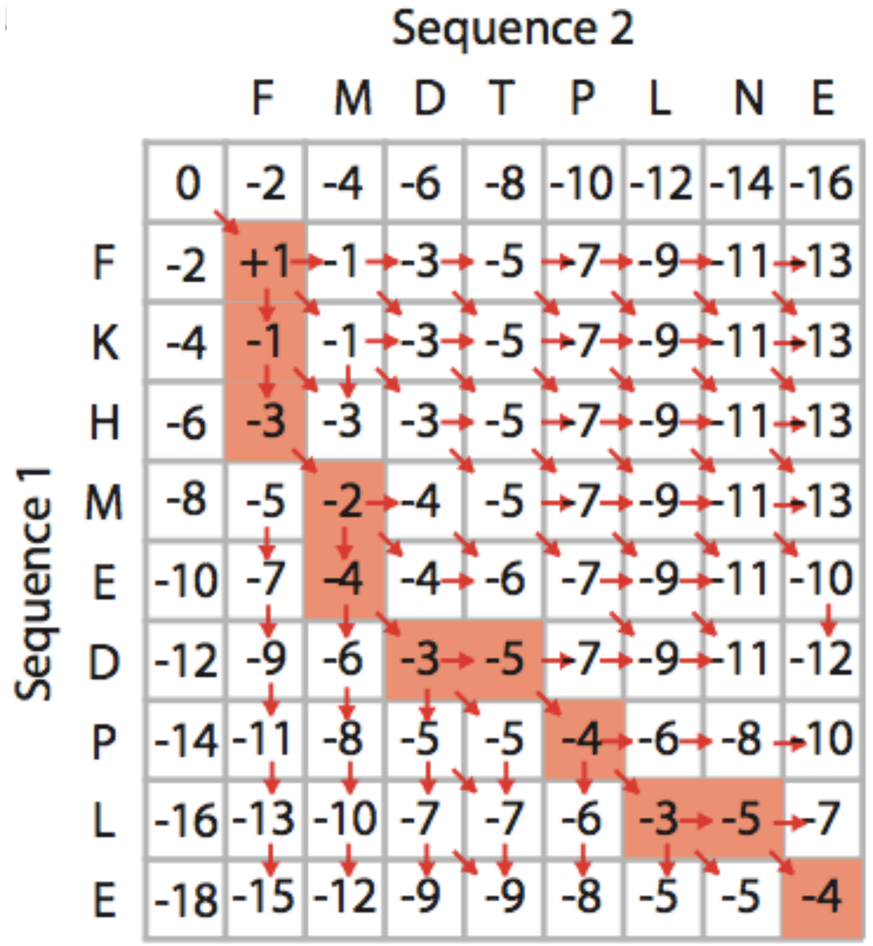
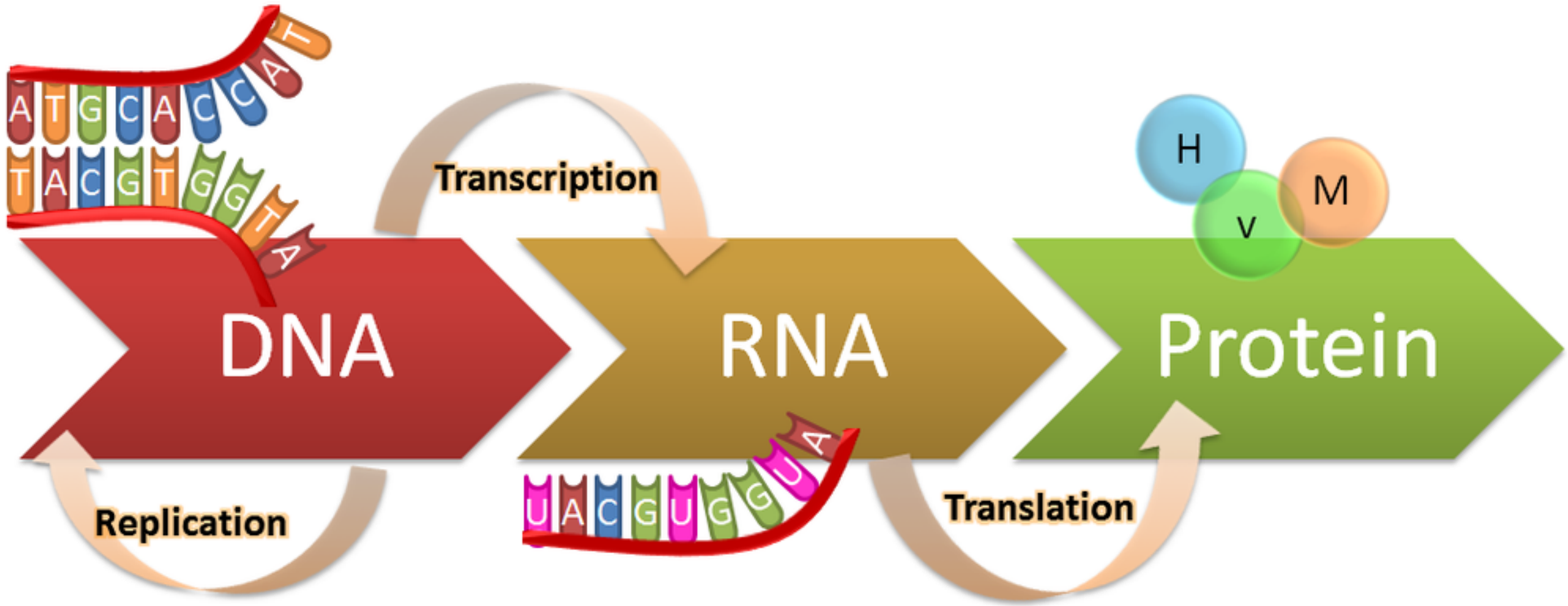
# Lecture 2 - Pairwise Sequence Alignment



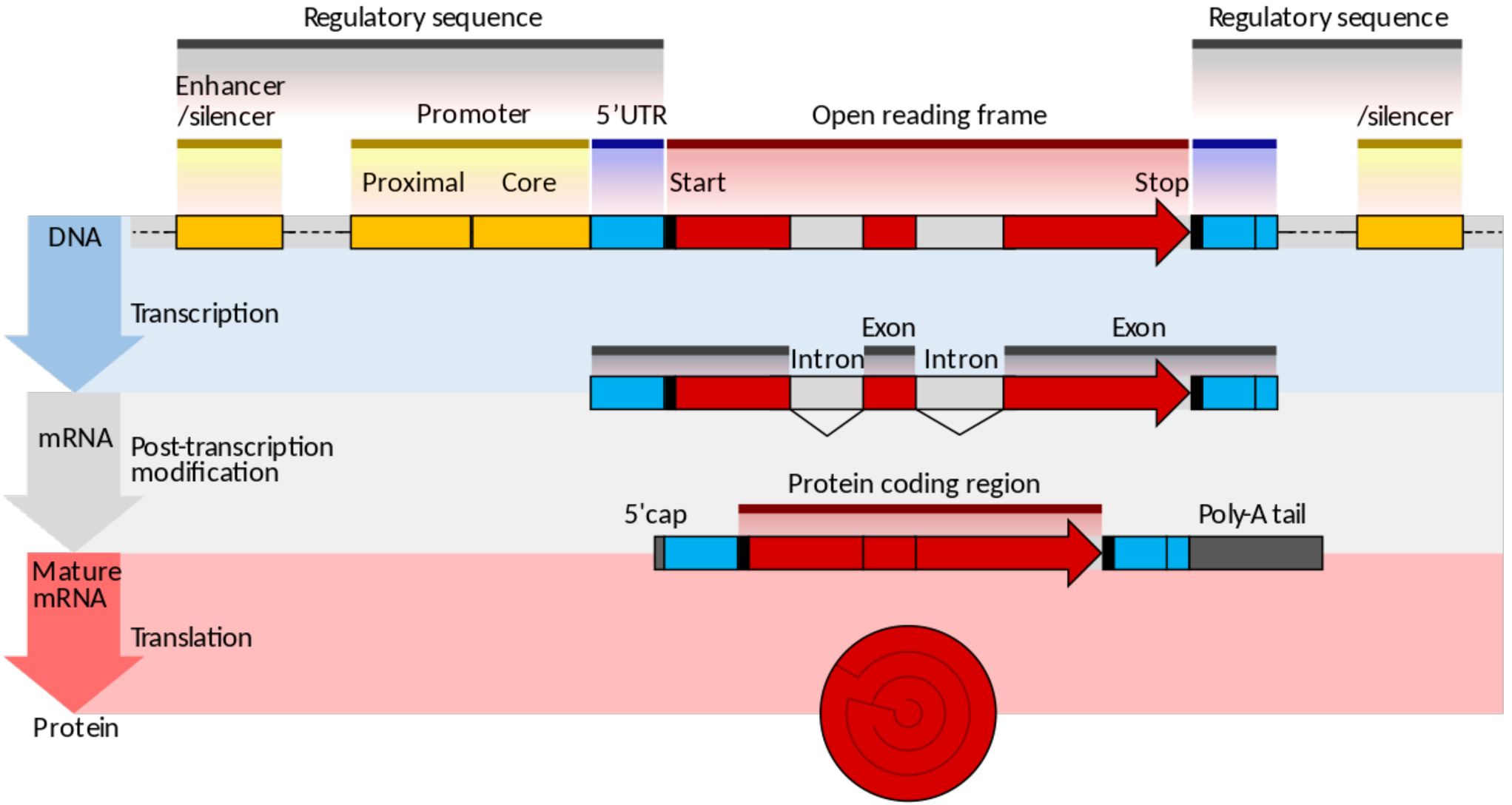
# Brief Refresher on the “Central Dogma”





ian.simpson@ed.ac.uk

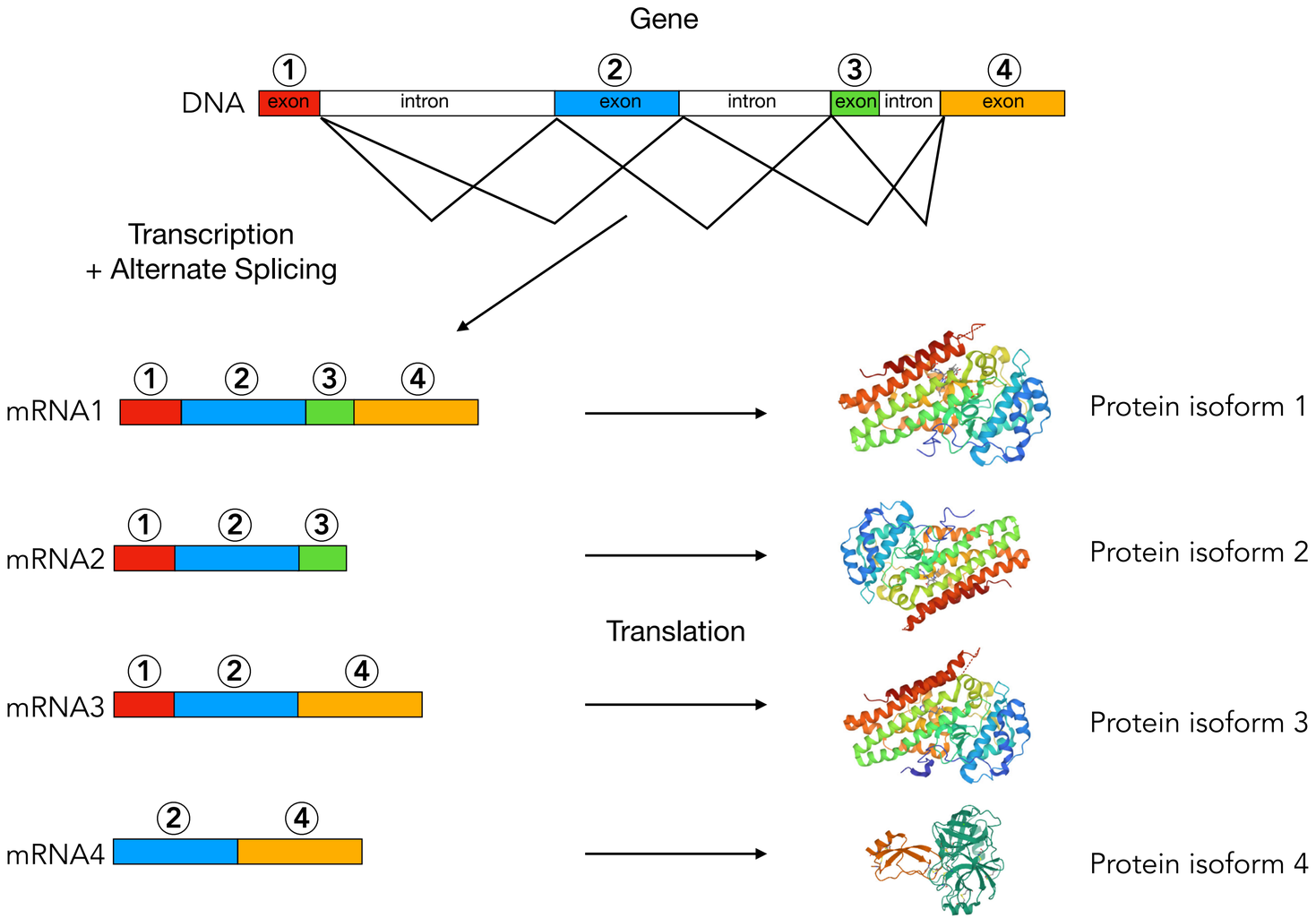
# Structure of a Eukaryotic Gene





ian.simpson@ed.ac.uk

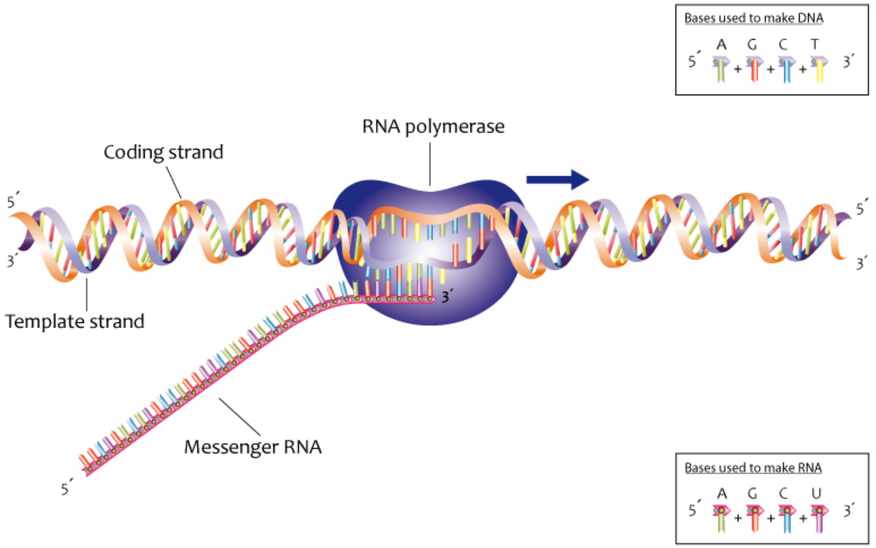
# One Gene, Many Transcripts, Many Proteins





ian.simpson@ed.ac.uk

# Transcription




DNA coding strand (aka Crick strand, strand +1)

