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

sequence patterns that repeat throughout sequence

Consensus Sequences
or Sequence Motifs

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

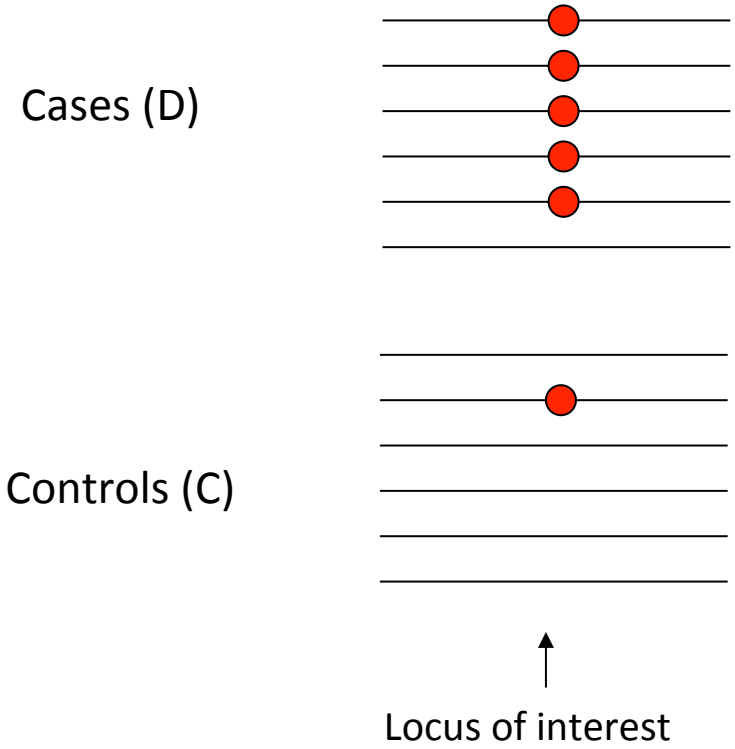
Sequences of Common
Structure or Function

Sequence Similarity

```
                10        20        30        40                50
Query VLSPADKTNVKAAWGKVGAHAGEVGAEALERMFLSFPTTKTYFPHF------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 betwen 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

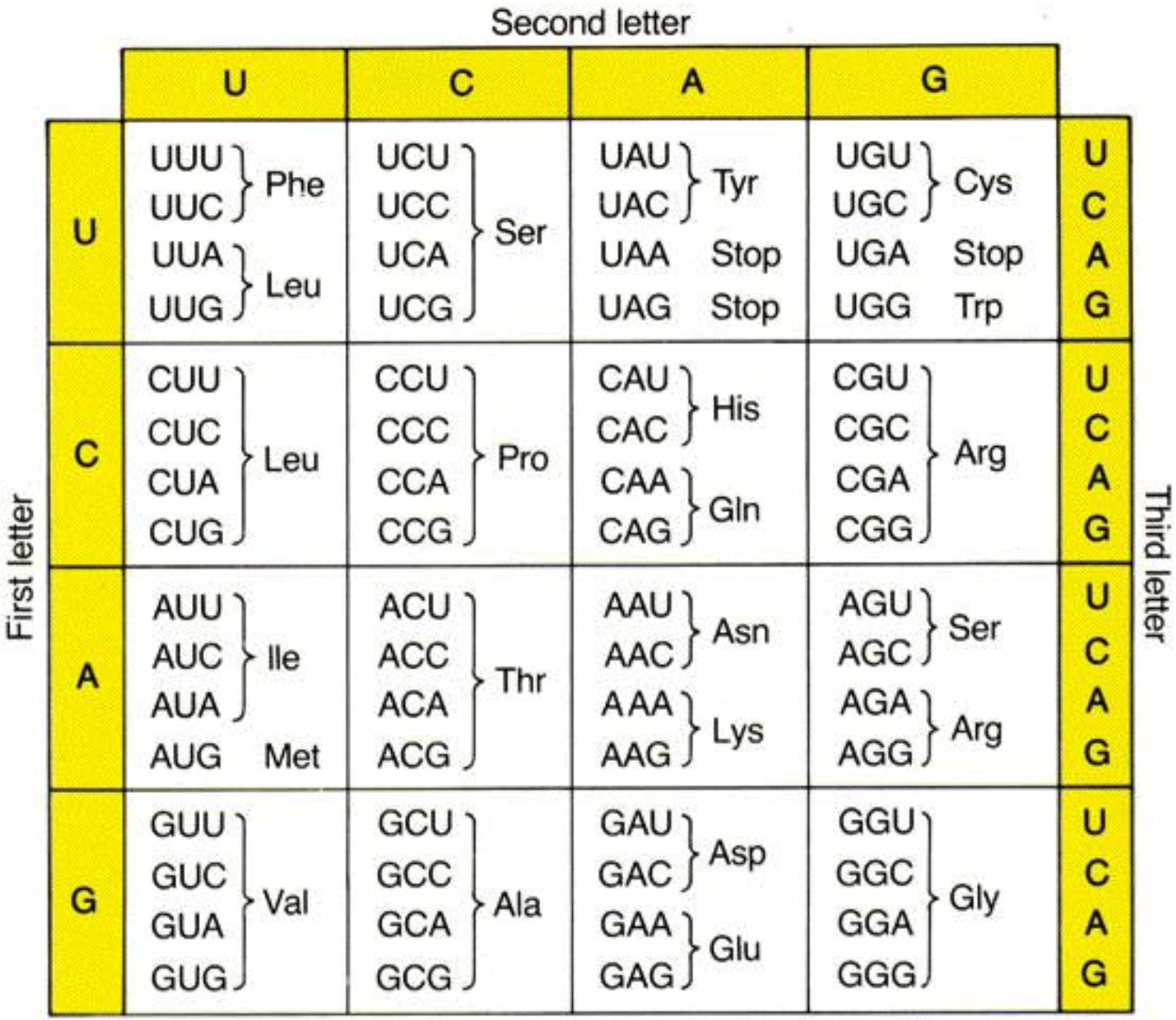
* Speciation is the origin of a new species capable of making a living in a new way from the species from which it arose. As part of this process it has also acquired some barrier to genetic exchange with the parent species.

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.

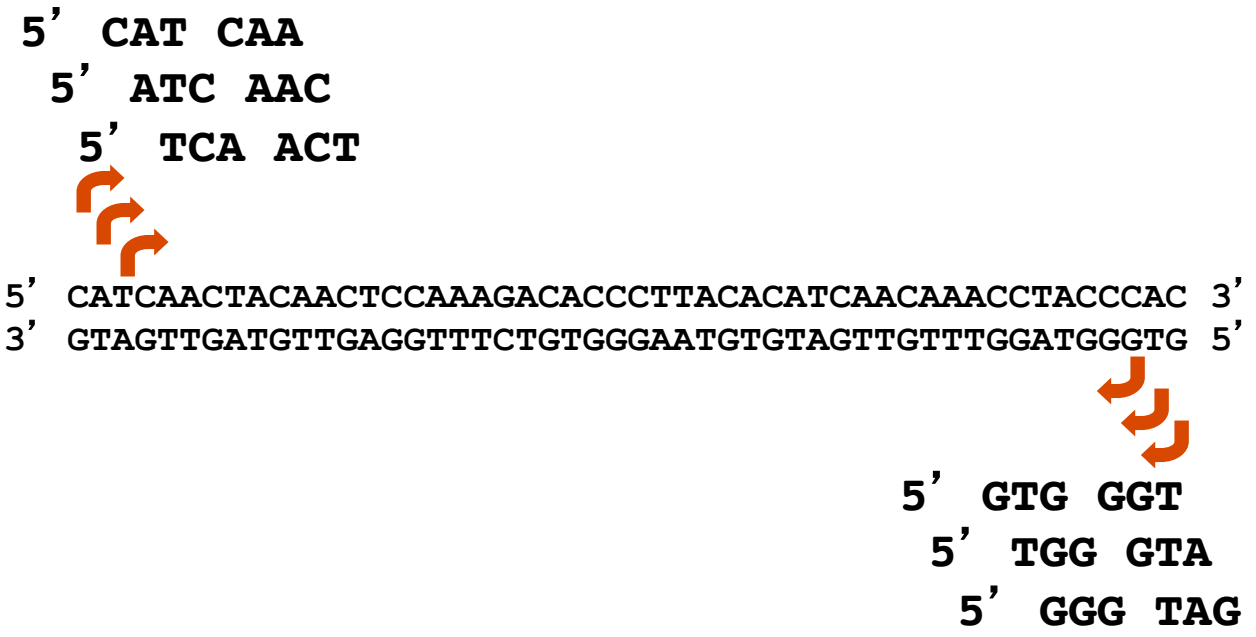
Pairwise alignment: protein sequences can be more informative than DNA

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

Pairwise alignment: protein sequences can be more informative than DNA

- DNA can be translated into six potential proteins



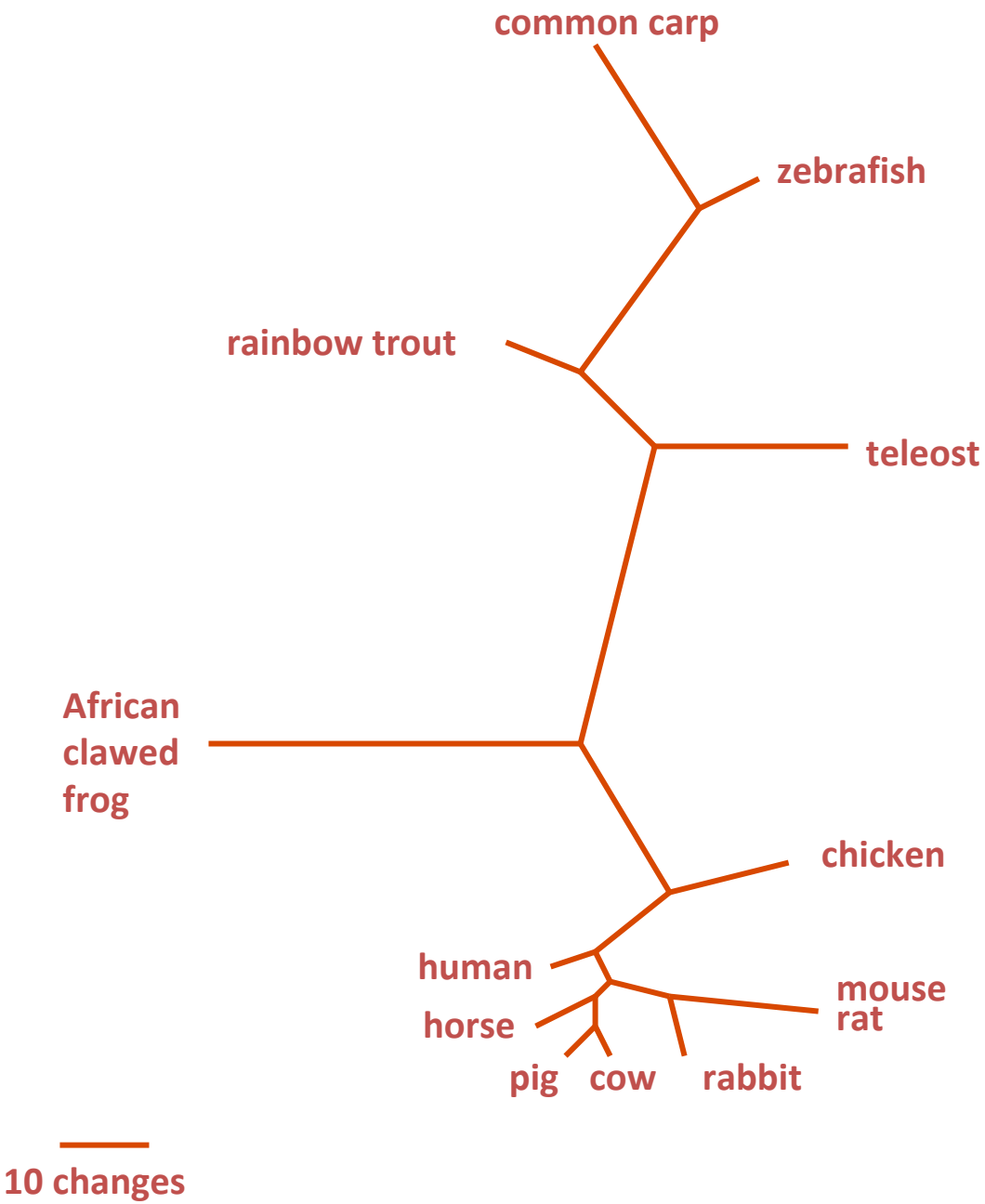
Pairwise alignment: protein sequences can be more informative than DNA

- 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

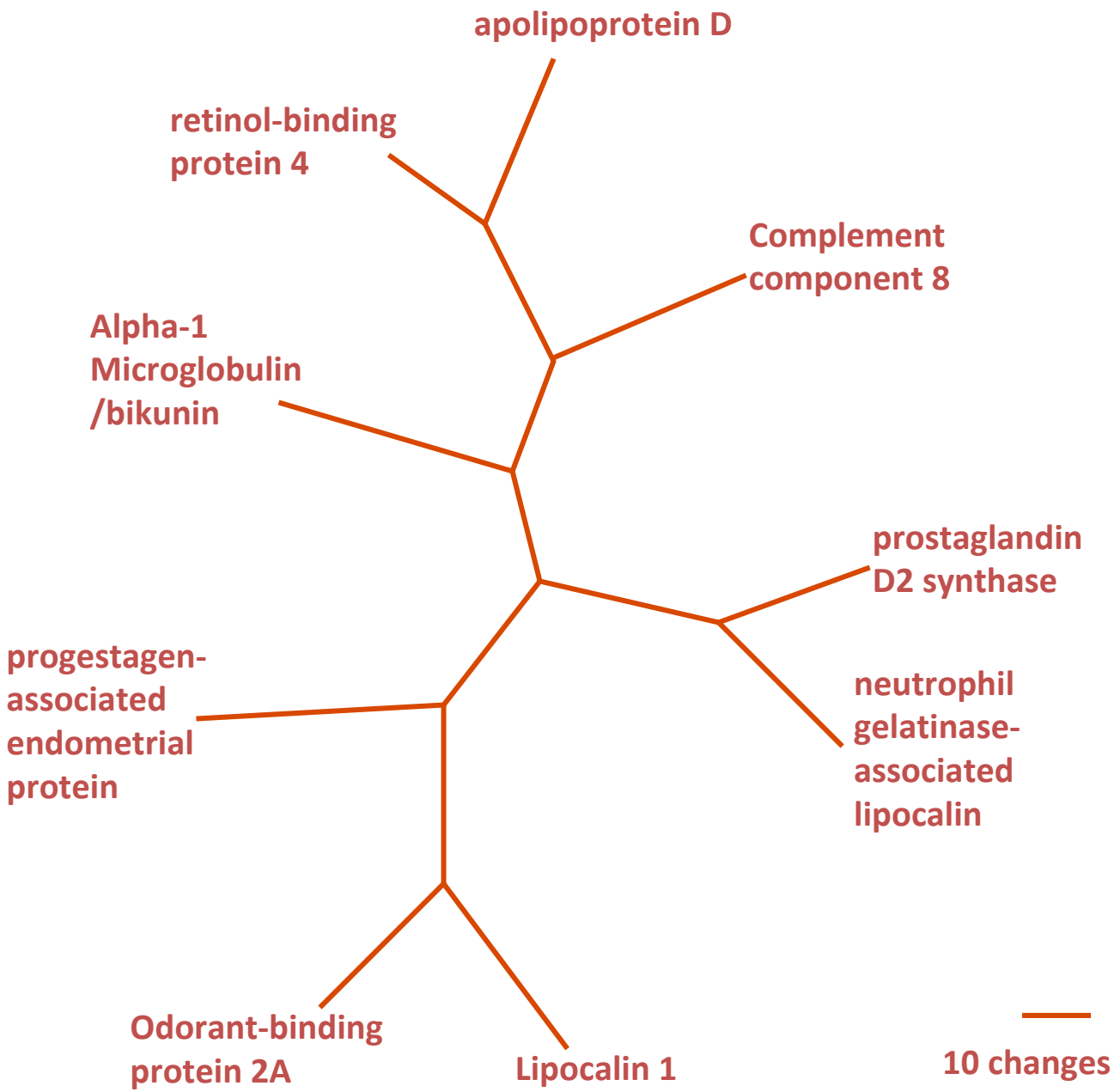
```
Query: 181 catcaactacaactccaaagacacccttacacccactaggatatcaacaaacctacccac 240
           |||||||| |||| |||||| ||||| | ||||||||||||||||||||||||||||||
Sbjct: 189 catcaactgcaaccccaaagccaccccct-cacccactaggatatcaacaaacctacccac 247
```

Definitions

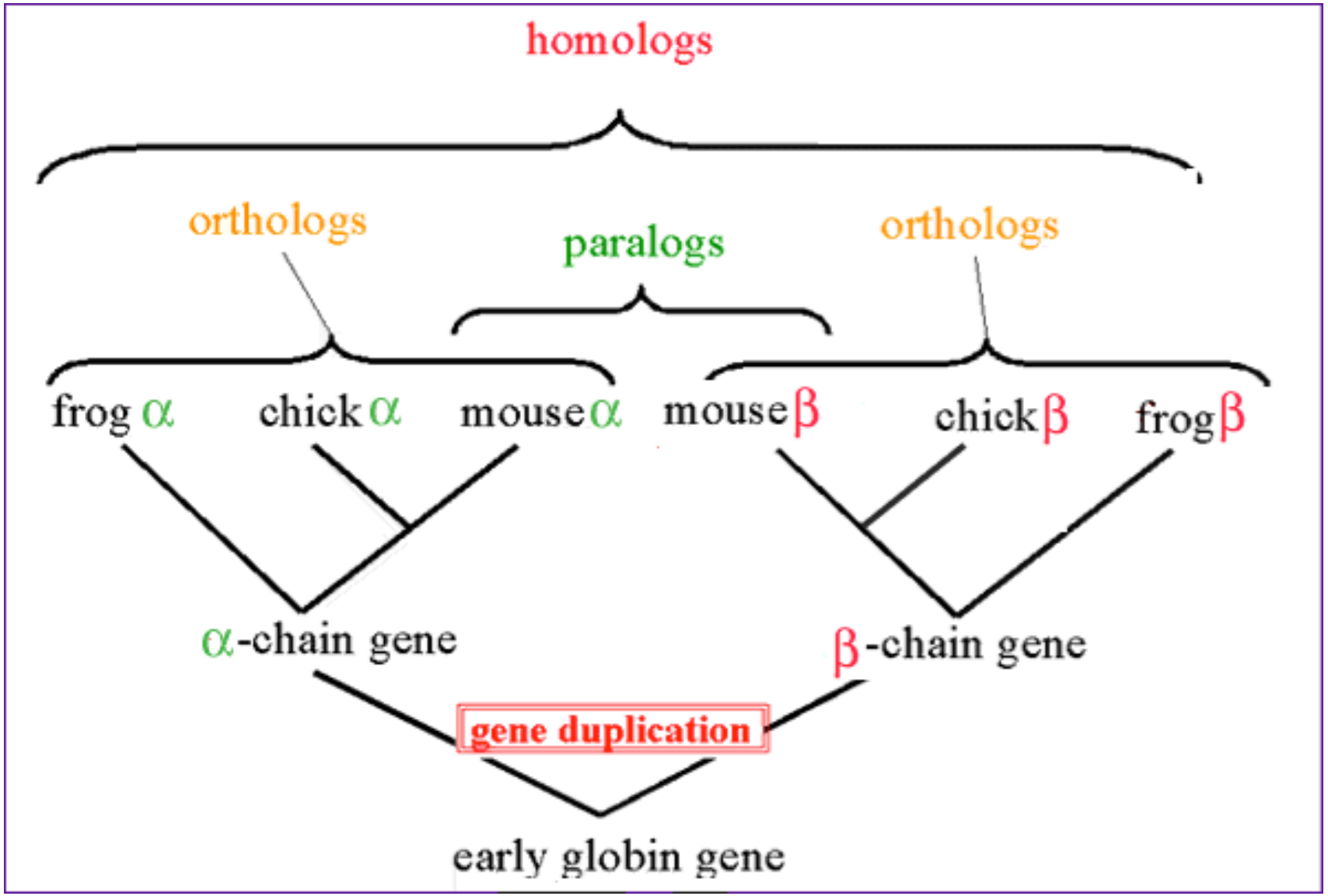
Homology
