

# Protein sequences

Alignment

Variation

# Challenges in Understanding Genetic Information



- Genetic information is redundant
- Structural information is redundant
- Genes and proteins are meta-stable
- Single genes have multiple functions
- Genes are one dimensional but function depends on three-dimensional structure

# Inferring Biological Function from Protein Sequence

Consensus Sequences  
or Sequence Motifs

Zinc Finger (C2H2 type)  
 $C \times \{2,4\}$   $C \times \{12\}$   $H \times \{3,5\}$   $H$

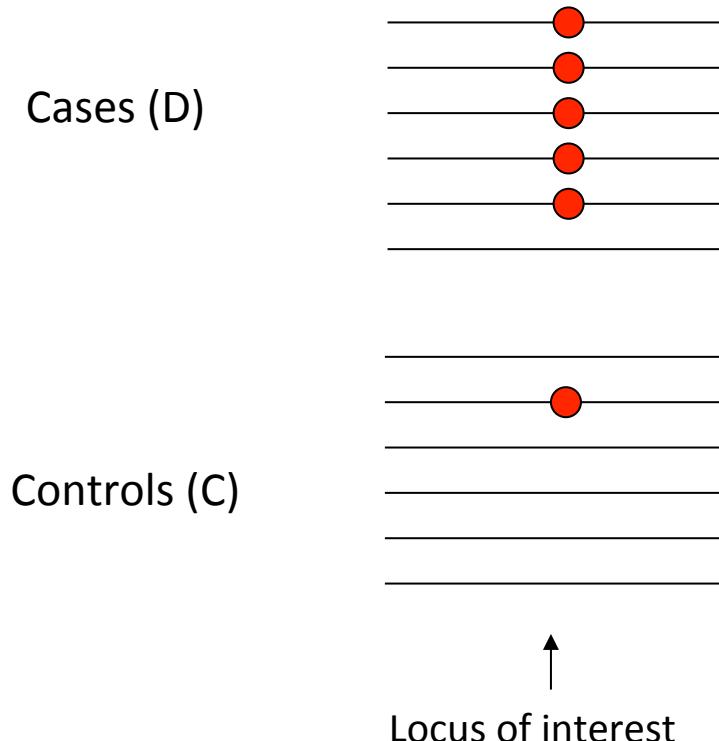
Sequences of Common  
Structure or Function

Sequence Similarity

	10	20	30	40	50	
Query	VLSPADKTNVKAAGWGKVGAHAGEVGAEALERMFLSFPTTKTYFPHF-----DLSHGS					
	:  :  :    :        :      : : :  :  :        :					
Match	HLTPEEKSAVTALWGKV--NVDEYGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN	10	20	30	40	50

# A Bioinformatics problem: How small is my P-value?

- The basic idea of association studies is to look for genetic differences between groups



It is easy to ask the question  
*“Is there a significant difference in the frequency of a mutation between groups?”*

# What we really want to ask

- “*Does any of the genome show an association with disease over and above any effect I might expect from the correlation between genotype and environmental risk?*”
- “*If so, what is the most likely position for the causal mutation(s)?*”

# Sequence alignment

# Sequence Alignment

- Sequence analysis is the process of making biological inferences from the known sequence of monomers in protein, DNA and RNA polymers.
- It is the task of locating equivalent regions of two or more sequences to maximize their similarity

## **Pairwise sequence alignment is the most fundamental operation of bioinformatics**

- Comparing DNA/protein sequences for
  - Similarity
  - Homology
- Prediction of function
- Construction of phylogeny
- Shotgun assembly
  - End-space-free alignment / overlap alignment
- Finding motifs
- Identifying shared domains

# Understanding evolutionary relationships

molecular

molecular

Nothing in biology makes sense except in the light of evolution



Dobzhansky, 1973

# Alignment goal:

- Align two sequences:

THISSEQUENCE

THISISASEQUENCE

THIS---SEQUENCE

THISISASEQUENCE

Find optimal superposition of the two sequences

# Alignment: Global vs Local

Compare:

GGQLAKEEAL

EGQPVEVLP

Local: Find best matches

GGQLAKEEAL

EGQ.PVEVL

Global – all residues aligned

GGQLAKEEAL.

EGQ..PVEVLP

# Some terminology

## **Homolog**

- \* A gene related to a second gene by descent from a common ancestral DNA sequence.  
The term, homolog, may apply to the relationship between genes separated by the event of speciation (see ortholog) or to the relationship between genes separated by the event of genetic duplication (see paralog).

## **Ortholog**

- \* Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. (See also Paralogs.).

## **Speciation**

- \* Speciation is the origin of a new species capable of making a living in a new way from the species from which it arose. As part of this process it has also acquired some barrier to genetic exchange with the parent species.

## **Paralog**

- \* Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

## **Pairwise alignment: protein sequences can be more informative than DNA**

- protein is more informative (20 vs 4 characters); many amino acids share related biophysical properties
- codons are degenerate: changes in the third position often do not alter the amino acid that is specified
- protein sequences offer a longer “look-back” time
- DNA sequences can be translated into protein, and then used in pairwise alignments

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

## Pairwise alignment: protein sequences can be more informative than DNA

---

- DNA can be translated into six potential proteins

5' CAT CAA  
5' ATC AAC  
5' TCA ACT



5' CATCAACTACAACTCCAAAGACACCCTTACACATCAACAAACCTACCCAC 3'  
3' GTAGTTGATGTTGAGGTTCTGTGGGAATGTGTAGTTGGATGGGTG 5'



5' GTG GGT  
5' TGG GTA  
5' GGG TAG

## Pairwise alignment: protein sequences can be more informative than DNA

- Many times, DNA alignments are appropriate
  - to confirm the identity of a cDNA
  - to study noncoding regions of DNA
  - to study DNA polymorphisms
  - example: Neanderthal vs modern human DNA

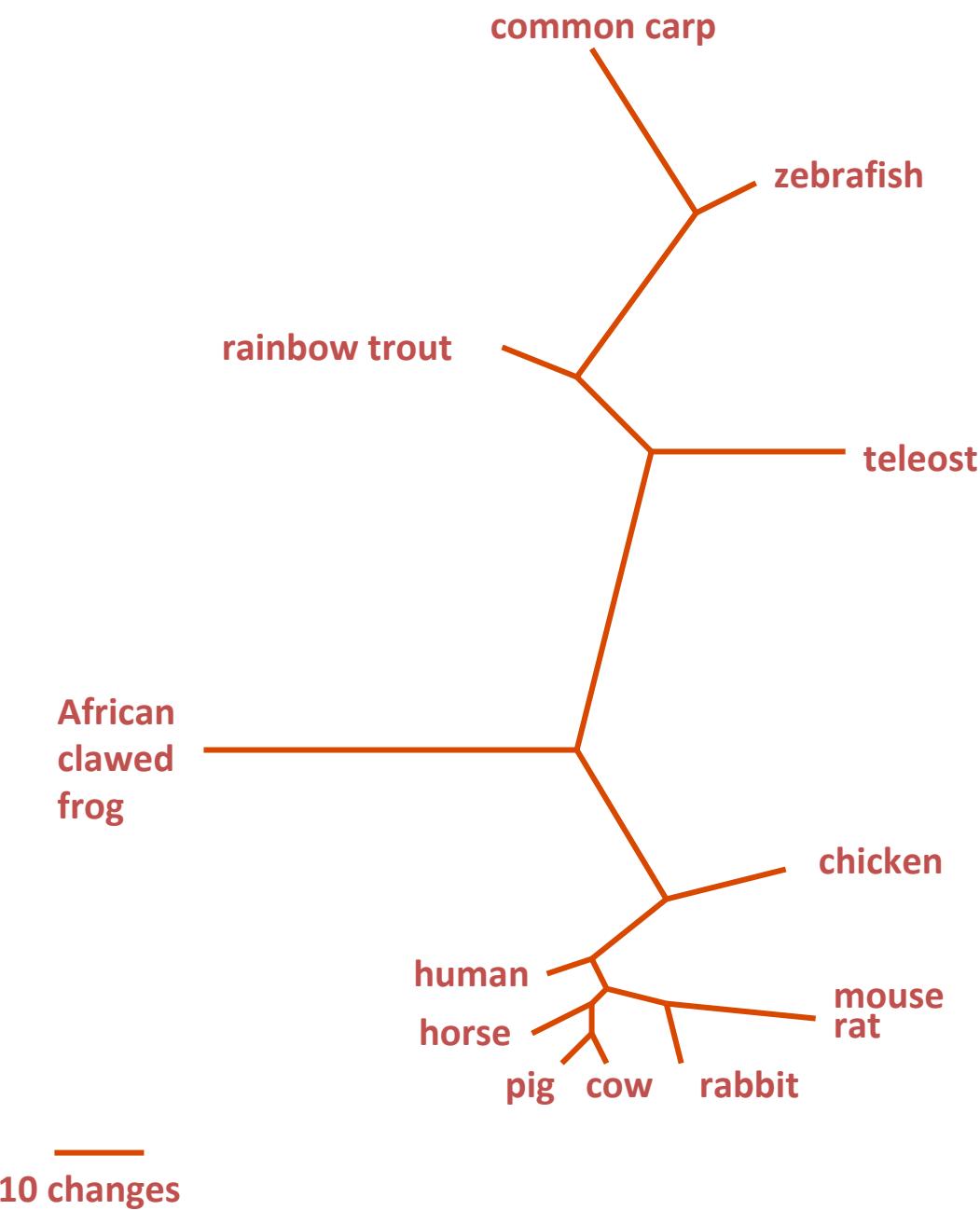
Query: 181 catcaactacaactccaaaggacacccttacaccccacttaggatatcaacaaacctacccac 240  
          ||||| ||||| ||||| ||||| ||||| | ||||||| ||||| ||||| ||||| ||||| |||||