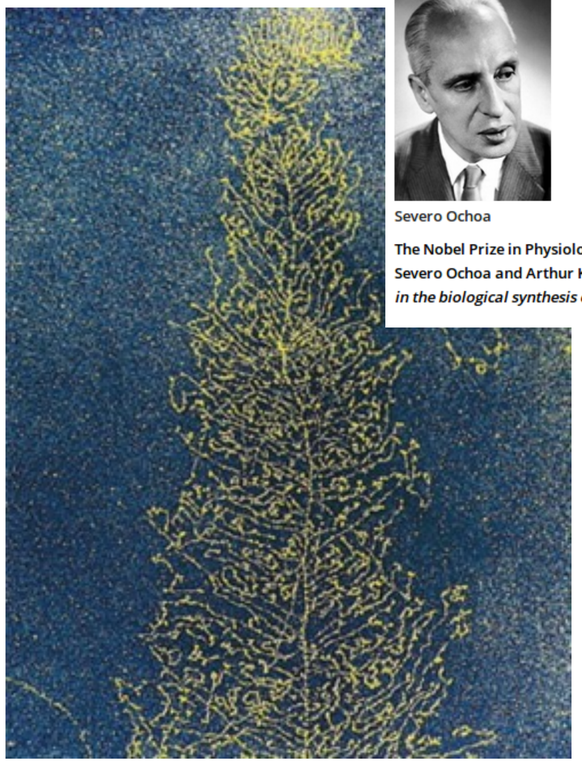
```
5′   ATGGCCATTGTAATGGGCCGCTGAAAGGGTGCCCGATAG   3′
     ||||||||||||||||||||||||||||||||||||||
3′   TACCGGTAACATTACCCGGCGACTTTCCCACGGGCTATC   5′
```

DNA template strand (aka Watson strand, strand –1)

↓ transcription

```
5′   AUGGCCAUUGUAAUGGGCCGCUGAAAGGGUGCCCGAUAG   3′
```

Single stranded messenger RNA



The Nobel Prize in Physiology or Medicine 1959
Severo Ochoa, Arthur Kornberg

## The Nobel Prize in Physiology or Medicine 1959

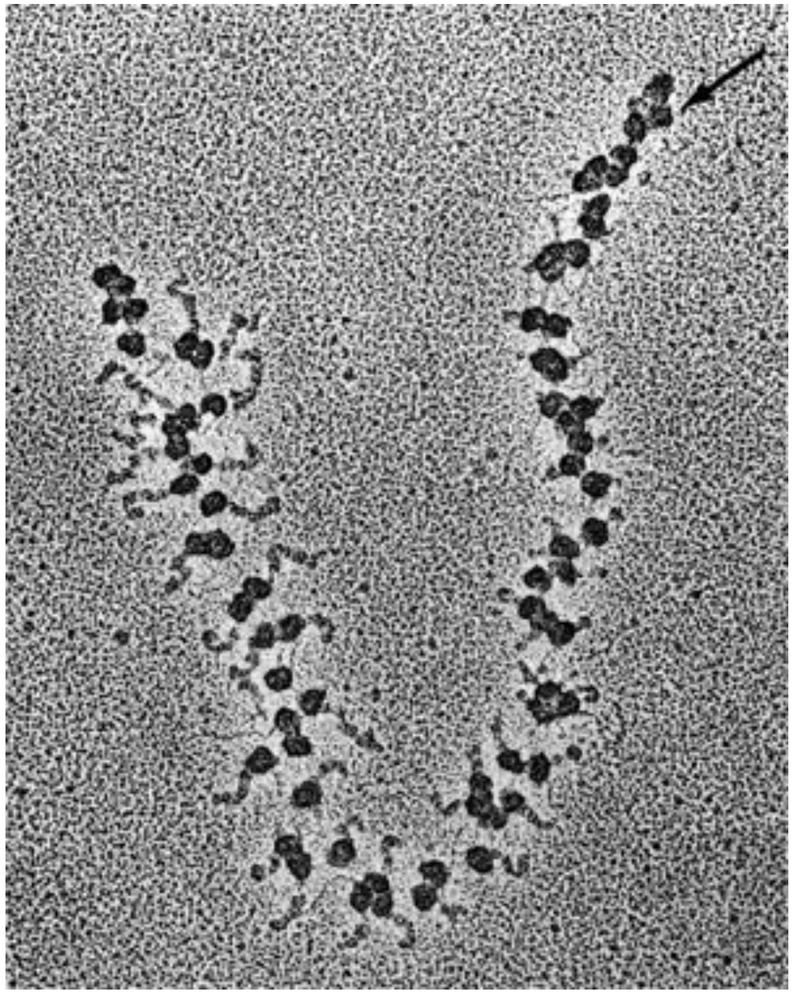

Severo Ochoa


Arthur Kornberg

The Nobel Prize in Physiology or Medicine 1959 was awarded jointly to Severo Ochoa and Arthur Kornberg *"for their discovery of the mechanisms in the biological synthesis of ribonucleic acid and deoxyribonucleic acid"*


ian.simpson@ed.ac.uk

# Translation



The Nobel Prize in Chemistry 2009
Venkatraman Ramakrishnan, Thomas A. Steitz, Ada E. Yonath

## The Nobel Prize in Chemistry 2009

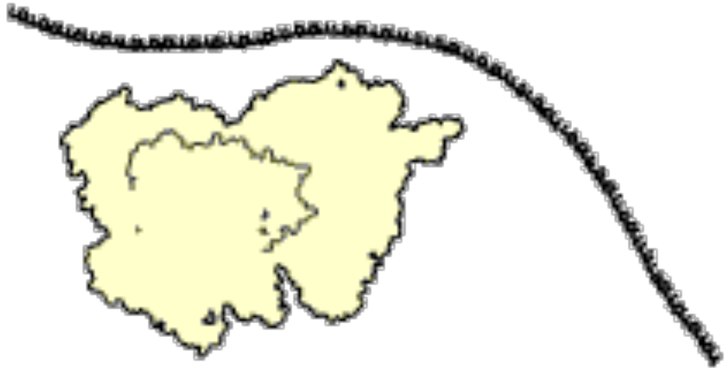
Photo: U. Montan
**Venkatraman Ramakrishnan**

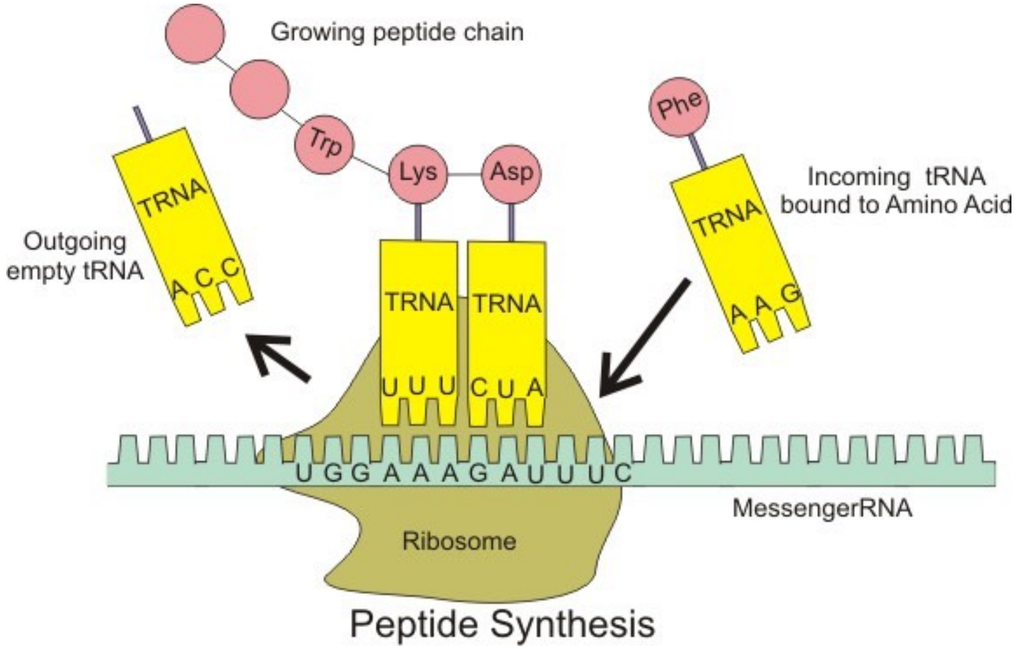
Photo: U. Montan
**Thomas A. Steitz**

Photo: U. Montan
**Ada E. Yonath**

**The Nobel Prize in Chemistry 2009 was awarded jointly to Venkatraman Ramakrishnan, Thomas A. Steitz and Ada E. Yonath *"for studies of the structure and function of the ribosome"*.**

# Translation






http://en.wikipedia.org/wiki/Translation_(biology)


ian.simpson@ed.ac.uk

# The Genetic Code

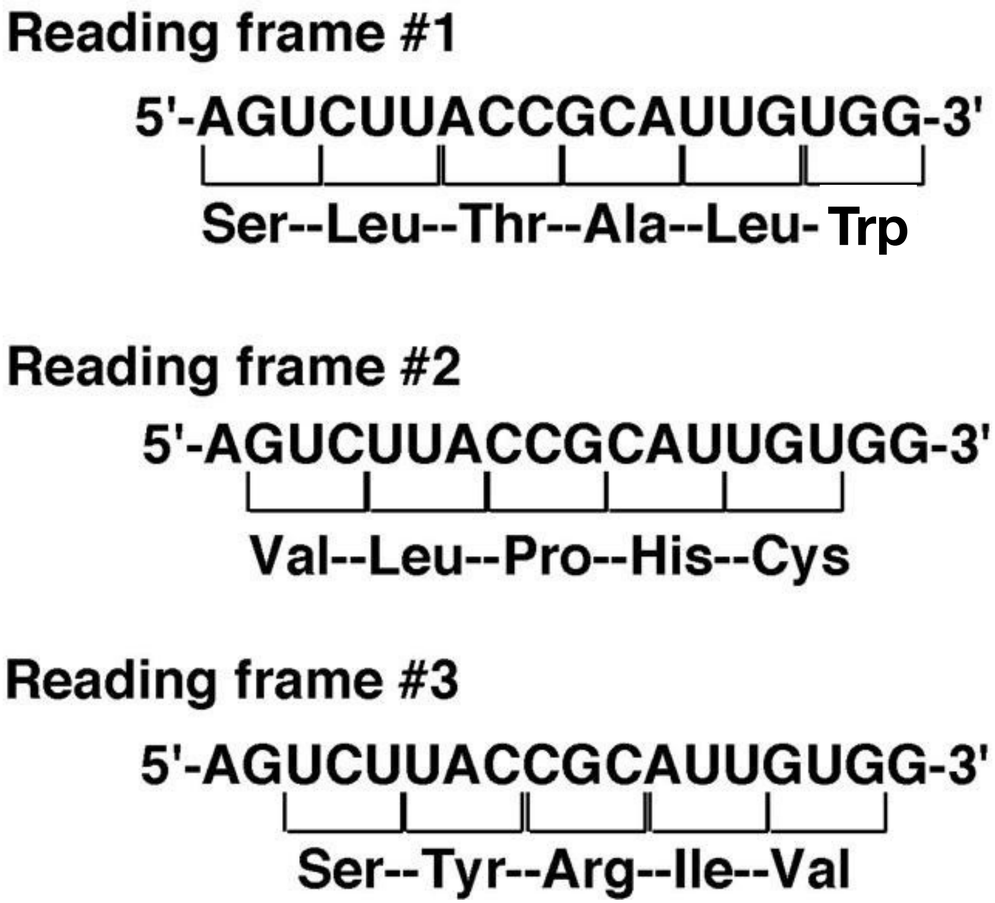
**Standard genetic code**

| 1st base | 2nd base | | | | | | | | 3rd base |
|---|---|---|---|---|---|---|---|---|---|
| | T | | C | | A | | G | | |
| T | TTT | (Phe/F) Phenylalanine | TCT | (Ser/S) Serine | TAT | (Tyr/Y) Tyrosine | TGT | (Cys/C) Cysteine | T |
| | TTC | | TCC | | TAC | | TGC | | C |
| | TTA | (Leu/L) Leucine | TCA | | TAA | Stop (*Ochre*) | TGA | Stop (*Opal*) | A |
| | TTG | | TCG | | TAG | Stop (*Amber*) | TGG | (Trp/W) Tryptophan | G |
| C | CTT | | CCT | (Pro/P) Proline | CAT | (His/H) Histidine | CGT | (Arg/R) Arginine | T |
| | CTC | | CCC | | CAC | | CGC | | C |
| | CTA | | CCA | | CAA | (Gln/Q) Glutamine | CGA | | A |
| | CTG | | CCG | | CAG | | CGG | | G |
| A | ATT | (Ile/I) Isoleucine | ACT | (Thr/T) Threonine | AAT | (Asn/N) Asparagine | AGT | (Ser/S) Serine | T |
| | ATC | | ACC | | AAC | | AGC | | C |
| | ATA | | ACA | | AAA | (Lys/K) Lysine | AGA | (Arg/R) Arginine | A |
| | ATG[A] | (Met/M) Methionine | ACG | | AAG | | AGG | | G |
| G | GTT | (Val/V) Valine | GCT | (Ala/A) Alanine | GAT | (Asp/D) Aspartic acid | GGT | (Gly/G) Glycine | T |
| | GTC | | GCC | | GAC | | GGC | | C |
| | GTA | | GCA | | GAA | (Glu/E) Glutamic acid | GGA | | A |
| | GTG | | GCG | | GAG | | GGG | | G |


ian.simpson@ed.ac.uk

# How To Read mRNA Sequences

Start: AUG
Stop: UAA, UAG, UGA





ian.simpson@ed.ac.uk

# Learning Objectives

- Define **homology** as well as **orthologs** and **paralogs**
- Explain how PAM (accepted point mutation) matrices are derived
- Contrast the utility of **PAM** and **BLOSUM** scoring matrices
- Define dynamic programming and explain how:
  Global (**Needleman–Wunsch**) &
  Local (**Smith–Waterman**)
  pairwise alignments are performed
- Perform pairwise alignment of protein or DNA sequences


ian.simpson@ed.ac.uk

# Pairwise Sequence Alignment Is Fundamental

- Determine whether two proteins (or genes) are related structurally or functionally
- Studying domain or motif similarities between proteins
- Genome Analysis
- The basis of many search strategies and algorithms including BLAST


ian.simpson@ed.ac.uk

# Comparing Protein vs DNA Sequences

## Protein

- protein is more informative (20 vs 4 characters);
  amino acids share related biophysical properties
- codons are degenerate: changes in the third position
  often do not alter the amino acid specified
- protein sequences offer a longer "look-back" time

## DNA

- to study noncoding regions of DNA
  introns or intergenic regions
- to study DNA polymorphisms
- genome sequencing relies on DNA analysis

**Standard genetic code**

| 1st base | 2nd base | | | | | | | | 3rd base |
|---|---|---|---|---|---|---|---|---|---|
| | T | | C | | A | | G | | |
| T | TTT | (Phe/F) Phenylalanine | TCT | (Ser/S) Serine | TAT | (Tyr/Y) Tyrosine | TGT | (Cys/C) Cysteine | T |
| | TTC | | TCC | | TAC | | TGC | | C |
| | TTA | (Leu/L) Leucine | TCA | | TAA | Stop (*Ochre*) | TGA | Stop (*Opal*) | A |
| | TTG | | TCG | | TAG | Stop (*Amber*) | TGG | (Trp/W) Tryptophan | G |
| C | CTT | | CCT | (Pro/P) Proline | CAT | (His/H) Histidine | CGT | (Arg/R) Arginine | T |
| | CTC | | CCC | | CAC | | CGC | | C |
| | CTA | | CCA | | CAA | (Gln/Q) Glutamine | CGA | | A |
| | CTG | | CCG | | CAG | | CGG | | G |
| A | ATT | (Ile/I) Isoleucine | ACT | (Thr/T) Threonine | AAT | (Asn/N) Asparagine | AGT | (Ser/S) Serine | T |
| | ATC | | ACC | | AAC | | AGC | | C |
| | ATA | | ACA | | AAA | (Lys/K) Lysine | AGA | (Arg/R) Arginine | A |
| | ATG[A] | (Met/M) Methionine | ACG | | AAG | | AGG | | G |
| G | GTT | (Val/V) Valine | GCT | (Ala/A) Alanine | GAT | (Asp/D) Aspartic acid | GGT | (Gly/G) Glycine | T |
| | GTC | | GCC | | GAC | | GGC | | C |
| | GTA | | GCA | | GAA | (Glu/E) Glutamic acid | GGA | | A |
| | GTG | | GCG | | GAG | | GGG | | G |


ian.simpson@ed.ac.uk

# Pairwise Alignment

"The process of lining up two sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology"


ian.simpson@ed.ac.uk

# Identity, Similarity & Conservation

## Homology

Similarity attributed to descent from a common ancestor.

