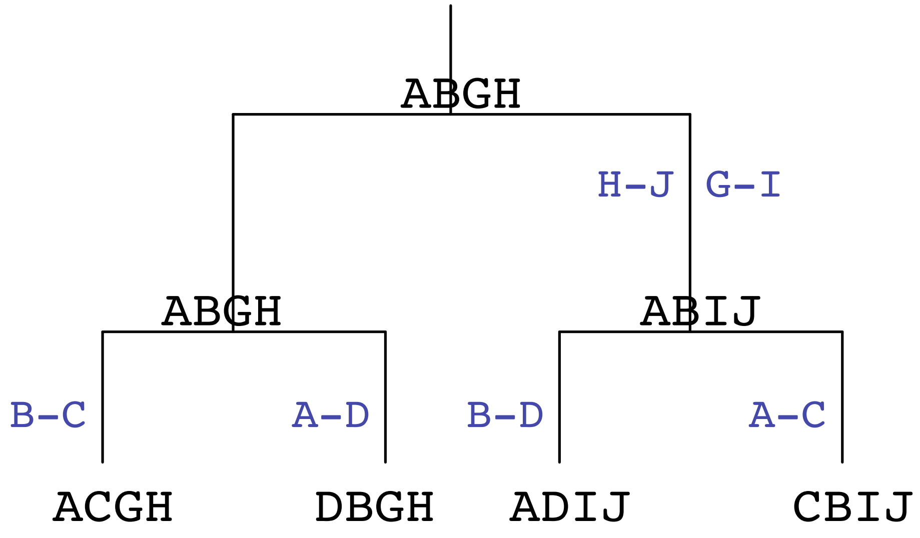
For each pair of different amino acids ($i$,$j$), the total number $a_{ij}$ of substitutions from $i$ to $j$ along the edges of the phylogenetic tree is calculated.

(they are indicated in blue on the previous slide)

# Derivation of the PAM matrices

Each edge of a given tree is associated with the ungapped alignment of the two sequences connected by this edge.

Thus, any tree shown above generates 6 alignments. For example the first phylogenetic tree generates the following alignments:



```
A B G H     A B G H     A B G H
A B G H     A B I J     A C G H

A B G H     A B I J     A B I J
D B G H     A D I J     C B I J
```

Those alignments can be used to assess the "relative mutability" of each amino acid.

# Derivation of the PAM matrices

### Relative mutability ($m_i$)

The relative mutability is defined by the ratio of the total number of times that amino acid *j* has changed in all the pair-wise alignments (in our case 6x4=24 alignments) to the number of times that *j* has occurred in these alignments, i.e.

alns / tree

# of trees

$$m_j = \frac{\textit{number of changes of } j}{\textit{number of occurrences of } j}$$

Relative amino acid mutability values $m_j$ *for our example*

| Amino acid | A | B | I | H | G | J | C | D |
|---|---|---|---|---|---|---|---|---|
| Changes (substitutions) | 8 | 8 | 4 | 4 | 4 | 4 | 8 | 8 |
| Frequency of occurrence | 40 | 40 | 24 | 24 | 24 | 24 | 8 | 8 |
| Relative mutability $m_j$ | 0.2 | 0.2 | 0.167 | 0.167 | 0.167 | 0.167 | 1 | 1 |

The relative mutability accounts for the fact that the different amino acids have different mutation rates. This is thus the probability to mutate.

# Derivation of the PAM matrices

## Relative mutability of the 20 amino acids

| aa | $m_i$ | aa | $m_i$ |
|---|---|---|---|
| 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 |

Values according Dayhoff (1978)

The value for Ala has been arbitrarily set at 100.

**Trp** and **Cys** are less mutable

Cys is known to have several unique, indispensable function (attachment site of heme group in cytochrome and of FeS clusters in ferredoxin). It also forms cross-links such as in chymotrypsin or ribonuclease.

Big groups like Trp or Phe are less mutable due to their particular chemistry. On the other extreme, the low mutability of Cys must be due to its unique smallness that is advatageous in many places.

**Asn**, **Ser**, **Asp** and **Glu** are most mutable

Although Ser sometimes functions in the active center, it more often performs a function of lesser importance, easily mimicked by several other amino acids of similar physical and chemical properties.

# Derivation of the PAM matrices

## Effective frequency (*$f_i$*)

The notion of effective frequency $f_i$ takes into account the difference in variability of the primary structure conservation in proteins with different functional roles. Two alignment blocks corresponding to 2 different families may contribute differently to $f_i$ even if the number of occurrence of amino acid *j* in these blocks is the same.

$$\begin{pmatrix} \textit{relative frequency of} \\ \textit{exposure to mutation} \end{pmatrix} = \begin{pmatrix} \textit{average composition} \\ \textit{of each group} \end{pmatrix} \times \begin{pmatrix} \textit{number of mutations in} \\ \textit{the corresponding tree} \end{pmatrix}$$

# Derivation of the PAM matrices

### Effective frequency ($f_i$)

The effective frequency is defined as

$$f_j = k \sum_b q_j^{(b)} N^{(b)}$$

where the sum is taken over all alignment blocks *b*
$q_j^{(b)}$ is the observed frequency of amino acid *j* in block *b*,
$N^{(b)}$ is the number of substitutions in a tree built for *b*
and the coefficient *k* is chosen the ensure that the sum of the frequences $f_j = 1$.

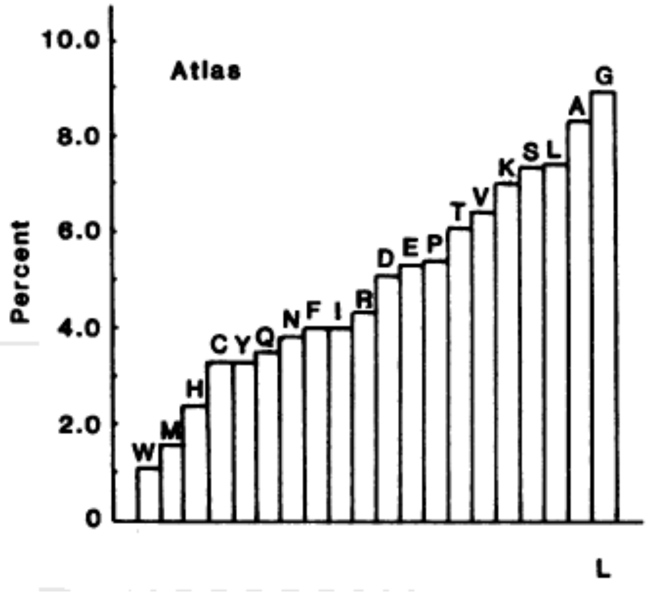
In our example, there is only one block, therefore the effective frequencies are equal to the compositional frequencies ($f_i = q_j$)

