# Scoring Matrices for Sequence Comparisons

DEKM book

Notes from Dr. Bino John

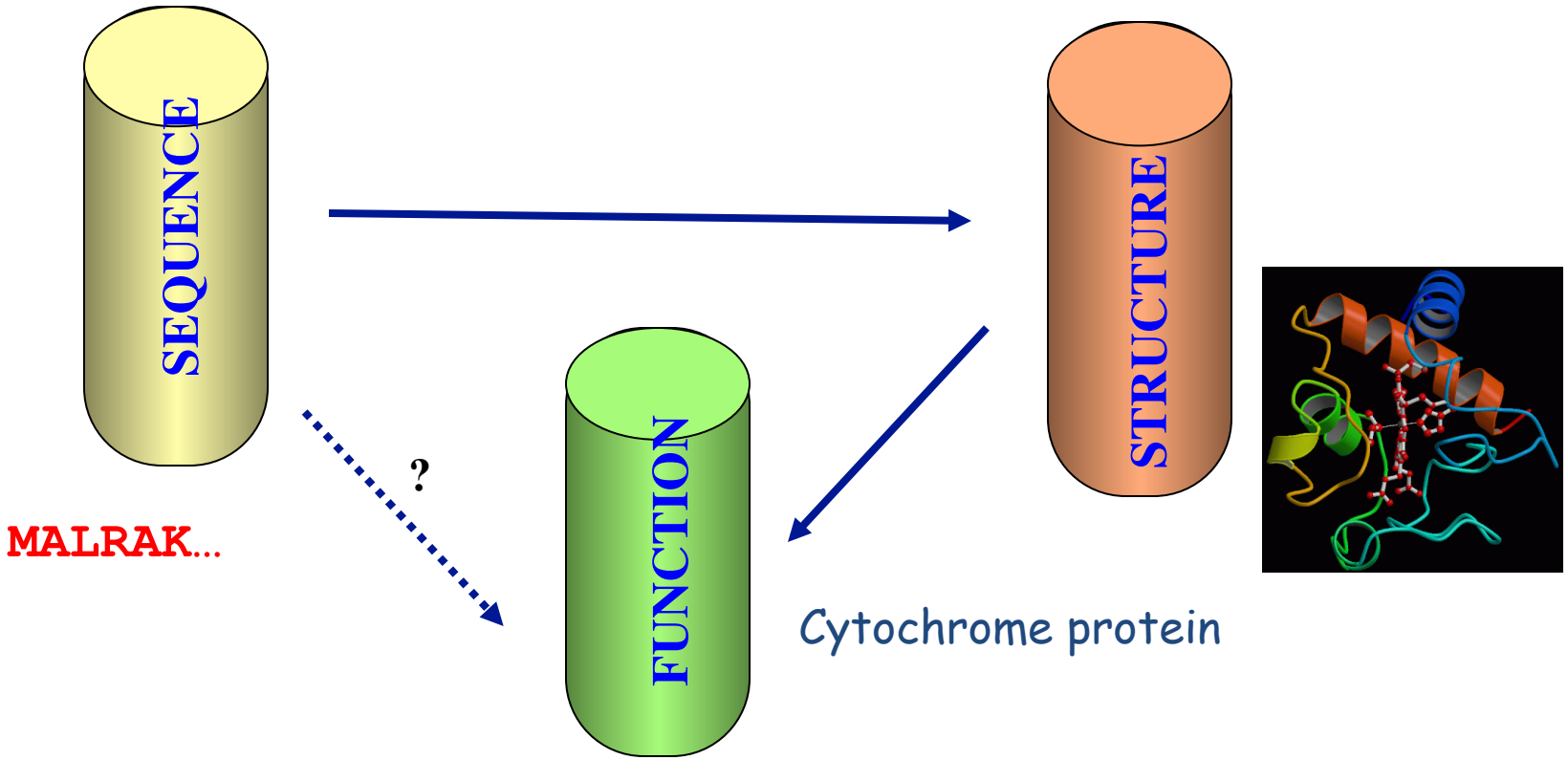
and Dr. Takis Benos

# Why compare sequences?

- Given a new sequence, infer its function based on similarity to another sequence

- Find important molecular regions – conserved across species

# Sequence -> Structure -> Function



# Important molecular regions conserved across species

Human (C11A_HUMAN; P05108) *vs. pig* (C11A_PIG; P10612)

```
Query: 1   MLAKGLPPRSVLVKGYQTFLSAPREGLGRLRVPTGEGAGISTRSPRPFNEIPSPGDNGWL 60
           MLA+GL  RSVLVKG Q FLSAPRE  G  RV TGEGA IST++PRPF+EIPSPGDNGW+
Sbjct: 1   MLARGLALRSVLVKGCQPFLSAPRECPGHPRVGTGEGACISTKTPRPFSEIPSPGDNGWI 60

Query: 61  NLYHFWRETGTHKVHLHHVQNFQKYGPIYREKLGNVESVYVIDPEDVALLFKSEGPNPER 120
           NLY FW+E GT K+H HHVQNFQKYGPIYREKLGN+ESVY+IDPEDVALLFK EGPNPER
Sbjct: 61  NLYRFWKEKGTQKIHYHHVQNFQKYGPIYREKLGNLESVYIIDPEDVALLFKFEGPNPER 120

Query: 121 FLIPPWVAYHQYYQRPIGVLLKKSAAWKKDRVALNQEVMAPEATKNFLPLLDAVSRDFVS 180
           + IPPWVAYHQ+YQ+P+GVLLKKS AWKKDR+ LN EVMAPEA KNF+PLLD VS+DFV
Sbjct: 121 YNIPPWVAYHQHYQKPVGVLLKKSGAWKKDRLVLNTEVMAPEAIKNFIPLLDTVSQDFVG 180

Query: 181 VLHRRIKKAGSGNYSGDISDDLFRFAFESITNVIFGERQGMLEEVVNPEAQRFIDAIYQM 240
           VLHRRIK+ GSG +SGDI +DLFRFAFESITNVIFGER GMLEE+V+PEAQ+FIDA+YQM
Sbjct: 181 VLHRRIKQQGSGKFSGDIREDLFRFAFESITNVIFGERLGMLEEIVDPEAQKFIDAVYQM 240

Query: 241 FHTSVPMLNLPPDLFRLFRTKTWKDHVAAWDVIFSKADIYTQNFYWELRQKGSVHHDYRG 300
           FHTSVPMLNLPPDLFRLFRTKTW+DHVAAWD IF+KA+ YTQNFYW+LR+K    ++Y G
Sbjct: 241 FHTSVPMLNLPPDLFRLFRTKTWRDHVAAWDTIFNKAEKYTQNFYWDLRRKRE-FNNYPG 299

Query: 301 MLYRLLGDSKMSFEDIKANVTEMLAGGVDTTSMTLQWHLYEMARNLKVQDMLRAEVLAAR 360
           +LYRLLG+ K+  ED+KANVTEMLAGGVDTTSMTLQWHLYEMAR+L VQ+MLR EVL AR
Sbjct: 300 ILYRLLGNDKLLSEDVKANVTEMLAGGVDTTSMTLQWHLYEMARSLNVQEMLREEVLNAR 359

Query: 361 HQAQGDMATMLQLVPLLKASIKETLRLHPISVTLQRYLVNDLVLRDYMIPAKTLVQVAIY 420
            QAQGD + MLQLVPLLKASIKETLRLHPISVTLQRYLVNDLVLRDYMIPAKTLVQVA+Y
Sbjct: 360 RQAQGDTSKMLQLVPLLKASIKETLRLHPISVTLQRYLVNDLVLRDYMIPAKTLVQVAVY 419

Query: 421 ALGREPTFFFDPENFDPTRWLSKDKNITYFRNLGFGWGVRQCLGRRIAELEMTIFLINML 480
           A+GR+P FF +P  FDPTRWL K++++ +FRNLGFGWGVRQC+GRRIAELEMT+FLI++L
Sbjct: 420 AMGRDPAFFSNPGQFDPTRWLGKERDLIHFRNLGFGWGVRQCVGRRIAELEMTLFLIHIL 479

Query: 481 ENFRVEIQHLSDVGTTFNLILMPEKPISFTFWPFNQEATQ 520
           ENF+VE+QH SDV T FNLILMP+KPI   F PFNQ+  Q
Sbjct: 480 ENFKVELQHFSDVDTIFNLILMPDKPIFLVFRPFNQDPLQ 519
```

