Lessons 3 & 4.
Contents

BIOinformatics = genes + proteins + informatics

(part of computational biology, biocomputing)

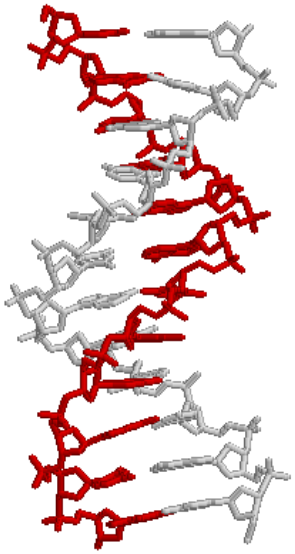
GENE: DNA segment which codes for a specific protein and determines an hereditary feature

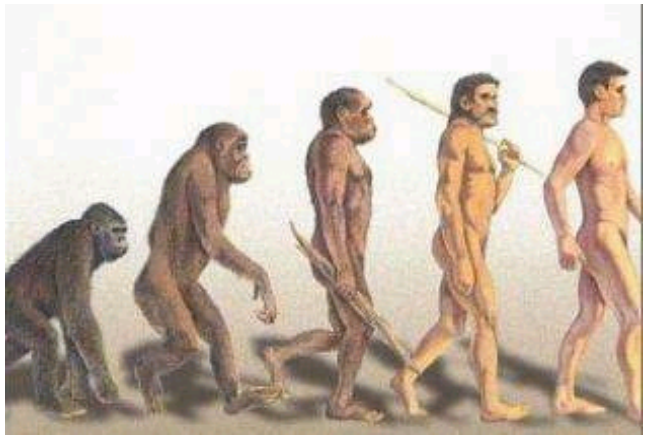
Building blocks: 4 nucleotides (ACGT/U)

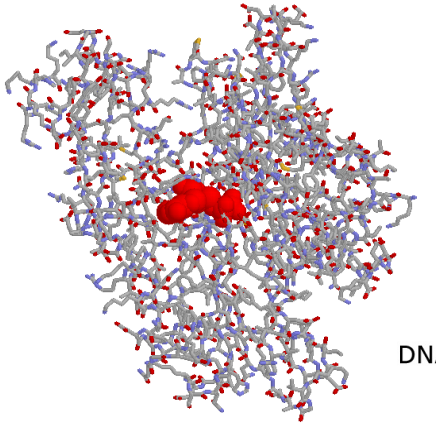
PROTEIN: expression product of a gene and ed EFFECTOR of the biochemical function whose information is stored in the gene

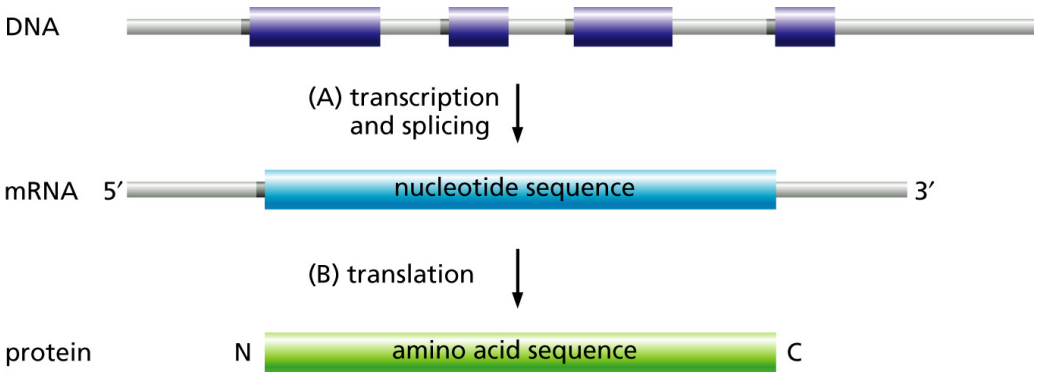
Building blocks: 20 amino acids

MOLECULAR EVOLUTION





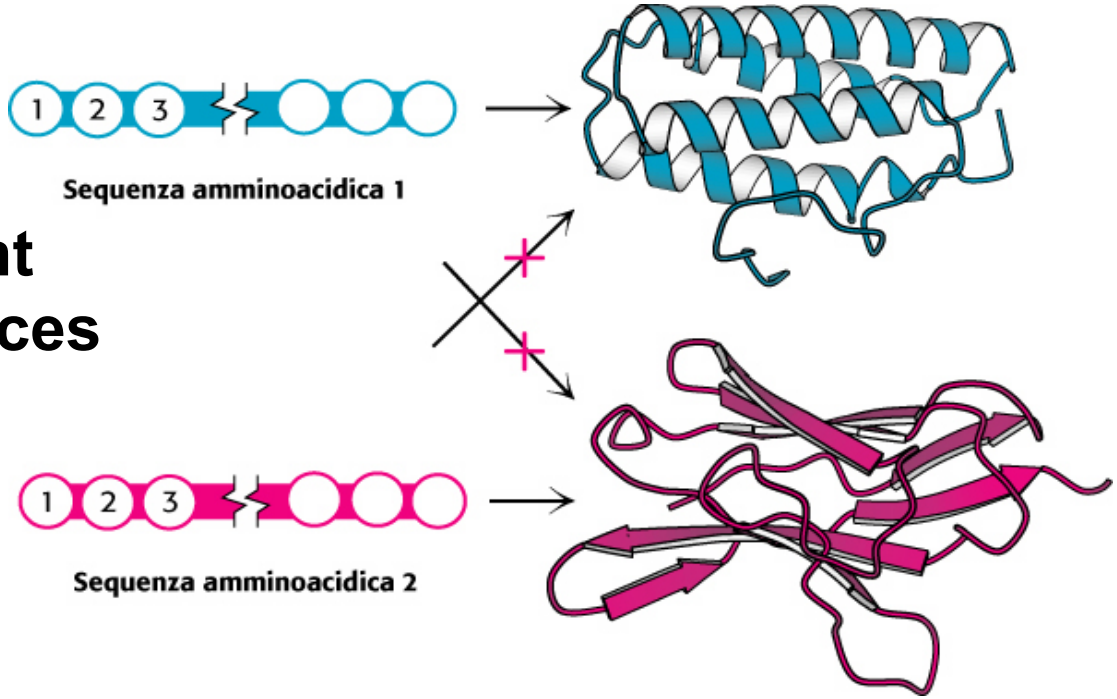




Protein structure

Proteins are made up of 20 different amino acids: ACDEFGHIKLMNPQRSTVWY

Different sequences



Different 3D structures and different biological functions!

Protein structure

Knowing the relationship between a protein structure and its function provides a fundamental understanding of how the protein works allowing to foresee how modifying the structure could affect the function

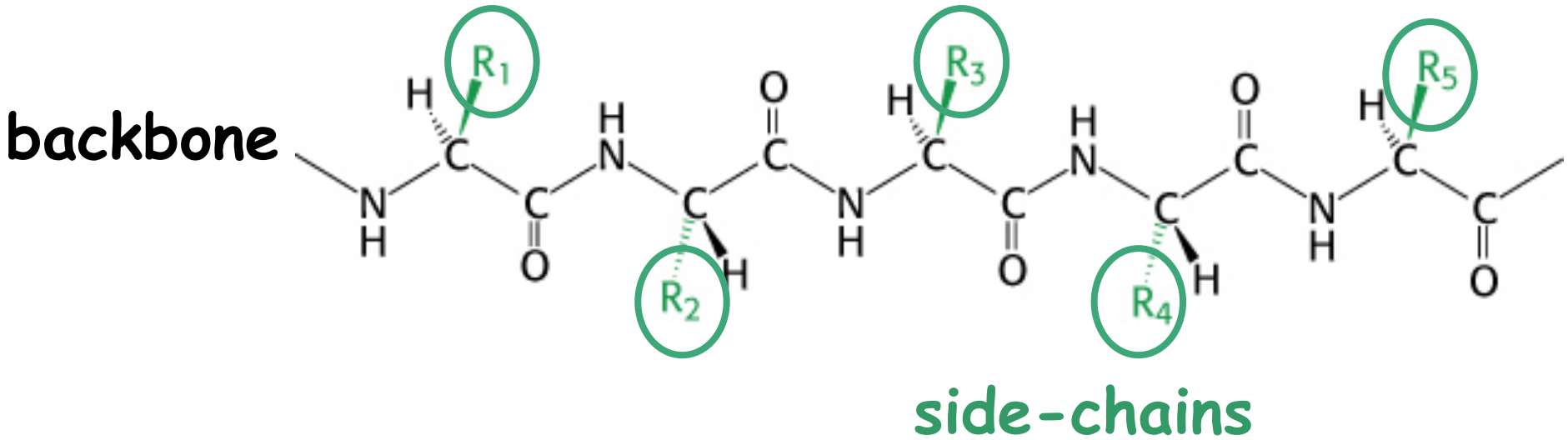
Most of the currently marketed pharmaceuticals act by interacting with proteins

The structure adopted by a protein is entirely determined by its amino acids sequence, however the rules that govern how a protein chain of a given sequence folds up are not yet fully understood

One of the main aims of Bioinformatics is to predict and analyze the structure of proteins and the relationship of the structure to the function

Protein structure

Proteins are made of 20 amino acids, covalently bonded by peptide bonds



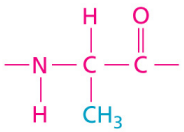
The 20 amino acids are made of C, N, O, H (S in case of Cys and Met) atoms

Their side chains differ in size and chemical nature

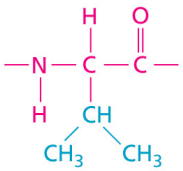
Protein structure

NONPOLAR SIDE CHAINS

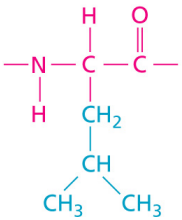
alanine
(Ala, or A)



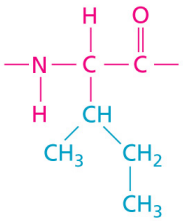
valine
(Val, or V)



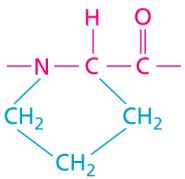
leucine
(Leu, or L)



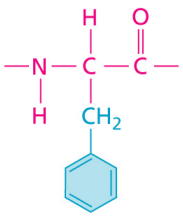
isoleucine
(Ile, or I)



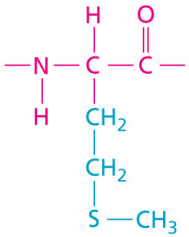
proline
(Pro, or P)



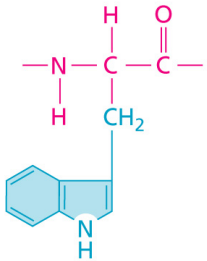
phenylalanine
(Phe, or F)



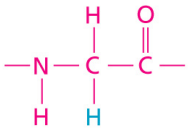
methionine
(Met, or M)



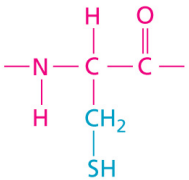
tryptophan
(Trp, or W)



glycine
(Gly, or G)



cysteine
(Cys, or C)

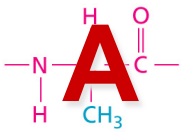


Protein structure

NONPOLAR SIDE CHAINS

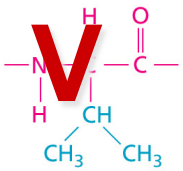
alanine

(Ala, or A)



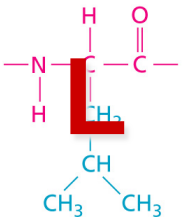
valine

(Val, or V)



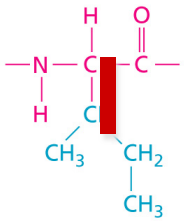
leucine

(Leu, or L)



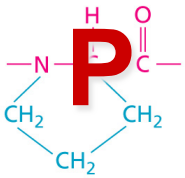
isoleucine

(Ile, or I)



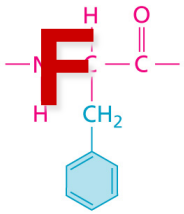
proline

(Pro, or P)



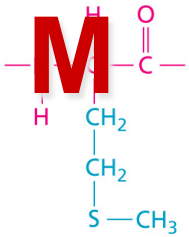
phenylalanine

(Phe, or F)



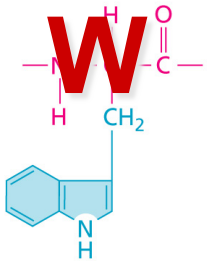
methionine

(Met, or M)



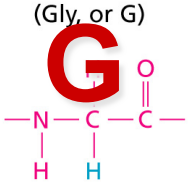
tryptophan

(Trp, or W)



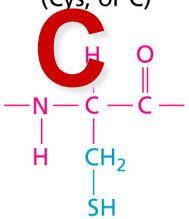
glycine

(Gly, or G)



cysteine

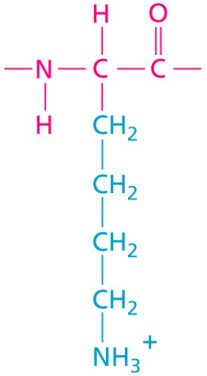
(Cys, or C)



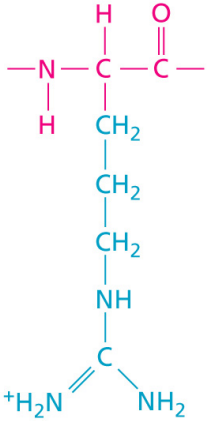
Protein structure

BASIC SIDE CHAINS

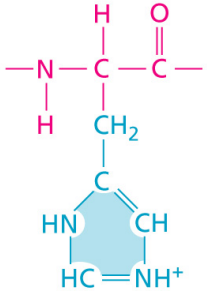
lysine
(Lys, or K)



arginine
(Arg, or R)

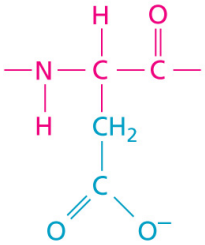


histidine
(His, or H)

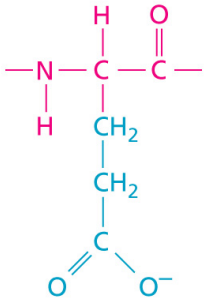


ACIDIC SIDE CHAINS

aspartic acid
(Asp, or D)

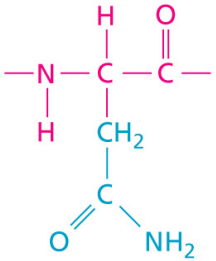


glutamic acid
(Glu, or E)

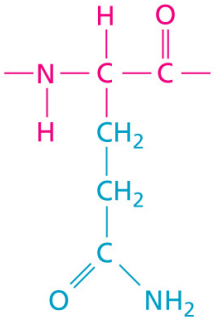


UNCHARGED POLAR SIDE CHAINS

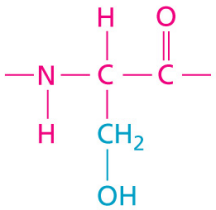
asparagine
(Asn, or N)



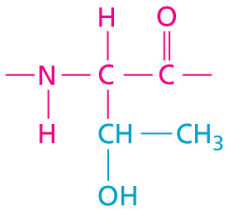
glutamine
(Gln, or Q)



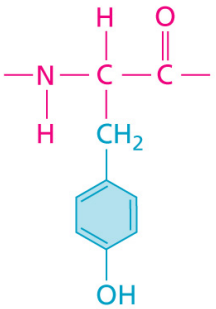
serine
(Ser, or S)



threonine
(Thr, or T)



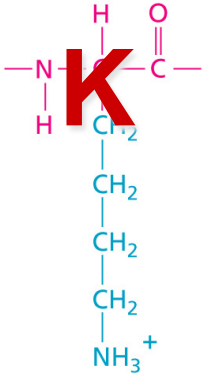
tyrosine
(Tyr, or Y)



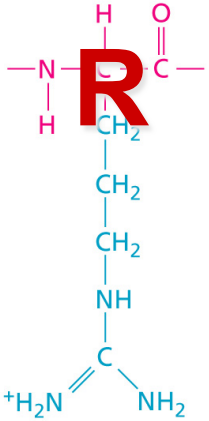
Protein structure

BASIC SIDE CHAINS

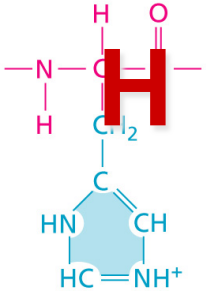
lysine
(Lys, or K)



arginine
(Arg, or R)

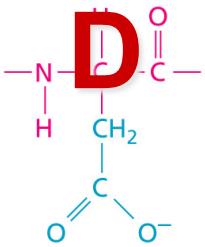


histidine
(His, or H)

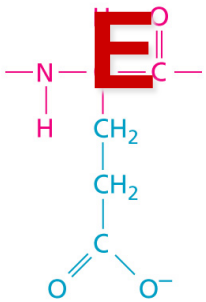


ACIDIC SIDE CHAINS

aspartic acid
(Asp, or D)

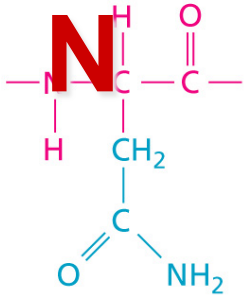


glutamic acid
(Glu, or E)

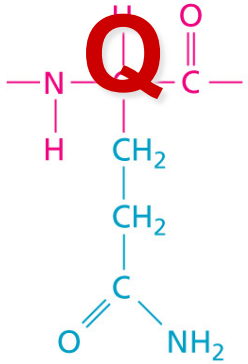


UNCHARGED POLAR SIDE CHAINS

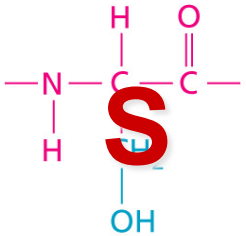
asparagine
(Asn, or N)



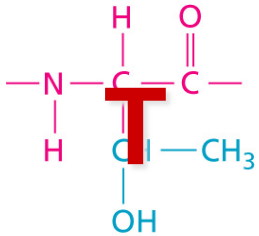
glutamine
(Gln, or Q)



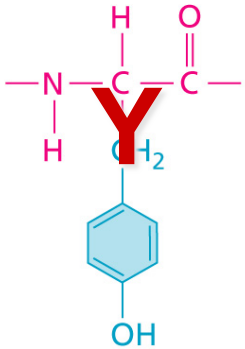
serine
(Ser, or S)



threonine
(Thr, or T)

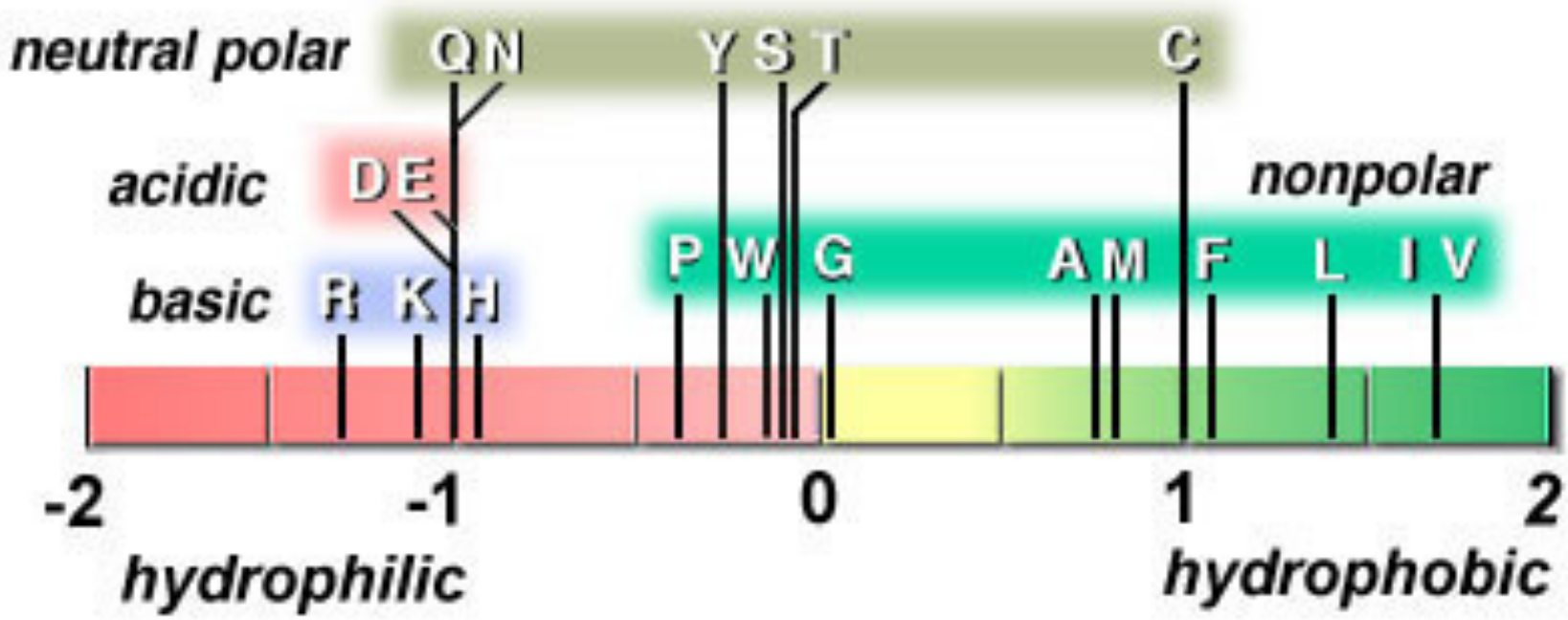


tyrosine
(Tyr, or Y)



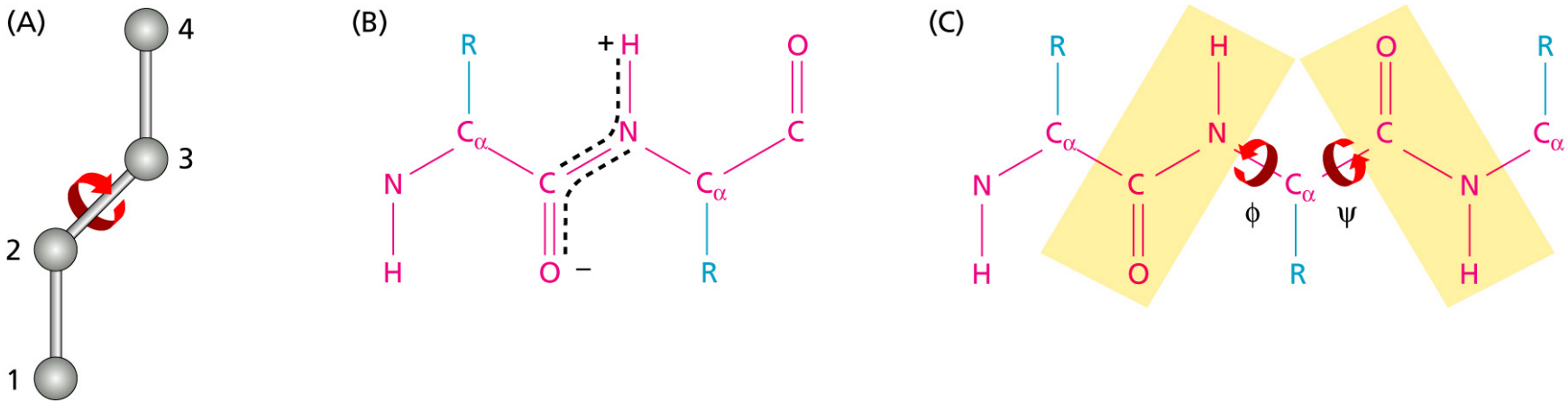
Protein structure

Amino acids hydrophilicity/hydrophobicity:



Protein structure

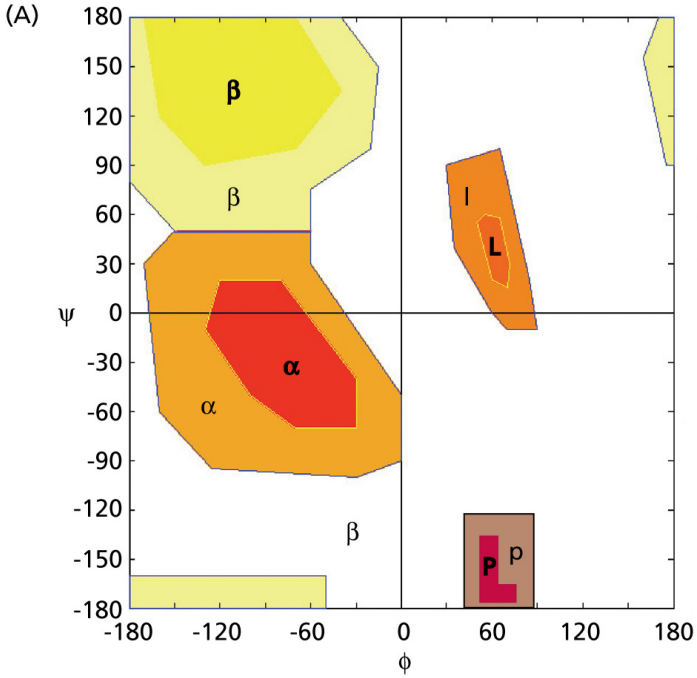
Peptide bonds are planar. However the bonds made by Cα with N and C are singular and give rise to two torsional angles per residue (ϕ and ψ, defined between -180° and +180°)



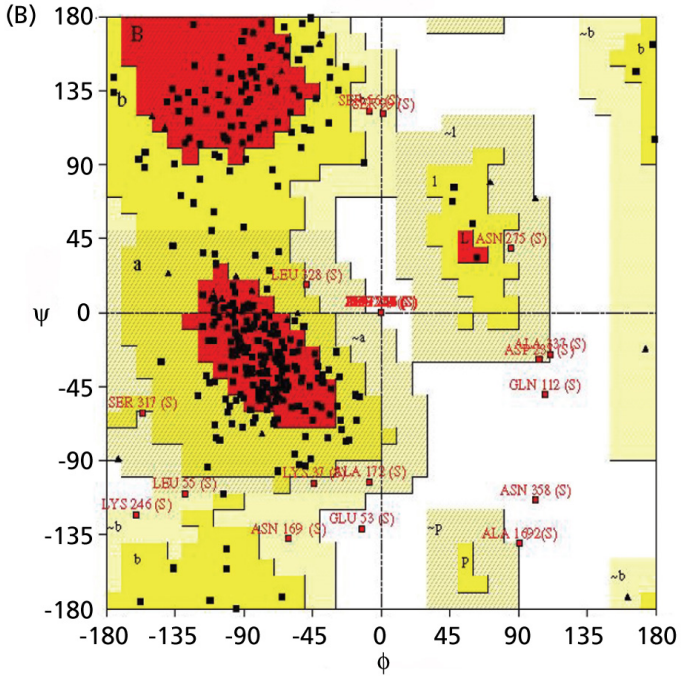
These torsion angles are the main source of flexibility for proteins

Protein structure

ϕ and ψ angles assume preferentially some values, as steric hindrance prevents certain combinations



Ramachandran plot, the darker the color, the more favorable the combination of angles