Similarity attributed to descent from a common ancestor.



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


homologs
orthologs
paralogs
orthologs
frog α
chick α
mouse α
mouse β
chick β
frog β
α-chain gene
β-chain gene
gene duplication
early globin gene

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/~wrp/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**K**ENFDKARFS**GTW**YA**MA**KKDPEGLF**L**QDNIV**A** 59
                 + **K**++ + ++  **GTW**++**MA**   +    **L** +   **A**
glycodelin:  23  QT**K**QDLELPKLA**GTW**HS**MA**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.

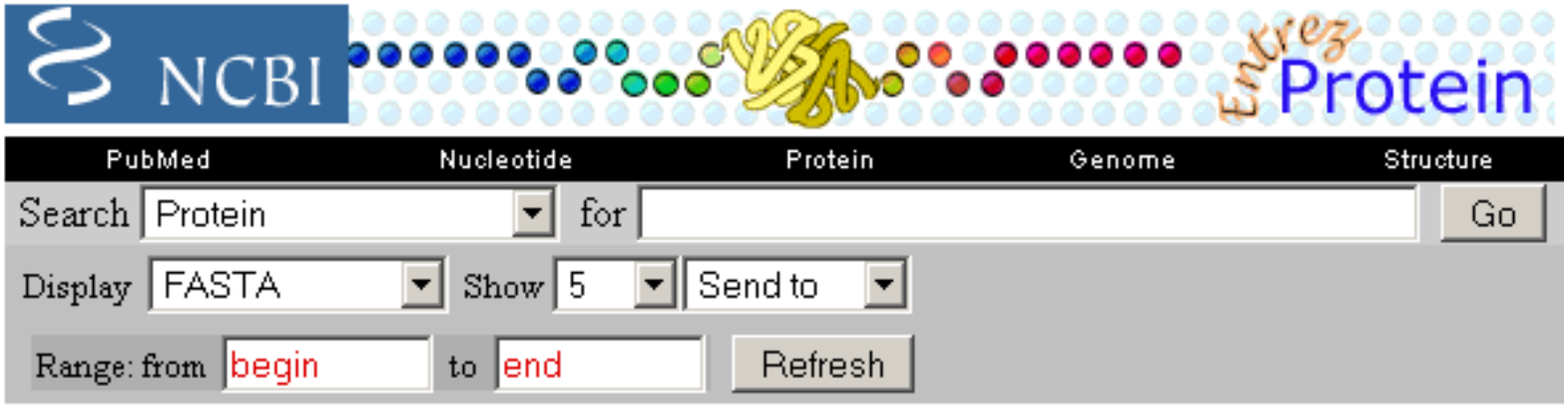
Conservation

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

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




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

Sequence Alignment

AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC

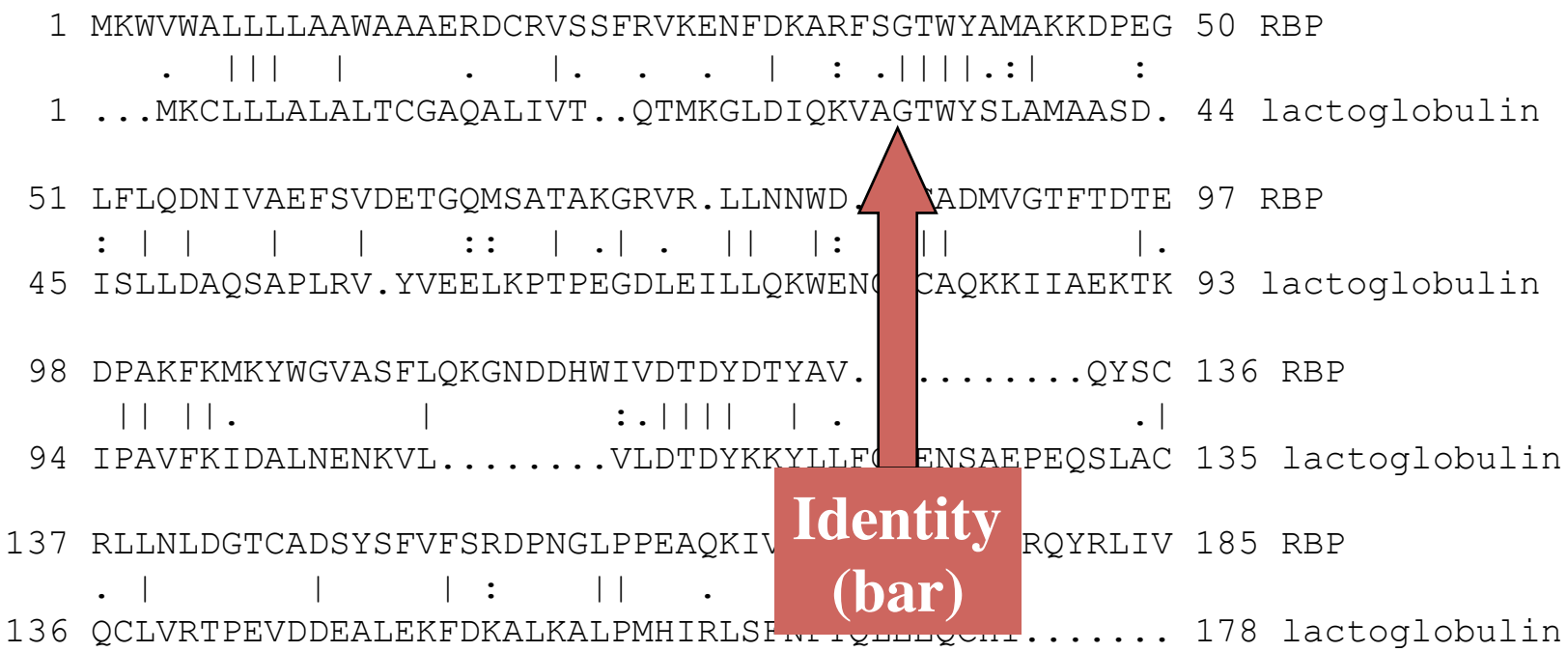
–AGG**CTATCAC**CT**GACC**T**C**CA**GG**C**CGA––TGCCC**–––
T**AG–CTATCAC––GACC**G**C––GG**T**CGA**TT**TGCCC**GAC

Definition

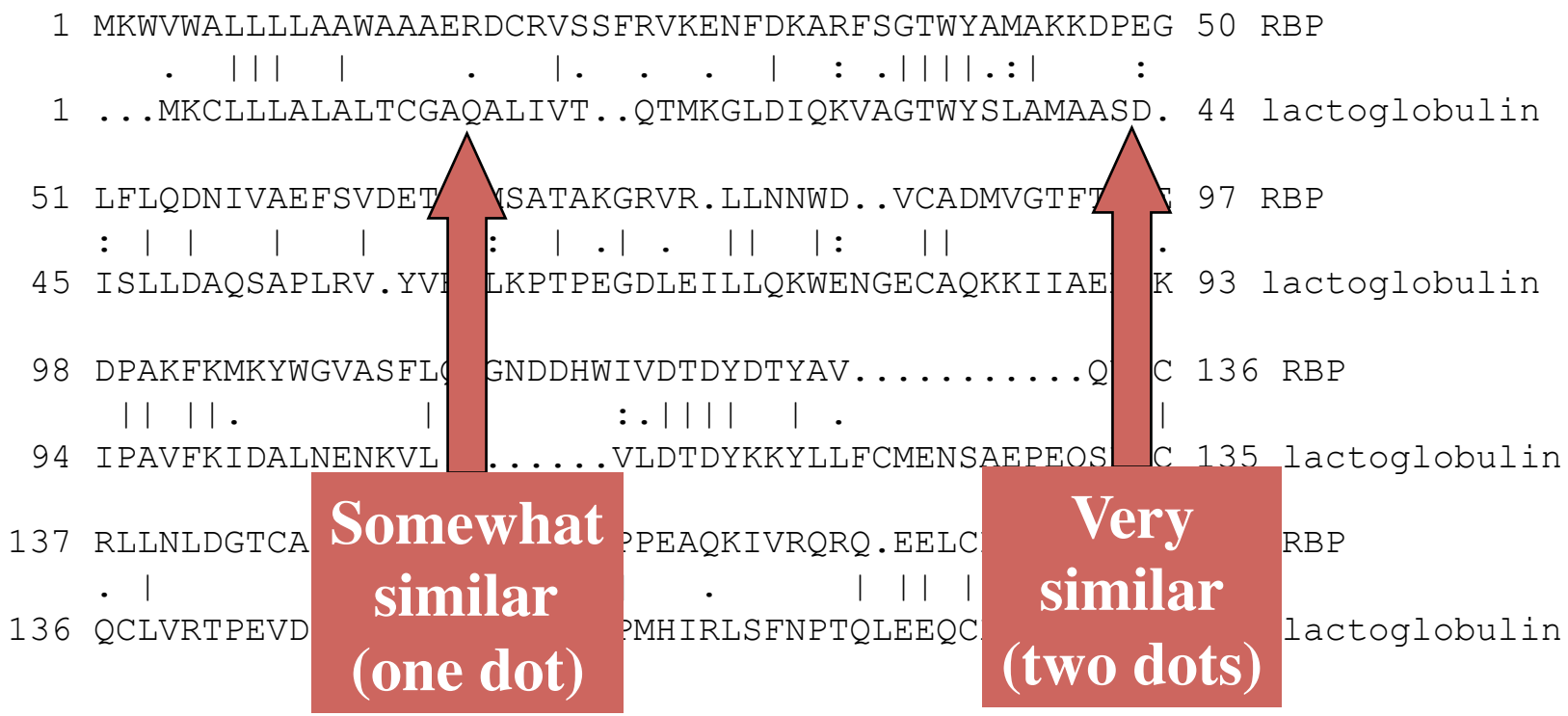
Given two strings $x = x_1x_2...x_M$, $y = y_1y_2...y_N$,

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 β-lactoglobulin



Pairwise alignment of retinol-binding protein and β-lactoglobulin



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.

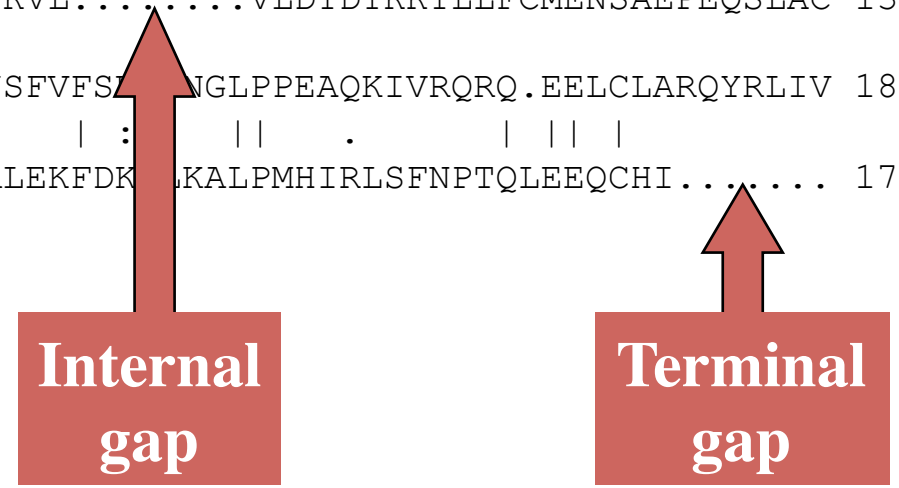
Pairwise alignment of retinol-binding protein and β-lactoglobulin

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

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

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV..........QYSC 136 RBP
     || ||.        |        :.||||  | .             .|
