Introduction and Basic Concepts

Laboratory of Bioinformatics I
Module 2

March 12, 2019

Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Department of Pharmacy and Biotechnology (FaBiT)
University of Bologna

Schedule and Materials

This module is a 60-hour course running for 6 weeks
March 12 - April 17, 2019

Schedule changes from week to week:

Tuesday, Thursday and Friday 14:00 - 17:00

In April more changes

Project submission deadline May 7, 2018

Course website

http://biofold.github.io/pages/courses/2019/lb1-2.html

Main Aims

- Knowledge of tools for sequence and structure analysis and their development

- Protein functional annotation

- Theoretical background of machine learning approaches

- Problem solving skills and development of basic tools.

Topics

- Protein Geometrical Features and Protein Structural Alignment
- Multiple Sequence Alignment
- Hidden Markov Models for Sequence Alignment
- Methods for Building Hidden Markov Models for Proteins
- Protein Structure and Mapping Problems
- Introduction to Statistical Methods and Machine Learning
- Development of Structure Prediction Methods

- Module Project: Model a Protein Domain HMM

Take Home Message

- Protein structure is more conserved than sequence. Proteins sharing high sequence identity usually share similar structures, as proven by pair-wise structural alignment procedures.

- When the identity level is high enough, it is possible to exploit the results of pair-wise sequence alignment for transferring structural information between proteins.

Structural Alignment

Given two sets of points $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots b_m)$ in Cartesian space, find the optimal subsets A(P) and B(Q) with |A(P)| = |B(Q)|, and find the optimal rigid body transformation G between the two subsets A(P) and B(Q) that minimizes a given distance metric D over all possible rigid body transformation G, i.e.

$$\min_{G}\{D[A(P) - G(B(Q))]\}$$

$$\mathbf{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(a_i - b_i)^2}{n}}$$

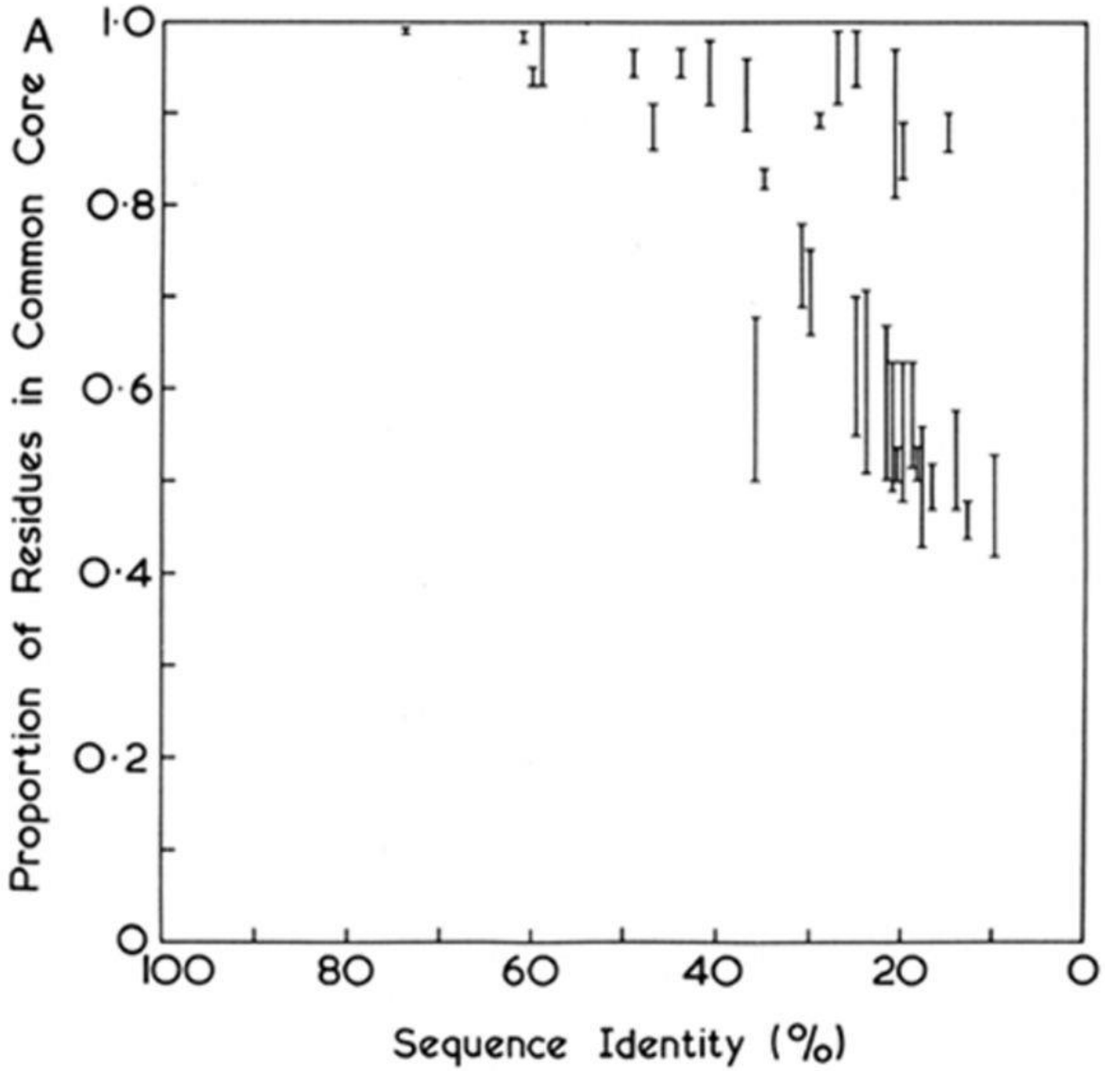
The two subsets A(P) and B(Q) define a "correspondence", and p = |A(P)| = |B(Q)| is called the correspondence length. Naturally, the correspondence length is maximal when A(P) and B(Q) are similar.

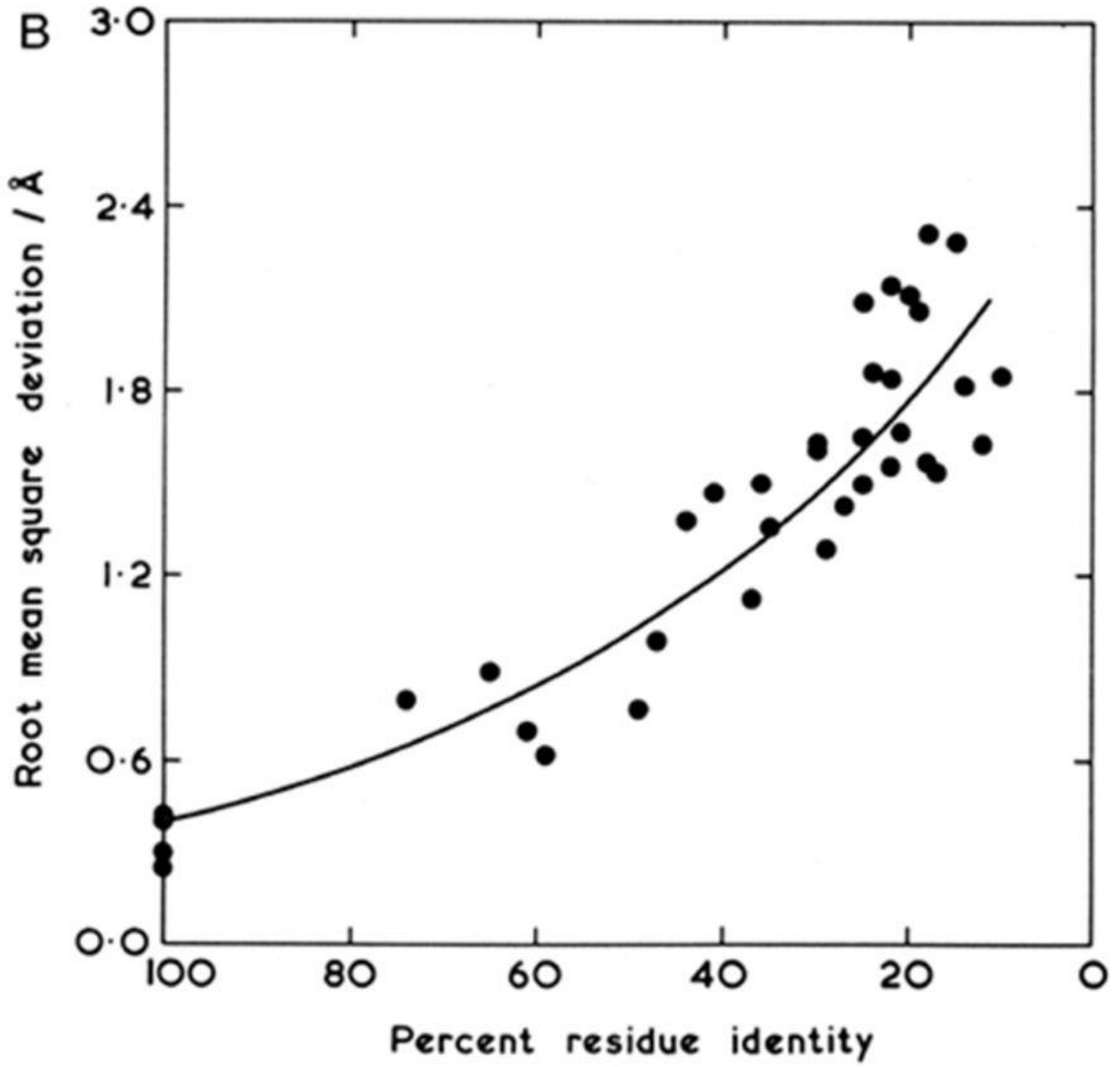
Therefore there are essentially two problems in structure alignment:

- Find the correspondence set (which is NP-hard), and
- Find the alignment transform (which is O(n)).

Bourne P. 2012

The Foundation of Structural Bioinformatics





Chotiha and Lesk, EMBO Journal 1986

Why Sequence Alignment?