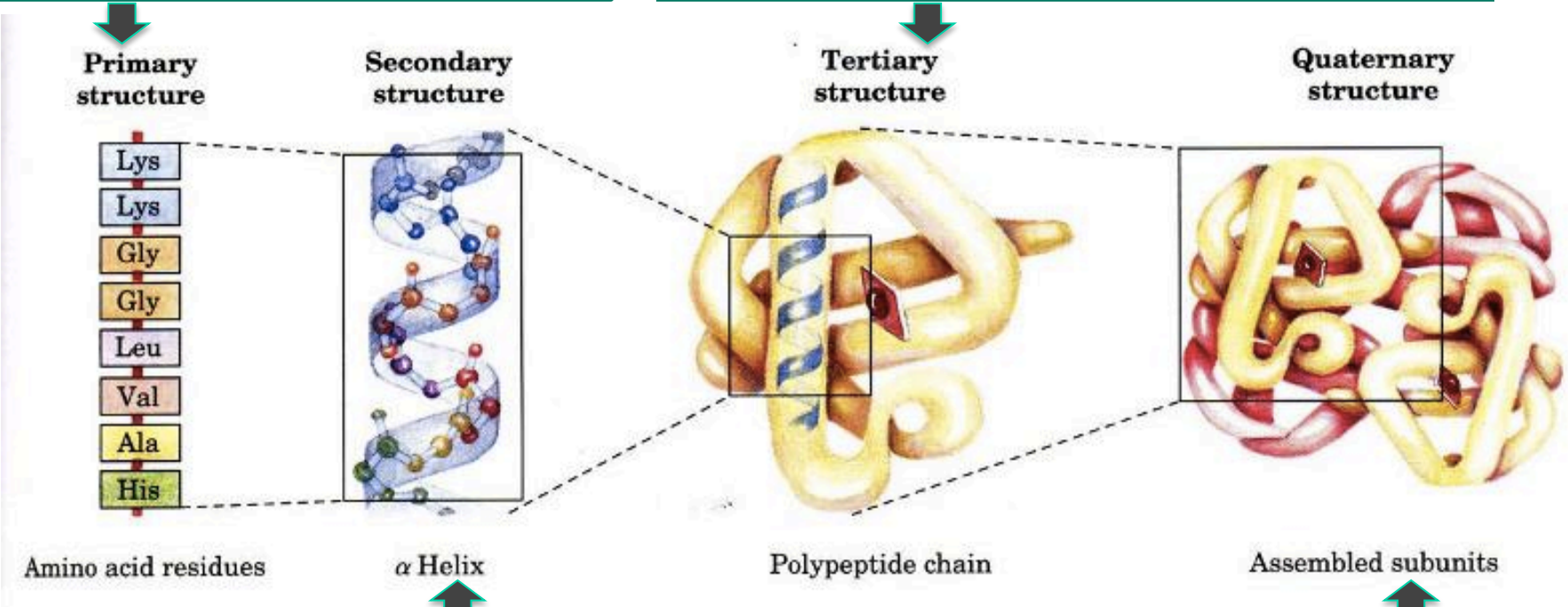
Example for a real structure, with outliers

Protein structure: hierarchy

There are four levels of protein structure to consider:

The protein sequence: types and order of the amino acids

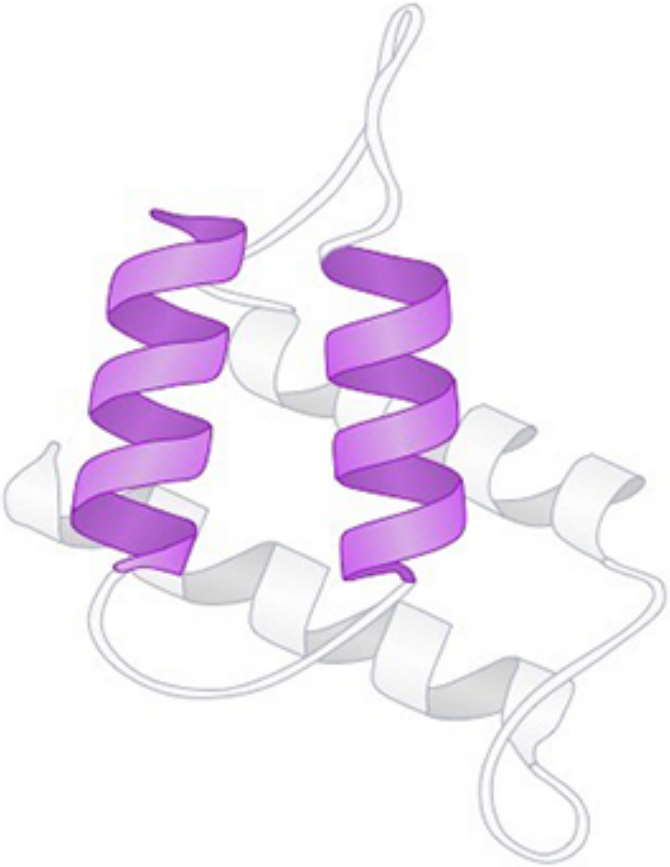
Folding and packing of the secondary structure elements to give the final 3D structure



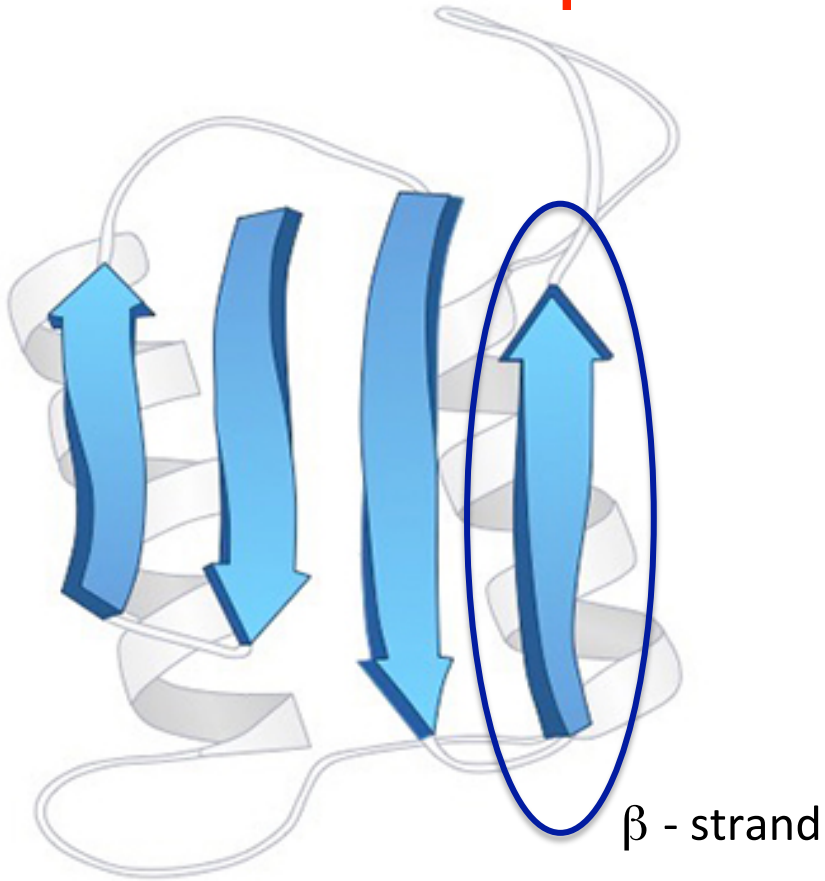
1st level of folding where parts of the protein fold to form local repetitive structures

Many functional proteins are formed by more protein chains (identical or not)

Protein secondary structure: α helices and β-strands



α - helices

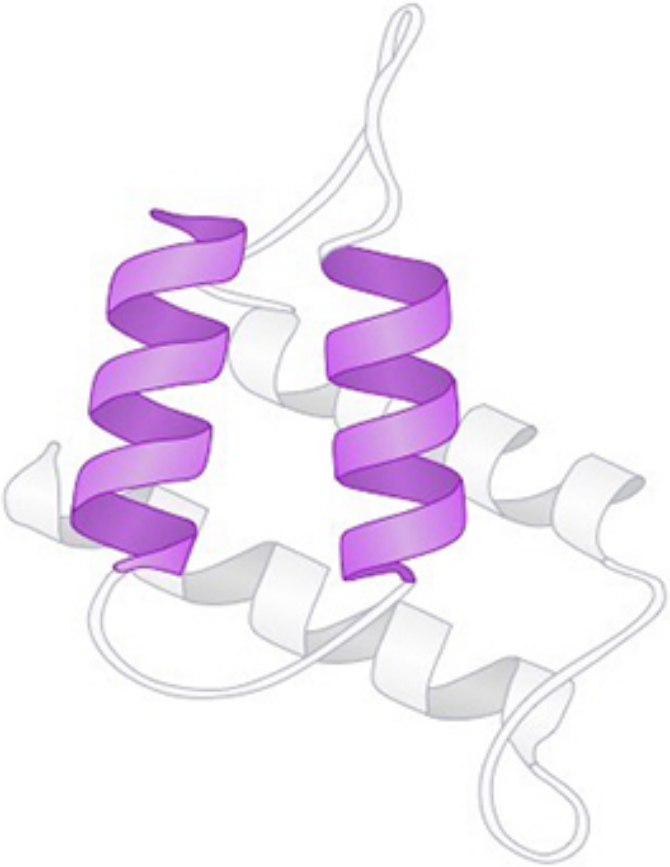


β - pleated sheets

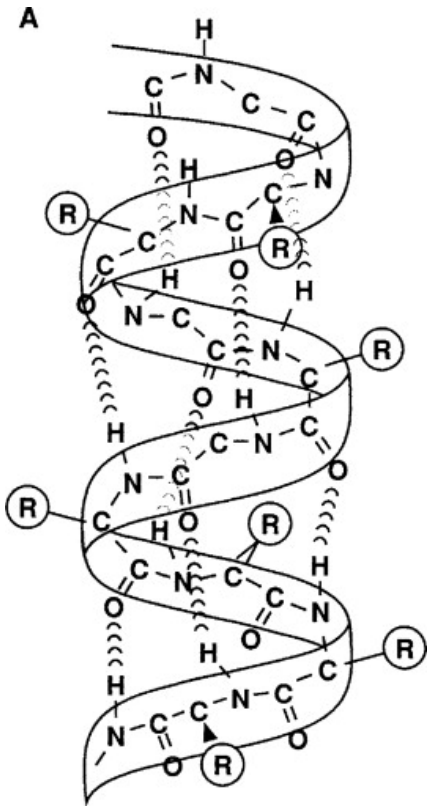
α-helices and β-strands are the only regular protein secondary structure motifs

α-helices and β-strands are connected by turns (ordered 3/4-residue motifs) or loops

Protein secondary structure: α helices and β-strands

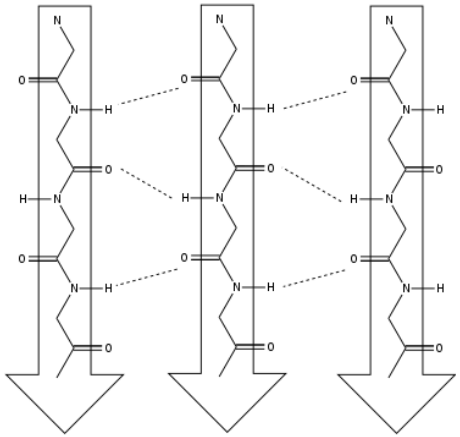


α - helices

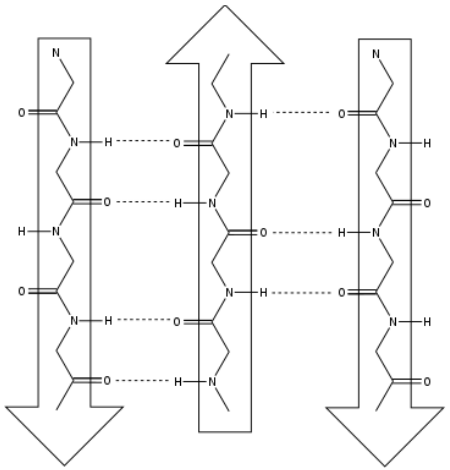


In a **α-helix** (right hand) conformation dihedral angles (φ, ψ) assume values around −60°/−45° and every backbone N−H group hydrogen bonds to the backbone C=O group of the amino acid located 4-residue upstream

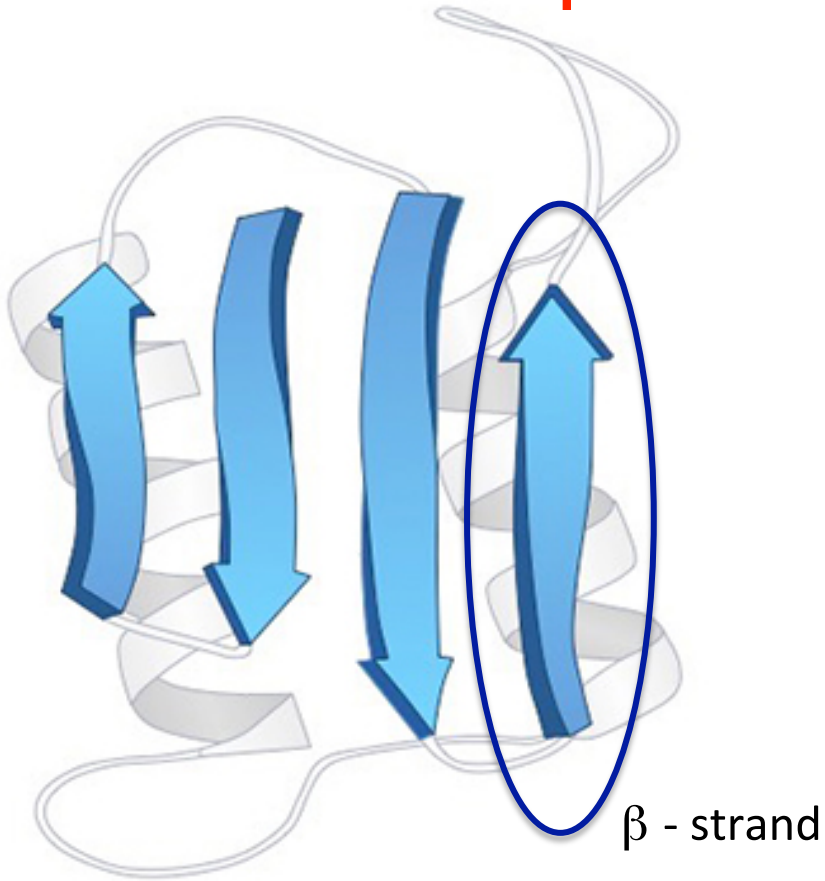
Protein secondary structure: α helices and β-strands



parallel



antiparallel



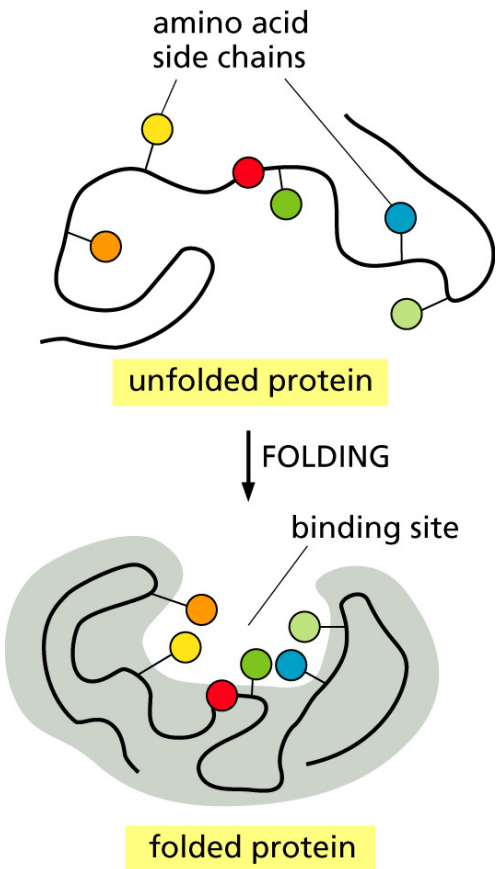
β - pleated sheets

A β-strand is a stretch of polypeptide chain typically 3 to 10 amino acids long with backbone in an extended conformation - dihedral angles (φ, ψ) values around −135°/+135°; they can form sheets where their backbone H-bond to that of another strand

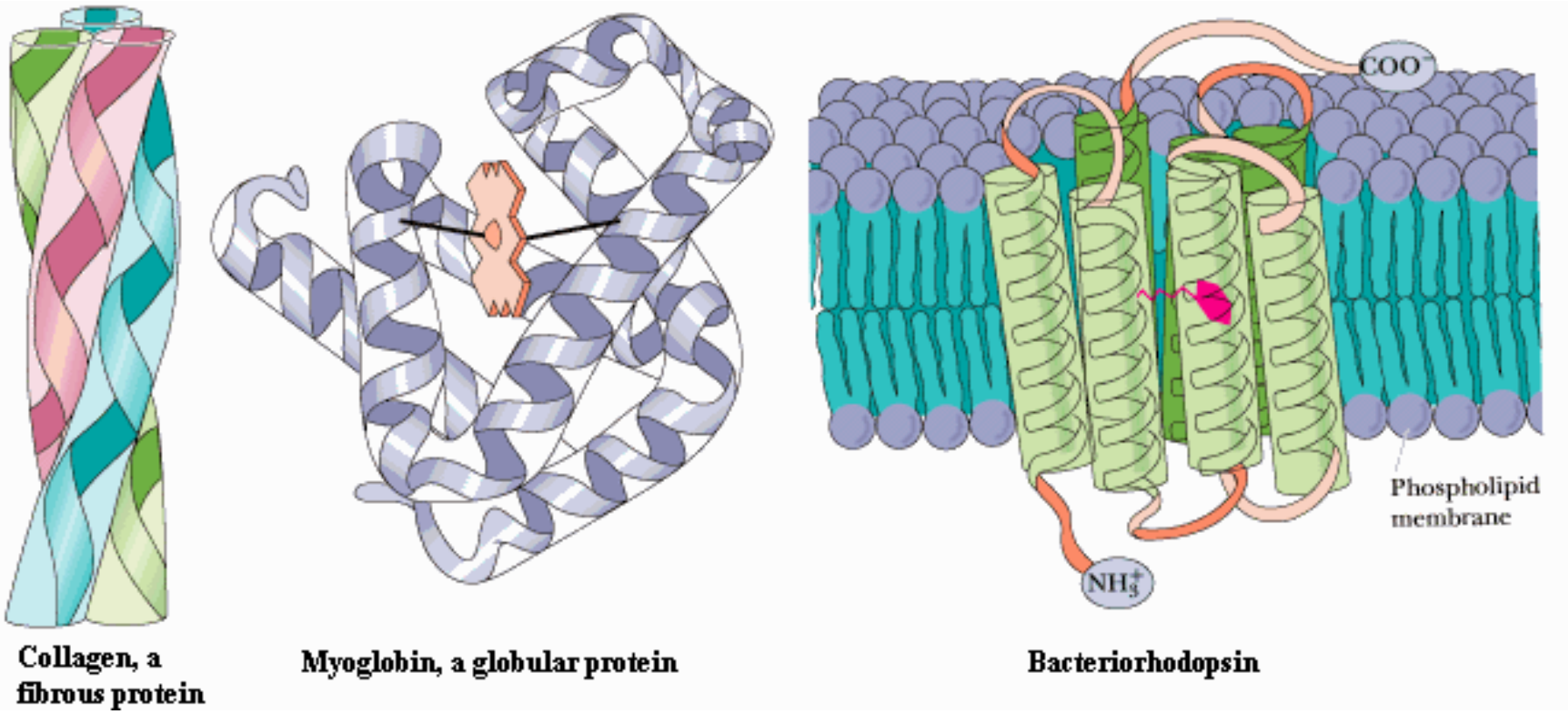
Protein structure

The folded state of a protein corresponds to a free energy minimum

Residues which are distant in sequence can come close in the folded structure to form a functional site, e.g. a binding/ catalytic site

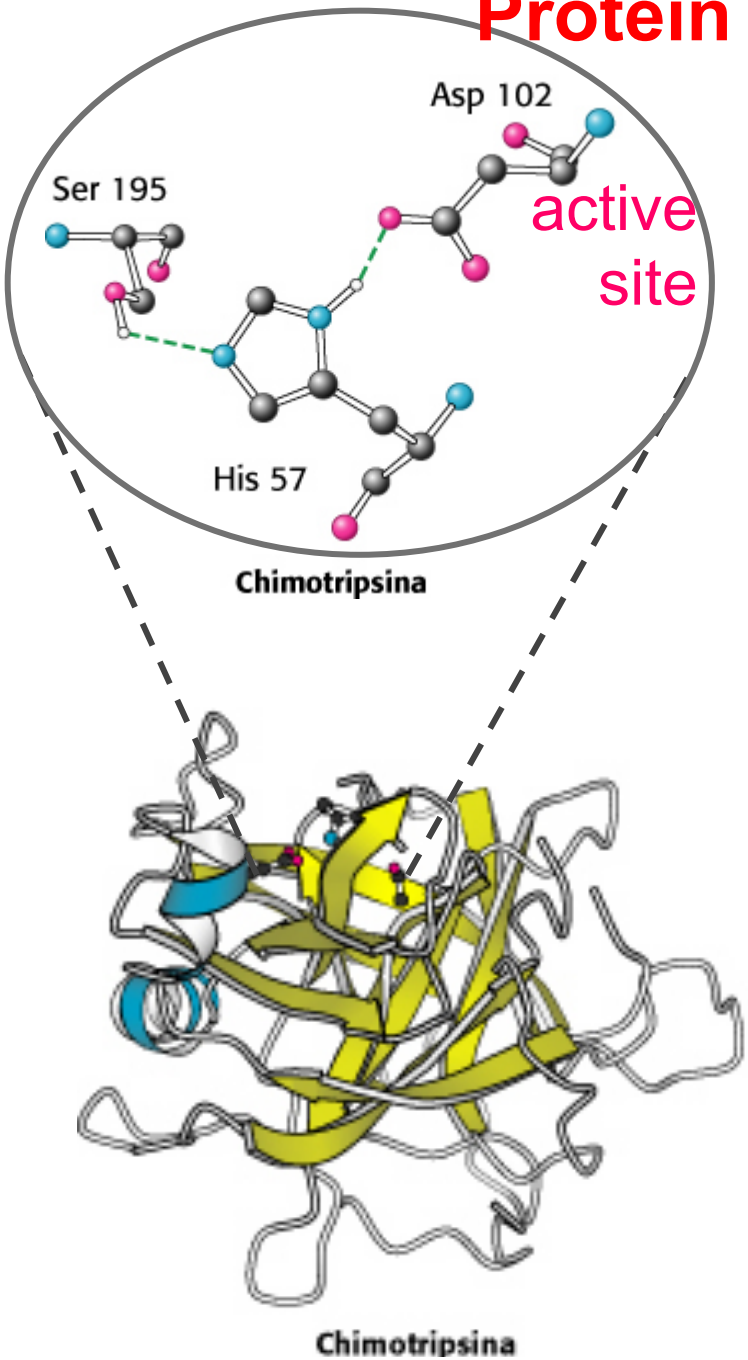


Protein structure: classes



structural globular membrane

Protein structure



Enzymes are globular proteins which catalyze reactions through an active site

Example:

Chymotrypsin is a digestive enzyme active in the small intestine where it contributes to proteins deigestion

It can break peptide bonds thanks to its active site (catalytic Ser-protease triad).