Sbjct: 189 catcaactgcaccccaaaggccaccct-cacccacttaggatatcaacaaacctacccac 247

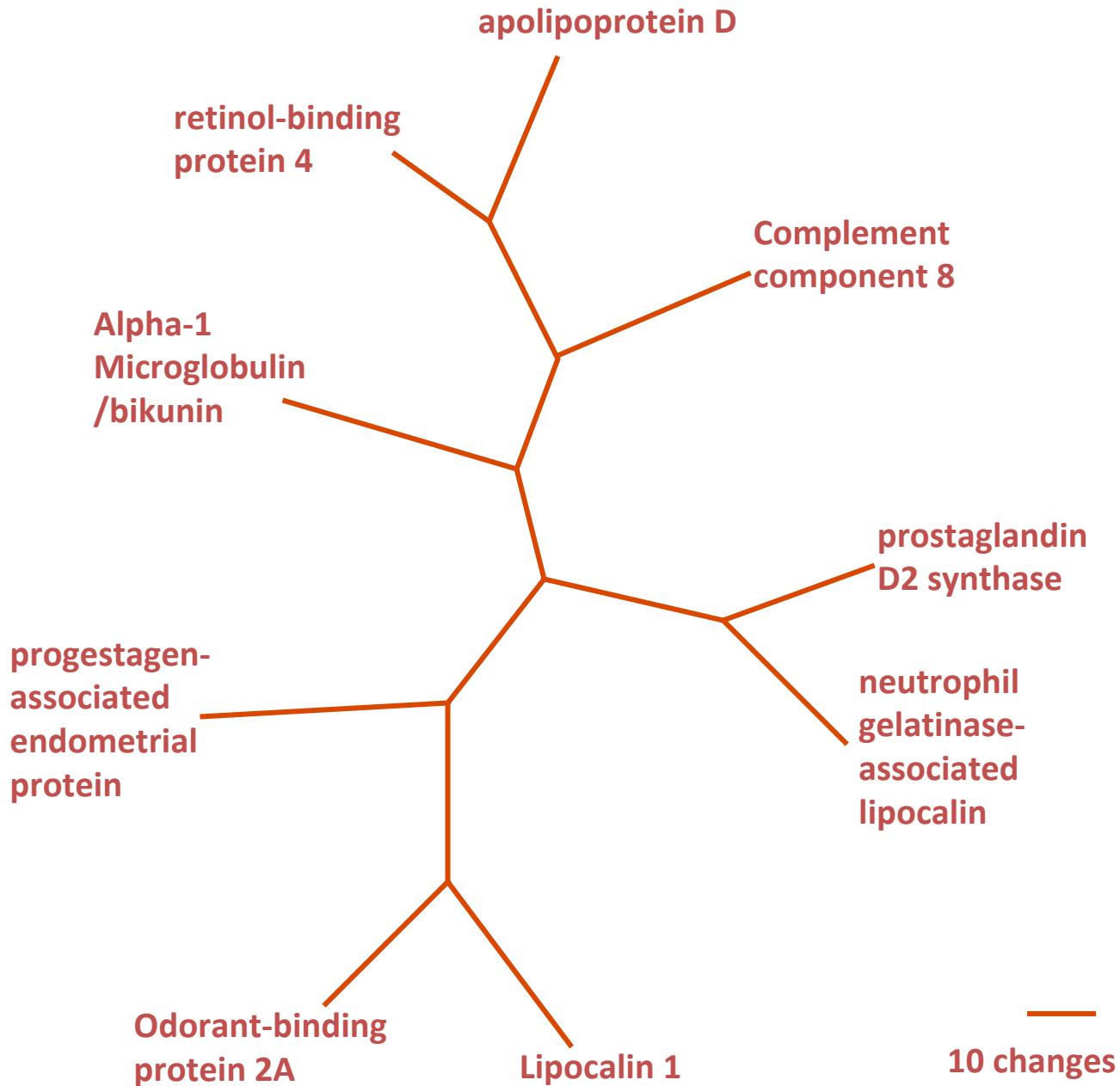
# **Definitions**

## **Homology**

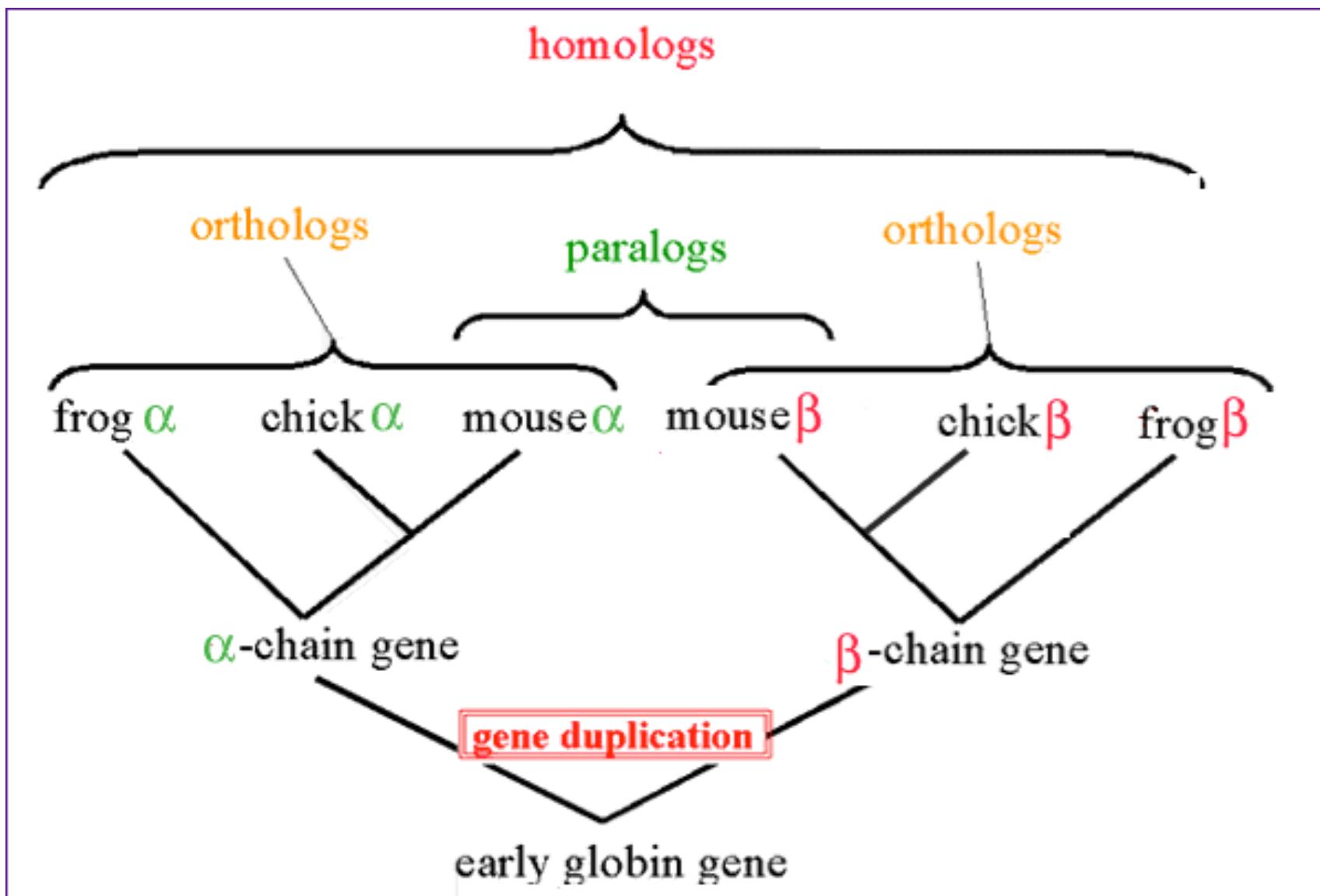
Similarity attributed to descent from a common ancestor.



**Orthologs:**  
members of a  
gene (protein)  
family in various  
organisms.  
**This tree shows**  
**RBP orthologs.**



**Paralogs:**  
members of a  
gene (protein)  
family within a  
species



# Alignment can reveal homology between sequences

- Orthologs
  - Divergence follows speciation
  - Similarity can be used to construct phylogeny between species
- Paralogs
  - Divergence follows duplication
- Xenologs
- ISMB tutorial on protein sequence comparison:  
<http://people.virginia.edu/~wfp/papers/ismb2000.pdf>

# Sequence Alignment

Procedure of comparing two (pairwise) or more (multiple) sequences by searching for a series of individual characters that are in the same order in the sequences

GCTAGTCAGATCTGACGCTA

| | | | | | | | | | |

TGGTCACATCTGCCGC

# Definitions

## Homology

Similarity attributed to descent from a common ancestor.

## Identity

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

RBP:            26    RV**KENFDKARFSGTWYAMA**KKDPEGLF**L**QDNIV**A** 59  
                  + **K++** + **++ GTW++MA** + **L** + **A**

glycodelin: 23    QT**KQDLELPKLA****GTWHSMA**MA-TNNIS**L**MATLK**A** 55

# **Definitions**

## **Similarity**

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

## **Identity**

The extent to which two sequences are invariant.

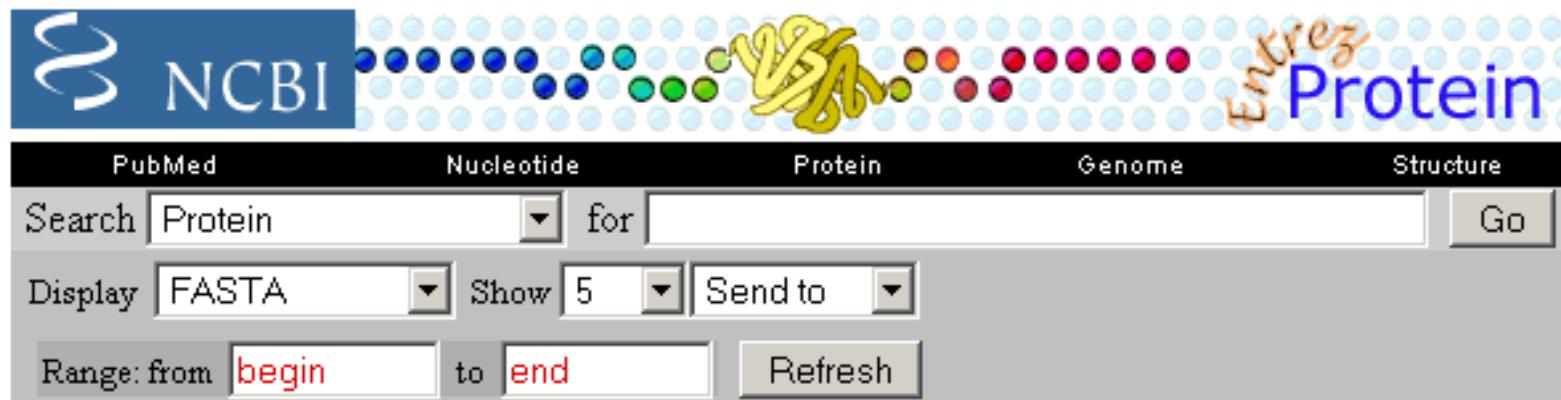
## **Conservation**

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

# Sequence alignments

[http://statgen.ncsu.edu/slse/  
animations/module1.html](http://statgen.ncsu.edu/slse/animations/module1.html)

**FASTA format:**  
**versatile, compact with**  
**>one header line followed by a string of nucleotides**  
**or amino acids in the single letter code**



1: [NP\\_000509](#). Reports beta globin [Homo...[gi:4504349]

```
>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDVGGEALGRLLVVYPWTQRFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

# Sequence Alignment

AGGCTATCACCTGACCTCCAGGCCGATGCC  
TAGCTATCACGACCGCGGTCGATTGCCCGAC

-**AGGCTATCACCTGACCTCCA**GGCCGA--TGCCC---  
**TAG-CTATCAC--GACC**GC--GGTCGA**TTTGCCC**GAC

## Definition

Given two strings     $x = x_1x_2\dots x_M$ ,     $y = y_1y_2\dots y_N$ ,

an alignment is an assignment of gaps to positions  
0, ..., M in x, and 0, ..., N in y, so as to line up each letter in one  
sequence with either a letter, or a gap  
in the other sequence

## Pairwise alignment of retinol-binding protein and $\beta$ -lactoglobulin

1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGT WYAMAKKDPEG 50 RBP	.			.	.	.	.		:	.	.	:			.	:
1 ...MKCLLLALALTCGAQALIVT..QTMKG LDIQKVAGTWYSLAMAASD. 44 lactoglobulin																
51 LFLQDNIVAEFSVDET GQMSATAKGRVR .LLNNWD .ADMVGTFTDTE 97 RBP	:					:		.	.		:					.
45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENCAQKKIIIAEKTK 93 lactoglobulin																
98 DPAKF KMKYWGVASFLQKG NDDHWIVDTDYDTYAV. .... QYSC 136 RBP		.		.				.	.		.					.
94 I PAVFKIDALNENKVL. .... VLDTDYKKYLLFC ENSAEPEQSLAC 135 lactoglobulin																
137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIV RQYRLIV 185 RBP	.				:		.									.
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNTQEEQCN. .... 178 lactoglobulin																

A red arrow points upwards from the bottom of the sequence alignment towards the top. A red box labeled "Identity (bar)" is positioned over the last two rows of the alignment, highlighting the conserved regions between the two proteins.

# Pairwise alignment of retinol-binding protein and $\beta$ -lactoglobulin

1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP	.			.	.	..	.		:	.   .	:	:	
1 ...MKCLLLALALTCGAQALIVT...QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin													
51 LFLQDNIVAEFSVDETMSATAKGRVR.LLNNWD..VCADMVGTFT 97 RBP	:					:		.	.		:		
45 ISLLDAQSAPLRV.YVILKPTPEGDLEILLQKWENGECAQKKIIAEK 93 lactoglobulin													
98 DPAKFKMKYWGVASFLQGNDDHWIVDTDYDTYAV.....Q 136 RBP		.				..			.				
94 I PAVFKIDALNENKVL.....VLDTDYKKYLLFCMENSAEPEOSC 135 lactoglobulin													
137 RLLNLDGTCA PPEAQKIVRQRQ.EELC 137 RBP	.					.							
136 QCLVRTPEVD PMHIRLSFNPTQLEEQC 136 lactoglobulin													

**Somewhat similar (one dot)**

**Very similar (two dots)**

# **Definitions**

## **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.

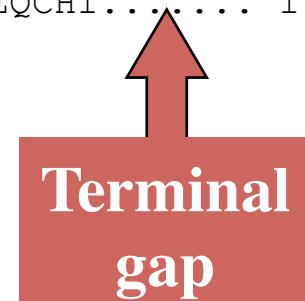
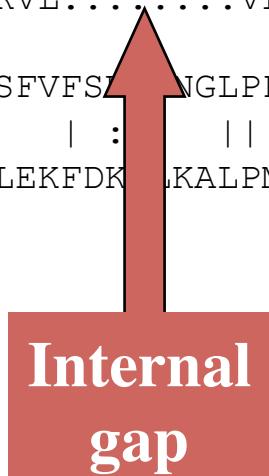
# Pairwise alignment of retinol-binding protein and $\beta$ -lactoglobulin