The measure of sequence similarity allow to make estimation about the structural similarity

Comparison of two sequences for measuring their similarity

- To define a distance between two sequences
- Develop an algorithm for finding the alignment with minimal distance
- To statistically evaluate the significance of the alignment

Sequence Distance Score

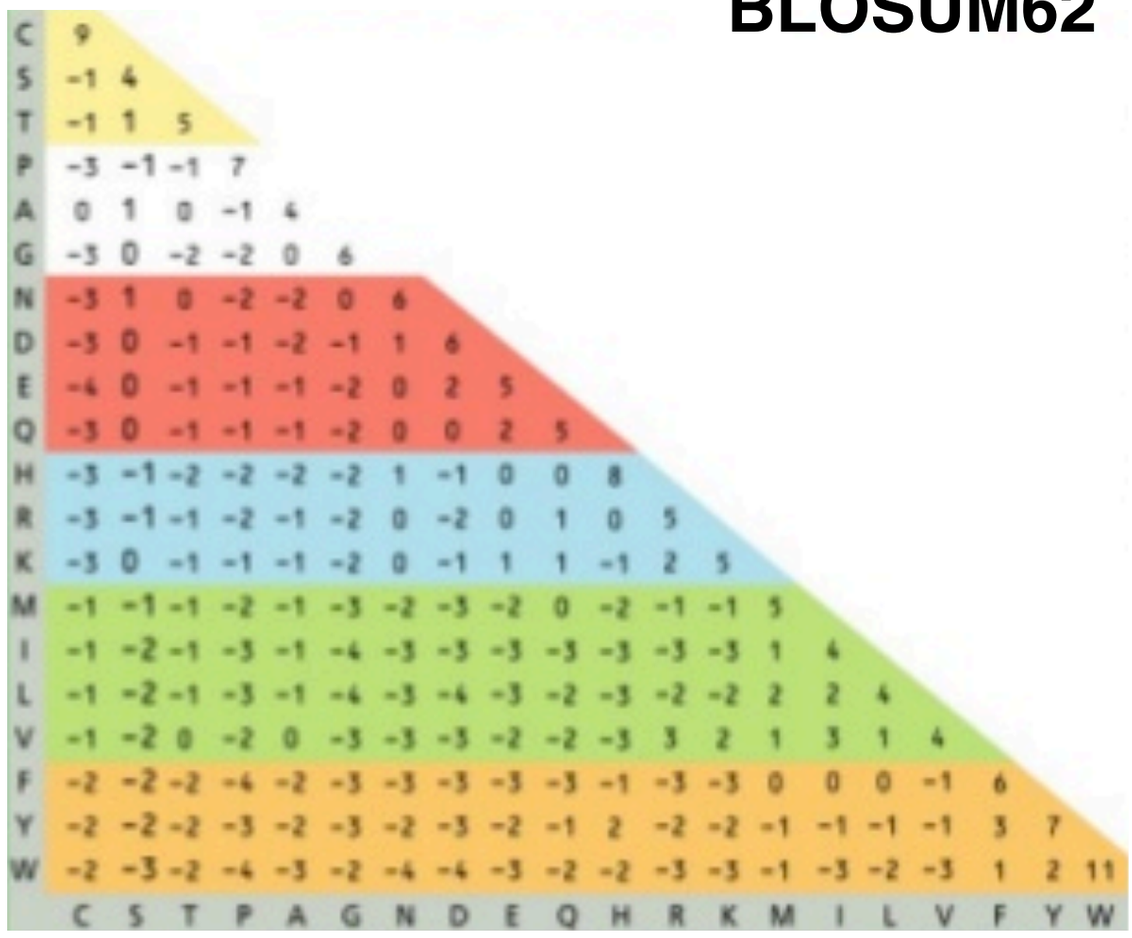
Which events do we consider?
Mutation

It is necessary to define a score for the substitution of residue i with residue j
Substitution Matrices *s(i,j)*

A: ALASVLIRLITRLYP
B: ASAVHLNRLITRLYP

$$Score(A,B) = \sum s(A^i, B^i)$$

BLOSUM62



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

Other events

Deletion and Insertion: some residues can be inserted or deleted during the evolution

A: ALASVLIRLIT--YP
B: ASAVHL---ITRLYP

$$Score(A,B) = \sum s(A^i, B^i) + \sigma(3) + \sigma(2)$$

The (negative) score of a gap depends only on the length

σ(n) = -nd linear
σ(n) = -d - (n-1) e (d: opening, e: extension)

Alignment Algorithms

Algorithms for finding the minimum distance between two sequences

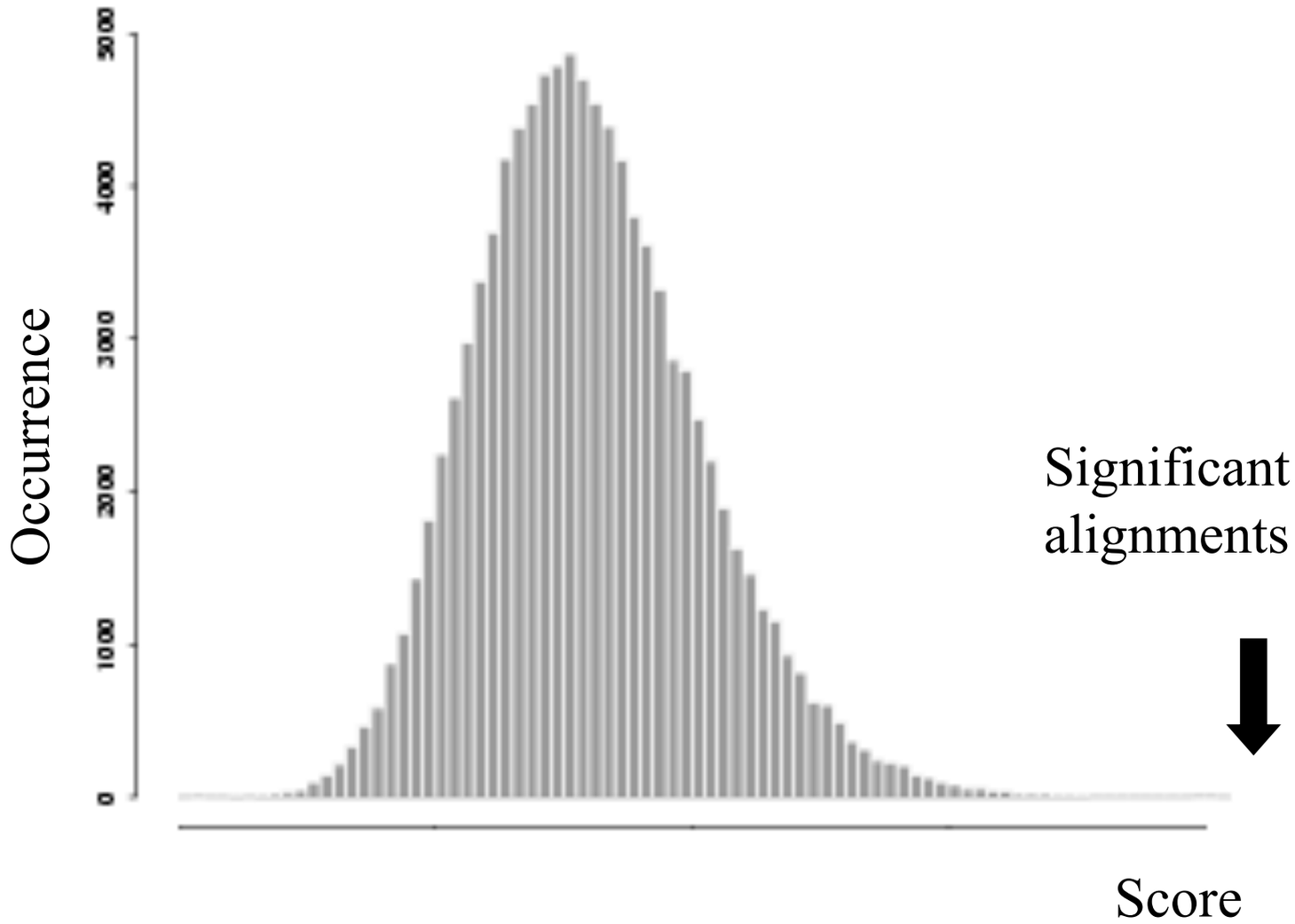
- Global alignment: Needleman-Wunsch: Global alignment-compare pairs of sequences on their whole length

- Local alignment: Smith-Waterman: Local alignment-compare pairs of sequences searching the most similar subsequences