Protein domains

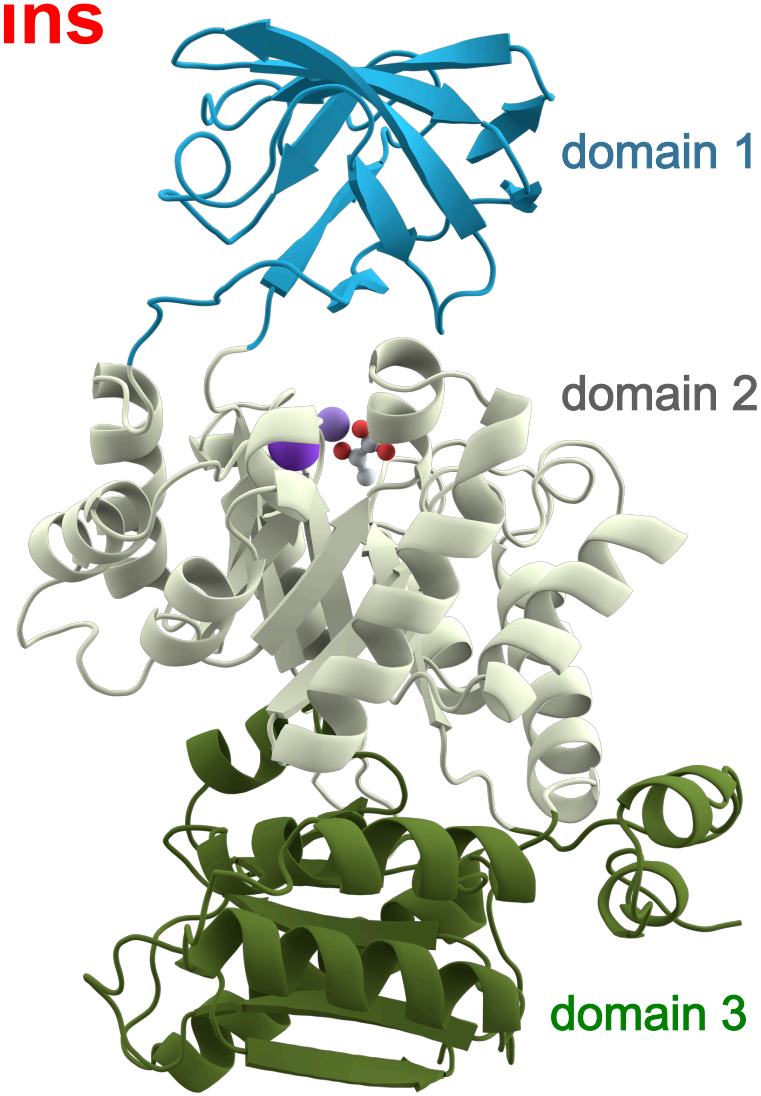
Many proteins consist of several domains

A protein domain is a region of a protein that is self-stabilizing and that folds independently from the rest

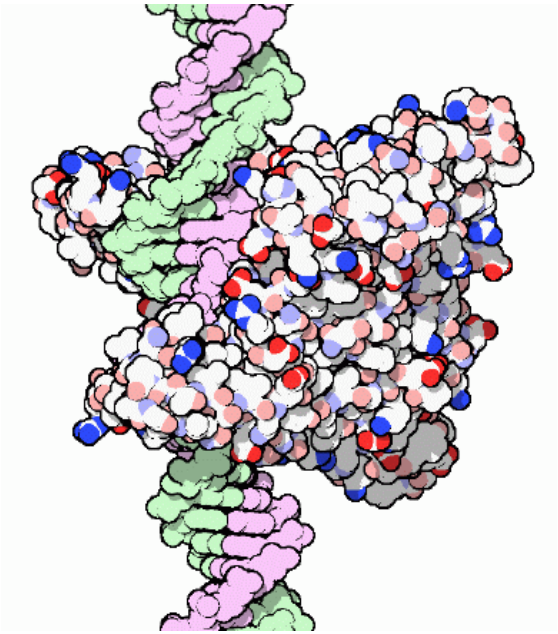
Domains usually form functional units

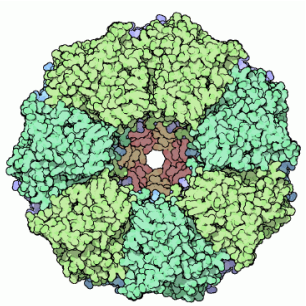
Domains vary in length from ≈50 amino acids up to ≈250 amino acids

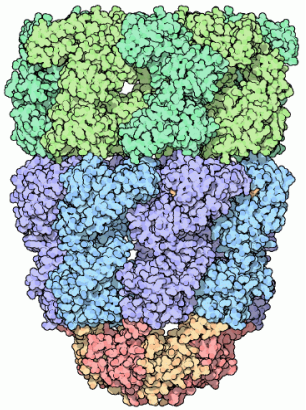
Molecular evolution uses domains as building blocks: a domain may appear in a variety of different proteins

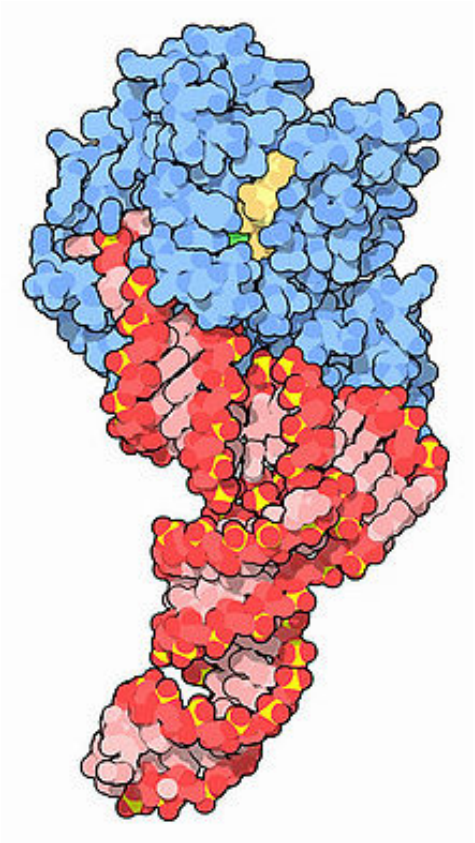


Example: pyruvate kinase contains an all-β nucleotide binding domain (blue), an α/β-substrate binding domain (grey) and an α/β-regulatory domain (green) connected by linkers

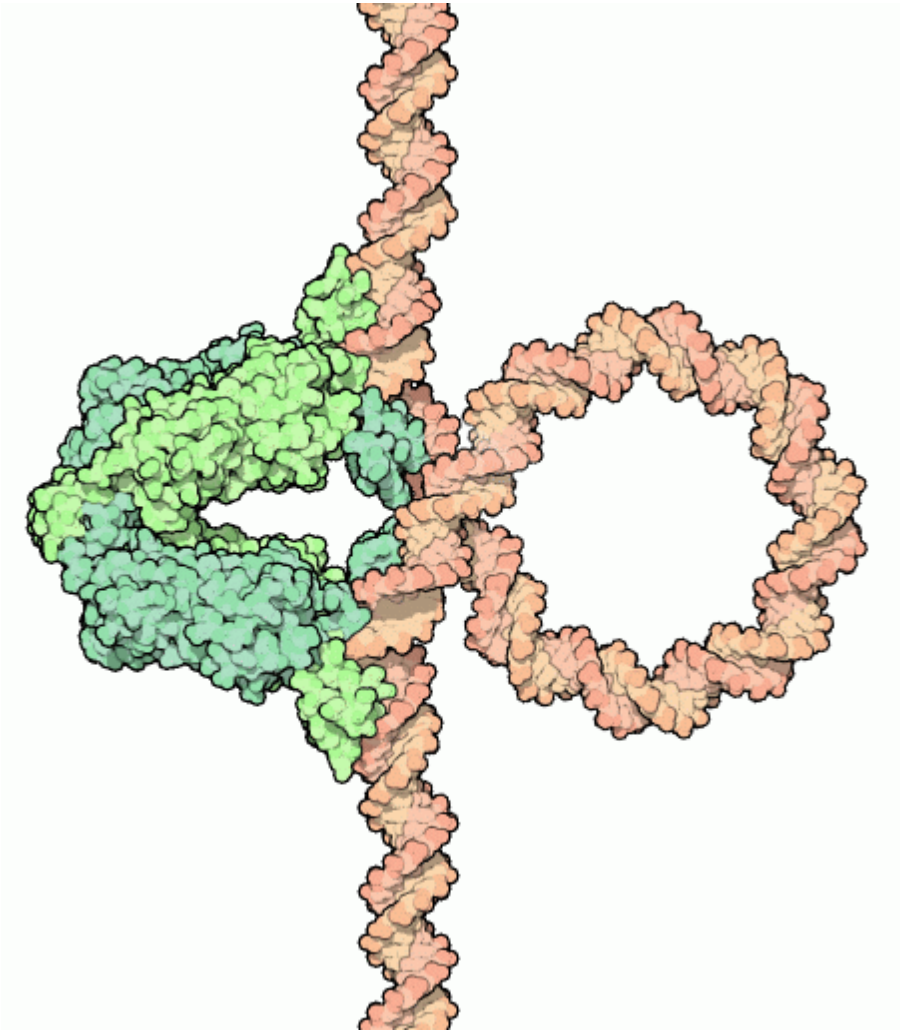








A specific protein fold allows it to perform its function!



lac repressor:
blocks transcription
of a specific DNA
region

Example of protein sequence

```
YQVRNSSGLYHVTNDCPNSSIVYEAADAILHTPGCVPCVREGNASRCWV
AVTPTVATRDGKLPTTQLRRHIDLLVGSATLCSALYVGDLCGSVFLVGQ
LFTFSPRRHWTTQDCNCSIYPGHITGHRMAWDMMMNWSPTAALVVAQLL
RIPQAILDMIAGAHWGVLAGIAYFSMVGNWAKVLVVLLLFAGVDAETHV
TGGSAGHTTAGLVRLLSPGAKQNIQLINTNGSWHINSTALNCNESLNTG
WLAGLFYHHKFNSSGCPERLASCRRLTDFAQGGGPISYANGSGLDERPY
CWHYPPRPCGIVPAKSVCGPVYCFTPSPVVVGTTDRSGAPTYSWGANDT
DVFVLNNTRPPLGNWFGCTWMNSTGFTKVCGAPPCVIGGVGNNTLLCPT
DCFRKHPEATYSRCGSGPWITPLLLLLALPQRAY
```

Structural/functional information is contained in the amino acid sequence of a protein chain

Proteins vary by the different combination of the 20 amino acids in their sequence

Example of protein sequence

```
YQVRNSSGLYHVTNDCPNSSIVYEAADAILHTPGCVPCVREGNASRCWV
AVTPTVATRDGKLPTTQLRRHIDLLVGSATLCSALYVGDLCGSVFLVGQ
LFTFSPRRHWTTQDCNCSIYPGHITGHRMAWDMMMNWSPTAALVVAQLL
RIPQAILDMIAGAHWGVLAGIAYFSMVGNWAKVLVVLLLFAGVDAETHV
TGGSAGHTTAGLVRLLSPGAKQNIQLINTNGSWHINSTALNCNESLNTG
WLAGLFYHHKFNSSGCPERLASCRRLTDFAQGGGPISYANGSGLDERPY
CWHYPPRPCGIVPAKSVCGPVYCFTPSPVVVGTTDRSGAPTYSWGANDT
DVFVLNNTRPPLGNWFGCTWMNSTGFTKVCGAPPCVIGGVGNNTLLCPT
DCFRKHPEATYSRCGSGPWITPLLLLLALPQRAY
```

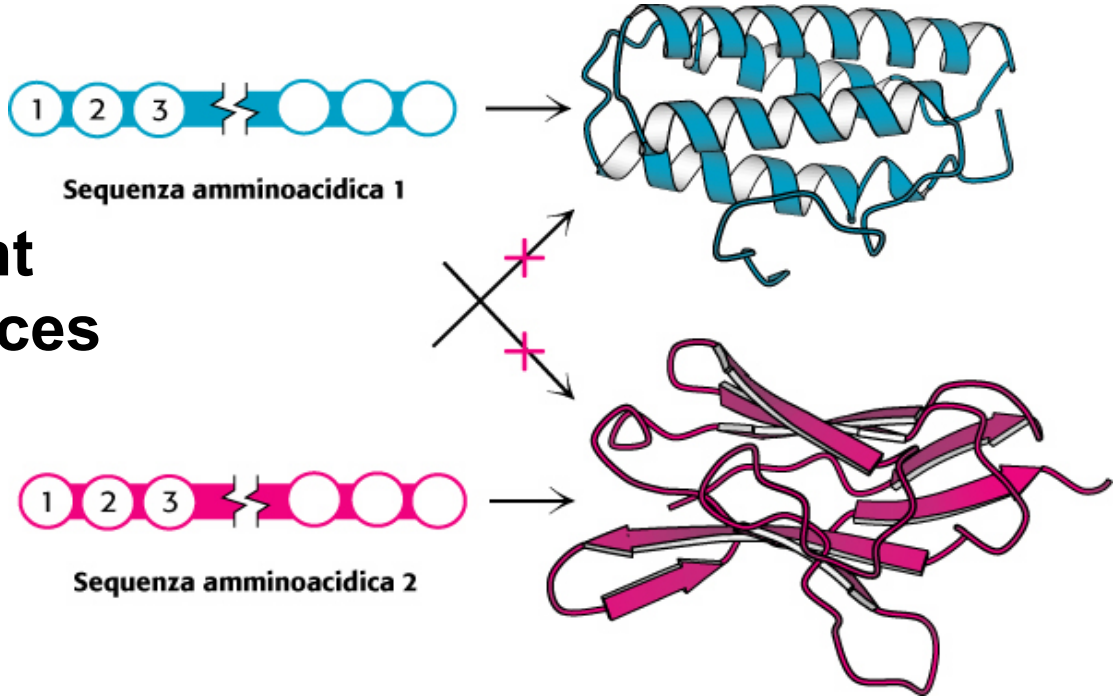
Structural/functional information is contained in the amino acid sequence of a protein chain

In principle, there can be 20^{n} different polypeptide chains of length ***n***: 20^{250} of length ***250*** (over 10^{325}), but only a tiny fraction of them exist (*again, think of evolution!*)

Protein structure

Proteins are made up of 20 different amino acids: ACDEFGHIKLMNPQRSTVWY

Different sequences



Different 3D structures and different biological functions!

There seems to be a limited number, in the order of thousands (10^3), of fold families, thus also proteins with different sequences may in principle fold similarly

“Informatics” problems with protein sequences

Storing and archiving protein sequences

Search for regularities and “patterns” (e.g. active sites)

Comparing protein sequences and measuring their similarity

“Informatics” problems with protein sequences

Storing and archiving protein sequences

Search for regularities and “patterns” (e.g. active sites)

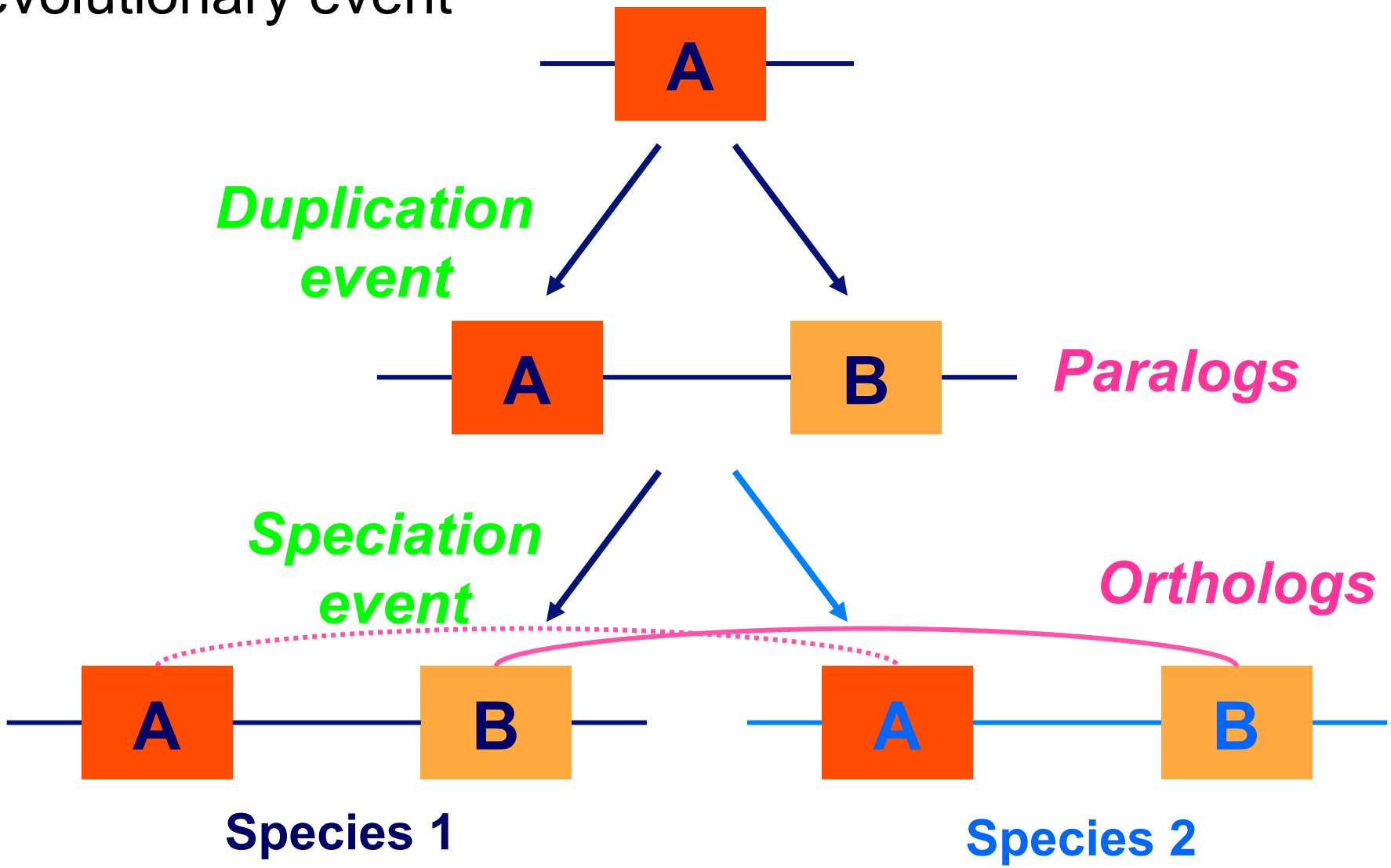
Comparing protein sequences and measuring their similarity

Similarity & homology

- Two sequences are similar if they can be aligned so that many corresponding (aligned) amino acids are identical or similar
- Technically two or more proteins may be defined homologous if they derive from a common ancestor
- Homology between two sequences *cannot be observed* but only inferred by their similarity in sequence or function
- The concept of similarity can be extended to 3D structures

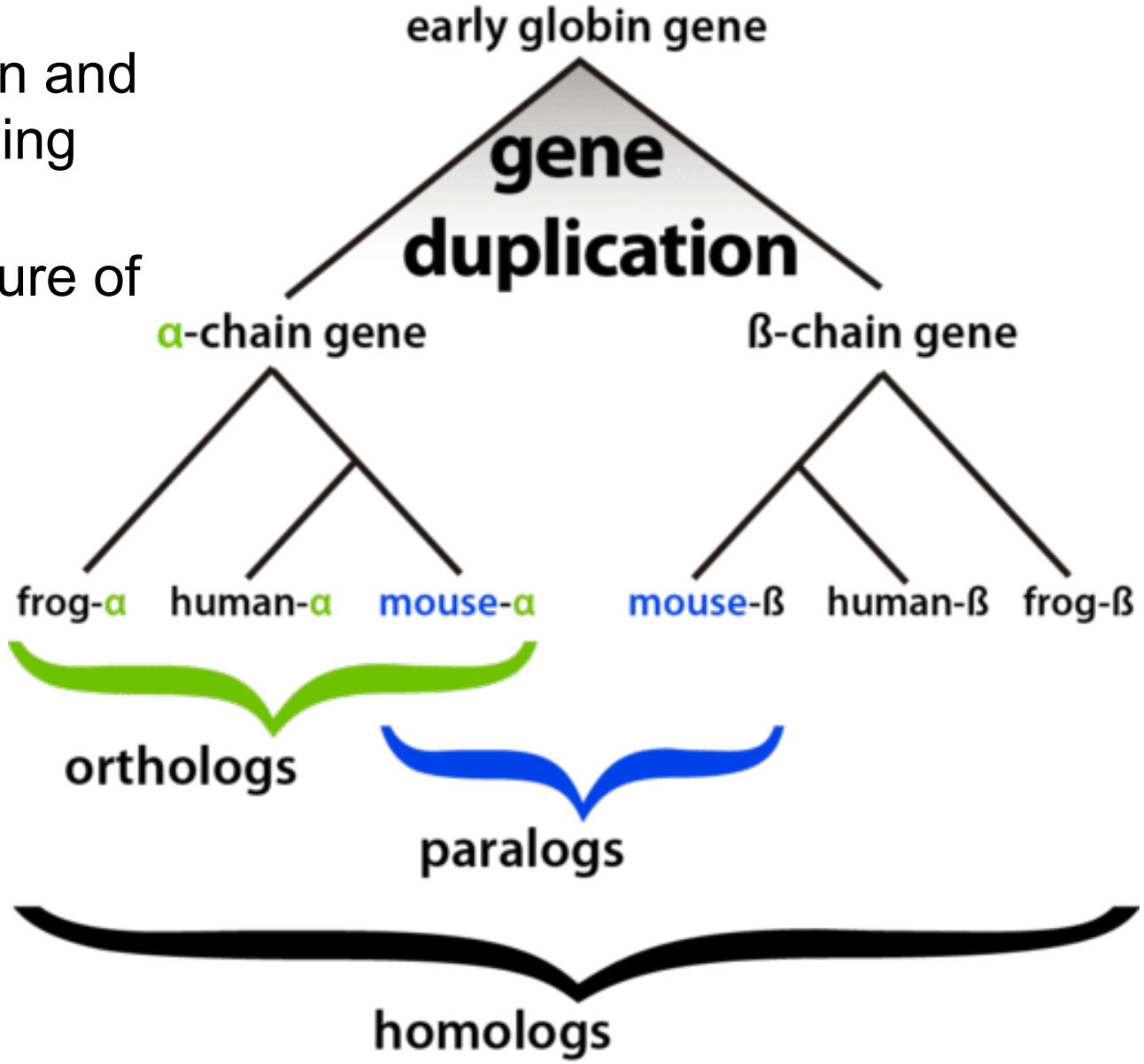
Homology

Two or more proteins may be defined homologous if they derive from a common ancestor through an evolutionary event



Homology

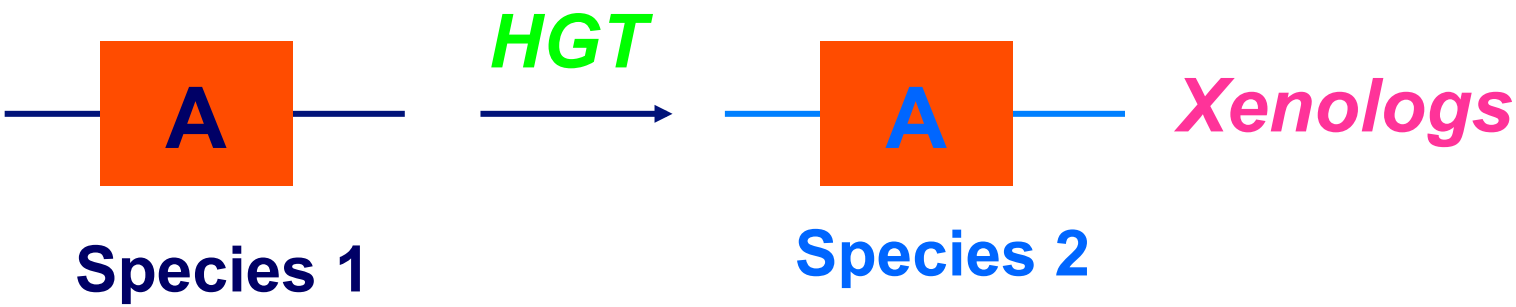
Example: α-globin and β-globin, composing together the quaternary structure of hemoglobin

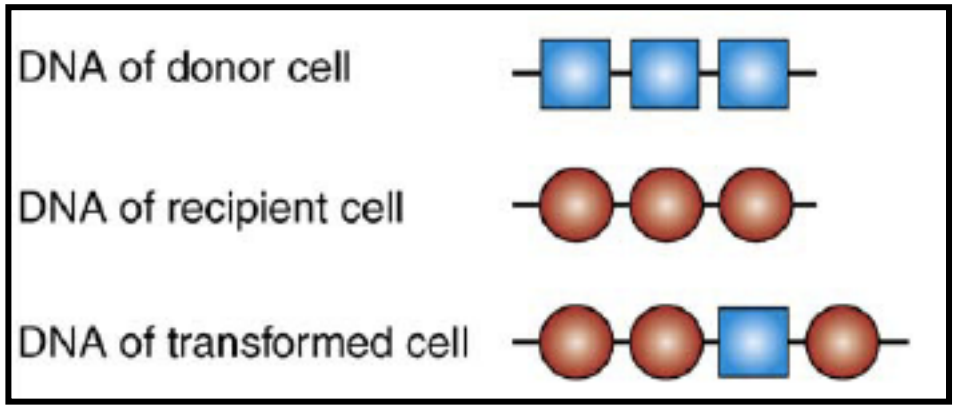


Homology

Two sequences are homologous if they derive from a common ancestor:

The **Horizontal Gene Transfer (HGT)** is the genetic material transfer between two different genomes



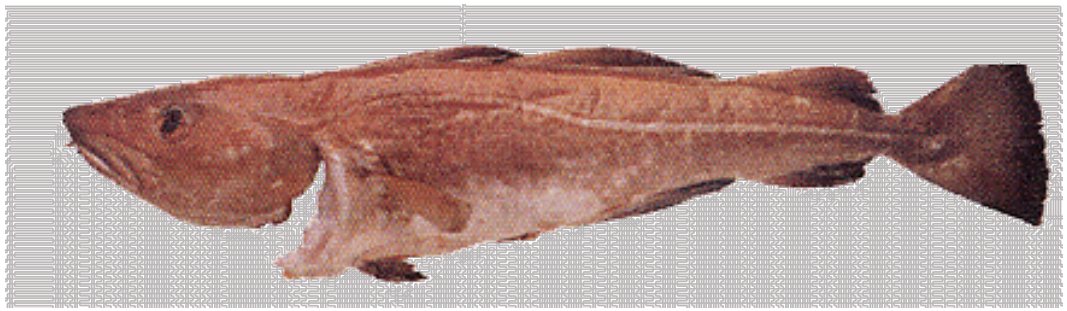


Homology

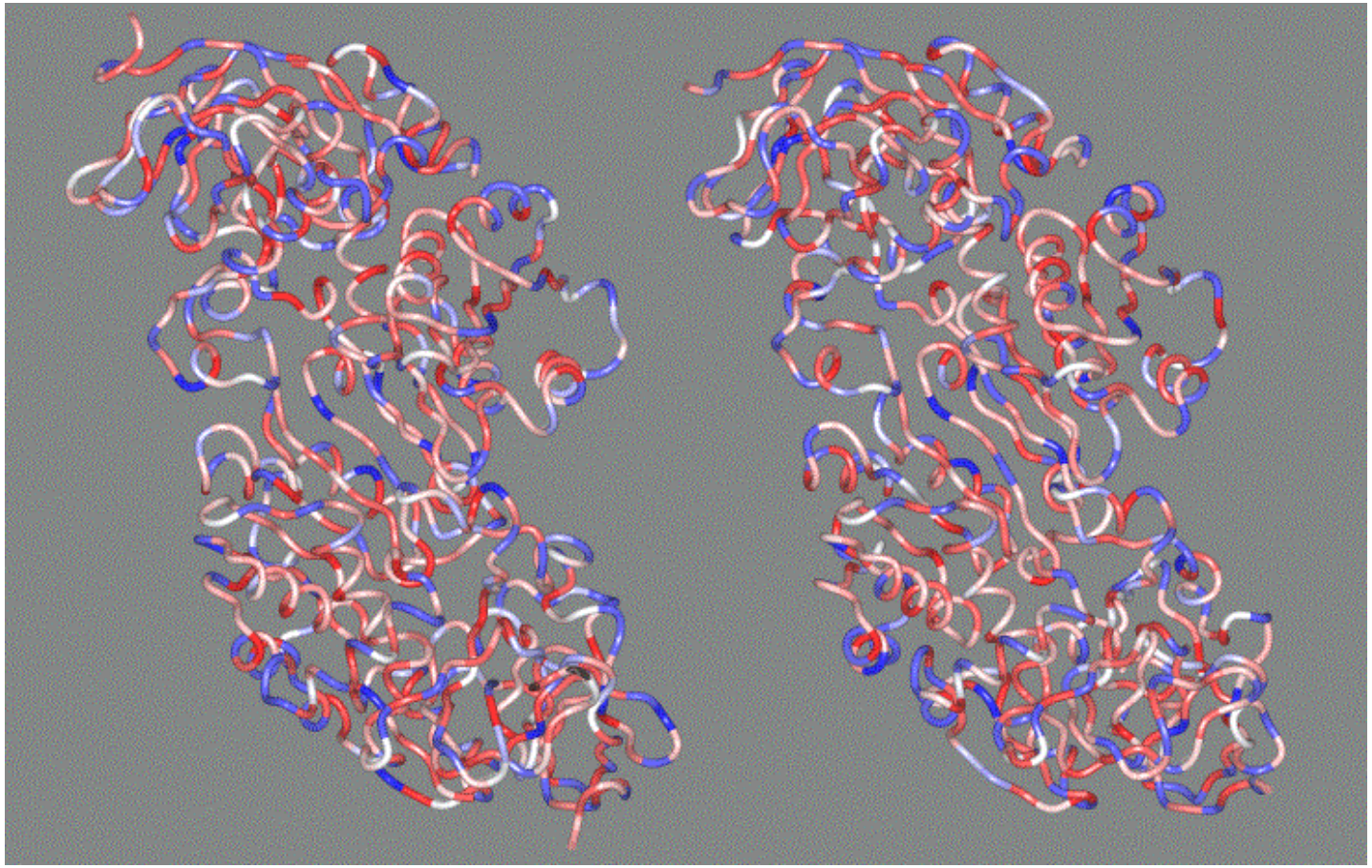
Homology between two proteins/genes can be deduced by their similarity in sequence, structure or function

Species = human

Species = codfish



Difference is self-apparent!