 94 IPAVFKIDALNENKVL.....VLDTDYKKYLLFCMENSAEPEQSLAC 135 lactoglobulin

137 RLLNLDGTCADSYSFVFS  NGLPPEAQKIVRQRQ.EELCLARQYRLIV 185 RBP
    . |       |     | :    ||   .      | || |
136 QCLVRTPEVDDEALEKFDK KALPMHIRLSFNPTQLEEQCHI..... 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 β-lactoglobulin

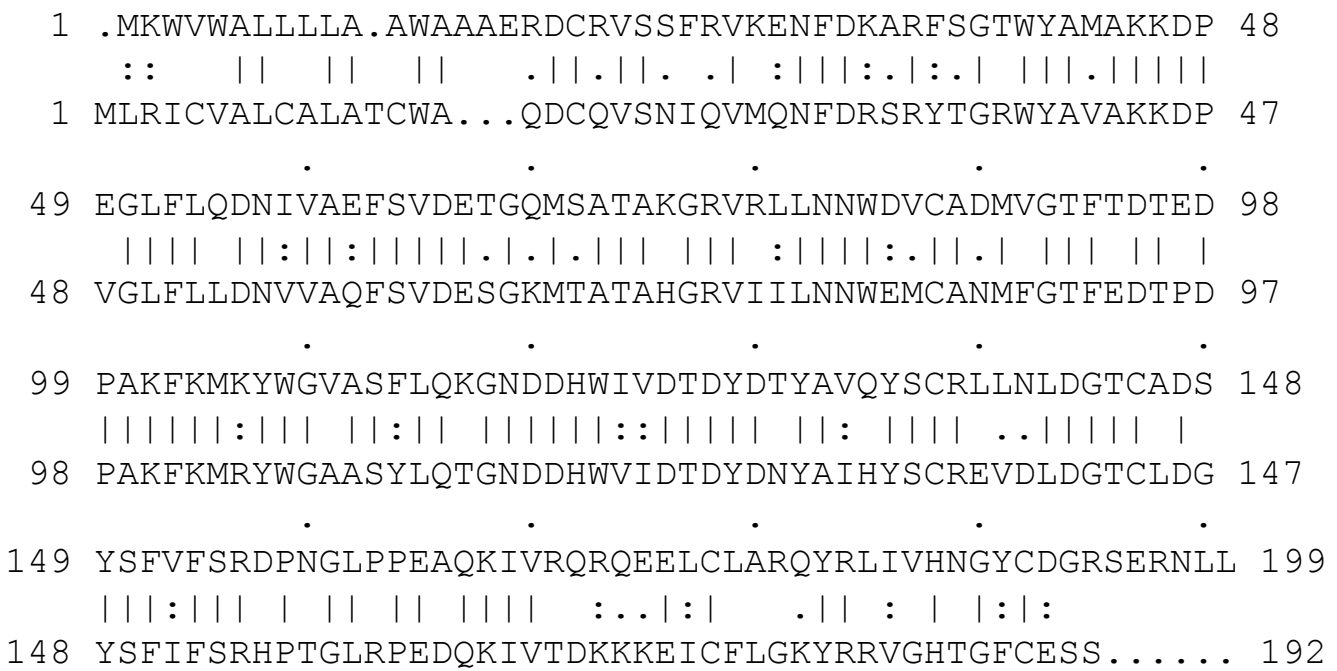
```
  1 MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEG 50 RBP
     .  |||  |     .   |.  .  .  |  : .||||.:|    :
  1 ...MKCLLLALALTCGAQALIVT..QTMKGLDIQKVAGTWYSLAMAASD. 44 lactoglobulin

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

 98 DPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAV...........QYSC 136 RBP
     || ||.        |        :.||||  | .             .|
 94 IPAVFKIDALNENKVL........VLDTDYKKYLLFCMENSAEPEQSLAC 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*)



Significant percent identity

Burkhard Rost: if more than 30% identisch, likely that proteins have similar structure!

- 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 ≠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        AATCTATA        AATCTATA
AAGATA           AAGATA           AAGATA
```

- Which one is better?

What is a good alignment?

AGGCTAGTT, AGCGAAGTTT

```
AGGCTAGTT-
AGCGAAGTTT
```

6 matches, 3 mismatches, 1 gap

```
AGGCTA-GTT-
AG-CGAAGTTT
```

7 matches, 1 mismatch, 3 gaps

```
AGGC-TA-GTT-
AG-CG-AAGTTT
```

7 matches, 0 mismatches, 5 gaps

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	**AATCTATA**	**AATCTATA**
AAGATA	**AAGATA**	**AAGATA**
4	1	3

Matches do not have to identical - similarity

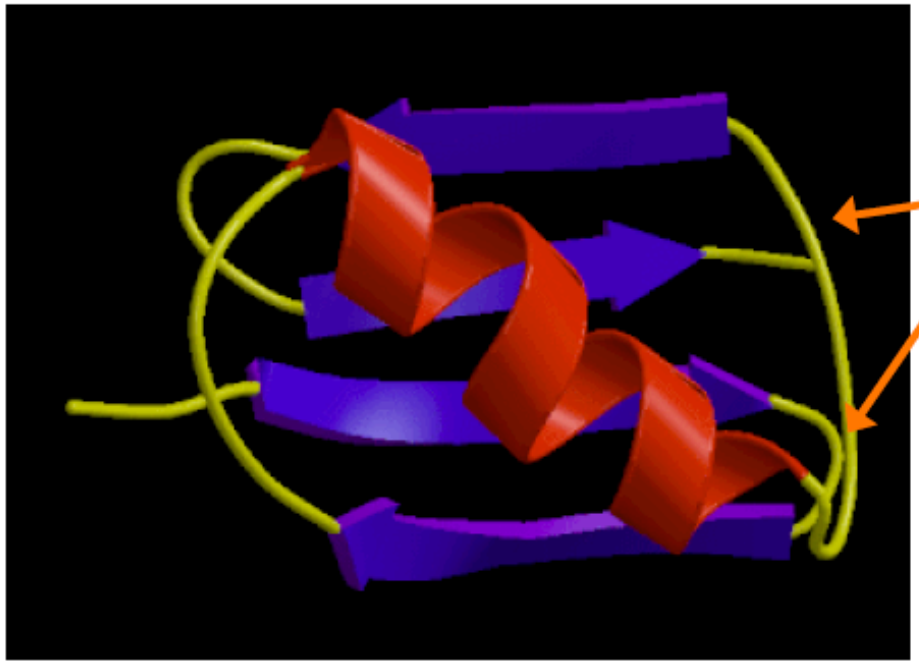
- Some amino acids resemble each other and can substitute functionally for each other.