# Important molecular regions conserved across species

Human (C11A_HUMAN; P05108) *vs.* zebrafish (Cyp11a1; Q8JH93)

```
Query: 34   TGEGAGISTRSPRPFNEIPSPGDNGWLNLYHFWRETGTHKVHLHHVQNFQKYGPIYREKL 93
            T  G       + +PFN+IP    N  L++  F +  G   VH   V NF+ +GPIYREK+
Sbjct: 27   TRSGRAPQNSTVQPFNKIPGRWRNSLLSVLAFTKMGGLRNVHRIMVHNFKTFGPIYREKV 86

Query: 94   GNVESVYVIDPEDVALLFKSEGPNPERFLIPPWVAYHQYYQRPIGVLLKKSAAWKKDRVA 153
            G  +SVY+I PED A+LFK+EG +P R  +  W AY  Y  +  GVLLK+  AWK DR+
Sbjct: 87   GIYDSVYIIKPEDGAILFKAEGHHPNRINVDAWTAYRDYRNQKYGVLLKEGKAWKTDRMI 146

Query: 154  LNQEVMAPEATKNFLPLLDAVSRDFVSVLHRRIKKAGSGNYSGDISDDLFRFAFESITNV 213
            LN+E++ P+    F+PLLD V +DFV+ ++++I+++G   ++ D++ DLFRF+ ES++ V
Sbjct: 147  LNKELLLPKLQGTFVPLLDEVGQDFVARVNKQIERSGQKQWTTDLTHDLFRFSLESVSAV 206

Query: 214  IFGERQGMLEEVVNPEAQRFIDAIYQMFHTSVPMLNLPPDLFRLFRTKTWKDHVAAWDVI 273
            ++GER G+L + ++PE Q FID +  MF T+ PML LPP L R   +  WK+HV AWD I
Sbjct: 207  LYGERLGLLLDNIDPEFQHFIDCVSVMFKTTSPMLYLPPGLLRSIGSNIWKNHVEAWDGI 266

Query: 274  FSKADIYTQNFYWELRQKGSVHHDYRGMLYRLLGDSKMSFEDIKANVTEMLAGGVDTTSM 333
            F++AD   QN + + ++    +  Y G+L  LL   K+S EDIKA+VTE++AGGVD+ +
Sbjct: 267  FNQADRCIQNIFKQWKENPEGNGKYPGVLAILLMQDKLSIEDIKASVTELMAGGVDSVTF 326

Query: 334  TLQWHLYEMARNLKVQDMLRAEVLAARHQAQGDMATMLQLVPLLKASIKETLRLHPISVT 393
            TL W LYE+AR   +QD LRAE+ AAR   +GDM  M++++PLLKA++KETLRLHP++++
Sbjct: 327  TLLWTLYELARQPDLQDELRAEISAARIAFKGDMVQMVKMIPLLKAALKETLRLHPVAMS 386

Query: 394  LQRYLVNDLVLRDYMIPAKTLVQVAIYALGREPTFFFDPENFDPTRWLSKDKNITYFRNL 453
            L RY+  D V+++Y IPA TLVQ+ +YA+GR+  FF  PE + P+RW+S ++   YF++L
Sbjct: 387  LPRYITEDTVIQNYHIPAGTLVQLGVYAMGRDHQFFPKPEQYCPSRWISSNRQ--YFKSL 444

Query: 454  GFGWGVRQCLGRRIAELEMTIFLINMLENFRVEIQHLSDVGTTFNLILMPEKPISFTFWP 513
            GFG+G RQCLGRRIAE EM IFLI+MLENFR+E Q   +V + F L+LMPEKPI  T  P
Sbjct: 445  GFGFGPRQCLGRRIAETEMQIFLIHMLENFRIEKQKQIEVRSKFELLLMPEKPIILTIKP 504

Query: 514  FN 515
             N
Sbjct: 505  LN 506
```

# Why compare sequences? Do more..

- Determine the evolutionary constraints at work
- Find mutations in a population or family of genes
- Find similar looking sequence in a database
- Find secondary/tertiary structure of a sequence of interest – molecular modeling using a template (homology modeling)

# How to compare sequences?

- We need to compare DNA or protein sequences, estimate their distances e.g., helpful for inferring molecular function by finding similarity with known function
- => we need 'good alignment' of sequences
- => we need: a measure of judging the quality of alignment in relation to other possible alignments, a scoring system.

# Sequence alignment

- Are two sequences related?
  - Align sequences or parts of them
  - Decide if alignment is by chance or evolutionarily linked?
- Issues:
  - What sorts of alignments to consider?
  - How to score an alignment and hence rank?
  - Algorithm to find good alignments
  - Evaluate the significance of the alignment

# Aligning two sequences

- **Problem:** Given two strings, S and T, find their "best" arrangement (global alignment) or the two largest substrings, s and t, with maximum similarity (local alignment).