# Derivation of the PAM matrices

Effective frequency of the 20 amino acids determined for the original alignment data (Dayhoff *et al.*, 1978)

| **Amino acid** | **Gly** | **Ala** | **Leu** | **Lys** | **Ser** | **Val** | **Thr** |
|---|---|---|---|---|---|---|---|
| Frequency *f* | 0.089 | 0.087 | 0.085 | 0.081 | 0.070 | 0.065 | 0.058 |
| **Amino acid** | **Pro** | **Glu** | **Asp** | **Arg** | **Asn** | **Phe** | **Gln** |
| Frequency *f* | 0.051 | 0.050 | 0.047 | 0.041 | 0.040 | 0.040 | 0.038 |
| **Amino acid** | **Ile** | **His** | **Cys** | **Tyr** | **Met** | **Trp** | |
| Frequency *f* | 0.037 | 0.034 | 0.033 | 0.030 | 0.015 | 0.010 | |

Source: Dayhoff, 1978



Distribution of amino acids found in 1081 peptides and proteins listed in the *Atlas of Protein Sequence and Structure* (1981).

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? *Science*. 214:149-59.

# Derivation of the PAM matrices

## Mutational probability matrix (*M*)

Let's define $M_{ij}$ the probability of the amino acid in column *j* having been substituted by an amino acid in row *i* over a given evolutionary time unit.

Non-diagonal elements of M:

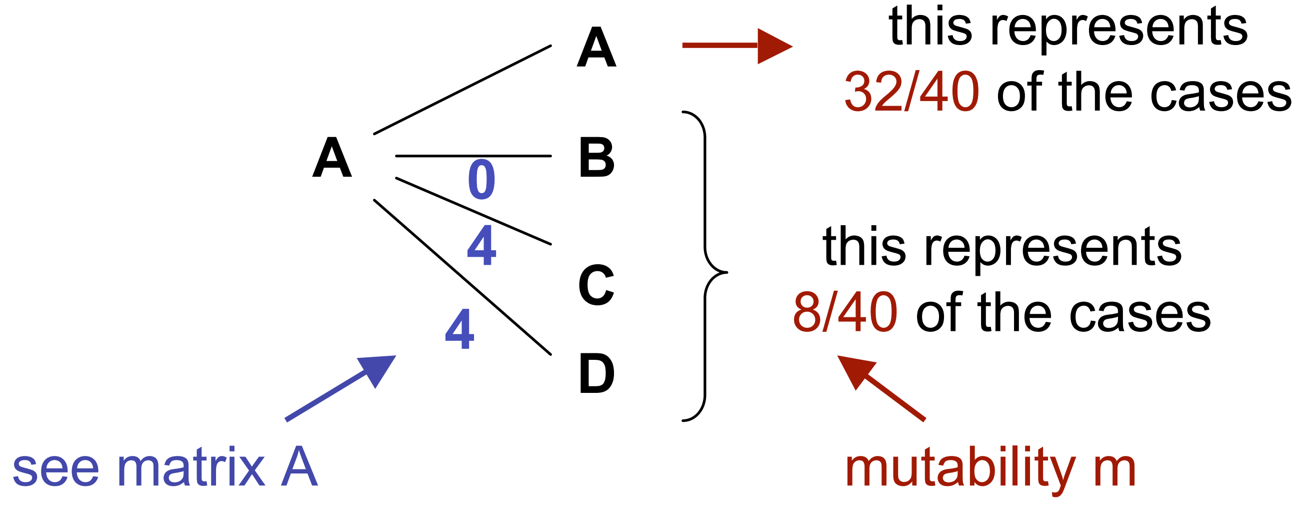
$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_k A_{kj}}$$

Diagonal elements of M:

$$M_{ii} = 1 - \lambda m_i$$

In these equations, m is the relative mutability and A is the matrix of accepted point mutations. The constant λ represents a degree of freedom that could be used to connect the matrix M with an evolutionary time scale.

**In our example:**



this represents 32/40 of the cases

this represents 8/40 of the cases

see matrix A

mutability m

If A is mutated, the probability that it is mutated into D is

$A_{DA}/(A_{BA}+A_{CA}+A_{DA}) = 4/8$

Thus the probability that A is mutated into D is:

$M_{DA} = 4/8 * 8/40 = 4/40$

and the probability that A is not mutated is:

$M_{AA} = 1 - 8/40 = 32/40$

# Derivation of the PAM matrices

## Mutational probability matrix (*M*)

Let's define $M_{ij}$ the probability of the amino acid in column *j* having been substituted by an amino acid in row *i* over a given evolutionary time unit.

Non-diagonal elements of M:

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_k A_{kj}}$$

Diagonal elements of M:

