Bioinformatics

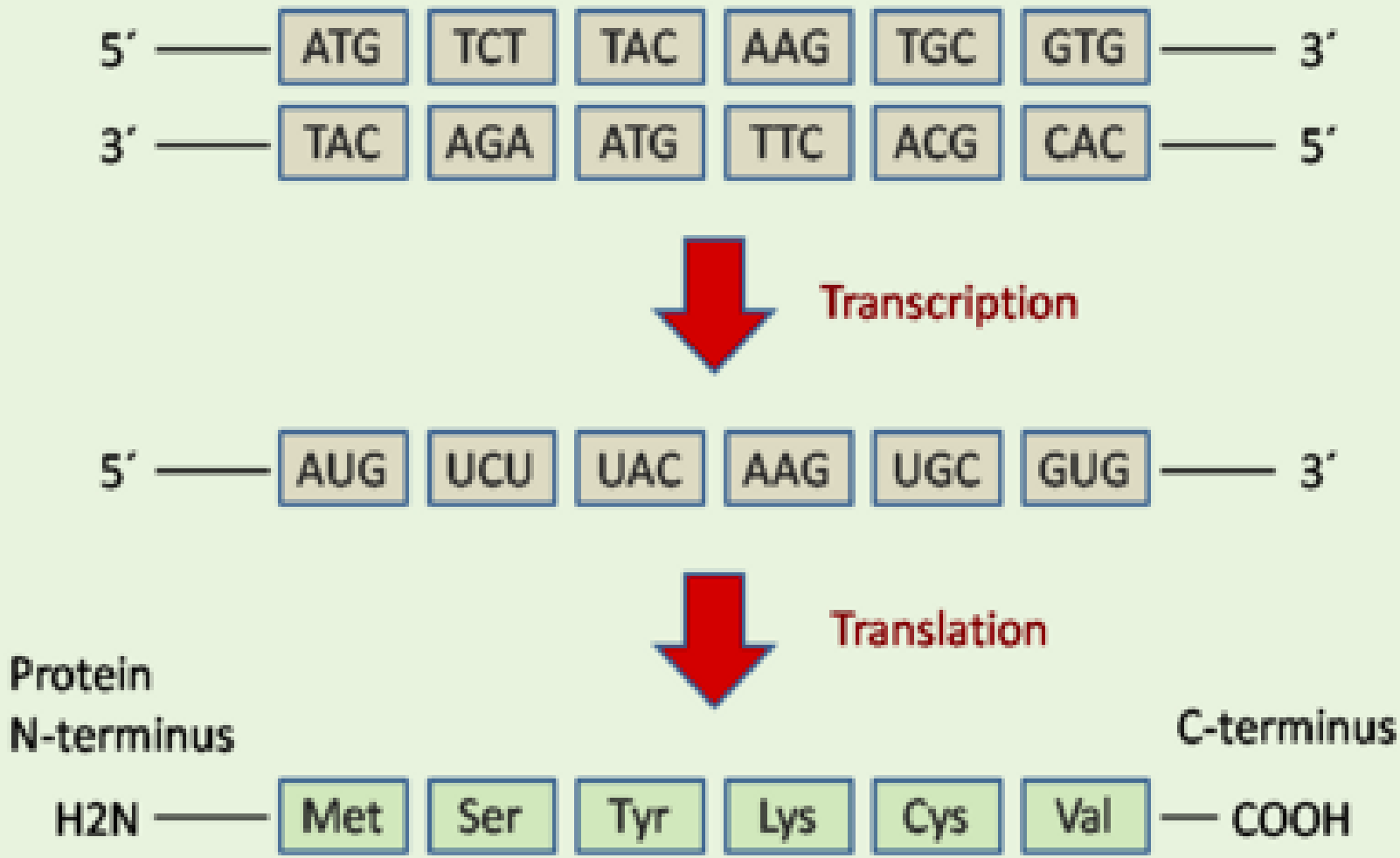
CS300

Substitution Matrices and Protein Alignments

Fall 2017

Oliver Bonham-Carter

The central dogma of molecular biology

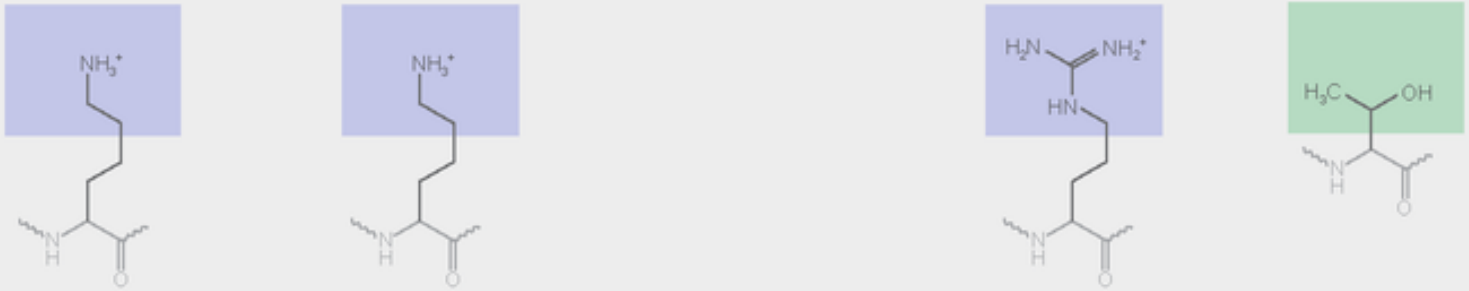


https://en.wikipedia.org/wiki/DNA_codon_table

Silent Mutations

- Redundant codons mean ~1/3 of DNA mutations often do not alter protein sequence

	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr



basic
polar

https://en.wikipedia.org/wiki/Silent_mutation

Silent Mutations

- Are these mutations really so subtle?
- Are there dangers involved?
  - While the protein may be fine, the RNA has still has dangerous folding issues
- Nature: *Silent Mutations Speak Up: Overlooked genetic changes could impact on disease*
  - http://www.nature.com/news/2006/061221/full/news061218-12.html

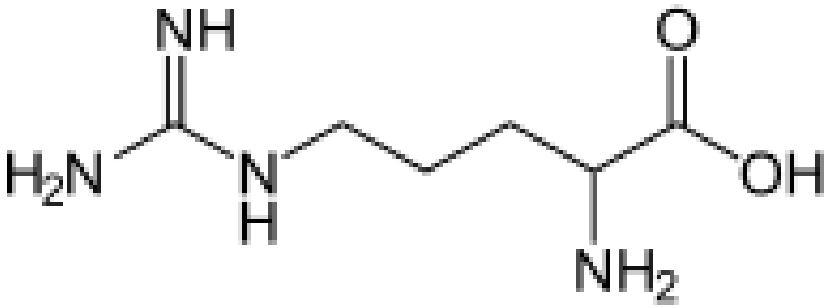
nature International weekly journal of science