Species = human **Species = codfish**

```
Uomo      - - A N E V I K C K A A V A W E A G K P L S I E E I E V A P P K A H E V R I K I I A T A V C H T D A Y T L S G A D P E G
Merluzzo  A T V G K V I K C K A A V A W E A N K P L V I E E I E V D V P H A N E I R I K I I A T G V C H T D L Y H L F E G K H K D

Uomo      C F P V I L G H E G A G I V E S V G E G V T K L K A G D T V I P L Y I P Q C G E C K F C L N P K T N L C Q K I R V T Q G
Merluzzo  G F P V V L G H E G A G I V E S V G P G V T E F Q P G E K V I P L F I S Q C G E C R F C Q S P K T N Q C V K G W A N E S

Uomo      K G L M P D G T S R F T C K G K T I L Y M G T S T F S E Y T V V A D I S A K I D P L A P L D K V C L L G C G I S T G
Merluzzo  P D V M S P K E T R F T C K G R K V L Q L G T S T F S Q Y T V V N Q I V A K I D P S A P L D T V C L L G C G V S T G

Uomo      Y G A A V N T A K L E P G S V C A V F G L G G V G L A V I M G C K V A G A S R I I G V D I N K D K F A R A K E F G A T E
Merluzzo  F G A A V N T A K V E P G S T C A V F G L G A V G L A A V M G C S A G A K R I I A V D L N P D K F E K A K V F G A T D

Uomo      C I N P Q D F S K P I Q E V L I E M T D G G V D Y S F C N S N V K V M R A A L E A C H K G W G V S V V V G V A A S G E
Merluzzo  F V N P N D H S E P I S Q V L S K M T N G G V D F S L C G N V G V M R N A L E S C L K G W G V S V L V G W T D L H D

Uomo      E I A T R P F Q L V T G R T W K G T A F G G W K S V E S V P K L V S E Y M S K K I K V D E F V T H N L S F D E I N K A F
Merluzzo  - V A T R P I Q L I A G R T W K G S M F G G F K G K D G V P K M V S Y L D K K V K L D E F I T H N M P L E S V - - - -

Uomo      E L M H S G K S N A T V V K L
Merluzzo  - - - - - - - - - - - - - - -
```

Is the difference self-apparent ?

```
Uomo     - - A N E V I K C K A A V A W E A G K P L S I E E I E V A P P K A H E V R I K I I A T A V C H T D A Y T L S G A D P E G
Merluzzo A T V G K V I K C K A A V A W E A N K P L V I E E I E V D V P H A N E I R I K I I A T G V C H T D L Y H L F E G K H K D

Uomo     C F P V I L G H E G A G I V E S V G E G V T K L K A G D T V I P L Y I P Q C G E C K F C L N P K T N L C Q K I R V T Q G
Merluzzo G F P V V L G H E G A G I V E S V G P G V T E F Q P G E K V I P L F I S Q C G E C R F C Q S P K T N Q C V K G W A N E S

Uomo     K G L M P D G T S R F T C K G K T I L H Y M G T S T F S E Y T V V A D I S V A K I D P L A P L D K V C L L G C G I S T G
Merluzzo P D V M S P K E T R F T C K G R K V L Q F L G T S T F S Q Y T V V N Q I A V A K I D P S A P L D T V C L L G C G V S T G

Uomo     Y G A A V N T A K L E P G S V C A V F G L G G V G L A V I M G C K V A G A S R I I G V D I N K D K F A R A K E F G A T E
Merluzzo F G A A V N T A K V E P G S T C A V F G L G A V G L A A V M G C H S A G A K R I I A V D L N P D K F E K A K V F G A T D

Uomo     C I N P Q D F S K P I Q E V L I E M T D G G V D Y S F E C I G N V K V M R A A L E A C H K G W G V S V V V G V A A S G E
Merluzzo F V N P N D H S E P I S Q V L S K M T N G G V D F S L E C V G N V G V M R N A L E S C L K G W G V S V L V G W T D L H D

Uomo     E I A T R P F Q L V T G R T W K G T A F G G W K S V E S V P K L V S E Y M S K K I K V D E F V T H N L S F D E I N K A F
Merluzzo - V A T R P I Q L I A G R T W K G S M F G G F K G K D G V P K M V K A Y L D K K V K L D E F I T H R M P L E S V - - - -

Uomo     E L M H S G K S I R T V V K I
Merluzzo - - - - - - - - - - - - - - -
```

Two sequences are similar if they can be *aligned* so that many corresponding (aligned) amino acids (or nucleotides) are identical or similar

With a sequence alignment we search for a correspondence between amino acids (or nucleotides) that most probably reflects the evolution of proteins (or genes)

Sequence alignment

What is the correspondence between amino acids (or nucleotides) which most likely reflects the evolution of two proteins (or genes)?

Sequence alignment

Aim: minimizing the evolutionary distance between sequences to be aligned, therefore minimizing differences (that is maximizing similarities) between the components (nucleotides or amino acids) of the sequences themselves

The ***hypothesis*** of ***most reasonable*** alignment is the one involving the lowest number of mutations to pass from one sequence to the other one

Sequence alignment

Applications

- Function recognition
  - assessing a significant similarity between two sequences is enough
- Phylogeny
  - Measuring the similarity on a quantitative basis is required
- Model building
  - Explicitly constructing the best possible alignment is required

Sequence alignment

A dot matrix or **dot-plot** provides un immediate view of the similarity between two sequences

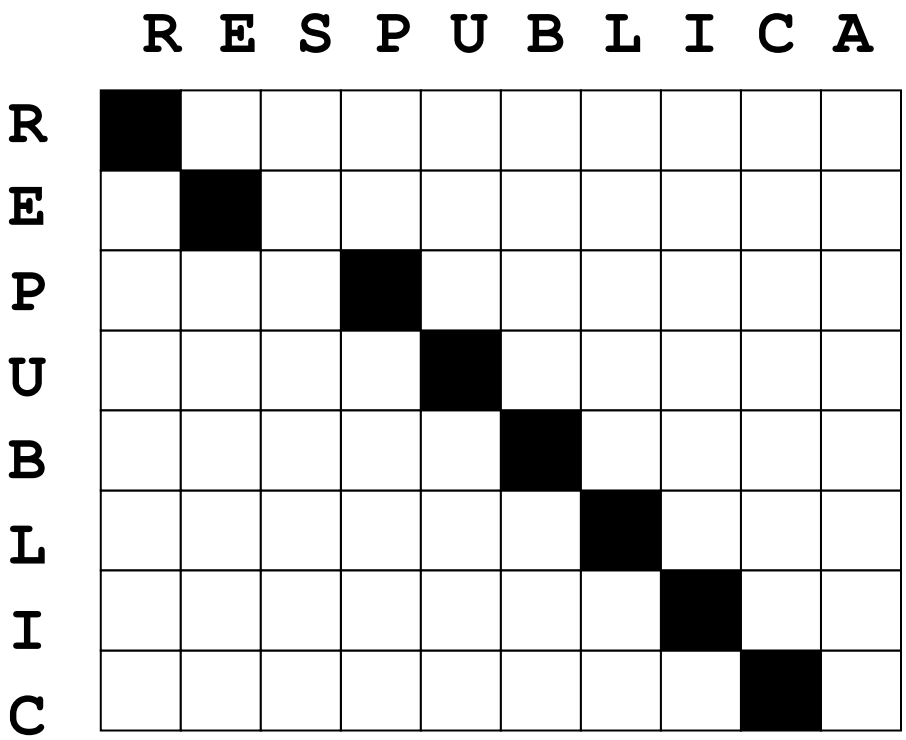
In a **dot-plot** a dot is reported in correspondence of two identical characters

Sequence alignment: dot-plot

Latin -IUSTITIA

Italian GIUSTIZIA

Sequence alignment: dot-plot

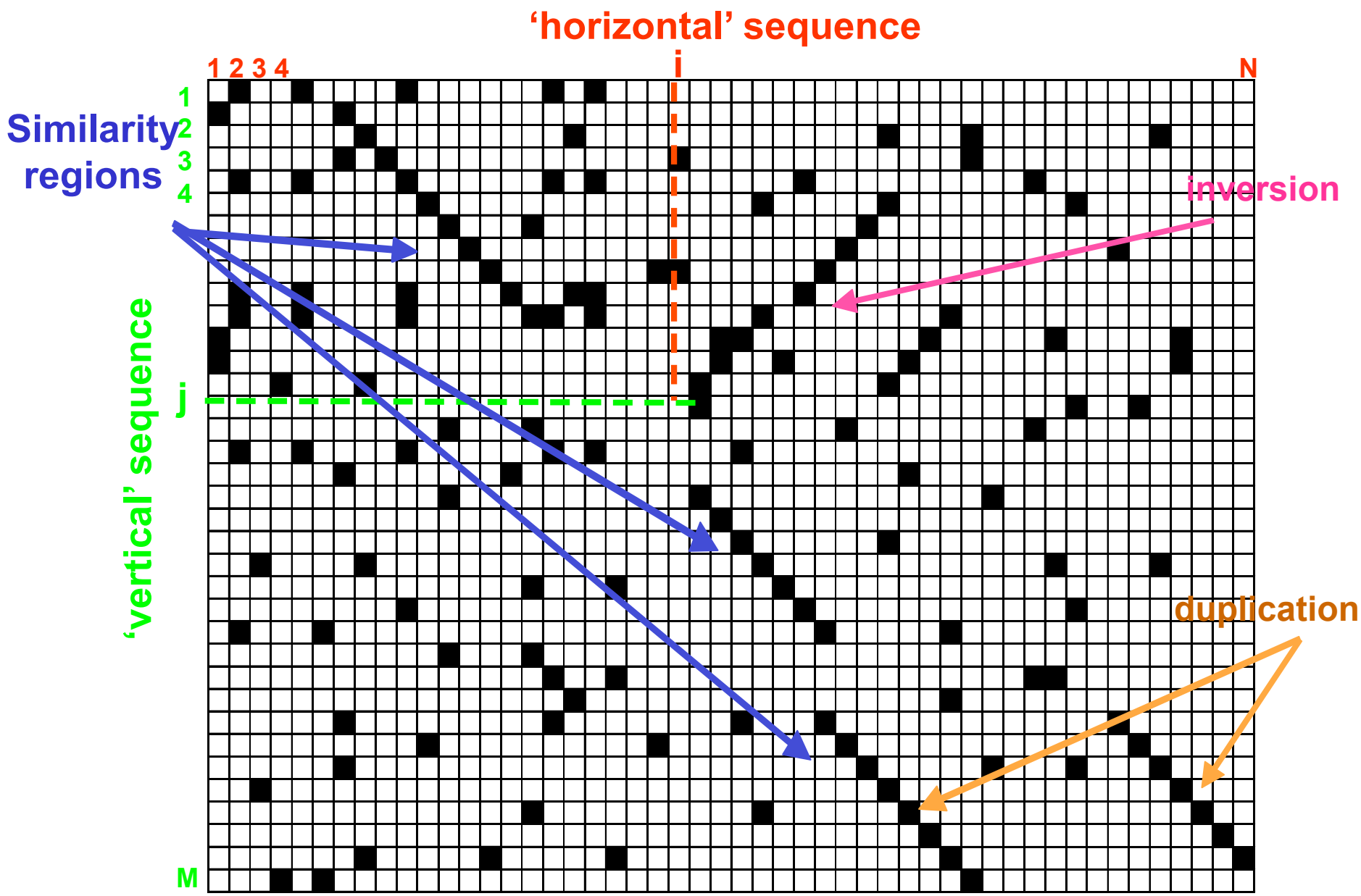


Latin **RESPUBLICA**

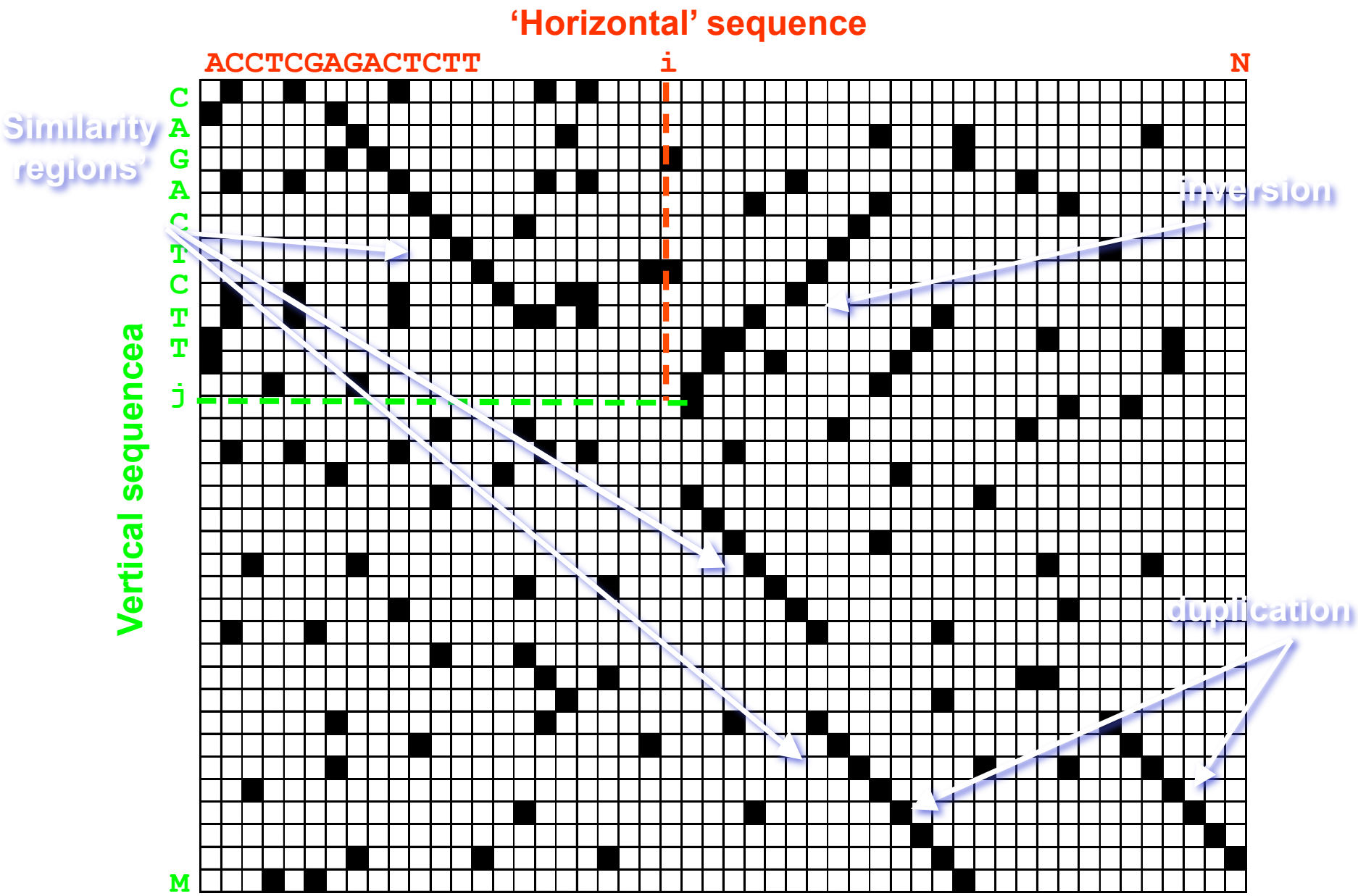
English **RE-PUBLIC-**

deletions

Dot plot


'horizontal' sequence
1 2 3 4
i
N
Similarity regions
1
2
3
4
inversion
'vertical' sequence
j
duplication
M

Dot plot


'Horizontal' sequence
ACCTCGAGACTCTT
i
N
Similarity 'regions'
Vertical sequencea
C
A
G
A
C
T
C
T
T
T
j
M
inversion
duplication

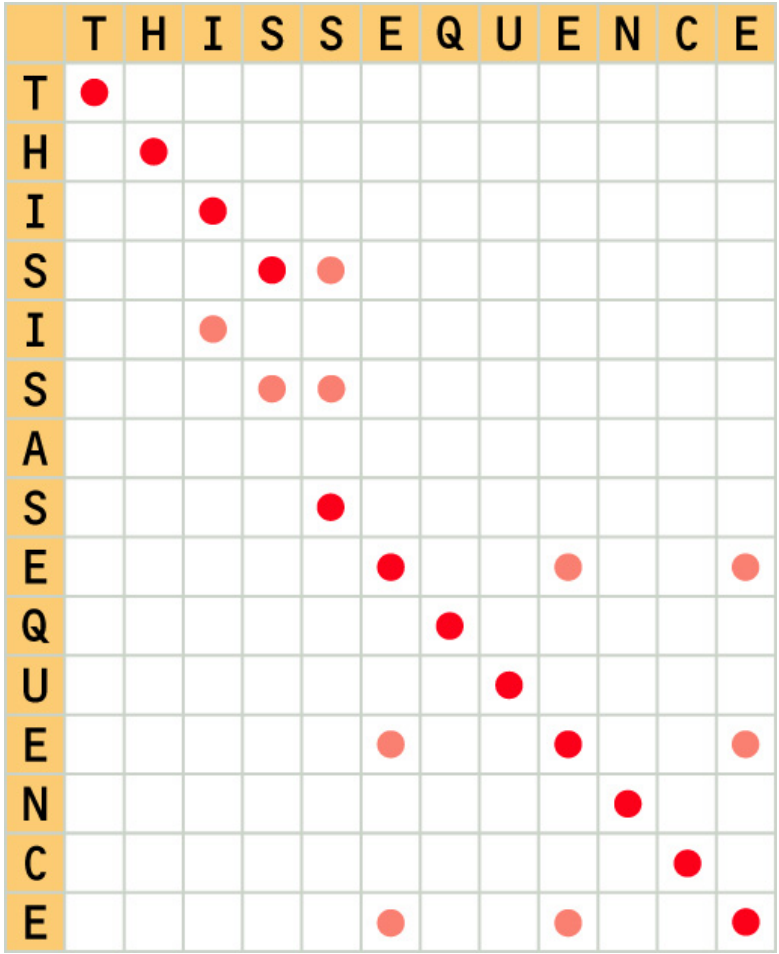
Sequence alignment: dot-plot

Sequence1 (12 charaters)

THISSEQUENCE

Sequence2 (15 charaters)

THISISASEQUENCE



Sequence alignment: dot-plot

Sequence1 (19 characters)

LAMIAPRIMASEQCREATA

Sequence2 (22 characters)

MIAALTRASEQDALLINEARE

	L	A	M	I	A	P	R	I	M	A	S	E	Q	C	R	E	A	T	A
M	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
I	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
S	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

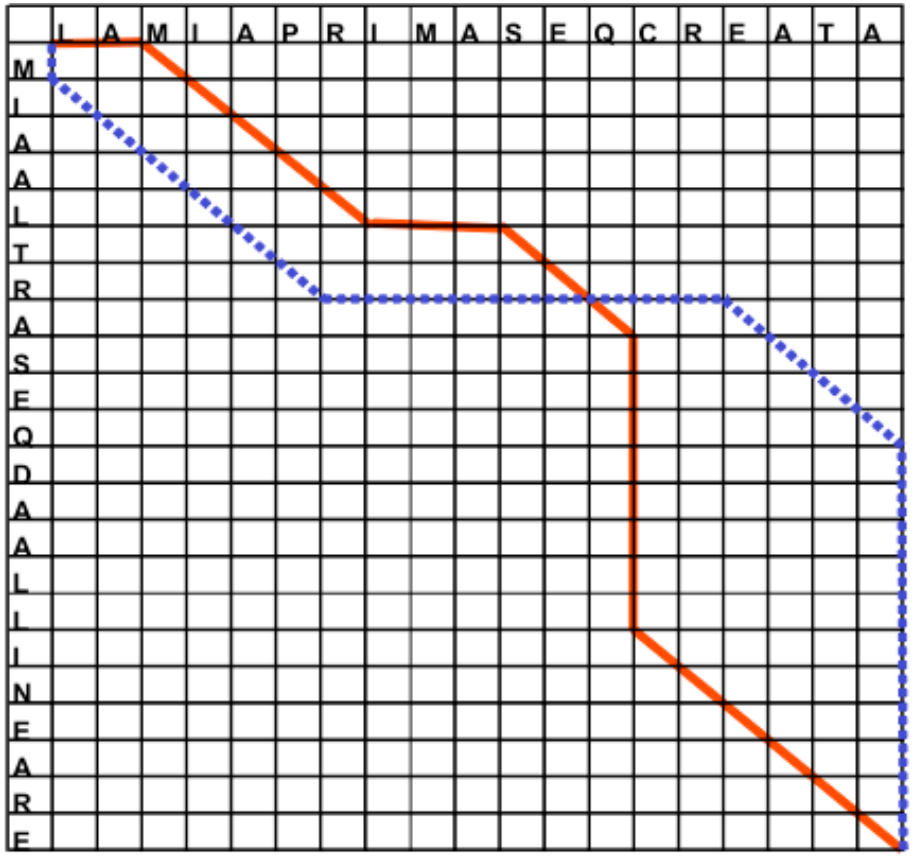
	L	A	M	I	A	P	R	I	M	A	S	E	Q	C	R	E	A	T	A
M	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
I	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
S	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0
A	0	1	0	0	1	0	0	0	0	1	0	0	0	0	0	0	1	0	1
R	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0

10 points

LAMIAP---RIMASEQ-------CREATA

--MIA-ALTR--ASEQDAALLIN--EARE

Sequence alignment




```
LAMIAPRIMASEQCREATA-----------
-MIAALTR-------ASEQDAALLINEARE
  LAMIAPRIMASEQ--------CREATA
  --MIAAL---TRASEQDAALLINEARE
```

Different paths through a dot-plot correspond to different alignments

The quality of an alignment is measured by giving it a quantitative score

Non only the **identity** between amino acids matter, but also the **similarity**

Not all amino acids substitutions are equally likely to occur

Sequence alignment

For obtaining an optimal alignment (the one with maximum score, not necessarily the correct one, reflecting the evolutionary process), we need:

1) A score for the substitution of amino acids/ nucleobases
2) Penalty for insertions/deletions (INDELs)
3) Algorithm to perform the alignment
4) Measure of the alignment significance

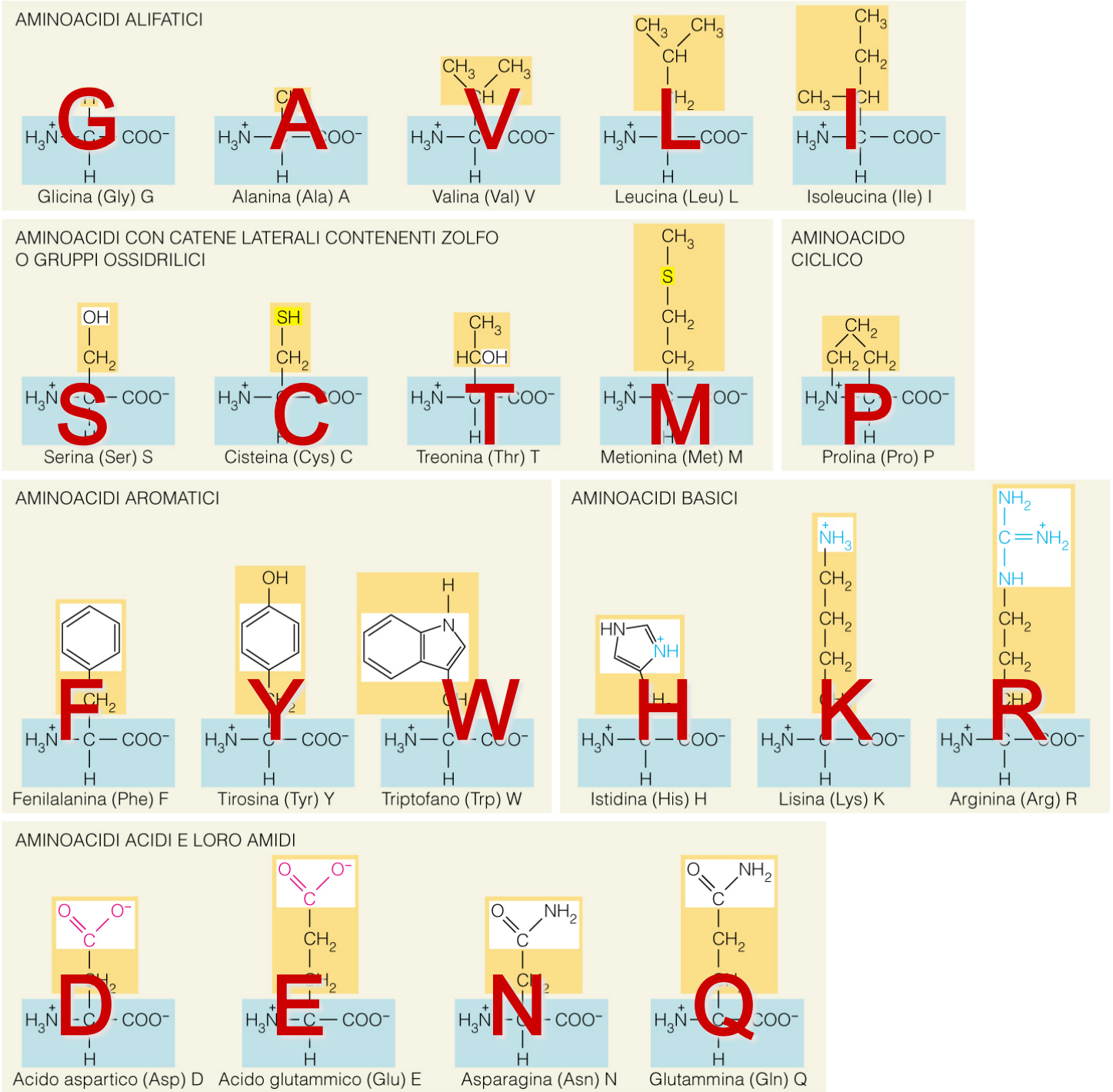
Sequence alignment – Part 1

For obtaining an optimal alignment (the one with maximum score, not necessarily the correct one, reflecting the evolutionary process), we need:

1) A score for the substitution of amino acids/ nucleobases
2) Penalty for insertions/deletions (INDELs)
3) Algorithm to perform the alignment
4) Measure of the alignment significance

1. *Score* for substitutions

- Identity (*1* or *0*) between nucleobases and amino acids
- Physico-chemical properties of amino acids
- Lowest number of nucleobases to be substituted to obtain the observed mutation
- <u>Substitution frequences observed in protein families (first proposed by Margaret Dayhoff in the ‘70s)</u>


AMINOACIDI ALIFATICI
G
Glicina (Gly) G
A
Alanina (Ala) A
V
Valina (Val) V
L
Leucina (Leu) L
I
Isoleucina (Ile) I
AMINOACIDI CON CATENE LATERALI CONTENENTI ZOLFO O GRUPPI OSSIDRILICI
S
Serina (Ser) S
C
Cisteina (Cys) C
T
Treonina (Thr) T
M
Metionina (Met) M
AMINOACIDO CICLICO
P
Prolina (Pro) P
AMINOACIDI AROMATICI
F
Fenilalanina (Phe) F
Y
Tirosina (Tyr) Y
W
Triptofano (Trp) W
AMINOACIDI BASICI
H
Istidina (His) H
K
Lisina (Lys) K
R
Arginina (Arg) R
AMINOACIDI ACIDI E LORO AMIDI
D
Acido aspartico (Asp) D
E
Acido glutammico (Glu) E
N
Asparagina (Asn) N
Q
Glutammina (Gln) Q