$$M_{ii} = 1 - \lambda m_i$$

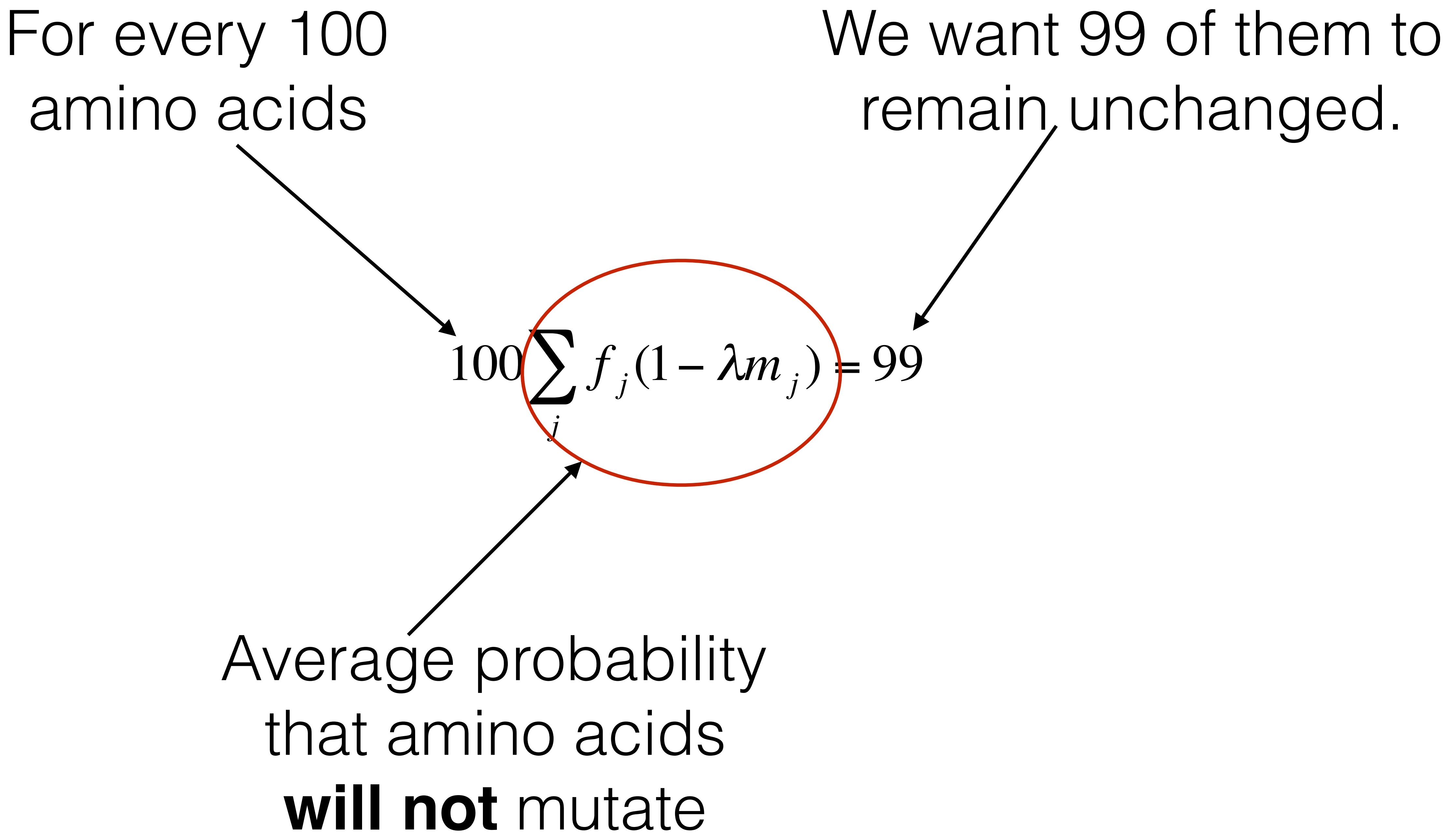
In these equations, m is the relative mutability and A is the matrix of accepted point mutations. The constant λ represents a degree of freedom that could be used to connect the matrix M with an evolutionary time scale.

The coefficient λ could be adjusted to ensure that a specific (small) number of substitutions would occur on average per hundred residues. This adjustement was done by Dayhoff *et al* in the following way. The expected number of amino acids that will remain inchanged in a protein sequence 100 amino acid long is given by:

$$100 \sum_j f_j M_{jj} = 100 \sum_j f_j (1 - \lambda m_j)$$

If only one substitution per residue is allowed, then λ is calculated from the equation:

$$100 \sum_j f_j (1 - \lambda m_j) = 99$$



For every 100 amino acids

We want 99 of them to remain unchanged.

$$100\sum_j f_j(1-\lambda m_j) = 99$$

Average probability that amino acids **will not** mutate

# Derivation of the PAM matrices

## Mutational probability matrix

In our example, $\lambda = 0.0261$ and the mutation probability matrix (PAM1) is:

| | A | B | C | D | G | H | I | J |
|---|---|---|---|---|---|---|---|---|
| **A** | 0.9948 | 0 | 0.0131 | 0.0131 | 0 | 0 | 0 | 0 |
| **B** | 0 | 0.9948 | 0.0131 | 0.0131 | 0 | 0 | 0 | 0 |
| **C** | 0.0026 | 0.0026 | 0.9740 | 0 | 0 | 0 | 0 | 0 |
| **D** | 0.0026 | 0.0026 | 0 | 0.9740 | 0 | 0 | 0 | 0 |
| **G** | 0 | 0 | 0 | 0 | 0.9957 | 0 | 0.0043 | 0 |
| **H** | 0 | 0 | 0 | 0 | 0 | 0.9957 | 0 | 0.0043 |
| **I** | 0 | 0 | 0 | 0 | 0.0043 | 0 | 0.9957 | 0 |
| **J** | 0 | 0 | 0 | 0 | 0 | 0.0043 | 0 | 0.9957 |

Note that *M* is a non-symmetric matrix.

# Derivation of the PAM matrices

## Mutational probability matrix derived by Dayhoff for the 20 amino acids

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 9867 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 | 6 | 4 | 2 | 6 | 2 | 22 | 35 | 32 | 0 | 2 | 18 |
| R | 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 8 | 0 | 1 |
| N | 4 | 1 | 9822 | 36 | 0 | 4 | 6 | 6 | 21 | 3 | 1 | 13 | 0 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| D | 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 1 |
| C | 1 | 1 | 0 | 0 | 9973 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 3 | 2 |
| Q | 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 23 | 1 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 0 | 0 | 1 |
| E | 10 | 0 | 7 | 56 | 0 | 35 | 9865 | 4 | 2 | 3 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 | 2 |
| G | 21 | 1 | 12 | 11 | 1 | 3 | 7 | 9935 | 1 | 0 | 1 | 2 | 1 | 1 | 3 | 21 | 3 | 0 | 0 | 5 |
| H | 1 | 8 | 18 | 3 | 1 | 20 | 1 | 0 | 9912 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 1 | 1 | 4 | 1 |
| I | 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 | 9872 | 9 | 2 | 12 | 7 | 0 | 1 | 7 | 0 | 1 | 33 |
| L | 3 | 1 | 3 | 0 | 0 | 6 | 1 | 1 | 4 | 22 | 9947 | 2 | 45 | 13 | 3 | 1 | 3 | 4 | 2 | 15 |
| K | 2 | 37 | 25 | 6 | 0 | 12 | 7 | 2 | 2 | 4 | 1 | 9926 | 20 | 0 | 3 | 8 | 11 | 0 | 1 | 1 |
| M | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 5 | 8 | 4 | 9874 | 1 | 0 | 1 | 2 | 0 | 0 | 4 |
| F | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 8 | 6 | 0 | 4 | 9946 | 0 | 2 | 1 | 3 | 28 | 0 |
| P | 13 | 5 | 2 | 1 | 1 | 8 | 3 | 2 | 5 | 1 | 2 | 2 | 1 | 1 | 9926 | 12 | 4 | 0 | 0 | 2 |
| S | 28 | 11 | 34 | 7 | 11 | 4 | 6 | 16 | 2 | 2 | 1 | 7 | 4 | 3 | 17 | 9840 | 38 | 5 | 2 | 2 |
| T | 22 | 2 | 13 | 4 | 1 | 3 | 2 | 2 | 1 | 11 | 2 | 8 | 6 | 1 | 5 | 32 | 9871 | 0 | 2 | 9 |
| W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 9976 | 1 | 0 |
| Y | 1 | 0 | 3 | 0 | 3 | 0 | 1 | 0 | 4 | 1 | 1 | 0 | 0 | 21 | 0 | 1 | 1 | 2 | 9945 | 1 |
| V | 13 | 2 | 1 | 1 | 3 | 2 | 2 | 3 | 3 | 57 | 11 | 1 | 17 | 1 | 3 | 2 | 10 | 0 | 2 | 9901 |