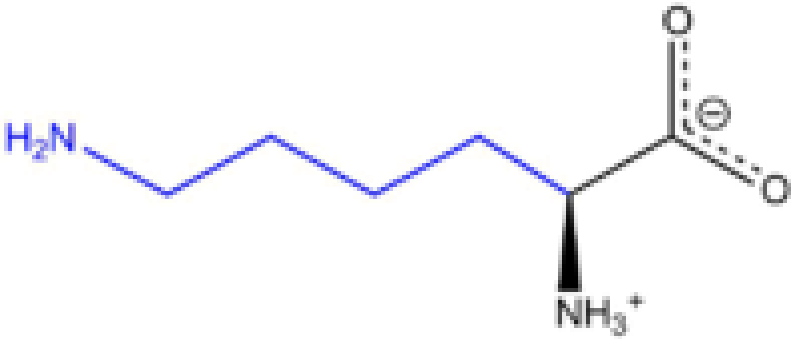
Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC } Phe UUA, UUG } Leu	UCU, UCC, UCA, UCG } Ser	UAU, UAC } Tyr UAA Stop UAG Stop	UGU, UGC } Cys UGA Stop UGG Trp	U C A G
C	CUU, CUC, CUA, CUG } Leu	CCU, CCC, CCA, CCG } Pro	CAU, CAC } His CAA, CAG } Gln	CGU, CGC, CGA, CGG } Arg	U C A G
A	AUU, AUC, AUA } Ile AUG Met	ACU, ACC, ACA, ACG } Thr	AAU, AAC } Asn AAA, AAG } Lys	AGU, AGC } Ser AGA, AGG } Arg	U C A G
G	GUU, GUC, GUA, GUG } Val	GCU, GCC, GCA, GCG } Ala	GAU, GAC } Asp GAA, GAG } Glu	GGU, GGC, GGA, GGG } Gly	U C A G