  - Identical - highest score
  - Similar - high score (I, L)
  - Different - lowest score (I, K)
- Get overall alignment score.

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

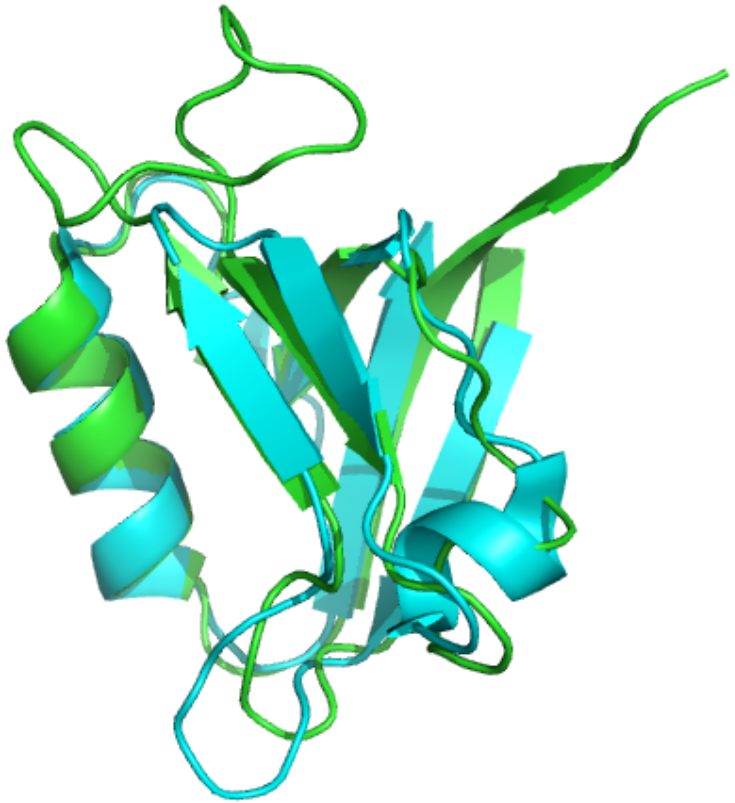
Insertions/Deletions and Protein Structure

- 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

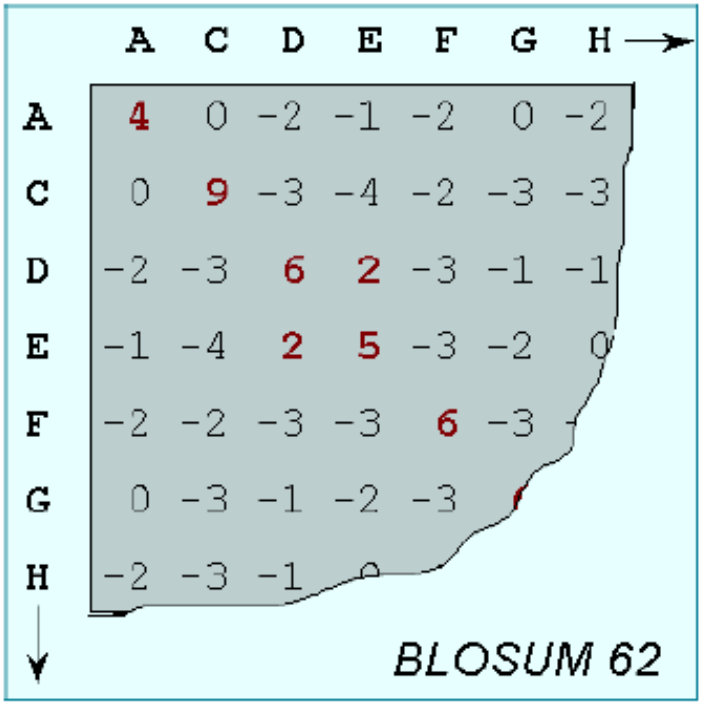
```
VAHV---D--DMPNALSALSDLHAHKL
AIQLQVTGVVVTDATLKNLGSVHVSKG
```

- we would score it by

  s(V,A) + s(A,I) + s(H,Q) + s(V,L) + 3*g* + s(D,G) + 2g …

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.



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

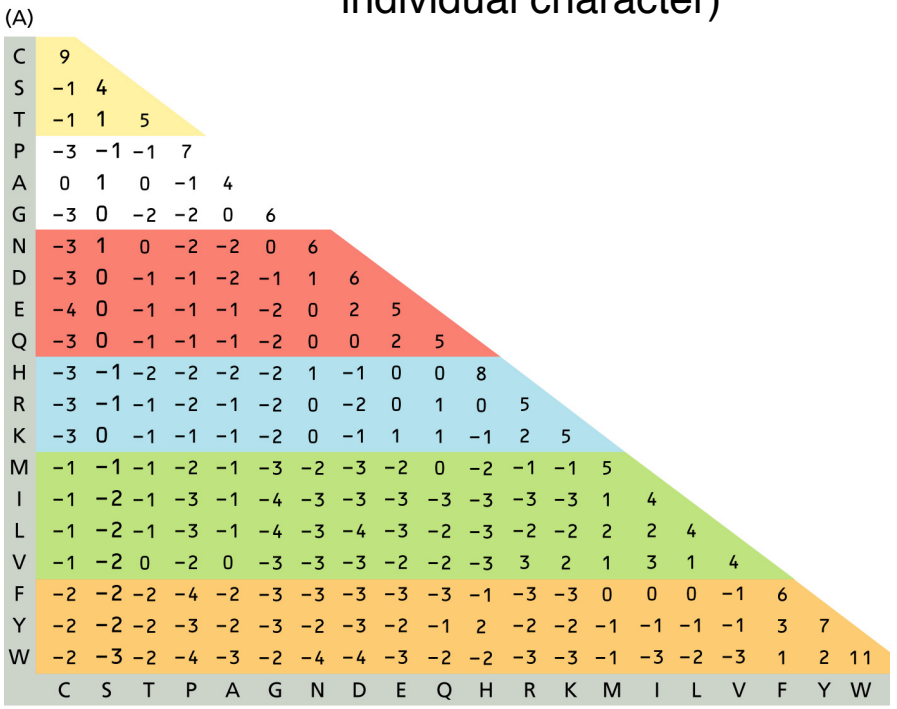
not all substitutions are as costly! For example Cis-Cis quite high, because important to form disulfide bonds! In addition tryptophane is the most costly because evolutionary & energetically costly (very individual character)

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

Overall score: 52 (Blosum)



BLOSSUM

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 arg ser val

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

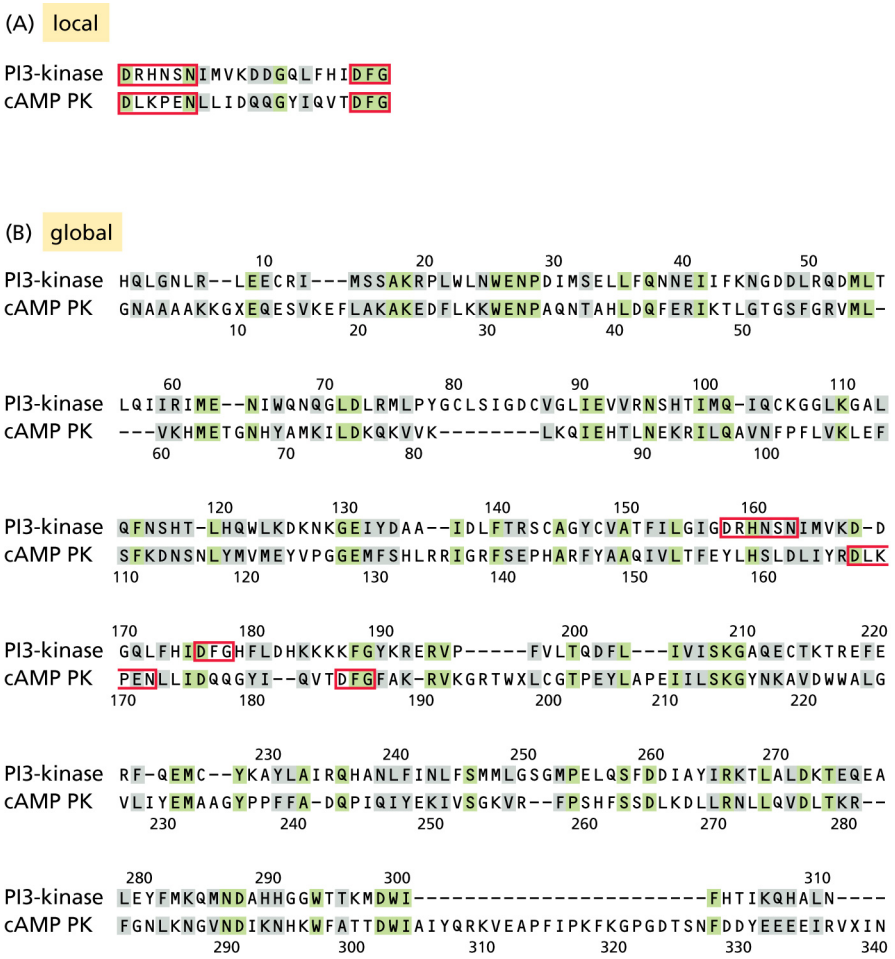
Two kinds of sequence alignment: global and local

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

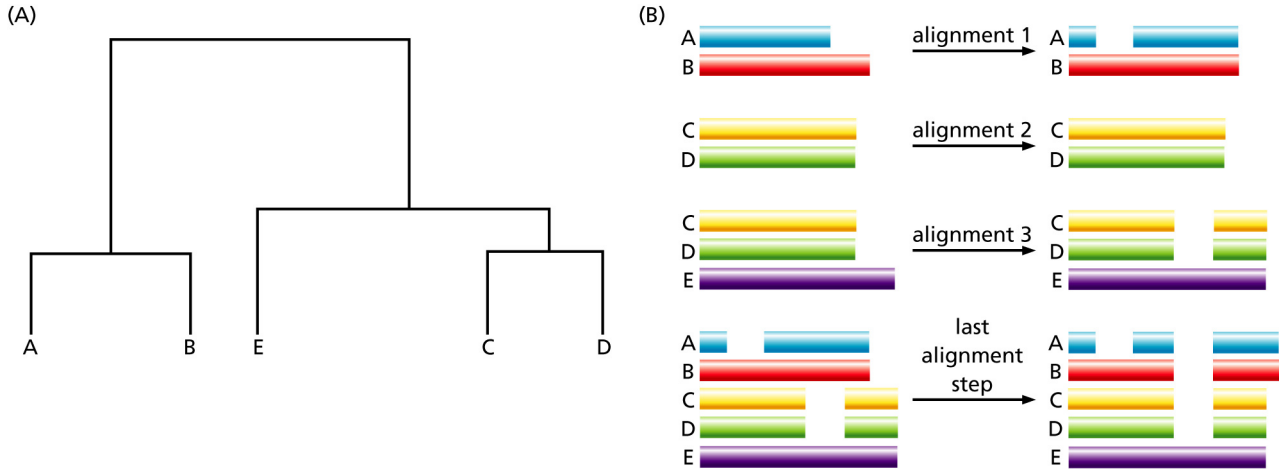


Needleman-Wunsch
(global alignment)

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

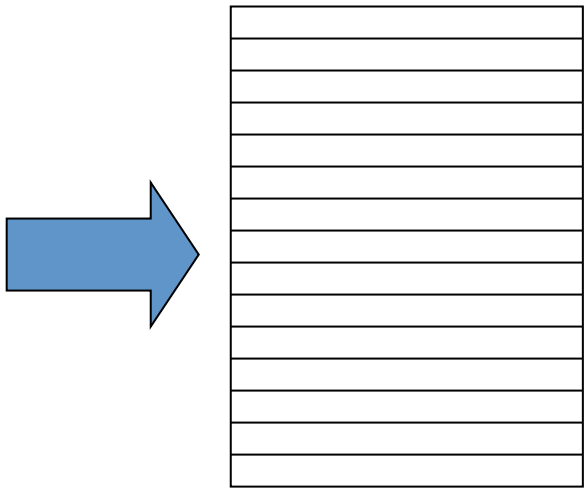
Multiple alignment

Improves accuracy of alignment for sequences of low similarity

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

ACTTTTGGTGACTGTAC



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 un-gapped (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 `IHAVEADREAM`

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 aligments 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.