G A V L I

S C T M P

F Y W H K R

D E N Q

ex. 1 Substitution Matrix

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

PAM 250

20x20 matrices

***P**oint **A**ccepted **M**utations*

A positive score means that a given *aa* substitution is favorable

A negative score means that a given *aa* substitution is unfavorable

ex. 2 Substitution Matrix

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

BLOSUM 62

<u>BLO</u>cks <u>SU</u>bstitution <u>M</u>atrix (from the derived BLOCKS database)

A positive score means that a given *aa* substitution is favorable

A negative score means that a given *aa* substitution is unfavorable

PAM e BLOSUM matrices report the $\log_2$ of:

$$\frac{f_{ij}}{f_i \times f_j}$$

where

f_{ij} represents the frequency with whom two amino acids are found aligned

ex. f_{AT} *is the number of times we observe an Ala aligned with a Thr, divided by the total number of pairs in an alignment*

f_i and f_j rapresent the frequencies with whom the two amino acids appear in an alignment

ex. f_A *and* f_T *are the number of times we observe an Ala or a Thr, divided by the total number of amino acids in an alignment*

Example of score calculation for the substitution of an Ala with a Thr

$$\frac{f_{ij}}{f_i \times f_j}$$

Seq1 CADGCFTL

Seq2 CTCGHILM

Seq3 TLCGHIAN

Tot aa = 3 x 8 = 24

Tot aligned aa pairs = 3 x 8 = 24

$f_{AT} = 2 / 24 = 0.083$

$f_A = 2 / 24 = 0.083$

$f_T = 3 / 24 = 0.12$

$f_A \times f_T = 0.010$

$$\frac{f_{AT}}{f_A \times f_T} = \frac{0.083}{0.010} = 8$$

$$\ln_2\left(\frac{f_{AT}}{f_A \times f_T}\right) = 3$$

Example of score calculation for the substitution of an Ala with a Thr

$$\frac{f_{ij}}{f_i \times f_j}$$

Seq1	CADGCFTL
Seq2	CTCGHILM
Seq3	TLCGHIAN

Tot aa = 3 x 8 = 24

Tot aligned aa pairs = 3 x 8 = 24

In a substitution matrix we would write 3 at the cross between Ala(A) and Thr(T)

$$\frac{f_{AT}}{f_A \times f_T} = \frac{0.083}{0.010} = 8$$

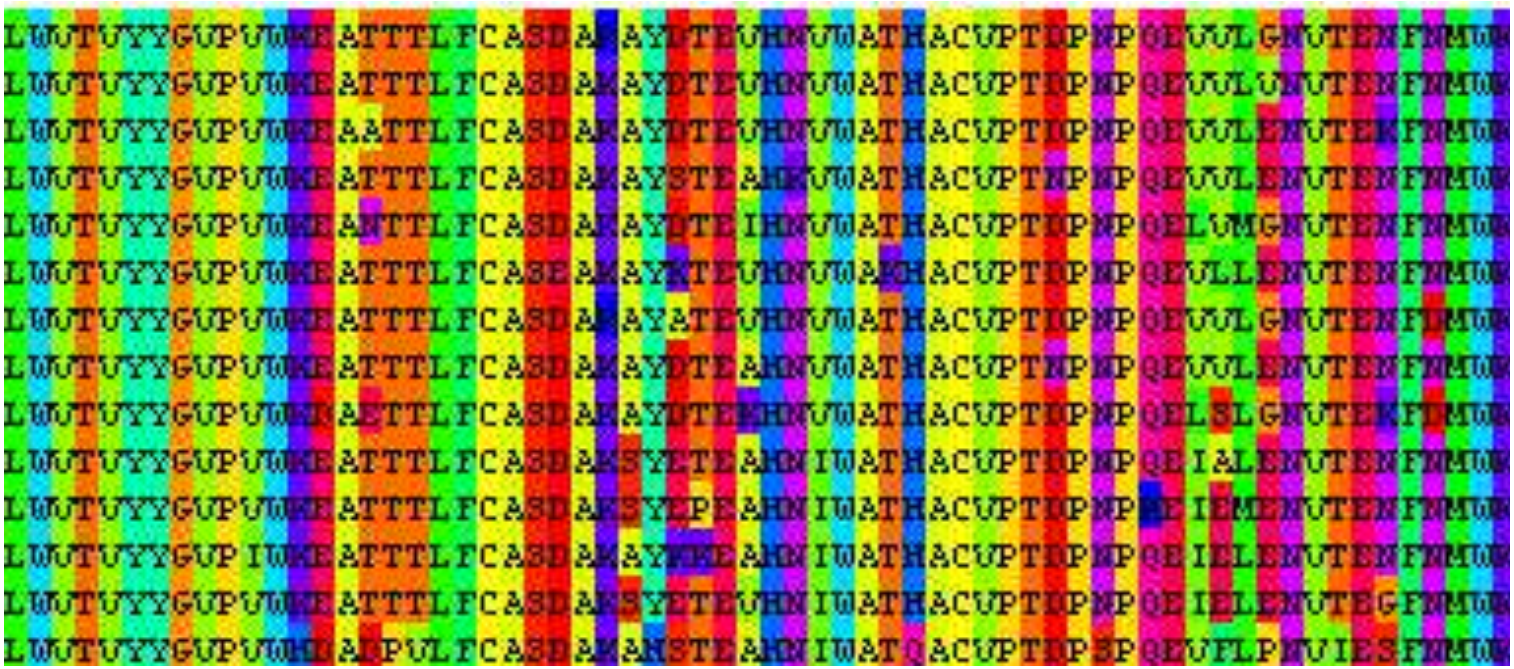
$$\ln_2\left(\frac{f_{AT}}{f_A \times f_T}\right) = 3$$

More rigorously…

$$s(ij) = \text{int}\left[k \cdot \ln_2 \frac{f_{ij}}{f_i \times f_j}\right]$$

We only take the integer

Site-specific matrices, based on empirical rules



PAM N: Percent/Point Accepted Mutations (where N is the number of accepted mutations every 100 aa)

BLOSUM N: BLOcks SUbstitution Matrix (where N is the maximum % of sequence identity between aligned homologs)

- **PAM 1** can be used to generate matrices for higher evolutionary distances:

  multiplying it again and again by itself.

PAM2 = PAM1 * PAM1

- **PAM250:** **etc etc**
  - 2,5 mutations per residue
  - Equivalent to 20% remaining matches between two sequences, that is l'80% of amino acid positions are changed.
  - It is the default matrix used in many analysis software.

ex. 2 Substitution Matrix

- **BLOSUM** matrices have been developed to align scarcely correlated sequences. They have largely replaced the PAM ones.
- They are obtained from the derived **BLOCKS** databank containing alignments of highly correlated protein regions, which can be aligned without gaps.
- **BLOSUM62:** is obtained from alignments of proteine sharing a maximum of 62 % sequence identity. It is largely used. (Corrispondes approximately to a PAM110).

ex. 2 Substitution Matrix

- **BLOSUM** matrices have been developed to align scarcely correlated sequences. They have largely replaced the PAM ones.
- They are obtained from the derived **BLOCKS** databank containing alignments of highly correlated protein regions, which can be aligned without gaps.
- **BLOSUM62:** is obtained from alignments of proteine sharing a maximum of 62 % sequence identity. It is largely used. (Corrispondes approximately to a PAM110).

More recently, matrices have been constructed using newer and larger data sets. The PET91 matrix, e.g., represents a new generation of Dayhoff-type matrices

ex. 1 Substitution Matrix

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

PAM 250

20x20 matrices

***P**oint **A**ccepted **M**utations*

ex. 1 Substitution Matrix

Ala

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

Ala residues are easily substituted by other *aa*

ex. 1 Substitution Matrix

Cys

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

Cys residues are not easily substituted

(they often give disulfide bonds)

ex. 1 Substitution Matrix

Lys

Arg

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

3

Arg & Lys tend to substitute each other

ex. 1 Substitution Matrix

Lys

Val

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

Polar & apolar *aa* do not tend to substitute each other

2. Scoring penalty for INDELs (gaps)

Substitution matrices have been derived from alignments that did not present insertions/ deletions (INDELs). Indels need therefore to be dealt with separately, on an empirical basis.

In aligning two sequences an lgorithm would tend to maximize the score (correspondence between identical or similar amino acids) by inserting a large number of gaps.

Is this the way which best reflects evolution?

2. Scoring penalty for INDELs (gaps)

We have indels (gaps) when a letter of a stretch of letters in one sequence is paired up with blanks spaces in another one

In nature INDEL events are often lethal (deleterious)

Therefore we need to penalize insertions and deletions. That means associating to them a negative score to be subtracted to the total score of the alignment .

2. Scoring penalty for INDELs (gaps)

In nature deletion of a series of contiguous nucleobases/amino acids is a more likely event than the independent deletion of the same number of nucleobases/amino acids in non contiguous positions

Let's distinguish the start of (introducing) a gap :

```
EGQTCA
AG-TCL
```

from the extension of (extending) a gap:

```
EGQQQTCA
AG---TCL
```

2. Scoring penalty for INDELs (gaps)

In nature deletion of a series of contiguous nucleobases/amino acids is a more likely event than the independent deletion of the same number of nucleobases/amino acids in non contiguous positions

Let's distinguish the start of a gap:

```
EGQTCA
AG-TCL
```

from the extension of a gap:

```
EGQQQTCA
AG---TCL
```

We penalize more a gap start than a gap extension

example:

-11 start

-1 extension

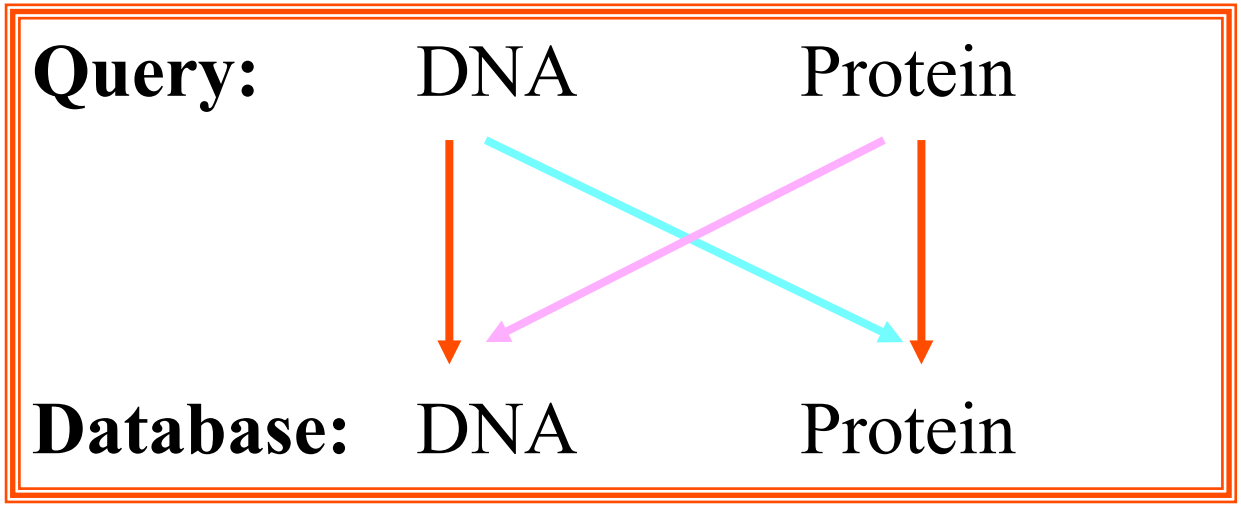
Sequence alignment – Part 1

For obtaining an optimal alignment (the one with maximum score, not necessarily the correct one, reflecting the evolutionary process), we need:

1) A score for the substitution of amino acids/ nucleobases
2) Penalty for insertions/deletions (INDELs)
3) Algorithm to perform the alignment
4) Measure of the alignment significance

Homology search in databases

- Protein *vs.* proteins
- Gene (tranlastion to aa) *vs.* proteins
- Gene *vs.* genes
- Protein *vs.* translation to *aa* of nucleotide sequences (all frames)



When we compare protein sequences we search for the best correspondence for 20 different amino acids

When we compare nucleotide sequences we search for the best correspondence for only 4 nucleotides (nucleobases)

When we compare protein sequences we search for the best correspondence for 20 different amino acids

When we compare nucleotide sequences we search for the best correspondence for only 4 nucleotides (nucleobases)

Probability of finding a good correspondence (high score alignment) by chance is higher for nucleotide sequences than for protein sequences

Furthermore, when we compare protein sequences we can take into account the similarity between amino acids

When we compare protein sequences we search for the best correspondence for 20 different amino acids

When we compare nucleotide sequences we search for the best correspondence for only 4 nucleotides (nucleobases)

Probability of finding a good correspondence (high score alignment) by chance is higher for nucleotide sequences than for protein sequences

Furthermore, when we compare protein sequences we can take into account the similarity between amino acids

When possible, comparing protein sequences has to be preferred!

How we can “fish” from databases potentially homologous sequences?

Exact algorithms (Smith-Waterman)

Exact, it provides the best alignment(s) for a pair of sequences.

Given 2 sequences: A of length n and B of length m,
Smith-Waterman takes n*m computational steps.

If we search for homologs of the query sequence A (n=200 aa)

In a database made of 10^6 sequences with m=200 aa

The number of computational steps is = 10^6 x 200 x 200 = ~10^{10}

10^3 steps per sec = 10^7 secs = 120 days = 4 months!

There is a need for approximate (heuristic) algorithms

Exact algorithms (Smith-Waterman)

Exact, it provides the best alignment(s) for a pair of sequences.

Given 2 sequences: A of length n and B of length m, Smith-Waterman takes n*m computational steps.

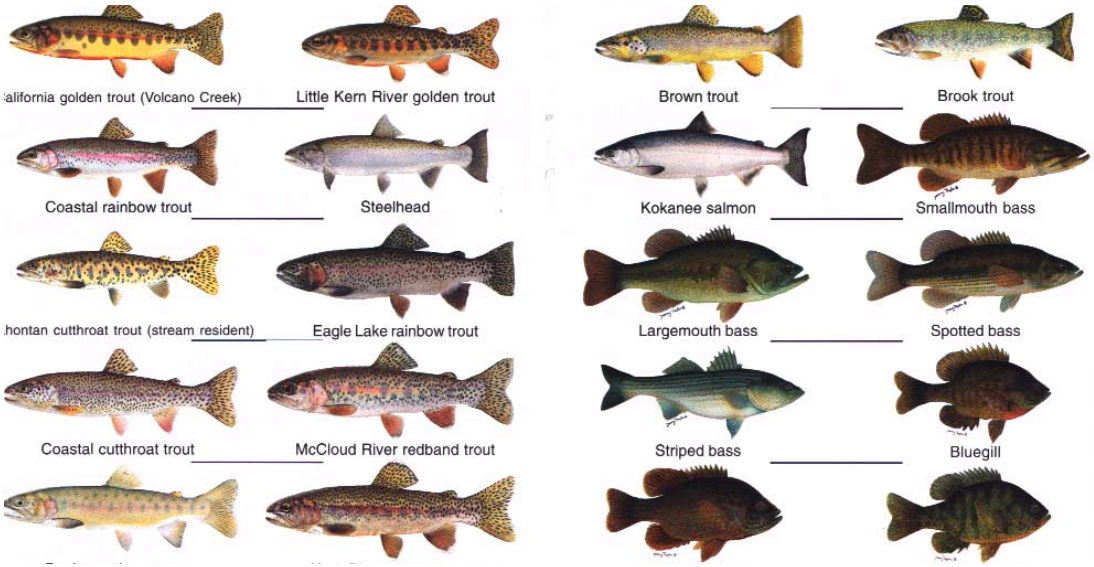
How do we discard irrelevant alignments?

Heuristic algorithms (BLAST, FASTA) are needed to discard most of the irrelevant alignments.

Software such as **FASTA** and **BLAST**, starting from a query sequence:

Software such as **FASTA** and **BLAST**, starting from a query sequence: 

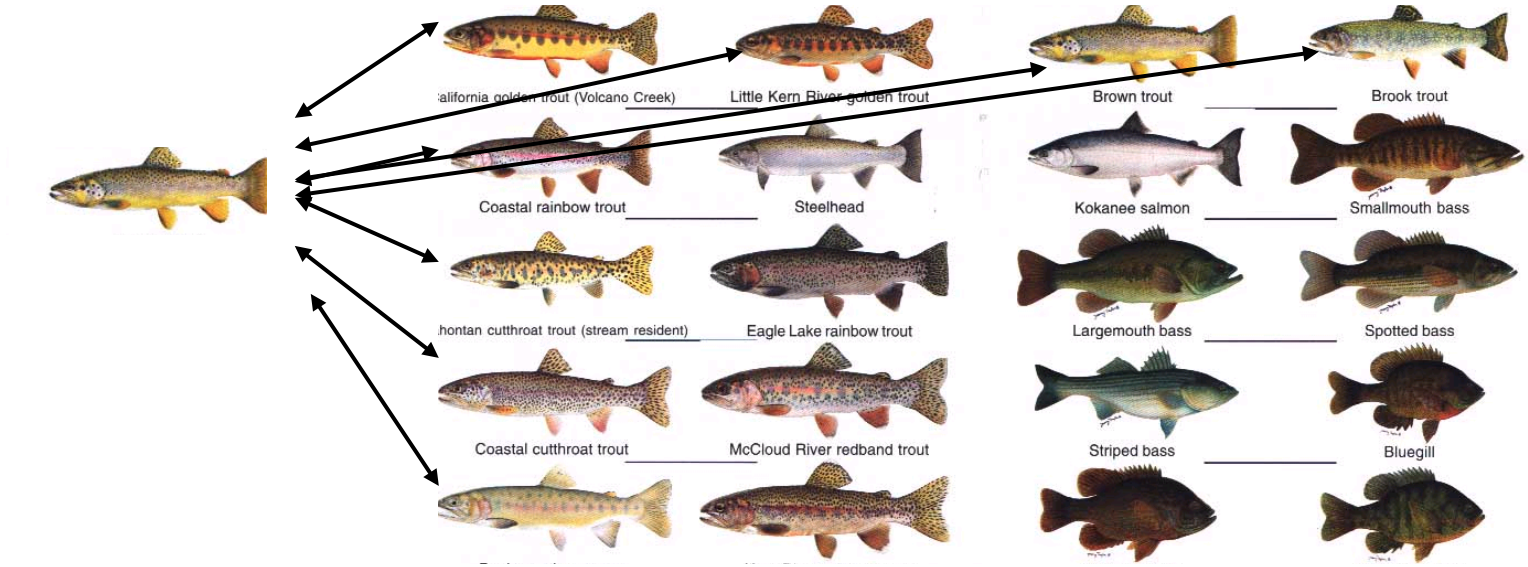
first "fish" from databases a subset of sequences which are potential homologs



Software such as **FASTA** and **BLAST**, starting from a query sequence:

first "fish" from databases a subset of sequences which are potential homologs

then perform the best alignment of each sequence in the subset with the query sequence



FASTA: example

query

ACDDEFGSATRMASTRK

↓

Data bank (DB)

Seq 1: LKDCDDAFSGSTLTLMRASRK

Seq 2: ACKRAEFSGSVTRMLSTRK

Seq 3: ACDDEFGLLLTRYTMASTRK

Step 1 **= Division of the sequence in 2-letter words (k-tuples).**

Possible words:

AC, CD, DD, DE, EF, FG, GS, SA, AT, TR, RM, MA, AS, ST, RK

Note. *A typical value of* ***k*** *for DNA is 6*

Step 2 = Table of word frequencies

Query: `ACDDEFGSATRMASTRK`

DB: **Seq 1** `LKDCDDAFSGSTLTLMRASRK`
Seq 2 `ACKRAEFSGSVTRMLSTRK`
Seq 3 `ACDDEFGLLLTRYTMASTRK`

Word	Query	Seq 1	Seq 2	Seq 3	Off1	Off2	Off3
AC	1	-	1	1	-	0	0
CD	2	4	-	2	2	-	0
DD	3	5	-	3	2	-	0
DE	4	-	-	4	-	-	0
EF	5	-	6	5	-	1	0
FG	6	-	-	6	-	-	0
GS	7	10	9	-	3	**2**	-
SA	8	-	-	-	-	-	-
AT	9	-	-	-	-	-	-
TR	10	-	12, 17	11	-	**2**, 7	1
RM	11	-	13	-	-	**2**	-
MA	12	-	-	15	-	-	3
AS	13	18	-	16	5	-	3
ST	14	11	16	17	-3	**2**	3
RK	16	20	18	19	4	**2**	3

Step 3 = Similarity *score* calculation *Init1* (based on the Table at the step 2)

Query: ACDDEFGSATRMASTRK

DB: Seq 1 LKDCDDAFSGSTLTLMRASRK
Seq 2 ACKRAEFSGSVTRMLSTRK
Seq 3 ACDDEFGLLLTRYTMASTRK

Query		A	C	D	D	E	F	G	S	A	T	R	M	A	S	T	R	K	
Pos		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	
								X	X		X	X			X	X	X	X	
Seq2	A	C	K	R	A	E	F	S	G	S	V	T	R	M	L	S	T	R	K	T
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Off									2			2				2		2		

Init1(seq2) based on this approximate alignment

Query: ACDDEFGSATRMASTRK

Table of word frequencies

Word	Query	Seq 1	Seq 2	Seq 3	Off1	Off2	Off3
AC	1	-	1	1	-	0	0
CD	2	4	-	2	2	-	0
DD	3	5	-	3	2	-	0
DE	4	-	-	4	-	-	0
EF	5	-	6	5	-	1	0
FG	6	-	-	6	-	-	0
GS	7	10	9	-	3	2	-
SA	8	-	-	-	-	-	-
AT	9	-	-	-	-	-	-
TR	10	-	12, 17	11	-	2, 7	1
RM	11	-	13	-	-	2	-
MA	12	-	-	15	-	-	3
AS	13	18	-	16	5	-	3
ST	14	11	16	17	-3	2	3
RK	16	20	18	19	4	2	3

Step 4 = Similarity *score* calculation *InitN* (based on the alignment at step 3 and on the Table at step 2)

Query: ACDDEFGSATRMASTRK

DB: Seq 1 LKDCDDAFSGSTLTLMRASRK
Seq 2 ACKRAEFSGSVTRMLSTRK
Seq 3 ACDDEFGLLLTRYTMASTRK

Query	A	C	-	D	D	E	F	-	G	S	A	T	R	M	A	S	T	R	K	
Pos	1	2		3	4	5	6		7	8	9	10	11	12	13	14	15	16	17	
	X	X				X	X		X	X		X	X			X	X	X	X	
Seq2	A	C	K	R	A	E	F	S	G	S	V	T	R	M	L	S	T	R	K	T
Pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Off	**0**					**1**			**2**			**2**				**2**		**2**		

InitN(seq2) = ***Init1***(seq2) + score(novel matches) – ***K***(gap)

Step 5 **= Final alignment of the sequences with the best *InitN score with the query sequence* and calculation of the final score *opt* (score for the novel, complete alignment)**

NOTE. **The <u>choice of the sequences subset</u> in the DB with whom the optimal alignment is finally performed is based on the approximate scores Init1 & InitN**

FASTA

1. Divides the query sequence in 2-letter words (k-tuples).
2. Finds these words in the database sequences and calculates the offset
3. Calculates the similarity of the ten regions with most identical words for each sequence in the DB (init1)
4. Calculates the similarity of the ten regions with most identical words including penalization for insertions & deletions
5. Accurately aligns the N sequences with best initN score → obtaining opt

How good an alignment is?

How good an alignment is?

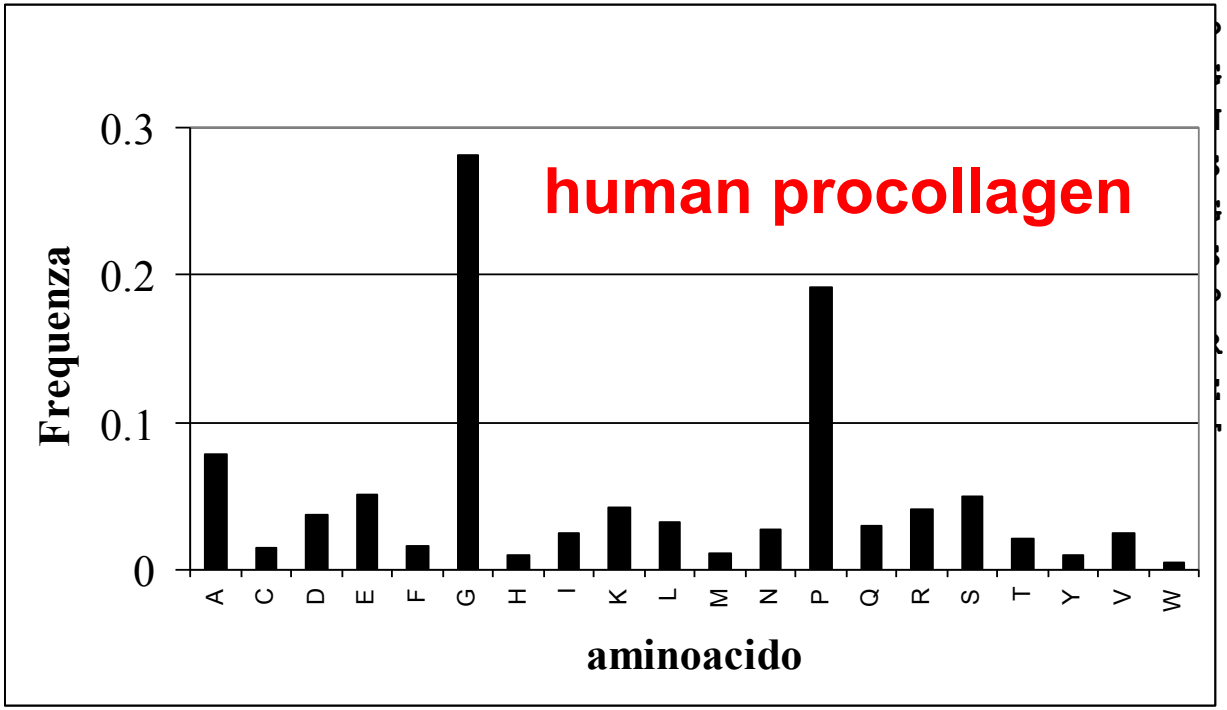
=

How better than a random alignment it is?