## Identity

The extent to which two (nucleotide or amino acid) sequences are invariant.

## Similarity

The extent to which nucleotide or protein sequences are related. It is based upon identity plus conservation.

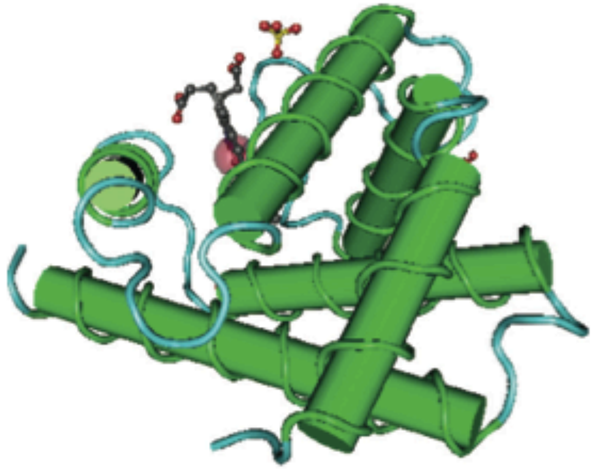
## Conservation

Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

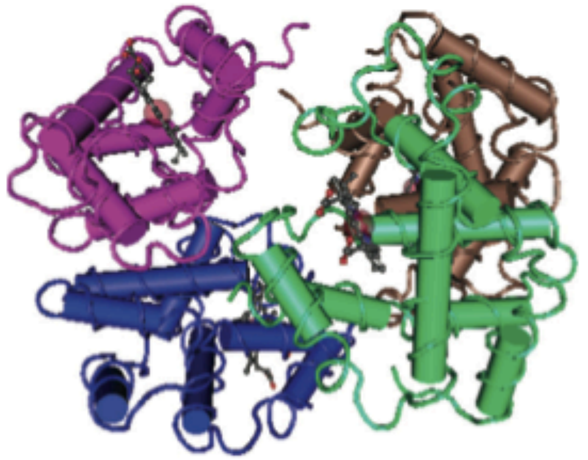

ian.simpson@ed.ac.uk

# Globin Homologs

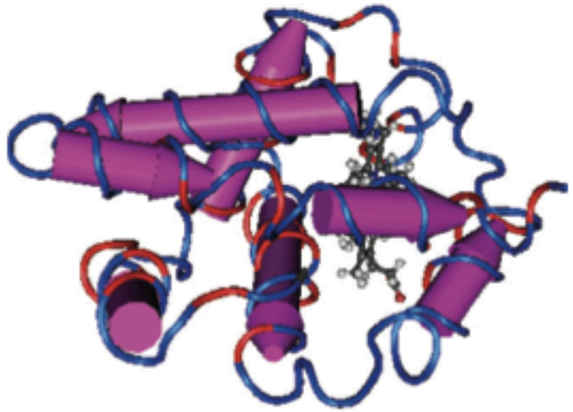
(a) Human myoglobin (3RGK)



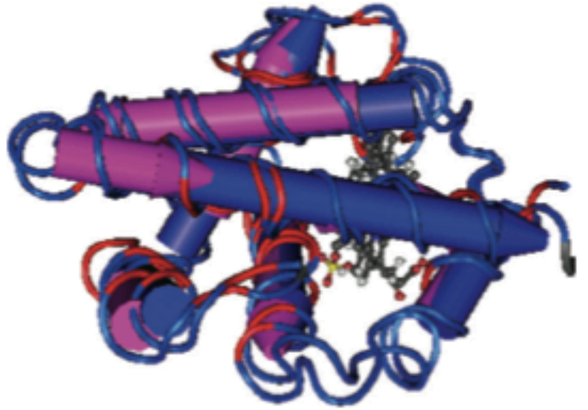
(b) Human hemoglobin tetramer (2H35)



(c) Human beta globin (subunit of 2H35)



(d) Pairwise alignment of beta globin and myoglobin





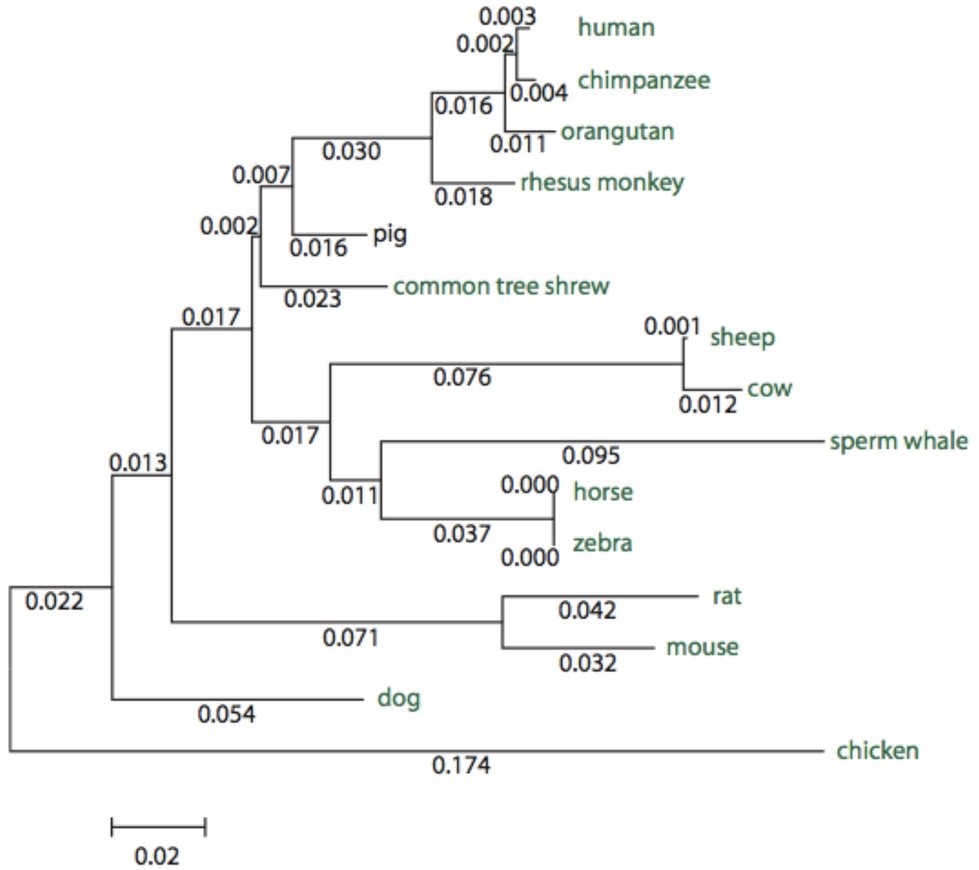
ian.simpson@ed.ac.uk

# Types of Homology

## Orthologs

Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function.

Globins in different species





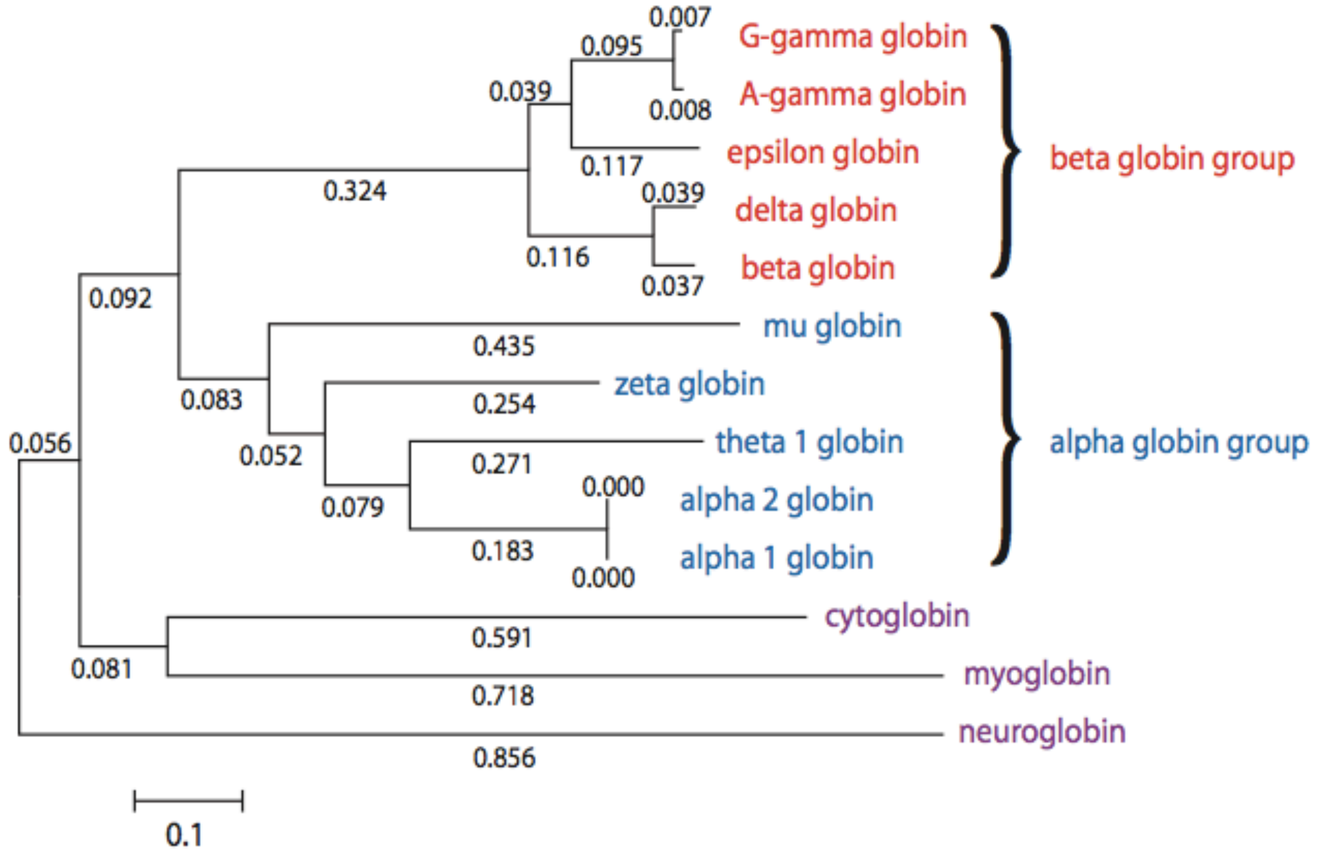
ian.simpson@ed.ac.uk

# Types of Homology

## Paralogs

Homologous sequences within a single species that arose by gene duplication.

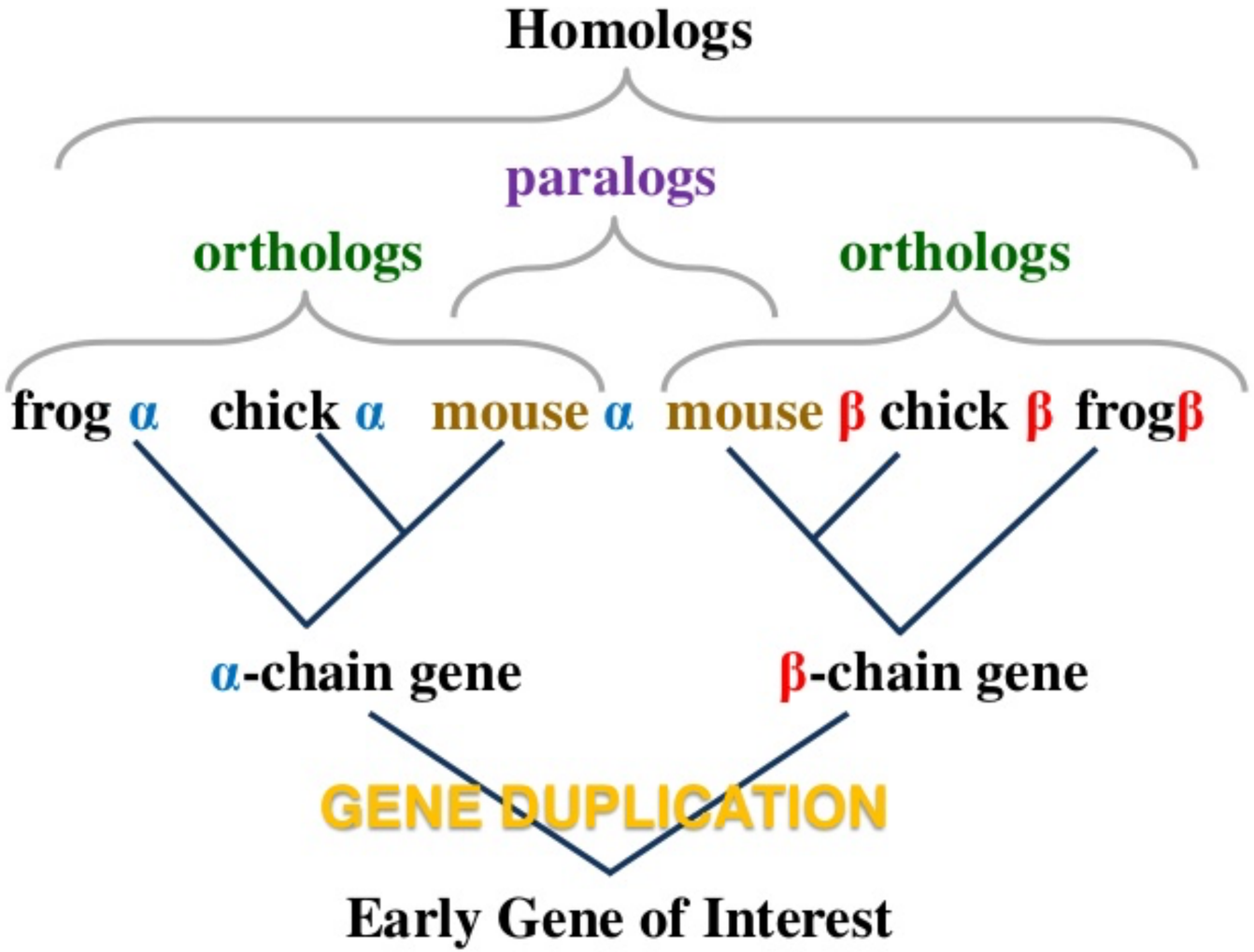
Globins in humans






ian.simpson@ed.ac.uk

# Orthologs and Paralogs Together





ian.simpson@ed.ac.uk

# General Approach to Pairwise Alignment

- Choose two sequences
- Select an algorithm that generates a score
- Allow gaps (insertions, deletions)
- Score reflects degree of similarity
- Alignments can be global or local
- Estimate probability that the alignment occurred by chance