- Aligned residue states:
  - Match - Mismatch - Gap (insertion/deletion)

- We need:
  - Scoring scheme for "similarity"
  - Alignment algorithm

# Global alignment

- Given two strings, *S* and *T*, find their relative alignment with the highest "score".

```
Seq. #1: G A A T T C A G T T A

Seq. #2: G G A T C G A
```

# Sequence comparison

- Look for evidence that the sequences had a common ancestor
- Mutations and selection cause divergence between sequences
- Mutational process: substitution
  - Change residues in a sequence
  - Insertions or deletions (indels) which add or remove residues
- Selection screens mutations, so that some substitutions occur more often than others

# Global alignment (cntd)

```
  G A A T T C A G T T A
  | |
G G A T C G A
```

2 matches, 4 mism., *0 gaps*

```
G A A T T C A G T T A
|   | |     |
G G A T C G A
```

4 matches, 3 mism., 0 gaps

```
G A A T T C - A G T T A
|   |   | |   |
G G A - T C G A
```

5 matches, 1 mism., 2 gaps

# Alignment: the problem (cntd)

Scoring schemes: three possible situations…

- Match
- Mismatch
- Gap
  - Linear
  - Convex
  - Affine

REWARD!!

Penalise

Penalise

How much??

# Sequence Alignment

- Methods to align
  - short genome segments
  - database of sequences
  - whole genomes and chromosomes
  - in the presence of large scale rearrangements
- Approximations for speed and storage

# Things to learn..

- Compare different scoring schemes
- Techniques for obtaining best scoring alignment of a given type
- Reduce computational burden
- Speed up database search techniques
- Evaluate approximations used in common database search programs
- Techniques for aligning DNA and protein sequences together
- Identify sequences of low complexity
- Identify significant alignments based on their score
- Techniques for aligning complex genomic sequences

# Circular logic in alignment and scoring

How do we now the what is the right distance without a good alignment?

How do we construct a good alignment without knowing what substitutions were made previously?

```
ATGCGT--GCAAGT
--GCGGCGGCAGTT
```

```
ATGCGT--GCAAGT-
--GCGGCGGC-AGTT
```

Less gaps but 7 matches

More gaps but 8 matches

Which is better?

# Substitution matrices for proteins

- We need to compare DNA or protein sequences, estimate their distances e.g., helpful for inferring molecular function by finding similarity with known function
- => we need 'good alignment' of sequences
- => we need: a measure of judging the quality of alignment in relation to other possible alignments, a scoring system.

# Additive Scoring Systems

- Look at each position of a given alignment, and give a score for the 'quality of the match'. Total/cumulative scores is the sum over all individual scores.
- e.g. two DNA sequences; match = +1; mismatch = -1; gap open = -3; gap extension = -1

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | a | a | g | t | t | t | c | - | - | t | t | g | |
| | a | a | a | c | t | c | c | t | c | t | t | g | |
| Base scores: | 1 | 1 | -1 | -1 | 1 | -1 | 1 | | | 1 | 1 | 1 | = 7 |
| Gap Penalty: | | | | | | | | -3 | -1 | | | | =-4 |

- Cumulative score = 7 – 4 = 3

# Scoring or substitution matrices

- Score (i,j) => AAs or nts (i,j) are aligned at any position

E.g.

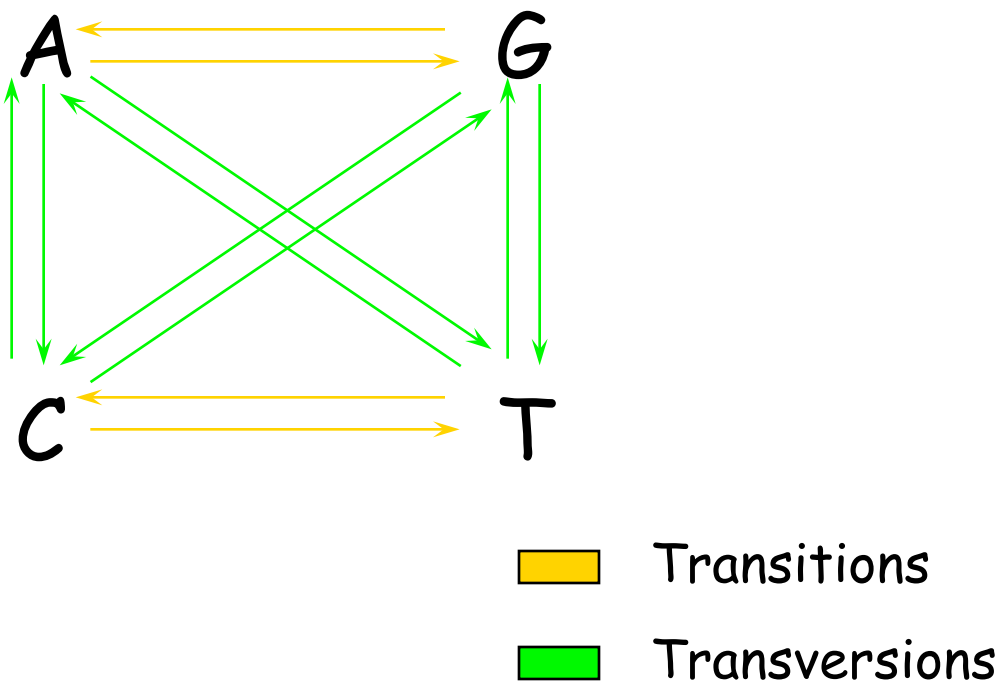
$$S = \begin{array}{c} \\ A \\ T \\ G \\ C \end{array} \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} s_{a,a} & s_{a,c} & s_{a,g} & s_{a,t} \\ s_{c,a} & s_{c,c} & s_{c,g} & s_{c,t} \\ s_{g,a} & s_{g,c} & s_{g,g} & s_{g,t} \\ s_{t,a} & s_{t,c} & s_{t,g} & s_{t,t} \end{pmatrix} \end{array} = \begin{array}{c} \begin{array}{cccc} A & T & G & C \end{array} \\ \begin{pmatrix} 1 & -1 & -1/2 & -1 \\ -1 & 1 & -1 & -1/2 \\ -1/2 & -1 & 1 & -1 \\ -1 & -1/2 & -1 & 1 \end{pmatrix} \end{array} \begin{array}{c} \\ A \\ T \\ G \\ C \end{array}$$

