Sequence Analysis 1

A. Concepts, conservation & substitution

Cheong Xin Chan (CX)

c.chan1@uq.edu.au

Australian Centre for Ecogenomics
School of Chemistry & Molecular Biosciences
The University of Queensland

SCIE2100 | BINF6000 | Bioinformatics 1 – Introduction

Lecture outline

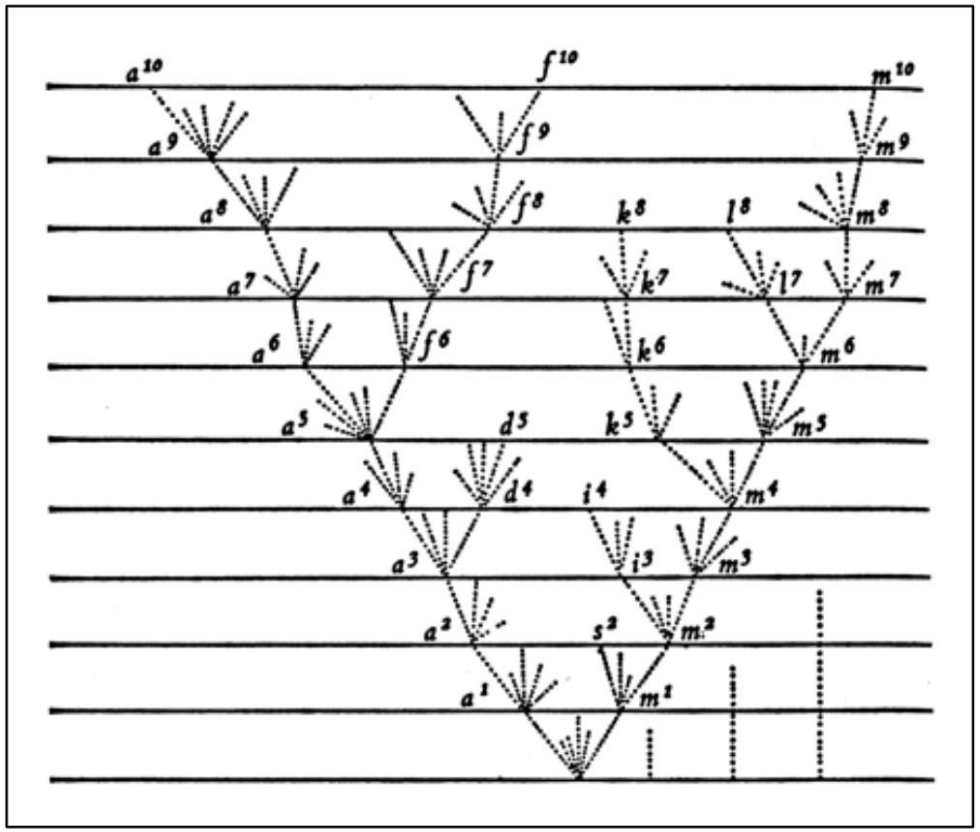
- **Concept of homology**
  - Homology and sequence similarity
- **Sequence change in evolution**
  - Random change versus biological evolution
  - Sequence change versus conservation
- **Quantifying sequence change**
  - Observed versus null probabilistic models
  - Log-odds score
  - Basic principles of PAM and BLOSUM matrices, and their differences

Concept of homology

Organisms related by genealogical descent with modification

Traits (and genes) are passed on from one generation to the next (**inheritance**)

Charles Darwin (1859). *On the Origin of Species*



time

Features derived from a common ancestor are said to be **homologous**. This applies to any feature – morphological or molecular (genes, RNAs, proteins).

Homology and sequence similarity

- **Sequence similarity (*or* sequence identity):** a measure of identical residues shared between two sequences, usually in an alignment (where identical/similar residues were aligned); the extent to which two sequences are **invariant**

Example:

```
Seq1  ACGTAGCTAGCTAGCTACCT
Seq2  ACGTAGCTAGCTAGCTAGCT
```

%identity between Seq1 & Seq2 **= 19/20 = 95%**

- High level of sequence similarity **usually**, but **not necessarily**, indicates evidence for homology
- Similar sequences may be homologs

Homology and sequence similarity

- **Conservation:** unchanged/invariant positions when comparing two sequences. At amino acid level, this can also refer to changes at an amino acid position that preserves the physicochemical property