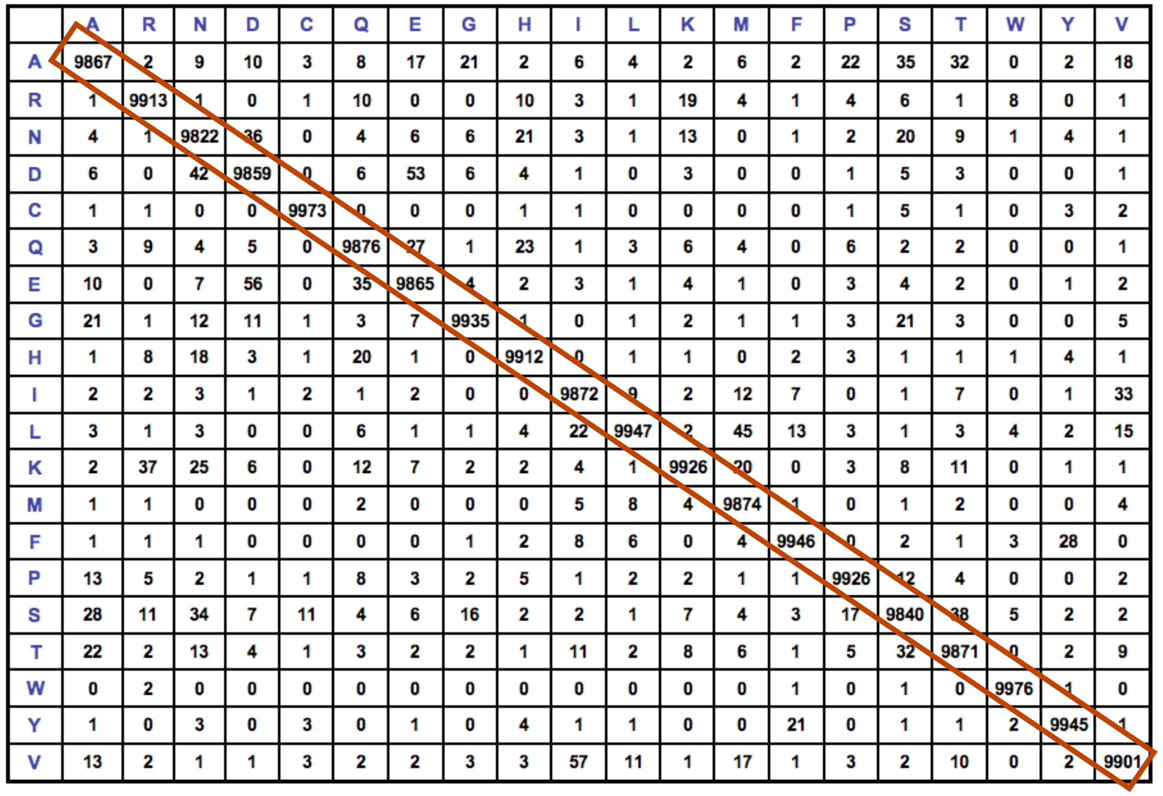
For clarity, the values have been multiplied by 10000

This matrix corresponds to an evolution time period giving 1 mutation/100 amino acids, and is refered to as the **PAM1 matrix.**

Source: Dayhoff, 1978

# Derivation of the PAM matrices

## Mutational probability matrix derived by Dayhoff for the 20 amino acids



| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 9867 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 | 6 | 4 | 2 | 6 | 2 | 22 | 35 | 32 | 0 | 2 | 18 |
| R | 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 8 | 0 | 1 |
| N | 4 | 1 | 9822 | 36 | 0 | 4 | 6 | 6 | 21 | 3 | 1 | 13 | 0 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| D | 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 1 |
| C | 1 | 1 | 0 | 0 | 9973 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 3 | 2 |
| Q | 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 23 | 1 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 0 | 0 | 1 |
| E | 10 | 0 | 7 | 56 | 0 | 35 | 9865 | 4 | 2 | 3 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 | 2 |
| G | 21 | 1 | 12 | 11 | 1 | 3 | 7 | 9935 | 1 | 0 | 1 | 2 | 1 | 1 | 3 | 21 | 3 | 0 | 0 | 5 |
| H | 1 | 8 | 18 | 3 | 1 | 20 | 1 | 0 | 9912 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 1 | 1 | 4 | 1 |
| I | 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 | 9872 | 9 | 2 | 12 | 7 | 0 | 1 | 7 | 0 | 1 | 33 |
| L | 3 | 1 | 3 | 0 | 0 | 6 | 1 | 1 | 4 | 22 | 9947 | 2 | 45 | 13 | 3 | 1 | 3 | 4 | 2 | 15 |
| K | 2 | 37 | 25 | 6 | 0 | 12 | 7 | 2 | 2 | 4 | 1 | 9926 | 20 | 0 | 3 | 8 | 11 | 0 | 1 | 1 |
| M | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 5 | 8 | 4 | 9874 | 1 | 0 | 1 | 2 | 0 | 0 | 4 |
| F | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 8 | 6 | 0 | 4 | 9946 | 0 | 2 | 1 | 3 | 28 | 0 |
| P | 13 | 5 | 2 | 1 | 1 | 8 | 3 | 2 | 5 | 1 | 2 | 2 | 1 | 1 | 9926 | 12 | 4 | 0 | 0 | 2 |
| S | 28 | 11 | 34 | 7 | 11 | 4 | 6 | 16 | 2 | 2 | 1 | 7 | 4 | 3 | 17 | 9840 | 38 | 5 | 2 | 2 |
| T | 22 | 2 | 13 | 4 | 1 | 3 | 2 | 2 | 1 | 11 | 2 | 8 | 6 | 1 | 5 | 32 | 9871 | 0 | 2 | 9 |
| W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 9976 | 1 | 0 |
| Y | 1 | 0 | 3 | 0 | 3 | 0 | 1 | 0 | 4 | 1 | 1 | 0 | 0 | 21 | 0 | 1 | 1 | 2 | 9945 | 1 |
| V | 13 | 2 | 1 | 1 | 3 | 2 | 2 | 3 | 3 | 57 | 11 | 1 | 17 | 1 | 3 | 2 | 10 | 0 | 2 | 9901 |

This matrix is the mutation probability matrix for an evolution time of **1 PAM**.