- DNA = 4 x 4
- Protein = 20 x 20

# Base mutations (general): definitions

Purines

Pyrimidines



# Nucleic acid Distance Matrices

## Nucleotide substitution matrices.

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

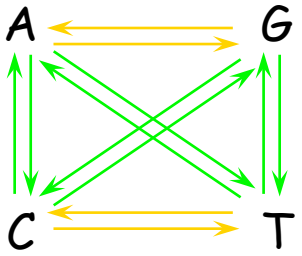
**Identity**

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

**BLAST**

| | A | T | C | G |
|---|---|---|---|---|
| A | 0 | 5 | 5 | 1 |
| T | 5 | 0 | 1 | 5 |
| C | 5 | 1 | 0 | 5 |
| G | 1 | 5 | 5 | 0 |

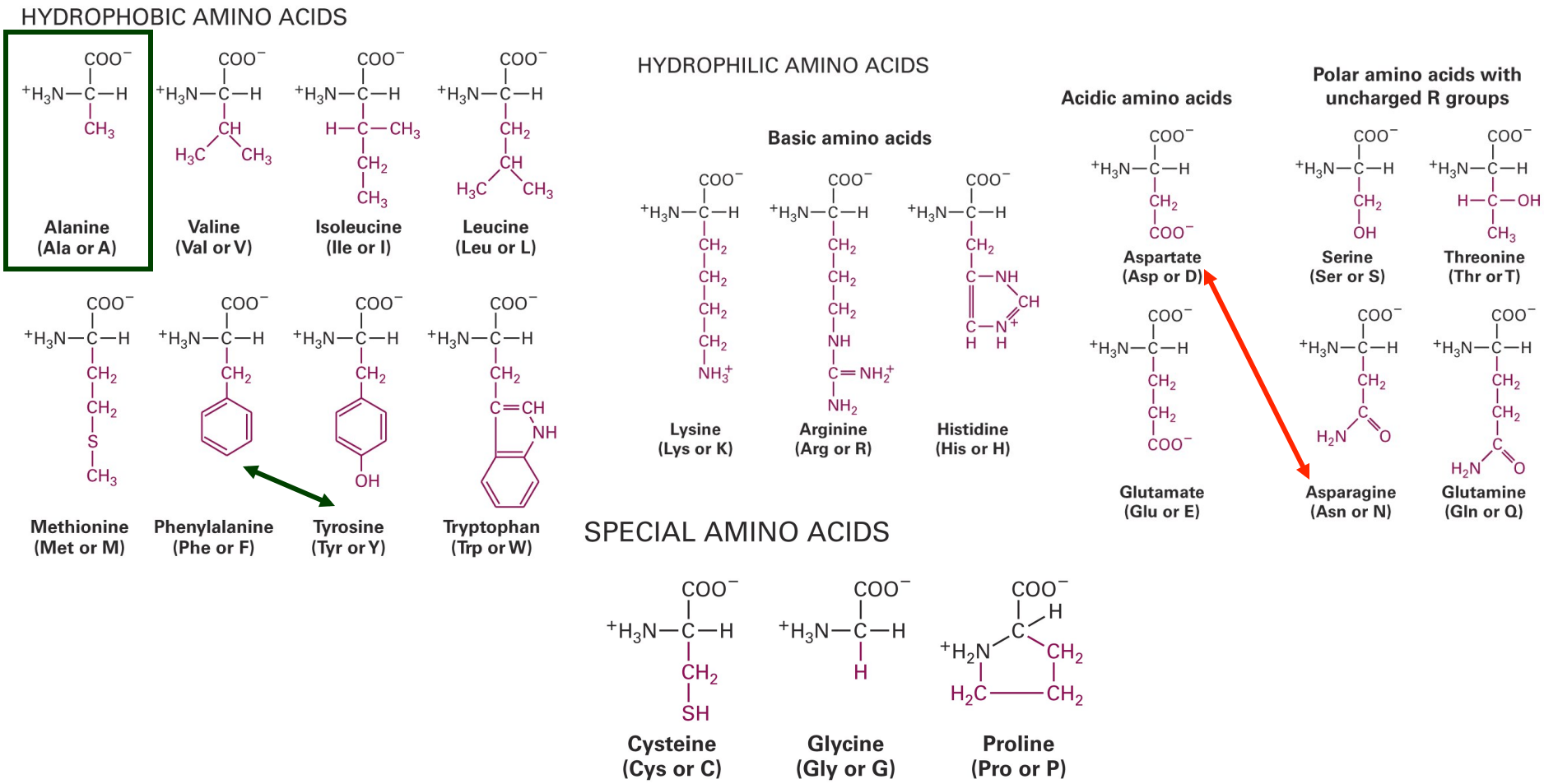
**Transition/ Transversion**



# Finding biologically meaningful scores

- DNA: simple scoring matrices are often effective
  - usually don't study very diverged DNA sequences
- Proteins: some substitutions are more likely to occur than others because chemical props are similar e.g., isoleucine for valine, serine for threonine (*conservative substitutions*)
  - To get better alignments, use scoring matrices derived from statistical analysis of protein data

# Conservative substitutions




*Slide courtesy: Serafim Batzoglou*

# Specs for deriving biologically meaningful scores

- Identically aligned AAs should have greater score than any substitution
- Conservative subs score > non-conservatives
- Scores should reflect evolutionary distances
  - Mouse and Rat => very similar sequences
  - Mouse and Yeast => divergent sequences
  - => scoring matrices are a function of evolutionary distance between sequences

# Sample substitution matrix

```
134 LQQGELDLVMTSDILPRSELHYSPMFDFEVRLVLAPDHPLASKTQITPEDLASETLLI
    |      |||           |         |          ||||||     |     || ||
137 LDSNSVDLVLMGVPPRNVEVEAEAFMDNPLVVIAPPDHPLAGERAISLARLAEETFVM
```

D:D = +6

D:R = -2

BLOSUM62

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

# Two frequently used matrices

- PAM (point accepted mutation) family
  - Markov chains and phylogenetic trees
    *for fitting evolutionary model*
  - Log-likelihood ratios
    *for getting a score from an estimated transition matrix*

- BLOSUM family
  - Log-likelihood ratios
    *for getting a score from an estimated transition matrix*

# 1. log-odds scoring

- What are the odds that an alignment is biologically meaningful?
- Random model: product of chance events
- Non-random model: two sequences derived from a common ancestor

# 2. log-odds scoring

What are the odds that this alignment is meaningful?

$$X_1X_2X_3 \ldots X_n$$
$$Y_1Y_2Y_3 \ldots Y_n$$

**Random model:** We're observing a chance event. The probability is

$$\prod_i p_{X_i} \prod_i p_{Y_i}$$

where $p_X$ is the frequency of $X$

**Alternative:** The two sequences derive from a common ancestor. The probability is

$$\prod_i q_{X_iY_i}$$

where $q_{XY}$ is the joint probability that $X$ and $Y$ evolved from the same ancestor.