```
1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTwyAMAKKDPEG 50 RBP
      . | | | | . . . | : . || | . : | : :
1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
  : | | | | | :: | . | . || | : || | | . .
45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

98 DPAKFKMKYWGVASFLQKGNDHWIVDTDYDTYAV.....QYSC 136 RBP
  || | | . | | : . || | | | . . | .
94 I PAVFKIDALNENKVL.....VLDTDYKKYLLFCMENSAEPEQLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSIANGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
  . | | | | : | | | . | | | | |
136 QCLVRTPEVDDEALEKFDKAKALPMHIRLSFNPTQLEEQCHI..... 178 lactoglobulin
```



# Gaps

- 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.

# Pairwise alignment of retinol-binding protein and $\beta$ -lactoglobulin

```
1 MKWVWALLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
      . ||| | . . | : .|||.::|   :
1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

51 LFLQDNIVAEFSVDETGQMSATAKGRVR.LLNNWD..VCADMVGTFTDTE 97 RBP
    : | | | | | :: | . | . || | : || | | .
45 ISLLDAQSAPLRV.YVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTK 93 lactoglobulin

98 DPAFKMKYWGVASFLQKGNDHWIVDTDYDTYAV.....QYSC 136 RBP
    || ||. | :. || | | | . . | .
94 IPAVFKIDALNENKVL.....VLDTDYKKYLLFCMENSAEPEQLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . | | : | | . | || |
136 QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEEQCHI..... 178 lactoglobulin
```

## Pairwise alignment of retinol-binding protein from human (top) and rainbow trout (*O. mykiss*)

```
1 .MKWVWALLLA.AWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 48
::    ||  ||  ||  .||.||. .| :|||:..|.:|  |||.||||||

1 MLRICVALCALATCWA...QDCQVSNIQVMQNFDRSRYTGRWYAVAKKDP 47
.

49 EGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVCADMVGTFTDTE 98
||||| ||:||:|||||.||.|||  ||| :|||||:..|||.||| ||| ||| |

48 VGLFLLDNVVAQFSVDESGMTATAHGRVIIILNNWEMCANMFGTFEDTPD 97
.

99 PAKFKMKYWGVASFLQKGNDHWIVDTDYDTYAVQYSCRLLNLDGTCADS 148
|||||||:|||  ||:||  |||||:||:|||||  ||:  |||| ..|||||  |

98 PAKFKMRYWGAASYLQTGNDDHWVIDTDYDNYAIHYSCREVDLDGTCLDG 147
.

149 YSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL 199
|||:|||  |  ||  ||  |||  :..|:|  .|| : |  |:|:

148 YSFIFSRHPTGLRPEDQKIVTDKKKEICFLGKYRRVGHTGFCESS..... 192
```

# Significant percent identity

Burkhard Rost:

- 90% of sequence pairs with >30% identity were structurally similar proteins
- 20-30% identity: **twilight zone** - homology may exist but cannot be assumed in the absence of other evidence.

Alignment can reveal homology  
between sequences

## The Role of Homology

- *homology*: similarity due to descent from a common ancestor
- often we can infer homology from similarity
- thus we can sometimes infer structure/function from sequence similarity

# Alignment can reveal homology between sequences

- Homology  $\neq$  similarity
- Similarity: sequences in question match to some degree
- Homology: similarity due to descent from a common ancestor
  - Implies common function (but not always!!)

# Differing rates of DNA evolution

- Functional/selective constraints (particular features of coding regions, particular features in 5' untranslated regions)
- Variation among different gene regions with different functions (different parts of a protein may evolve at different rates).
- Within proteins, variations are observed between
  - surface and interior amino acids in proteins (order of magnitude difference in rates in haemoglobins)
  - charged and non-charged amino acids
  - protein domains with different functions
  - regions which are strongly constrained to preserve particular functions and regions which are not
  - different types of proteins -- those with constrained interaction surfaces and those without

# Scoring alignments

- Quality of an alignment is measured by a score
  - Simplest: measure identity