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

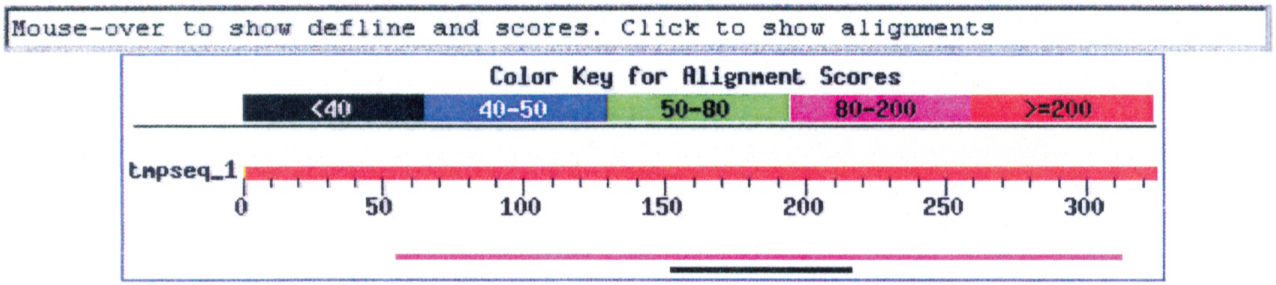
BLAST output

(A)

sp|P32871|P11A_BOVIN	PHOSPHATIDYLINOSITOL 3-KINASE CATALYTI...	680	0.0
sp|P42336|P11A_HUMAN	PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...	676	0.0
sp|P42337|P11A_MOUSE	PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...	674	0.0
sp|P42338|P11B_HUMAN	PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...	338	9e-93
sp|O35904|P11D_MOUSE	PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...	332	7e-91
sp|O00329|P11D_HUMAN	PHOSPHATIDYLINOSITOL 3-KINASE CATALYT...	331	2e-90
sp|P47473|RIR1_MYCGE	RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE A...	34	0.59

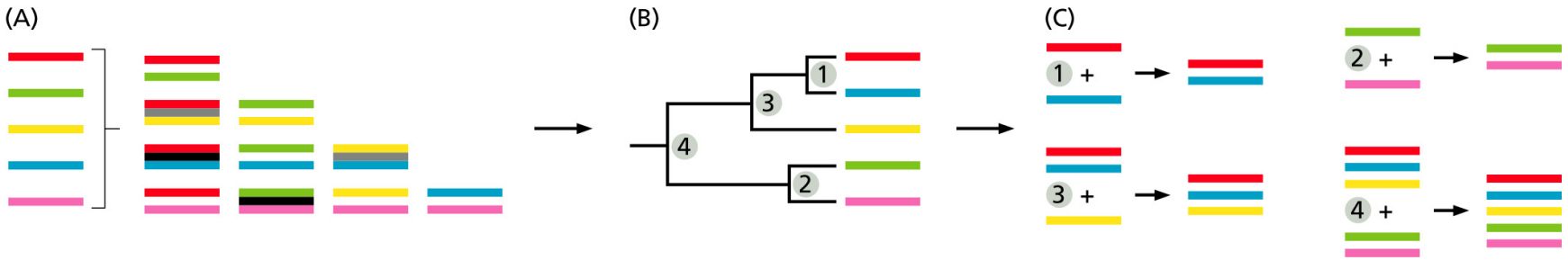
(B)

Distribution of 2 Blast Hits on the Query Sequence



dbj|BAB10275.1|	(AB008266) phosphatidylinositol 4-kinase [A...	111	3e-25
dbj|BAB11344.1|	(AB011477) AtRAD3 [Arabidopsis thaliana]	38	0.008

Progressive alignment



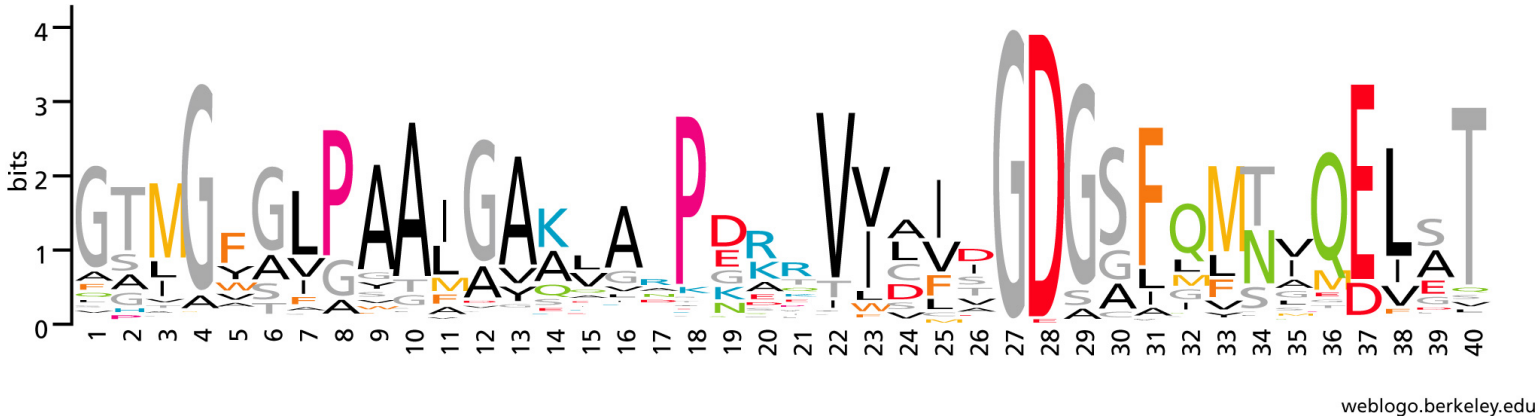
Find seq
most similar
to each other

Build guide tree

Create intmdt ali

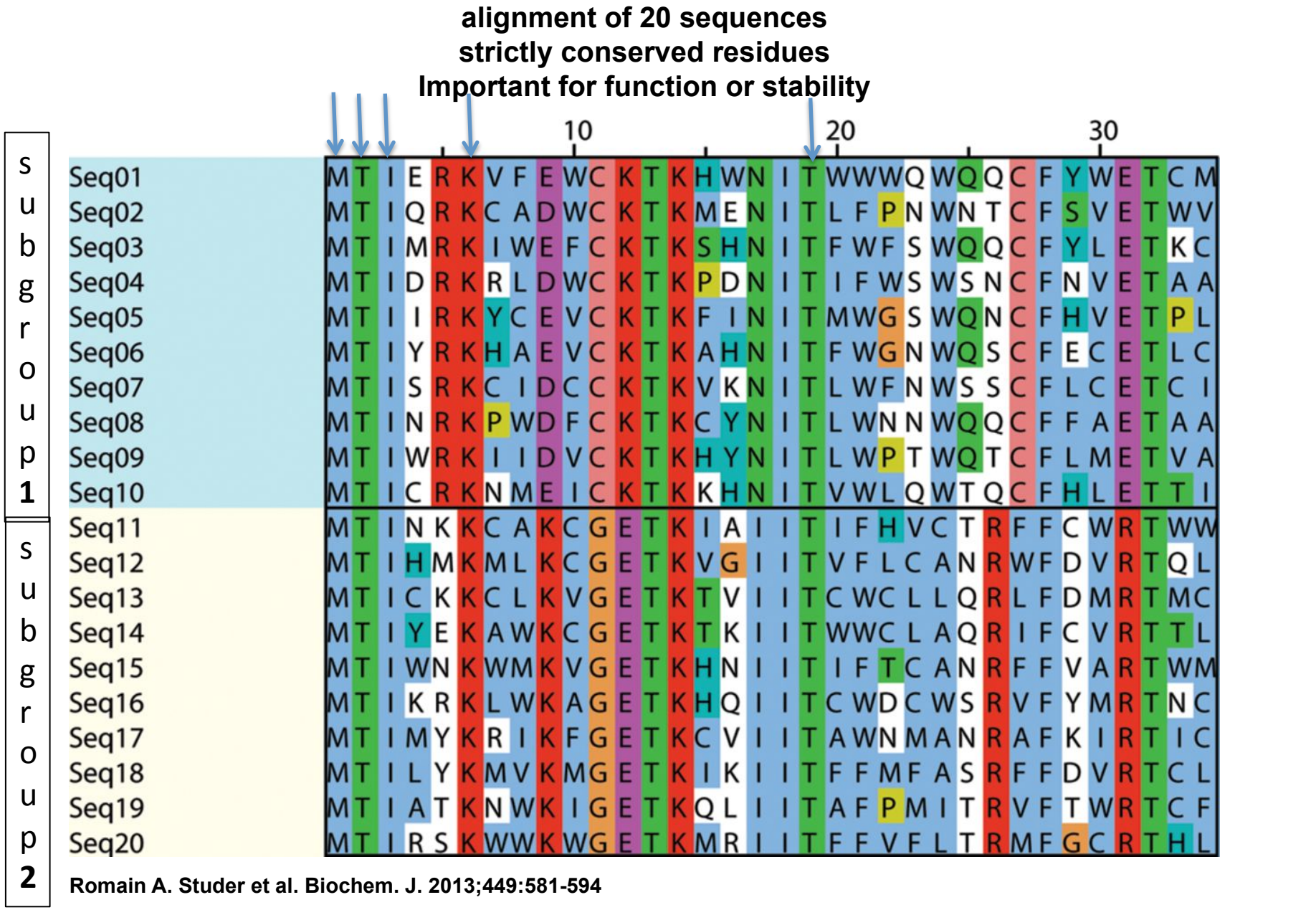
Representing a profile as a logo

- 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]

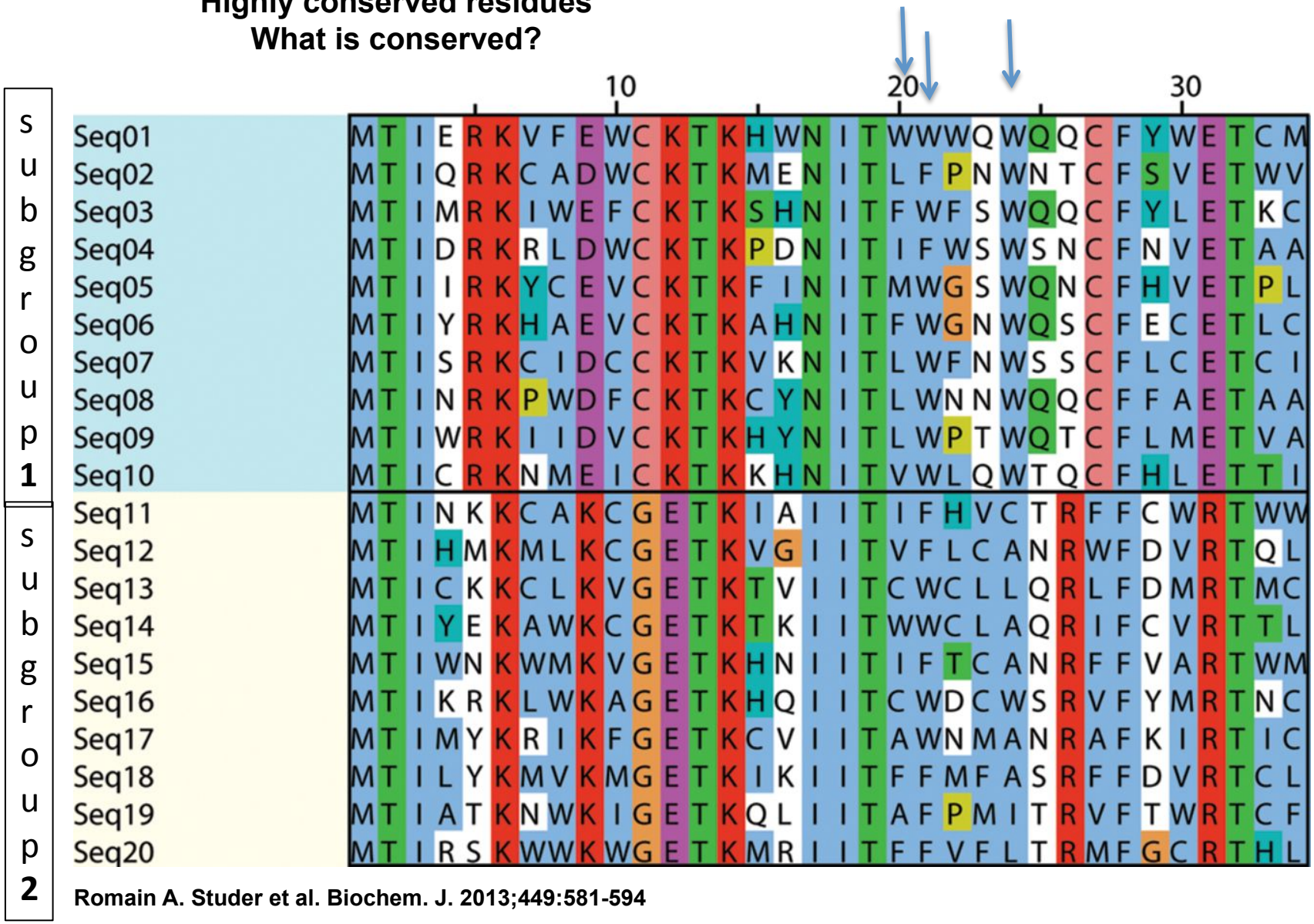


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green
His – turquoise, Pro – yellow, Gly – orange, Cys - pink

PORTLAND PRESS

alignment of 20 sequences
Highly conserved residues
What is conserved?