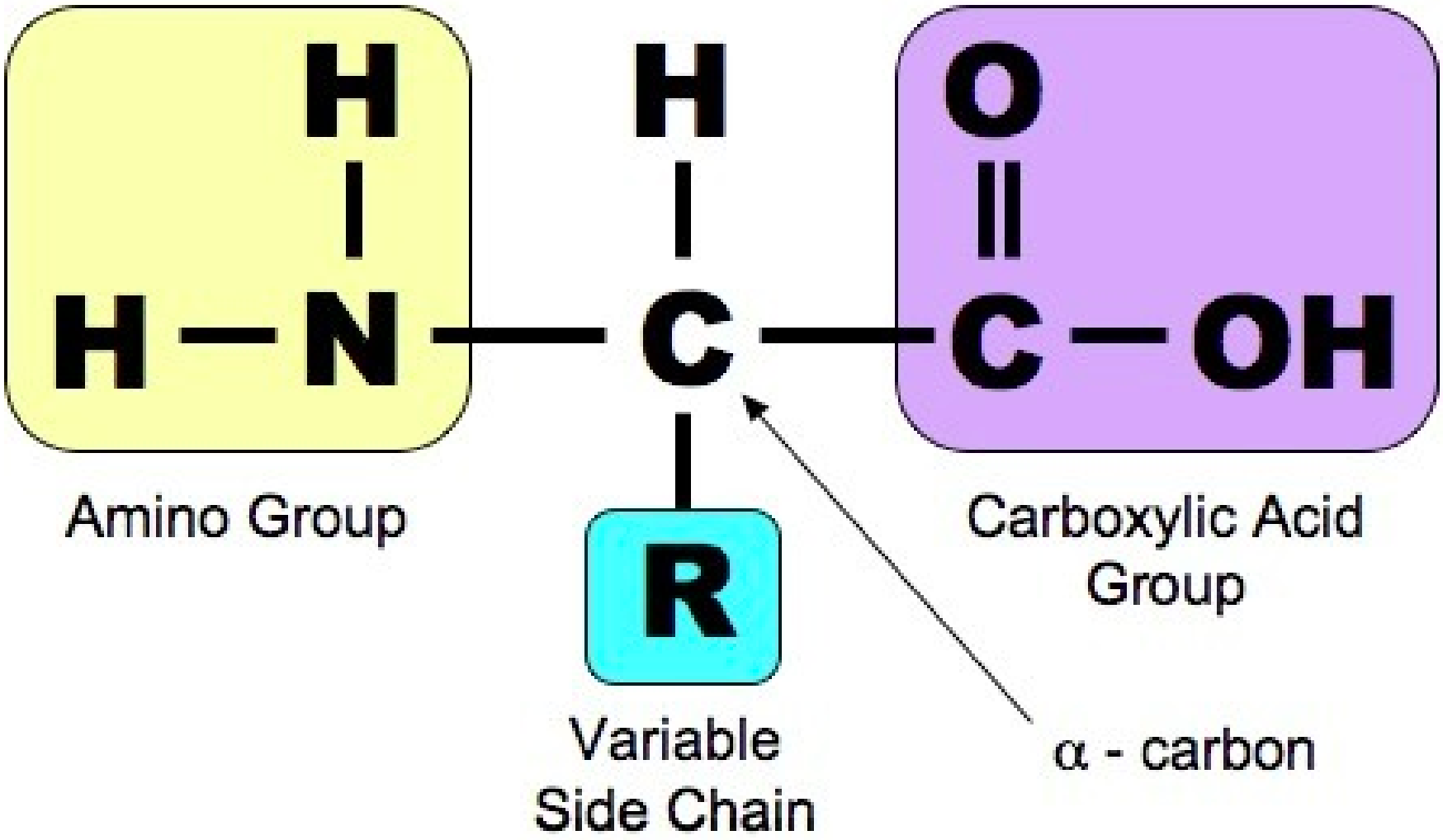
Alphabetical Interests

- With a larger protein “alphabet” (20 amino acids), it is much less likely to get matches by chance.
- Matches are likely to be statistically significance
- Amino acid changes are not equally harmful to protein structure
  - Chemical complexes being replaced by similar chemical complex.
  - Ex: Arginine (Arg) and Lysine (Lys)

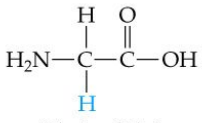


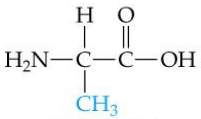


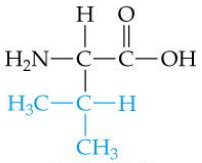
Amino Acid Substitutions

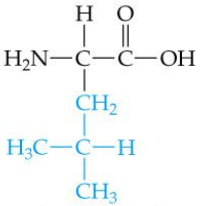


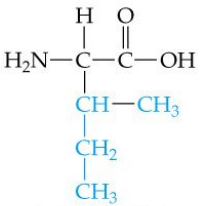
Amino Acid Substitutions

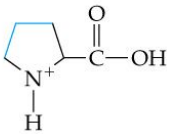

Glycine (Gly)

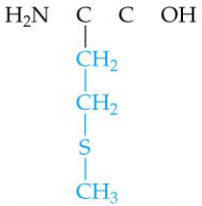

Alanine (Ala)

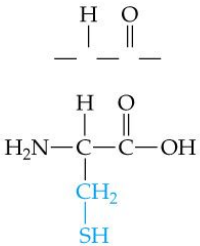

Valine (Val)

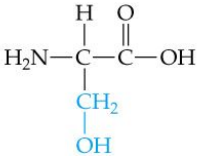

Leucine (Leu)

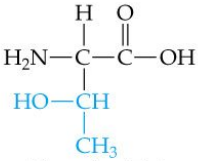

Isoleucine (Ile)

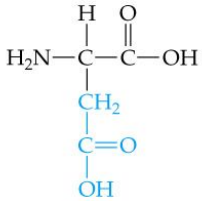

Proline (Pro)

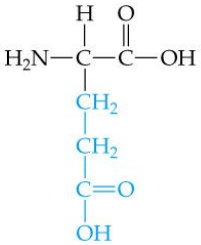

Methionine (Met)

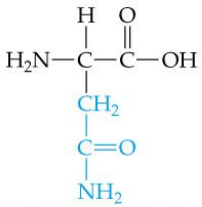

Cysteine (Cys)

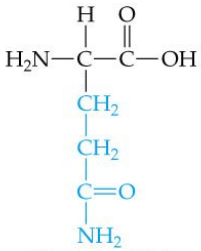

Serine (Ser)

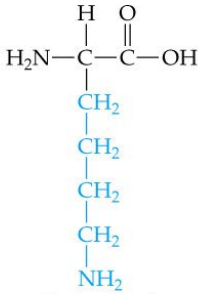

Threonine (Thr)