Phenylalanine (**F**) Valine (**V**) are hydrophobic

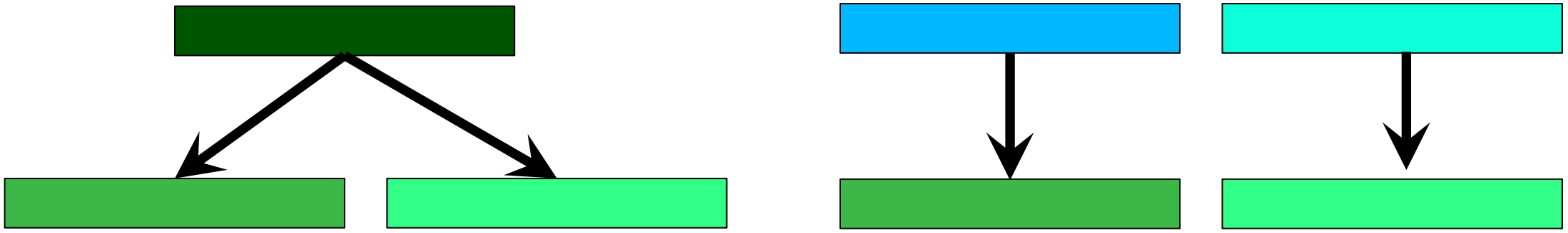
```
Seq1 ANYWQFPDGIYYEGCS
Seq2 ANYWQVPDGIHYEGCS
Seq3 ANYWQFPDGIYYEGCS
     *****:*****:******
```

Tyrosine (**Y**) and Histidine (**H**) are polar

Homology is an evolutionary relationship that either exists or not (i.e. it is **all-or-nothing**, there are no “degrees of homology”). We may be able to quantify how confident we are in believing that two molecules/sequences are homologous, but they are nonetheless either homologous or non-homologous.

Sequence change in evolution

Two alternative explanations to observed sequence similarity:



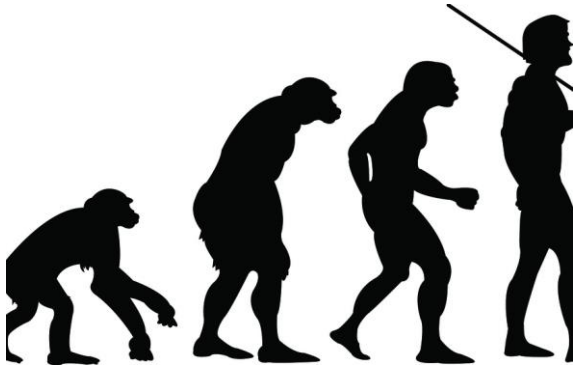
Evolution

similarity is due to shared ancestry

Random

similarity is coincidental by pure chance

Compare the probabilities of the two hypotheses:



$P_{evolution}$ *versus* P_{random}

what we observed — the null model

Random sequence change

What are the chances of randomly matching 2 basepairs (bp) in a human-size genome (3 x 10^9 bp)?

Total genome size

$$\frac{3{,}000{,}000{,}000}{4^2} = 187{,}500{,}000$$

number of possible matches

4 possible nucleotides

Length of matched region

$$\frac{187{,}500{,}000}{3{,}000{,}000{,}000} \times 100 = 6.25\%$$

Probability of a match to occur at random

Random sequence change

*What are the chances of randomly matching **N** basepairs (bp) in a human-size genome (3 x 10^9 bp)?*

Length of match in bp (N)	**# possible matches ($3 \times 10^9 / 4^N$)**	**Probability (P_{random})**
1	750,000,000	25.0 %
2	187,500,000	6.25 %
10	2,861	9.5 x 10^{-5} %
20	2.73 x 10^{-3}	9.1 x 10^{-11} %
300	7.23 x 10^{-172}	2.4 x 10^{-179} %

Random change or biological evolution?

Chimpanzee genome size ≈ human genome size ≈ **3Gbp**.
Consider a **300bp**-region in chimpanzee and in human:

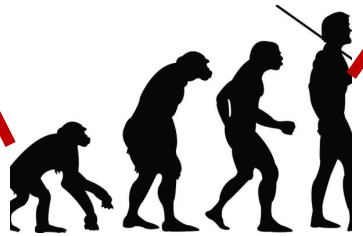
%identity
= $296/300 \times 100\%$
= **98.67%**

```
GGCTGTCATCACTTAGACCTCACCCTGTGG
AGCCACACCCTAGGGTTGGCCAATCTACTC
CCAGGAGCAGGGAGGGCAGGAGCCAGGGCT
GGGCATAAAAGTCAGGGCAGAGCCATCTAT
TGCTTACATTTGCTTCTGACACAACTGTGT
TCACTAGCAACCTCAAACAGACACCATGGT
GCATCTGACTCCTGAGGAGAAGTCTGCCGT
TACTGCCCTGTGGGGCAAGGTGAACGTGGA
TGAAGTTGGTGGTGAGGCCCTGGGCAGGTT
GGTATCAAGGTTACAAGACAGGTTTAAGGA
```

```
GGCTGTCATCACTTAGACCTCACCCTGTGG
AGCCACACCCTAGGGTTGGCCAATCTACTC
CCAGGAGCAGGGAGGGCAGGAGCCAGGGCT
GGGCATAAAAGTCAGGGCAGAGCCATCTAT
TGCTTACATTTGCTTCTGACACAACTGTGT
TCACTAGCAACCTCAAACAGACACCATGGT
ACCGCTGACTCCTGAGGAGAAGTCTGCCGT
TACTGCCCTGTGGGGCAAGGTGAACGTGGA
TGAAGTTGGTGGTGAGGCCCTGGGCAGGTT
GGTATCAAGGTTACAAGACAGGCTTAAGGA
```

Example

P_{random} of 296bp identical match in a 3Gbp genome
= 6.17×10^{-177} %



Do you think human and chimpanzee share a common ancestry?