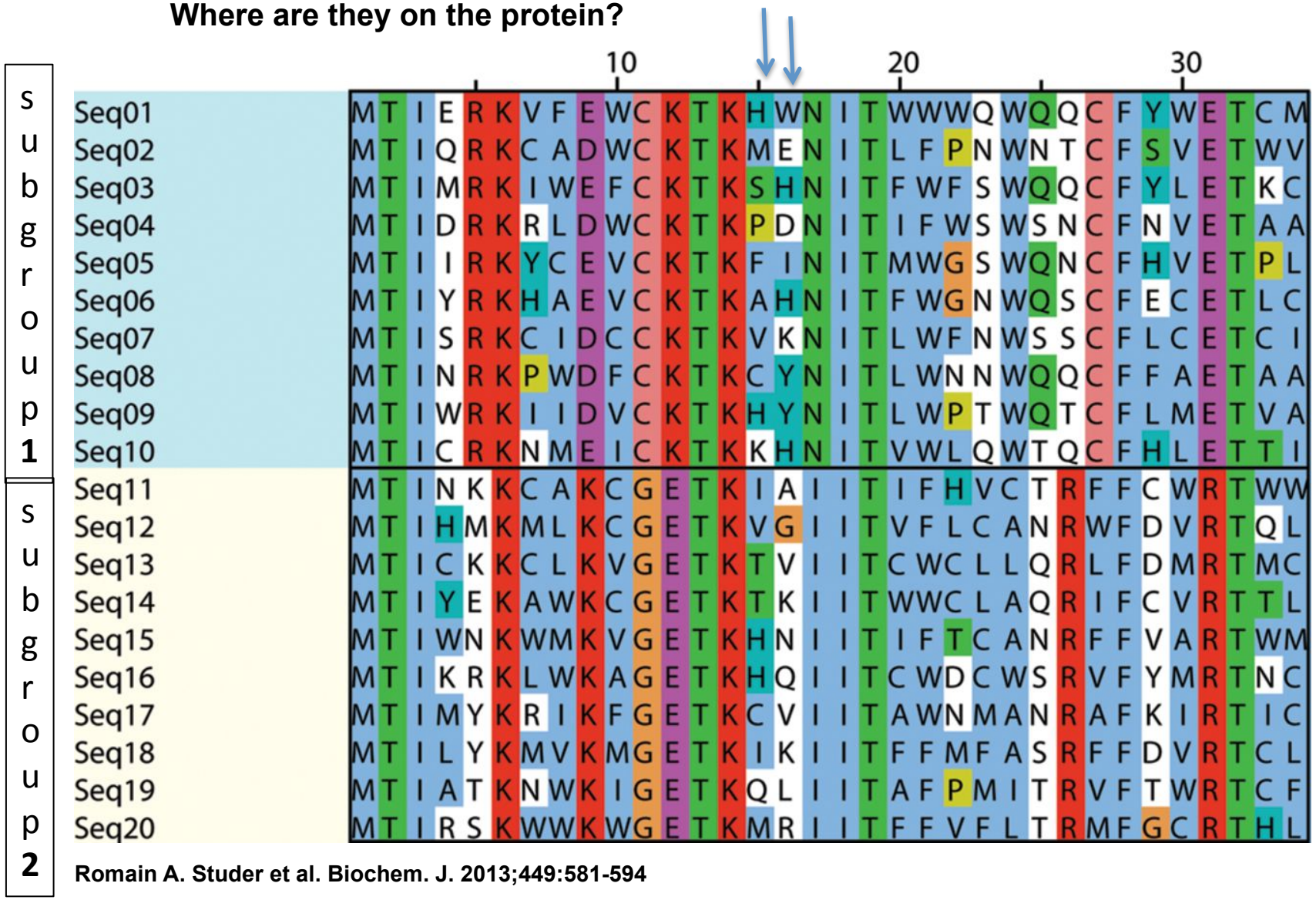
Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green
His – turquoise, Pro – yellow, Gly – orange, Cys - pink

PORTLAND PRESS

can be changed into whatever… probably on surface, since “easier” to change

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

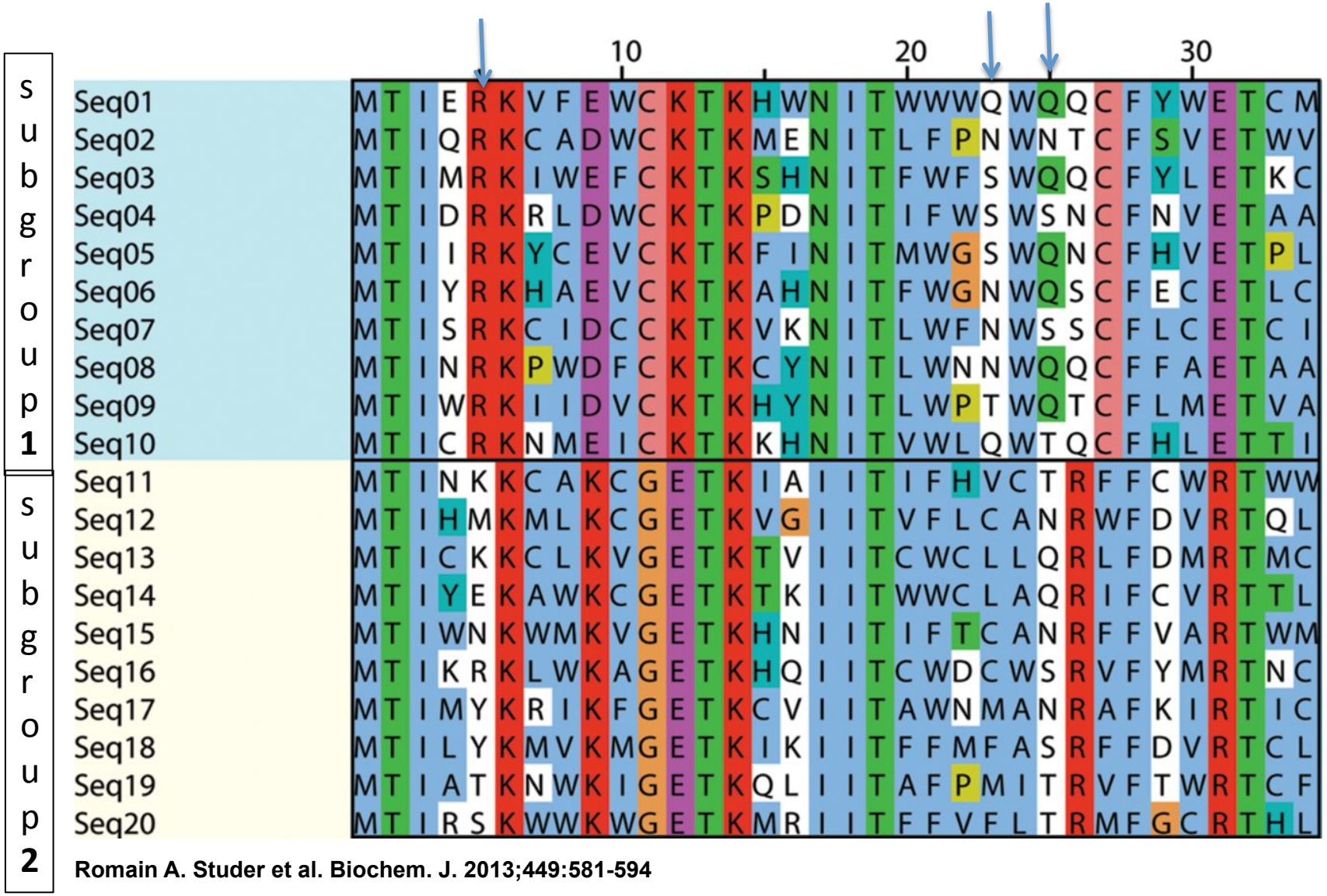


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green
His – turquoise, Pro – yellow, Gly – orange, Cys - pink

PORTLAND PRESS

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
His – turquoise, Pro – yellow, Gly – orange, Cys – pink.

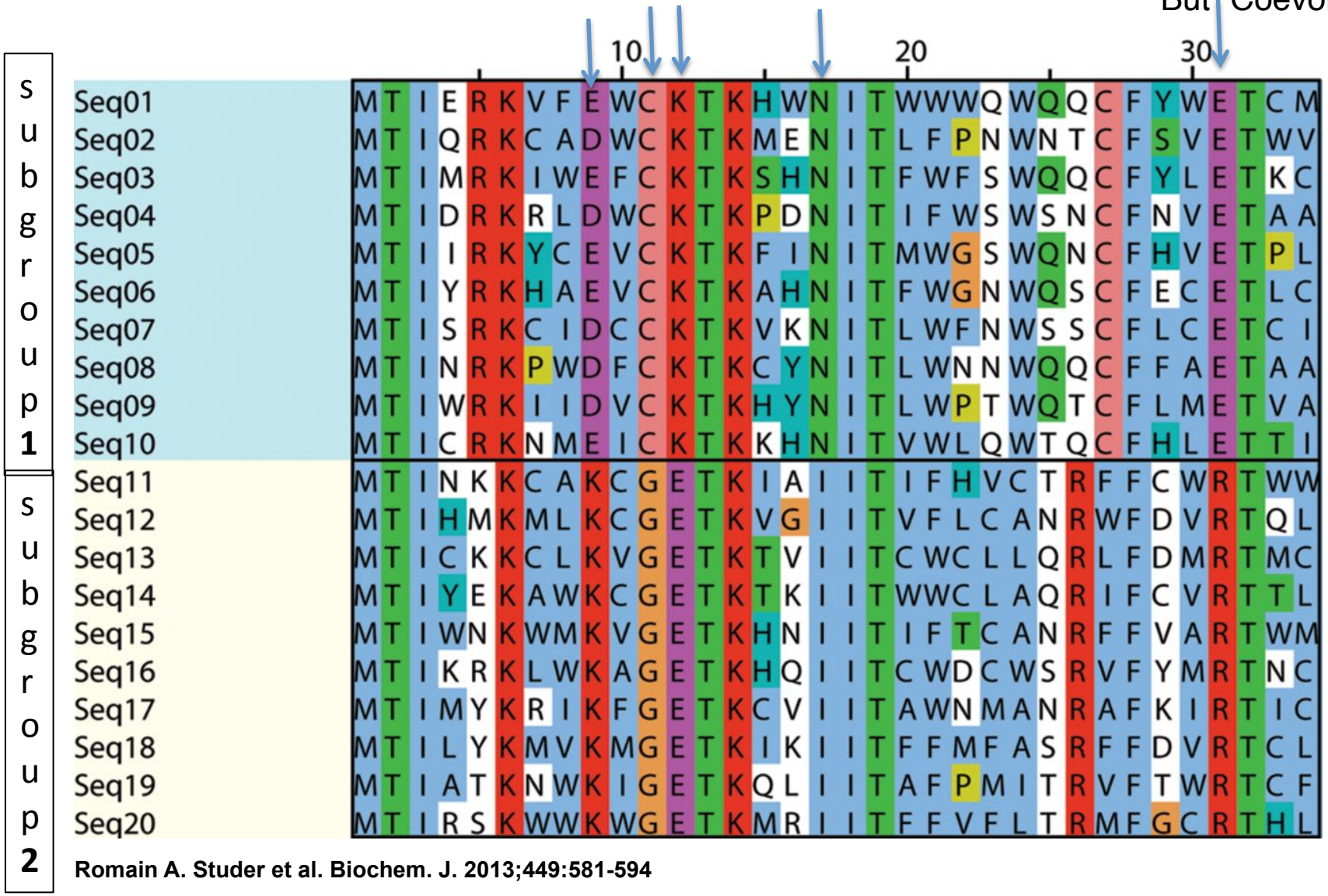
PORTLAND PRESS

replacement of E and D is acceptable, since both are acidic!