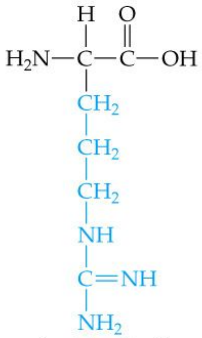

Aspartic acid (Asp)

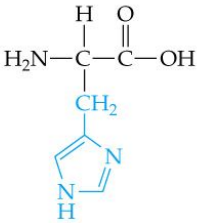

Glutamic acid (Glu)

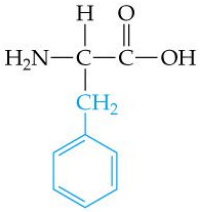

Asparagine (Asn)

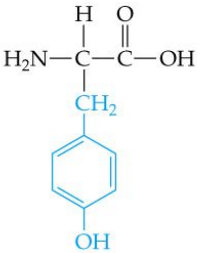

Glutamine (Glu)

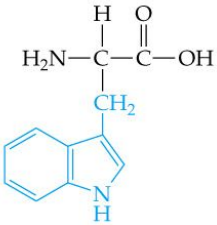

Lysine (Lys)


Arginine (Arg)


Histidine (His)


Phenylalanine (Phe)

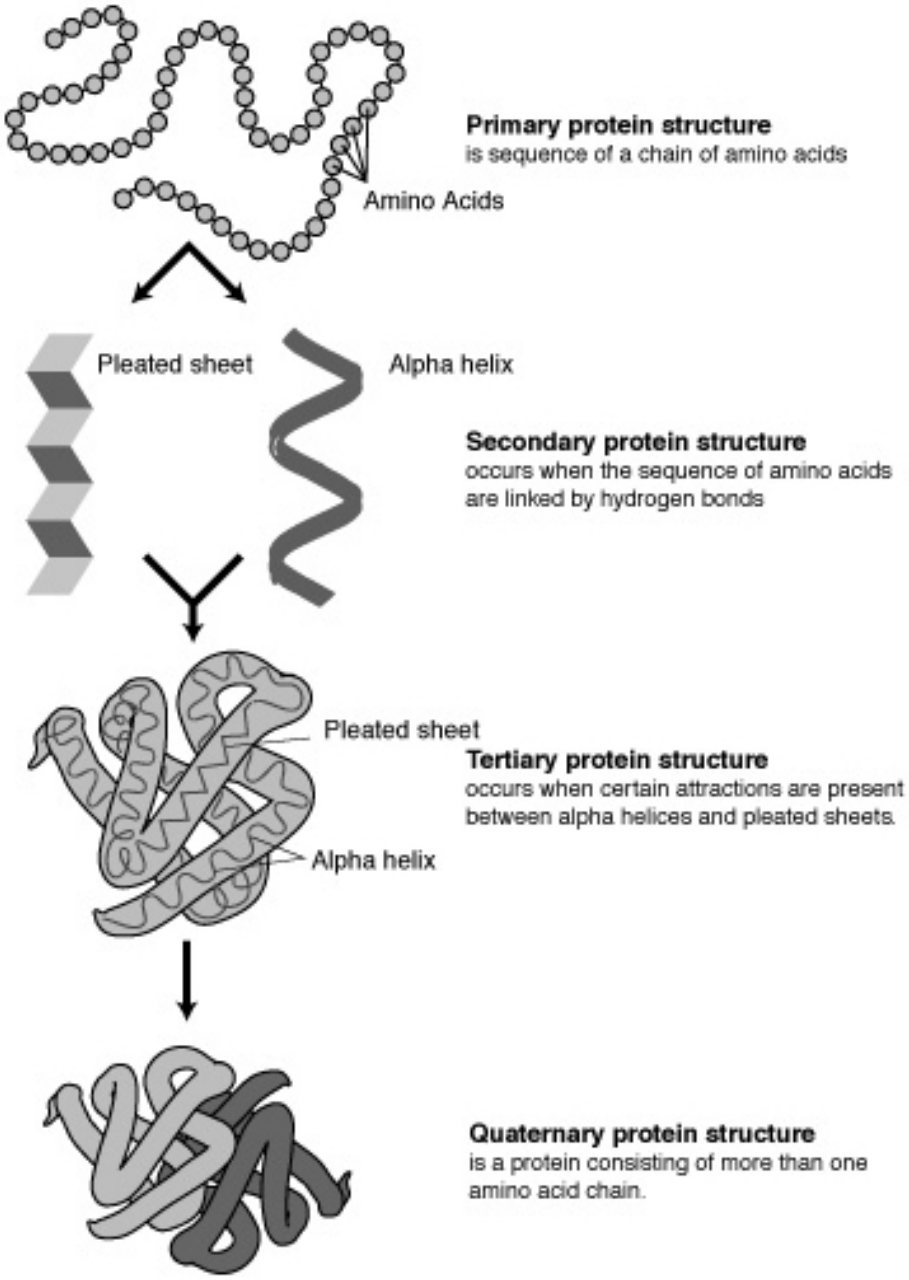

Tyrosine (Tyr)


Tryptophan (Trp)

Amino Acid Components

- Similarity of amino acids means
  - Similar *physicochemical properties* (Physics + chemistry)
    - Polar vs nonpolar
    - Hydrophobic vs hydrophilic
    - Positive electric charge vs negative electric charge
    - Basic vs Acidic
- Amino Acid Table: http://www.bio.davidson.edu/courses/genomics/jmol/aatable.html
- Roles in Protein Structures
- http://www.proteinstructures.com/Structure/Structure/amino-acids.html

Amino Acids Determine Protein's Shape and Function



The hierarchy of protein structure. Public domain image from The National Genome Research Institute