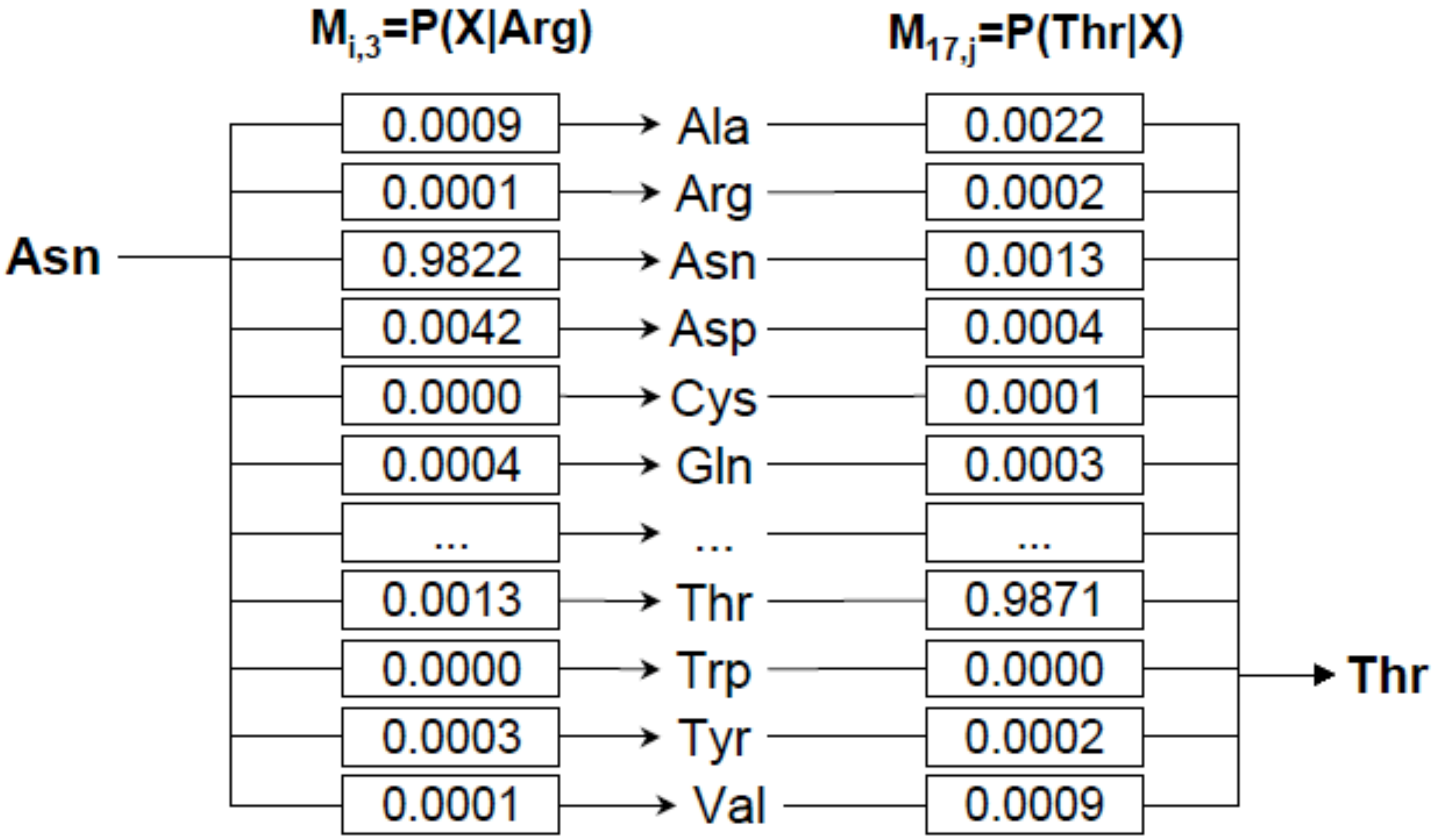
The **diagonal** represents the probability to still observe the same residue after 1 PAM. Therefore the diagonal represents the **99% of the case of non-mutation**.

Note that this does not mean that there was no mutation during this time interval. Indeed, the conservation of a residue could reflect either a conservation during the whole period, or a succession of two or more mutations ending at the initial residue

Source: J. van Helden

# Derivation of the PAM matrices

## From PAM1 to PAM2



Asn → Thr

| $M_{i,3}$=P(X\|Arg) |  | $M_{17,j}$=P(Thr\|X) |
|---|---|---|
| 0.0009 | Ala | 0.0022 |
| 0.0001 | Arg | 0.0002 |
| 0.9822 | Asn | 0.0013 |
| 0.0042 | Asp | 0.0004 |
| 0.0000 | Cys | 0.0001 |
| 0.0004 | Gln | 0.0003 |
| ... | ... | ... |
| 0.0013 | Thr | 0.9871 |
| 0.0000 | Trp | 0.0000 |
| 0.0003 | Tyr | 0.0002 |
| 0.0001 | Val | 0.0009 |

P(Asn -> Thr)= P(Asn -> Ala -> Thr) + P(Asn -> Arg -> Thr) + ... + P(Asn -> Val -> Thr)
= (0.0009)(0.0001) + (0.0001)(0.0002) + ... + (0.0001)(0.009)

line 3 of PAM1 column 17 of PAM1

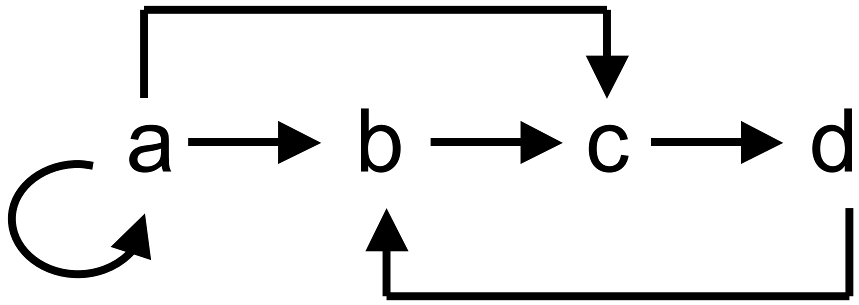
**=> Matrix product: PAM2 = PAM1 x PAM1**

Source: J. van Helden

# Derivation of the PAM matrices

## From PAM1 to PAM2, PAM100, PAM250, etc...

**Remark** (from graph theory)



| | a | b | c | d |
|---|---|---|---|---|
| **a** | 1 | 1 | 1 | 0 |
| **b** | 0 | 0 | 1 | 0 |
| **c** | 0 | 0 | 0 | 1 |
| **d** | 0 | 1 | 0 | 0 |

Matrix **Q** indicates the number of paths going from one node to another in 1 step

| | a | b | c | d |
|---|---|---|---|---|
| **a** | 1 | 1 | 2 | 1 |
| **b** | 0 | 0 | 0 | 1 |
| **c** | 0 | 1 | 0 | 1 |
| **d** | 0 | 1 | 1 | 1 |

