

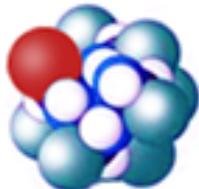
# Introduction and Basic Concepts

**Laboratory of Bioinformatics I  
Module 2**

March 13, 2017

**Emidio Capriotti**

<http://biofold.org/>



**Biomolecules  
Folding and  
Disease**

Department of Biological, Geological,  
and Environmental Sciences (BiGeA)  
University of Bologna



# Schedule and Materials

This module is a 60-hour course running for 8 weeks  
March 13 - May 4, 2017

Regular schedule:

Monday 10:00 - 13:00

Tuesday and Thursday 14:00 - 17:00

Course website

<http://biofold.org/courses/2017/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  $A = (a_1, a_2, \dots, a_n)$  and  $B = (b_1, b_2, \dots, b_m)$  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.

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

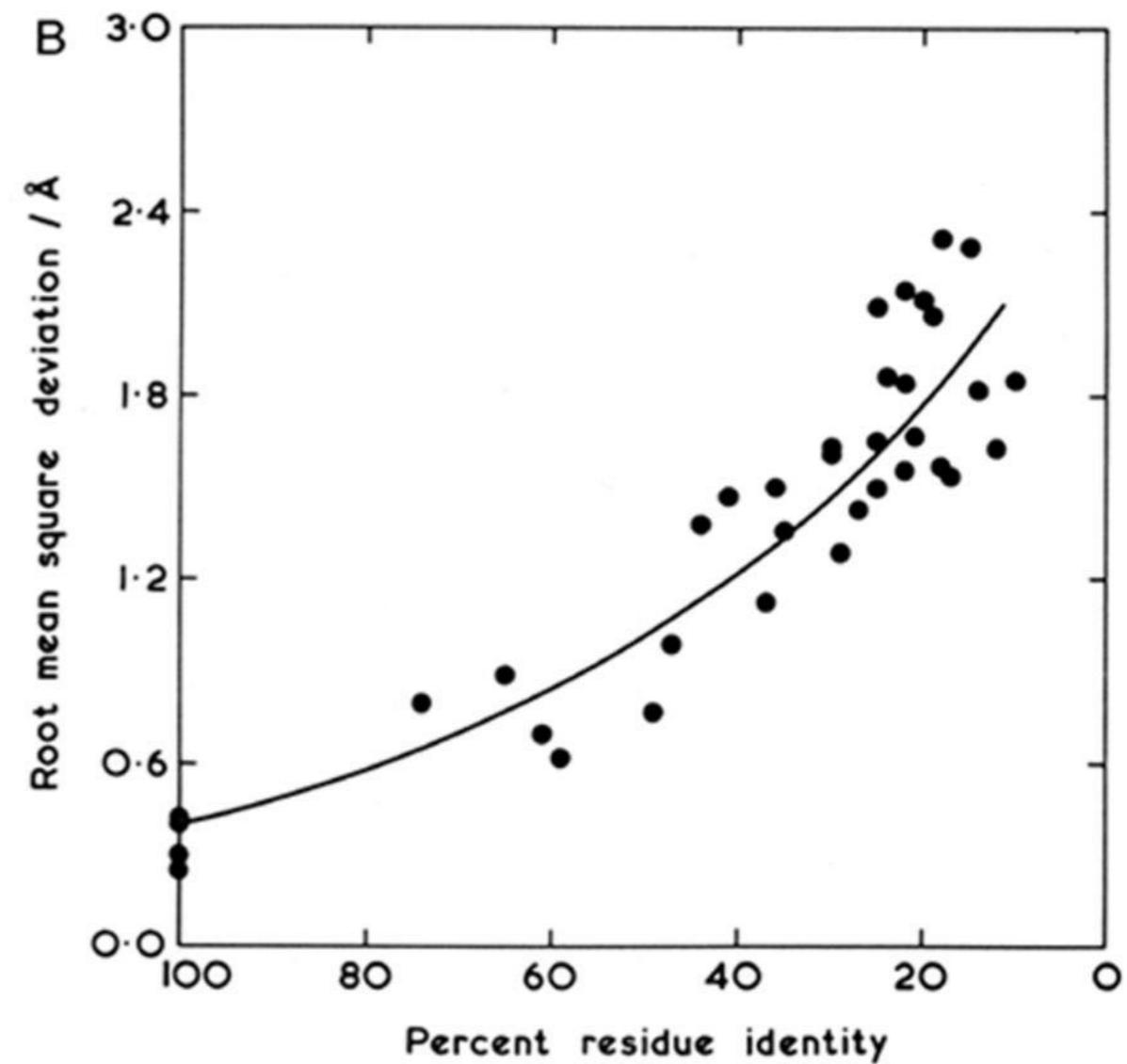
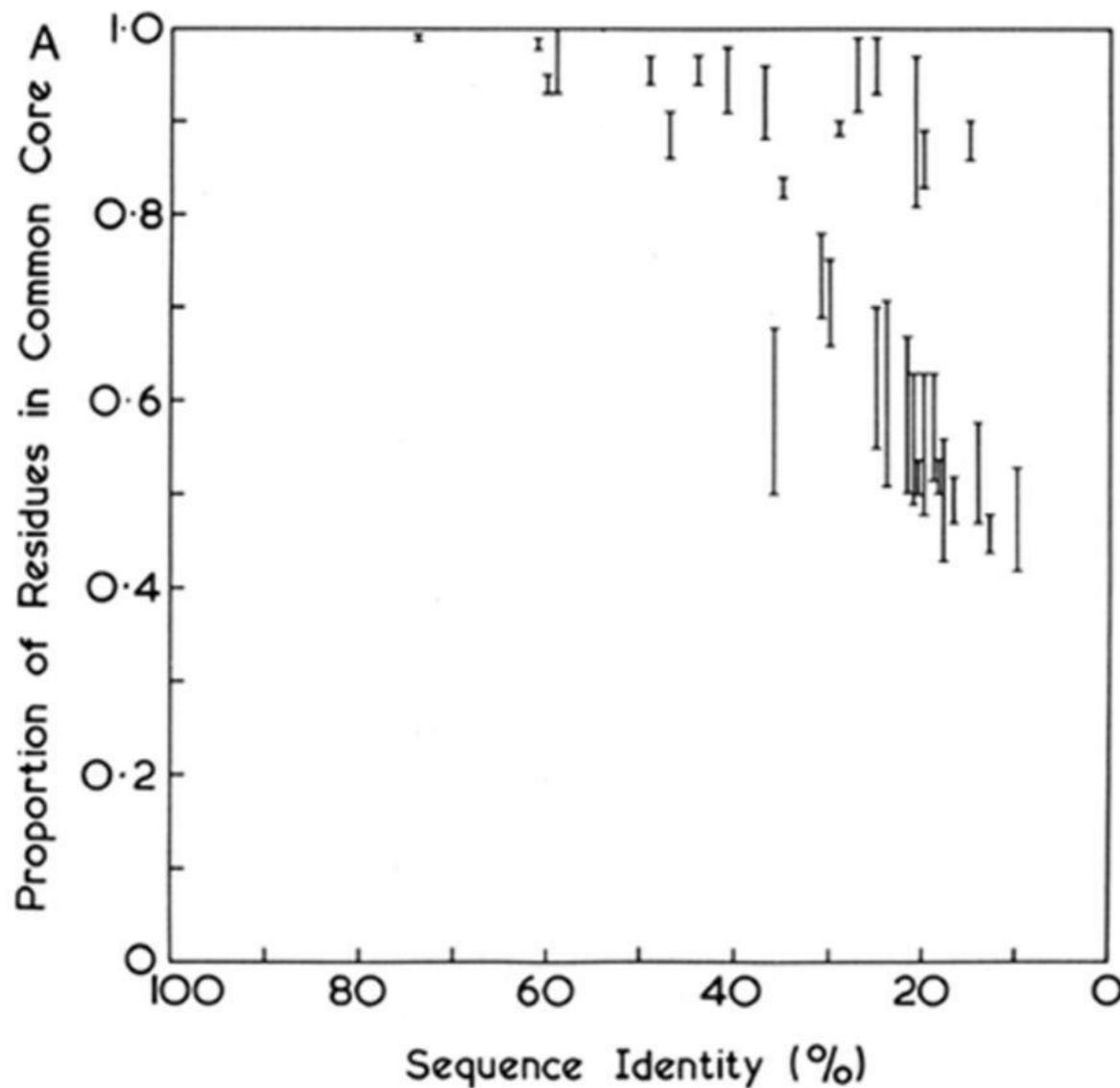
$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^n (a_i - b_i)^2}{n}}$$

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  $O(n)$ ).