Scoring Amino Acid Substitutions

- Could we quantify sequence by physicochemical properties? (yes!)

Table 5.1 Hydrophobicity values for the 20 amino acids. A more positive value represents a more hydrophobic amino acid.

Amino Acid	Hydrophobicity	Amino Acid	Hydrophobicity	Amino Acid	Hydrophobicity
D	–3.5	Y	–1.3	I	4.5
K	–3.9	N	–3.5	C	2.5
H	–3.2	L	3.8	A	1.8
T	–0.7	E	–3.5	S	–0.8
V	4.2	R	–4.5	G	–0.4
F	2.8	W	–0.9	P	–1.6
M	1.9	Q	–3.5		

Scoring Amino Acid Substitutions

Better to study evolution of real proteins from closely related organisms

Minimizes likelihood that an observed difference represents a series of >1 individual mutations

Species A – Ala

Species B – Ile

Ala x Ile – 1 mutation

Ala x Pro x Ser x Ile – 3 mutations

A Model of Evolutionary Change in Proteins, Dayhoff et al., 1978

Global Pairwise Alignment

Observed frequency of each possible amino acid substitution:

$$10 \log_{10} (M_{ij}/f_j)$$

- M_{ij} - the probability of a mutation replacing amino *i* with *j*
- f_j - the frequency of amino acid *j* in a large set of sequences