```
T H I S - -     S E Q U E N C E  
T H I S I S A S E Q U E N C E
```

68.75% (11 matches over 16 positions, including gaps)

Is this significant - i.e. is 30% identity over a long alignment the same as 30% identity over a short sequence?

# A simple alignment

- Let us try to align two short nucleotide sequences:
  - AATCTATA and AAGATA
- Without considering any gaps (insertions/deletions) there are 3 possible ways to align these sequences

AATCTATA

AAGATA

AATCTATA

AAGATA

AATCTATA

AAGATA

- Which one is better?

# What is a good alignment?

AGGCTAGTT , AGCGAAGTTT

AGGCTAGTT-      6 matches, 3 mismatches, 1 gap  
AGCGAAGTTT

AGGCTA-GTT-      7 matches, 1 mismatch, 3 gaps  
AG-CGAAGTTT

AGGC-TA-GTT-      7 matches, 0 mismatches, 5 gaps  
AG-CG-AAGTTT

# Scoring the alignments

- We need to have a scoring mechanism to evaluate alignments
  - match score
  - mismatch score
- We can have the total score as:  
$$\sum_{i=1}^n \text{match or mismatch score at position } i$$
- For the simple example, assume a match score of 1 and a mismatch score of 0:

AATCTATA

AAGATA

4

AATCTATA

AAGATA

1

AATCTATA

AAGATA

3

# Matches do not have to identical - similarity

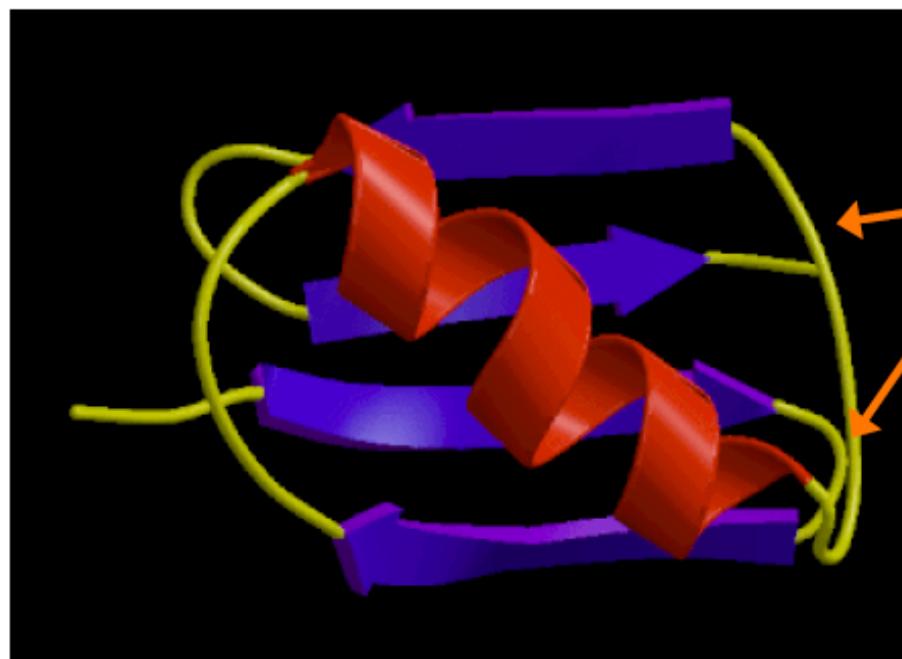
- Some amino acids resemble each other and can substitute functionally for each other.
  - Identical - highest score
  - Similar - high score (I, L)
  - Different - lowest score (I, K)
- Get overall alignment score.

# Typical score matrix

- DNA
  - Match = +1
  - Mismatch = -3
  - Gap penalty = -5
  - Gap extension penalty = -2
- Protein sequences
  - Similarity score
  - Gap open penalty = -11
  - Gap extension = -1

# Insertions/Deletions and Protein Structure

- Why is it that two “similar” sequences may have large insertions/deletions?
  - some insertions and deletions may not significantly affect the structure of a protein



*loop structures:* insertions/deletions here not so significant

# Sequence Alignment in PDZ domains



cyan:syntenin (1nte)

green:erbin (2h3l)

Insert gaps  
- dynamic programming

# Score matrix

- Instead of having a single match/mismatch score for every pair of nucleotides or amino acids, consider chemical, physical, evolutionary relationships:
  - E.g.
    - alanine vs. valine or alanine vs. lysine? Alanine and valine are both small and hydrophobic, but lysine is large and charged.
    - which substitutions occur more in nature?
- Assign scores to each pair of symbol
  - Higher score means more similarity

# Scoring an Alignment

- the score of an alignment is the sum of the scores for pairs of aligned characters plus the scores for gaps
- example: given the following alignment

**VAHV---D--DMPNALSALSSDLHAHKL**

**AIQLQVTGVVVTDATLKNLGSVHVSKG**

- we would score it by
- $$s(V,A) + s(A,I) + s(H,Q) + s(V,L) + 3g + s(D,G) + 2g \dots$$