# The Foundation of Structural Bioinformatics



# 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: ASAVALNRLITRLYP

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

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

# 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

$$\sigma(n) = -nd \text{ linear}$$

$$\sigma(n) = -d - (n-1)e \quad (d: \text{opening}, e: \text{extension})$$

# Alignment Algorithms

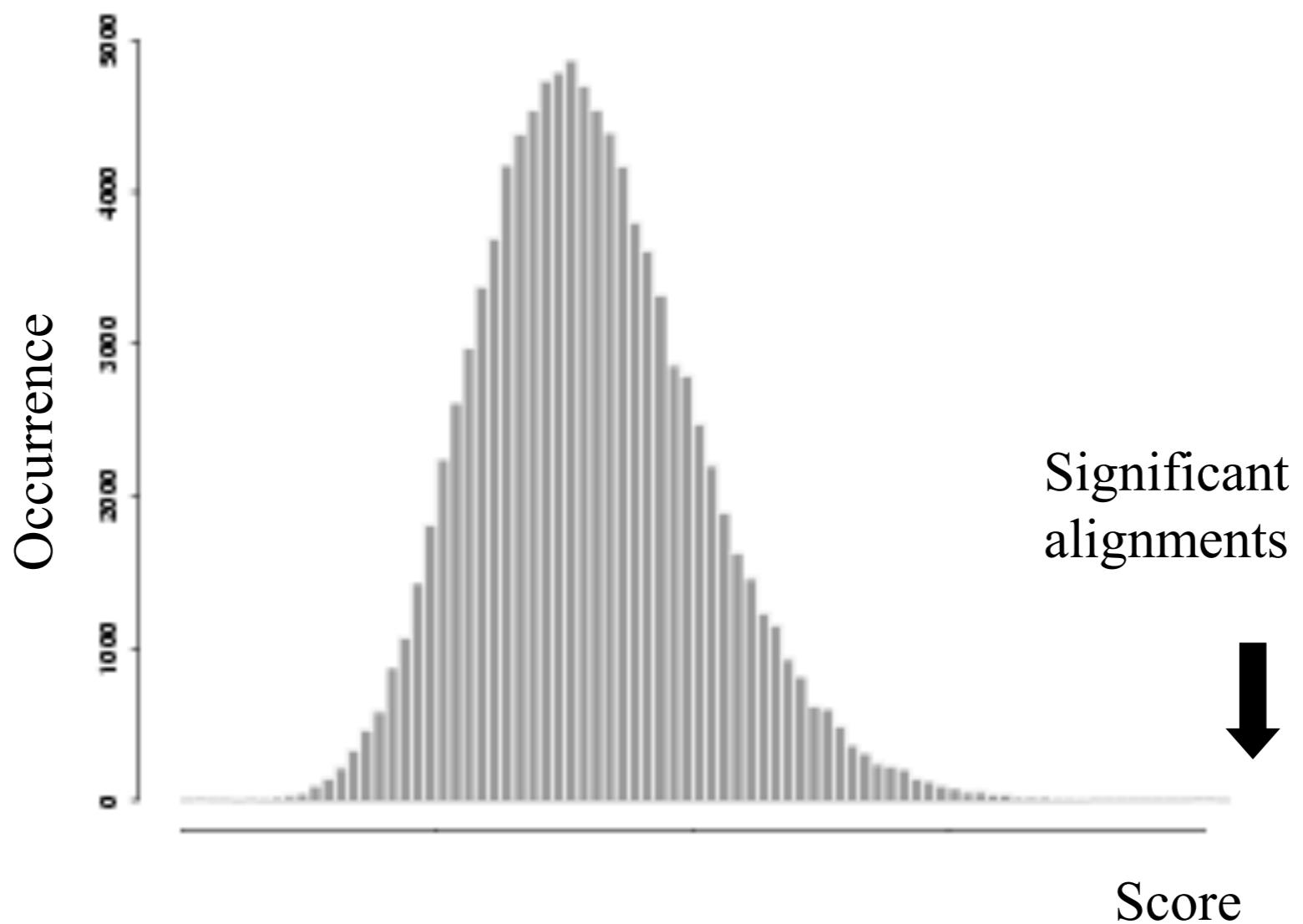
Algorithms for finding the **minimum distance** between two sequences

- **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

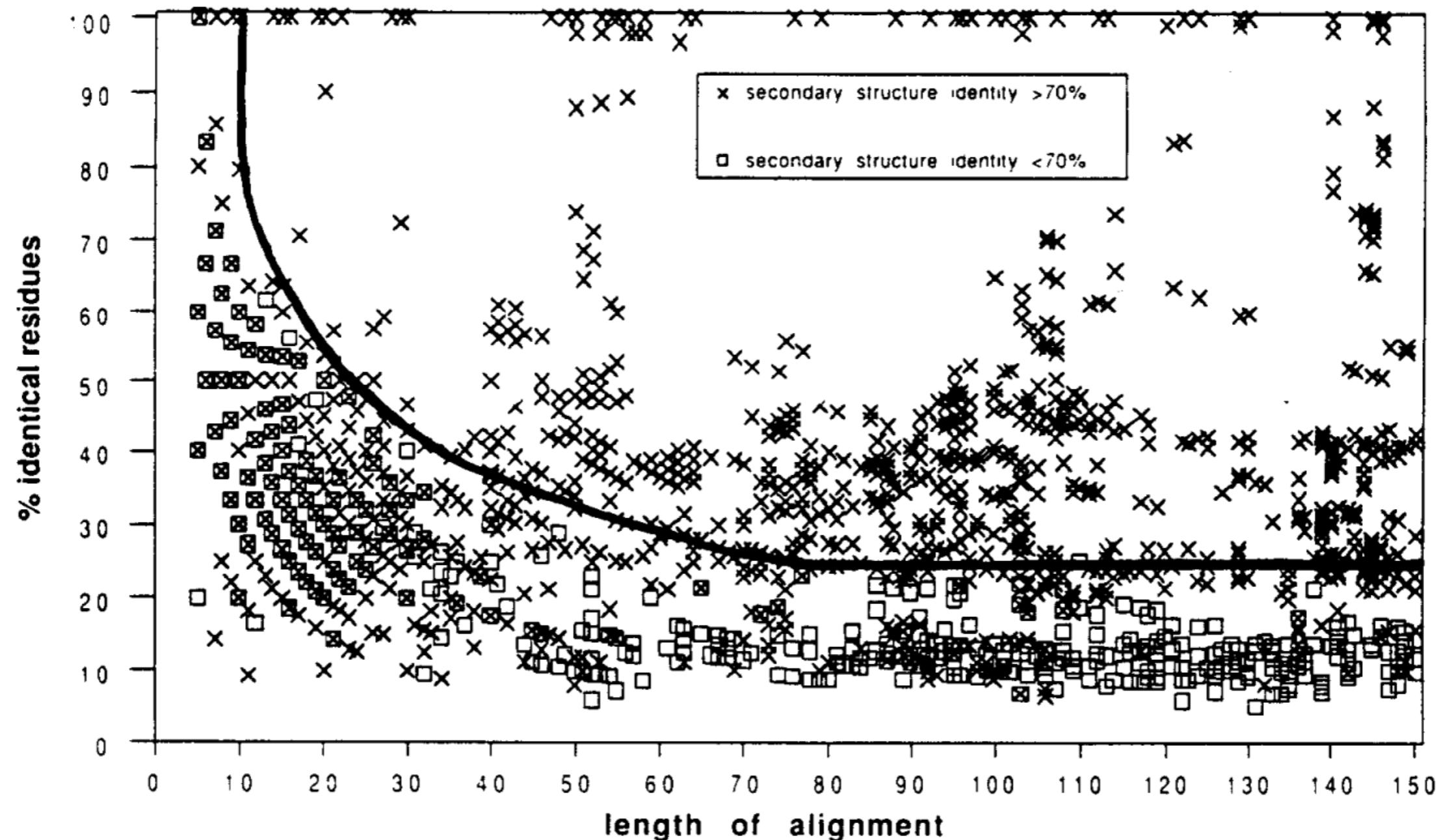
Given an alignment with score  $S$ , is it significant?