Sequence change *versus* conservation

DNA/RNA sequence
↓
Protein sequence
↓
Tertiary structure
↓
Function

- Selective pressure for **divergence** e.g. genetic diversity (to increase viability), e.g. fast-evolving genes
- Substitutions, insertions, deletions, translocations, genetic transfers/exchange, etc.
- **Adaptation** due to changes in environments (biotic and abiotic stressors)
- Copy/replication errors

- Selective pressure for **preservation of critical gene function**, protein structure and function, including non-coding sequences and regulatory elements
- Especially true for critical machineries and slow-evolving genes e.g. housekeeping genes, ribosomal RNA genes (phylogenetic markers)

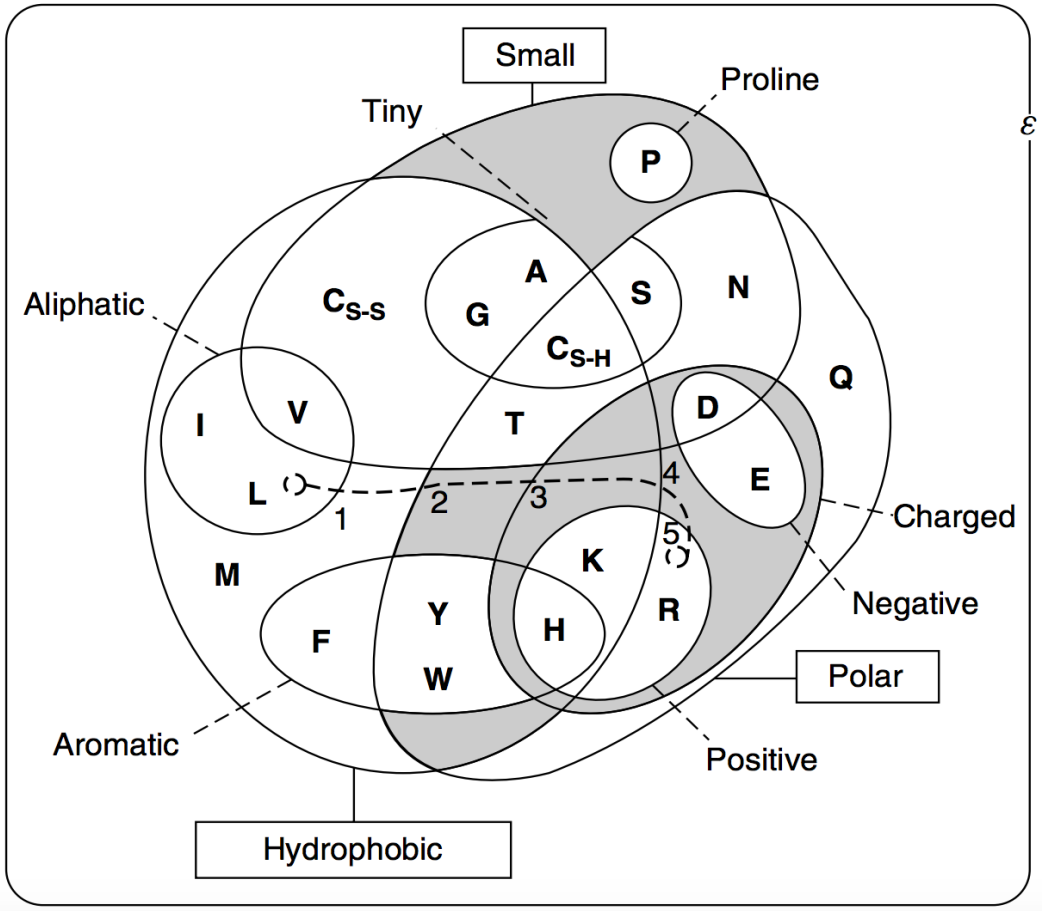
Quantifying sequence change

Homology and sequence conservation is commonly observed at the **protein** level. *Why?*

- **Codon degeneracy**: nearly one-third of the bases in coding regions are under a weak (if any) selection
- **Greater information content**: 20 amino acids versus 4 bases

Among a set of homologous sequences: **observed substitution frequency** for each amino acid can be used to quantify sequence change

- Commonly weighted based on shared **physicochemical properties** of amino acids