# What's a scoring matrix?

- Substitution matrices are used for amino acid alignments. These are matrices in which each possible residue substitution is given a score reflecting the probability that it is related to the corresponding residue in the query.
- A unitary matrix is used for DNA pairs because each position can be given a score of +1 if it matches and a score of zero if it does not.

	A	C	D	E	F	G	H →
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	
G	0	-3	-1	-2	-3		
H	-2	-3	-1	0			

BLOSUM 62

# **SUBSTITUTION MATRICES**

- In aligning two protein sequences, some method must be used to score the alignment of one residue against another. Substitution matrices contain such values
- PAM (Dayhoff) & BLOSSUM

# Substitution matrix

BLOSUM62

Color coding:  
physicochemical  
properties

20x20 matrix

Each cell occupied by a score representing the likelihood that that pair of aa will occupy same position through homology.

T	H	I	S	S	E	Q	U	E	N	C	E
T	H	A	T	S	E	Q	U	E	N	C	E
5	8	-1	1	4	5	5	0	5	6	9	5

(A)

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

Overall score: 52 (Blosum)

# BLOSUM

BLOSUM matrices are based on local alignments.

BLOSUM stands for blocks substitution matrix.

BLOSUM62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.

Derived from local, ungapped alignments of distantly related sequences

The BLOSUM series of matrices were created by Steve Henikoff and colleagues (PNAS 89:10915).

# The relative mutability of amino acids

Dayhoff et al. described the “relative mutability” of each amino acid as the probability that amino acid will change over a small evolutionary time period. The total number of changes are counted (on all branches of all protein trees considered), and the total number of occurrences of each amino acid is also considered. A ratio is determined.

Relative mutability  $\propto$  [changes] / [occurrences]

Example:

sequence 1	ala his val ala
sequence 2	ala argserval

For **ala**, relative mutability = [1] / [3] = 0.33

For **val**, relative mutability = [2] / [2] = 1.0

# The relative mutability of amino acids

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Note that alanine is normalized to a value of 100.

Trp and cys are least mutable. **WHY?**

Asn and ser are most mutable.

# Types of alignment

- Global (Needleman & Wunsch)
  - Strings of similar size
    - Genes with a similar structure
    - Larger regions with a preserved order (syntenic regions)
- Local (Smith & Waterman)
  - Finding similar regions among
    - Dissimilar regions
    - Sequences of different lengths

# **Two kinds of sequence alignment: global and local**

global alignment algorithm of Needleman and Wunsch (1970).

Local alignment algorithm of Smith and Waterman (1981).

BLAST, a heuristic version of Smith-Waterman (Altschull).

# Global vs Local Alignments

(A) local

PI3-kinase DRHNSNIMVKDDGQLFHI<sup>DFG</sup>  
cAMP PK DLKPENLLIDQQGYIQVT<sup>DFG</sup>

(B) global

PI3-kinase HQLGNLRL--LEE<sup>CRI</sup>--MSSAKRPLWLWNWENPDIMSEL<sup>L</sup>FQNNEIIFKNGDDLQDMLT  