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



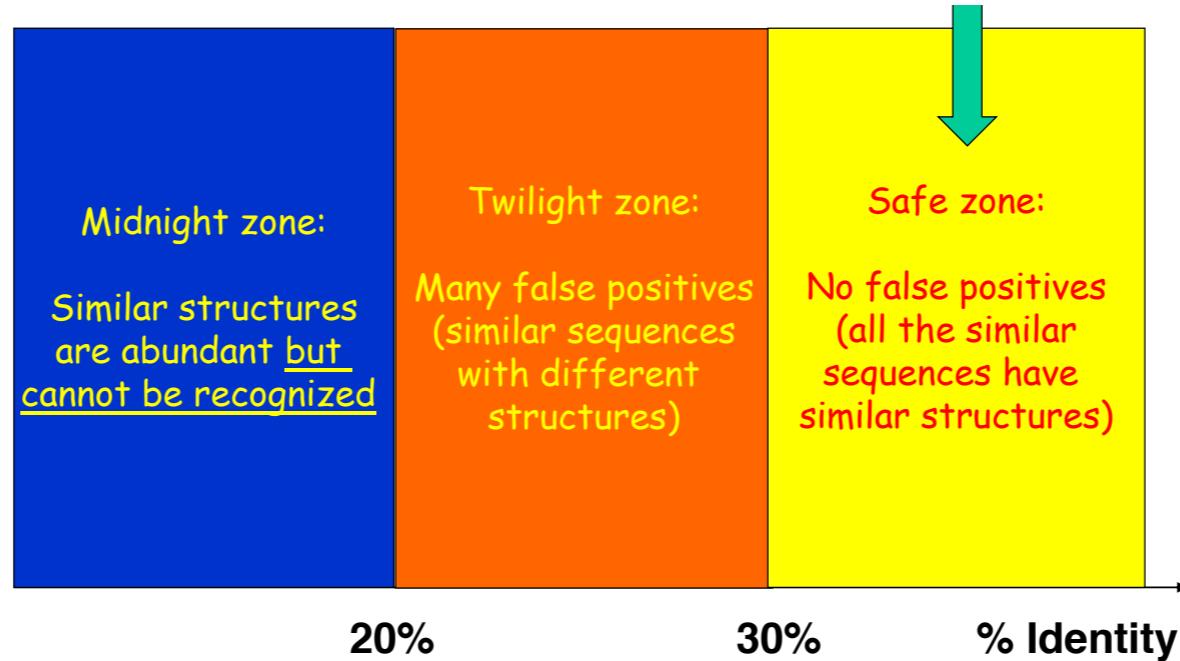
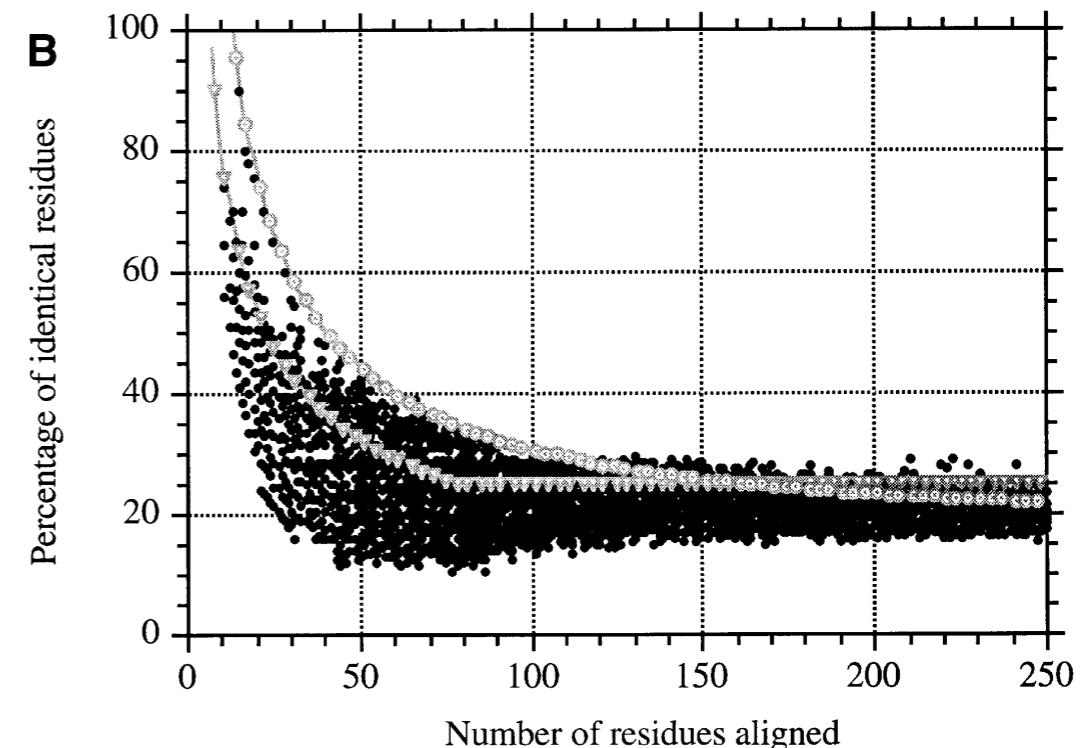
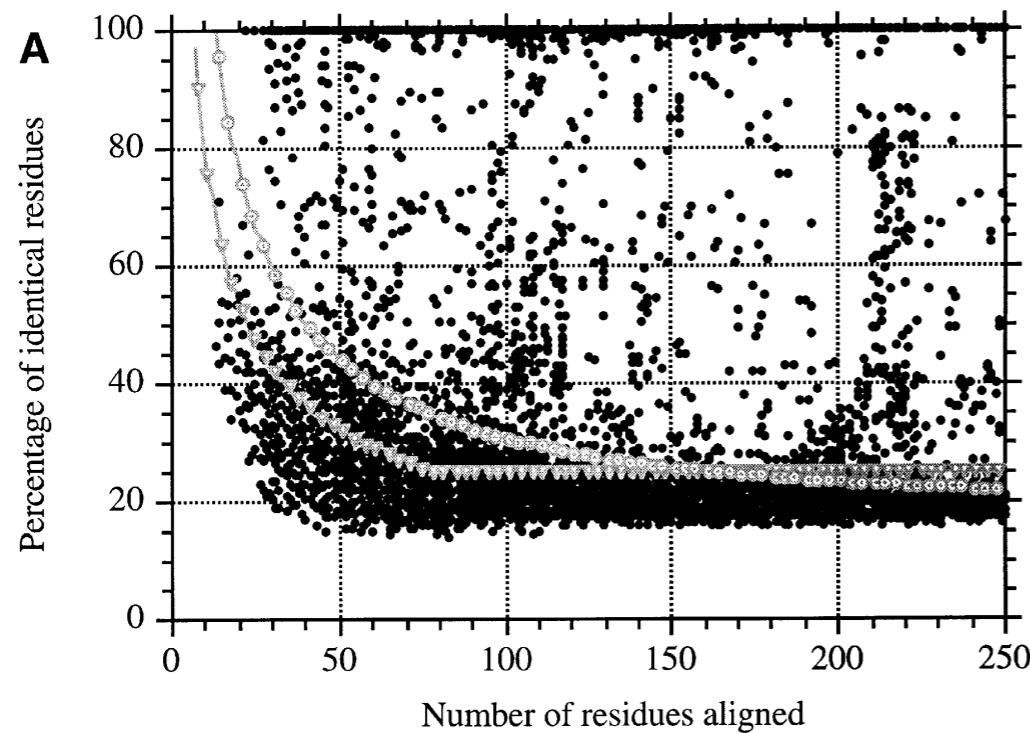
# Structural Homology