- (Unrelated) sequences which give a random alignment:
  - Non-homologous sequences
  - Shuffled sequences
  - Randomly generated sequences
  - Low complexity sequences

Low complexity

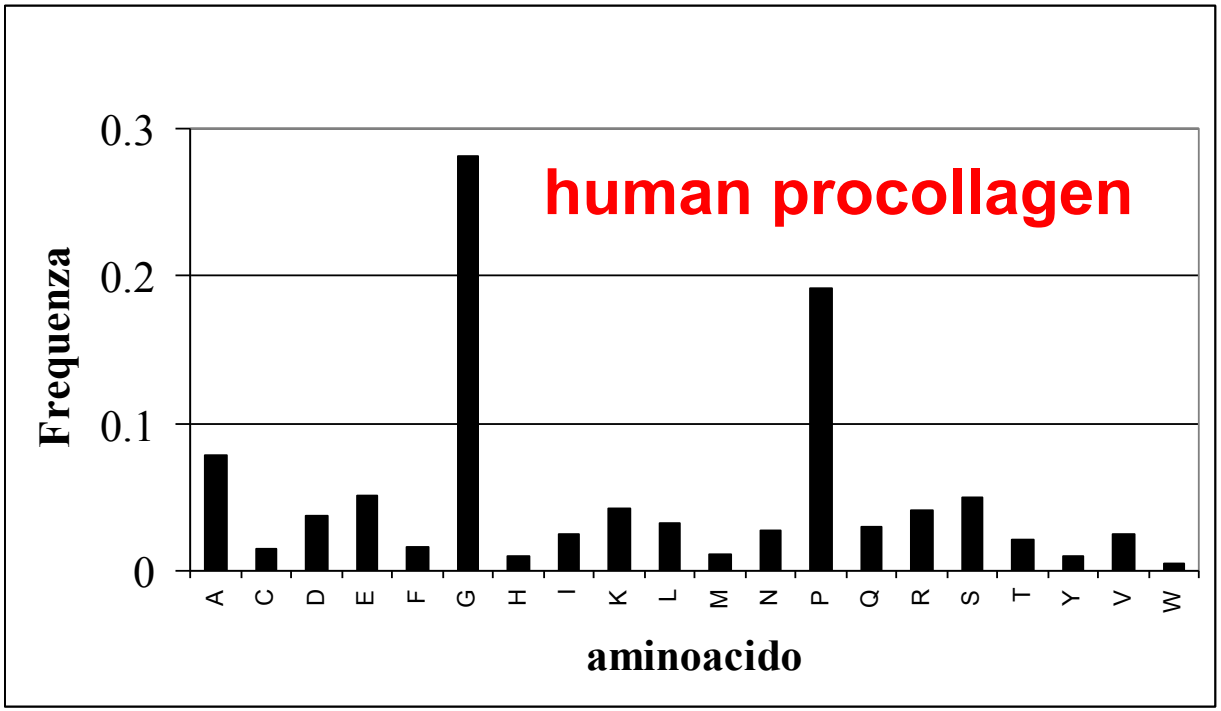
```
MMSFVQKGSW LLLALLHPTI ILAQQEAVEG GCSHLGQSYA DRDVWKPEPC QICVCDSGSV LCDDIICDDQ
ELDCPNPEIP FGECCAVCPQ PPTAPTRPPN GQGPQGPKGD PGPPGIPGRN GDPGIPGQPG SPGSPGPPGI
CESCPTGPQN YSPQYDSYDV KSGVAVGGLA GYPGPAGPPG PPGPPGTSGH PGSPGSPGYQ GPPGEPGQAG
PSGPPGPPGA IGPSGPAGKD GESGRPGRPG ERGLPGPPGI KGPAGIPGFP GMKGHRGFDG RNGEKGETGA
PGLKGENGLP GENGAPGPMG PRGAPGERGR PGLPGAAGAR GNDGARGSDG QPGPPGPPGT AGFPGSPGAK
GEVGPAGSPG SNGAPGQRGE PGPQGHAGAQ GPPGPPGING SPGGKGEMGP AGIPGAPGLM GARGPPGPAG
ANGAPGLRGG AGEPGKNGAK GEPGPRGERG EAGIPGVPGA KGEDGKDGSP GEPGANGLPG AAGERGAPGF
RGPAGPNGIP GEKGPAGERG APGPAGPRGA AGEPGRDGVP GGPGMRGMPG SPGGPGSDGK PGPPGSQGES
GRPGPPGPSG PRGQPGVMGF PGPKGNDGAP GKNGERGGPG GPGPQGPPGK NGETGPQGPP GPTGPGGDKG
DTGPPGPQGL QGLPGTGGPP GENGKPGEPG PKGDAGAPGA PGGKGDAGAP GERGPPGLAG APGLRGGAGP
PGPEGGKGAA GPPGPPGAAG
GPAGQPGDKG EGGAPGLPGI
VAGPPGGSGP AGPPGPQGVK
TGAPGSPGVS GPKGDAGQPG
GKPGANGLSG ERGPPGPQGL
PPGPVGPAGK SGDRGESGPA
PGPAGQQGAI GSPGPAGPRG
GAPGPCCGGV GAAAIAGIGG
NCRDLKFCHP ELKSGEYWVD
SMDGGFQFSY GNPELPEDVL
KAEGNSKFTY TVLEDGCTKH
```



Low complexity

Low complexity regions in protein sequences have a highly biased amino acid composition, often repeats of proline, alanine, serine, glycine, leucine, and glutamic acid

Especially abundant in eukaryotic proteins



Are they homologous sequences?

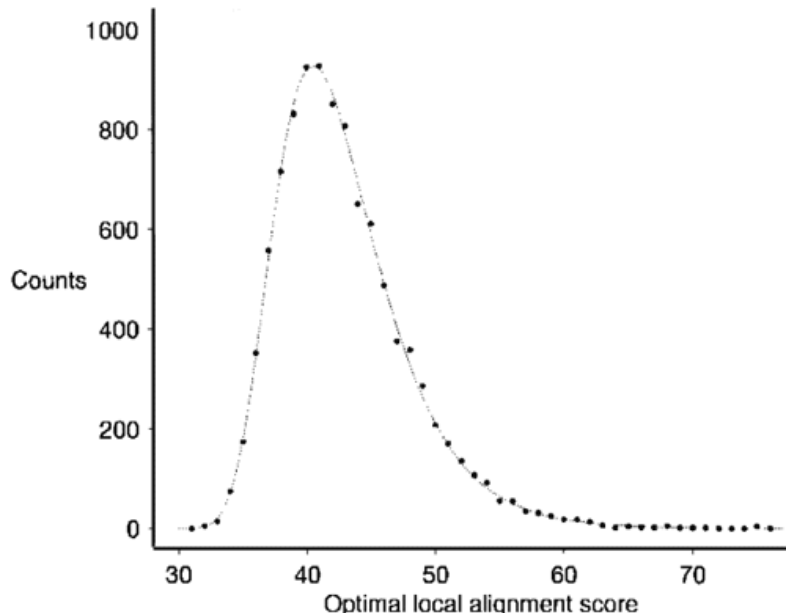
Evaluating the significance of the alignment

a) Generating a large number of random sequences with the same composition of the query seq ("shuffled" sequences)

b) Ripeting the similarity search on random of the DBs using as a query each of the random sequence

c) Calculating corresponding *opt* scores, their average value M_{random} and their standard devation σ_{random}

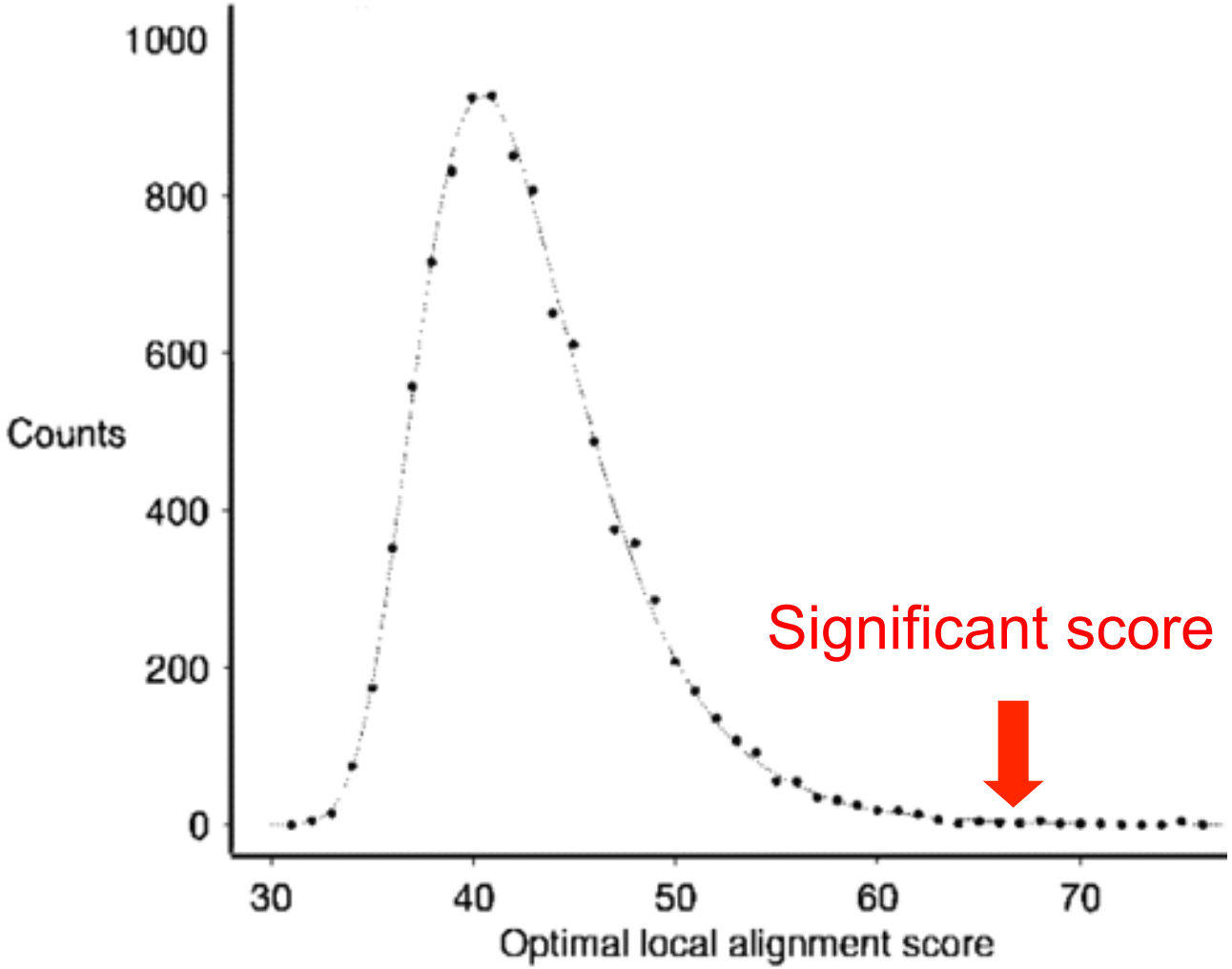
Distribution of random scores



Are they homologous sequences?

Evaluating the significance of the alignment

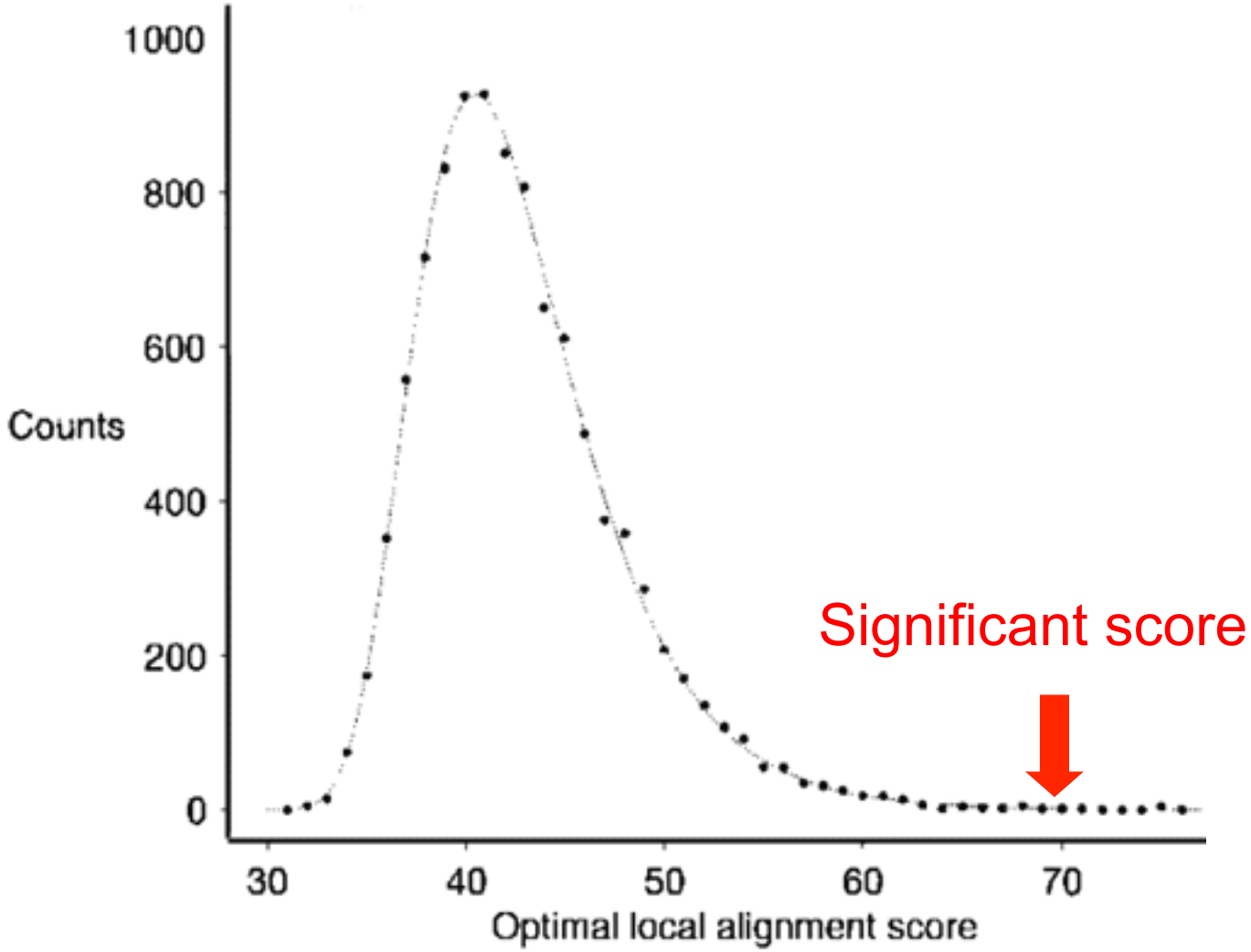
Two sequences can be considered homologous if the optimal score (opt) for their alignment falls off the random scores distribution



Are they homologous sequences?

Evaluating the significance of the alignment

4. Calculating the *Z-score* and the expectation value (*E-value)* for the aligment of the query sequence with its putative homologs



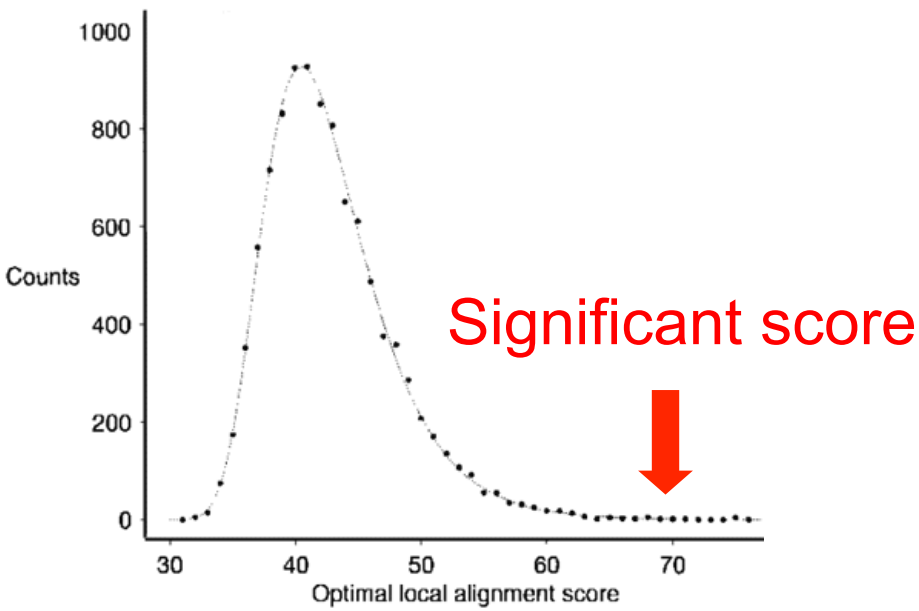
Are they homologous sequences?

Evaluating the significance of the alignment

4. Calculating the ***Z-score*** and the expectation value (*E-value)* for the aligment of the query sequence with its putative homologs

Z-score = number of standard deviations which separate the *query score (opt)* from the average of the random scores

$$\text{Z-score } (S) = (opt_{query} - M_{random}) / \sigma_{random}$$



Z-score » 4

➔ opt_{query} off the random distribution

average

$$M_{\text{random}} = \frac{\sum_i (opt_i)}{n}$$

standard deviation

$$\sigma_{\text{random}} = \sqrt{\frac{\sum_i (opt_i - M_{\text{random}})}{n - 1}}$$

Are they homologous sequences?

Evaluating the significance of the alignment

E-value = *expectation value*: number of alignments with a score **≥** S (o opt) that would be expected by chance by searching a complete database of size *n* (length of all sequences)

Indicates how probable is finding a score S by chance

$$\textit{E-value} = \kappa mn \exp(-\lambda S)$$

Are they homologous sequences?

Evaluating the significance of the alignment

E-value = *expectation value*: number of alignments with a score ≥ S (o opt) that would be expected by chance by searching a complete database of size *n* (length of all sequences)

The lower, the better!

Statistical parameters, dipending on the matrix and the DB

$$E\text{-}value = \kappa mn \exp(-\lambda S)$$

size of the query (m) size of the database (n)

Are they homologous sequences?

Evaluating the significance of the alignment

E-value = *expectation value*: number of alignments with a score **≥** S (o opt) that would be expected by chance by searching a complete database

The typical threshold for a good E−value from a FASTA/BLAST search is $E=10^{-5}$ *or lower*

$$\textit{E-value} = \kappa mn \exp(-\lambda S)$$

The **probability** of having by chance an alignment with a score ≥ S is given by:

$$P = 1-e^{(-\kappa mn \exp(-\lambda S))}$$

Are they homologous sequences?

Evaluating the significance of the alignment

It is possible to normalize the score:

$$S' = (\lambda S - \ln \kappa) / \ln 2$$

Bit-score

⇩

$$E\text{-}value = mn\, 2^{-S'}$$

The bit-score is independent from query sequence length and database size

Thus, it is possible to compare directly the obtained bit-scores from searches in different databases and with different matrices

FASTA

Protein | Nucleotide | Genomes | Proteomes | Whole Genome Shotgun | Web services | Help & Documentation | Also in this section

Feedback

Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database)

STEP 1 - Select your databases

PROTEIN DATABASES

1 Database Selected — X Clear Selection

- UniProt Knowledgebase (The UniProt Knowledgebase includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL)
- UniProtKB/Swiss-Prot (The manually annotated section of UniProtKB)
- UniProtKB/Swiss-Prot isoforms (The manually annotated isoforms of UniProtKB/Swiss-Prot)
- UniProtKB/TrEMBL (The automatically annotated section of UniProtKB)
- UniProtKB Reference Proteomes plus Swiss-Prot
- UniProtKB COVID-19
- **UniProtKB Taxonomic Subsets**
- **UniProt Clusters**
- **Patents**
- **Structures**
- **Other Protein Databases**

STEP 2 - Enter your input sequence

Enter or paste a PROTEIN sequence in any supported format:

```
>NP_001382996.1 putative keratin-associated protein 4-16 [Homo sapiens]
MCSSKMPCSPSASSLCAASPPNCCHPSCCQTTCCRTTSCSHSCSVSSCCRPQCCHSVCCQPTCCRPSCCQ
TTCCRTTCCHPSCCVSSCCRPQCCHSVCFQPTCCHPSCCISSSCCPSCCESSCCCPCCCLRPVCGRVSCH
VTCYHPTCVISTCPHPLCCASPPLPLPFPSPPVPLPFFLSLALPSPPRPSPPLLSPVLIPSPSPSPSLPS
LSPPLPSPPLPSPHFPSVNPKSMLQ
```

or Upload a file: Choose File no file selected

Use a example sequence | Clear sequence | See more example inputs

no file selected

STEP 3 - Set your parameters

PROGRAM: FASTA

MATRIX	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION	EXPECTATION
BLOSUM50	-10	-2	2	10	0 (default)

DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES
N/A	no	none	Regress

SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs
50	50	START-END	START-END	no

SCORE FORMAT	ANNOTATION FEATURES
Default	no

STEP 4 - Submit your job

☐ Be notified by email *(Tick this box if you want to be notified by email when the results are available)*

Submit

Protein | Nucleotide | Genomes | Proteomes | Whole Genome Shotgun | Web services | Also in this section | Feedback

Tools > Sequence Similarity Searching > FASTA

Results for job fasta-I20220325-093358-0760-32105718-p2m

Summary Table | Tool Output | Visual Output | Functional Predictions | Submission Details

Selection:

Select All | Invert | Clear

Apply to selection:

Annotations:

Show | Hide

Alignments:

Show | Hide

Entries:

Download in

fasta

format

Tools:

Launch

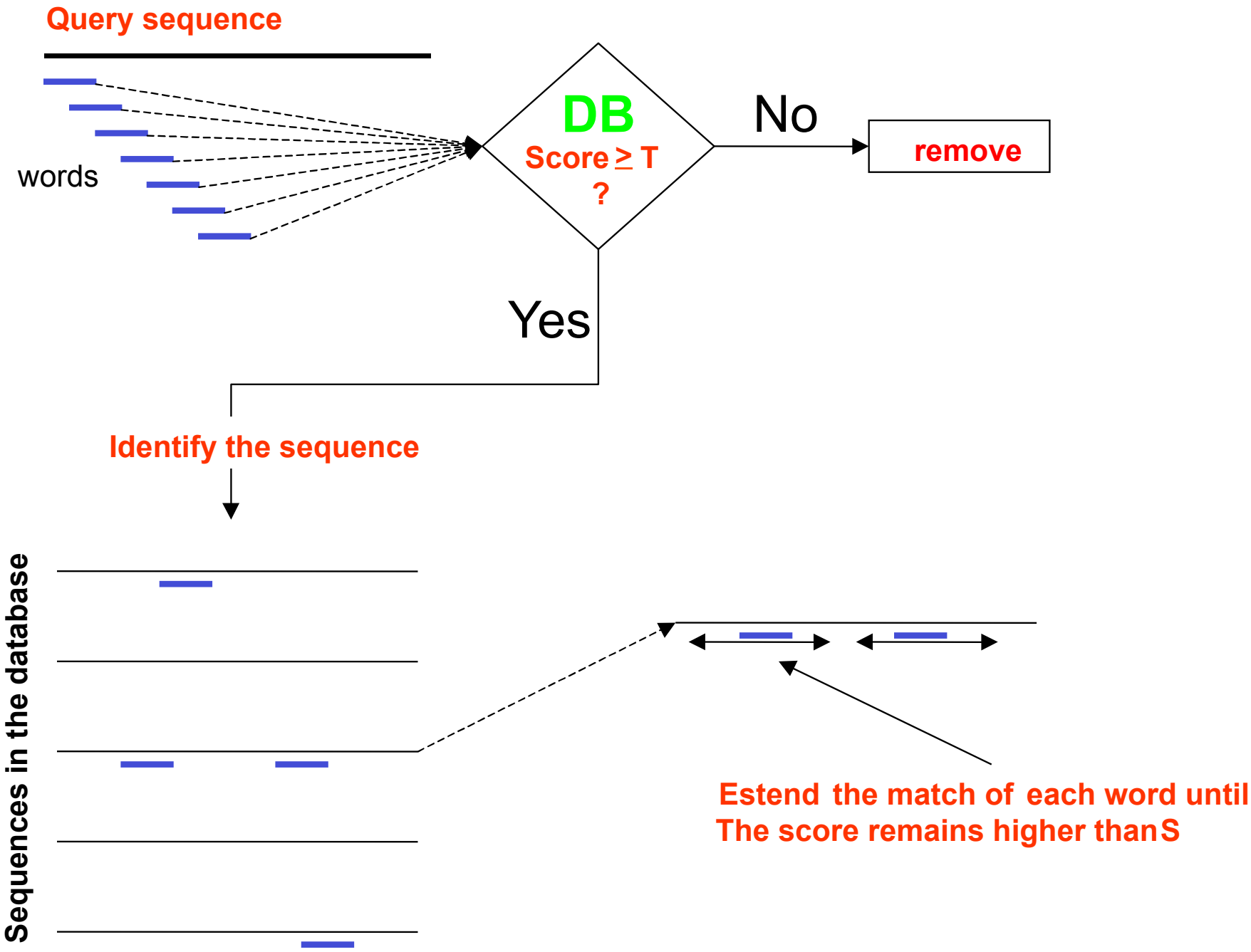
Clustal Omega

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
1	SP:G5E9R7	Putative keratin-associated protein 4-16 OS=Homo sapiens OX=9606 GN=KRTAP4-16 PE=5 SV=1 *Cross-references and related information in:* Gene expression, Nucleotide sequences, Genomes & metagenomes, Literature, Samples & ontologies, Protein families, Protein expression data, Protein sequences	235	219.0	100.0	100.0	3.8E-56
2	SP:Q9BQ66	Keratin-associated protein 4-12 OS=Homo sapiens OX=9606 GN=KRTAP4-12 PE=1 SV=1 *Cross-references and related information in:* Gene expression, Bioactive molecules, Nucleotide sequences, Genomes & metagenomes, Literature, Samples & ontologies, Molecular interactions, Protein families, Protein expression data, Protein sequences, Diseases	201	133.6	81.4	89.4	1.6E-30
3	SP:Q9BYR2	Keratin-associated protein 4-5 OS=Homo sapiens OX=9606 GN=KRTAP4-5 PE=1 SV=4 *Cross-references and related information in:* Gene expression, Bioactive molecules, Nucleotide sequences, Genomes & metagenomes, Literature, Samples & ontologies, Molecular interactions, Protein families, Protein expression data, Reactions & pathways, Protein sequences, Diseases	181	120.4	68.9	82.9	1.4E-26
4	SP:Q9BYQ8	Keratin-associated protein 4-9 OS=Homo sapiens OX=9606 GN=KRTAP4-9 PE=2 SV=2 *Cross-references and related information in:* Gene expression, Bioactive molecules, Nucleotide sequences, Genomes & metagenomes, Literature, Samples & ontologies, Protein families, Protein expression data, Reactions & pathways, Protein sequences, Diseases	210	119.7	69.9	83.1	2.5E-26
5	SP:Q9BYQ5	Keratin-associated protein 4-6 OS=Homo sapiens OX=9606 GN=KRTAP4-6 PE=2 SV=4 *Cross-references and related information in:* Gene expression, Bioactive molecules, Nucleotide sequences, Genomes & metagenomes, Literature, Samples & ontologies, Protein families, Protein expression data, Reactions & pathways, Protein sequences, Diseases	205	117.4	64.6	71.2	1.2E-25
6	SP:Q9BYQ6	Keratin-associated protein 4-11 OS=Homo sapiens OX=9606 GN=KRTAP4-11 PE=1 SV=2 *Cross-references and related information in:* Gene expression, Bioactive molecules, Nucleotide sequences	195	117.3	62.6	72.1	1.3E-25

```
>>SP:Q9BQ66 KR412_HUMAN Keratin-associated protein 4-12
OS=Homo sapiens OX=9606 GN=KRTAP4-12 PE=1 SV=1 (201 aa)
 initn: 1297 init1: 1126 opt: 1142  Z-score: 683.6  bits: 133.6 E(566996): 1.6e-30
Smith-Waterman score: 1142; 81.4% identity (89.4% similar) in 161 aa overlap (2-161:40-198)

                                                 10         20         30
NP_001                                  MCSSKMPCSPSA-SSLCAASPPNCCHPSCCQ
                                         :  .  : :.  .:.:    :.::.:::::
SP:Q9B CSDQGCGLENCCRPSCCQTTCCRTTCCRPSCCVSSCCRPQCCQSVCCQ--PTCCRPSCCQ
      10         20         30         40         50          60

               40         50         60         70         80         90
NP_001 TTCCRTTSCSHSCSVSSCCRPQCCHSVCCQPTCCRPSCCQTTCCRTTCCHPSCCVSSCCR
       ::::::: :  :: :::::::::.:::::::::::::::::::::::::.::::::::::
SP:Q9B TTCCRTTCCRPSCCVSSCCRPQCCQSVCCQPTCCRPSCCQTTCCRTTCCRPSCCVSSCCR
        70         80         90        100        110        120

              100        110        120        130        140        150
NP_001 PQCCHSVCFQPTCCHPSCCISSSCCPSCCESSCCCPCCCLRPVCGRVSCHVTCYHPTCVI
       ::::.::: :::::.::::::::::::::::::: ::::::::::::::::.:::.:::::
SP:Q9B PQCCQSVCCQPTCCRPSCCISSSCCPSCCESSCCRPCCCLRPVCGRVSCHTTCYRPTCVI
       130        140        150        160        170        180

              160        170        180        190        200        210
NP_001 STCPHPLCCASPPLPLPFPSPPVPLPFFLSLALPSPPRPSPPLLSPVLIPSPSPSPSLPS
       ::::.::::::
SP:Q9B STCPRPLCCASSCC
       190        200
```