# 3. log-odds scoring

**Odds ratio:**

$$\frac{\prod_i q_{X_i Y_i}}{\prod_i p_{X_i} \prod_i p_{Y_i}} = \prod_i \frac{q_{X_i Y_i}}{p_{X_i} p_{Y_i}}$$

**Log-odds ratio (score):** $S = \sum_i s(X_i, Y_i)$

**where** $s(X,Y) = \log\left(\frac{q_{XY}}{p_X p_Y}\right)$

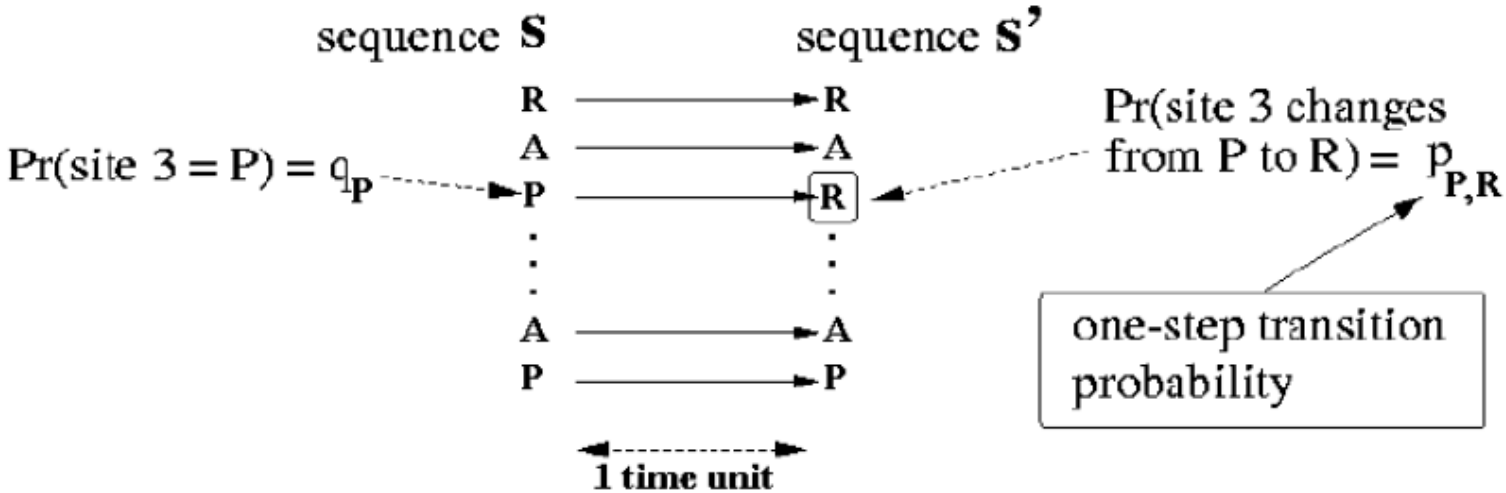
**is the score for *X*, *Y*. The *s*(*X*,*Y*)'s define a scoring matrix**

# Point/Percent Accepted Mutations (PAM)

- PAM Markov transition matrix
  - Table of estimated transition probabilities for the underlying evolutionary model
- PAM substitution matrix
  - Table of scores for all possible pairs of AAs

# PAM Model

- Each site evolves as a Markov chain
- Independent of others
- All Markov chains have the same transition matrix
- Dayhoff et al (1978) estimated one-step
- transitions



# PAM1 transition matrix

- Expect 1% of the AAs to undergo accepted point mutations.
  - Align protein sequences that are at least 85% identical
  - Reconstruct phylogenetic trees and infer ancestral sequences
  - Count AA replacements that occur along the tree..i.e. count mutations accepted by natural selection
  - Use the counts to estimate probabilities for replacements

First, for any pair (j, k) define

$$a_{jk} = \frac{C_{j \to k}}{\sum_{m=1..20} C_{j \to m}}$$

This is the observed relative frequency for the substitution j $\rightarrow$ k.

The $a_{j,k}$'s are estimated probabilities.

*These probabilities were then scaled for calculating 1 PAM probabilities:*

For j $\neq$ k,

$$p_{j,k} = c \, . a_{j,k}$$

and

$$p_{j,j} = 1 - \sum_{k=1..20, k \neq j} p_{j,k}$$

... but why the scaling factor c?

**We want to choose a value of c, such that 1% of the amino acids are expected to undergo accepted point mutations during one time unit.**

**Such a time unit is called an evolutionary distance of 1 PAM.**

To determine $c$, it *suffices to consider one of the sites* in the sequence, i.e. we consider only one of the parallel Markov chains.

Let $Z_n =$ *the amino acid present at the site considered at time* $n$, $n \geq 0$. (hence $1 \leq Z_n \leq 20$, since the AA's are coded as 1 to 20).

**The probability that the site will change after 1 PAM time unit** (i.e. after one step) is given by

$$\mathbf{P}(Z_1 \neq Z_0) = \sum_{j=1}^{20} \mathbf{P}(Z_0 = j, Z_1 \neq j)$$

$$= \sum_{j=1}^{20} \mathbf{P}(Z_1 \neq j | Z_0 = j) \cdot \mathbf{P}(Z_0 = j) \approx \sum_{j=1}^{20} \mathbf{P}(Z_1 \neq j | Z_0 = j) \cdot q_j,$$

where $q_j$ is the *observed frequency of the amino acid no.* $j$ in the original blocks of aligned proteins.