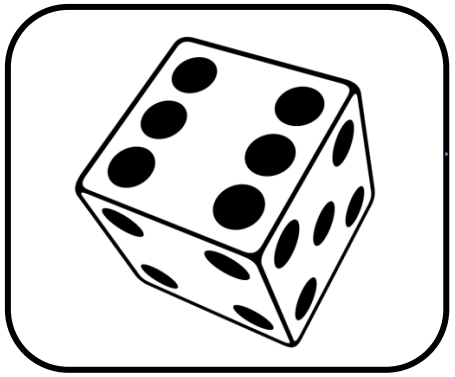


Betts & Russell (2003) Chapter 14. *Bioinformatics for Geneticists*.In: Barnes MR & Cray IC (Eds). John Wiley & Sons. Chapter 13 in the 2nd Edition (QH430 .B375 2007)

Probability in an unbiased null model

On a die:

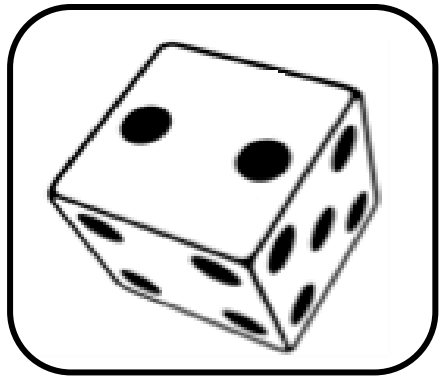
Landing a **6**?



1/6

P(‘6’) = 1/6

Landing a **6** then a **2**?



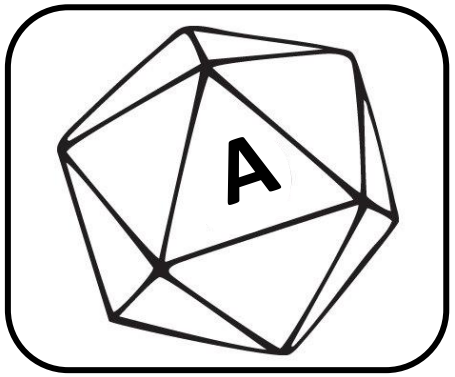
1/36

P(‘6’ ∧ ‘2’) = 1/6 × 1/6
= 1/36

Probability in an unbiased null model

Imagine a 20-face die (each face represents an amino acid)

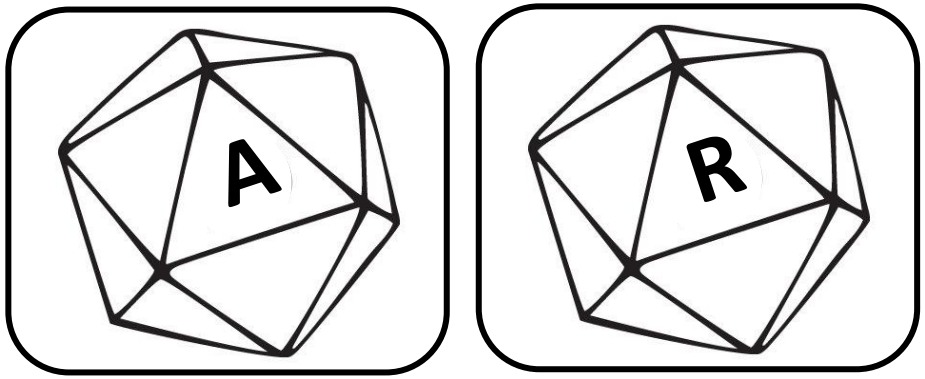
Observing an **A**?
(Alanine)