Alignment Significance

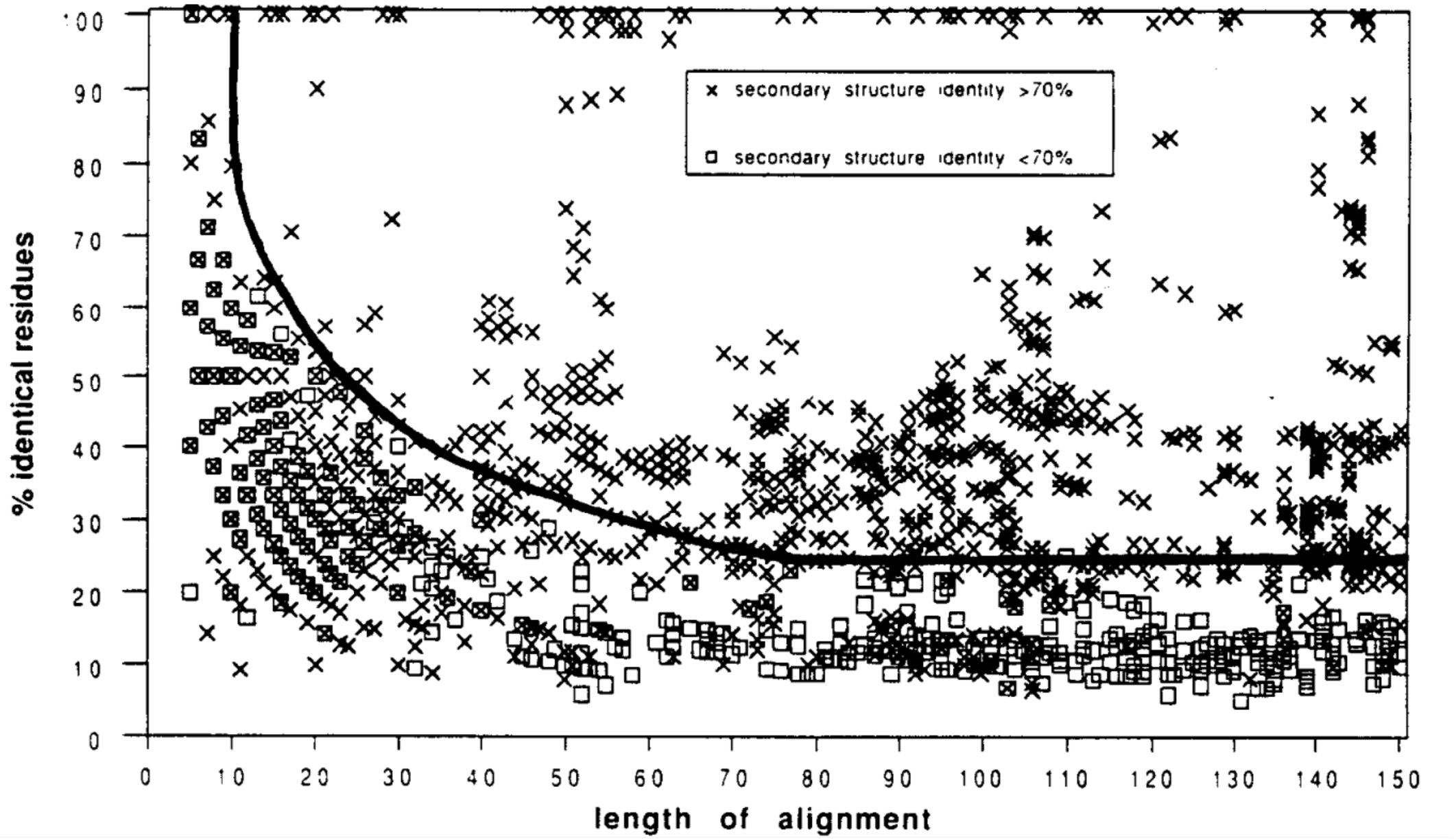
Given an alignment with score S, is it significant?

Significance can be evaluated by comparing with the score distribution of random alignments



Structural Homology

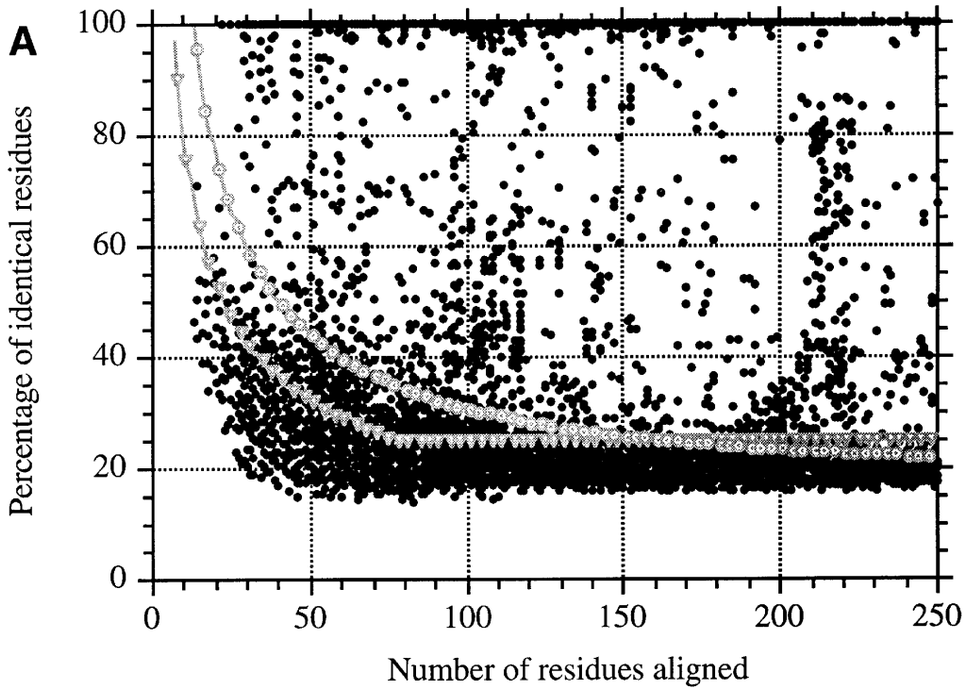
Based on the database of homology-derived secondary structure of proteins (HSSP).
Define the relation between sequence similarity, structure similarity, and alignment length.

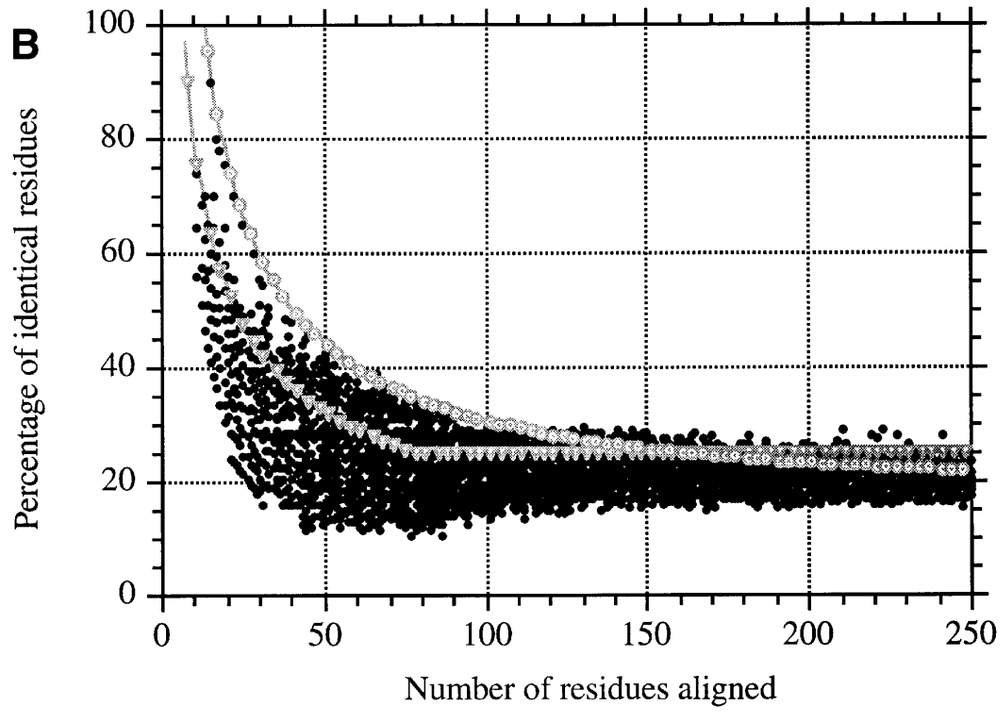


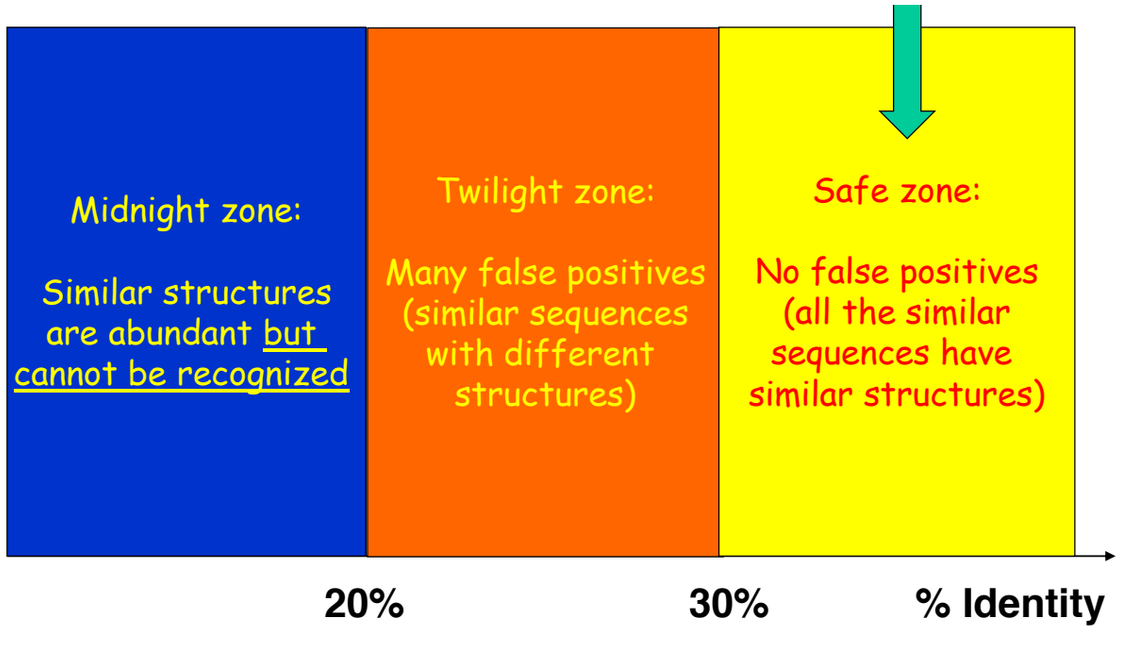
Sander and Schneider, Proteins 1991

Twilight Zone

In the region above 20% of sequence identity, 90% of alignments correspond to homologous protein; while below 25% only 10%.