Matrix $\mathbf{Q}^2$ indicates the number of paths going from one node to another in 2 steps

| | a | b | c | d |
|---|---|---|---|---|
| **a** | ... | ... | ... | ... |
| **b** | ... | ... | ... | ... |
| **c** | ... | ... | ... | ... |
| **d** | ... | ... | ... | ... |

Matrix $\mathbf{Q}^n$ indicates the number of paths going from one node to another in *n* steps

Source: J. van Helden

# Derivation of the PAM matrices

## From PAM1 to PAM2, PAM100, PAM250, etc...

**Similarly:**

PAM1 gives the probability to observe the changes *i* → *j* per 100 mutations

$PAM2 = PAM1^2$ gives the probability to observe the changes *i* → *j* per 200 mutations

$PAM100 = PAM1^{100}$ gives the probability to observe the changes *i* → *j* per 10 000 mutations

$PAM250 = PAM1^{250}$ gives the probability to observe the changes *i* → *j* per 25 000 mutations

$PAMn = PAM1^n$ gives the probability to observe the changes *i* → *j* per 100x*n* mutations.

**Convergence:** it can be verified that

$PAM\infty = PAM1^\infty$ converges to the observed frequencies:

$$\lim_{n\to\infty} M^n = \begin{pmatrix} f_A & f_A & \cdots & f_A \\ f_R & f_R & \cdots & f_R \\ \cdots & \cdots & & \cdots \\ f_V & f_V & \cdots & f_V \end{pmatrix}$$

Dayhoff *et al.* (1978) checked this convergence by computing $M^{2034}$.

# Derivation of the PAM matrices

## PAM250 derived by Dayhoff for the 20 amino acids

| | 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 | 4 | 5 | 4 | 15 | 10 | 4 | 10 | 5 | 5 | 5 | 72 | 4 | 17 |

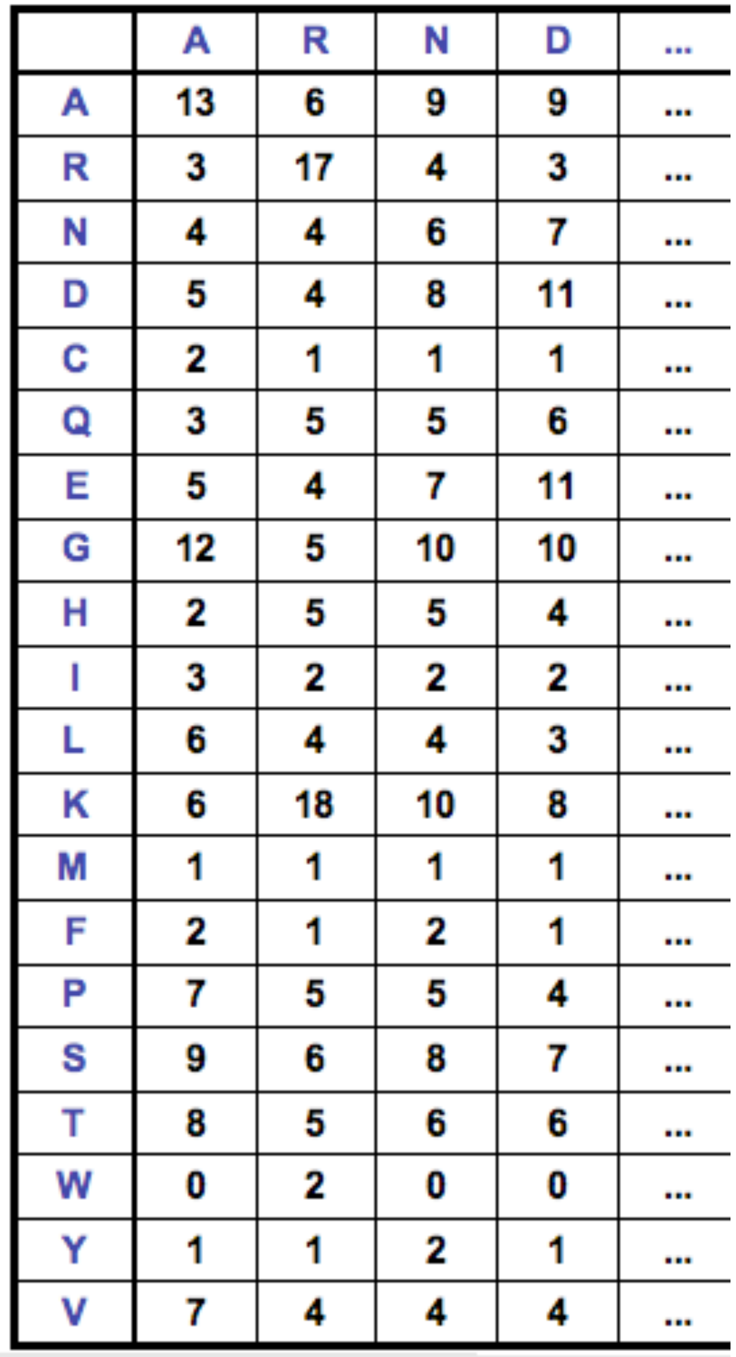
For clarity, the values have been multiplied by 100

This matrix corresponds to an evolution time period giving 250 mutation/100 amino acids (i.e. an evolutionary distance of 250 PAM), and is refered to as the **PAM250 matrix.**

Source: Dayhoff, 1978

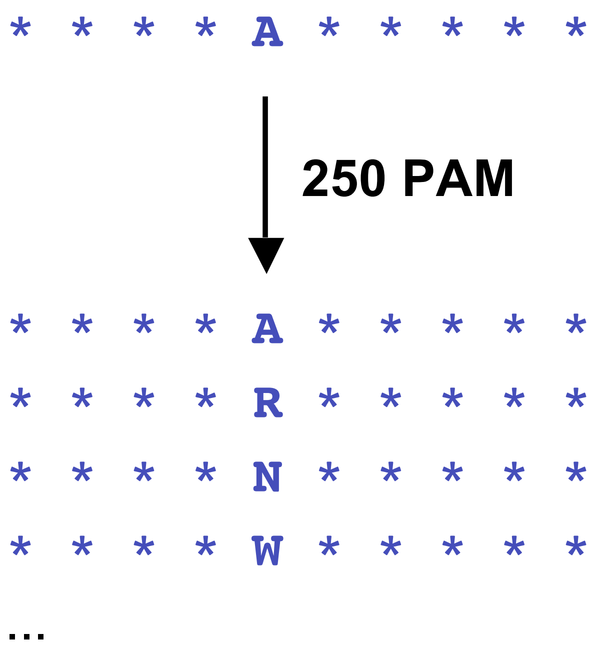
# Derivation of the PAM matrices

## Interpretation of the PAM250 matrix



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

In comparing 2 sequences at this evolutionary distance (250 PAM), there is:



**probability of 13%**

**probability of 3%**

**probability of 4%**

**probability of 0%**

Source: Dayhoff, 1978

# Derivation of the PAM matrices

## From probabilities to scores

So far, we have obtained a **probability matrix**, but we would like a **scoring matrix**.