cAMP PK GNAAAAKKGXEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVML-  
10 20 30 40 50  
10 20 30 40 50

PI3-kinase LQIIRIME--NIWQNQGLDLRMLPYGCLSIGDCVGLIEVVVRNSHTIMQ-IQCKGGLKGA<sup>L</sup>  
cAMP PK ---VKHMETGNHYAMKILDQKVKVVK-----LKQIEHTLN<sup>E</sup>NEKRILQAVNFPFLVKLEF  
60 70 80 90 100 110  
60 70 80 90 100

PI3-kinase QFNSH<sup>T</sup>-LHQWLKDKNKGEIYDA--IDL<sup>F</sup>TRSCAGYC<sup>V</sup>ATFIL<sup>G</sup>IGDRHNSNIMVKD-D  
cAMP PK SFKDNSNL<sup>Y</sup>MVM<sup>E</sup>YVPG<sup>G</sup>EMFSHLRRIGRFSEPHARFYAAQIVLT<sup>F</sup>EYLHSLDLIYRDLK  
120 130 140 150 160  
110 120 130 140 150 160

PI3-kinase GQLFH<sup>I</sup>DFG<sup>H</sup>FLDHKKKKFGYKRERV<sup>P</sup>----FVL<sup>T</sup>QDFL---IVISKGAQECTKTREFE  
cAMP PK PEN<sup>L</sup>LDQQGYI--QVT<sup>D</sup>FGFAK-RVKGRTWXLCGTPEYLAPEIILSKGYNKAVDW<sup>W</sup>ALG  
170 180 190 200 210 220  
170 180 190 200 210 220

PI3-kinase RF-QEMC--YKAYLAIRQHANLFINLF<sup>S</sup>MMLGSGMP<sup>E</sup>LQS<sup>F</sup>DDIAYIRKT<sup>L</sup>ALDKTEQEA  
cAMP PK VLIYEMAAGYPPFFA-DQPIQIYEKIVSGKVR--FPSHFSSDLKDLLRNLLQVDLTKR--  
230 240 250 260 270  
230 240 250 260 270 280

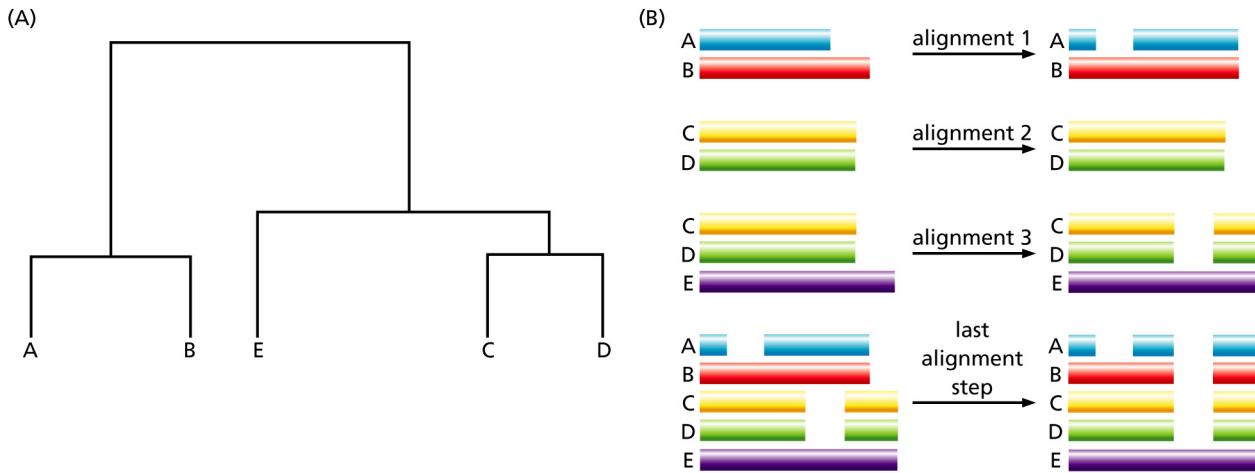
PI3-kinase LEYFMKQMNDAHGGWTTKMDWI-----FHTIKQHALN---  
cAMP PK EGNLKNGVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEEIRVXIN  
280 290 300 310 320 330 340  
290 300 310 320 330 340

Needleman-Wunsch  
(global alignment)

Smith-Waterman  
(local alignment)

Low overall similarity (18%)  
--> functionally important regions not matched!!

# Pairwise vs multiple alignments



Multiple sequence alignment (CLUSTALW):

Pairwise align all pairs (AB, AC, AD...)

Perform cluster analysis - rank in a tree according to similarity

Align most similar sequences in pairs, then align these to next closest sequence

# Multiple alignment

Improves accuracy of alignment for sequences of low similarity

(A) p110 $\alpha$   
cAMP-kinase

TFILGIGDRHNSNIMVKDDG-QLFHI	DFGHFLDHKKKKFGYKRERVPFVLT--QDFLIVI	142
QIVLTFEYLHSLDLIYRDLKPE	NLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPE	179

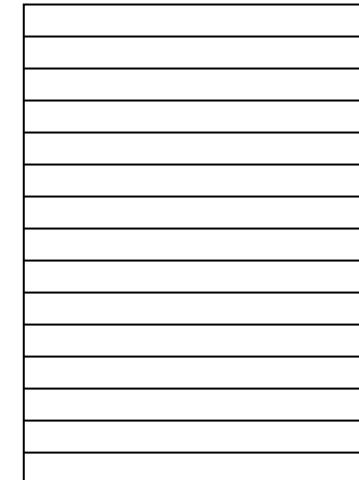
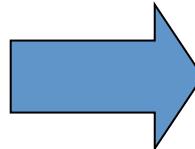
(B) p110 $\beta$   
p110 $\delta$   
p110 $\alpha$   
p110 $\gamma$   
p110\_dicti  
cAMP-kinase

SYVLGIG-----	DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT	136	
TYVLGIG-----	DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT	136	
TFILGIG-----	DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT	135	
TFVLGIG-----	DRHNDNIMITETGNLFHI	DFGHILGNYKSFLGINKERVPFVLT	135
TYVLGIG-----	DRHNDNLMVTKGGRLFHI	DFGHFLGNYKKFGFKRERAPFVFT	135
QIVLTFEYLHSLDLIYRDLKPE	NLLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA	177	

# Sequence Alignment vs. Database

- **Task:** Given a query sequence and millions of database records, find the record that has an optimal alignment with the query

ACTTTGGTGACTGTAC



# Sequence Alignment vs. Database

- **Tool:** Given two sequences, there exists an algorithm to find the best alignment.