One wants the probability that the site will change after 1 PAM to be equal to 0.01. (*That implies an average change of* 1%.)

$$
\begin{aligned}
0.01 &= \sum_{j=1}^{20} \mathbf{P}(Z_1 \neq j | Z_0 = j) \cdot q_j \\
&= \sum_{j=1}^{20} \Big( \sum_{k \neq j} \mathbf{P}(Z_1 = k | Z_0 = j) \Big) \cdot q_j \\
&\approx \sum_{j=1}^{20} \Big( \sum_{k \neq j} p_{j,k} \Big) \cdot q_j \\
&= \sum_{j=1}^{20} \Big( \sum_{k \neq j} c \cdot a_{j,k} \Big) \cdot q_j \\
&= c \cdot \sum_{j=1}^{20} \sum_{k \neq j} q_j \cdot a_{j,k}.
\end{aligned}
$$

That is, we want

$$0.01 = c \cdot \sum_{j=1}^{20} \sum_{k \neq j} q_j \cdot a_{j,k}.$$

Therefore, using the estimated probabilities $q_j$ and $a_{j,k}$, just put

$$c = \frac{0.01}{\sum_{j=1}^{20} \sum_{k \neq j} q_j \cdot a_{j,k}}.$$

---

Thus, with this choice for $c$, the *PAM transition matrix* is obtained ('one-step', i.e. for the evolutionary distance of 1 PAM) .

*How can this transition matrix be turned into a scoring matrix?*

# From transition matrix to scores

Consider two given protein sequences $\boldsymbol{s} = a_1 a_2 \cdots a_n$ and $\boldsymbol{s'} = b_1 b_2 \cdots b_n$ (at a evolutionary distance of 1 PAM, say).

The score for aligning $\boldsymbol{s}$ with $\boldsymbol{s'}$ is generated *by comparing two different hypothesis* $H_0$ and $H_A$:

- $H_0$: $\boldsymbol{s}$ *and* $\boldsymbol{s'}$ *<u>are not</u> evolutionarily related (i.e. a chance alignment).*
- $H_A$: $\boldsymbol{s}$ *and* $\boldsymbol{s'}$ *<u>are</u> evolutionarily related (i.e.* $\boldsymbol{s'}$ *depends on* $\boldsymbol{s}$ *via the Markov model).*

*Under $H_0$*, we have a chance alignment

$$\begin{aligned} \boldsymbol{s}: &\quad a_1 a_2 \cdots a_n \\ \boldsymbol{s'}: &\quad b_1\, b_2 \cdots b_n \end{aligned}$$

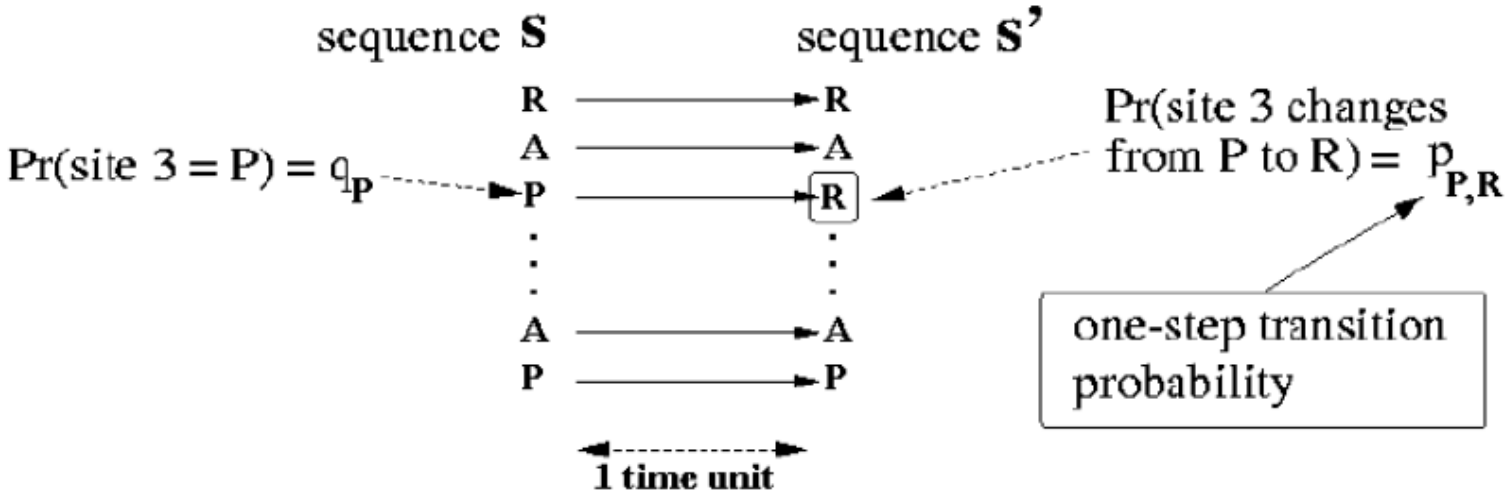
That is, all sites in both sequences are randomly generated, all sites independent of each other.

Amino acid $j$ appears with probability $q_j$.

The probability for getting this *chance* alignment is equal to

$$\begin{aligned} \mathbf{P}_{H_0}(\textit{the alignment}) &= \Big(\prod_{i=1}^{n} q_{a_i}\Big) \cdot \Big(\prod_{i=1}^{n} q_{b_i}\Big) \\ &= \prod_{i=1}^{n} (q_{a_i} \cdot q_{b_i}). \end{aligned}$$

*Under $H_A$*, the sites in the sequences are dependent, according to the Markov model described earlier.




Example: $\mathbf{P}_{H_A}(\textit{align } P \textit{ and } R \textit{ in a given site}) = q_P \cdot p_{P,R}$.

Since the different sites evolve independently of each other, we get

$$\mathbf{P}_{H_A}(\textit{the alignment}) = \prod_{i=1}^{n} (q_{a_i} \cdot p_{a_i,b_i}).$$

In principle, we want our score to reflect the 'chance' (or the *odds*) that with $\boldsymbol{s}$ and $\boldsymbol{s'}$ we have aligned evolutionarily related sequences (i.e. basically we want a high score if the odds are high that we have aligned related sequences).

A natural choice for the score is then a comparison of the probabilities under $H_A$ and $H_0$, respectively:

The **likelihood ratio**:

$$\begin{aligned}
\text{alignment score} &= \frac{\mathbf{P}_{H_A}(\textit{the alignment})}{\mathbf{P}_{H_0}(\textit{the alignment})} \\
&= \frac{\prod_{i=1}^{n}(q_{a_i} \cdot p_{a_i,b_i})}{\prod_{i=1}^{n}(q_{a_i} \cdot q_{b_i})} \\
&= \prod_{i=1}^{n} \frac{q_{a_i} \cdot p_{a_i,b_i}}{q_{a_i} \cdot q_{b_i}} = \prod_{i=1}^{n} \frac{p_{a_i,b_i}}{q_{b_i}}.
\end{aligned}$$

Or, equivalently, but better for theoretical reasons, one can use the **log likelihood ratio** (Dayhoff et al.: "the **log odds ratio**"):

$$\text{alignment score} = \log\left(\frac{\mathbf{P}_{H_A}(\textit{the alignment})}{\mathbf{P}_{H_0}(\textit{the alignment})}\right)$$

$$= \log\left(\prod_{i=1}^{n} \frac{p_{a_i,b_i}}{q_{b_i}}\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{p_{a_i,b_i}}{q_{b_i}}\right).$$

The entry $(a, b)$ in the **PAM substitution matrix** is then of the form

$$S_{a,b} = \log\left(\frac{p_{a,b}}{q_b}\right)$$

(or rounded to the nearest integer for convenience).