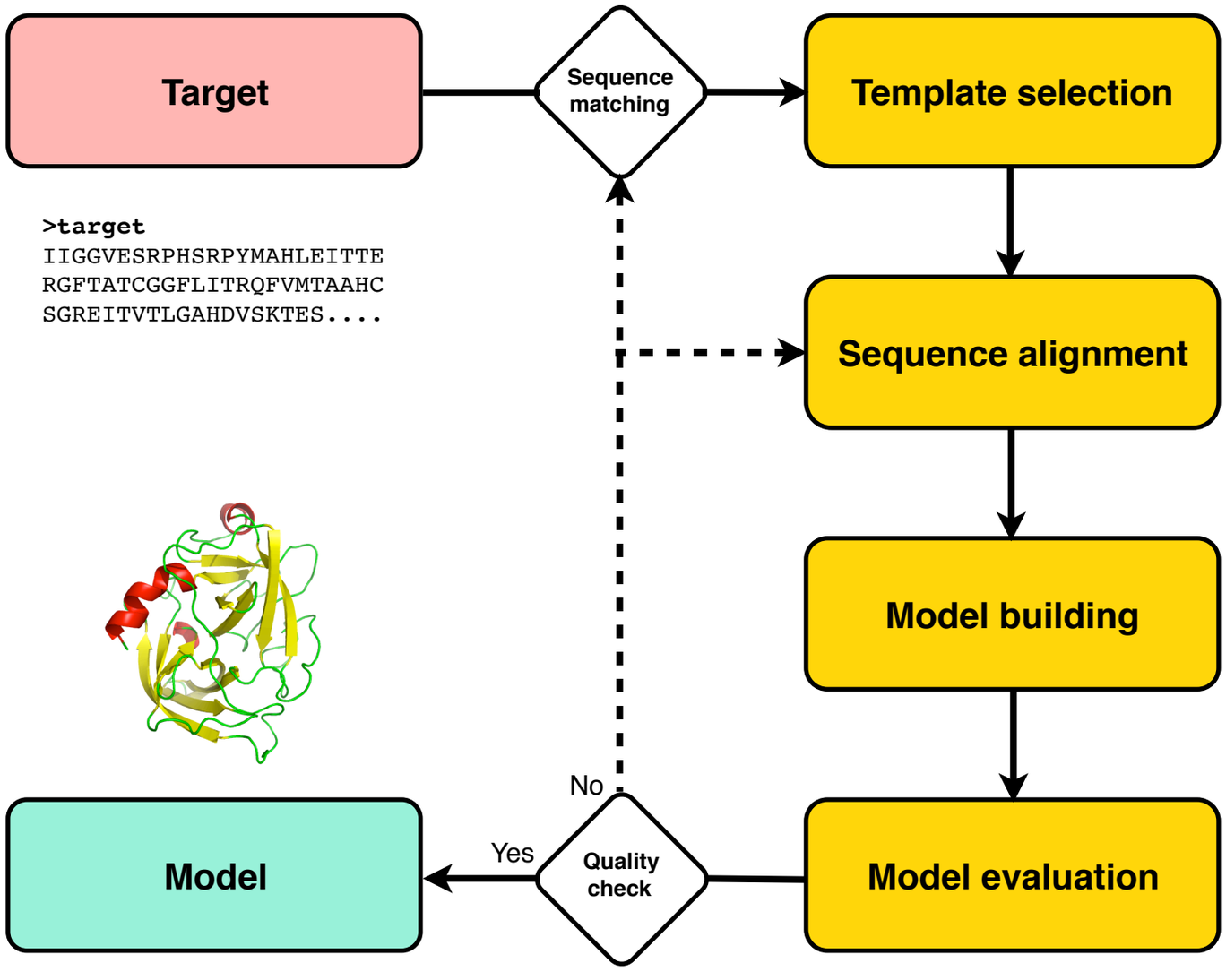


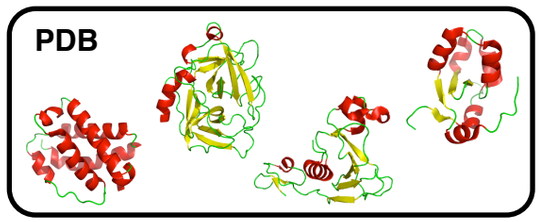


Rost, Proteins 1999

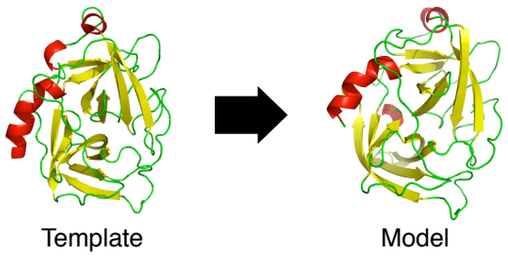
Comparative Modeling

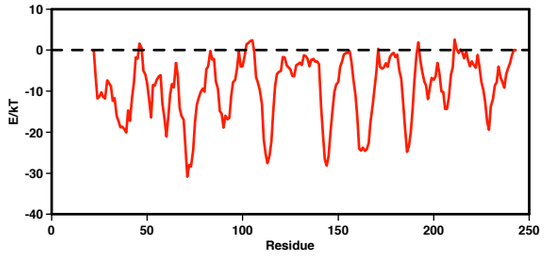
Flow chart of Comparative Modeling






```
target   IIGGVESRPHSRPYMAHLEI
3RP2A    IIGGVESIPHSRPYMAHLDI
target   TTERGFTATCGGFLITRQ..
3RP2A    VTEKGLRVICGGFLISRQ..
```





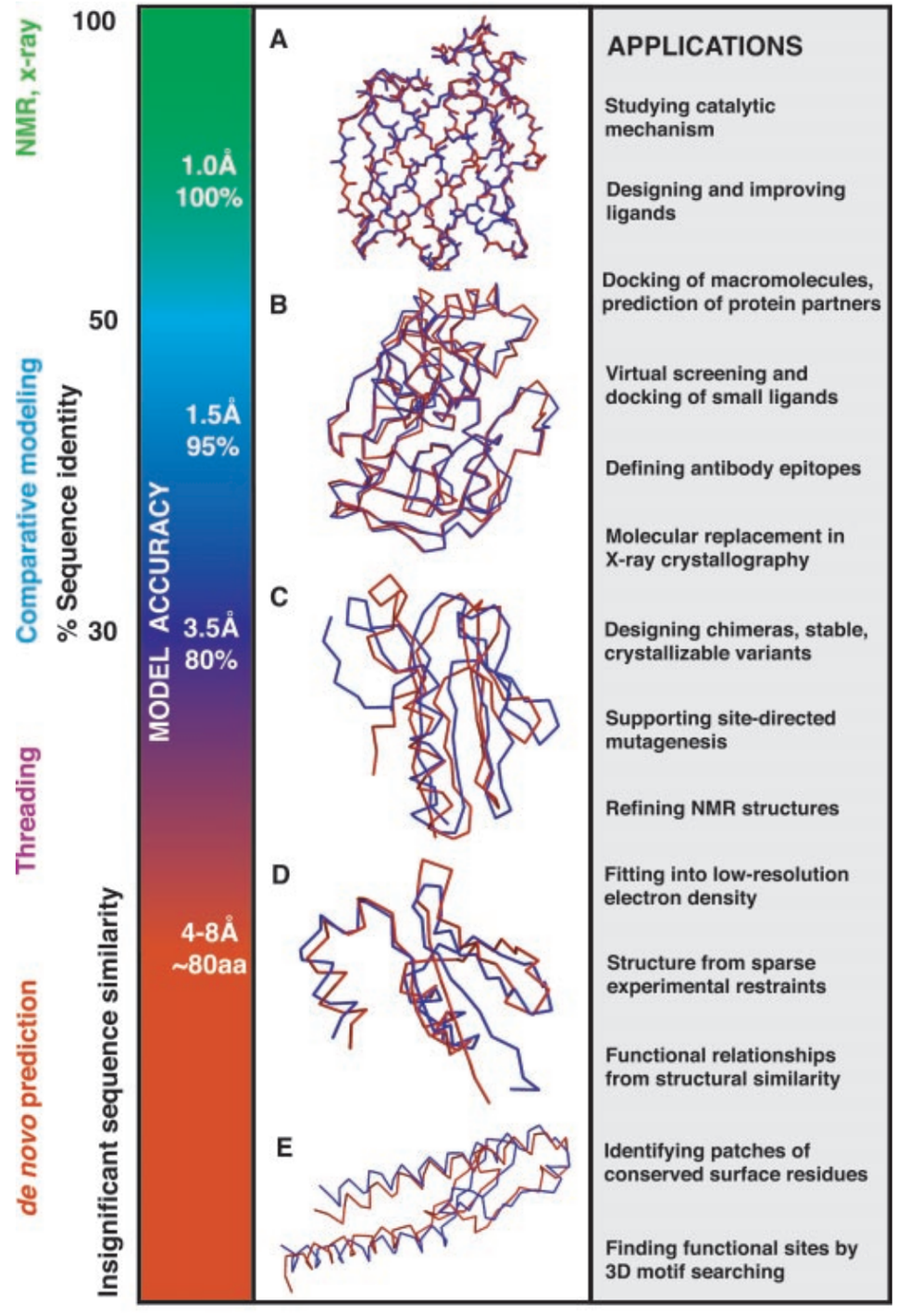
Use of Predicted Structures

Depending off the sequence similarity with the template the predicted structure can be used for different purposes

- Comparative Modeling

- Threading

- *Ab initio* or De novo predictions



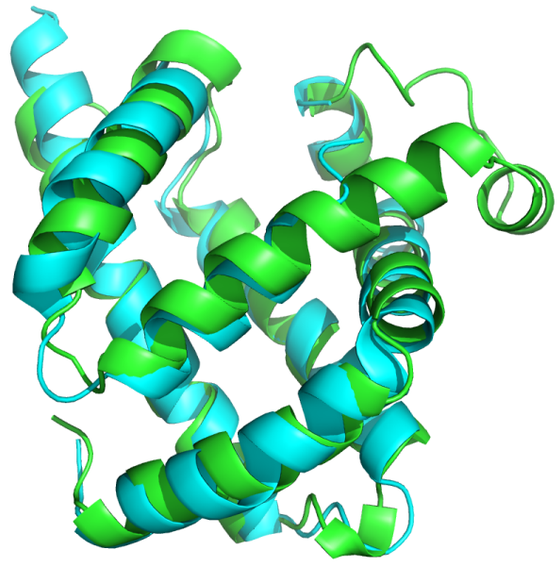
Baker and Sali (2001). Science, 294: 93-96.