A **score** should reflect the significance of an alignment occurring as a result of an evolutionary process with respect to what we could expect by chance.

A score should involve the ratio between the probability derived from non-random (evolutionary) to random models:

$$r_n(i,j) = \frac{M^n_{ji}}{f_j} = \frac{P_{ji,n}}{f_i f_j}$$

← probability to see a pair (i,j) due to evolution

← probability to see a pair (i,j) by chance

The matrix $\boldsymbol{M_{ji}^n}$ is the mutational probability matrices at PAM distance *n*. Matrices $M^1$ and $M^{250}$ have been shown before.

$\boldsymbol{P_{ji,n} = f_i M_{ji}^n}$ is the probability that two aligned amino acids have diverged from a common ancestor *n*/2 PAM unit ago, assuming that the substitutions follow a Markov process (for details, see Borodovsky & Ekisheva, 2007).

Note that ***R*** (the odd-score or relatedness matrix) is a symmetric matrix.

# Derivation of the PAM matrices

## Log-odd scores

In practice, we often use the log-odd scores defined by

$$s_n(i,j) = \log \frac{M_{ji}^{n}}{f_j} = \log \frac{P_{ji,n}}{f_i f_j}$$

This definition has convenient practical consequences:

A **positive score** ($s_n > 0$) characterizes the accepted mutations
A **negative score** ($s_n < 0$) characterizes the unfavourable mutations

Another property of the log-odd scores is that they can be added to produce the score of an alignment:

```
T   A   H   G   K
Y   S   D   G   D
```

$S_{alignment}$ = s(T,Y) + s(A,S) + s(H,D) + s(G,G) + s(K,D)

# Derivation of the PAM matrices

## PAM250 matrix (log-odds)

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

Source: J. van Helden

# PAM matrices: exercise

The original PAM250 substitution matrix scores a substitution of *Gly* by *Arg* by a negative score -3 (decimal logarithm and scaling factor 10 are used, with rounding to the nearest neighbour). The average frequency of *Arg* in the protein sequence database is 0.041. Use this information as well as the method described above to estimate the probability that *Gly* will be substituted by *Arg* after a PAM250 time period.

Source: Borodovsky & Ekisheva (2007)

# PAM matrices: exercise

The original PAM250 substitution matrix scores a substitution of *Gly* by *Arg* by a negative score -3 (decimal logarithm and scaling factor 10 are used, with rounding to the nearest neighbour). The average frequency of *Arg* in the protein sequence database is 0.041. Use this information as well as the method described above to estimate the probability that *Gly* will be substituted by *Arg* after a PAM250 time period.

The element $s_{ij}$ of the PAM250 substitution matrix and the frequency of amino acid $j$ ($f_j$) in a protein sequence database are connected by the following formula:

$$s_{ij} = \left( 10 \log \frac{P(i \to j \text{ in } 250 \text{ } PAM)}{f_j} \right)$$

Therefore, the probability of substitution of *Gly* by *Arg* is:

$$P(Gly \to Arg \text{ in } 250 \text{ } PAM) = 0.041 \times 10^{-0.3} = 0.0205$$

Source: Borodovsky & Ekisheva (2007)

# Derivation of the PAM matrices

## Scoring an alignment

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

A scoring matrix like PAM250 can be used to score an alignment

T A H G K

Y S D G D

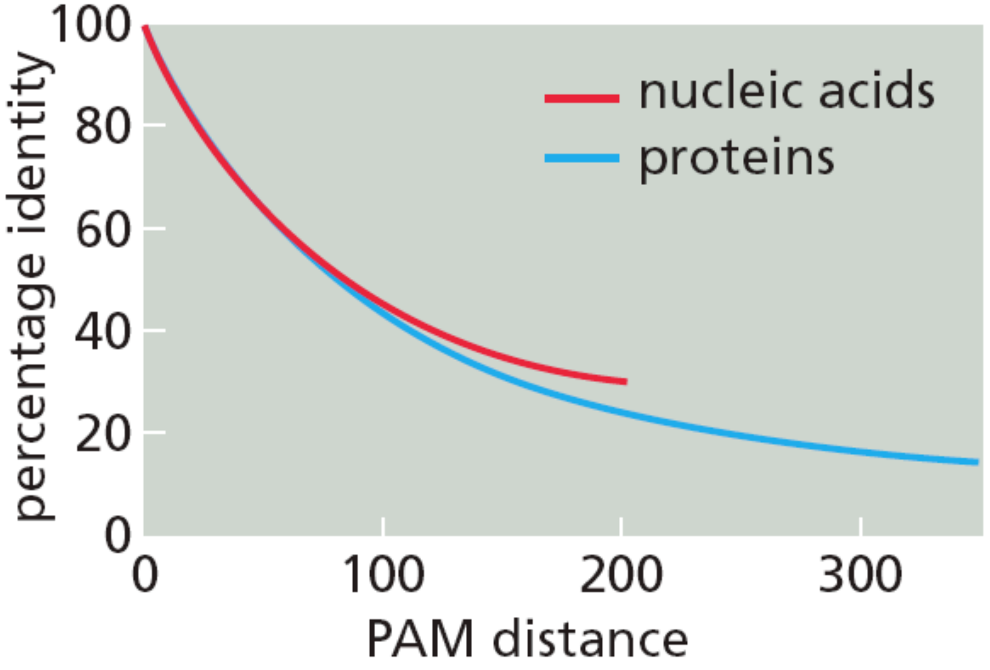
$$
\begin{aligned}
S_{\text{alignment}} &= s(T,Y) + s(A,S) + s(H,D) + s(G,G) + s(K,D) \\
&= -3 + 1 + 1 + 5 + 0 \\
&= 4
\end{aligned}
$$

# Choosing the appropriate PAM matrix

**How to choose the appropriate PAM matrix?**

Correspondance between the observed percent of amino acid difference *d* between the evolutionary distance *n* (in PAM) between them:

$$100 \sum_j f_j M_{jj}^n = 100 - d$$



**twilight zone**

(detection limit)

| ***identity (%)*** | ***difference d (%)*** | ***PAM index n*** |
|---|---|---|
| 99 | 1 | 1 |
| 95 | 5 | 5 |
| 90 | 10 | 11 |
| 85 | 15 | 17 |
| 80 | 20 | 23 |
| 75 | 25 | 30 |
| 70 | 30 | 38 |
| 60 | 40 | 56 |
| 50 | 50 | 80 |
| 40 | 60 | 112 |
| 30 | 70 | 159 |
| 20 | 80 | 246 |
| **14** | **86** | **350** |

# Choosing the appropriate PAM matrix

**How to choose the appropriate PAM matrix?**

Altschul SF(1991) Amino acid substitution matrices from an information theoretic perspective. *J Mol Biol.* 219:555-65.