- **Naïve Solution:** Apply algorithm to each of the records, one by one

# Sequence Alignment vs. Database

- **Problem:** An *exact algorithm* is too slow to run millions of times (even linear time algorithm will run slowly on a huge DB)
- **Solution:**
  - Run in parallel (expensive).
  - Use a fast (*heuristic*) method to discard irrelevant records. Then apply the *exact algorithm* to the remaining few.

# Sequence Alignment vs. Database

## General Strategy of Heuristic Algorithms:

- Homologous sequences are expected to contain ungapped (at least) short segments (probably with substitutions, but without ins/dels)
- Preprocess DB into some fast access data structure of short segments.

# **BLAST** (Basic Local Alignment Search Tool)

## Approximate Matches

### **BLAST:**

Words are allowed to contain inexact matching.

### **Example:**

In the polypeptide sequence I HAVE A DREAM

The 4-long word HAVE starting at position 2 may match  
HAVE, RAVE, HIVE, HALE, ...

BLAST is a heuristic application of Smith  
Waterman (local alignment).

# BLAST (Basic Local Alignment Search Tool)

- BLAST was developed and is maintained by [a group at the National Center for Biotechnology Information \(NCBI\)](#).
- **Local alignments**  
BLAST tries to find patches of regional similarity, rather than trying to find the best alignment between your entire query and an entire database sequence
- **Ungapped alignments**  
Alignments generated with BLAST do not contain gaps. BLAST's speed and statistical model depend on this, but in theory it reduces sensitivity. However, BLAST will report multiple local alignments between your query and a database sequence.

# HOW TO SCORE IN BLAST

- **Raw Score**

The score of an alignment,  $S$ , calculated as the sum of substitution scores. Substitution scores are given by a look-up table (e.g PAM, BLOSUM). Gap scores are typically calculated as the sum of  $G$ , the gap opening penalty and  $L$ , the gap extension penalty. For a gap of length  $n$ , the gap cost would be  $G+Ln$ . The choice of gap costs,  $G$  and  $L$  is empirical, but it is customary to choose a high value for  $G$  and a low value for  $L$ .

## E value

Expectation value. The number of different alignments with scores equivalent to or better than  $S$  that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

$$E = Kmn e^{-\lambda S} \quad (1)$$

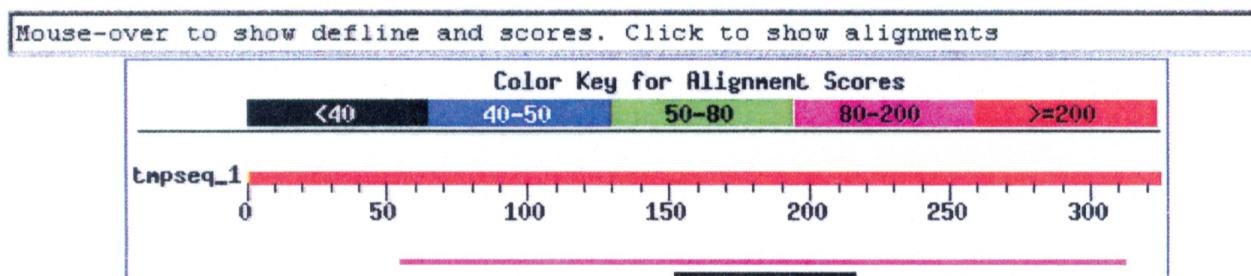
# BLAST output

(A)

<a href="#">sp P32871 P11A BOVIN</a>	PHOSPHATIDYLINOSITOL 3-KINASE CATALYTIC SUBUNIT	680	0.0
<a href="#">sp P42336 P11A HUMAN</a>	PHOSPHATIDYLINOSITOL 3-KINASE CATALYTIC SUBUNIT	676	0.0
<a href="#">sp P42337 P11A MOUSE</a>	PHOSPHATIDYLINOSITOL 3-KINASE CATALYTIC SUBUNIT	674	0.0
<a href="#">sp P42338 P11B HUMAN</a>	PHOSPHATIDYLINOSITOL 3-KINASE CATALYTIC SUBUNIT	338	9e-93
<a href="#">sp O35904 P11D MOUSE</a>	PHOSPHATIDYLINOSITOL 3-KINASE CATALYTIC SUBUNIT	332	7e-91