ian.simpson@ed.ac.uk

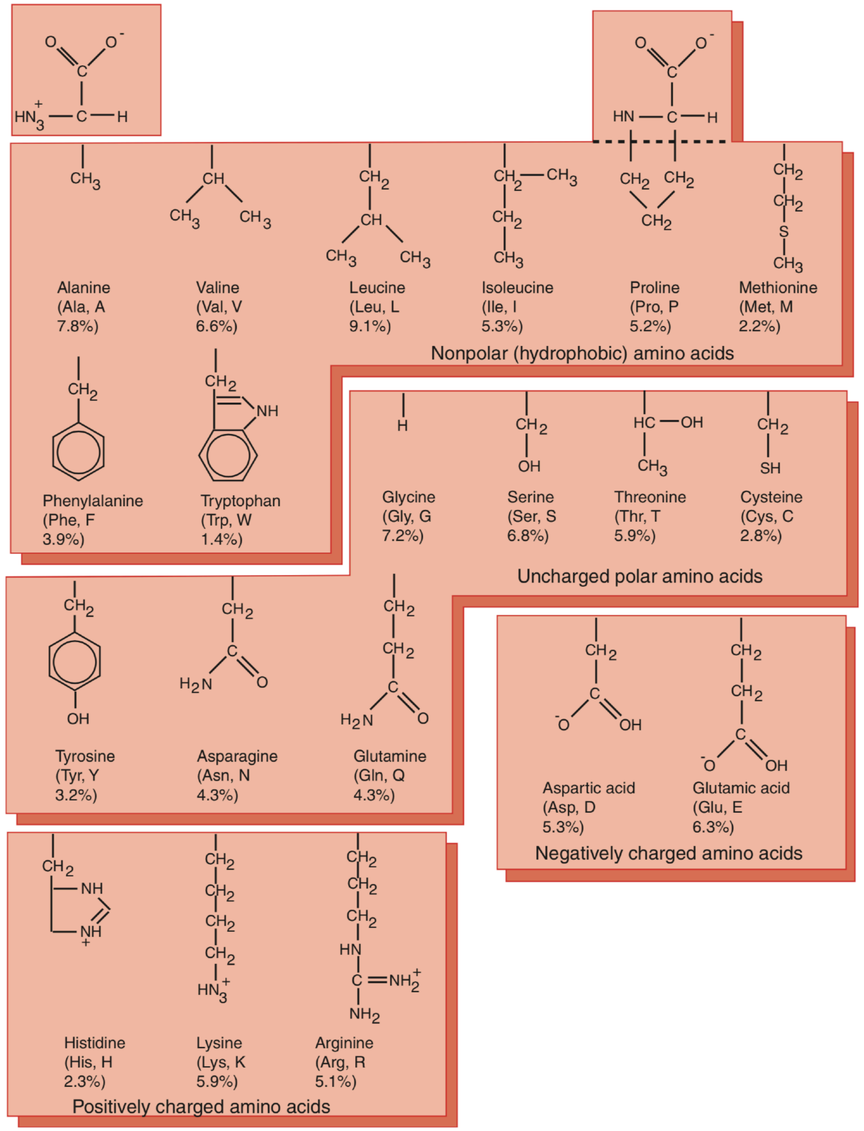
# Human Beta-Globin & Myoglobin

```
#=======================================
#
# Aligned_sequences: 2
# 1: NP_000509.1
# 2: NP_005359.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 155
# Identity:      37/155 (23.9%)
# Similarity:    58/155 (37.4%)
# Gaps:           9/155 ( 5.8%)
# Score: 110.5
#
#
#=======================================

NP_000509.1         1 MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFGD     48
                      :.|:..|...|..:||||..|  ..|.|.|.||...:|.|...|:.|..
NP_005359.1         1 -MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKH     49

NP_000509.1        49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH     98
                      |.:.|.:..:..:|.||..||.|....|....:.:.....|::.|..|..
NP_005359.1        50 LKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHK     99

NP_000509.1        99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-    147
                      :..:...:...::.||......:|....|.|..|.:......:|..|.
NP_005359.1       100 IPVKYLEFISECIIQVLQSKHPGDFGADAQGAMNKALELFRKDMASNYKE    149

NP_000509.1       148 -----    147

NP_005359.1       150 LGFQG    154


#---------------------------------------
#---------------------------------------
```



ian.simpson@ed.ac.uk

# Amino Acids Share Properties





ian.simpson@ed.ac.uk

# Genetic Code

**Standard genetic code**

| 1st base | 2nd base | | | | | | | | 3rd base |
|---|---|---|---|---|---|---|---|---|---|
| | T | | C | | A | | G | | |
| T | TTT | (Phe/F) Phenylalanine | TCT | (Ser/S) Serine | TAT | (Tyr/Y) Tyrosine | TGT | (Cys/C) Cysteine | T |
| | TTC | | TCC | | TAC | | TGC | | C |
| | TTA | (Leu/L) Leucine | TCA | | TAA | Stop (*Ochre*) | TGA | Stop (*Opal*) | A |
| | TTG | | TCG | | TAG | Stop (*Amber*) | TGG | (Trp/W) Tryptophan | G |
| C | CTT | | CCT | (Pro/P) Proline | CAT | (His/H) Histidine | CGT | (Arg/R) Arginine | T |
| | CTC | | CCC | | CAC | | CGC | | C |
| | CTA | | CCA | | CAA | (Gln/Q) Glutamine | CGA | | A |
| | CTG | | CCG | | CAG | | CGG | | G |
| A | ATT | (Ile/I) Isoleucine | ACT | (Thr/T) Threonine | AAT | (Asn/N) Asparagine | AGT | (Ser/S) Serine | T |
| | ATC | | ACC | | AAC | | AGC | | C |
| | ATA | | ACA | | AAA | (Lys/K) Lysine | AGA | (Arg/R) Arginine | A |
| | ATG[A] | (Met/M) Methionine | ACG | | AAG | | AGG | | G |
| G | GTT | (Val/V) Valine | GCT | (Ala/A) Alanine | GAT | (Asp/D) Aspartic acid | GGT | (Gly/G) Glycine | T |
| | GTC | | GCC | | GAC | | GGC | | C |
| | GTA | | GCA | | GAA | (Glu/E) Glutamic acid | GGA | | A |
| | GTG | | GCG | | GAG | | GGG | | G |

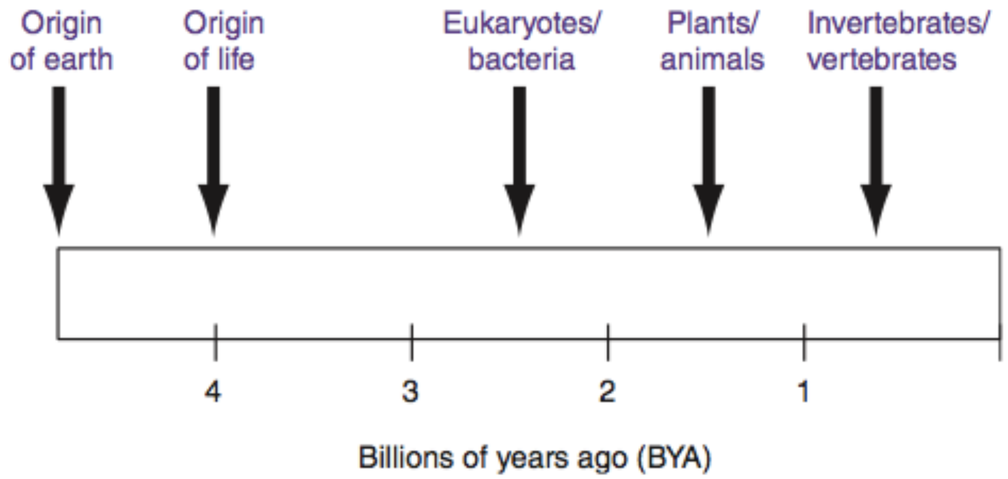

ian.simpson@ed.ac.uk

# Gaps in Alignments

- Positions at which a letter is paired with a null are called gaps
- Gap scores are typically negative
- Since a single mutational event may cause the insertion or deletion of more than one residue, the presence of a gap is ascribed more significance than the length of the gap. Thus there are separate penalties for **gap creation** and **gap extension**


ian.simpson@ed.ac.uk

# Pairwise Alignment & Evolution




- When two proteins (or DNA sequences) are homologous they share a common ancestor
- When we align globins from human and a plant we can imagine their common ancestor, a single celled organism that lived 1.5 billion years ago, and we can infer that ancient globin sequence
- Through pairwise alignment we can look back in time at sequence evolution


ian.simpson@ed.ac.uk

# Accepted Point Mutations (PAMs) in Protein Families

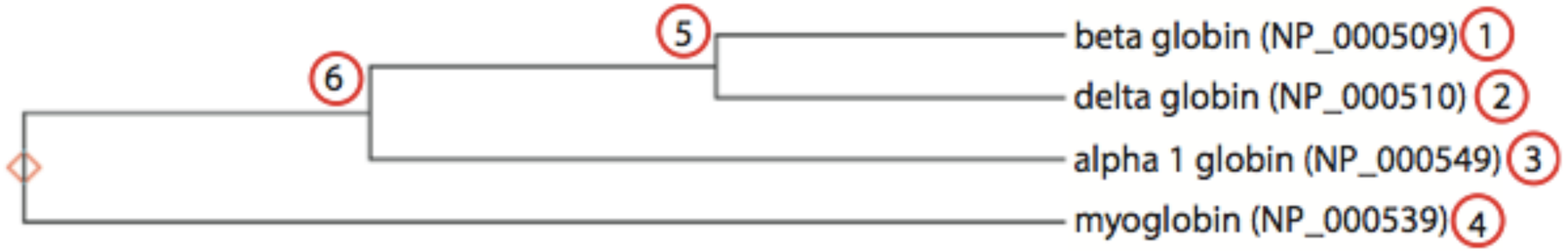
| PROTEIN | PAMS PER 100 MILLION YEARS |
|---|---|
| Immunoglobulin (Ig) kappa chain C region | 37 |
| Kappa casein | 33 |
| Epidermal growth factor | 26 |
| Serum albumin | 19 |
| Hemoglobin alpha chain | 12 |
| Myoglobin | 8.9 |
| Nerve growth factor | 8.5 |
| Trypsin | 5.9 |
| Insulin | 4.4 |
| Cytochrome c | 2.2 |
| Glutamate dehydrogenase | 0.9 |
| Histone H3 | 0.14 |
| Histone H4 | 0.10 |

Margaret Dayhoff and colleagues developed scoring matrices in the 1960s and 1970s. They defined PAMs as "accepted point mutations." Some protein families evolve very slowly (e.g. histones change little over 100 million years); others (such as kappa casein) change very rapidly.


ian.simpson@ed.ac.uk

# PAMs Are Defined Relative to a Common Ancestor

```
beta globin      MVHLTPEEKSAVTALWGKV
delta globin     MVHLTPEEKTAVNALWGKV
alpha 1 globin   MV.LSPADKTNVKAAWGKV
myoglobin        .MGLSDGEWQLVLNVWGKV
5                MVHLSPEEKTAVNALWGKV
6                MVHLTPEEKTAVNALWGKV
```




Dayhoff et al. evaluated amino acid changes. They applied an evolutionary model to compare changes between for example @1 and @2 and an inferred common ancestor @5


ian.simpson@ed.ac.uk

# Amino Acid Frequency

- If 20 amino acids occurred in nature at equal frequencies, each would be observed 5% of the time
- Some are more common (G, A, L, K) and some rare (C, Y, M, W).

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

- blue=6 codons; red=1 codon in the genetic code


ian.simpson@ed.ac.uk

# Amino Acid Substitutions

| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly | H His | I Ile | L Leu | K Lys | M Met | F Phe | P Pro | S Ser | T Thr | W Trp | Y Tyr | V Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | | | | | | | | | | | | | |
| R | 30 | | | | | | | | | | | | | | | | | | | |
| N | 109 | 17 | | | | | | | | | | | | | | | | | | |
| D | 154 | 0 | 532 | | | | | | | | | | | | | | | | | |
| C | 33 | 10 | 0 | 0 | | | | | | | | | | | | | | | | |
| Q | 93 | 120 | 50 | 76 | 0 | | | | | | | | | | | | | | | |
| E | 266 | 0 | 94 | 831 | 0 | 422 | | | | | | | | | | | | | | |
| G | 579 | 10 | 156 | 162 | 10 | 30 | 112 | | | | | | | | | | | | | |
| H | 21 | 103 | 226 | 43 | 10 | 243 | 23 | 10 | | | | | | | | | | | | |
| I | 66 | 30 | 36 | 13 | 17 | 8 | 35 | 0 | 3 | | | | | | | | | | | |
| L | 95 | 17 | 37 | 0 | y | 75 | 15 | 17 | 40 | 253 | | | | | | | | | | |
| K | 57 | 477 | 322 | 85 | 0 | 147 | 104 | 60 | 23 | 43 | 39 | | | | | | | | | |
| M | 29 | 17 | 0 | 0 | 0 | 20 | 7 | 7 | 0 | 57 | 207 | 90 | | | | | | | | |
| F | 20 | 7 | 7 | 0 | 0 | 0 | 0 | 17 | 20 | 90 | 167 | 0 | 17 | | | | | | | |
| P | 345 | 67 | 27 | 10 | 10 | 93 | 40 | 49 | 50 | 7 | 43 | 43 | 4 | 7 | | | | | | |
| S | 772 | 137 | 432 | 98 | 117 | 47 | 86 | 450 | 26 | 20 | 32 | 168 | 20 | 40 | 269 | | | | | |
| T | 590 | 20 | 169 | 57 | 10 | 37 | 31 | 50 | 14 | 129 | 52 | 200 | 28 | 10 | 73 | 696 | | | | |
| W | 0 | 27 | 3 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 13 | 0 | 0 | 10 | 0 | 17 | 0 | | | |
| Y | 20 | 3 | 36 | 0 | 30 | 0 | 10 | 0 | 40 | 13 | 23 | 10 | 0 | 260 | 0 | 22 | 23 | 6 | | |
| V | 365 | 20 | 13 | 17 | 33 | 27 | 37 | 97 | 30 | 661 | 303 | 17 | 77 | 10 | 50 | 43 | 186 | 0 | 17 | |
| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly | H His | I Ile | L Leu | K Lys | M Met | F Phe | P Pro | S Ser | T Thr | W Trp | Y Tyr | V Val |