BLAST
(Basic Local Alignment Search Tool)

1. Divides the query sequence in words (*default*, 3 aa)
2. Compares each word with regions of same size in the DB sequences and computes the *score*
3. If the *score* is ≥ a threshold value ***T*** below which the similarity is considered too low, extends the aligned region searching for high similarity regions (score above a second threshold value ***S***), stopping when the score cannot be improved anymore


Query sequence
words
DB
Score ≥ T
?
No
remove
Yes
Identify the sequence
Sequences in the database
Estend the match of each word until
The score remains higher than S

BLAST Algorithm, Step 1

- Given a **word of length *w*** (usually 3 for proteins) & a scoring/substitution matrix (es. BLOSUM62):

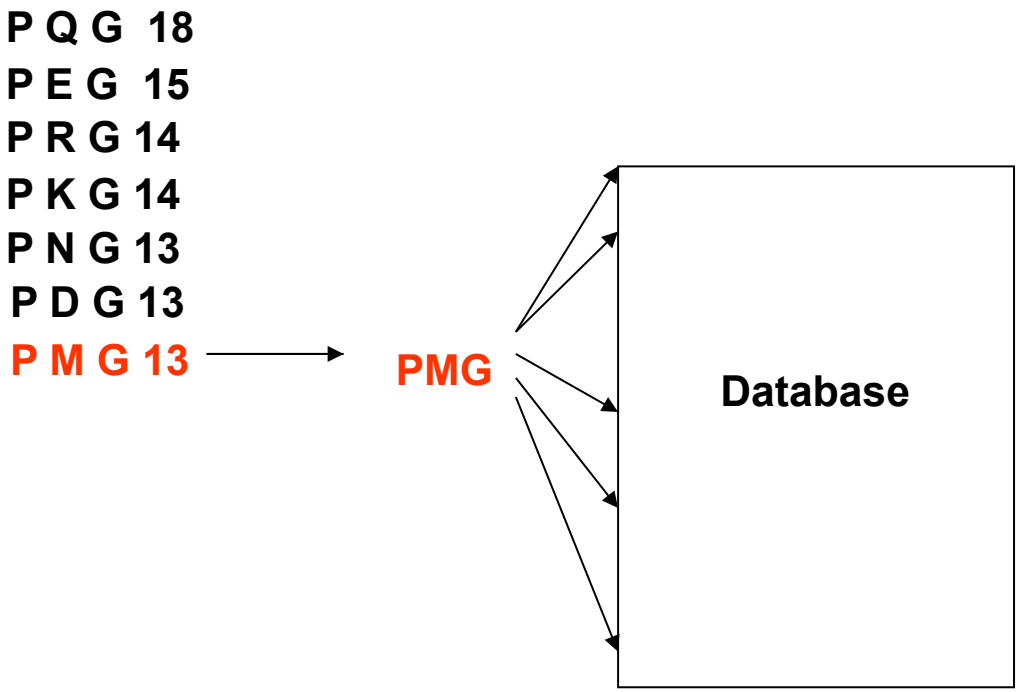
  Create a list of all the words (*w*-letters) that give a **score >*T*** sewhen compared with the query words of length *w*-

Query Sequence **L N K C K T P Q G Q R L V N Q**

	P Q G 18	**Word**
	P E G 15	**Neighborhood Words**
	P R G 14	
	P K G 14	
	P N G 13	
	P D G 13	
	P M G 13	
Below Threshold (T=13)	**P Q A 12**	
	P Q N 12	
	etc.	

BLAST Algorithm, Step 2

- Identifies all the **positions in the database** where there is a word sufficiently similar (hit list).



BLAST Algorithm, Step 3

- The software attempts to **extend** the alignment in both directions adding pairs of residues. Residues are added until the score cannot be improved anymore. It considers only alignments with a score above the threshold value (S).

High-scoring Segment Pair (HSP)

High Scoring Segment Pairs

BLAST & **FASTA** differ in the way they “fish” putative homologs from the DB (similarity/identity).

BLAST & **FASTA** differ in the way they “fish” putative homologs from the DB (similarity/identity)

Furthermore a fundamental difference between **BLAST** & **FASTA** is in the way they compute the distribution of random scores:

FASTA computes it each time a novel query is submitted for the search in a given DB

BLAST uses distributions precomputed on each DB for ensembles of random sequences of standard composition

BLAST e **FASTA** differ in the way they “fish” putative homologs from the DB (similarity/identity)

Furthermore a fundamental difference between **BLAST** & **FASTA** is in the way they compute the distribution of random scores:

FASTA computes it each time a novel query is submitted for the search in a given DB

BLAST uses distributions precomputed on each DB for ensembles of random sequences of standard composition

For this reason BLAST “masks” the regions of query sequence at low complexity

```
Query  635  XXXVGLSHLGVVPPHQRGSPSSXXXXX--XXXQHQRALNYSXXXXXXXXXXXXXXXXXXX  692
            H   GL   G++PP    + ++A AAA  AA+     +   +L A +         G
Sbjct  307  HSVPGLMSPGIIPPTGLTAAAAAAAAATNAAIAEAMKVKKIKLEAMSNYHASNNQHGADS  366

Query  635  hqTVGLSHLGVVPPHQRGSPSSeeaaa--aaQHQRALNYSQlaaaaavangaavgggav  692
            H   GL   G++PP    + ++A AAA  AA+     +   +L A +         G
Sbjct  307  HSVPGLMSPGIIPPTGLTAAAAAAAAATNAAIAEAMKVKKIKLEAMSNYHASNNQHGADS  366
```

old BLAST format (X)

novel BLAST format (small letters)

BLAST

Specialized searches

SmartBLAST

Find proteins highly similar to your query

Primer-BLAST

Design primers specific to your PCR template

Global Align

Compare two sequences across their entire span (Needleman-Wunsch)

CD-search

Find conserved domains in your sequence

IgBLAST

Search immunoglobulins and T cell receptor sequences

VecScreen

Search sequences for vector contamination

CDART

Find sequences with similar conserved domain architecture

Multiple Alignment

Align sequences using domain and protein constraints

MOLE-BLAST

BLAST

NIH U.S. National Library of Medicine NCBI Sign in to NCBI

SMARTBLAST

Home Help

blastp

Enter Protein Query Sequence

Smart Blast searches a protein query against the landmark database

Enter one protein accession, gi, or FASTA sequence Clear

```
>NP_001382996.1 putative keratin-associated protein 4-16 [Homo sapiens]
MCSSKMPCSPSASSLCAASPPNCCHPSCCQTTCCRTTSCSHSCSVSSCCRPQCCHSVCCQPTC
CRPSCCQTTCCRTTCCHPSCCVSSCCRPQCCHSVCFQPTCCHPSCCISSSCCPSCCESSCCCPC
CCLRPVCGRVSCHVTCYHPTCVISTCPHPLCCASPPLPLPFPSPPVPLPFFLSLALPSPPRPSPPLLS
PVLIPSPSPSPSLPS
```

BLAST Show results in a new window

BLAST is a registered trademark of the National Library of Medicine

Support center Mailing list

NCBI

National Center for Biotechnology Information, U.S. National Library of Medicine

8600 Rockville Pike, Bethesda MD, 20894 USA

BLAST

Best hits

Additional BLAST Hits

Select: All None Selected:0

Alignments GenPept

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	putative keratin-associated protein 4-16 [Homo sapiens]	350	350	100%	6e-120	100.00%	NP_001382996.1
☐	KRTAP4-16 isoform 1 [Pan troglodytes]	211	211	93%	2e-65	84.62%	PNI34085.1
☐	keratin-associated protein 4-8 isoform X2 [Pongo abelii]	167	167	67%	1e-48	76.03%	XP_024091030.1
☐	keratin-associated protein 4-9-like isoform X2 [Nomascus leucogenys]	164	164	67%	7e-47	75.50%	XP_012353014.1
☐	keratin-associated protein 4-8 isoform X9 [Pan paniscus]	164	164	66%	1e-46	77.14%	XP_034799243.1
☐	keratin-associated protein 4-11 isoform X9 [Pan troglodytes]	159	159	66%	2e-45	77.14%	XP_024206089.1
☐	keratin-associated protein 4-9-like isoform X4 [Nomascus leucogenys]	157	157	57%	5e-45	82.40%	XP_004091436.1
☐	keratin-associated protein 4-9-like isoform X1 [Nomascus leucogenys]	159	159	67%	5e-45	73.08%	XP_004091435.1
☐	keratin-associated protein 4-11 isoform X6 [Pan paniscus]	157	157	66%	8e-44	69.68%	XP_034799240.1
☐	keratin-associated protein 4-9-like isoform X3 [Nomascus leucogenys]	155	155	67%	1e-43	74.17%	XP_003278340.2
☐	keratin-associated protein 4-9 isoform X3 [Pan paniscus]	155	155	66%	4e-43	68.75%	XP_034799237.1
☐	keratin-associated protein 4-11 isoform X5 [Pan troglodytes]	153	153	66%	1e-42	69.68%	XP_003953071.3
☐	keratin-associated protein 4-11 isoform X10 [Pan paniscus]	152	152	57%	3e-42	79.20%	XP_034799244.1
☐	keratin-associated protein 4-12 isoform X4 [Cavia porcellus]	150	219	67%	6e-42	70.21%	XP_023419679.1

BLAST

GenPept ▼ Next ▲ Previous Descriptions

KRTAP4-16 isoform 1, partial [Pan troglodytes]

Sequence ID: PNI34085.1 **Length:** 228 **Number of Matches:** 1

Range 1: 1 to 208 GenPept ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
211 bits(538)	2e-65()	Compositional matrix adjust.	176/208(85%)	180/208(86%)	12/208(5%)	

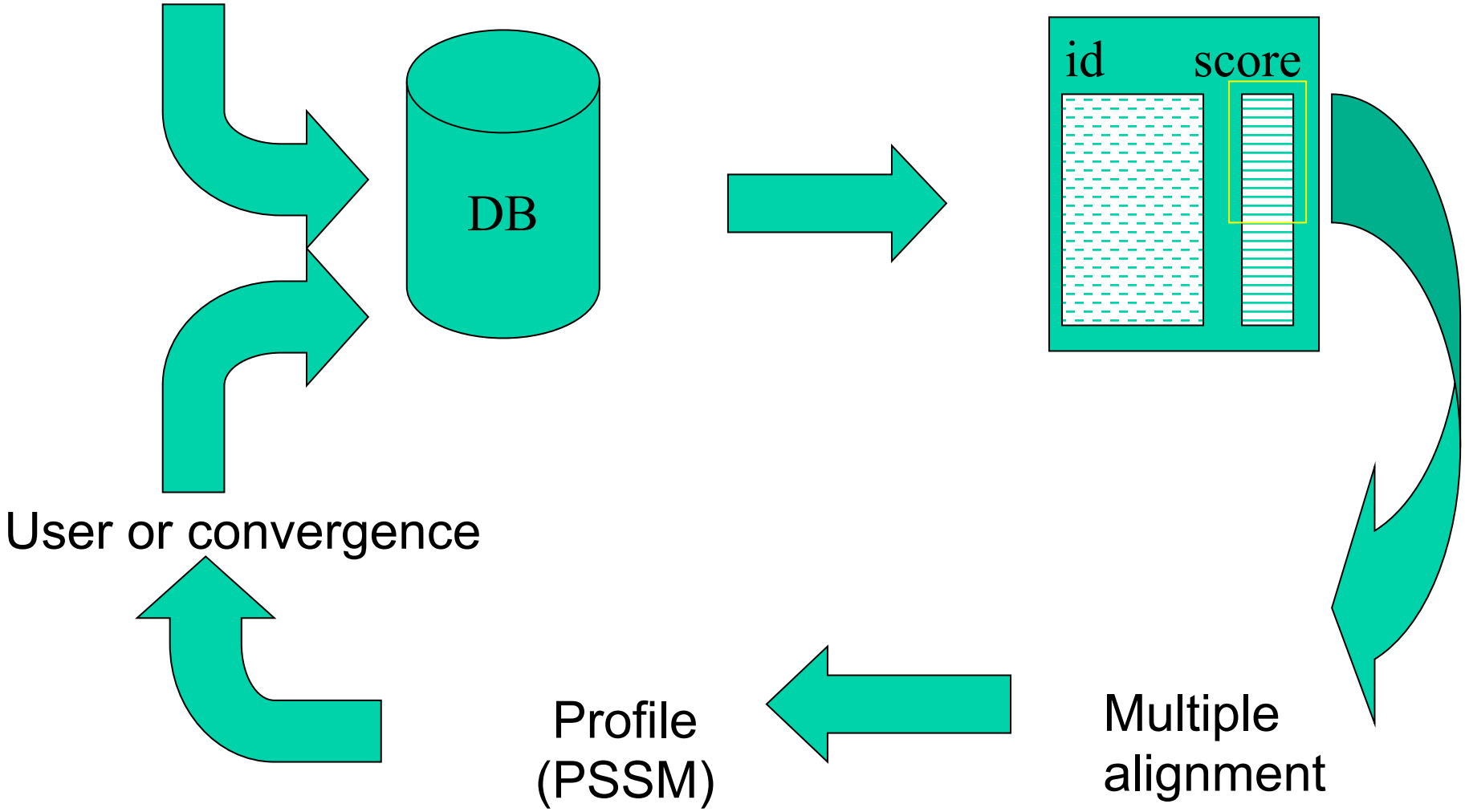
```
Query  4    SKMPCSPSASSLCAASPPNCCHPSCCQTTCCRTTSCSHSCSVSSCCRPQCCHSVCCQPTC  63
            SKMPCSPSASS+CAASPPNCCHPS CQTTCCRTTSC HSCSVSSCCRPQCC SVCCQPTC
Sbjct  1    SKMPCSPSASSVCAASPPNCCHPSSCQTTCCRTTSCPHSCSVSSCCRPQCCQSVCCQPTC  60

Query  64   CRPSCCQTTCCRTTCCHPSCCVSSCCRPQCCHSVCFQPTCCHPS-----CCISSSCCPSC  118
            C PSCCQTTCCR TCCHPSCCVSSCCRPQCCHSVC QPTCC P+     CCISSSCCPSC
Sbjct  61   CHPSCCQTTCCRMTCCHPSCCVSSCCRPQCCHSVCCQPTCCRPTCCRPSCCISSSCCPSC  120

Query  119  CESSCCCPCCCLRPVCGRVSCHVTCYHPTCVISTCPHPLCCA-------SPPLPLPFPSP  171
            CESSCCCP CCLRPVCGRVSCH+TCYHPTCVISTCP PLCCA       SP  PLP PSP
Sbjct  121  CESSCCCPFCCLRPVCGRVSCHITCYHPTCVISTCPRPLCCASPPLPLLSPSPPLPLPSP  180

Query  172  PVPLPFFLSLALPSPPRPSPPLLSPVLI  199
             +PLPFFLSLALPSPP  SPPLLS VLI
Sbjct  181  TLPLPFFLSLALPSPPHTSPPLLSTVLI  208
```

Related Information

NewGenome Data Viewer - aligned genomic context

PSI-blast

(*Position-Specific Iterative blast)*



PSI-blast

(*Position-Specific Iterative blast)*

PSI-BLAST is a variation of BLAST that uses features of a particular protein family to identify related sequences in a protein database

In PSI-BLAST a **profile** or **position-specific scoring matrix** (PSSM) of a set of sequences is constructed from a multiple alignment of the highest-score hits found by the initial BLAST search

In the PSSM a high score is assigned to a highly conserved residue at a certain position while a negative score is assigned to other residues at that position

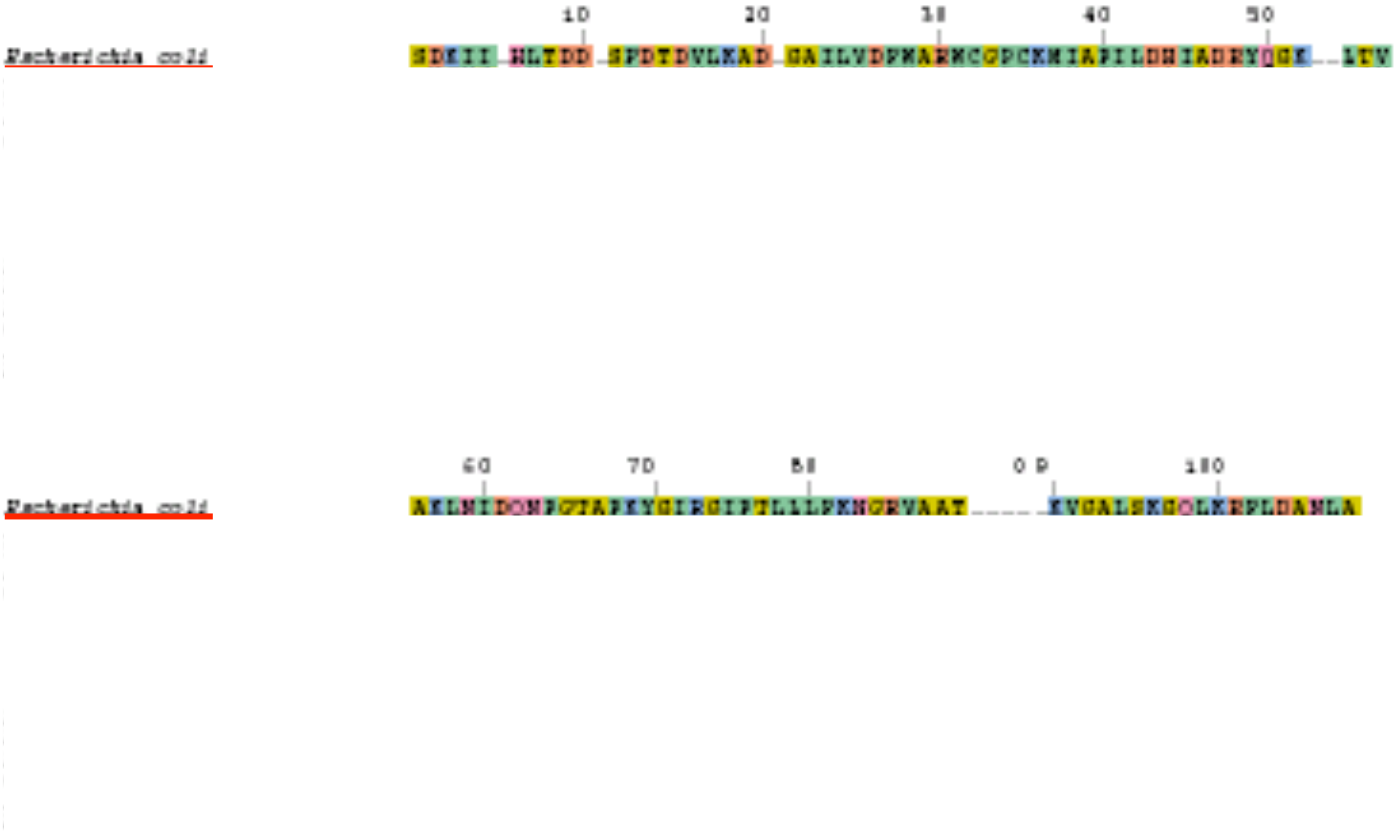
The profile generated is used to replace the substitution matrix in a subsequent BLAST search