<a href="#">sp Q00329 P11D HUMAN</a>	PHOSPHATIDYLINOSITOL 3-KINASE CATALYTIC SUBUNIT	331	2e-90
<a href="#">sp P47473 RIR1 MYCGE</a>	RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE A SUBUNIT	34	0.59

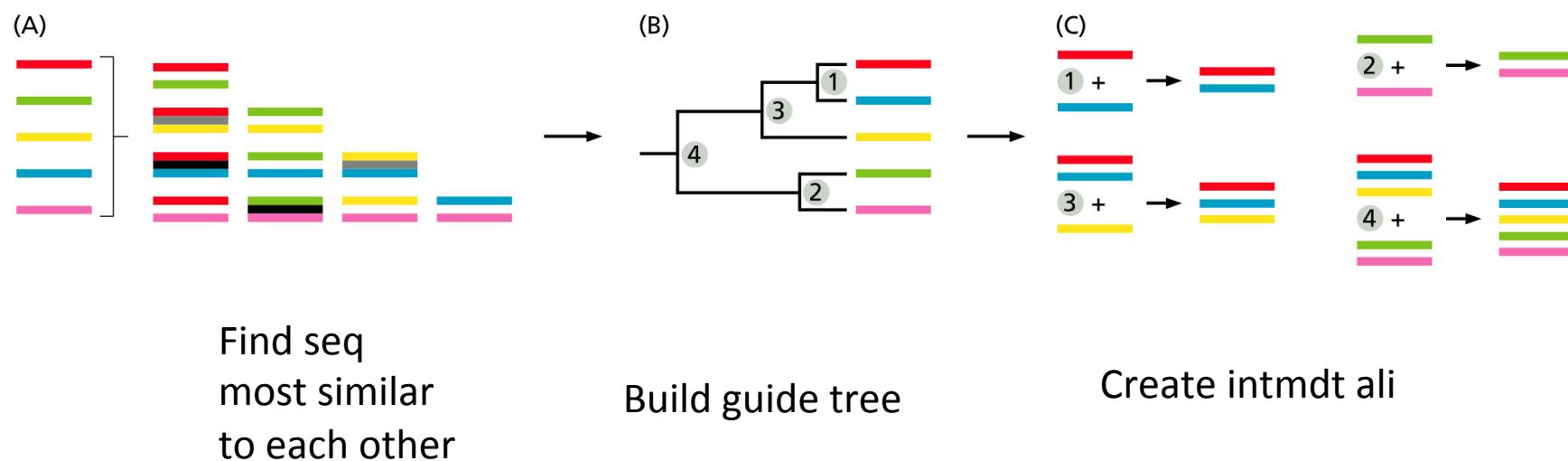
(B)

## Distribution of 2 Blast Hits on the Query Sequence



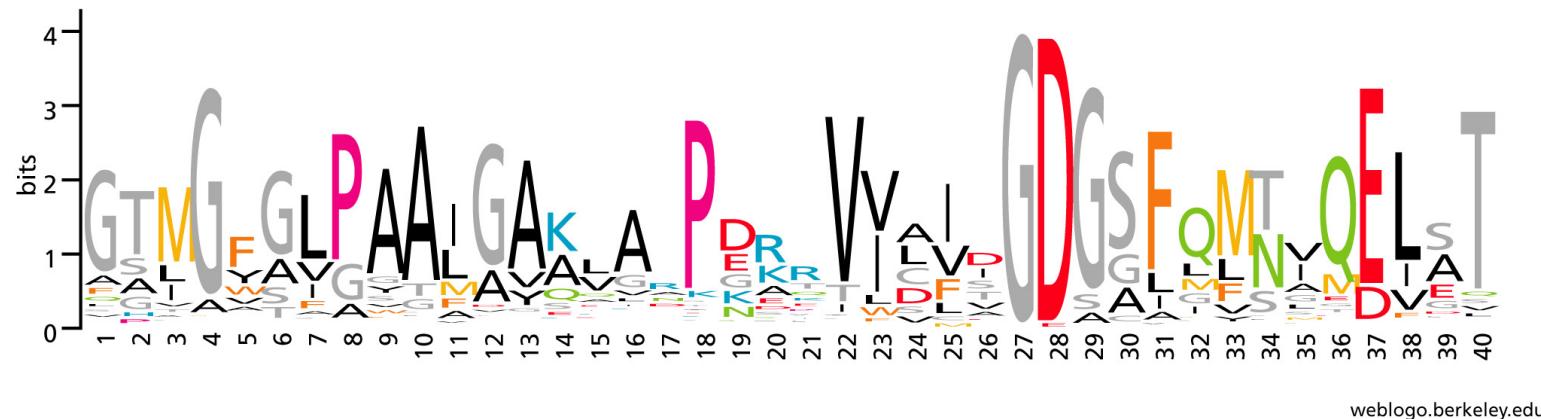
<a href="#">dbj BAB10275.1 </a>	(AB008266) phosphatidylinositol 4-kinase [A...	111	3e-25
<a href="#">dbj BAB11344.1 </a>	(AB011477) AtRAD3 [Arabidopsis thaliana]	38	0.008

# Progressive alignment



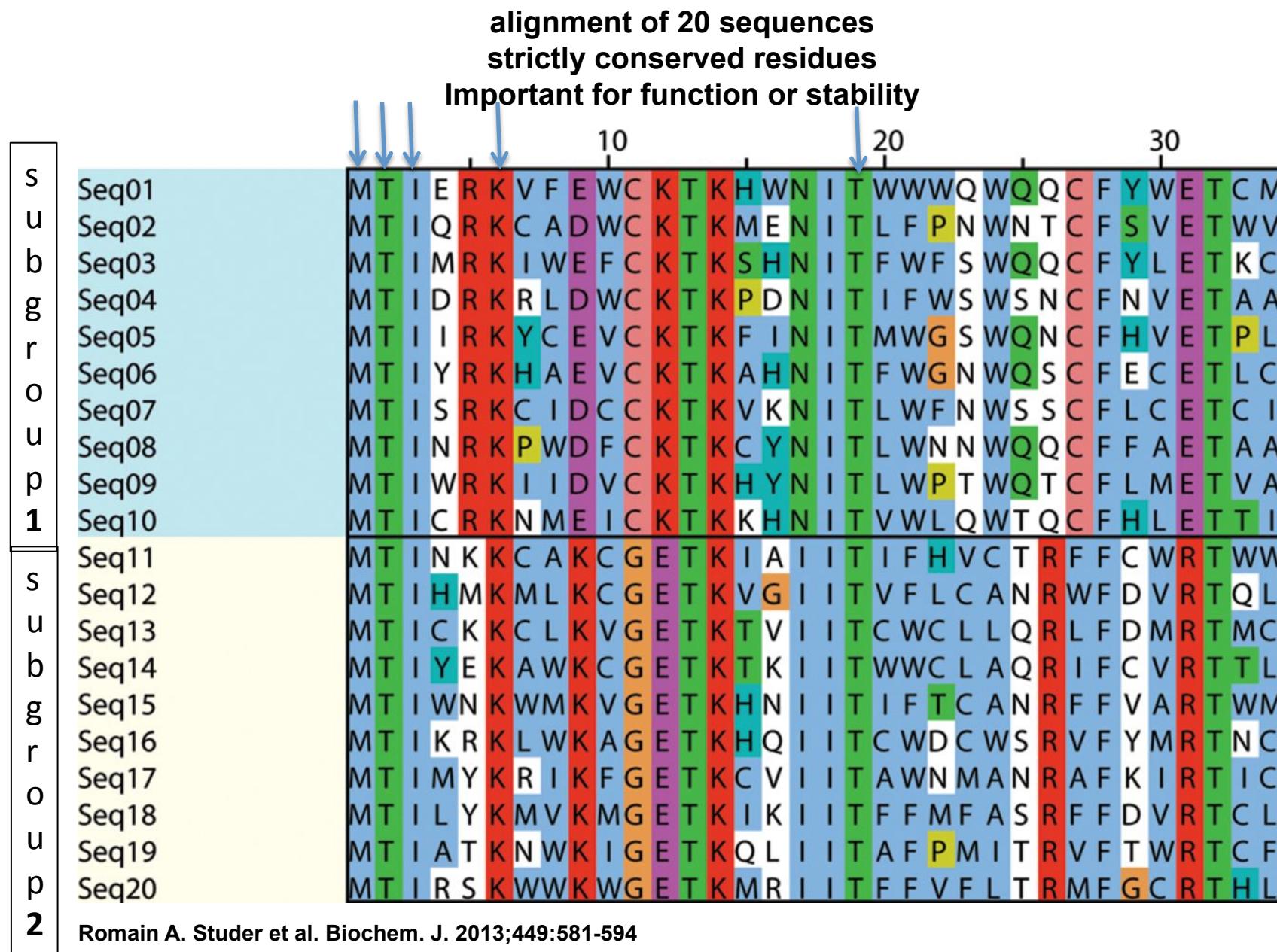
# Representing a profile as a logo

- Contribution of each residue to a position



Phosphate binding pattern from Prosite:

[LIVMF]-[GSA]-x(5)-P-x(4)-[LIVMFYW]-x-[LIVMF]-x-G-D-[GSA]-[GSAC]

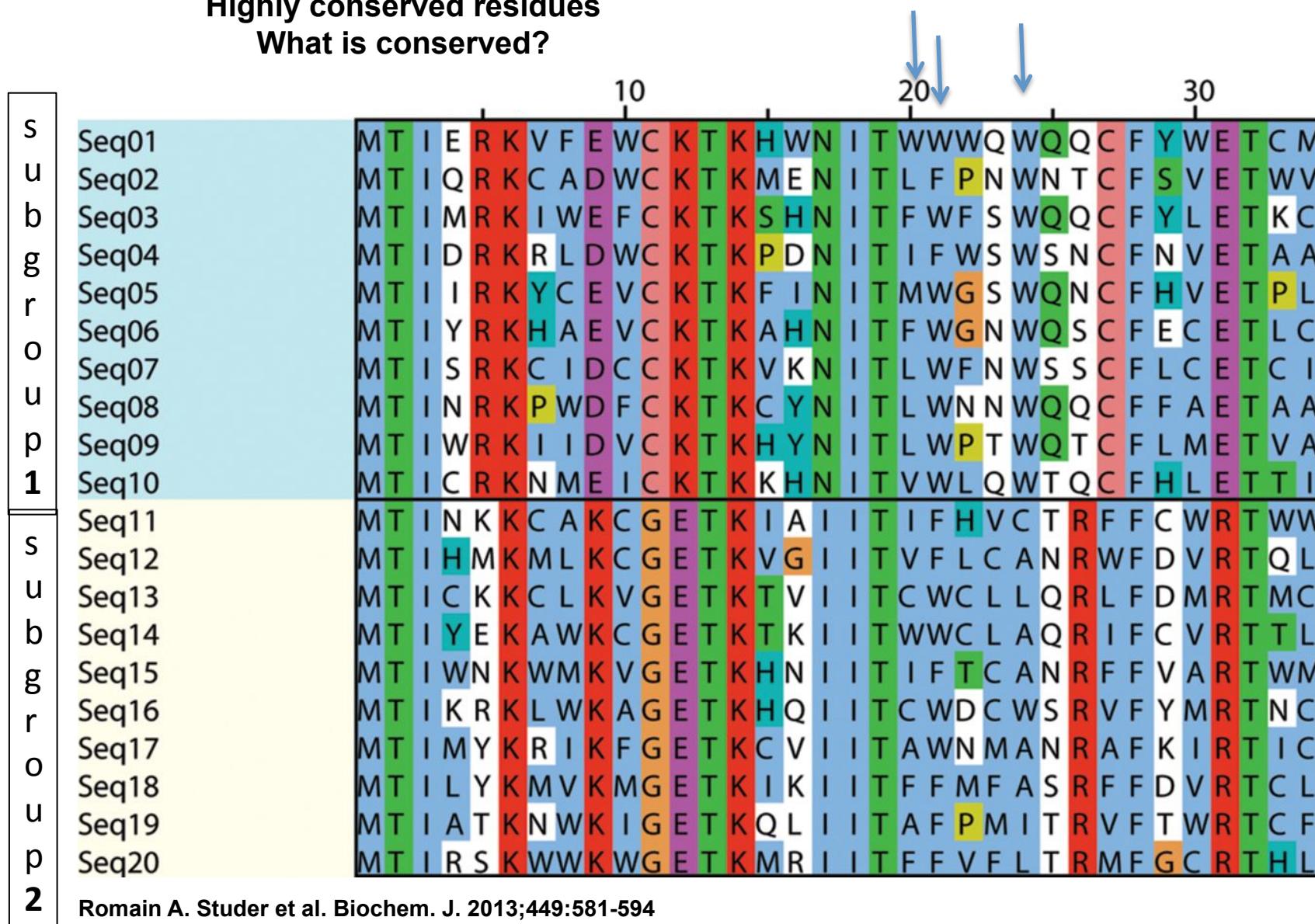


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green

©2013 by Portland Press Ltd His – turquoise, Pro – yellow, Gly – orange, Cys - pink

alignment of 20 sequences  
 Highly conserved residues  
 What is conserved?

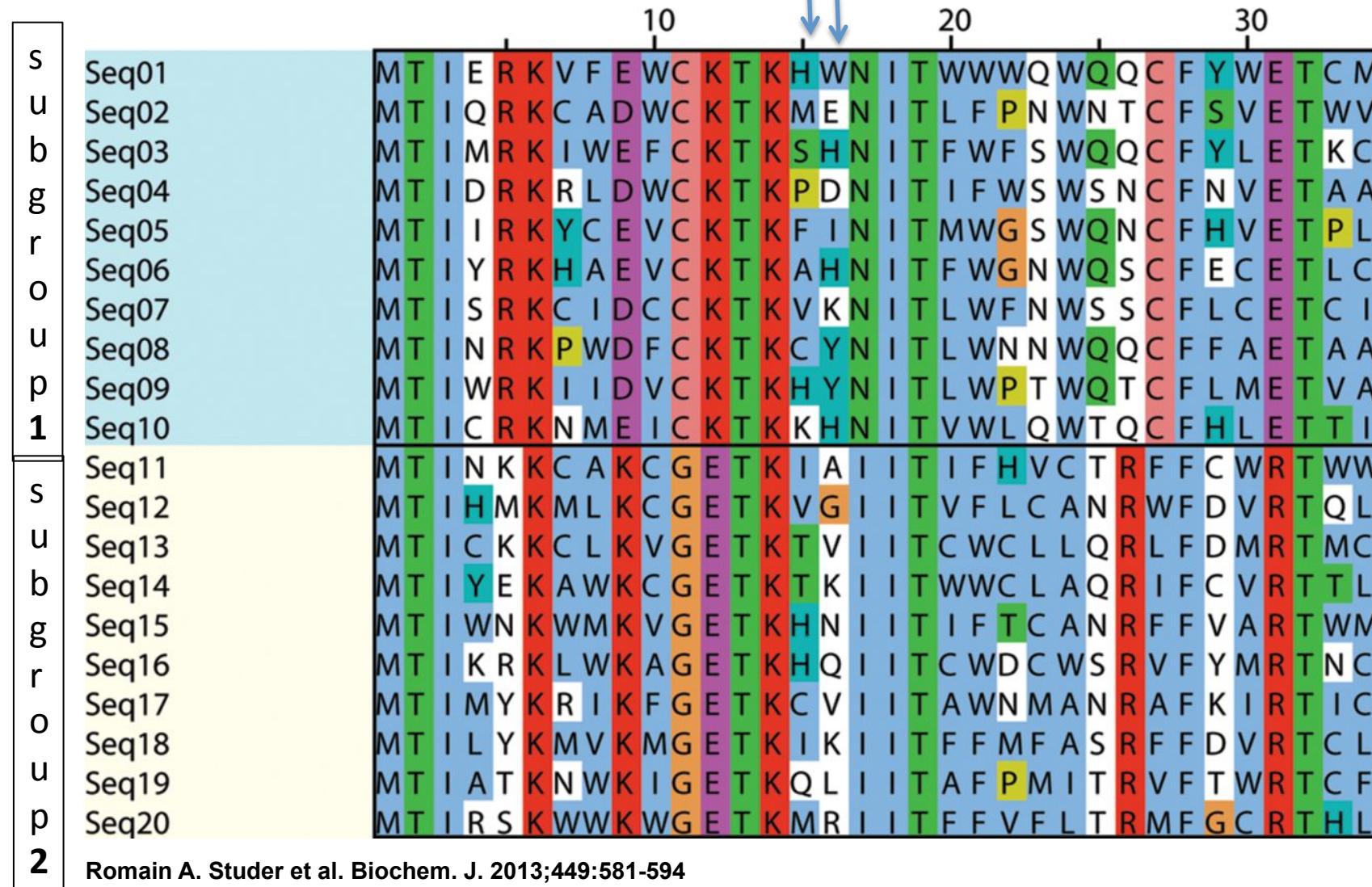


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green

©2013 by Portland Press Ltd His – turquoise, Pro – yellow, Gly – orange, Cys - pink

**alignment of 20 sequences**  
**Fully relaxed**  
**Where are they on the protein?**

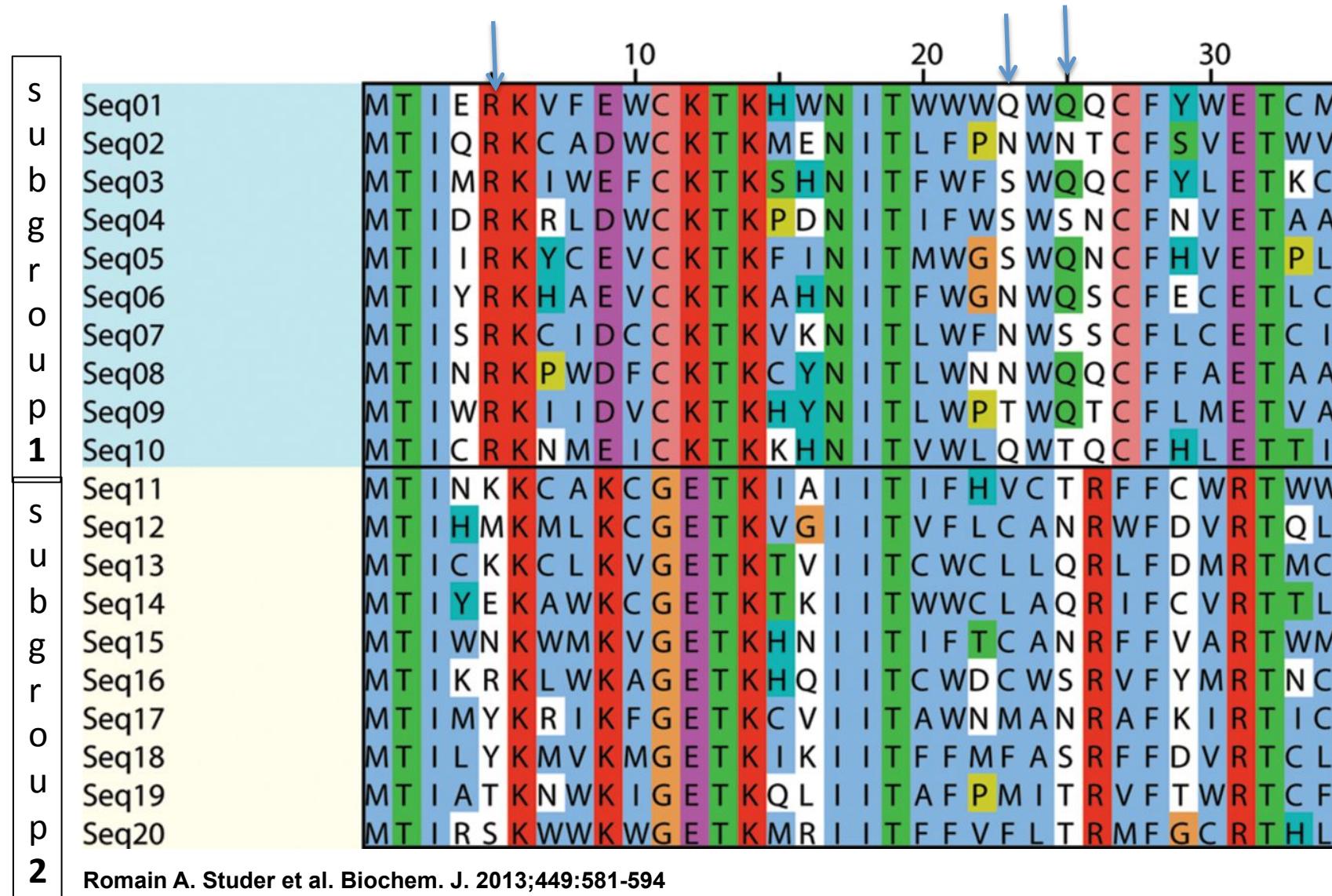


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green

©2013 by Portland Press Ltd His – turquoise, Pro – yellow, Gly – orange, Cys - pink

**alignment of 20 sequences**  
**Compare two subgroups**

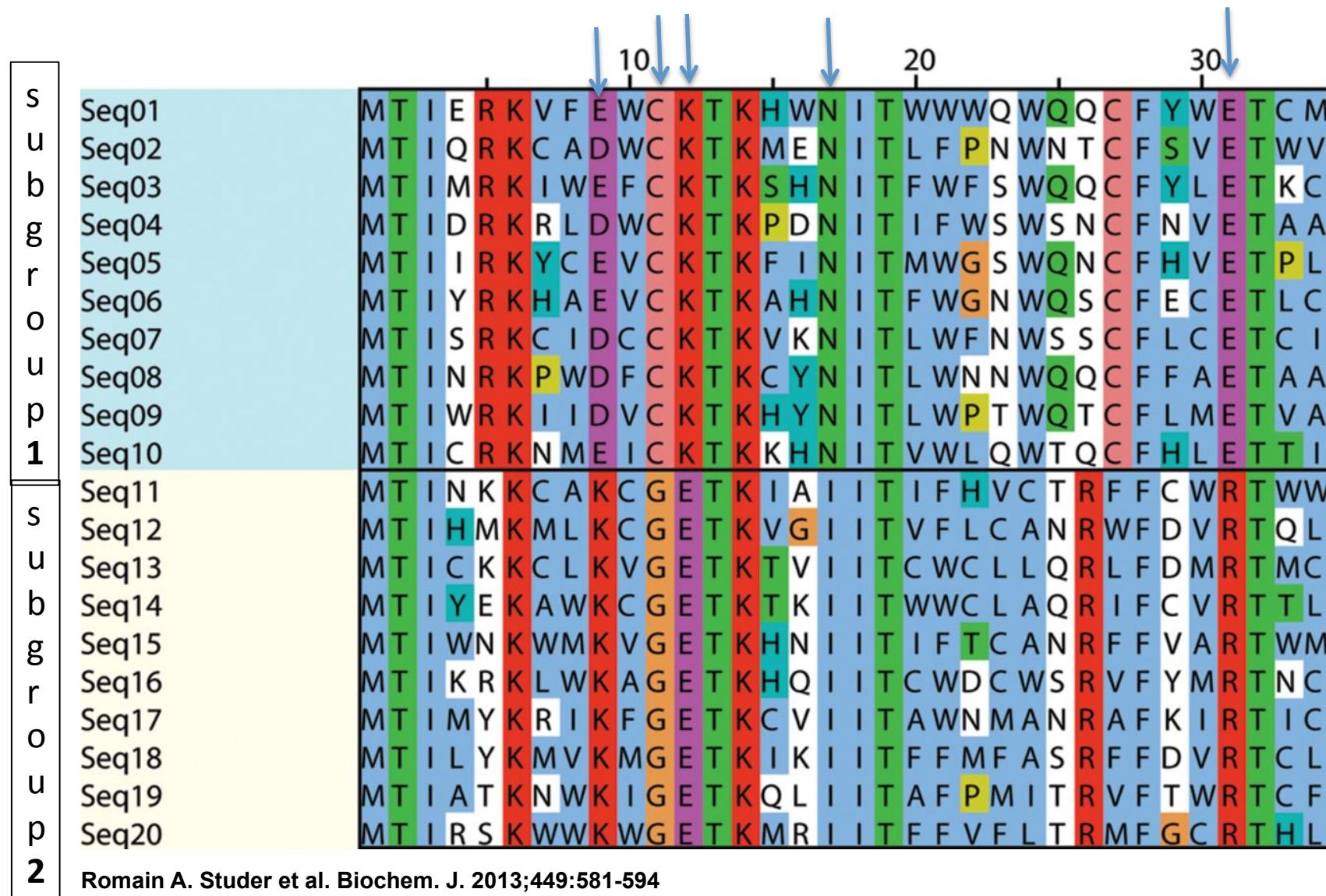


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green

©2013 by Portland Press Ltd His – turquoise, Pro – yellow, Gly – orange, Cys – pink.

**alignment of 20 sequences**  
**Physicochemical properties**

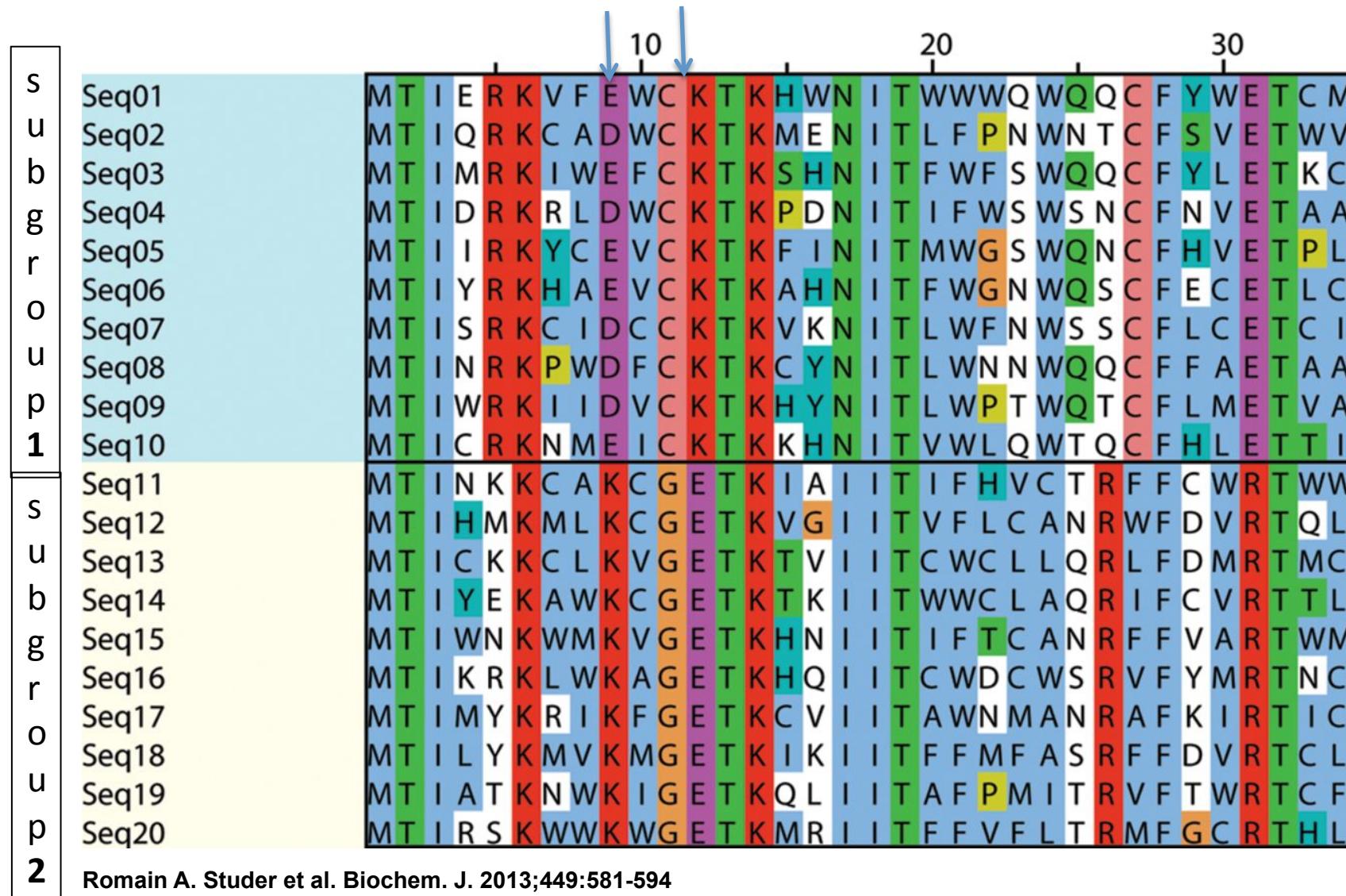


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green

©2013 by Portland Press Ltd His – turquoise, Pro – yellow, Gly – orange, Cys – pink.

**alignment of 20 sequences**  
**Coevolution**

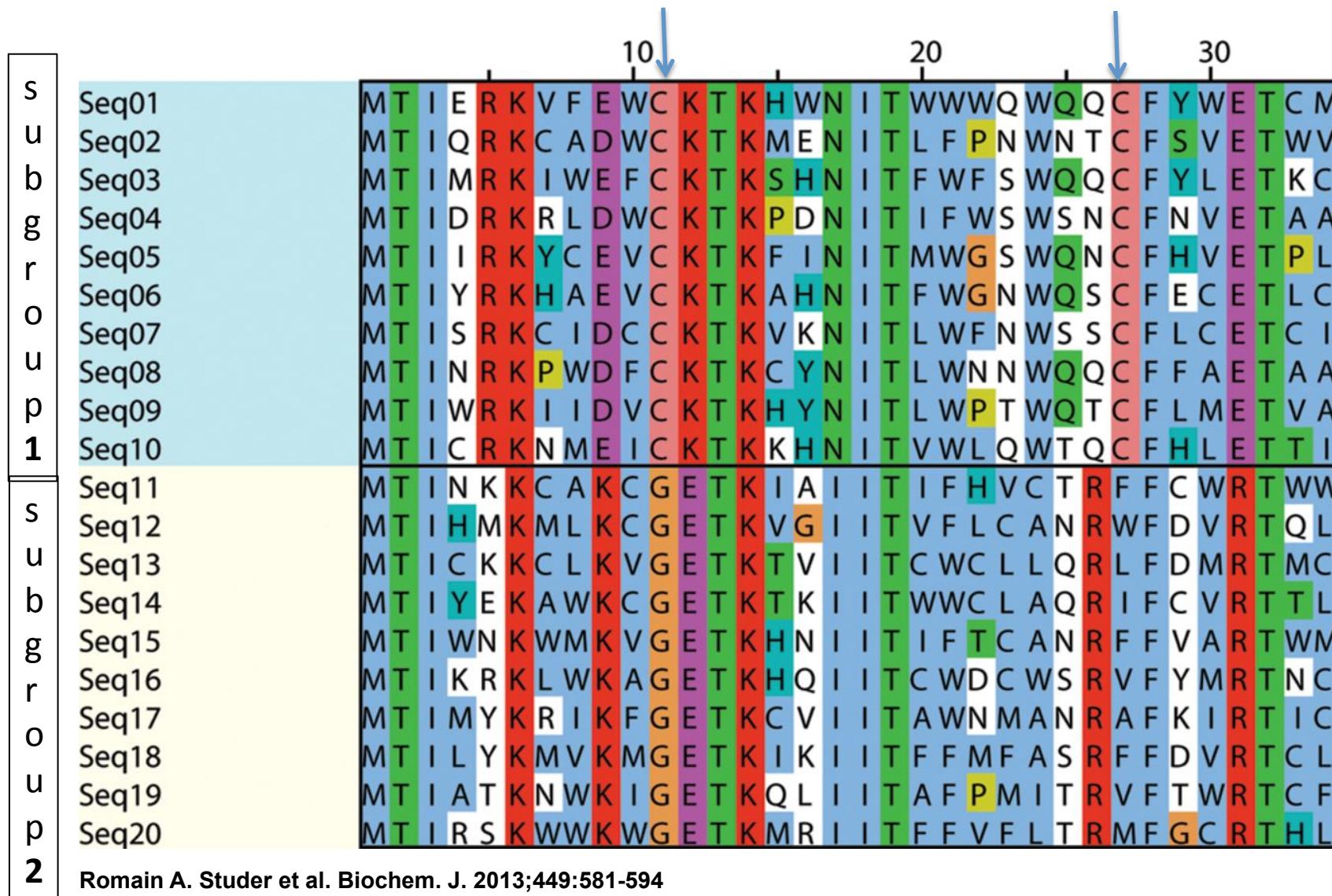


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green

©2013 by Portland Press Ltd His – turquoise, Pro – yellow, Gly – orange, Cys – pink.

**alignment of 20 sequences**  
**Coevolution**

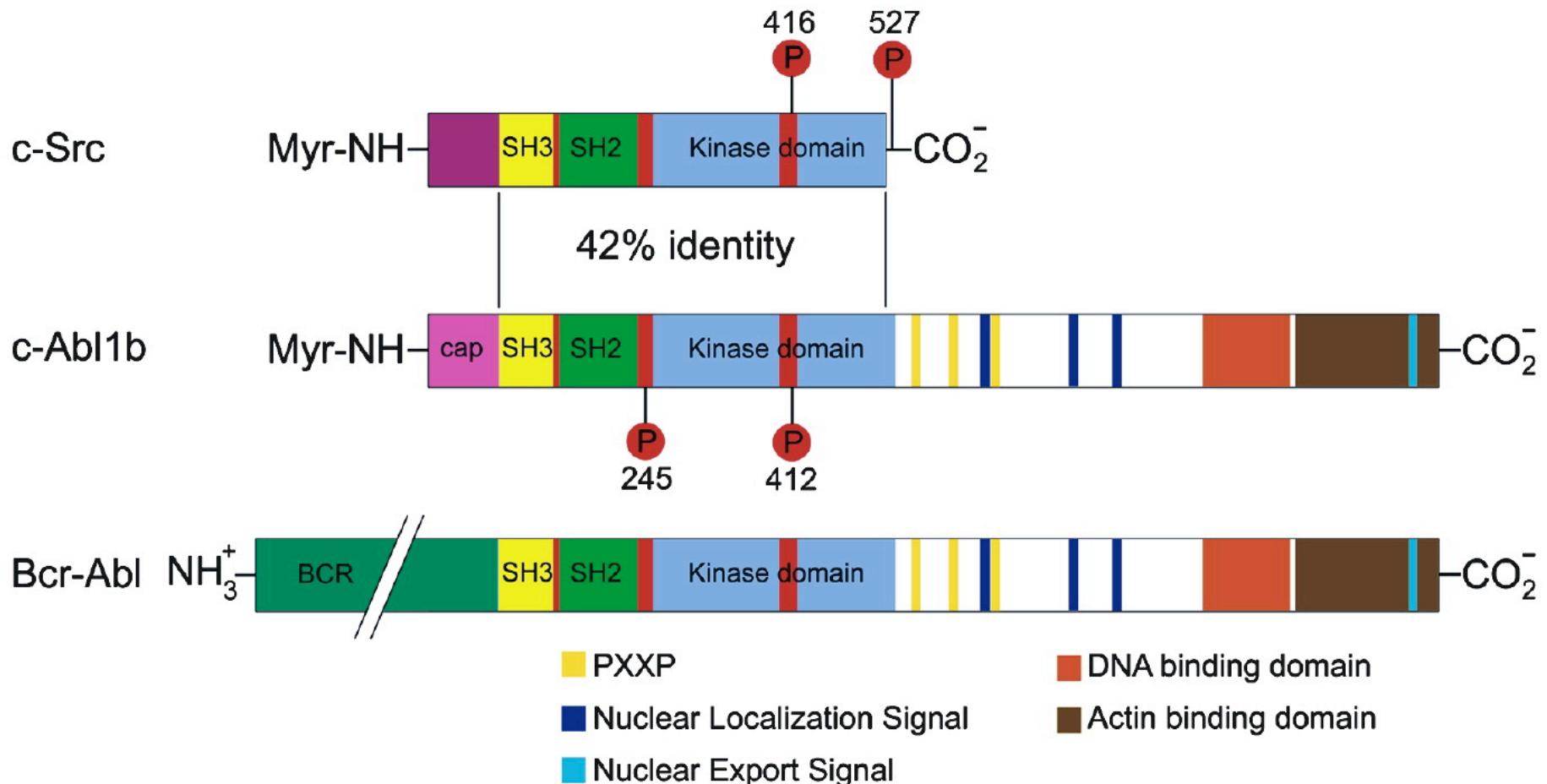


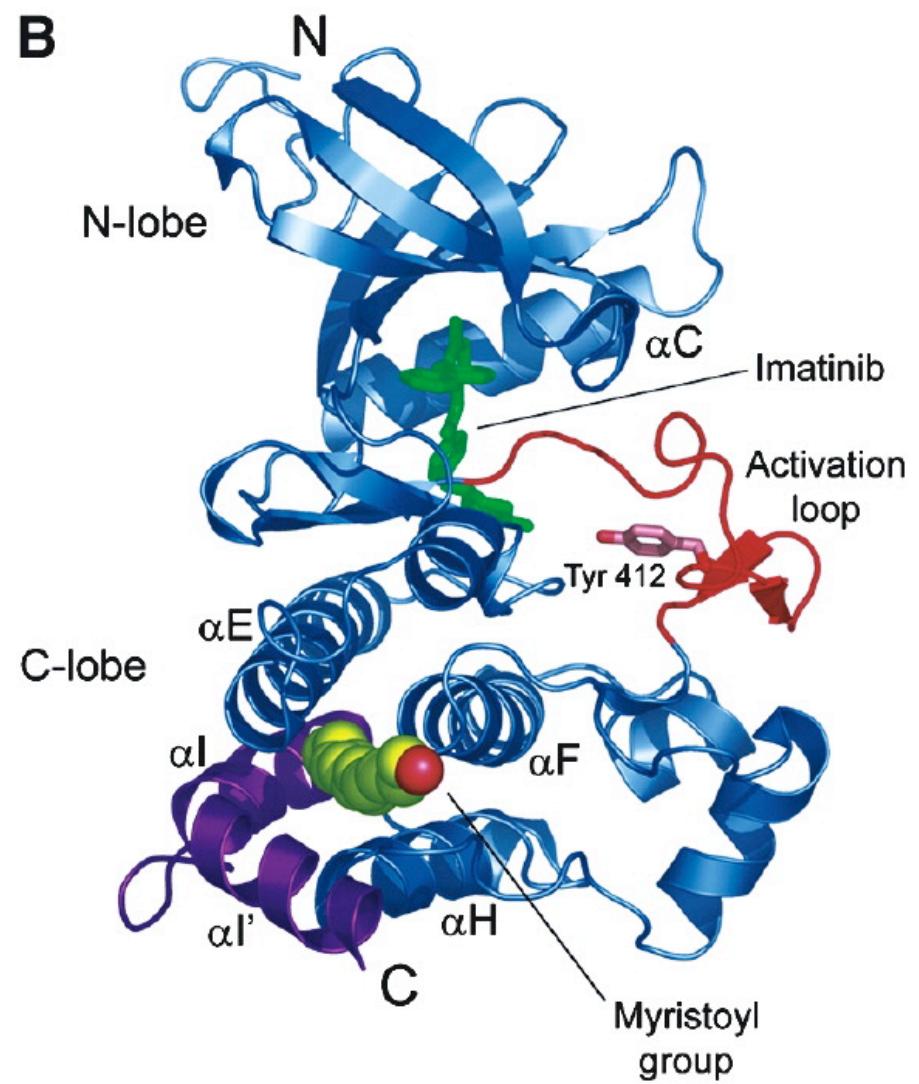
Romain A. Studer et al. Biochem. J. 2013;449:581-594

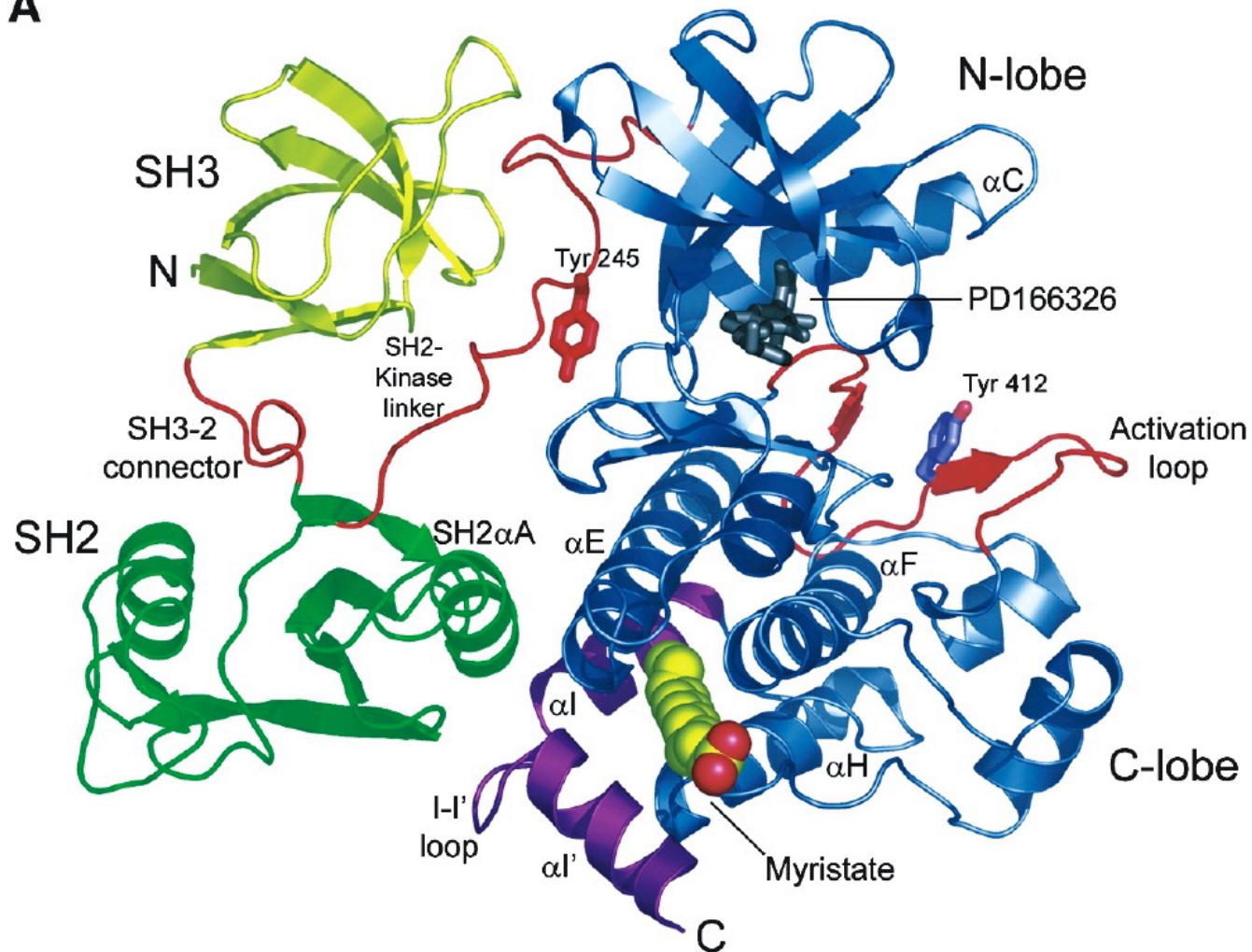
Hydrophobic – blue, basic – red, acidic – purple, polar – green

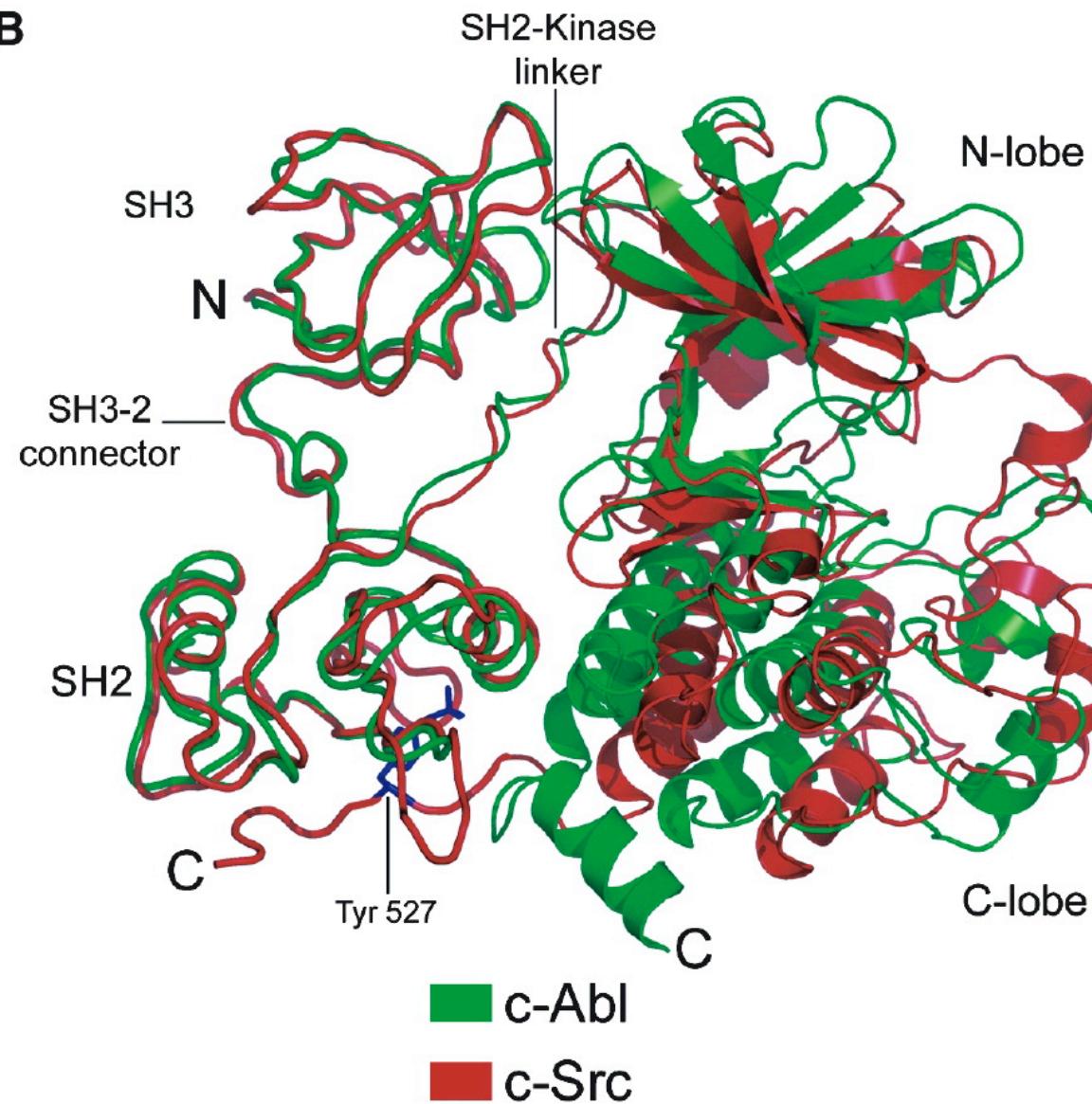
©2013 by Portland Press Ltd His – turquoise, Pro – yellow, Gly – orange, Cys – pink.

# Abl kinase





**A**

**B**

# Using the EMBL-EBI resources

- <https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/introduction-public-genetic>
- Go over the four case studies
- The first case study was related to a gene of interest. For your protein of interest, repeat the steps.