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



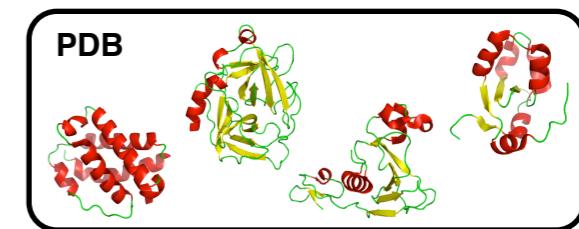
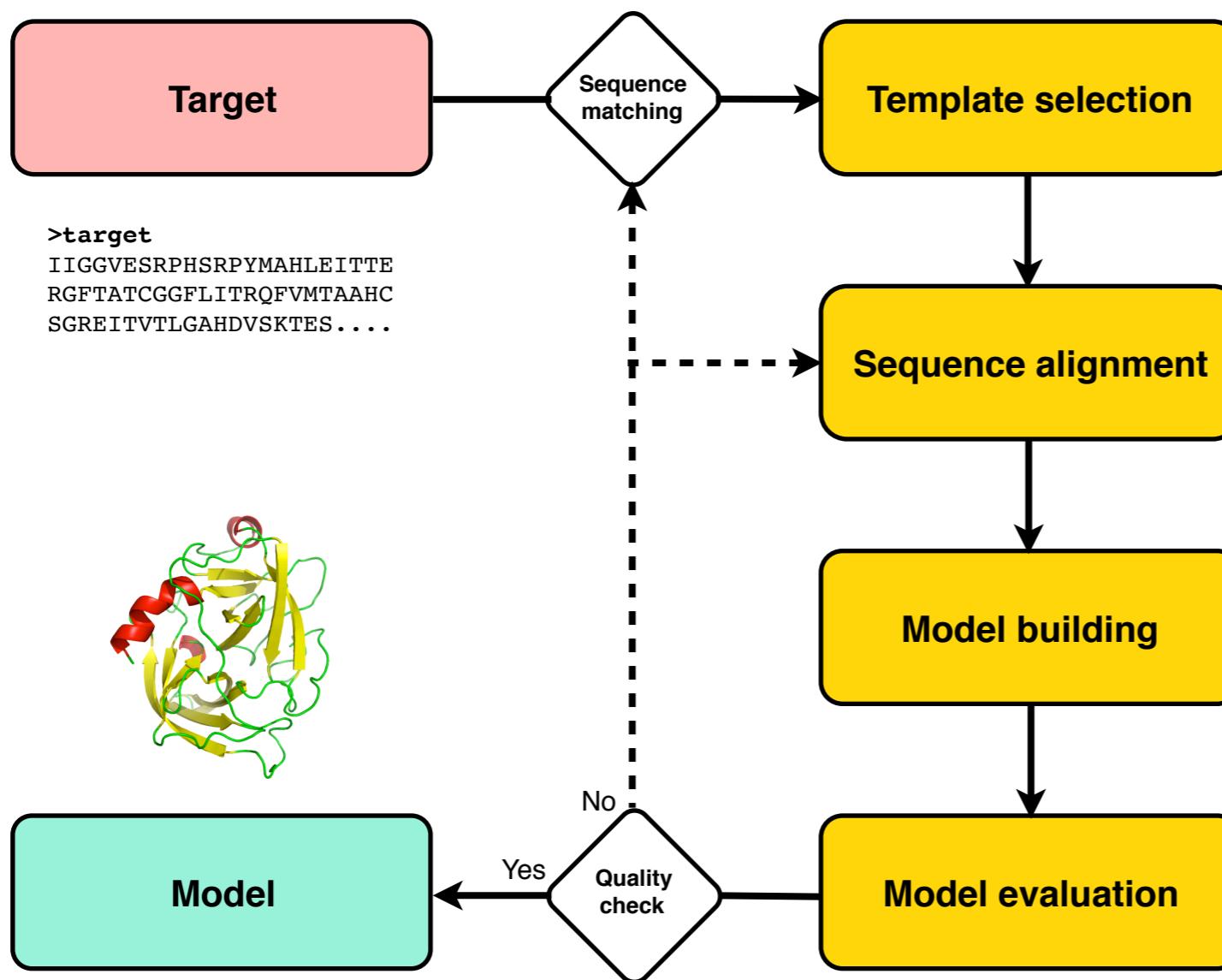
# Twilight Zone

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

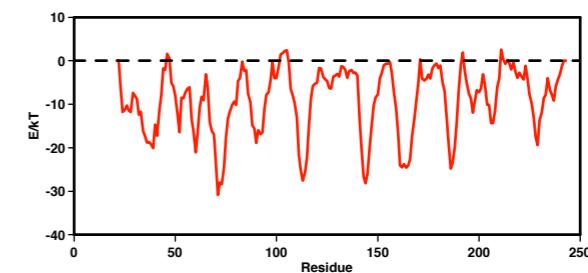
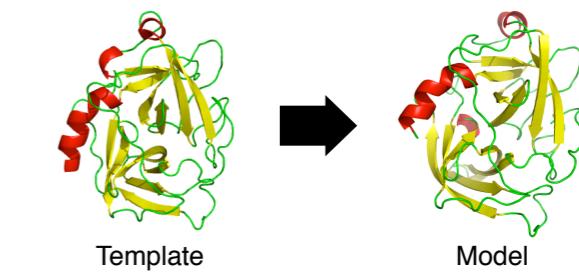


# Comparative Modeling

Flow chart of Comparative Modeling



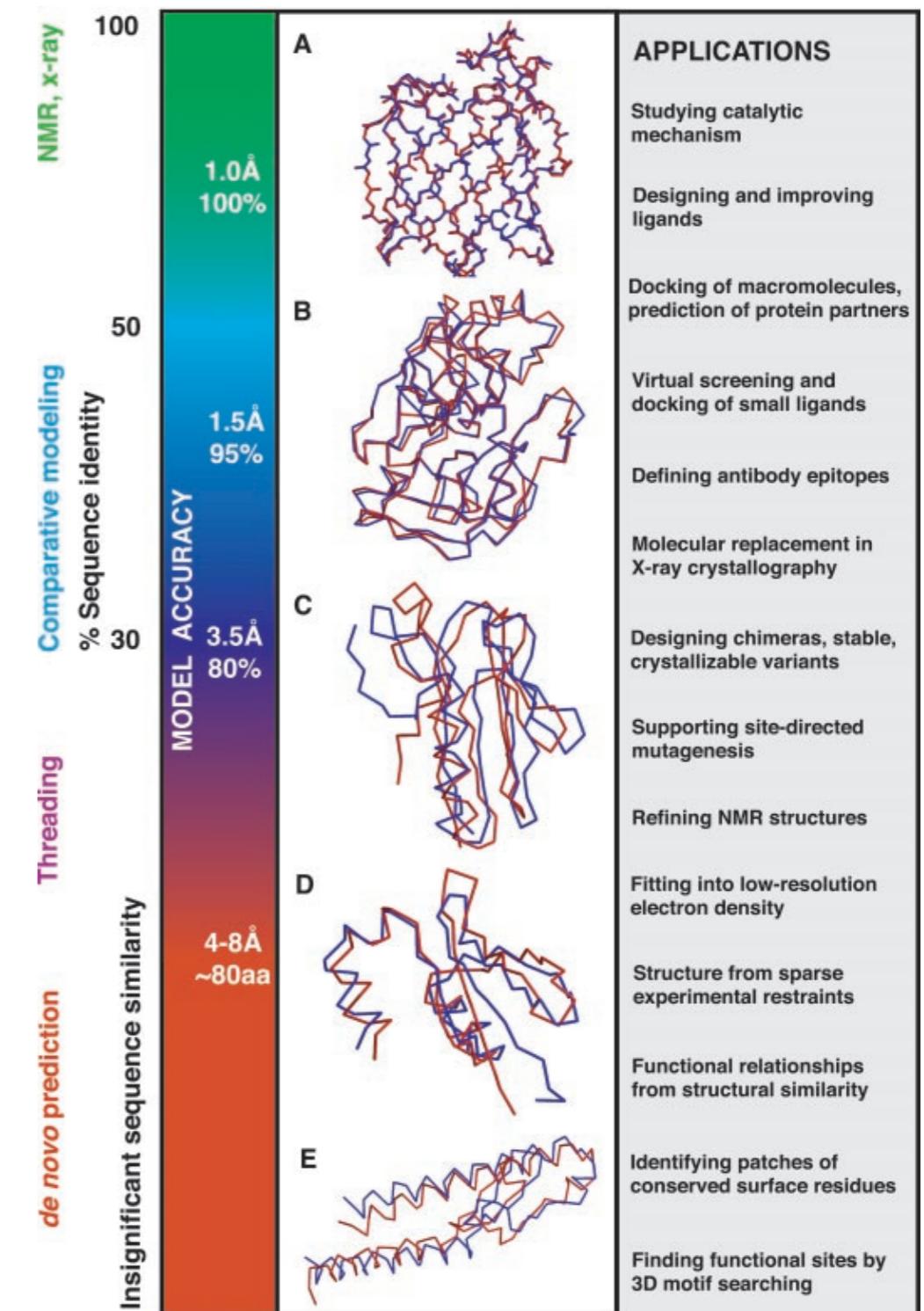
target IIGGVESRPHSRPYMAHLEI  
3RP2A IIGGVESIPHSPYMAHLDI  
target TTERGFTATCGGFLITRQ..  
3RP2A VTEKGLRVICGGFLISRQ..