Dayhoff survey of 1572 observed substitutions
(original amino acid in columns and changes in rows)


ian.simpson@ed.ac.uk

# Substitution Matrix

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


ian.simpson@ed.ac.uk

# PAM Matrices

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


ian.simpson@ed.ac.uk

# Mutation Probability Matrix for the Evolutionary Distance of 1 PAM

| Replacement amino acid | Original amino acid: A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly | H His | I Ile | L Leu | K Lys | M Met | F Phe | P Pro | S Ser | T Thr | W Trp | Y Tyr | V Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 98.7 | 0.0 | 0.1 | 0.1 | 0.0 | 0.1 | 0.2 | 0.2 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 0.2 | 0.4 | 0.3 | 0.0 | 0.0 | 0.2 |
| R | 0.0 | 99.1 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 |
| N | 0.0 | 0.0 | 98.2 | 0.4 | 0.0 | 0.0 | 0.1 | 0.1 | 0.2 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.2 | 0.1 | 0.0 | 0.0 | 0.0 |
| D | 0.1 | 0.0 | 0.4 | 98.6 | 0.0 | 0.1 | 0.5 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| C | 0.0 | 0.0 | 0.0 | 0.0 | 99.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| Q | 0.0 | 0.1 | 0.0 | 0.1 | 0.0 | 98.8 | 0.3 | 0.0 | 0.2 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| E | 0.1 | 0.0 | 0.1 | 0.6 | 0.0 | 0.4 | 98.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| G | 0.2 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.1 | 99.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 |
| H | 0.0 | 0.1 | 0.2 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 99.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| I | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 98.7 | 0.1 | 0.0 | 0.2 | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.3 |
| L | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.2 | 99.5 | 0.0 | 0.5 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 |
| K | 0.0 | 0.4 | 0.3 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 99.3 | 0.2 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 |
| M | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 98.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| F | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 | 99.5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 | 0.0 |
| P | 0.1 | 0.1 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 99.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 |
| S | 0.3 | 0.1 | 0.3 | 0.1 | 0.1 | 0.0 | 0.1 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 0.2 | 98.4 | 0.4 | 0.1 | 0.0 | 0.0 |
| T | 0.2 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.1 | 0.1 | 0.0 | 0.1 | 0.3 | 98.7 | 0.0 | 0.0 | 0.1 |
| W | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 99.8 | 0.0 | 0.0 |
| Y | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.0 | 99.5 | 0.0 |
| V | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.6 | 0.1 | 0.0 | 0.2 | 0.0 | 0.0 | 0.0 | 0.1 | 0.0 | 0.0 | 99.0 |

- This mutation probability matrix includes original amino acids (columns) and replacements (rows)
- The diagonals show that at a distance of 1 PAM most residues remain the same about 99% of the time
- Note how cysteine (C) and tryptophan (W) undergo few substitutions, and asparagine (N) many


ian.simpson@ed.ac.uk

# PAM250 and Other PAM Matrices

| | | | |
|---|---|---|---|
| NP_002037.2 | 164 | IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII | 207 |
| XP_001162057.1 | 164 | IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNII | 207 |
| NP_001003142.1 | 162 | IHDHFGIVEGLMTTVHAITATQKTVDGPSGKMWRDGRGAAQNII | 205 |
| XP_893121.1 | 168 | IHDNFGIMEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII | 211 |
| XP_576394.1 | 162 | IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII | 205 |
| NP_058704.1 | 162 | IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII | 205 |
| XP_001070653.1 | 162 | IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII | 205 |
| XP_001062726.1 | 162 | IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII | 205 |
| NP_989636.1 | 162 | IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGAAQNII | 205 |
| NP_525091.1 | 161 | INDNFEIVEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII | 204 |
| XP_318655.2 | 161 | INDNFGILEGLMTTVHATTATQKTVDGPSGKLWRDGRGAAQNII | 204 |
| NP_508535.1 | 170 | INDNFGIIEGLMTTVHAVTATQKTVDGPSGKLWRDGRGAGQNII | 213 |
| NP_595236.1 | 164 | INDTFGIEEGLMTTVHATTATQKTVDGPSKKDWRGGRGASANII | 207 |
| NP_011708.1 | 162 | INDAFGIEEGLMTTVHSLTATQKTVDGPSHKDWRGGRTASGNII | 205 |
| XP_456022.1 | 161 | INDEFGIDEALMTTVHSITATQKTVDGPSHKDWRGGRTASGNII | 204 |
| NP_001060897.1 | 166 | IHDNFGIIEGLMTTVHAITATQKTVDGPSSKDWRGGRAASFNII | 209 |

Consider a multiple alignment of glyceraldehyde 3-phosphate protein sequences. Some substitutions are observed in columns (arrowheads). These give us insight into changes tolerated by natural selection.


ian.simpson@ed.ac.uk

# PAM250 and Other PAM Matrices


- Now consider the alignment of distantly related kappa caseins
- There are few conserved column positions, and some columns (double arrowheads) have five different residues among the 7 proteins
- We want to design a scoring system that is tolerant of distantly related proteins, if the scoring system is too strict the divergent sequences may be penalised so heavily that authentic homologs are not identified or aligned.


ian.simpson@ed.ac.uk

# PAM Extremes

original amino acid

replacement amino acid

| PAM0 | A | R | N | D | C | Q | E | G |
|---|---|---|---|---|---|---|---|---|
| A | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| R | 0 | 100 | 0 | 0 | 0 | 0 | 0 | 0 |
| N | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 100 | 0 | 0 | 0 | 0 |
| C | 0 | 0 | 0 | 0 | 100 | 0 | 0 | 0 |
| Q | 0 | 0 | 0 | 0 | 0 | 100 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 |
| G | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 100 |

original amino acid

replacement amino acid

| PAM∞ | A | R | N | D | C | Q | E | G |
|---|---|---|---|---|---|---|---|---|
| A | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 | 8.7 |
| R | 4.1 | 4.1 | 4.1 | 4.1 | 4.1 | 4.1 | 4.1 | 4.1 |
| N | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 |
| D | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 |
| C | 3.3 | 3.3 | 3.3 | 3.3 | 3.3 | 3.3 | 3.3 | 3.3 |
| Q | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 | 3.8 |
| E | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 |
| G | 8.9 | 8.9 | 8.9 | 8.9 | 8.9 | 8.9 | 8.9 | 8.9 |

NB - remember amino acid frequencies!


ian.simpson@ed.ac.uk

# PAM250 for Proteins That Share ~20% Identity

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


ian.simpson@ed.ac.uk

# Mutation Probability Matrix -> Relatedness Odds Matrix

$$R_{ij} = \frac{M_{ij}}{f_i}$$

- A relatedness odds matrix reports the probability that amino acid *j* will change to *i* in a homologous sequence.
- The numerator models the observed change. The denominator *fi* is the probability of amino acid residue *i* occurring in the second sequence by chance.
- A positive value indicates a replacement happens more often than expected by chance. A negative value indicates the replacement is not favoured
- We actually use log-odds; easier to use for a scoring system. They allow us to sum the scores of aligned residues (rather than having to multiply them).


ian.simpson@ed.ac.uk

# Mutation Probability Matrix to a Log-Odds Matrix

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

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

The score S for an alignment of residues a,b is given by:

$S(a,b) = 10 \log_{10} (M_{ab}/p_b)$

As an example, for tryptophan,

$S(trp,trp) = 10 \log_{10} (0.55/0.010) = 17.4$

$$s_{i,j} = 10 \times \log\left(\frac{q_{i,j}}{p_{i,j}}\right)$$


ian.simpson@ed.ac.uk

# Meaning of the Values in a Log Odds Matrix

- A score of +2 indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance
- A score of 0 is neutral
- A score of –10 indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one tenth as frequent as the chance alignment of these amino acids.


ian.simpson@ed.ac.uk

# Log-Odds Matrix for PAM250

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 6 | | | | | | | | | |
| -3 | 5 | | | | | | | | |
| 4 | 0 | 6 | | | | | | | |
| 2 | -5 | 0 | 9 | | | | | | |
| -3 | -1 | -2 | -5 | 6 | | | | | |
| -3 | 0 | -2 | -3 | 1 | 2 | | | | |
| -2 | 0 | -1 | -3 | 0 | 1 | 3 | | | |
| -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | | |
| -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | |
| 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |
| **L** | **K** | **M** | **F** | **P** | **S** | **T** | **W** | **Y** | **V** |

This is a useful matrix for comparing distantly related proteins

Note alignment of two tryptophan (W) residues earns +17 and a W to T mismatch is -5.


ian.simpson@ed.ac.uk

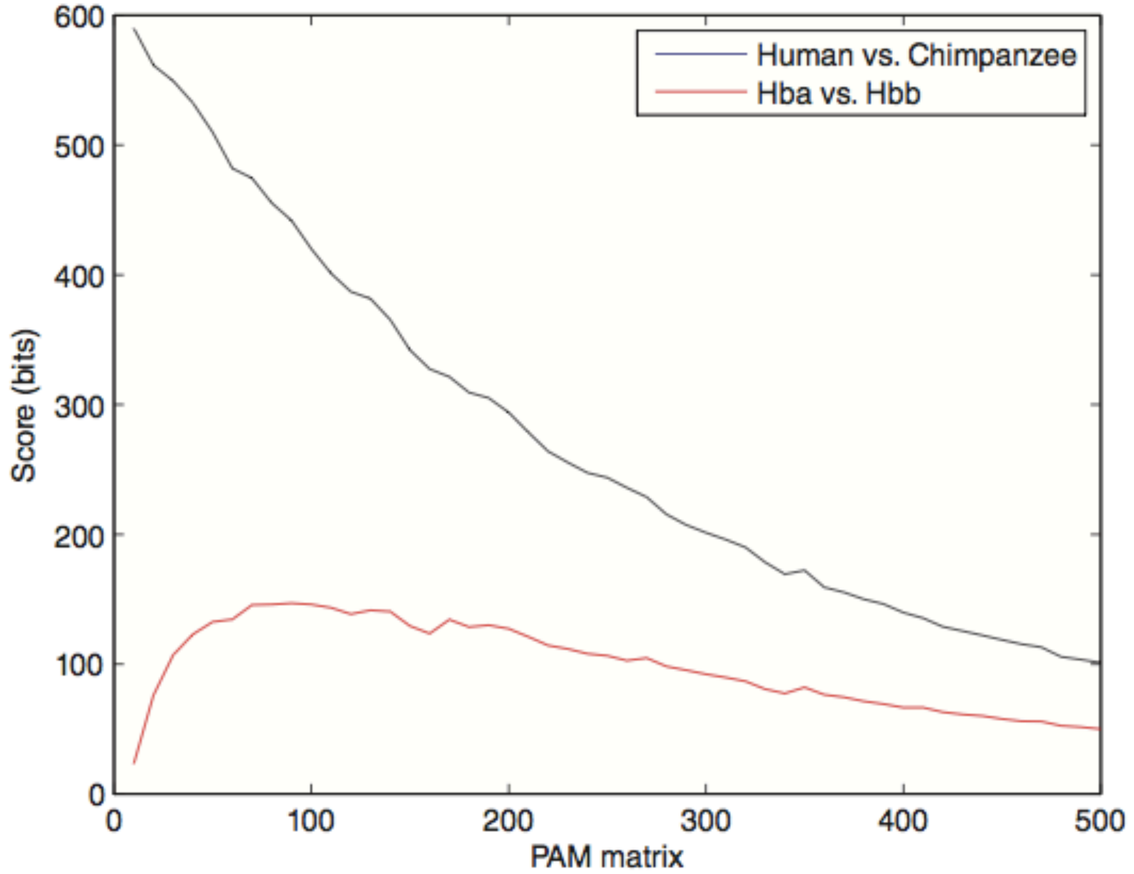
# Log-Odds Matrix for PAM10

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

This is an example of a scoring matrix with “severe” penalties. A match of W to W earns +13, but a mismatch (e.g. W aligned to T) has a score of -19, far lower than in PAM250.


ian.simpson@ed.ac.uk

# Effect of Scoring Matrix on Scores




Take home message - PAM10 scores closely related sequences well and distantly related ones poorly

Look at score for distantly related proteins (e.g. beta globin versus alpha globin) and note that PAM10 or similar matrices assign very low scores. This effect is not seen for very closely related proteins (e.g. a chimp vs. human globin).


ian.simpson@ed.ac.uk

# BLOSUM62 Scoring Matrix

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

BLOSUM62 is the default scoring matrix at the NCBI BLAST site.