1/20

$$P(\text{'A'}) = p_A = 1/20$$

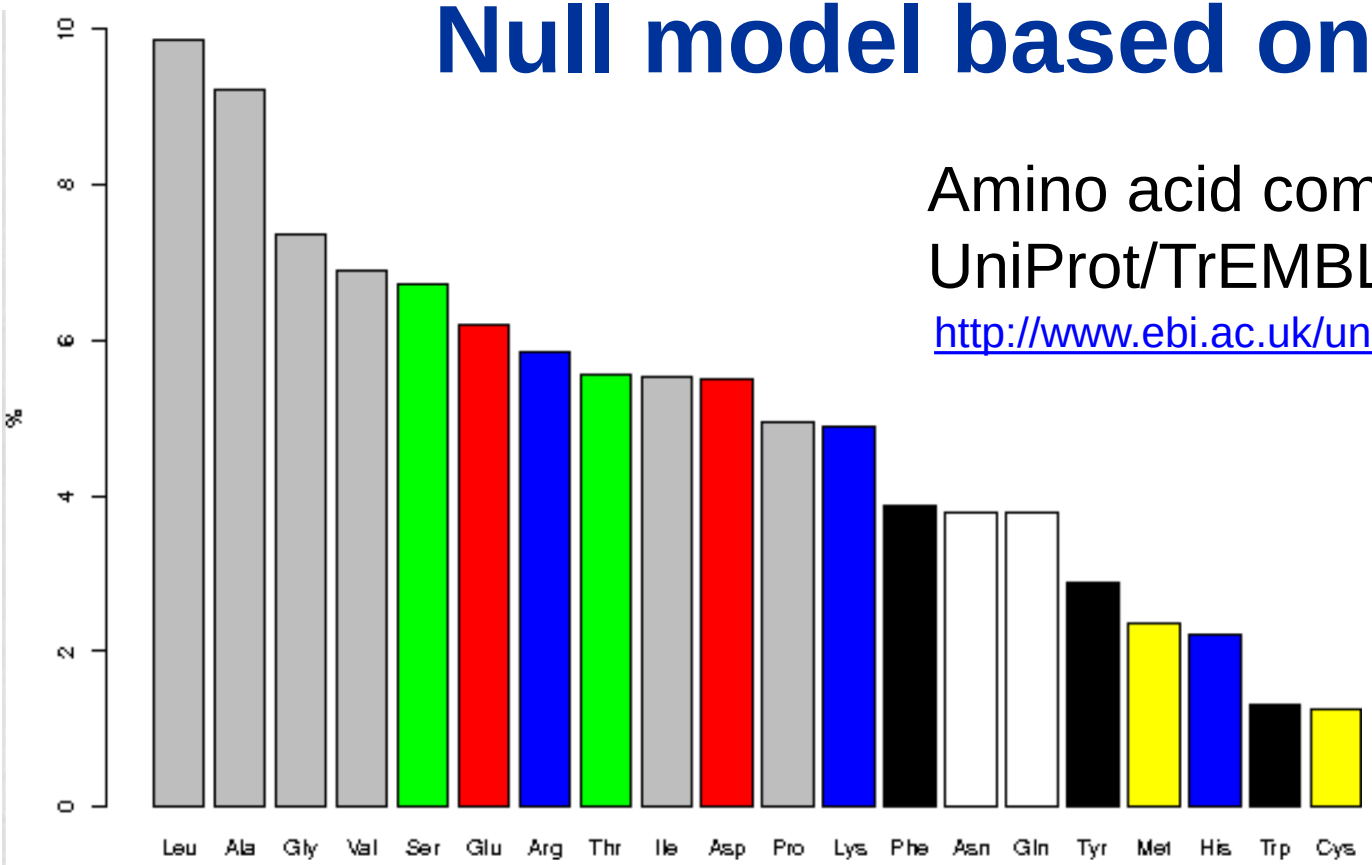
Observing **A** in one but **R** in another?
(Arginine)



1/400

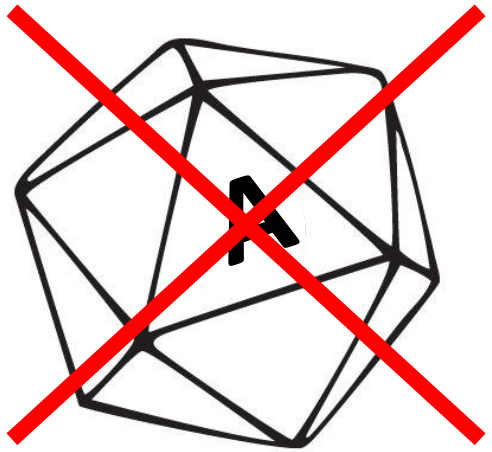
$$P(\text{'A'} \wedge \text{'R'}) = p_A \times p_R = 1/20 \times 1/20$$
$$= 1/400$$

Null model based on empirical data



Amino acid composition (%) in UniProt/TrEMBL (Feb 2021)

http://www.ebi.ac.uk/uniprot/TrEMBLstats



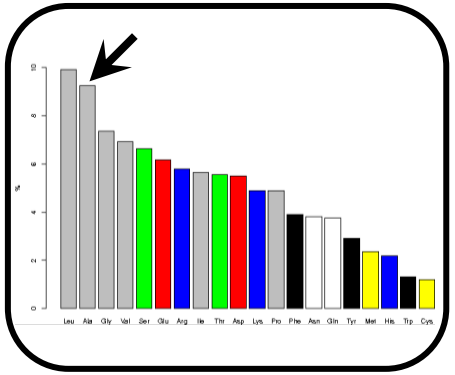
$p_S = 0.0672$

Amino acid composition (%)

Ala (A) 9.23	Gln (Q) 3.77	Leu (L) 9.86	Ser (S) 6.72
Arg (R) 5.84	Glu (E) 6.20	Lys (K) 4.89	Thr (T) 5.54
Asn (N) 3.78	Gly (G) 7.35	Met (M) 2.35	Trp (W) 1.30
Asp (D) 5.48	His (H) 2.20	Phe (F) 3.88	Tyr (Y) 2.87
Cys (C) 1.24	Ile (I) 5.53	Pro (P) 4.94	Val (V) 6.91