But K with E change unlikely, since K is negatively and E is positively charged!

But "Coevolution" possible!

alignment of 20 sequences
Physicochemical properties

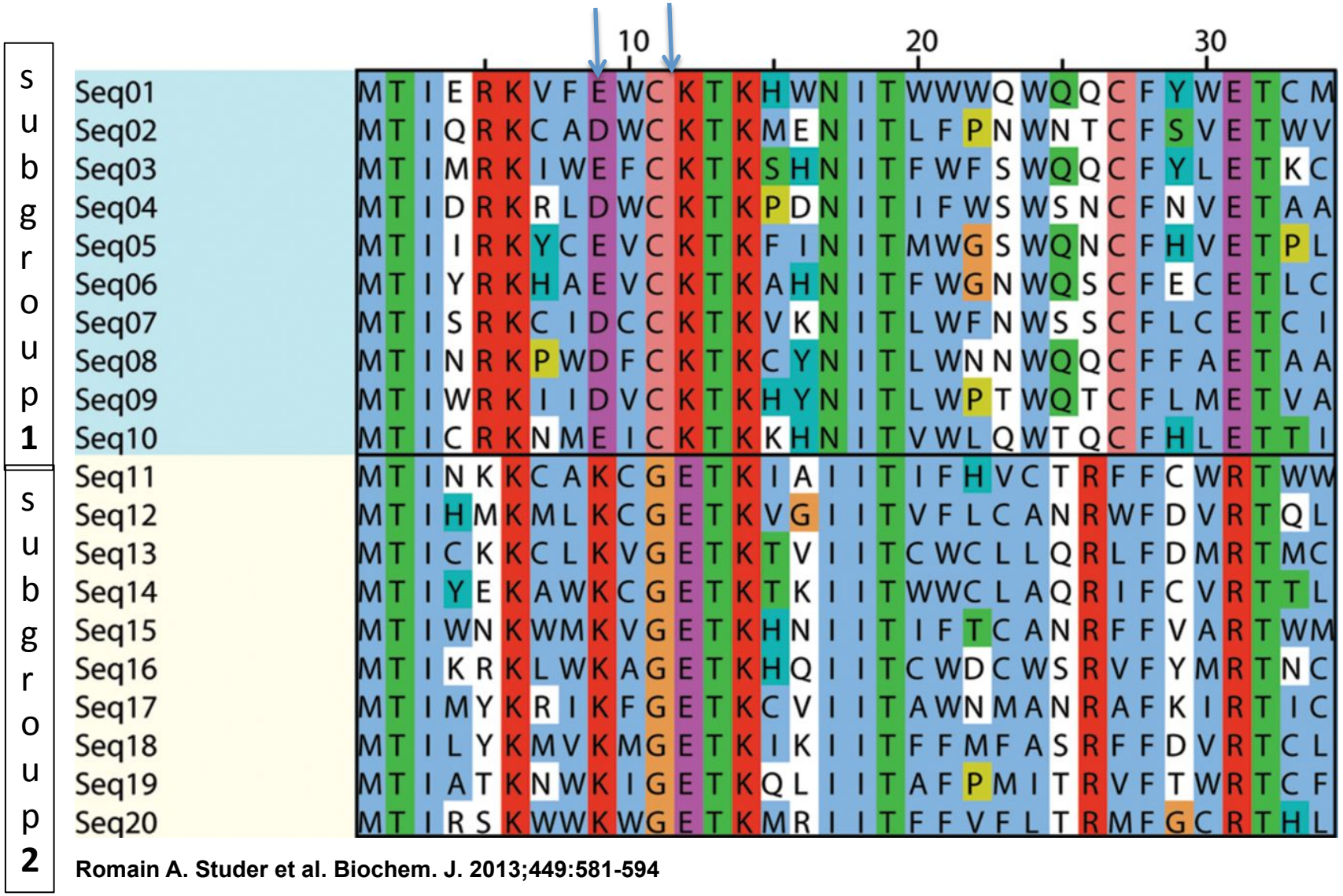


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green
His – turquoise, Pro – yellow, Gly – orange, Cys – pink.

PORTLAND PRESS

alignment of 20 sequences
Coevolution

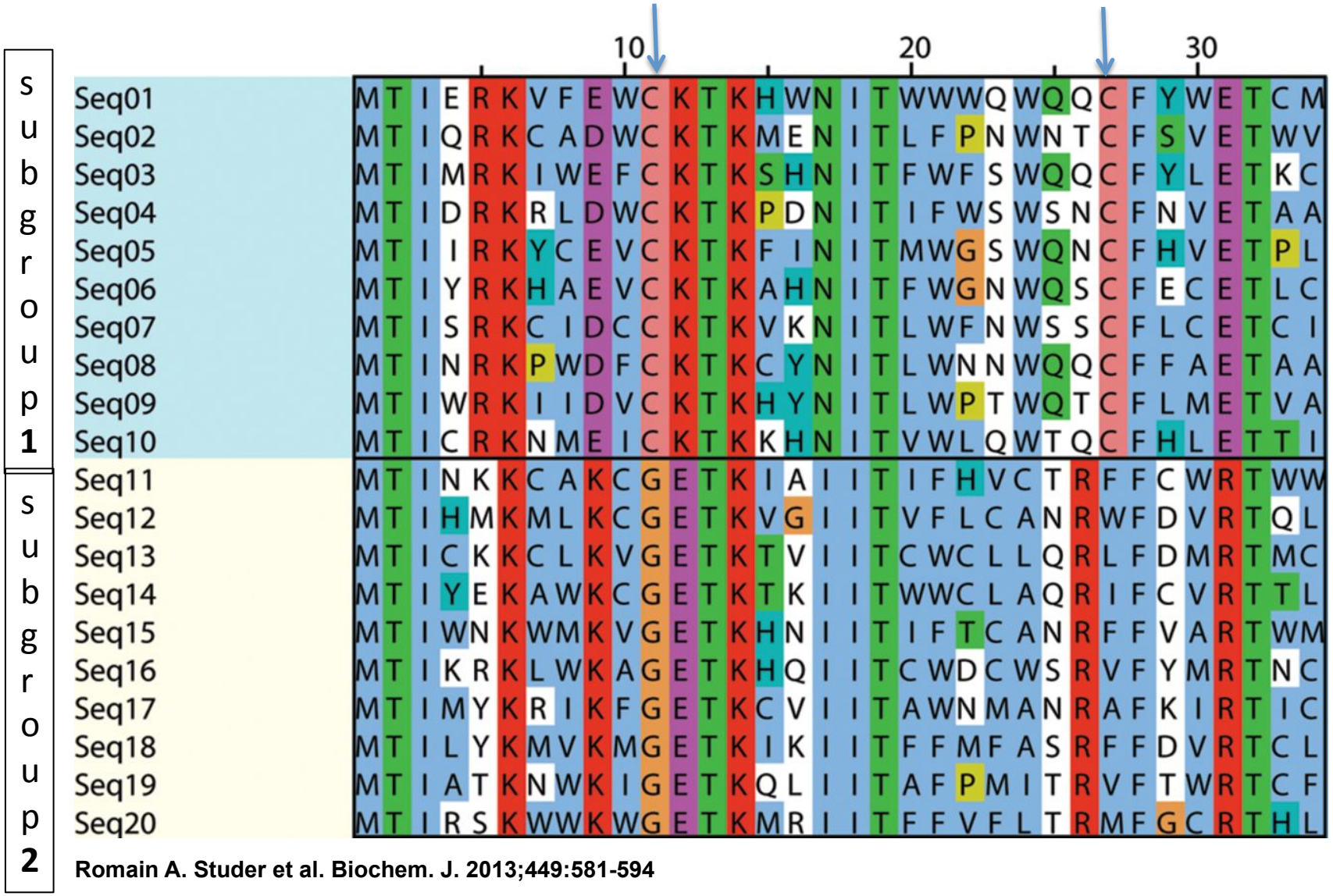


Romain A. Studer et al. Biochem. J. 2013;449:581-594

Hydrophobic – blue, basic – red, acidic – purple, polar – green
His – turquoise, Pro – yellow, Gly – orange, Cys – pink.