**Commonly multiplied by a power of 10 to deal with decimals**

Due to the logarithm, we have obtained an *additive* scoring system in a natural way:

alignment:

$$\begin{array}{ll} \boldsymbol{s}: & a_1 a_2 \cdots a_n \\ \boldsymbol{s'}: & b_1\, b_2 \cdots b_n \end{array}$$

$$\boxed{\text{Total score: } S(alignment) = \sum_{i=1}^{n} S_{a_i,b_i}.}$$

***

*Adding the scores* for each position is equivalent to *multiplying the probabilities* (due to the logarithm)!

$$\begin{aligned} S(alignment) &= \log\Big(\frac{\mathbf{P}_{H_A}(the\ alignment)}{\mathbf{P}_{H_0}(the\ alignment)}\Big) \\ &= \log\Big(\frac{(q_{a_1}\cdot p_{a_1,b_1})\cdot(q_{a_2}\cdot p_{a_2,b_2})\cdots(q_{a_n}\cdot p_{a_n,b_n})}{(q_{a_1}\cdot q_{b_1})\cdot(q_{a_2}\cdot q_{b_2})\cdots(q_{a_n}\cdot q_{b_n})}\Big) \\ &= \sum_{i=1}^{n}\log\Big(\frac{p_{a_i,b_i}}{q_{b_i}}\Big) = \sum_{i=1}^{n} S_{a_i,b_i}. \end{aligned}$$

Moreover,

$$S_{a,b} = \log\left(\frac{p_{a,b}}{q_b}\right) < 0$$

if

$$\frac{p_{a,b}}{q_b} < 1 \iff \frac{q_a \cdot p_{a,b}}{q_a \cdot q_b} < 1 \iff q_a \cdot p_{a,b} < q_a \cdot q_b$$

(*i.e.* $S_{a,b} < 0$ *if it is more likely to see* $a$ *and* $b$ *aligned against each other in a random alignment than to see* $a$ *and* $b$ *aligned in a comparison of two related sequences (at 1 PAM distance)*).

Otherwise,

$$S_{a,b} = \log\left(\frac{p_{a,b}}{q_b}\right) \geq 0.$$

# PAM*n* substitution

For sequences having an evolutionary distance of $n$ PAM units.

*Careful: "n PAM units" does not mean that we expect n% of the amino acids to differ... because substitutions can occur at the same site many times!!*

Let $P$ be the 1 PAM transition matrix. As always with Markov chains: the $n$-step transition probabilities $p_{a,b}^{(n)}$ are given as the entries in

$$P^n.$$

The scores are

$$S_{a,b}^{(n)} = \log\left(\frac{p_{a,b}^{(n)}}{q_b}\right).$$

# PAM Matrices

- Family of matrices PAM 80, PAM 120, PAM 250
- The number at the end indicates the evolutionary distance between the sequences on which the matrix is based
- Greater numbers denote greater distances
  - PAM250 is for more distant proteins than PAM80

# PAM250

Table 1 – The log odds matrix for 250 PAMs (multiplied by 10)

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

# PAM

## Assumptions of the PAM model:

- Replacement at any site depends only on the a.a. on that site, given the mutability table.

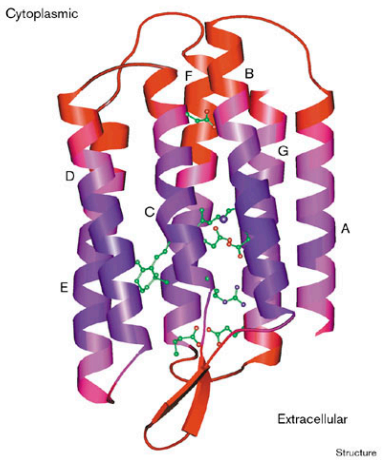
- Sequences in the training set (and those compared) have average a.a. composition.

# Sources of error in PAM

- Many proteins depart from the average a.a. composition.

- The a.a. composition can vary even within a protein (e.g. transmembrane proteins).

- A.a. positions are not “mutated” equally probably; especially in long evolutionary distances.



# Sources of error in PAM (cntd)

- A.a. residues are not equally prone to mutations.
- Rare replacements are observed too infrequently and…
- …errors in PAM1 are magnified in PAM250.

# AA matrices: BLOSUM

**Blo**cks **Su**bstitution **M**atrices (BLOSUM):

- Log-likelihood matrix (Henikoff & Henikoff, 1992)
- BLOCKS database of aligned sequences used as primary source set.

# BLOSUM matrices

- Different BOLSUM*n* matrices are calculated independently from BLOCKS (ungapped local alignments)
- BLOSUM*n* is based on a cluster of BLOCKS of sequences that share at least *n* percent identity
- BLOSUM62 represents closer sequences than BLOSUM45

# BLOSUM matrices designed to find conserved regions of proteins

- BLOCKS database contains large number of ungapped multiple local alignments of conserved regions of proteins
- Alignments include distantly related sequences in which multiple base substitutions at the same position could be observed

- BLOCKS alignments used to derive BLOSUM matrices include sequences that are much less similar to each other than those used by Dayhoff, but whole evolutionary homology can be confirmed through intermediate sequences
- Alignments created without phylogenetic tree

- Direct comparison of aligned residues does not model real substitutions, because the sequences have evolved from a common ancestor and not from each other.
- Unfortunately, large variation in sequences prevents tree construction..
- If the alignment is correct, aligned residues will be related by their evolutionary history and the alignment is expected to contain useful info on substitution preferences.
- Also, direct sequence comparison can often be for testing “significance” of similarity