# 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



# Remote homologs

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

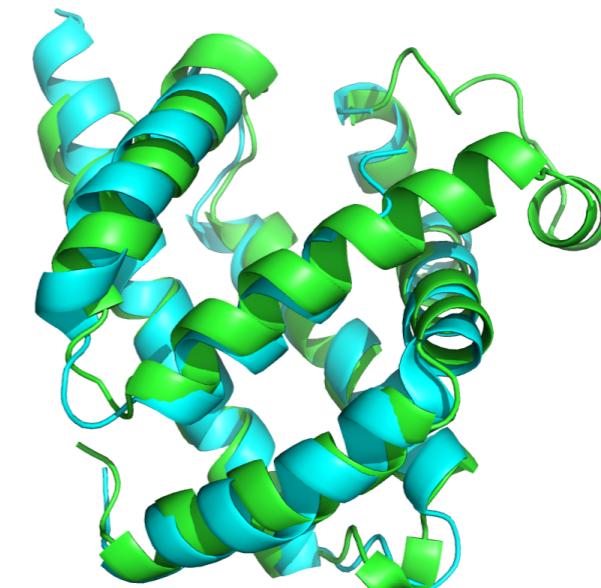
This **DO NOT** exclude that sequences sharing lower identity have similar structures.

**Example:**

Sperm Whale Myoglobin (1JP6:A)

Bacterial Haemoglobin (1VHB:A)

RMSD = 0.18 nm, Identity: 12%



Pairs of 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?

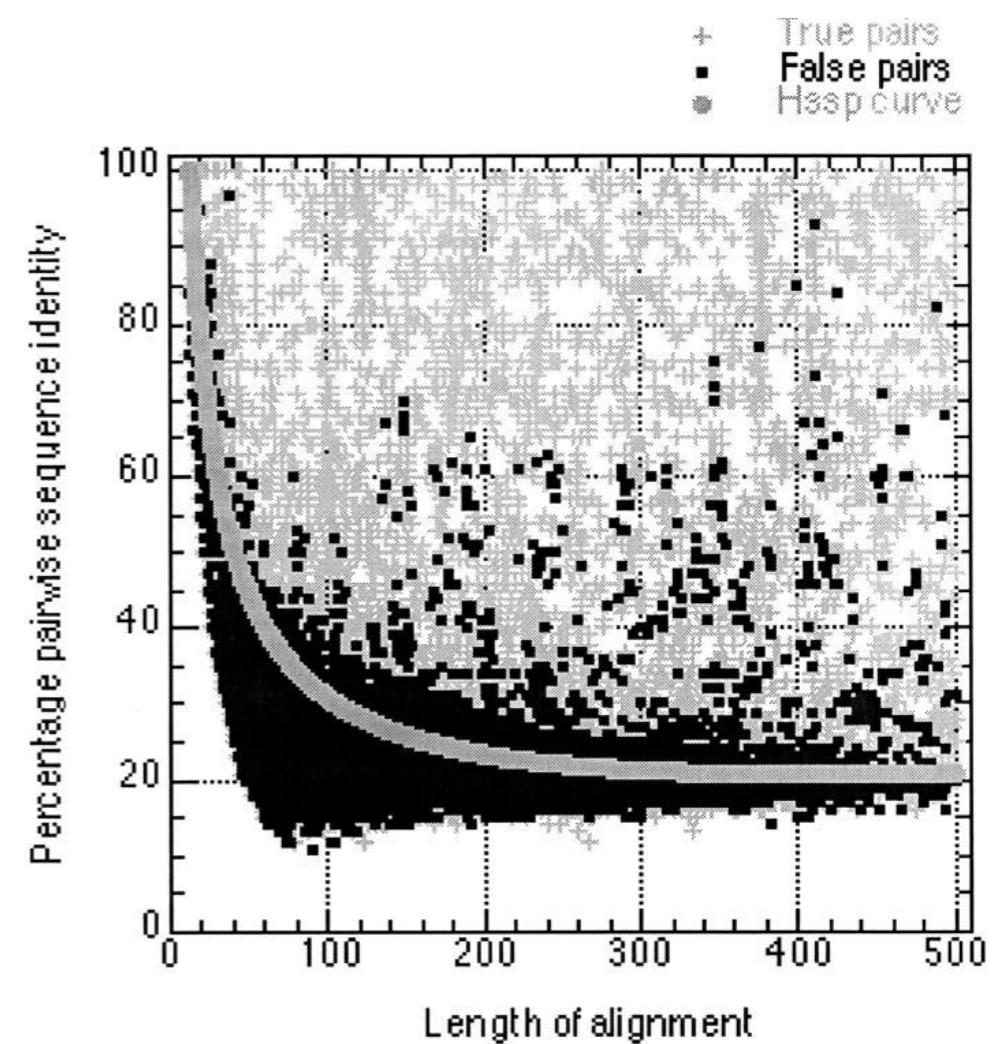
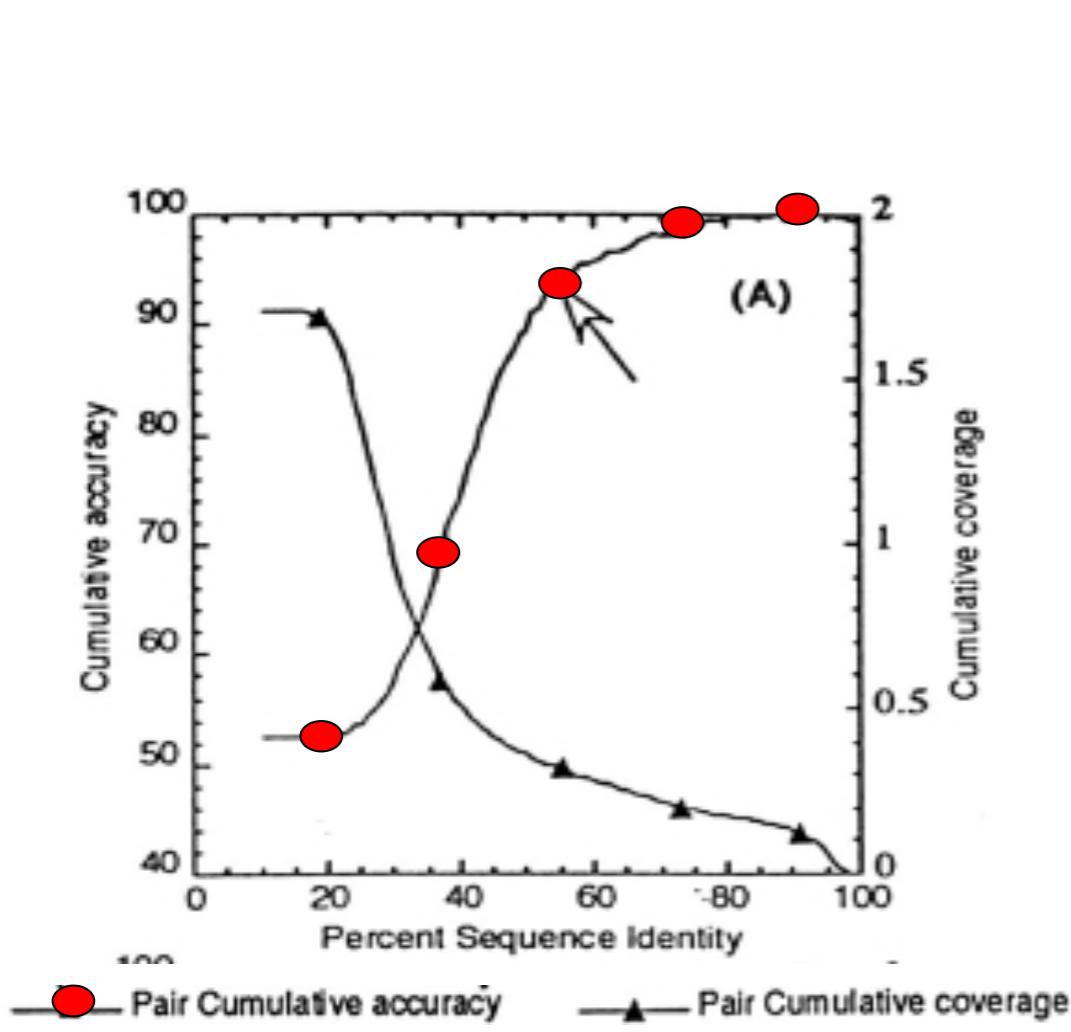
**Solution:** Define a the sequence similarity threshold that allow a reliable transfer of annotation features.