Remote homologs

Sequences longer than 100 residues and sharing more the 30% of residues have similar structures (for shorter sequences the level of identity must be higher).

This DO NOT exclude that sequences sharing lower identity have similar structures.

Example:
Sperm Whale Myoglobin (1JP6:A)
Bacterial Haemoglobin (1VHB:A)
RMSD = 0.18 nm, Identity: 12%



Pairs proteins with similar structure and low sequence identity are referred as “remote homologs”

aligned by TM-align

Sequence Identity Inference

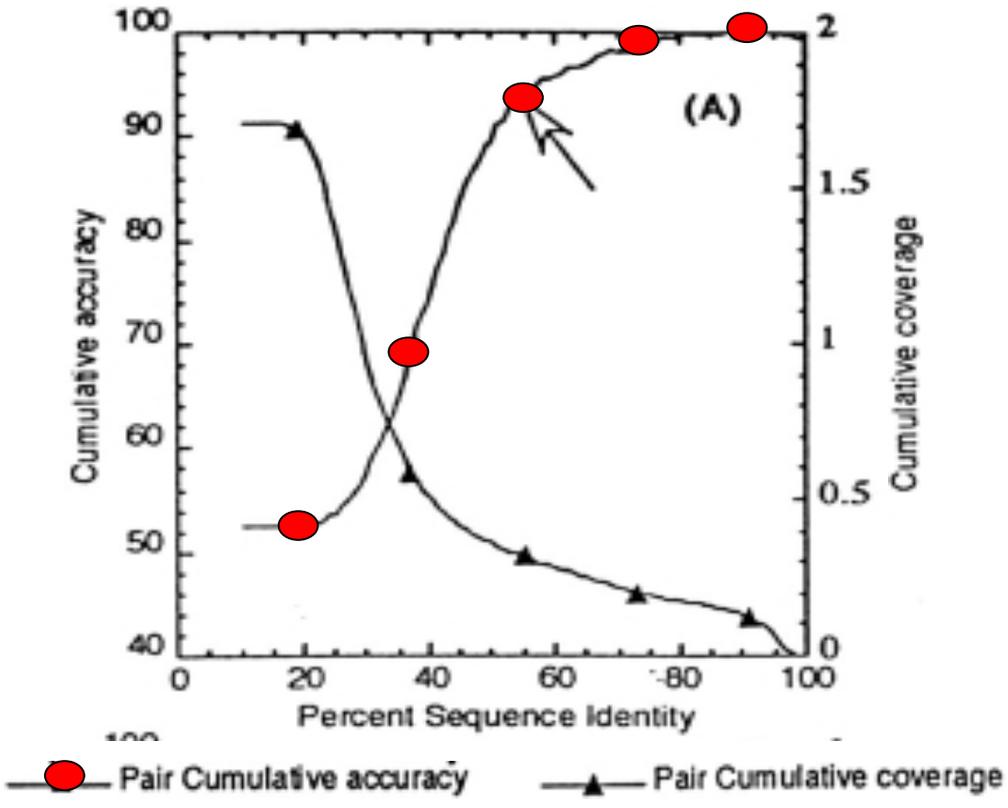
Can we use sequence similarity to predict other features of an unknown protein?

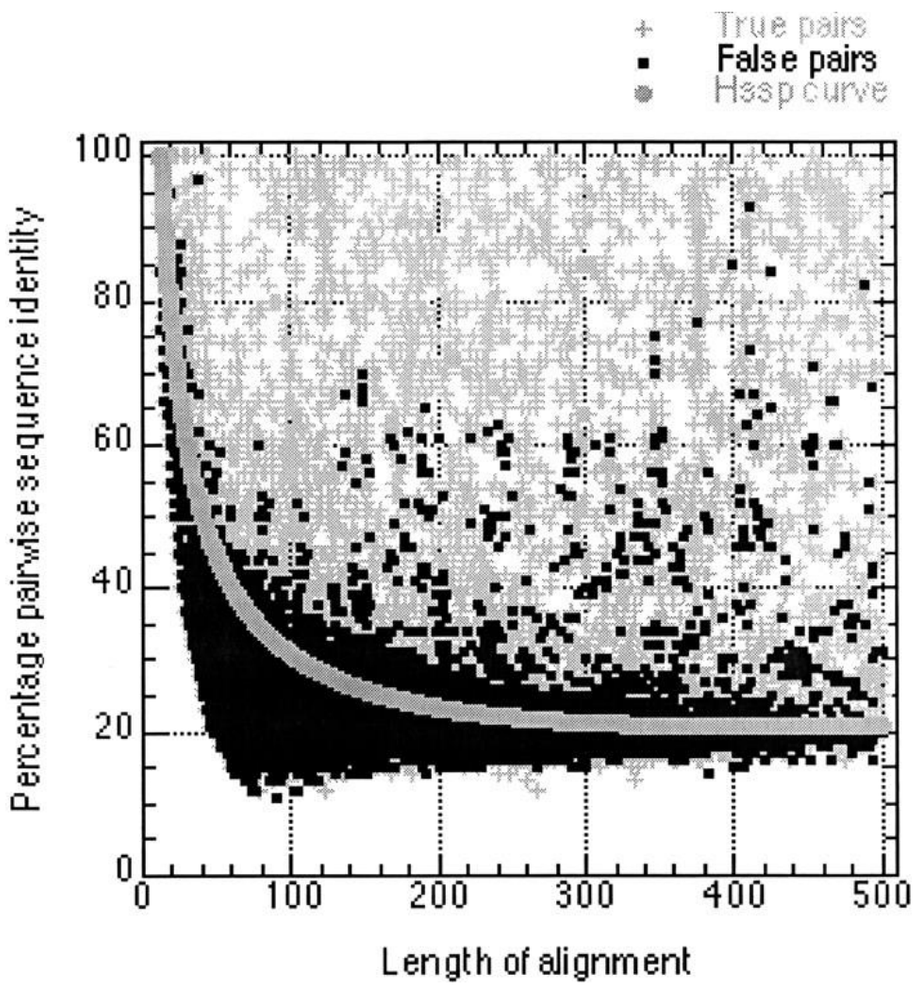
Solution: Define a the sequence similarity threshold that allow a reliable transfer of annotation features.

In other words we need to find the problem specific twilight region

Subcellular Localization

Sequence identity for reliably transferring subcellular localization is higher than that required for transferring structure.





Rost and Nair, Protein Sci. 2002

A false positive

```
               10        20        30        40        50        60
sp|Q9S MEFEKIKVINPVVEMDGDEMTRVIWKFIKDKLIFPFLELDIKYFDLGLPNRDFTDDKVTI
       : :::::: ::.:::::::::::::: :::::: ::.:::::::::::.:: :::::::
sp|Q9S MAFEKIKVANPIVEMDGDEMTRVIWKSIKDKLITPFVELDIKYFDLGLPHRDATDDKVTI
               10        20        30        40        50        60

               70        80        90       100       110       120
sp|Q9S ETAEATLKYNVAIKCATITPDEARVREFGLKKMWRSPNGTIRNILNGTVFREPIICRNIP
       :.:::: ::::::::::::::.:: :::::.:::::::::::::::::::::::.:.:
sp|Q9S ESAEATKKYNVAIKCATITPDEGRVTEFGLKQMWRSPNGTIRNILNGTVFREPIICKNVP
               70        80        90       100       110       120

              130       140       150       160       170
sp|Q9S RLVPGWTKPICIGRHAFGDQYRATDLIVNEPGKLKLVFEPSGSSQKTEFEVFNFTG-GGV
       .:::::::::::::::::::::::: ... :::: ..::  :.. ::: :::.::: :::
sp|Q9S KLVPGWTKPICIGRHAFGDQYRATDAVIKGPGKLTMTFE--GKDGKTETEVFTFTGEGGV
              130       140       150        160       170

     180       190       200       210       220       230
sp|Q9S ALAMYNTDESIRAFAESSMYTAYQKKWPLYLSTKNTILKIYDGRFKDIFQEVYEANWRSK
       :.:::::::::::::..:: :::.:::::::::::::::: ::::::::::::::.:.::
sp|Q9S AMAMYNTDESIRAFADASMNTAYEKKWPLYLSTKNTILKKYDGRFKDIFQEVYEASWKSK
      180       190       200       210       220       230

     240       250       260       270       280       290
sp|Q9S YEAAGIWYEHRLIDDMVAYAMKSEGGYVWACKNYDGDVQSDFLAQGYGSLGMMTSVLVCP
       :.::::::::::::::::::.:::::::::::::::::::::::::.::::.::::::::
sp|Q9S YDAAGIWYEHRLIDDMVAYALKSEGGYVWACKNYDGDVQSDFLAQGFGSLGLMTSVLVCP
      240       250       260       270       280       290

     300       310       320       330       340       350
sp|Q9S DGKTIEAEAAHGTVTRHYRVHQKGGETSTNSIASIFAWSRGLAHRAKLDSNAALLSYTEK
       ::::::::::::::::::.:::::::::::::::::::::.:::::::::::.:: ::..:::
sp|Q9S DGKTIEAEAAHGTVTRHFRVHQKGGETSTNSIASIFAWTRGLAHRAKLDDNAKLLDFTEK
      300       310       320       330       340       350

     360       370       380       390       400       410
sp|Q9S LEAACMGTVESGKMTKDLALLIHGAKVRRDQYVNTEEFIDAVAWELKRRLLGNNSRL
       :::::.:::::::::::::::.:::.:. :: :.:::::::::: :::.::
sp|Q9S LEAACVGTVESGKMTKDLALIIHGSKLSRDTYLNTEEFIDAVAAELKERL-----NA
      360       370       380       390       400             410
```