Probability in a little-more-realistic null model

Observing **A** in a sequence



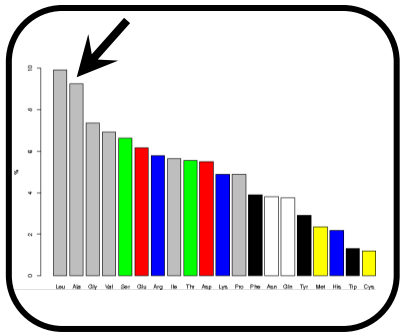
0.0923

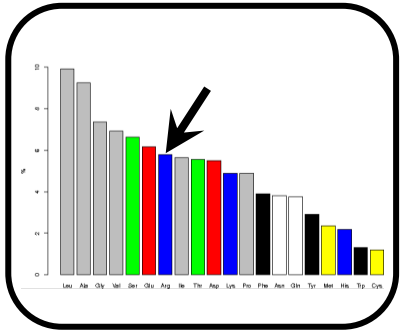
$P(\text{'A'}) = p_A = 0.0923$

sequence **X** ...**A**...
sequence **Y** ...**R**...

Aligned position:
X_i: position i in sequence X
Y_j: position j in sequence Y

Observing **A** at X_i and **R** at Y_j simply by chance





0.0054

$$P(X_i = \text{'A'} \wedge Y_j = \text{'R'}) = p_A \times p_R$$
$$= 0.0923 \times 0.0584$$
$$= 0.0054$$

Log-odds score for amino acid substitution

p_a : **prior** probability of observing residue *a*
p_b : **prior** probability of observing residue *b*
q_{ab}: **joint** probability of observing residue *a* and residue *b* in the same column
γ : a scaling factor (e.g. $\gamma = 10$ if $\log_{10}$ scale is used)

Log-odds score
for ***a*** substituted by ***b***

$$S_{ab} = \gamma \log\left(\frac{q_{ab}}{p_a \cdot p_b}\right)$$

$S_{ab} > 0$ (**positive**): **more** frequent than expected ($q_{ab} > p_a \times p_b$)
$S_{ab} < 0$ (**negative**): **less** frequent than expected ($q_{ab} < p_a \times p_b$)

Consider the **A–R** substitution in these two sequences:

```
Seq1 YPSVPFSAGP
Seq2 YPVLPFSRGP
```

Example

$P_{observed}$ for **A**→**R** substitution = q_{AR}
P_{null} for **A**→**R** substitution = $p_A \times p_R$

Substitution matrix

captures the propensity of any amino acids to be substituted by another amino acid due to biological reasons

Example

+2 indicates that the **N→N** replacement occurs 1.58 times more frequent than expected by chance

$$10 \cdot \log_{10}(x) = 2$$

$$\log_{10}(x) = 0.2$$

$$x = 10^{0.2} = 1.58$$

Log-odds scores

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

Why do we use log-odds score?

- 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)
- The sum of log-odds is equivalent to product of probabilities

Example

```
seqA QRVYPSVPFSAGP
seqB LRKYPVLPFSAGP
```

$$\boldsymbol{S}_{AB} = S_{\mathbf{QL}} + S_{\mathbf{RR}} + S_{\mathbf{VK}} + S_{\mathbf{YY}} + S_{\mathbf{PP}} + \ldots$$

Point Accepted Mutation (PAM) matrices

PAM matrices (Dayhoff, Schwartz & Orcutt, 1978)

- based on **1,572** observed mutations in **71 families** of **closely related** proteins

An **accepted point-mutation** is a single-residue mutation that was incorporated into the protein (and passed to its progeny), thus it:

(a) **did not disrupt** the protein function or

(b) was **beneficial** to the organism (e.g. in evolutionary terms, it increased the fitness of the species)