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.



# A false positive

|        |                   |                  |                |                |                 |                  |                |                  |
|--------|-------------------|------------------|----------------|----------------|-----------------|------------------|----------------|------------------|
| sp Q9S | MEFEKIKVINP<br>:: | VVEMDGDEM<br>::  | TRVIWKFI<br>:: | KDKLIFPF<br>:: | LELDIYF<br>::   | DLGLPNRDFT<br>:: | DDKV<br>::     | TI               |
| 10     | 20                | 30               | 40             | 50             | 60              |                  |                |                  |
| sp Q9S | MAFEKIKVAN<br>::  | PIVEMDGDEM<br>:: | TRVIWKS<br>::  | SIKDKLIT<br>:: | PVELDIYF<br>::  | DLGLPHR<br>::    | DATDDK<br>::   | V<br>::          |
| 10     | 20                | 30               | 40             | 50             | 60              |                  |                |                  |
| sp Q9S | ETAEATL<br>::     | KYNVAIK<br>::    | CATITP<br>::   | DEARVREF<br>:: | GLKKMWRSP<br>:: | PNGTIRN<br>::    | NILNGTVF<br>:: | REPIICRN<br>::   |
| 70     | 80                | 90               | 100            | 110            | 120             |                  |                |                  |
| sp Q9S | ESAEATKK<br>::    | YNVAIK<br>::     | CATITP<br>::   | DEGRVTEF<br>:: | GLKQMWRSP<br>:: | PNGTIRN<br>::    | NILNGTVF<br>:: | REPIICKN<br>::   |
| 70     | 80                | 90               | 100            | 110            | 120             |                  |                |                  |
| sp Q9S | RLVPGWT<br>::     | KPKICIGR<br>::   | HAFGDQYR<br>:: | ATDLIVNE<br>:: | PGKLKL<br>::    | VFEPSGS<br>::    | SQKTEFEV<br>:: | FVNFTG-GGV<br>:: |
| 130    | 140               | 150              | 160            | 170            |                 |                  |                |                  |
| sp Q9S | KLVPGWT<br>::     | KPKICIGR<br>::   | HAFGDQYR<br>:: | ATDAVIKG<br>:: | PGKLTMTF<br>::  | --GKDGT<br>::    | ETEVFTFT<br>:: | GEGGV<br>::      |
| 130    | 140               | 150              | 160            | 170            |                 |                  |                |                  |
| sp Q9S | 180               | 190              | 200            | 210            | 220             | 230              |                |                  |
| sp Q9S | ALAMYNT<br>::     | DESIRAF<br>::    | AESSMYT<br>::  | AYQKKW<br>::   | PLYLST<br>::    | KNTILKI<br>::    | YDGRFKD<br>::  | IFQE<br>::       |
| 180    | 190               | 200              | 210            | 220            | 230             |                  |                |                  |
| sp Q9S | 240               | 250              | 260            | 270            | 280             | 290              |                |                  |
| sp Q9S | YEAA<br>::        | GIWYE<br>::      | HRLIDDM<br>::  | VAYAMK<br>::   | SEGGYV<br>::    | WACKNY<br>::     | DGDVQSDF<br>:: | LAQGYG<br>::     |
| sp Q9S | YDAAGI<br>::      | WIWYE<br>::      | HRLIDDM<br>::  | VAYALK<br>::   | SEGGYV<br>::    | WACKNY<br>::     | DGDVQSDF<br>:: | LAQGFG<br>::     |
| 240    | 250               | 260              | 270            | 280            | 290             |                  |                |                  |
| sp Q9S | 300               | 310              | 320            | 330            | 340             | 350              |                |                  |
| sp Q9S | DGKTIE<br>::      | AEAAHGT<br>::    | TVTRHY<br>::   | RHQKG<br>::    | GETSTNS<br>::   | IASIFAWS<br>::   | RGLAHR<br>::   | AKLDSNA<br>::    |
| 300    | 310               | 320              | 330            | 340            | 350             |                  |                |                  |
| sp Q9S | 360               | 370              | 380            | 390            | 400             | 410              |                |                  |
| sp Q9S | LEAACMGT<br>::    | VESGK<br>::      | MTKDL<br>::    | ALLIHG<br>::   | AKVRRD<br>::    | QYVNTE<br>::     | EFIGDA<br>::   | VAWELK<br>::     |
| 360    | 370               | 380              | 390            | 400            | 410             |                  |                |                  |
| sp Q9S | LEAACVGT<br>::    | VESGK<br>::      | MTKDL<br>::    | ALLIHG<br>::   | SKLSR<br>::     | DTYLNTE<br>::    | EFIGDA<br>::   | AAELKERL<br>::   |
| 360    | 370               | 380              | 390            | 400            | 410             |                  |                |                  |

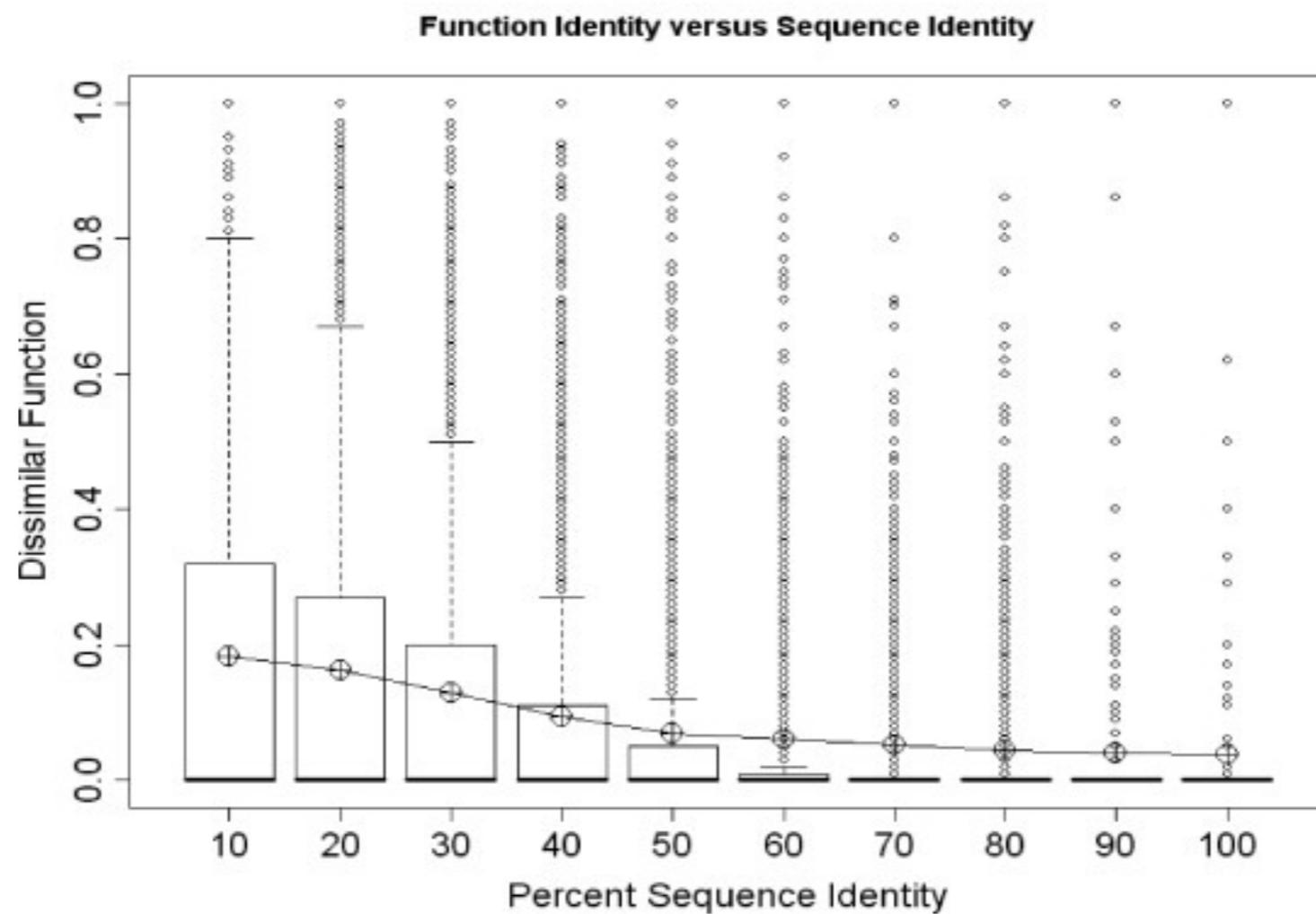
**Q9SLK0 (ICDHX\_ARATH):**  
**Peroxisomal** isocitrate dehydrogenase