# BLOSUM50 scoring matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | J | Z | X | * |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 5 | -2 | -1 | -2 | -1 | -1 | -1 | 0 | -2 | -1 | -2 | -1 | -1 | -3 | -1 | 1 | 0 | -3 | -2 | 0 | -2 | -2 | -1 | -1 | -5 |
| R | -2 | 7 | -1 | -2 | -4 | 1 | 0 | -3 | 0 | -4 | -3 | 3 | -2 | -3 | -3 | -1 | -1 | -3 | -1 | -3 | -1 | -3 | 0 | -1 | -5 |
| N | -1 | -1 | 7 | 2 | -2 | 0 | 0 | 0 | 1 | -3 | -4 | 0 | -2 | -4 | -2 | 1 | 0 | -4 | -2 | -3 | 5 | -4 | 0 | -1 | -5 |
| D | -2 | -2 | 2 | 8 | -4 | 0 | 2 | -1 | -1 | -4 | -4 | -1 | -4 | -5 | -1 | 0 | -1 | -5 | -3 | -4 | 6 | -4 | 1 | -1 | -5 |
| C | -1 | -4 | -2 | -4 | 13 | -3 | -3 | -3 | -3 | -2 | -2 | -3 | -2 | -2 | -4 | -1 | -1 | -5 | -3 | -1 | -3 | -2 | -3 | -1 | -5 |
| Q | -1 | 1 | 0 | 0 | -3 | 7 | 2 | -2 | 1 | -3 | -2 | 2 | 0 | -4 | -1 | 0 | -1 | -1 | -1 | -3 | 0 | -3 | 4 | -1 | -5 |
| E | -1 | 0 | 0 | 2 | -3 | 2 | 6 | -3 | 0 | -4 | -3 | 1 | -2 | -3 | -1 | -1 | -1 | -3 | -2 | -3 | 1 | -3 | 5 | -1 | -5 |
| G | 0 | -3 | 0 | -1 | -3 | -2 | -3 | 8 | -2 | -4 | -4 | -2 | -3 | -4 | -2 | 0 | -2 | -3 | -3 | -4 | -1 | -4 | -2 | -1 | -5 |
| H | -2 | 0 | 1 | -1 | -3 | 1 | 0 | -2 | 10 | -4 | -3 | 0 | -1 | -1 | -2 | -1 | -2 | -3 | 2 | -4 | 0 | -3 | 0 | -1 | -5 |
| I | -1 | -4 | -3 | -4 | -2 | -3 | -4 | -4 | -4 | 5 | 2 | -3 | 2 | 0 | -3 | -3 | -1 | -3 | -1 | 4 | -4 | 4 | -3 | -1 | -5 |
| L | -2 | -3 | -4 | -4 | -2 | -2 | -3 | -4 | -3 | 2 | 5 | -3 | 3 | 1 | -4 | -3 | -1 | -2 | -1 | 1 | -4 | 4 | -3 | -1 | -5 |
| K | -1 | 3 | 0 | -1 | -3 | 2 | 1 | -2 | 0 | -3 | -3 | 6 | -2 | -4 | -1 | 0 | -1 | -3 | -2 | -3 | 0 | -3 | 1 | -1 | -5 |
| M | -1 | -2 | -2 | -4 | -2 | 0 | -2 | -3 | -1 | 2 | 3 | -2 | 7 | 0 | -3 | -2 | -1 | -1 | 0 | 1 | -3 | 2 | -1 | -1 | -5 |
| F | -3 | -3 | -4 | -5 | -2 | -4 | -3 | -4 | -1 | 0 | 1 | -4 | 0 | 8 | -4 | -3 | -2 | 1 | 4 | -1 | -4 | 1 | -4 | -1 | -5 |
| P | -1 | -3 | -2 | -1 | -4 | -1 | -1 | -2 | -2 | -3 | -4 | -1 | -3 | -4 | 10 | -1 | -1 | -4 | -3 | -3 | -2 | -3 | -1 | -1 | -5 |
| S | 1 | -1 | 1 | 0 | -1 | 0 | -1 | 0 | -1 | -3 | -3 | 0 | -2 | -3 | -1 | 5 | 2 | -4 | -2 | -2 | 0 | -3 | 0 | -1 | -5 |
| T | 0 | -1 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -1 | -1 | -1 | -1 | -2 | -1 | 2 | 5 | -3 | -2 | 0 | 0 | -1 | -1 | -1 | -5 |
| W | -3 | -3 | -4 | -5 | -5 | -1 | -3 | -3 | -3 | -3 | -2 | -3 | -1 | 1 | -4 | -4 | -3 | 15 | 2 | -3 | -5 | -2 | -2 | -1 | -5 |
| Y | -2 | -1 | -2 | -3 | -3 | -1 | -2 | -3 | 2 | -1 | -1 | -2 | 0 | 4 | -3 | -2 | -2 | 2 | 8 | -1 | -3 | -1 | -2 | -1 | -5 |
| V | 0 | -3 | -3 | -4 | -1 | -3 | -3 | -4 | -4 | 4 | 1 | -3 | 1 | -1 | -3 | -2 | 0 | -3 | -1 | 5 | -3 | 2 | -3 | -1 | -5 |
| B | -2 | -1 | 5 | 6 | -3 | 0 | 1 | -1 | 0 | -4 | -4 | 0 | -3 | -4 | -2 | 0 | 0 | -5 | -3 | -3 | 6 | -4 | 1 | -1 | -5 |
| J | -2 | -3 | -4 | -4 | -2 | -3 | -3 | -4 | -3 | 4 | 4 | -3 | 2 | 1 | -3 | -3 | -1 | -2 | -1 | 2 | -4 | 4 | -3 | -1 | -5 |
| Z | -1 | 0 | 0 | 1 | -3 | 4 | 5 | -2 | 0 | -3 | -3 | 1 | -1 | -4 | -1 | 0 | -1 | -2 | -2 | -3 | 1 | -3 | 5 | -1 | -5 |
| X | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -5 |
| * | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |

# Substitution matrices: a comparison

## PAM *vs* BLOSUM

- PAM is based on closely related sequences, thus is biased for short evolutionary distances where number of mutations are scalable

- PAM is based on globally aligned sequences, thus includes conserved and non-conserved positions; BLOSUM is based on conserved positions only

# Substitution matrices: a comparison (cntd)

## PAM *vs* BLOSUM *(cntd)*

- Lower PAM/higher BLOSUM matrices identify shorter local alignments of highly similar sequences

- Higher PAM/lower BLOSUM matrices identify longer local alignments of more distant sequences

# Substitution matrices: a comparison (cntd)

## PAM *vs* BLOSUM

| BLOSUM 80 | BLOSUM 62 | BLOSUM 45 |
|---|---|---|
| PAM 1 | PAM 120 | PAM 250 |
| *Less divergent* ← | | → *More divergent* |

- Matrices of choice:
  - BLOSUM62: the all-weather matrix
  - PAM250: for distant relatives