Q9SLK0 (ICDHX_ARATH):
Peroxisomal isocitrate dehydrogenase

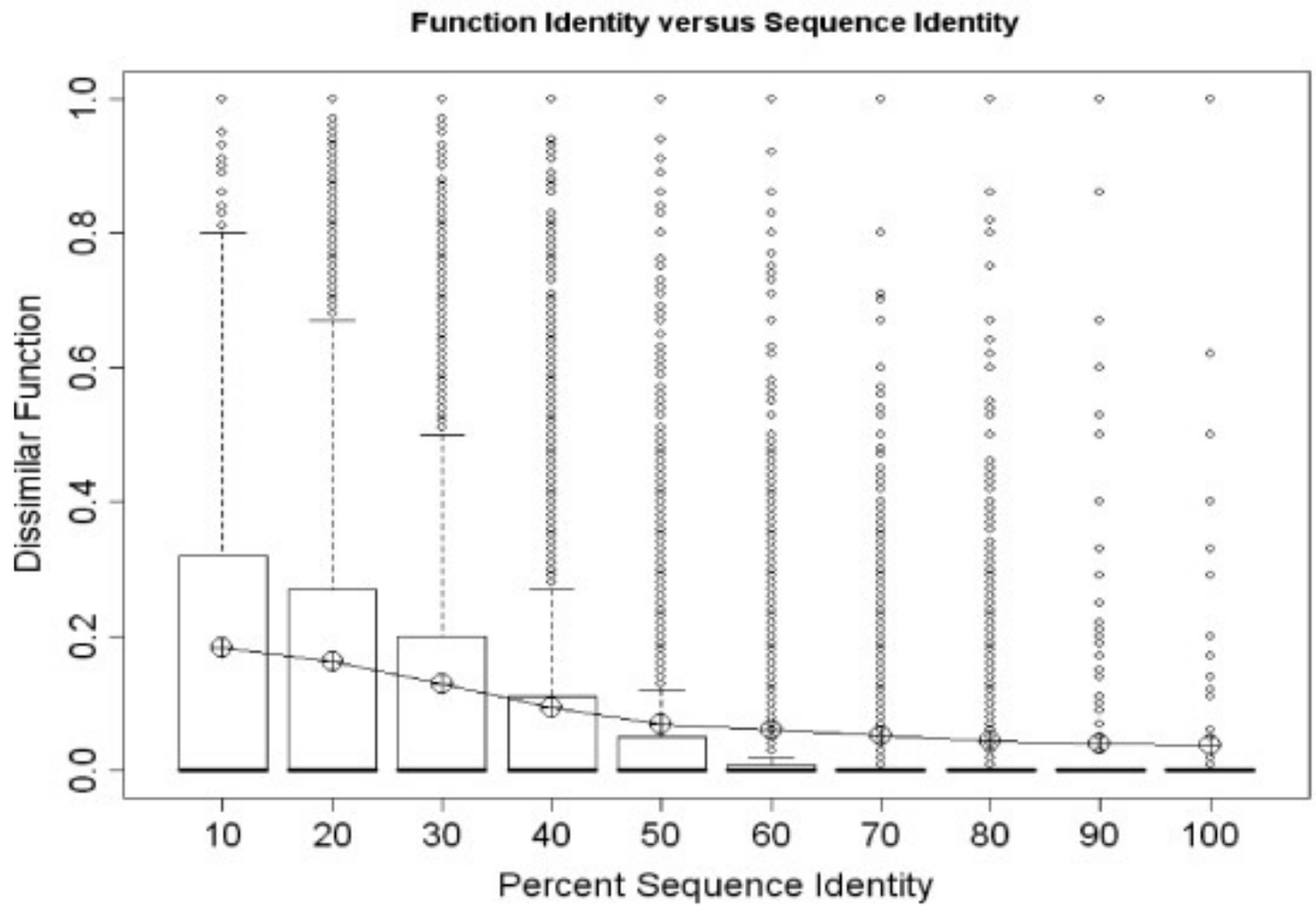
Q9SRZ6 (ICDHC_ARATH):
Cytosolic isocitrate dehydrogenase

84.2% identity (93.3% similar) in 417 aa overlap

Lalign at ExPASy

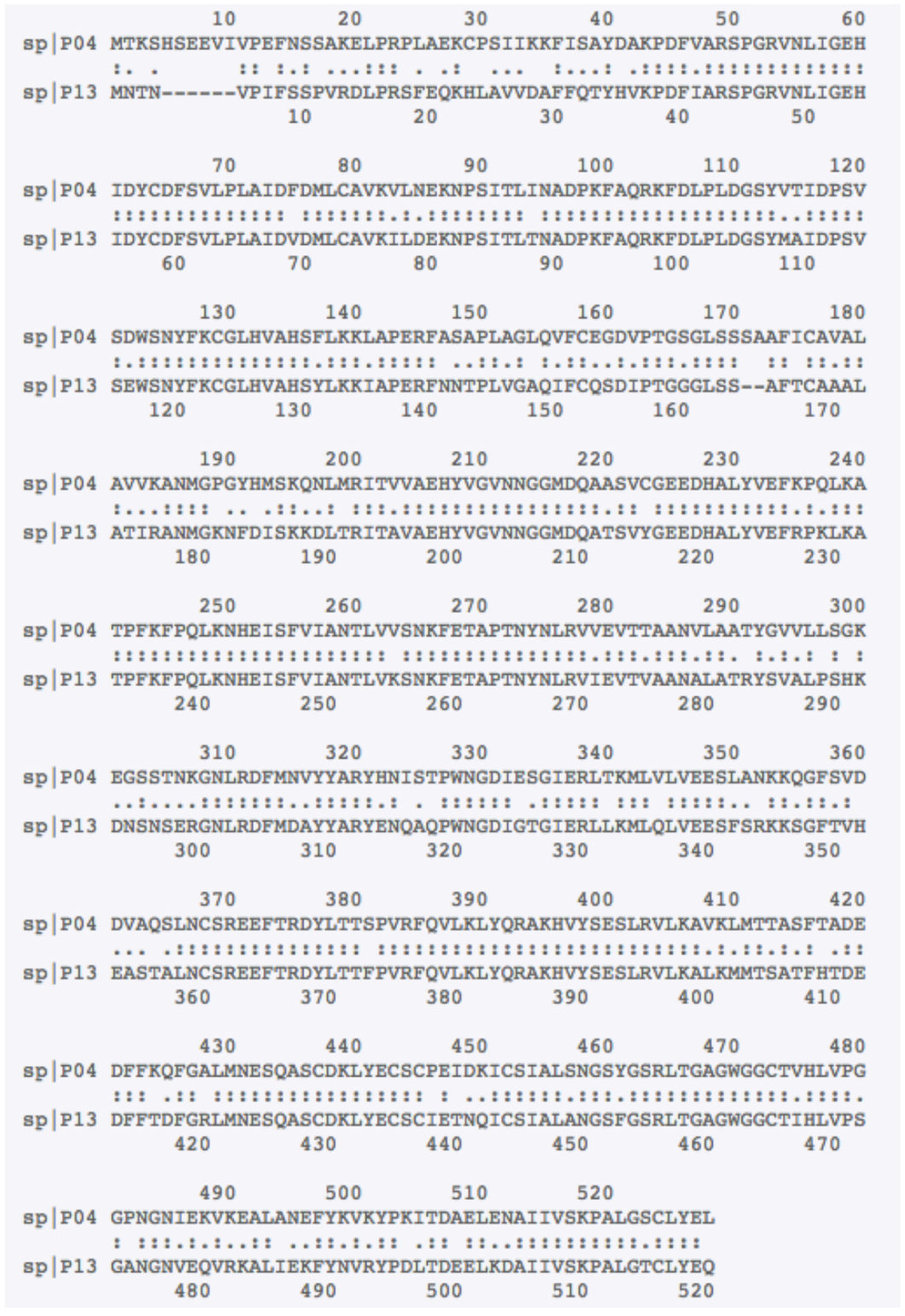
Functional Annotation

Sequence identity for can be used for functional annotation measuring the identity and similarity between Gene Ontology terms.



Sangar V et al. BMC Bioinformatics 2007

Dissimilar functions



P04385 (GAL1_YEAST) Galactokinase

Catalytic activity
ATP + alpha-D-galactose = ADP + alpha-D-galactose 1-phosphate.