PORTLAND PRESS

alignment of 20 sequences
Coevolution

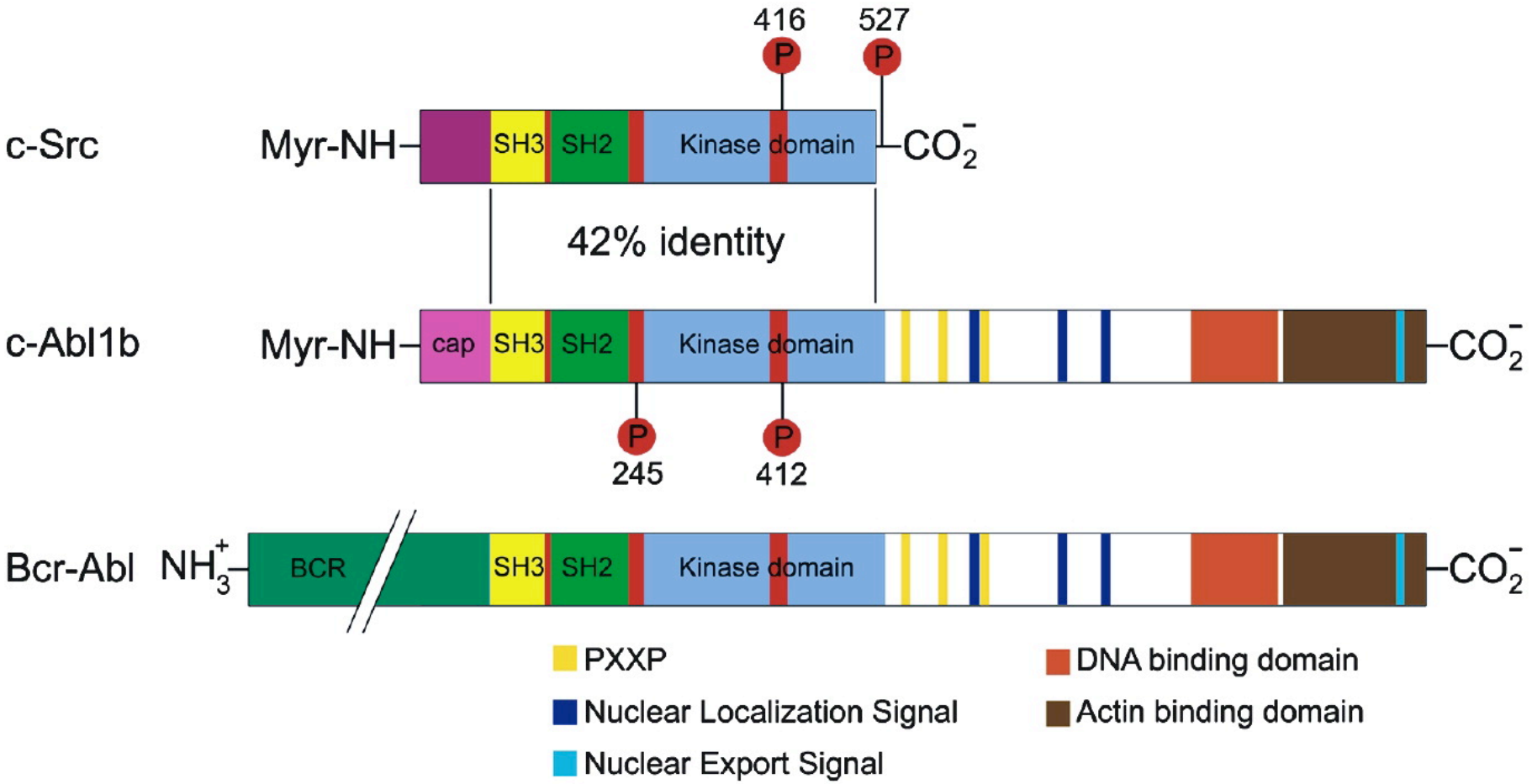


Romain A. Studer et al. Biochem. J. 2013;449:581-594

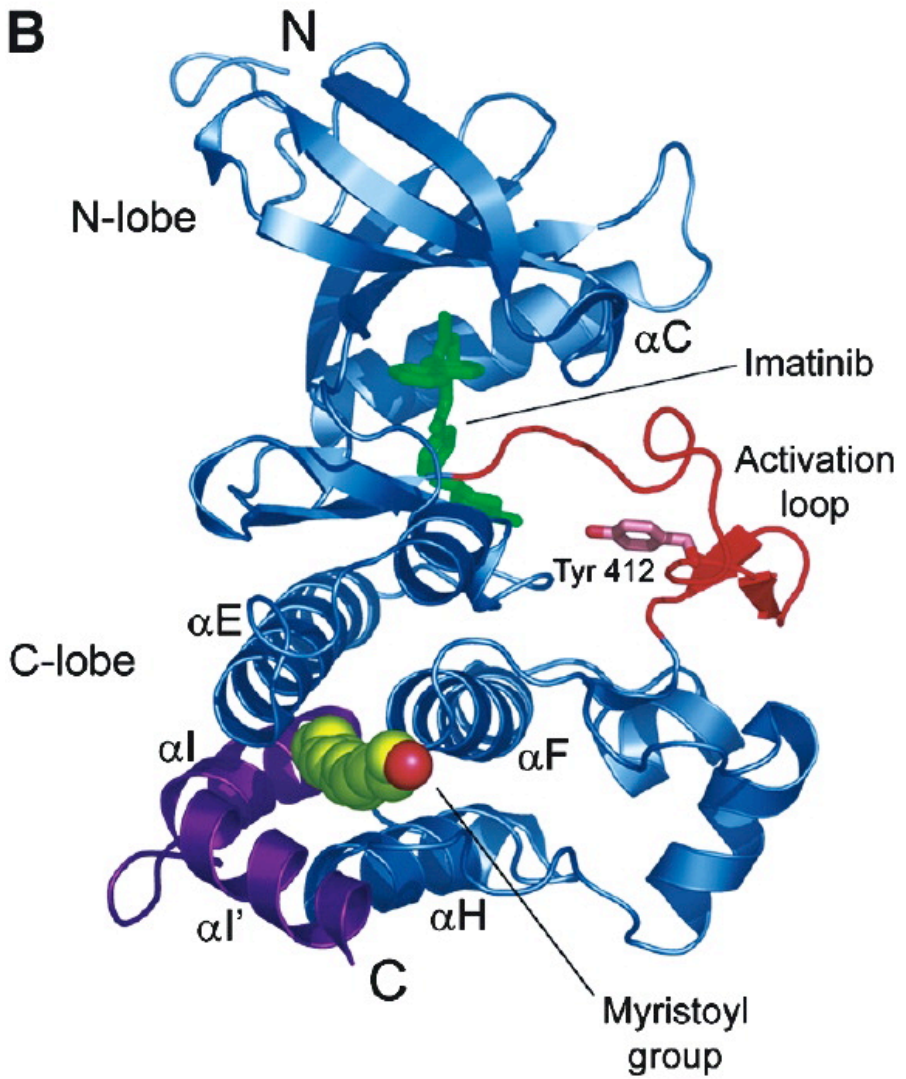
Hydrophobic – blue, basic – red, acidic – purple, polar – green
His – turquoise, Pro – yellow, Gly – orange, Cys – pink.

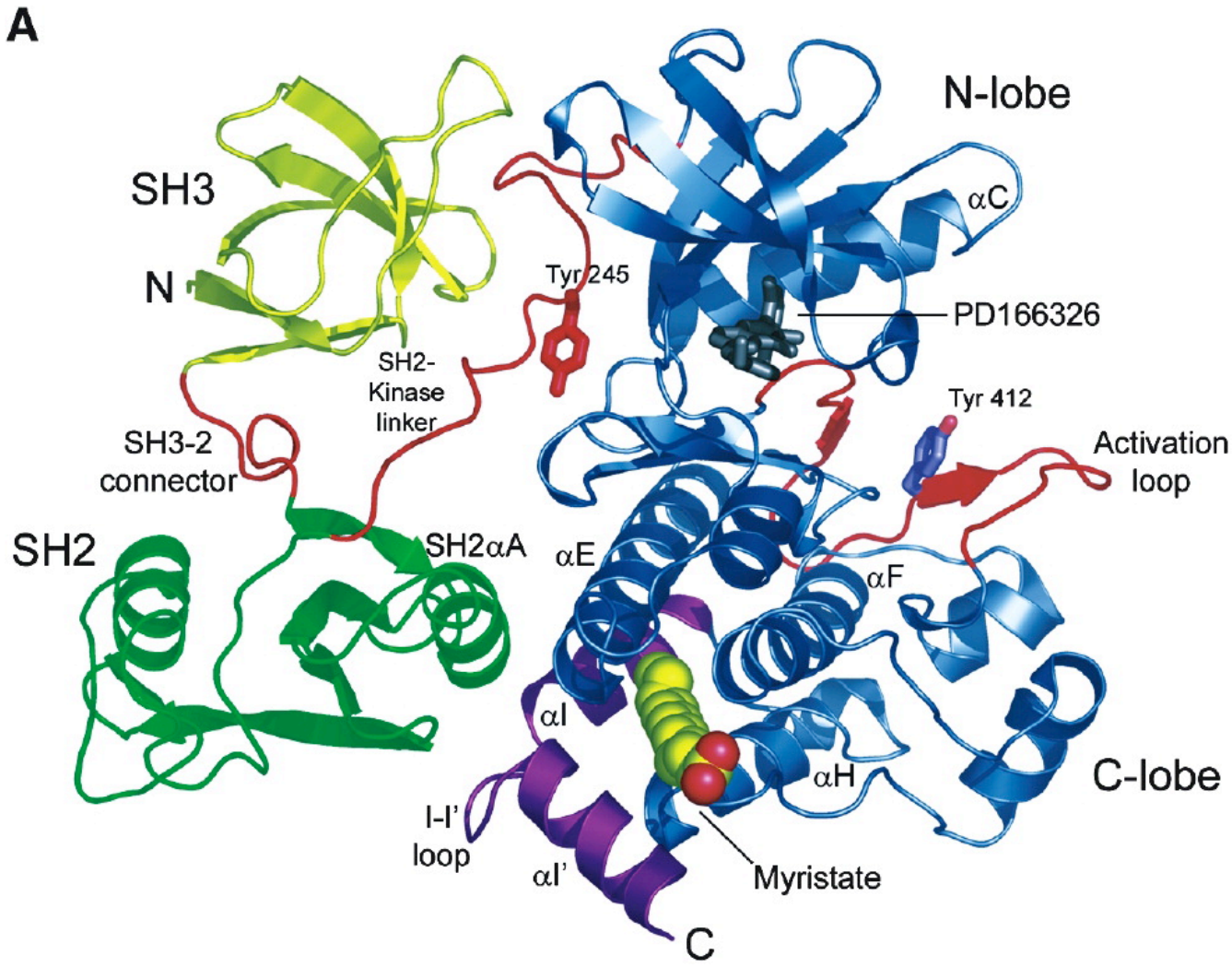
PORTLAND PRESS

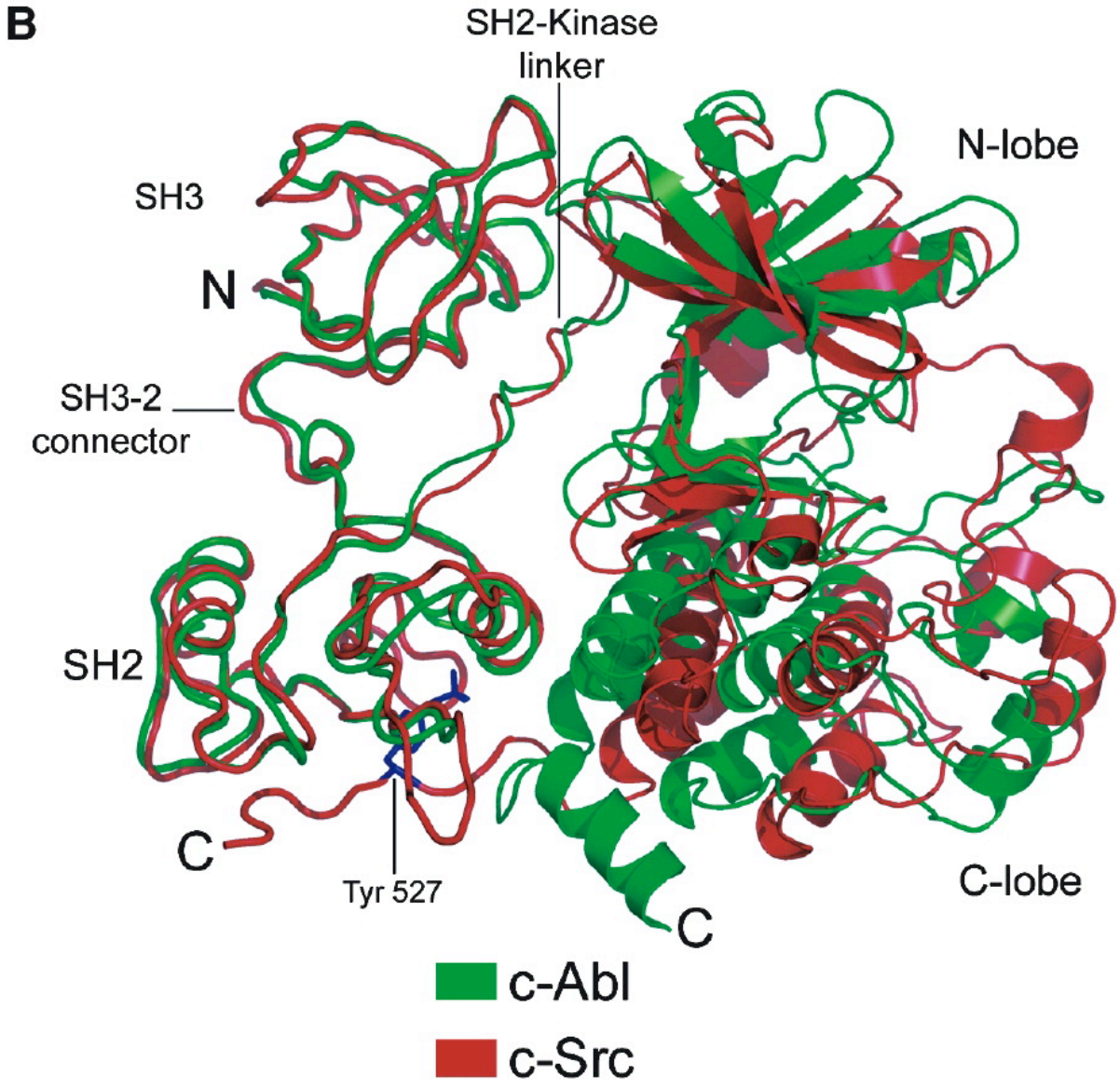
Abl kinase


416
527
P
P
c-Src
Myr-NH
SH3
SH2
Kinase domain
CO_2^-
42% identity
c-Abl1b
Myr-NH
cap
SH3
SH2
Kinase domain
CO_2^-
P
P
245
412
Bcr-Abl
NH_3^+
BCR
SH3
SH2
Kinase domain
CO_2^-
PXXP
Nuclear Localization Signal
Nuclear Export Signal
DNA binding domain
Actin binding domain

Kinases in general have the same shape!




A
SH3
N
SH2-
Kinase
linker
Tyr 245
N-lobe
αC
PD166326
Tyr 412
Activation
loop
SH3-2
connector
SH2
SH2αA
αE
αF
αI
αH
C-lobe
I-I'
loop
αI'
Myristate
C


B
SH2-Kinase
linker
N-lobe
SH3
N
SH3-2
connector
SH2
C
Tyr 527
C
C-lobe
c-Abl
c-Src

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.