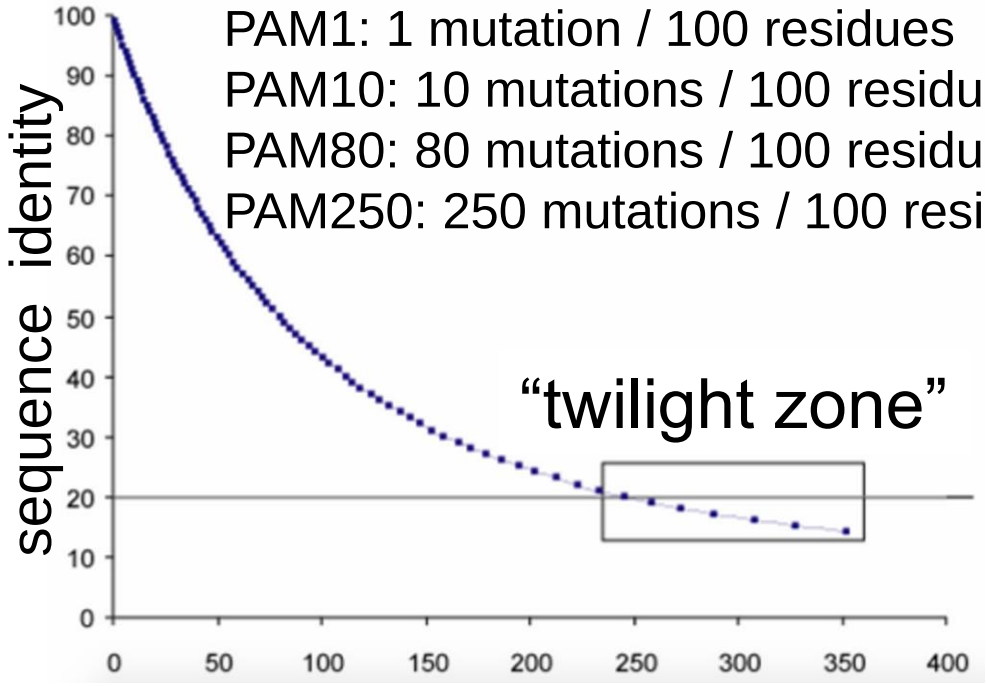
Margaret O. Dayhoff
(1925–1982)

PAM units

1 PAM unit: 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**. It measures the **rate of divergence**, i.e. the **evolutionary distance**.

- **PAM unit** between two sequences is **not necessarily the same** as percent difference in sequence identity
- Single position may undergo > 1 mutation, which could also result in no change observed in the sequence, e.g. **Y** ⟶ **H** ⟶ **Y**

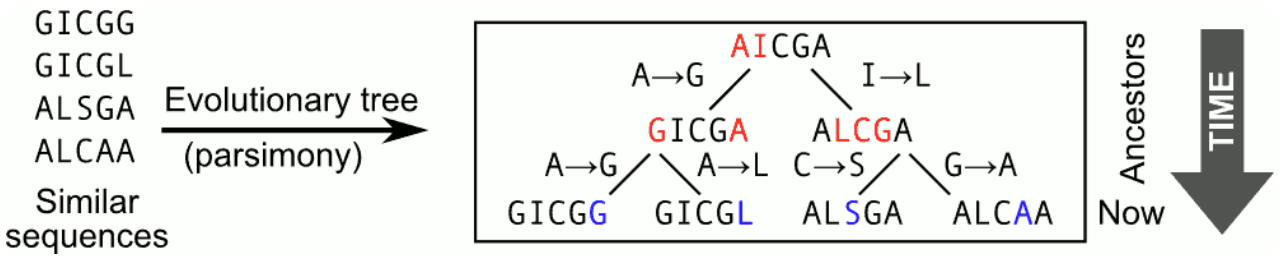
PAM0: 0 mutation / 100 residues
PAM1: 1 mutation / 100 residues
PAM10: 10 mutations / 100 residues
PAM80: 80 mutations / 100 residues
PAM250: 250 mutations / 100 residues



PAM unit

Construction of PAM*n* substitution matrices

1. for PAM1 matrix, protein sequences with >85% identity are used
2. count of amino acid replacements are recorded along branches of a phylogenetic tree



Count matrix

3. transition probability for each pair of amino acids is calculated based on the count matrix and the occurrence of these amino acids in the dataset

Transition probability matrix

4. matrices of other PAM units are extrapolated from PAM1 *via* matrix multiplication: **PAM2** = $(\text{PAM1})^2$; **PAM250** = $(\text{PAM1})^{250}$
5. probabilities can be transformed into log-odds scores

Scoring matrix

count matrix

	A	R	N	D	C
A	9867	2	9	10	3
R	1	9913	1	0	1
N	4	1	9822	36	0
D	6	0	42	9859	0
C	1	1	0	0	9973

PAM1

probability matrix

	A	R	N	D	C
A	13	6	9	9	5
R	3	17	4	3	2
N	4	4	6	7	2
D	5	4	8	11	1
C	2	1	1	1	52

PAM250

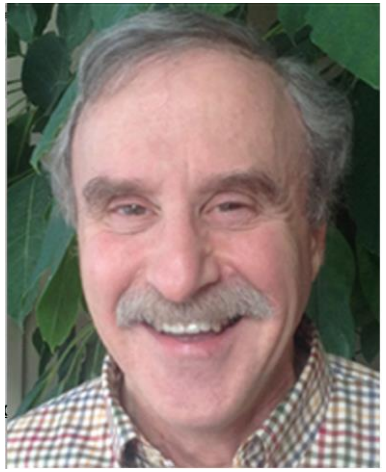
scoring matrix

A	2				
R	-2	6			
N	0	0	2		
D	0	-1	2	4	
C	-2	-4	-4	-5	12

PAM250

BLOck SUbstitution Matrix (BLOSUM)

- Based on **clustering** of **distantly related** proteins
- **Blocks** database consists of >2000 **locally aligned** (*blocks*) of conserved regions from >500 groups of distantly related proteins
- Observed amino acid frequencies derived based on aligned blocks (no phylogenetic trees)




Steven & Jorja Henikoff

- In **BLOSUM*n* matrices**, sequences with identity **> *n*%** are clustered
- Scores are derived from inter-cluster differences (among sequences sharing **< *n*% identity**).