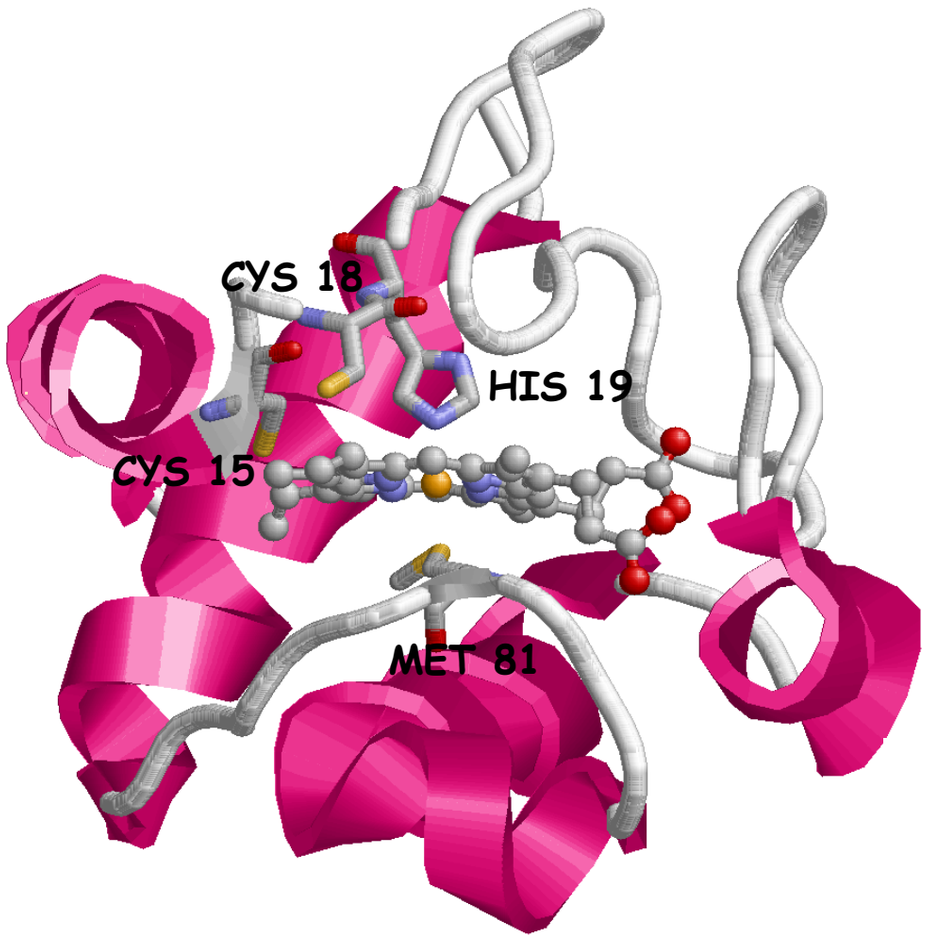
P13045 (GAL3_YEAST) Protein GAL3
The GAL3 regulatory function is required for rapid induction of the galactose system.

72.9% identity (90.5% similar) in 528 aa overlap

Lalign at ExPASy

Case Study

Electron carrier protein. The oxidized form of the cytochrome c heme group can accept an electron from the heme group of the cytochrome c1 subunit of cytochrome reductase. Cytochrome c then transfers this electron to the cytochrome oxidase complex, the final protein carrier in the mitochondrial electron-transport chain.



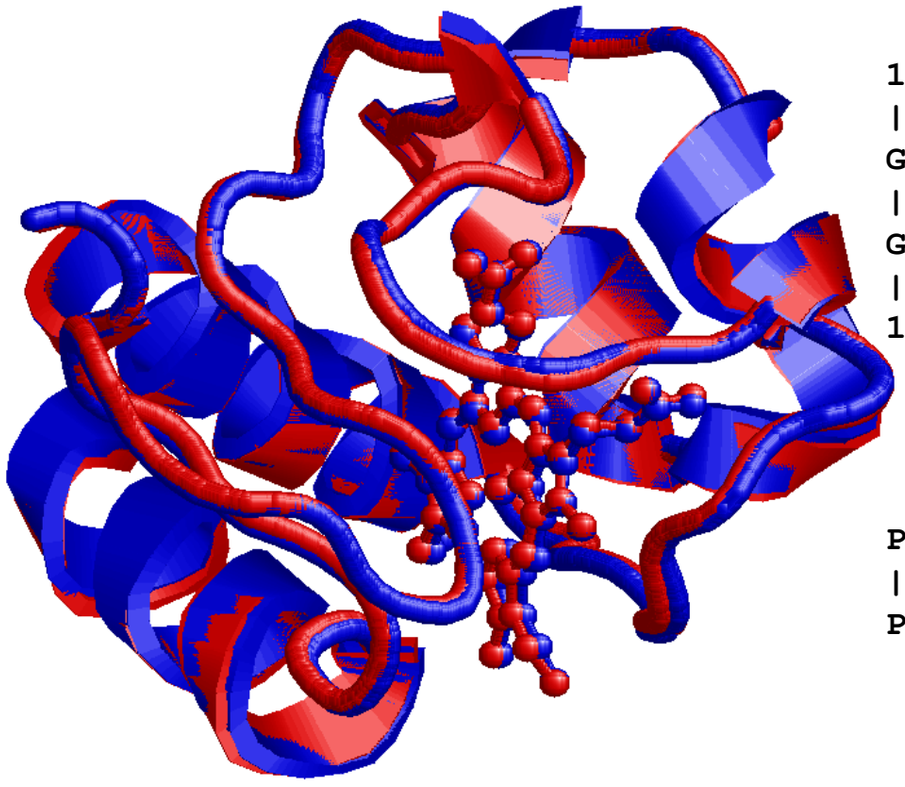
PDB: 3zcf:A

Feature key	Position(s)	Length	Description
Binding site[i]	15 – 15	1	Heme (covalent)
Binding site[i]	18 – 18	1	Heme (covalent)
Metal binding[i]	19 – 19	1	Iron (heme axial ligand)
Metal binding[i]	81 – 81	1	Iron (heme axial ligand)

Homo vs Horse

Human Cytochrome C – Uniprot:P99999. PDB: 3ZCF:A
Equine Cytochrome C – Uniprot: P00004. PDB 3O20:A

Structural alignment:
RMSD= 0.035 nm
88% sequence identity



```
1:A                  20:A                  40:A                  60:A
|         |    .    |    .    |    .    |    .    |    .    |    .    |
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLEN
||||||||||:.||.||||||||||||||||||||||||||||:.||.|||||||.|.|:||||||||
GDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMEYLEN
|         |    .    |    .    |    .    |    .    |    .    |    .    |
1:A                  20:A                  40:A                  60:A

          80:A                  100:A
     .    |    .    |    .    |
PKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
||||||||||||.|||||.||.||||||||||||
PKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE
     .    |    .    |    .    |
          80:A                  100:A
```

Sequence vs Structure

In this case the sequence alignment is the same of the structural alignment and the positions of the binding sites are conserved.

Sequence alignment:
88% sequence identity
IDENTICAL TO STRUCTURAL ALIGNMENT

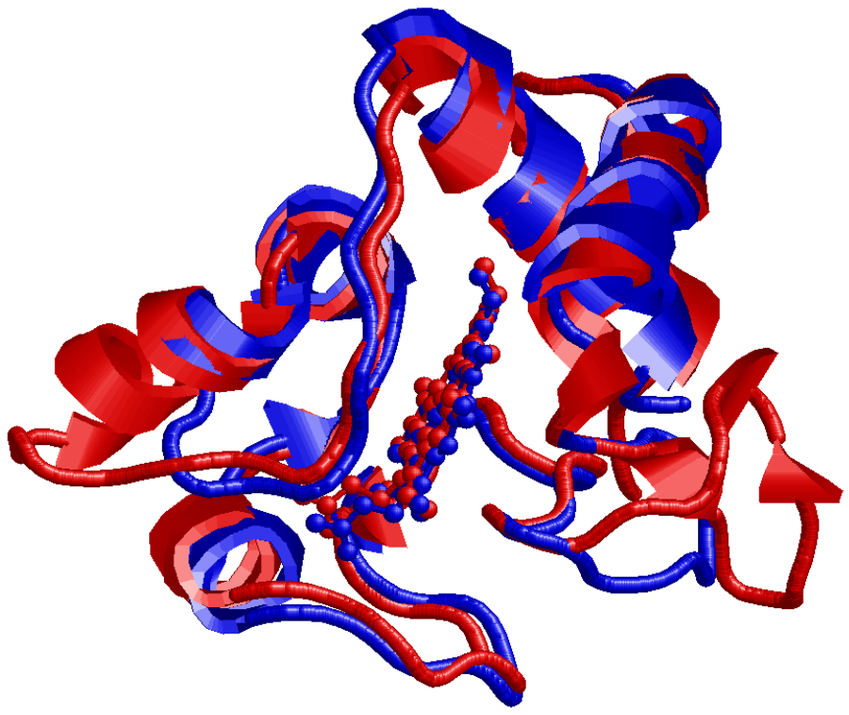
```
88.6% identity (95.2% similar) in 105 aa overlap (1-105:1-105)

                10        20        30        40        50        60
Homo  MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIW
      ::::::::::..::.:::::::::::::::::::::::::::::::.:: ::::::: :
Horse MGDVEKGKKIFVQKCAQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITW
                10        20        30        40        50        60

                70        80        90       100
Homo  GEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
       :.:::::::::::::::::.::::: :: ::::::::::::
Horse KEETLMEYLENPKKYIPGTKMIFAGIKKKTEREDLIAYLKKATNE
                70        80        90       100
```

Lalign at ExPASy

Homo vs Rhodobacter Sph.

Human Cytochrome C — Uniprot:P99999. PDB: 3ZCF:A
Cytochrome C2 Rhodobacter Sph. – Uniprot: P0C0X8. PDB 1CXC:A

Structural alignment:
RMSD= 0,18 nm
28% sequence identity



```
1:A                  20:A                        40:A
|          |    .    |             .    |    .    |    .     |    .
GDVEKGKKIFIMKCSQCHTVEKGG-------KHKTGPNLHGLFGRKTGQAPGYS-YTAANKNKG---IIW
||.|.|.|.|. .|..||.:....        ..||||||:|..||..|....:. |....|..|   :.|
GDPEAGAKAFN-QCQTCHVIVDDSGTTIAGRNAKTGPNLYGVVGRTAGTQADFKGYGEGMKEAGAKGLAW
|      |     .    |    .    |    .    |    .    |    .    |    .    |
3:A                20:A                40:A                60:A


60:A                           80:A                  100:A
|    .    |    .              |    .    |    .    |
GEDTLMEYLENPKKYI--------PGTKMIFVGIKKKEERADLIAYLKKATNE
.|:...:|.:.|.|::         ...||.| .:||..:...:.|||.....
DEEHFVQYVQDPTKFLKEYTGDAKAKGKMTF-KLKKEADAHNIWAYLQQVAVR
   .    |    .    |    .    |     .    |    .    |
        80:A                100:A                120:A
```

Sequence vs Structure (I)

In this case the sequence alignment can be used for homology modeling after a refinement of the alignment because one binding site is not conserved.