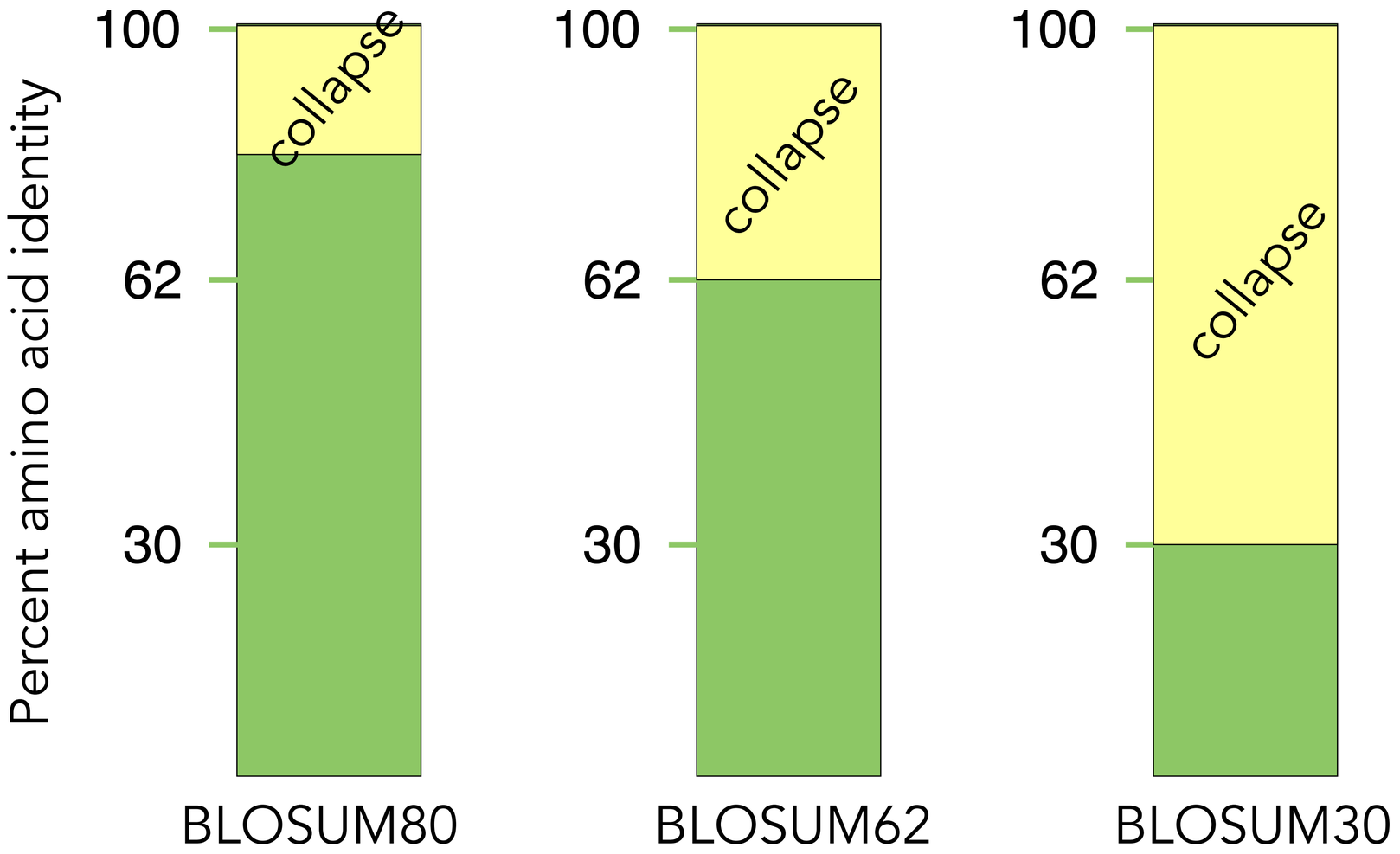

ian.simpson@ed.ac.uk

# BLOSUM Matrices

- BLOSUM matrices are based on local alignments
- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins
- BLOSUM stands for blocks substitution matrix
- BLOSUM62 is a matrix calculated from comparisons of sequences with at least 62% similarity
- BLOSUM62 is the default matrix in BLAST 2.0


ian.simpson@ed.ac.uk

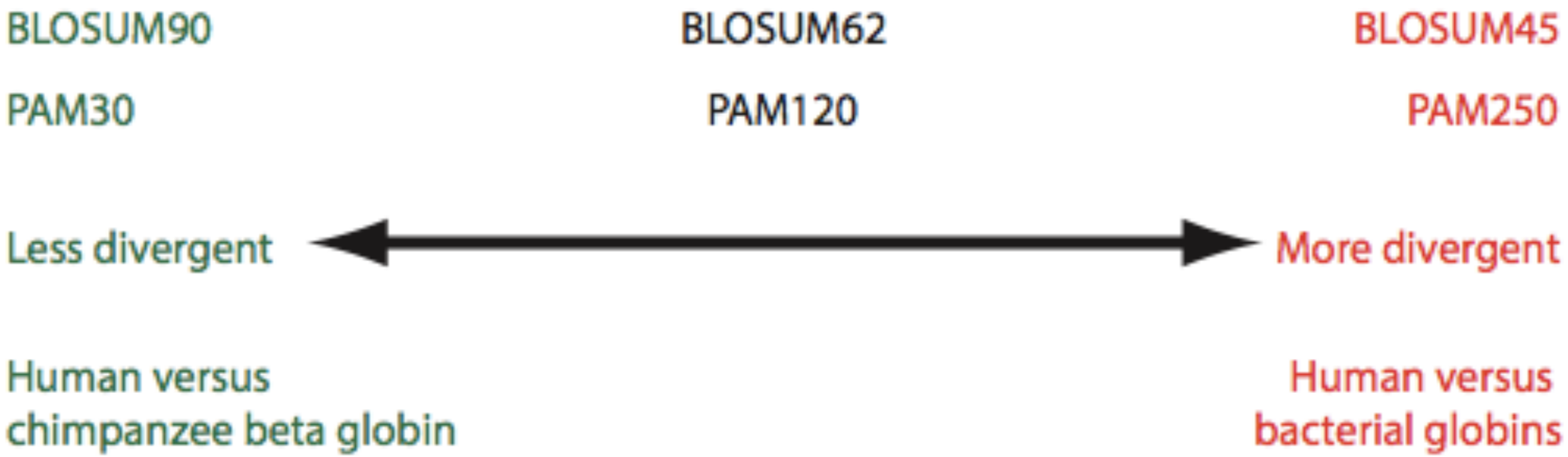
# BLOSUM Matrices





ian.simpson@ed.ac.uk

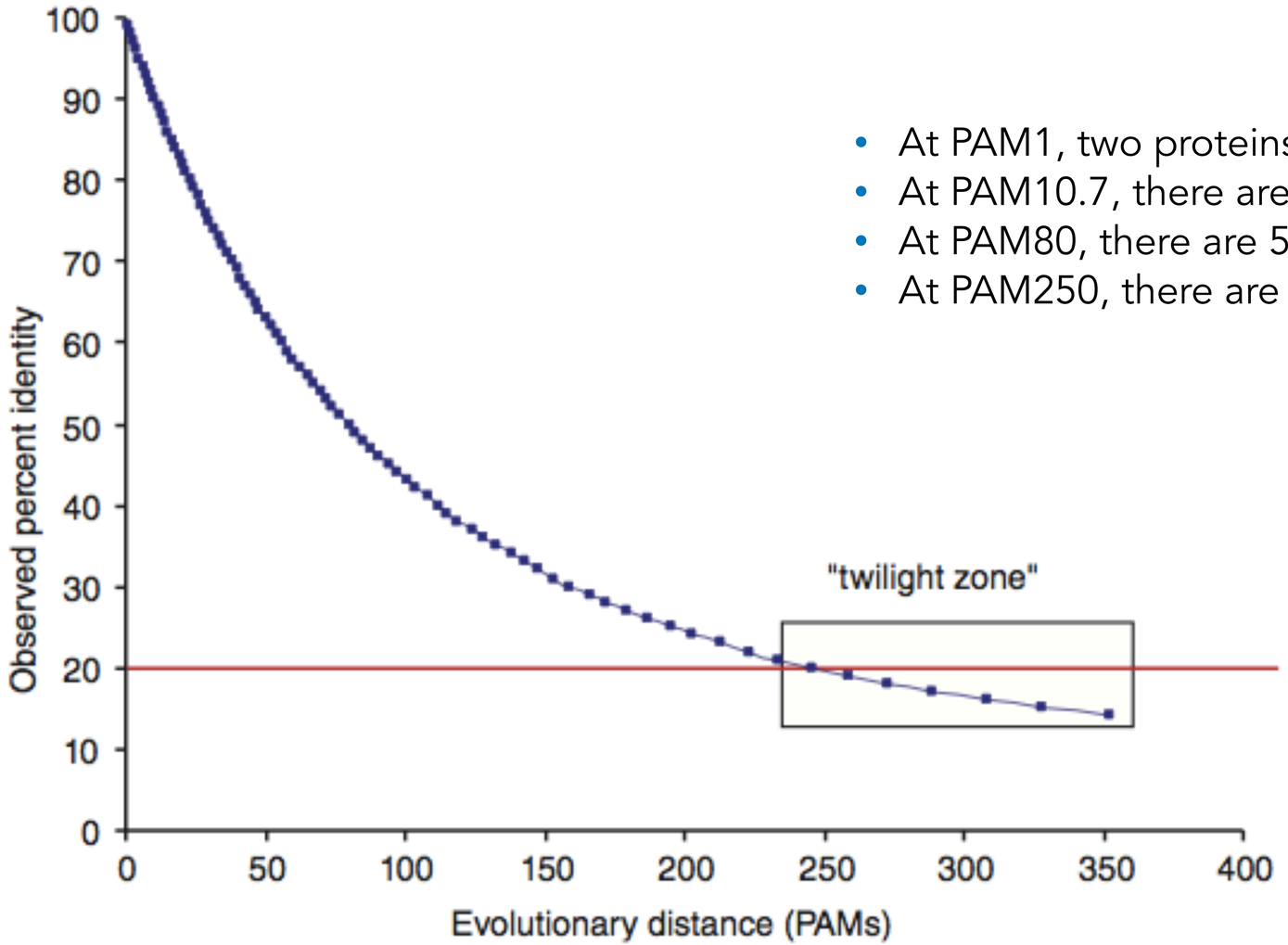
# Summary of PAM and BLOSUM Matrices




A higher PAM number, and a lower BLOSUM number, tends to correspond to a matrix tuned to more divergent proteins.


ian.simpson@ed.ac.uk

# Experiment With Randomly Diverging Protein Sequences




- At PAM1, two proteins are 99% identical
- At PAM10.7, there are 10 differences per 100 residues
- At PAM80, there are 50 differences per 100 residues
- At PAM250, there are 80 differences per 100 residues


ian.simpson@ed.ac.uk

# Global & Local Sequence Alignment

Global alignment algorithm - Needleman and Wunsch (1970)

Local alignment algorithm - Smith and Waterman (1981)

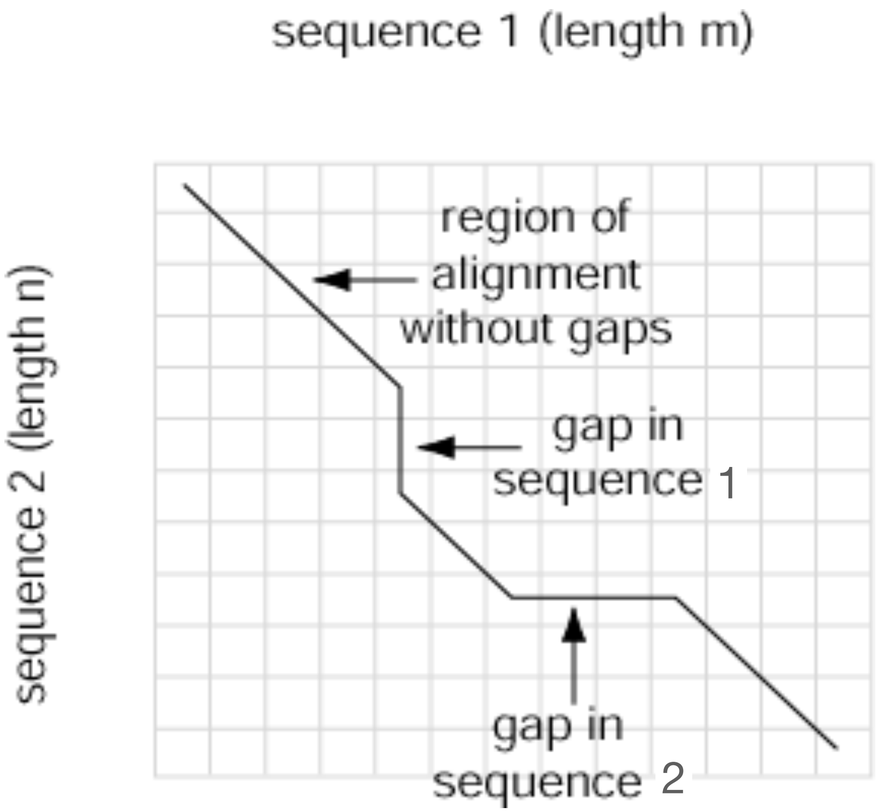

ian.simpson@ed.ac.uk

# Global Alignment - Needleman and Wunsch (1970)

- Two sequences can be compared in a matrix along x- and y-axes.
- If they are identical, a path along a diagonal can be drawn
- Find the optimal subpaths, and add them up to achieve the best score
  - adding gaps when needed
  - allowing for conservative substitutions
  - choosing a scoring system (simple or complicated)
- N-W is guaranteed to find optimal alignment(s)


ian.simpson@ed.ac.uk

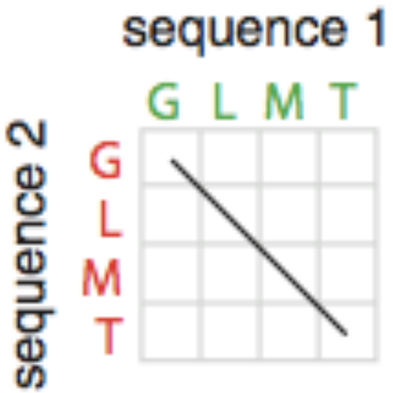
# Four Possible Events During Alignment



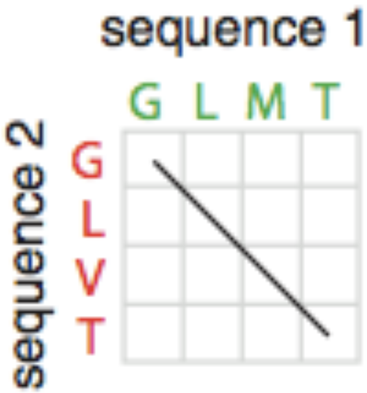

- identity - stay on diagonal
- mismatch - stay on diagonal
- gap in one sequence - move vertically
- gap in the other sequence - move horizontally