BLOSUM: clustering based on %identity

Example

```
AKLGGREAVE
AKLIGREAVE
DKIGGHPAIE
DNIGGQPAIE
DKIGGQPAIE
EKLGGTTAVD
EKLGGTTAMK
EKLGGTAAVQ
EKLGGQAAVQ
YEAIGEELLS
```

Cluster at % identity ≥ **80**

```
AKLGGREAVE
AKLIGREAVE
```

```
DKIGGHPAIE
DNIGGQPAIE
DKIGGQPAIE
```

Identity for each possible pairwise sequences within each cluster ≥ **80%**

```
EKLGGTTAVD
EKLGGTTAMK
EKLGGTAAVQ
EKLGGQAAVQ
```

These four clusters form the basis for BLOSUM**80** matrix

```
YEAIGEELLS
```

BLOSUM: deriving transition probability

Calculate q_{QN} for BLOSUM**50**

Example

- **3** possible pairs of clusters (**1–2**, **1–3** and **2–3**)
- **5 amino acid residues** in length

$\boldsymbol{A}$: total number of aligned pairs **= 3 × 5 = 15**

$\boldsymbol{B}$: total **Q**⟷**N** substitution frequency **between** each cluster-pair

1	ATCK**Q** ATCR**N** ASCK**N** SSCR**N**	between **1–2**	¼ is **Q** in **1** compares to ½ is **N** in **2** ¾ is **N** in **1** compares to ½ is **Q** in **2**	$\left(\frac{1}{4}\times\frac{1}{2}\right)+\left(\frac{3}{4}\times\frac{1}{2}\right)=\frac{4}{8}=\mathbf{\frac{1}{2}}$
2	SDCE**Q** SECE**N**	between **2–3**	½ is **Q** in **2** compares to **0** is **N** in **3** ½ is **N** in **2** compares to **1/1** is **Q** in **3**	$\left(\frac{1}{2}\times 0\right)+\left(\frac{1}{2}\times\frac{1}{1}\right)=\mathbf{\frac{1}{2}}$
3	TECR**Q**	between **1–3**	¼ is **Q** in **1** compares to **0** is **N** in **3** ¾ is **N** in **1** compares to **1/1** is **Q** in **3**	$\left(\frac{1}{4}\times 0\right)+\left(\frac{3}{4}\times\frac{1}{1}\right)=\mathbf{\frac{3}{4}}$

clustered at %identity ≥ 50

$$\boldsymbol{B}\ \text{Total} = \frac{1}{2}+\frac{1}{2}+\frac{3}{4} = \frac{4+4+6}{8} = \frac{\mathbf{14}}{\mathbf{8}}$$

$$q_{QN} = \frac{B}{A} = \frac{14}{8} \div 15 = \mathbf{0.1167}$$

BLOSUM62 matrix

	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

- small and polar
- small and non-polar
- polar or acidic
- basic
- large and hydrophobic
- aromatic

similar to PAM120

Figure 4.4 Textbook (pg. 83)

PAM matrices

- Uses **closely related** proteins
- Based on an **explicit** evolutionary model (i.e. replacements counted on the branches of a phylogenetic tree)
- mutations observed throughout a **global** alignment
- All mutations are **counted the same**
- **Higher** PAM units denote **larger evolutionary distance**

BLOSUM80 PAM1 **Less divergent**	BLOSUM62 PAM120	BLOSUM45 PAM250 **More divergent**

BLOSUM matrices

- Uses **evolutionarily divergent** proteins
- Based on an **implicit** model of evolution (no trees)
- Based only on highly conserved regions in series of **local** alignments without gaps
- Uses groups of sequences within which **not all mutations are counted the same**
- **Larger** numbers in the BLOSUM matrix naming scheme denote higher sequence similarity (and thus **smaller evolutionary distance**)

These matrices could be too simplistic. Further developments result in more-realistic substitution models based on more-extensive protein data, e.g. **JTT** and **WAG**.

Reflection

What is homology, and how does it relate to shared sequence similarity?

Biologically, why would DNA/proteins sequences be conserved, or different from one another?

How do we quantify and model sequence change?

Why do we prefer log-odds score to probability values?

How can we model substitutions based on empirical data?

What are PAM and BLOSUM matrices, and how do they differ?