Structural alignment:
RMSD= 0,18 nm
28% sequence identity

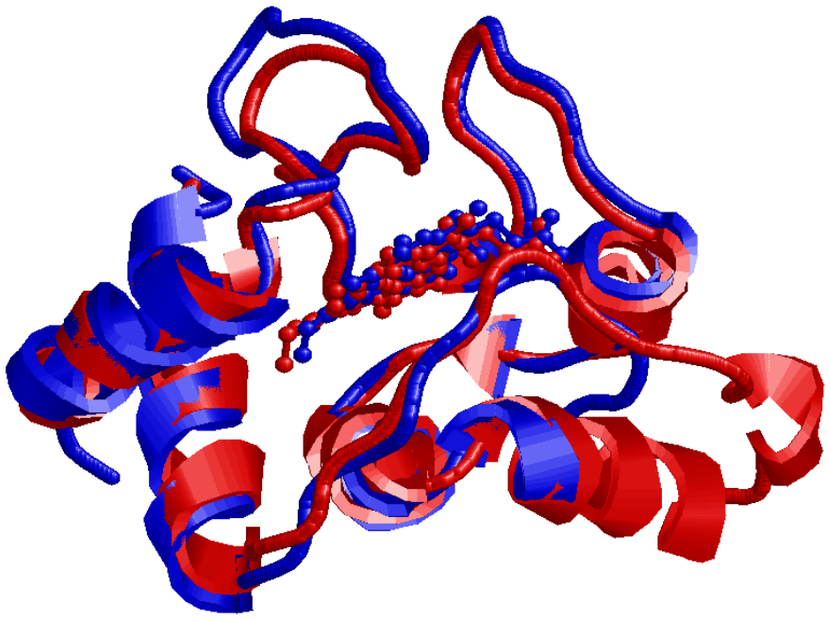
```
Global without end-gap score: 111; 29.3% identity (56.1% similar) in 123 aa
                10        20                  30        40         50
sp|P99  MGDVEKGKKIFIMKCSQCHTVEK-------GGKHKTGPNLHGLFGRKTG-QAPGYSYTA
         :: : : : :  .:. ::..          : . :::::::.:. :: .: ::   .:
sp|P0C QEGDPEAGAKAF-NQCQTCHVIVDDSGTTIAGRNAKTGPNLYGVVGRTAGTQADFKGYGE
               10          20        30        40         50

                  60        70                  80        90        100
sp|P99 ANKN---KGIIWGEDTLMEYLENPKKYIP-------GTKMIFVGIKKKEERADLIAYLKK
       . :.   ::. : :. ...:...: :..          .   .   .::. .   .. ::::..
sp|P0C GMKEAGAKGLAWDEEHFVQYVQDPTKFLKEYTGDAKAKGKMTFKLKKEADAHNIWAYLQQ
      60        70        80        90       100       110

sp|P99 ATNE
       .. .
sp|P0C VAVRP
     120
```

Lalign at ExPASy

Homo vs Rhodobacter Pal.

Human Cytochrome C - Uniprot:P99999. PDB: 3ZCF:A
Cytochrome C2 Rhodopseudomons pal. – Uniprot: P00091. PDB 1I8O:A

Structural alignment:
RMSD= 0,13 nm
29% sequence identity



```
1:A                  20:A                  40:A                        60:A
|         |    .    |    .    |    .    |    .    |    .         |    .
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKG---IIWGEDTLMEY
.|...|..:|. .|..||..   .|...||.|.|..|||.|.|.|:.|...|.|.|   ::|..|.:..|
xDAKAGEAVFK-QCMTCHRA---DKNMVGPALAGVVGRKAGTAAGFTYSPLNHNSGEAGLVWTADNIVPY
|         |    .         |    .    |    .    |    .    |    .    |    .
1:A                  20:A                  40:A                        60:A


                                80:A                 100:A
  |    .                        |    .    |    .    |
LENPKKYIP--------------GTKMIFVGIKKKEERADLIAYLKKAT
|..|..::.                .|||.| .:...::|.|.:|||....
LADPNAFLKKFLTEKGKADQAVGVTKMTF-KLANEQQRKDVVAYLATLK
   |    .    |    .    |    .     |    .    |
              80:A                 100:A
```

Sequence vs Structure (II)

In this case the sequence alignment needs to be fixed homology to because all the binding site shifted.

Structural alignment:
RMSD= 0,13 nm
29% sequence identity

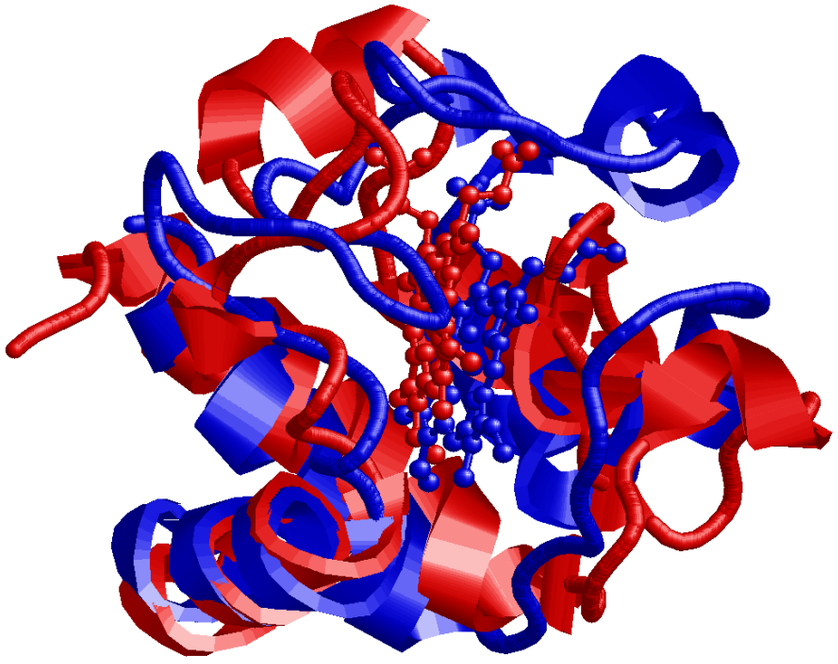
```
Global without end-gap score: 152; 28.7% identity (63.0% similar) in 108 aa
                                        10        20        30
sp|P99                              MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGL
                                     :.. :. .:    .:: : ... :. .:: : :.
sp|P00 MVKKLLTILSIAATAGSLSIGTASAQDAKAGEAVF----KQCMTCHRADKNMVGPALGGV
              10        20        30             40        50

         40        50            60        70        80        90
sp|P99 FGRKTGQAPGYSYTAANKNKG---IIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERA
        ::::.: : :..:.  :.:.:   ..:  :....::..:. ..   : ...   : .. .
sp|P00 VGRKAGTAAGFTYSPLNHNSGEAGLVWTADNIINYLNDPNAFL---KKFLTDKGKADQAV
         60        70        80        90           100       110

           100
sp|P99 DLIAYLKKATNE
        .   .   : .::
sp|P00 GVTKMTFKLANEQQRKDVVAYLATLK
           120       130
```

Lalign at ExPASy

Homo vs Arabidopsis

Human Cytochrome C - Uniprot:P99999. PDB: 3ZCF:A
Cytochrome C6A Arabidopsis Thaliana – Uniprot: Q93VA3. PDB 2CE0:A

Structural alignment:
RMSD= 0,35 nm
13% sequence identity



```
1:A                  20:A                  40:A                  60:A
|        |    .    |    .    |         .    |    .    |    .    |    .
GDVEKGKKIFIMKCSQCHTVEKGGKHKTGP---NLHG-LFGRKTGQAPGYSYTAANKNKGIIWGEDTLME
.|:::|..:|...|..||...  |...  .   .|.. ...|.                   .:..|:.:..
LDIQRGATLFNRACAACHDTG--GNII--QPGATLFTKDLERN-----------------GVDTEEEIYR
|      |    .    |      .      |    .    |                        .    |
3:A               20:A                  40:A


                        80:A                 100:A
            |    .    |    .    |    .    |
YLE---------NPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE
...         ..|....|.......:... |...|..:.|....:
VTYFGKGRMPGFGEKCTPRGQCTFGPRLQDE-EIKLLAEFVKFQADQ
   .    |    .    |    .    |    .    |    .
        60:A                 80:A
```

Sequence vs Structure (III)

In this case the sequence alignment is significantly different form the structural alignment.

Structural alignment:
RMSD= 0,35 nm
13% sequence identity

```
Global without end-gap score: 3; 20.0% identity (43.8% similar) in 105 aa
                                                  10        20        30
Homo                                        MGDVEKGKKIFIMKCSQCHTVEKGGKHKTG
                                              :...:  .:   :  :: .  :.  . :
A.Thal DFLLKKLAPPLTAVLLAVSPICFPPESLGQTLDIQRGATLFNRACIGCHDT-GGNIIQPG
               50        60        70        80        90        100

                40        50        60        70        80        90
sp|P99 PNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKE
        .:     ...:       :  .  .   .:.  .  . :   :  :  .  : : . ..
sp|Q93 ATLFTKDLERNGVD-----TEEEIYRVTYFGKGRMPGFGE---KCTPRGQCTF-GPRLQD
               110             120       130           140        150

;              100
sp|P99 ERADLIAYLKKATNE
       :.  :.: . :   .
sp|Q93 EEIKLLAEFVKFQADQGWPTVSTD
             160       170
```

Lalign at ExPASy

Search for Better Alignment

Why is it not sufficient to align sequences (when identity is low) to recover information, not even for “important” residues?

Sequence alignments are «general» and treat each position in the same way
There is no knowledge on the «important» sites

How can we detect the “important” residues starting from protein structures (even when information on catalytic sites is not available)?

Compare multiple structures and analyze the conservation of residues

How can we align sequences constraining the alignment of important residues?

Compare multiple sequences and check for the conservation of patterns Use alignment frameworks able to introduce positional dependences.