ian.simpson@ed.ac.uk

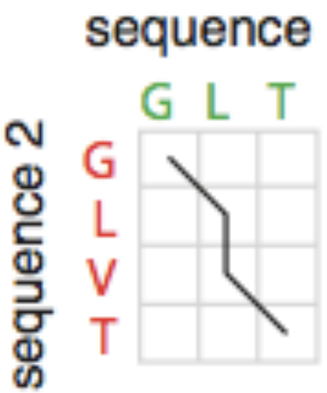
# Four Possible Events During Alignment



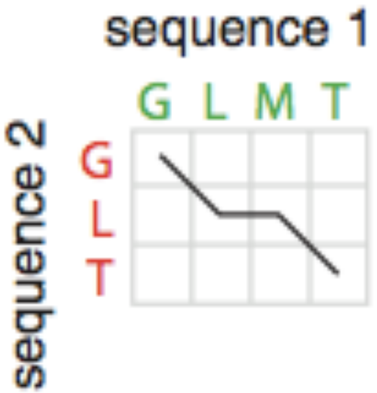

1 GLMT
2 GLMT




1 GLMT
2 GLVT



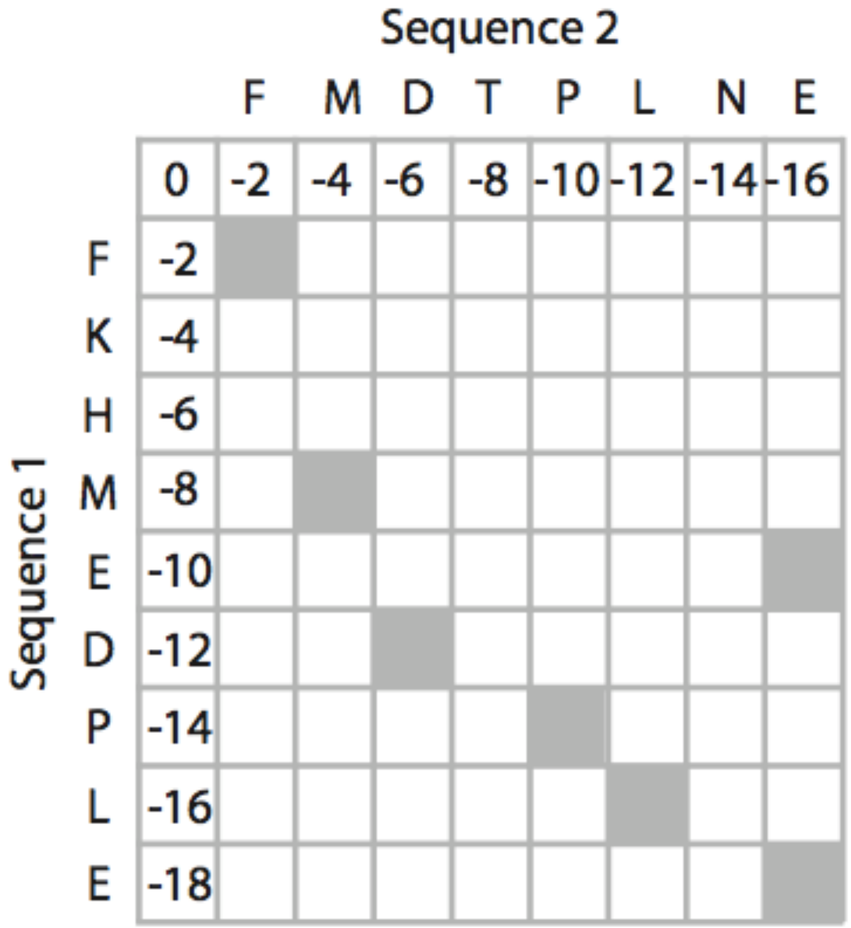

1 GL-T
2 GLVT




1 GLMT
2 GL-T


ian.simpson@ed.ac.uk

# Global Pairwise Alignment Using Needleman-Wunsch




- Identify positions of identity (shaded grey)
- NB gap-penalty = -2


ian.simpson@ed.ac.uk

# Global Pairwise Alignment Using Needleman-Wunsch

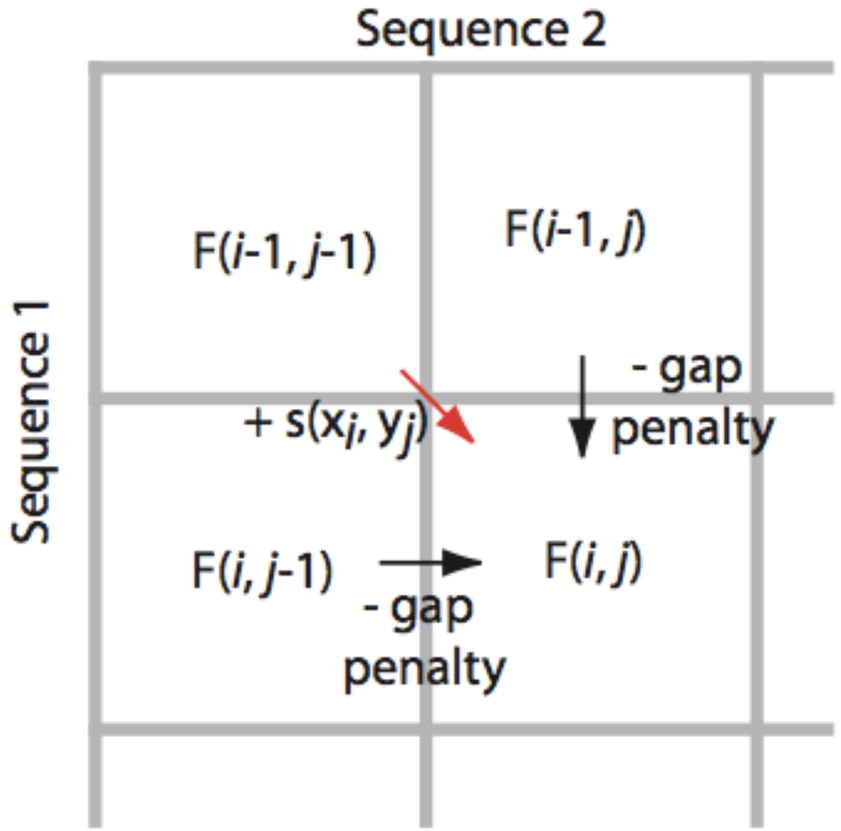
$$\text{Score} = \text{Max} \begin{cases} F(i-1, j-1) + s(x_i, y_i) \\ F(i-1, j) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

Score (this example) = +1 (match)
-2 (mismatch)
-2 (gap penalty)

Define an overall score that maximises cumulative scores at each position of the pairwise alignment,  allowing for substitutions and gaps in either sequence.


ian.simpson@ed.ac.uk

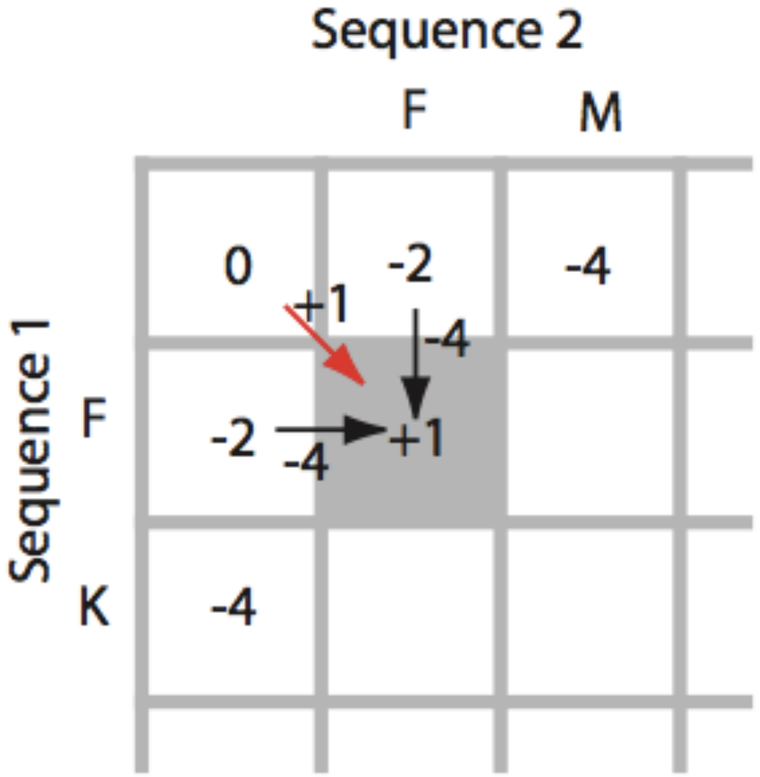
# Global Pairwise Alignment Using Needleman-Wunsch




To decide how to align sequences 1 and 2 in the box at lower right, decide what the scores are beginning at upper left (not requiring a gap), or beginning from the left or top (each requiring a gap penalty).


ian.simpson@ed.ac.uk

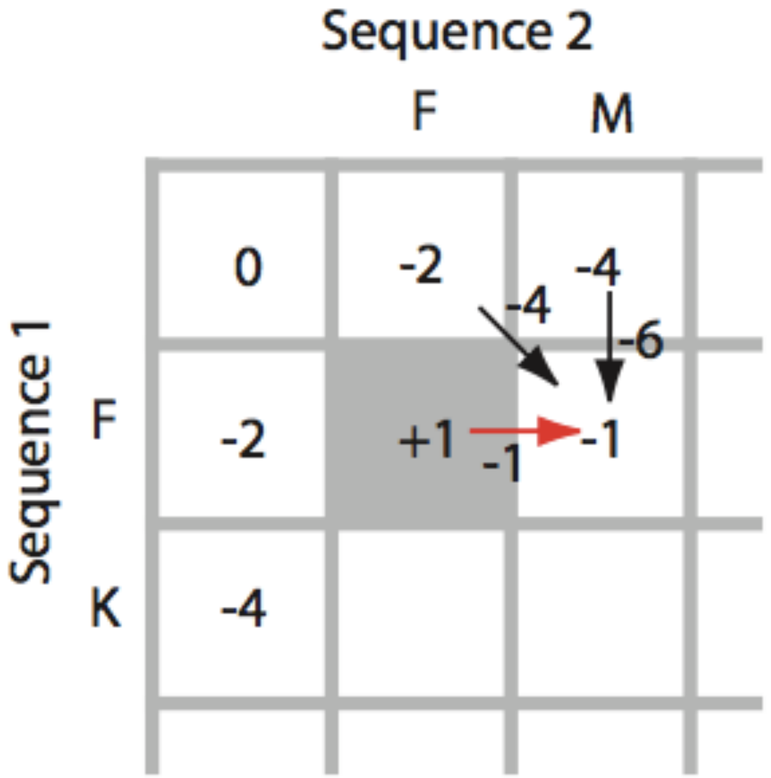
# Global Pairwise Alignment Using Needleman-Wunsch




Here the best score involves +1 (proceed from upper left to grey, lower right square).
If we instead select an alignment involving a gap the score would be worse (-4).


ian.simpson@ed.ac.uk

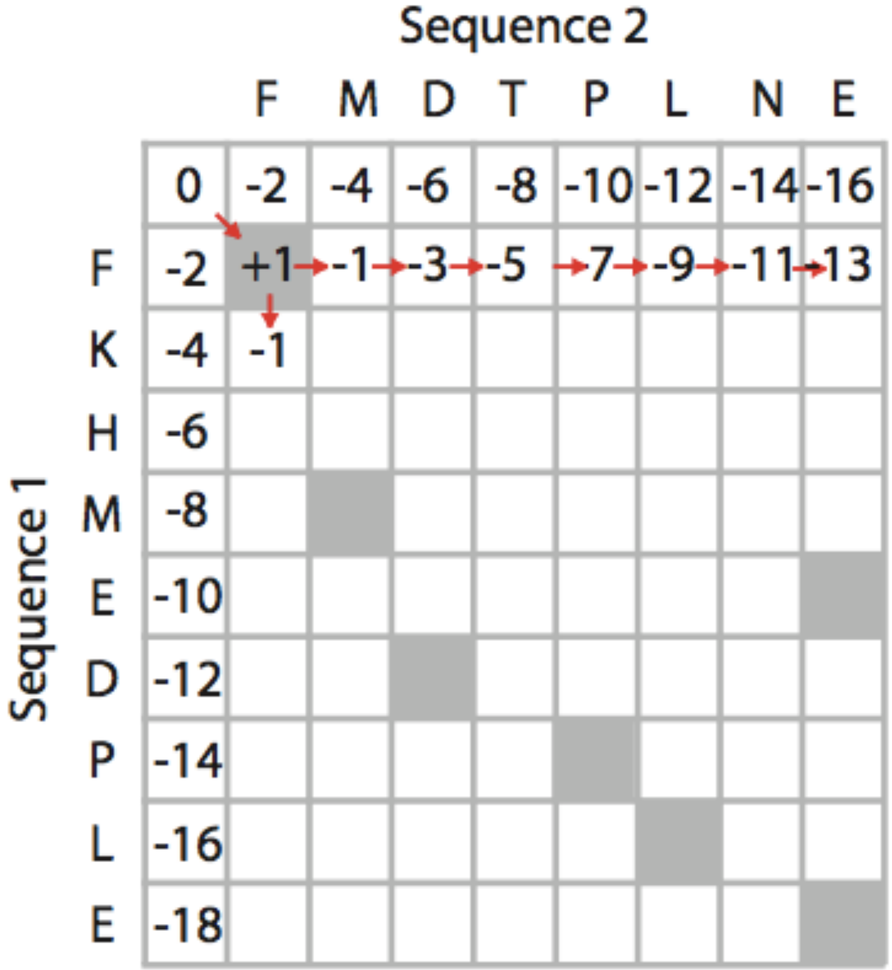
# Global Pairwise Alignment Using Needleman-Wunsch




Proceed to calculate the optimal score for the next position


ian.simpson@ed.ac.uk

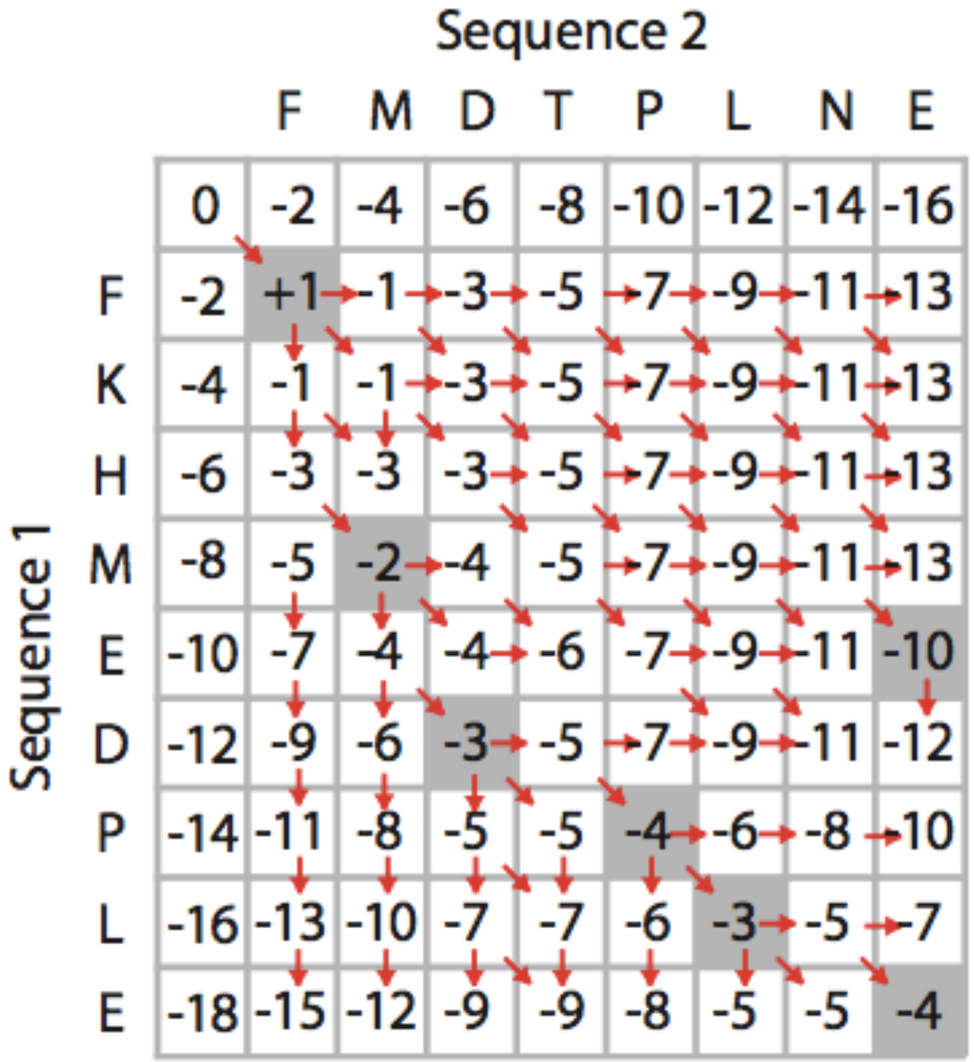
# Global Pairwise Alignment Using Needleman-Wunsch