**Q9SRZ6 (ICDHC\_ARATH):**  
**Cytosolic** isocitrate dehydrogenase

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

# Functional Annotation

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



# Dissimilar functions

# P04385 (GAL1\_YEAST) Galactokinase

## Catalytic activity

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

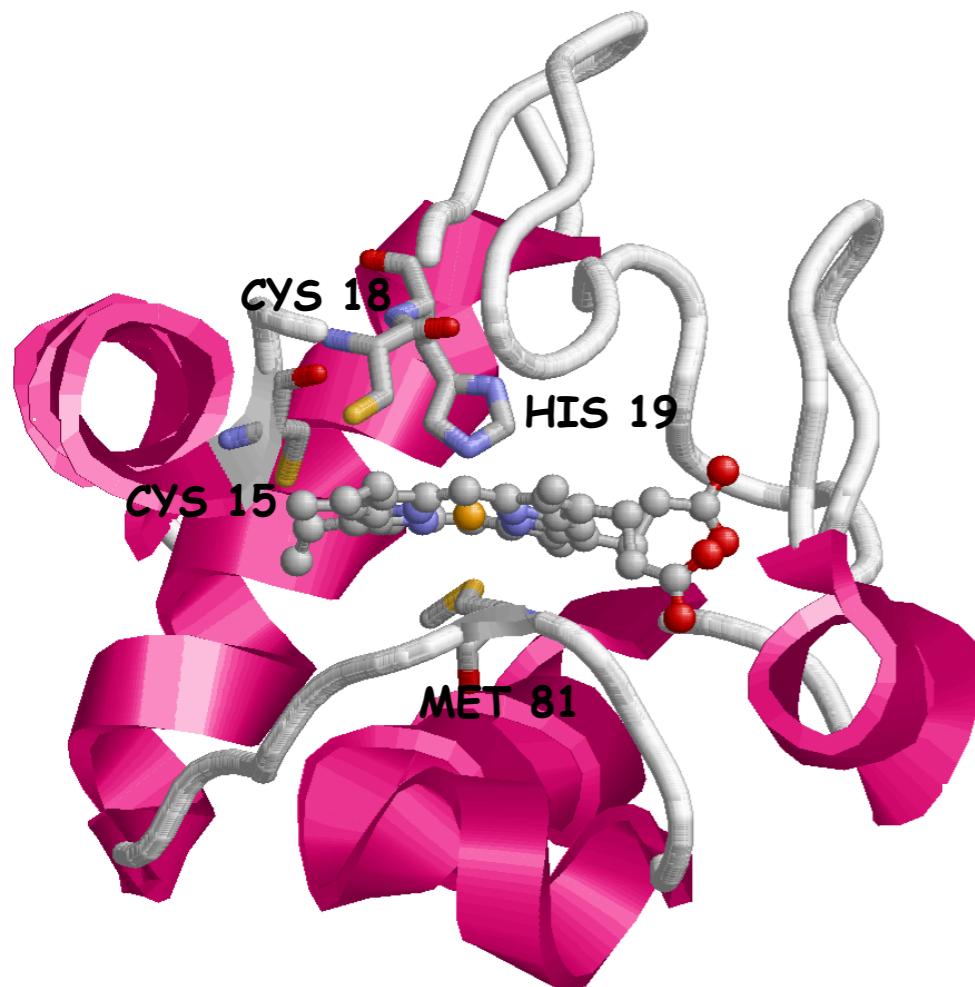
# P13045 (GAL3\_YEAST) Protein GAL3

The GAL3 regulatory function is required for rapid induction of the galactose system.

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

# 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.



| Feature key                | Position(s)             | Length | Description              |
|----------------------------|-------------------------|--------|--------------------------|
| Binding site <sup>i</sup>  | <a href="#">15 – 15</a> | 1      | Heme (covalent)          |
| Binding site <sup>i</sup>  | <a href="#">18 – 18</a> | 1      | Heme (covalent)          |
| Metal binding <sup>i</sup> | <a href="#">19 – 19</a> | 1      | Iron (heme axial ligand) |
| Metal binding <sup>i</sup> | <a href="#">81 – 81</a> | 1      | Iron (heme axial ligand) |

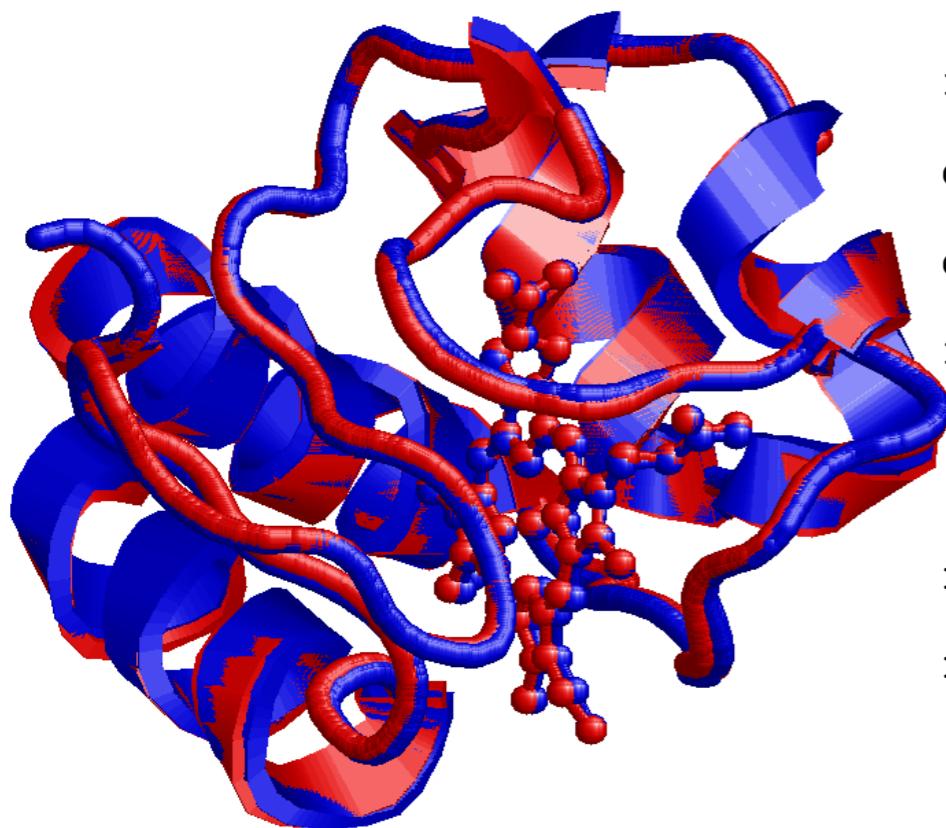
PDB: 3zcf:A

# 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 |
|                                                                       | .           | .      | .      |
| GDVEKGKKIFIMK <b>CSQCH</b> TVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWG | EDETLMEYLEN |        |        |
| :   :   .                                                             | :   .       | :   .  | :   .  |
| GDVEKGKKIFVQ <b>KCAQCH</b> TVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITW  | KEETLMEYLEN |        |        |
|                                                                       | .           | .      | .      |
| 80 : A                                                                | 100 : A     |        |        |
| .                                                                     | .           | .      | .      |
| PKKYIPG <b>TKM</b> IFVGIKKKEERADLIAYLKKATNE                           |             |        |        |
| :   :   .                                                             | :   .       | :   .  | :   .  |
| PKKYIPG <b>TKM</b> IFAGIKKKTEREDLIAYLKKATNE                           |             |        |        |
| .                                                                     | .           | .      | .      |
| 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)