The PAM Matrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	2																			
Arg	R	–1	5																		
Asn	N	0	0	3																	
Asp	D	0	–1	2	5																
Cys	C	–1	–1	–1	–3	11															
Gln	Q	–1	2	0	1	–3	5														
Glu	E	–1	0	1	4	–4	2	5													
Gly	G	1	0	0	1	–1	–1	0	5												
His	H	–2	2	1	0	0	2	0	–2	6											
Ile	I	0	–3	–2	–3	–2	–3	–3	–3	–3	4										
Leu	L	–1	–3	–3	–4	–3	–2	–4	–4	–2	2	5									
Lys	K	–1	4	1	0	–3	2	1	–1	1	–3	–3	5								
Met	M	–1	–2	–2	–3	–2	–2	3	3	–2	3	3	–2	6							
Phe	F	–3	–4	–3	–5	0	–4	–5	–5	0	0	2	–5	0	8						
Pro	P	1	–1	–1	–2	–2	0	–2	–1	0	–2	0	–2	–2	–3	6					
Ser	S	1	–1	1	0	1	–1	–1	1	–1	–1	–2	–1	–1	–2	1	2				
Thr	T	2	–1	1	–1	–1	–1	–1	–1	–1	1	–1	–1	0	–2	1	1	2			
Trp	W	–4	0	–5	–5	1	–3	–5	–2	–3	–4	–2	–3	–3	–1	–4	–3	–4	15		
Tyr	Y	–3	–2	–1	–2	2	–2	–4	–4	4	–2	–1	–3	–2	5	–3	–1	–3	0	9	
Val	V	1	–3	–2	–2	–2	–3	–2	–2	–3	4	2	–3	2	0	–1	–1	0	–3	–3	

PAM matrices

- Point Accepted Mutation
- Family of matrices PAM 1, PAM 80, PAM 120, PAM 250
- The number with a PAM matrix (the *n* in PAM *n*) represents the evolutionary distance between the sequences on which the matrix is based

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

PAM vs BLOSUM

- General Use
  - PAM 120
  - BLOSUM 62*
- Closely Related Species
  - PAM 60
  - BLOSUM 80
- Distantly Related Species
  - PAM 250
  - BLOSUM 45

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

*BLOSUM 62 – used by BLAST – computed by choosing blocks of local alignments more than 62% identical

Blast Subst Matrices

- Scoring for possible residue pair alignment
- Different substitution matrices are for detecting similarities according to degrees of divergence.
- BLOSUM-62 matrix good for detecting most weak protein similarities
- Provisional table of recommended substitution matrices and gap costs for various query lengths is

Query Length	Substitution Matrix	Gap Costs
<35	PAM-30	(9,1)
35-50	PAM-70	(10,1)
50-85	BLOSUM-80	(10,1)
85	BLOSUM-62	(10,1)

https://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html

BLOSUM matrix
Heinkoff and Heinkoff, 1992

- **BLOcks SUbstition Matrix** - Blocks of local alignments

$$S_{ij} = \left(\frac{1}{\lambda}\right) \log \left(\frac{p_{ij}}{q_i * q_j}\right)$$

- p_{ij} - probability j replacing i
- q_i and q_j - probabilities of finding the amino acids i and j in any protein sequence
- λ - scaling factor, set such that the matrix contains easily computable integer values.

- BLOSUM # - # = minimum % similarity of sequences compared

Needleman-Wunsch Algorithm: Nucleotide Alignment – Chap 3

- Create N x M matrix
- Place each sequence along one axis
- Place score 0 at the up-left corner
- Fill in 1st row & column with gap penalty multiples
- Fill in the matrix with max value of 3 possible moves:
  - Vertical move: Score + gap penalty
  - Horizontal move: Score + gap penalty
  - Diagonal move: Score + match/mismatch score
- The optimal alignment score is in the lower-right corner
- To reconstruct the optimal alignment, trace back where the max at each step came from, stop when hit the origin.

Needleman-Wunsch Algorithm:
Protein Alignment – Chap 5

- Create N x M matrix
- Place each sequence along one axis
- Place score 0 at the up-left corner
- Fill in 1st row & column with gap penalty multiples
- Fill in the matrix with max value of 3 possible moves:
  - Vertical move: Score + gap penalty
  - Horizontal move: Score + gap penalty
  - Diagonal move: Score + **match/mismatch score from sub. matrix**
- The optimal alignment score is in the lower-right corner
- To reconstruct the optimal alignment, trace back where the max at each step came from, stop when hit the origin.

Blast-Off!!

- Let’s blast some protein sequences
- https://blast.ncbi.nlm.nih.gov/Blast.cgi#dtr_Query_98931