Homologs search for *E. coli* thioredoxin

query

PSI-BLAST

1ª iteration

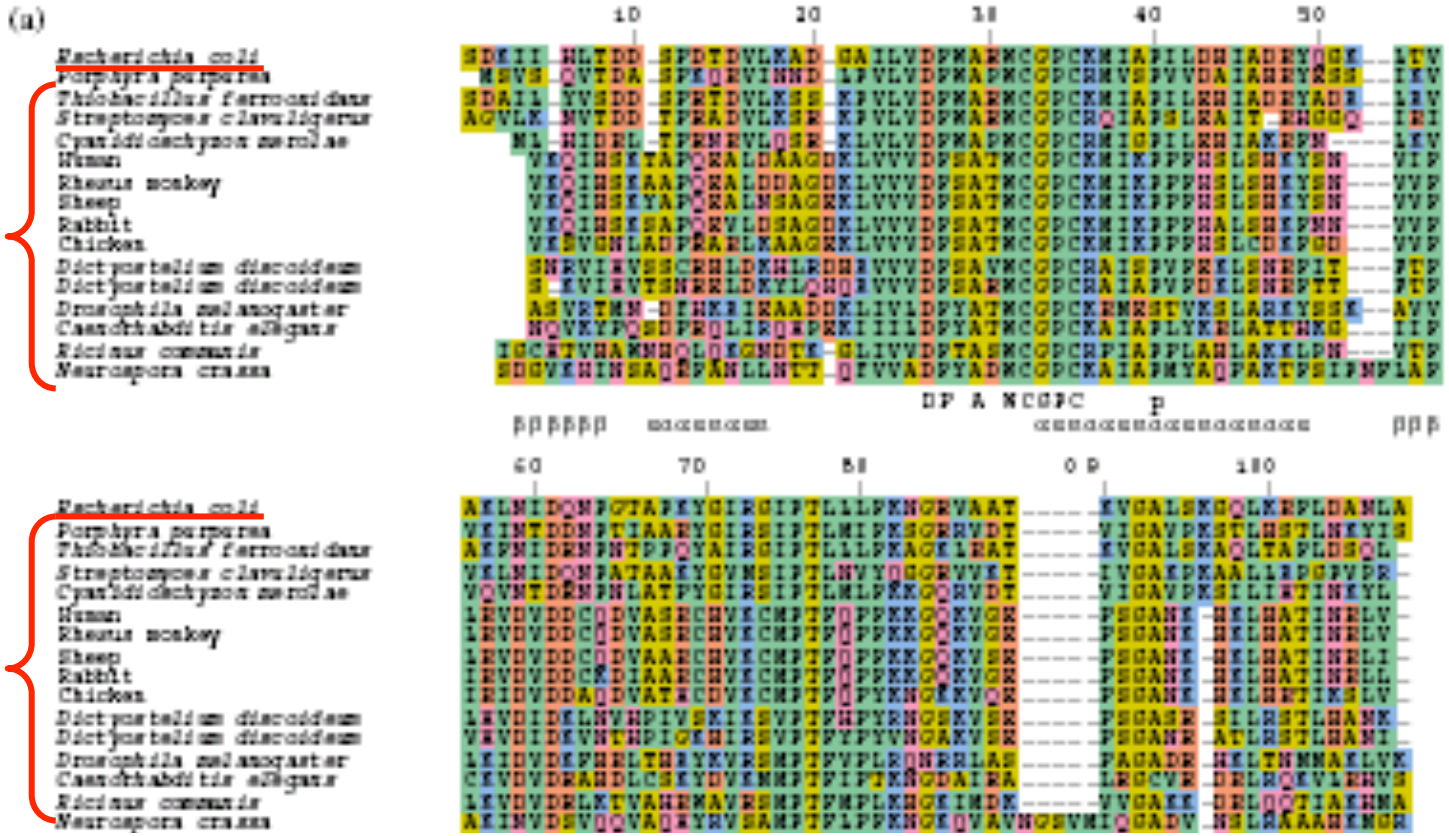


Homologs search for *E. coli* thioredoxin

results

PSI-BLAST

1[a] iteration

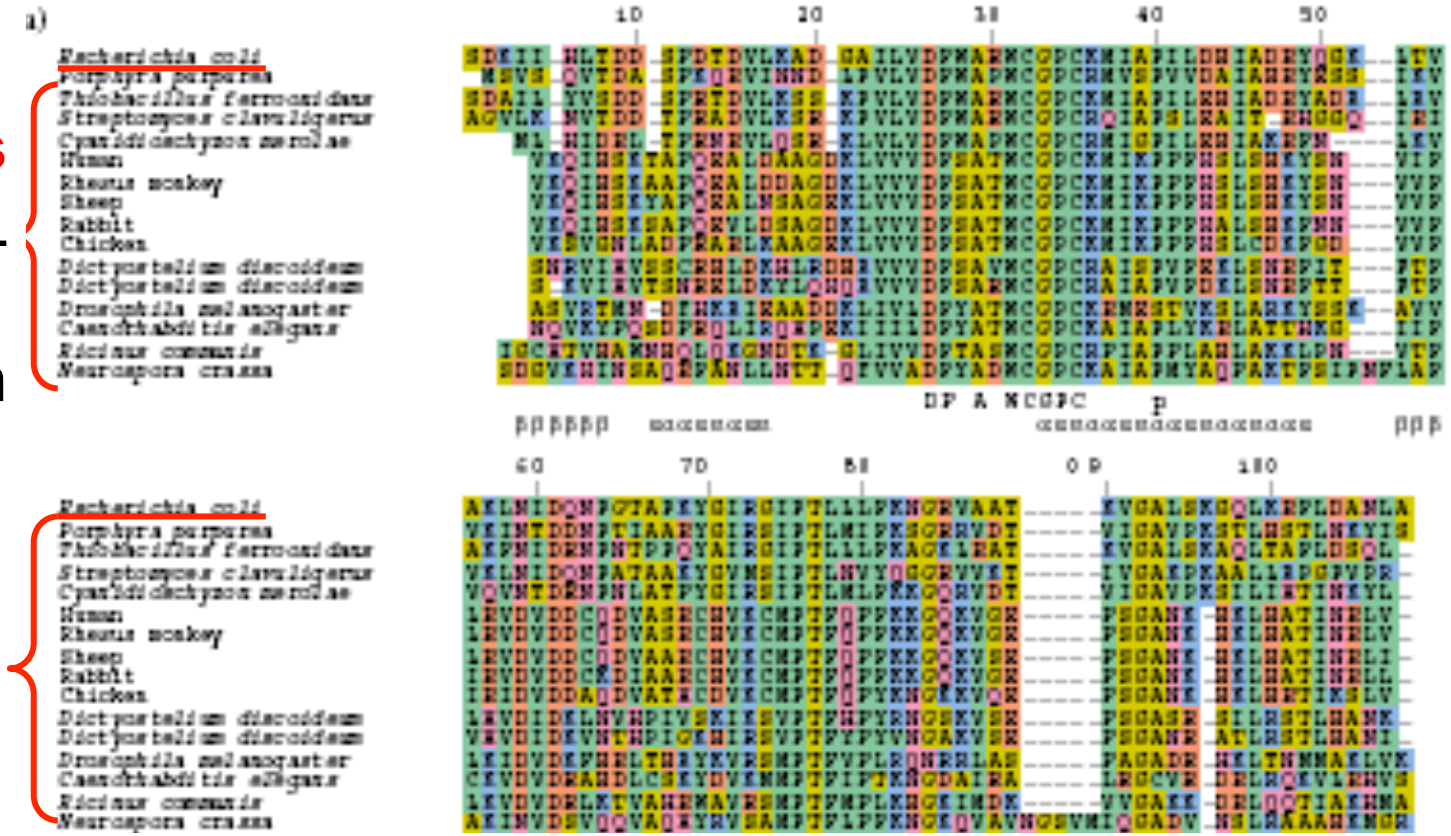


Homologs search for *E. coli* thioredoxin

results

PSI-BLAST

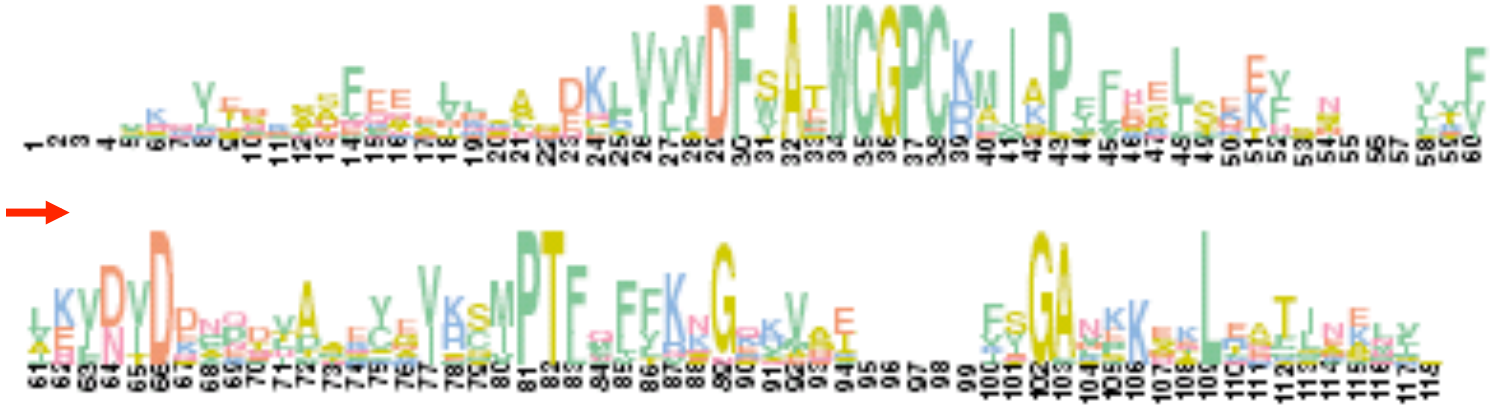
1ª iteration



Profile of results

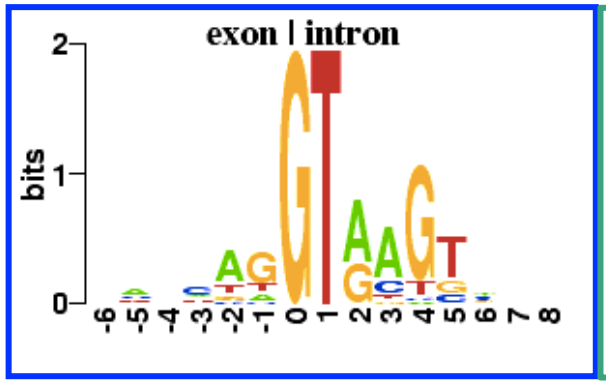
PSI-BLAST

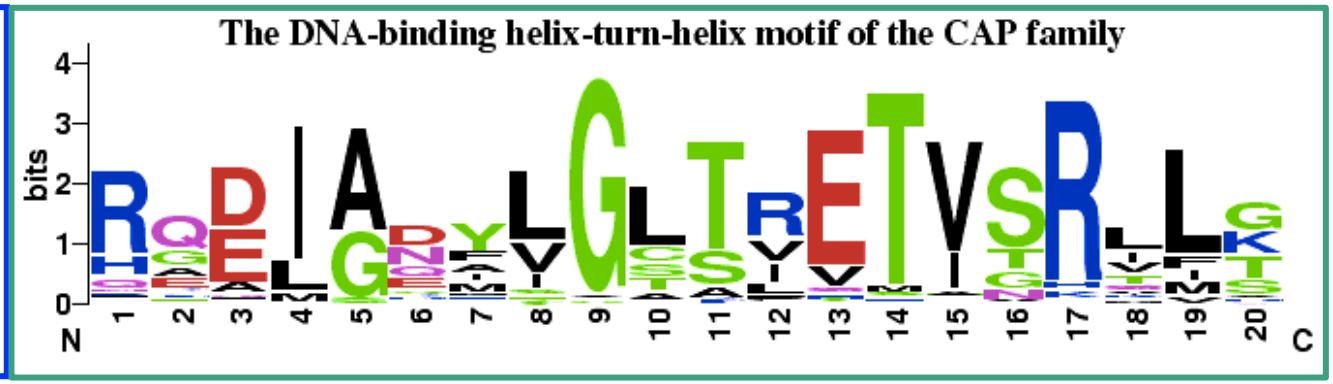
1ª iteration



Sequence logos

Profiles of multiple sequence alignments can be represented graphically in the form of sequence logos, easily showing the residue preference or conservation at particular positions, which point to a functional role





Examples from Web Logo

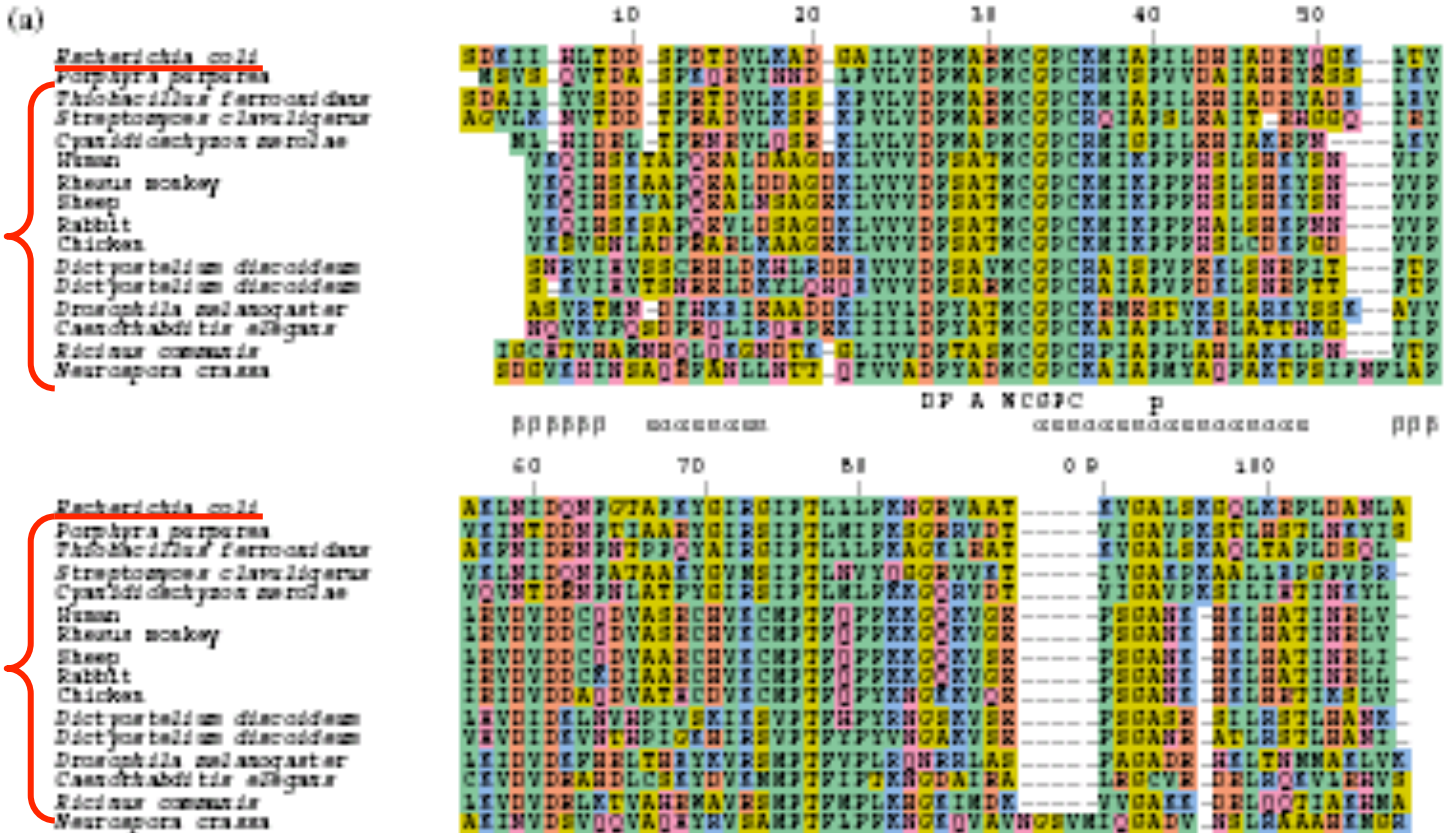
https://weblogo.berkeley.edu/logo.cgi

Homologs search for *E. coli* thioredoxin

results

PSI-BLAST

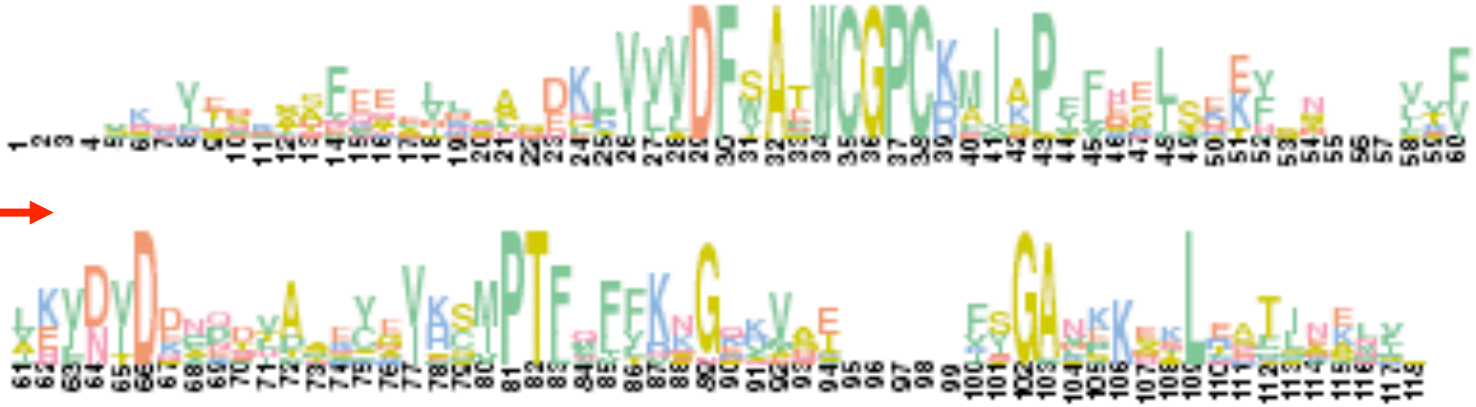
1ª iteration

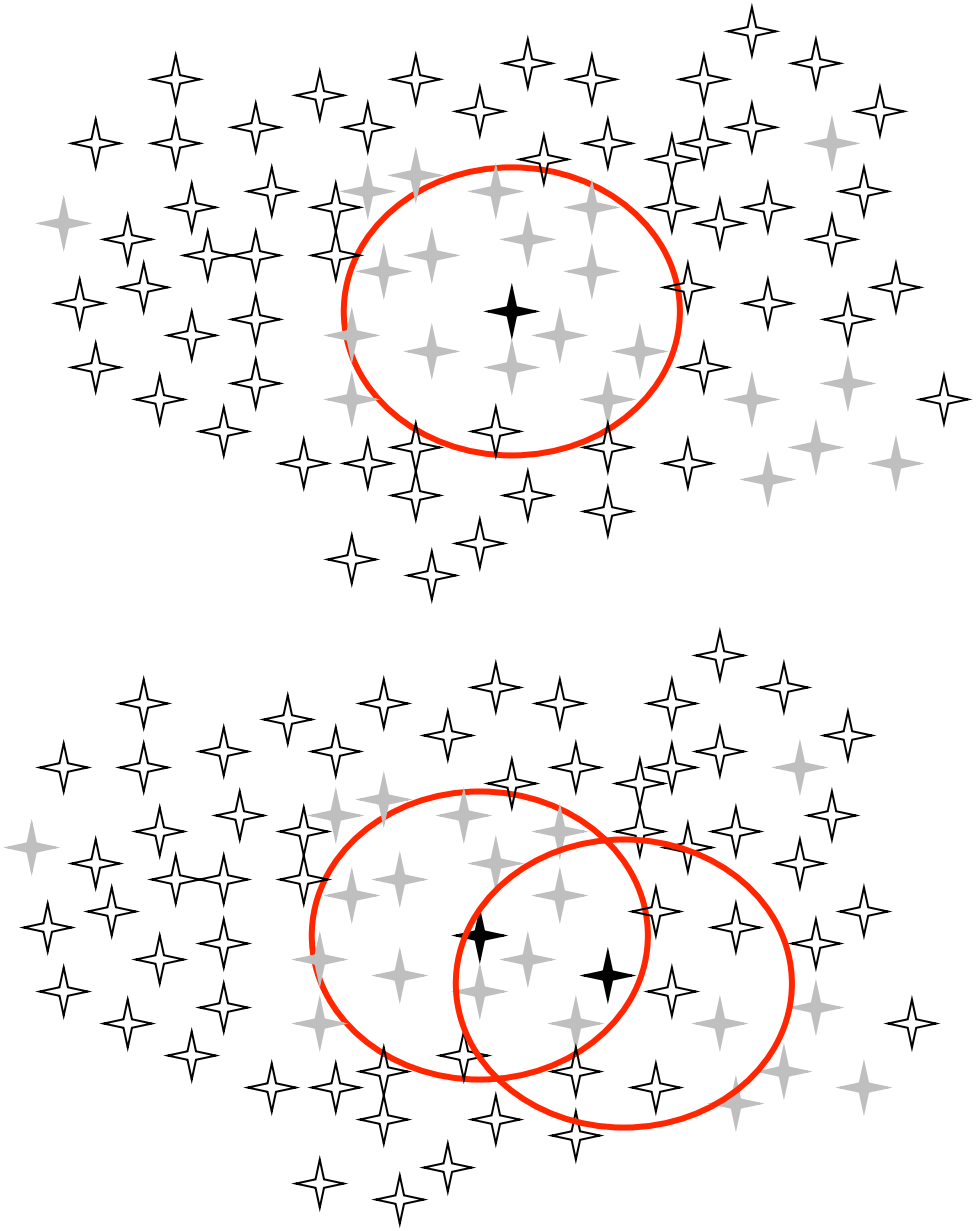


query

PSI-BLAST →

2ª iteration





For sequence identities below 30 % PSI-BLAST allows to correctly identify a three-fold higher number of homologs as as compared to BLAST

Edit View Favorites Tools Help

Search Favorites Media

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

NEW	☑	gi|22477501|gb|AAH37071.1|	Tec protein [Mus musculus] >gi|742...	310	9e-83	G
NEW	☑	gi|6006272|dbj|BAA84765.1|	src-like A-2 [Branchiostoma belcheri]	310	1e-82	
NEW	☑	gi|312891|emb|CAA39196.1|	protein-tyrosine kinase [Mus muscul...	310	1e-82	G
NEW	☑	gi|47226186|emb|CAG08333.1|	unnamed protein product [Tetraodon n	310	1e-82	
NEW	☑	gi|223232|prf||0610239A	protein src	310	1e-82	
NEW	☑	gi|18858911|ref|NP_571179.1|	IL2-inducible T-cell kinase [Dan...	309	2e-82	G
NEW	☑	gi|76660299|ref|XP_880094.1|	PREDICTED: similar to Tyrosine-p...	308	3e-82	G
NEW	☑	gi|6002439|dbj|BAA84738.1|	src-like C [Eptatretus burgeri]	308	4e-82	
NEW	☑	gi|157304|gb|AAA28489.1|	dsrc peptide	308	4e-82	
NEW	☑	gi|76647278|ref|XP_884232.1|	PREDICTED: similar to c-src tyro...	308	5e-82	G
NEW	☑	gi|66803|pir||TVFFS	protein-tyrosine kinase (EC 2.7.1.112) sr...	308	5e-82	
NEW	☑	gi|55627234|ref|XP_518694.1|	PREDICTED: similar to protein-ty...	307	9e-82	G
NEW	☑	gi|55586597|ref|XP_513246.1|	PREDICTED: hypothetical protein XP_	306	2e-81	G
NEW	☑	gi|6002445|dbj|BAA84742.1|	src-like B [Branchiostoma belcheri]	305	3e-81	
NEW	☑	gi|55651650|ref|XP_514571.1|	PREDICTED: hemopoietic cell kinase	305	3e-81	G
NEW	☑	gi|157176|gb|AAA28443.1|	dash peptide	305	4e-81	
NEW	☑	gi|72065446|ref|XP_795344.1|	PREDICTED: similar to c-src tyro...	303	1e-80	G
NEW	☑	gi|76625623|ref|XP_881439.1|	PREDICTED: similar to fyn proto-onc	302	2e-80	G
NEW	☑	gi|10835732|pdb|1FPU|B	Chain B, Crystal Structure Of Abl Kina...	302	2e-80	S
NEW	☑	gi|6006273|dbj|BAA84741.1|	src-like A-1 [Branchiostoma belcheri]	302	3e-80	
NEW	☑	gi|74000753|ref|XP_544774.2|	PREDICTED: similar to c-src tyro...	298	3e-79	G
NEW	☑	gi|88192844|pdb|2F4J|A	Chain A, Structure Of The Kinase Domai...	298	4e-79	S
NEW	☑	gi|5453034|gb|AAD43407.1|	protein tyrosine kinase TecIIA [Mus mu	296	1e-78	G
NEW	☑	gi|74217137|dbj|BAE43394.1|	unnamed protein product [Mus musculu	295	3e-78	G
NEW	☑	gi|6002441|dbj|BAA84739.1|	src-like A [Lethenteron reissneri]	294	7e-78	
NEW	☑	gi|47207768|emb|CAF90506.1|	unnamed protein product [Tetraodon n	293	9e-78	
NEW	☑	gi|74000751|ref|XP_867112.1|	PREDICTED: similar to c-src tyro...	293	1e-77	G
NEW	☑	gi|16197923|gb|AAL13726.1|	LD03455p [Drosophila melanogaster]	292	3e-77	

Run PSI-Blast iteration 2

Internet

Edit View Favorites Tools Help

Back Search Favorites Media

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

		gi|18150824|dbj|BAA81712.3| protein tyrosine kinase [Ephydatia f	534	4e-150	
		gi|55586597|ref|XP_513246.1| PREDICTED: hypothetical protein XP_	532	2e-149	G
NEW		gi|76668168|ref|XP_610012.2| PREDICTED: similar to Cytoplasmi...	524	3e-147	G
NEW		gi|74007139|ref|XP_548870.2| PREDICTED: similar to Cytoplasmi...	523	5e-147	G
		gi|55651650|ref|XP_514571.1| PREDICTED: hemopoietic cell kinase	520	4e-146	G
NEW		gi|16741712|gb|AAH16652.1| BMX non-receptor tyrosine kinase [...	520	6e-146	G
NEW		gi|33303803|gb|AAQ02415.1| BMX non-receptor tyrosine kinase [syn	518	2e-145	
		gi|8569366|pdb|1QPJ|A Chain A, Crystal Structure Of The Lymph...	516	1e-144	S
NEW		gi|6753196|ref|NP_033889.1| BMX non-receptor tyrosine kinase ...	515	2e-144	G
		gi|2781151|pdb|3LCK| The Kinase Domain Of Human Lymphocyte K...	514	3e-144	S
NEW		gi|3002963|gb|AAC08966.1| Etk/Bmx cytosolic tyrosine kinase [Hom	513	5e-144	G
		gi|55627234|ref|XP_518694.1| PREDICTED: similar to protein-ty...	513	8e-144	G
NEW		gi|62666805|ref|XP_346303.2| PREDICTED: similar to protein ty...	510	6e-143	G
		gi|47224264|emb|CAG09110.1| unnamed protein product [Tetraodon n	510	7e-143	
		gi|8569441|pdb|1QPD|A Chain A, Structural Analysis Of The Lym...	510	8e-143	S
		gi|72065446|ref|XP_795344.1| PREDICTED: similar to c-src tyro...	509	1e-142	G
		gi|1334146|emb|CAA27235.1| unnamed protein product [Mus musculus	505	3e-141	
		gi|10436623|dbj|BAB14871.1| unnamed protein product [Homo sap...	503	7e-141	G
NEW		gi|72042967|ref|XP_780086.1| PREDICTED: similar to CG8049-PB,...	502	1e-140	G
NEW		gi|76647272|ref|XP_884145.1| PREDICTED: similar to c-src tyro...	497	4e-139	G
		gi|6002443|dbj|BAA84740.1| src-like B [Lethenteron reissneri]	496	8e-139	
		gi|223232|prf||0610239A protein src	492	1e-137	
NEW		gi|26332200|dbj|BAC29830.1| unnamed protein product [Mus musculu	491	2e-137	G
		gi|6002437|dbj|BAA84737.1| src-like B [Eptatretus burgeri]	487	6e-136	
NEW		gi|76057623|emb|CAJ19682.1| dominant-negative kinase-deficien...	483	6e-135	G
NEW		gi|775208|gb|AAC50287.1| p56lck >gi|1585504|prf||2201317A protei	481	3e-134	G
		gi|6006272|dbj|BAA84765.1| src-like A-2 [Branchiostoma belcheri]	475	2e-132	
		gi|157176|gb|AAA28443.1| dash peptide	475	2e-132	

Run PSI-Blast iteration 3

Done Internet

Edit View Favorites Tools Help

Back Search Favorites Media

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

		gi|55586597|ref|XP_513246.1|	PREDICTED: hypothetical protein XP_	532	2e-149	G
NEW		gi|76668168|ref|XP_610012.2|	PREDICTED: similar to Cytoplasmi...	524	3e-147	G
NEW		gi|74007139|ref|XP_548870.2|	PREDICTED: similar to Cytoplasmi...	523	5e-147	G
		gi|55651650|ref|XP_514571.1|	PREDICTED: hemopoietic cell kinase	520	4e-146	G
NEW		gi|16741712|gb|AAH16652.1|	BMX non-receptor tyrosine kinase [...	520	6e-146	G
NEW		gi|33303803|gb|AAQ02415.1|	BMX non-receptor tyrosine kinase [syn	518	2e-145	
		gi|8569366|pdb|1QPJ|A	Chain A, Crystal Structure Of The Lymph...	516	1e-144	S
NEW		gi|6753196|ref|NP_033889.1|	BMX non-receptor tyrosine kinase ...	515	2e-144	G
		gi|2781151|pdb|3LCK|	The Kinase Domain Of Human Lymphocyte K...	514	3e-144	S
NEW		gi|3002963|gb|AAC08966.1|	Etk/Bmx cytosolic tyrosine kinase [Hom	513	5e-144	G
		gi|55627234|ref|XP_518694.1|	PREDICTED: similar to protein-ty...	513	8e-144	G
NEW		gi|62666805|ref|XP_346303.2|	PREDICTED: similar to protein ty...	510	6e-143	G
		64|emb|CAG09110.1|	unnamed protein product [Tetraodon n	510	7e-143	
		1|pdb|1QPD|A	Chain A, Structural Analysis Of The Lym...	510	8e-143	S
		46|ref|XP_795344.1|	PREDICTED: similar to c-src tyro...	509	1e-142	G
		5|emb|CAA27235.1|	unnamed protein product [Mus musculus	505	3e-141	
		gi|10436623|dbj|BAB14871.1|	unnamed protein product [Homo sap...	503	7e-141	G
NEW		gi|72042967|ref|XP_780086.1|	PREDICTED: similar to CG8049-PB,...	502	1e-140	G
NEW		gi|76647272|ref|XP_884145.1|	PREDICTED: similar to c-src tyro...	497	4e-139	G
		gi|6002443|dbj|BAA84740.1|	src-like B [Lethenteron reissneri]	496	8e-139	
		|prf||0610239A	protein src	492	1e-137	
		00|dbj|BAC29830.1|	unnamed protein product [Mus musculu	491	2e-137	G
		7|dbj|BAA84737.1|	src-like B [Eptatretus burgeri]	487	6e-136	
		23|emb|CAJ19682.1|	dominant-negative kinase-deficien...	483	6e-135	G
NEW		gi|775208|gb|AAC50287.1|	p56lck >gi|1585504|prf||2201317A protei	481	3e-134	G
		gi|6006272|dbj|BAA84765.1|	src-like A-2 [Branchiostoma belcheri]	475	2e-132	
		gi|157176|gb|AAA28443.1|	dash peptide	475	2e-132	

Result 2[a] iteration

Result 1[a] iteration

Run PSI-Blast iteration 3

Done Internet

Are homologs found for all the protein sequences?

Unique sequences, i.e. sequences with no significant match in homologs searches (BLAST hit with E-value $>10^{-3}$ or $> 10^{-5}$ for alignments of < 80 residues) are referred to as orphan ORFs or ORFans and, in particular, singleton ORFans

The percentage of singleton ORFans in each newly sequenced genome can be as high as 60%

In addition to these unique ORFans, a large fraction of ORFs in each genome has homologs only in the same genome or in closely related genomes. These ORFs are referred to as paralogous and orthologous ORFans, respectively

Lessons 3 & 4.
Contents

1. Introduction to proteins. Different sequences correspond to different 3D structures. Specific structures determine specific functions.

2. Sequence alignment. We search for those that best reflect the evolutionary path. Based on sequence similarity we can infer homology (an evolutionary relationship)

3. Substitutions & gaps. Substitution matrices allow to assign a score for the correspondence of different amino acids. It is necessary to penalize insertions and deletions (INDELs).

4. Homology search in databases. BLAST & FASTA identify a subset of sequences from the databanks, align them to the query and compare the obtained score to a distribution of random scores.