- PAM120 matrix is the most appropriate for database searches
- PAM200 matrix is the most appropriate for comparing two specific proteins with suspected homology

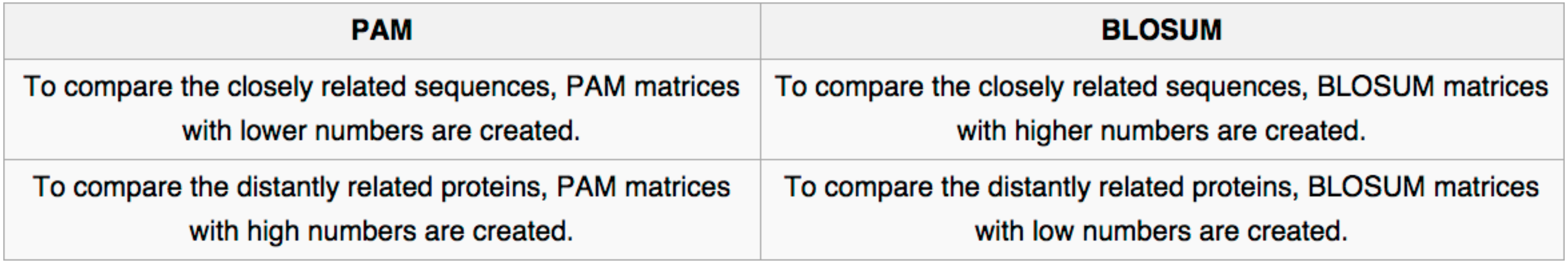
**Remark:**

In the PAM matrices, the **index** indicates the percentage of substitution per position.

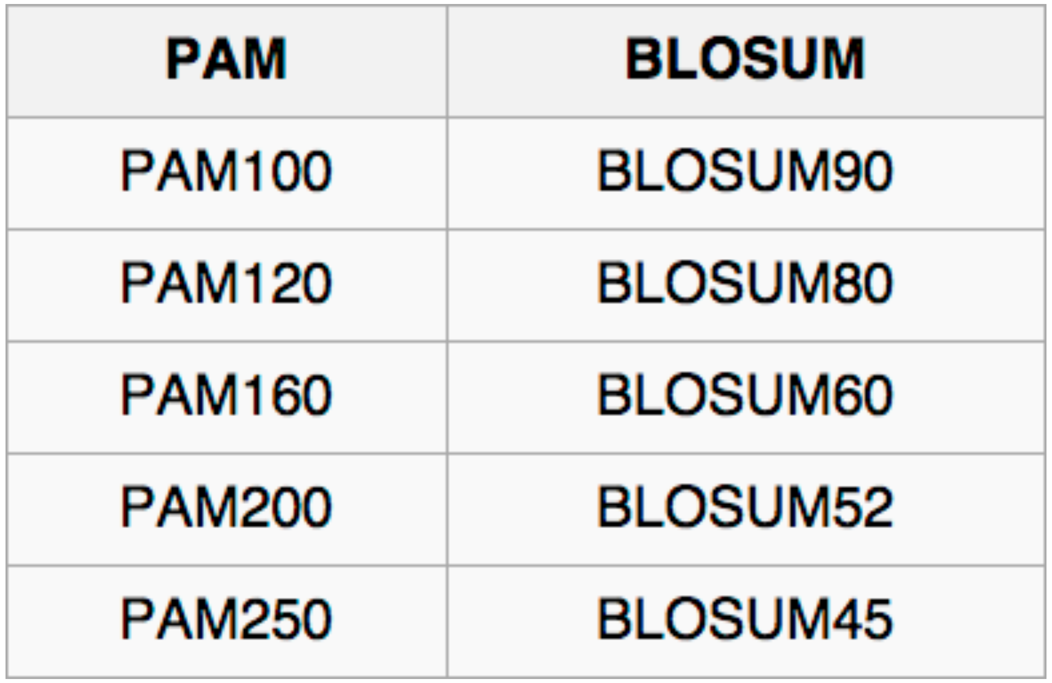
**Higher indexes** are more appropriate for **more distant** proteins (PAM250 better than PAM100 for distant proteins).

# Other Scoring Matrices

## PAM vs. BLOSUM



| PAM | BLOSUM |
|---|---|
| To compare the closely related sequences, PAM matrices with lower numbers are created. | To compare the closely related sequences, BLOSUM matrices with higher numbers are created. |
| To compare the distantly related proteins, PAM matrices with high numbers are created. | To compare the distantly related proteins, BLOSUM matrices with low numbers are created. |



| PAM | BLOSUM |
|---|---|
| PAM100 | BLOSUM90 |
| PAM120 | BLOSUM80 |
| PAM160 | BLOSUM60 |
| PAM200 | BLOSUM52 |
| PAM250 | BLOSUM45 |

from: http://en.wikipedia.org/wiki/BLOSUM, http://en.wikipedia.org/wiki/Point_accepted_mutation

# Other Scoring Matrices

## PAM vs. BLOSUM

| PAM | BLOSUM |
|---|---|
| Based on global alignments of closely related proteins. | Based on local alignments of protein segments. |
| PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergent | BLOSUM 62 is calculated from comparisons of sequences no less than 62% identical |
| Other PAM matrices are extrapolated from PAM1 | Other BLOSUM matrices are not extrapolated, but computed based on observed alignments at different identity percentage |
| Larger numbers in name denote larger evolutionary distance | Larger numbers in name denote higher sequence similarity (& therefore smaller evolutionary distance) |
| Based on explicit, Markovian, model of evolution | Not based on any explicit model of evolution, but learned empirically from alignments |

from: http://en.wikipedia.org/wiki/BLOSUM

# What about gap penalties?

Despite some work+, the setting of gap penalties is still much more arbitrary than the selection of a substitution matrix.

> ✭Gap penalty values are designed to reduce the score when an alignment has been disturbed by indels. The value should be small enough to allow a previously accumulated alignment to continue with an insertion of one of the sequences, but should not be so large that this previous alignment score is removed completely.

Changing the gap function can have significant effects on reported alignments. People often resort to “defaults” to avoid having to justify a choice.

+Reese, J. T., and William R. Pearson. "Empirical determination of effective gap penalties for sequence comparison." Bioinformatics 18.11 (2002): 1500-1507.

✭ http://en.wikipedia.org/wiki/Gap_penalty