Continue filling in the matrix


ian.simpson@ed.ac.uk

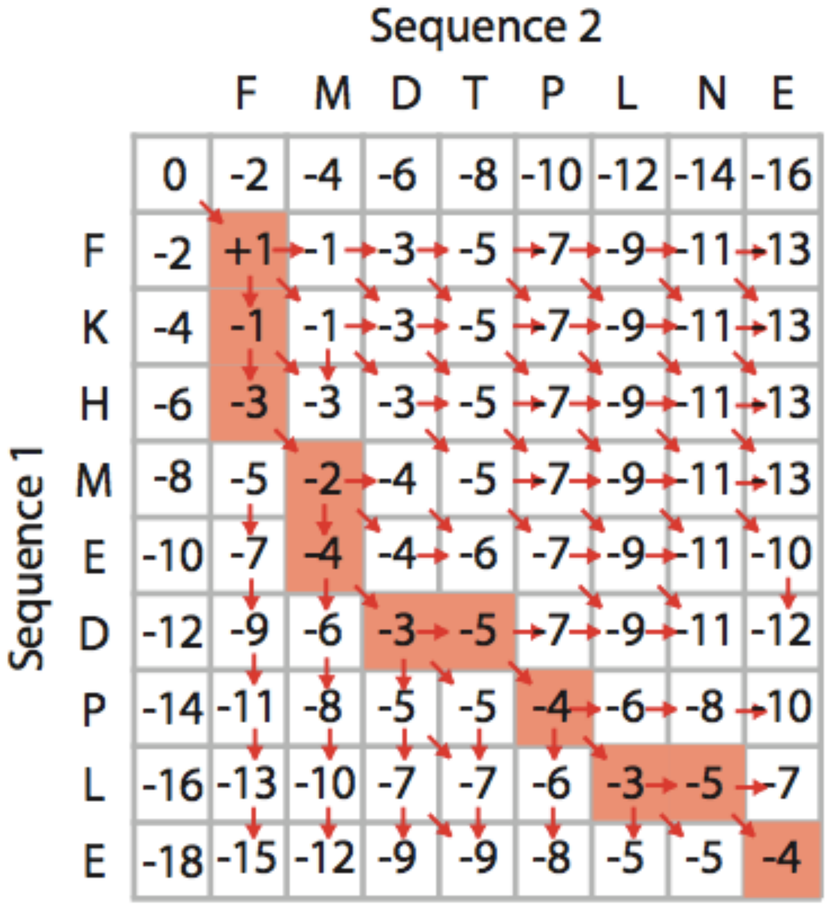
# Global Pairwise Alignment Using Needleman-Wunsch




Complete filling in the matrix


ian.simpson@ed.ac.uk

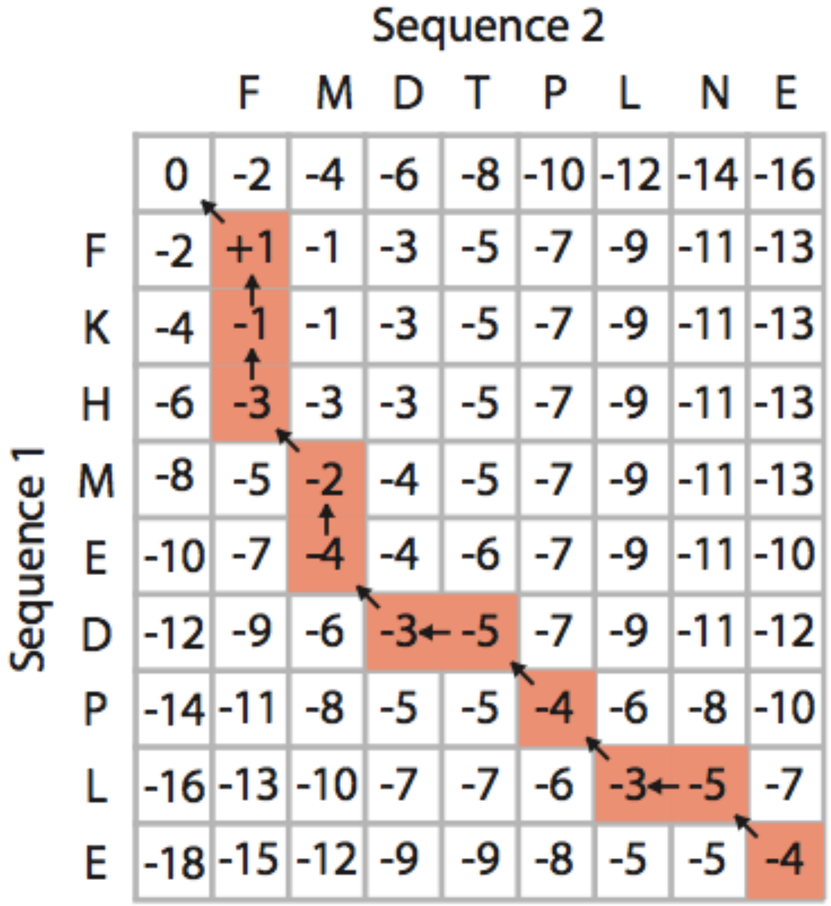
# Global Pairwise Alignment Using Needleman-Wunsch




Highlighted cells indicate the optimal path (best scores), indicating how the two sequences should be aligned


ian.simpson@ed.ac.uk

# Global Pairwise Alignment Using Needleman-Wunsch




Equivalent representation, showing the traceback procedure:
begin at the bottom right cell and proceed to the top left


ian.simpson@ed.ac.uk

# Global Pairwise Alignment Using Needleman-Wunsch

Really nice interactive tool:
https://bioboot.github.io/bimm143_W20/class-material/nw/

Sequence 2

| Sequence 1 | | F | M | D | T | P | L | N | E |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | -2 | -4 | -6 | -8 | -10 | -12 | -14 | -16 |
| F | -2 | +1 | -1 | -3 | -5 | -7 | -9 | -11 | -13 |
| K | -4 | -1 | -1 | -3 | -5 | -7 | -9 | -11 | -13 |
| H | -6 | -3 | -3 | -3 | -5 | -7 | -9 | -11 | -13 |
| M | -8 | -5 | -2 | -4 | -5 | -7 | -9 | -11 | -13 |
| E | -10 | -7 | -4 | -4 | -6 | -7 | -9 | -11 | -10 |
| D | -12 | -9 | -6 | -3 | -5 | -7 | -9 | -11 | -12 |
| P | -14 | -11 | -8 | -5 | -5 | -4 | -6 | -8 | -10 |
| L | -16 | -13 | -10 | -7 | -7 | -6 | -3 | -5 | -7 |
| E | -18 | -15 | -12 | -9 | -9 | -8 | -5 | -5 | -4 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | +1 | -1 | -3 | -2 | -4 | -3 | -5 | -4 | -3 | -5 | -4 |
| Sequence 1 | F | K | H | M | E | D | - | P | L | - | E |
| Sequence 2 | F | - | - | M | - | D | T | P | L | N | E |


ian.simpson@ed.ac.uk

# Global Pairwise Alignment Using Needleman-Wunsch

- N-W is guaranteed to find optimal alignments, although the algorithm does not search all possible alignments

- It is an example of a dynamic programming algorithm:

  an optimal path (alignment) is identified by incrementally extending optimal subpaths

- Thus, a series of decisions is made at each step of the alignment to find the pair of residues with the best score


ian.simpson@ed.ac.uk

# Global Pairwise Alignment Using Needleman-Wunsch

- Global alignment (Needleman-Wunsch) extends from one end of each sequence to the other
- Local alignment finds optimally matching regions within two sequences ("subsequences")
- Local alignment is almost always used for database searches such as BLAST.

  It is useful to find domains (or limited regions of homology) within sequences
- Smith and Waterman (1981) solved the problem of performing optimal local sequence alignment
- Other methods (BLAST, FASTA) are faster but less thorough.


ian.simpson@ed.ac.uk

# Global Alignment vs Local Alignment



ian.simpson@ed.ac.uk

# Local Alignment Using Smith-Waterman

Set up a matrix between two proteins (size m+1, n+1)

No values in the scoring matrix can be negative! $S \geq 0$

The score in each cell is the maximum of four values:

- s(i-1, j-1) + the new score at [i,j] (a match or mismatch)
- s(i,j-1) – gap penalty
- s(i-1,j) – gap penalty
- zero ← this is not in Needleman-Wunsch


ian.simpson@ed.ac.uk

# Local Alignment Using Smith-Waterman

Sequence 1

| Sequence 2 | | C | A | G | C | C | U | C | G | C | U | U | A | G |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| A | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 |
| A | 0.0 | 0.0 | 1.0 | 0.7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.7 |
| U | 0.0 | 0.0 | 0.0 | 0.7 | 0.3 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.7 |
| G | 0.0 | 0.0 | 0.0 | 1.0 | 0.3 | 0.0 | 0.0 | 0.7 | 1.0 | 0.0 | 0.0 | 0.7 | 0.7 | 1.0 |
| C | 0.0 | 1.0 | 0.0 | 0.0 | 2.0 | 1.3 | 0.3 | 1.0 | 0.3 | 2.0 | 0.7 | 0.3 | 0.3 | 0.3 |
| C | 0.0 | 1.0 | 0.7 | 0.0 | 1.0 | 3.0 | 1.7 | 1.3 | 1.0 | 1.3 | 1.7 | 0.3 | 0.0 | 0.0 |
| A | 0.0 | 0.0 | 2.0 | 0.7 | 0.3 | 1.7 | 2.7 | 1.3 | 1.0 | 0.7 | 1.0 | 1.3 | 1.3 | 0.0 |
| U | 0.0 | 0.0 | 0.7 | 1.7 | 0.3 | 1.3 | 2.7 | 2.3 | 1.0 | 0.7 | 1.7 | 2.0 | 1.0 | 1.0 |
| U | 0.0 | 0.0 | 0.3 | 0.3 | 1.3 | 1.0 | 2.3 | 2.3 | 2.0 | 0.7 | 1.7 | 2.7 | 1.7 | 1.0 |
| G | 0.0 | 0.0 | 0.0 | 1.3 | 0.0 | 1.0 | 1.0 | 2.0 | 3.3 | 2.0 | 1.7 | 1.3 | 2.3 | 2.7 |
| A | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0.3 | 0.7 | 0.7 | 2.0 | 3.0 | 1.7 | 1.3 | 2.3 | 2.0 |
| C | 0.0 | 1.0 | 0.0 | 0.7 | 1.0 | 2.0 | 0.7 | 1.7 | 1.7 | 3.0 | 2.7 | 1.3 | 1.0 | 2.0 |
| G | 0.0 | 0.0 | 0.7 | 1.0 | 0.3 | 0.7 | 1.7 | 0.3 | 2.7 | 1.7 | 2.7 | 2.3 | 1.0 | 2.0 |
| G | 0.0 | 0.0 | 0.0 | 1.7 | 0.7 | 0.3 | 0.3 | 1.3 | 1.3 | 2.3 | 1.3 | 2.3 | 2.0 | 2.0 |

scoring:
match = +1
mismatch = -0.3
gap = -1.3 (l=1)

local

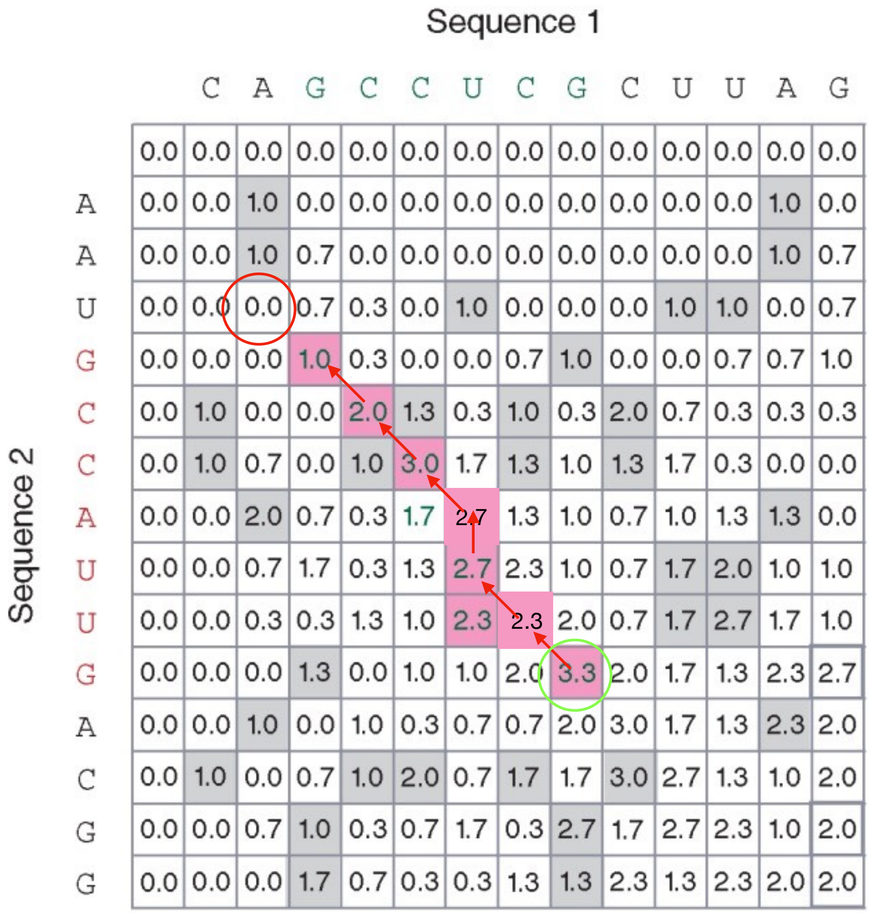
```
sequence 1   GCC-UCG
sequence 2   GCCAUUG
```

global

```
sequence 1   CA-GCC-UCGCUUAG
sequence 2   AAUGCCAUUGACG-G
```


ian.simpson@ed.ac.uk

# Local Alignment Using Smith-Waterman




scoring:
match = +1
mismatch = -0.3
gap = -1.3 (l=1)

local
```
sequence 1   GCC-UCG
sequence 2   GCCAUUG
```

global
```
sequence 1   CA-GCC-UCGCUUAG
sequence 2   AAUGCCAUUGACG-G
```


ian.simpson@ed.ac.uk

# Rapid, Heuristic Versions of Smith-Waterman: FASTA and BLAST

- Smith-Waterman is very rigorous and it is guaranteed to find an optimal alignment
- Smith-Waterman is slow. It requires computer space and time proportional to the product of the two sequences being aligned (or the product of a query against an entire database)
- Gotoh (1982) and Myers and Miller (1988) improved the algorithms so both global and local alignment require less time and space
- FASTA and BLAST provide rapid alternatives to S-W


ian.simpson@ed.ac.uk

# Next Week

## Week 3 - Basic Local Alignment & Search Tool (BLAST)

If you would like  to read ahead then please look at the following chapter, if not we will cover this in next week's lecture
This is available from the Bio1 course "Resource List"


BOOK **Bioinformatics and functional genomics**

Pevsner, Jonathan, 1961-, 3rd ed.., Chichester, West Sussex, UK ; Hoboken, NJ, USA, John Wiley and Sons, Incorporated, 2015

***Note:*** *Read Chapter 3, "Pairwise Sequence Alignment"*

Add tags to item

**Complete**



ian.simpson@ed.ac.uk