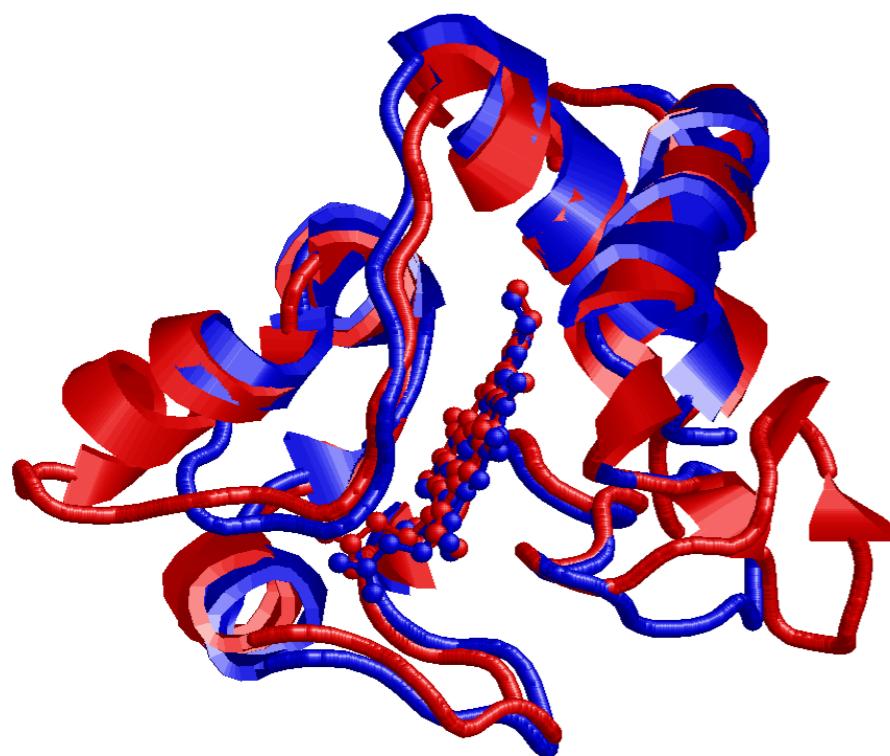
|       |                                                                       |         |         |         |         |         |
|-------|-----------------------------------------------------------------------|---------|---------|---------|---------|---------|
|       | 10                                                                    | 20      | 30      | 40      | 50      | 60      |
| Homo  | MGDVEKGKKIFIMK <u>CSQCH</u> TVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIW  |         |         |         |         |         |
|       | :::::::                                                               | ::::::: | ::::::: | ::::::: | ::::::: | ::::::: |
| Horse | MGDVEKGKKIFVQK <u>CAQCH</u> TVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITW |         |         |         |         |         |
|       | 10                                                                    | 20      | 30      | 40      | 50      | 60      |
|       | 70                                                                    | 80      | 90      | 100     |         |         |
| Homo  | GEDTLMEYLENPKKYIPGTM <u>I</u> FVGIKKKKEERADLIAYLKKATNE                |         |         |         |         |         |
|       | :::::::                                                               | ::::::: | ::::::: | ::::::: | ::::::: | ::::::: |
| Horse | KEETLMEYLENPKKYIPGTM <u>I</u> FAGIKKKTEREDLIAYLKKATNE                 |         |         |         |         |         |
|       | 70                                                                    | 80      | 90      | 100     |         |         |

# 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



# Sequence vs Structure (I)

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

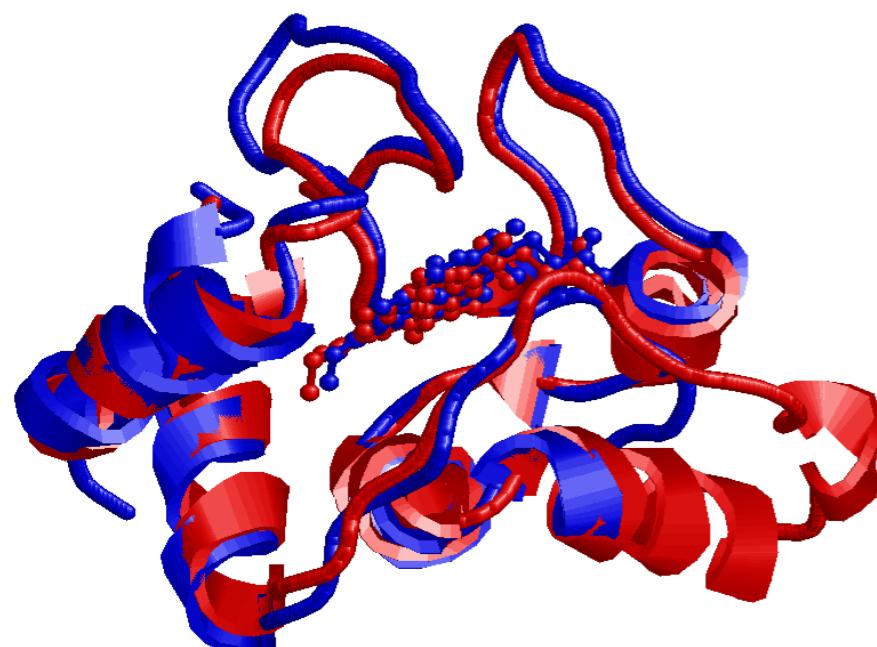
Structural alignment:  
RMSD= 0,18 nm  
28% sequence identity

# Homo vs Rhodobacter Pal.

# Human Cytochrome C - Uniprot:P99999. PDB: 3ZCF:A

# Cytochrome C2 Rhodopseudomonas pal. – Uniprot: P00091. PDB 1I8O:A

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



# 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

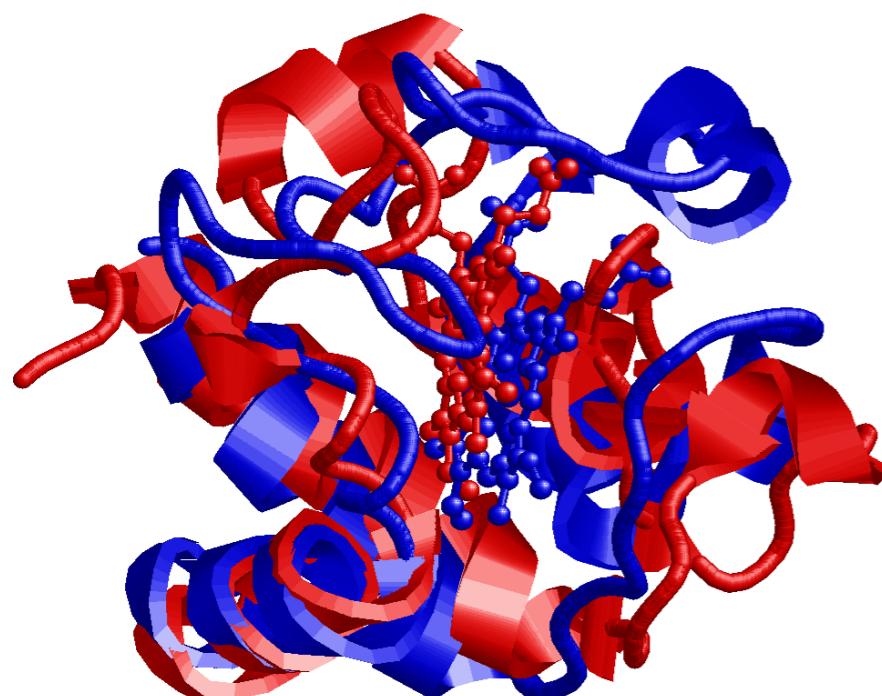
|          |                                                        |                     |                     |                     |                     |     |
|----------|--------------------------------------------------------|---------------------|---------------------|---------------------|---------------------|-----|
| sp   P99 | 10                                                     | 20                  | 30                  |                     |                     |     |
|          | <u>MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGL</u>            |                     |                     |                     |                     |     |
|          | ... : . . :                                            | . . . : . . .       | . . . : . . .       | . . . : . . .       | . . . : . . .       |     |
| sp   P00 | <u>MVKKLLTILSIAATAGSLSIGTASAQDAKAGEAVF</u> ---         |                     |                     |                     |                     |     |
|          | 10                                                     | 20                  | 30                  | 40                  | 50                  |     |
|          | <u>KQCMTCHRADKNMVGPA</u> LGGV                          |                     |                     |                     |                     |     |
|          | 40                                                     | 50                  | 60                  | 70                  | 80                  | 90  |
| sp   P99 | <u>FGRKTGQAPGYSYTAANKNKG</u> ---                       |                     |                     |                     |                     |     |
|          | : . . : . . . : . . .                                  | . . : . . . : . . . | . . : . . . : . . . | . . : . . . : . . . | . . : . . . : . . . |     |
| sp   P00 | <u>VGRKAGTAAGFTYSPLNHNSGEAGLVWTADNIINYLNDPNAFL</u> --- |                     |                     |                     |                     |     |
|          | 60                                                     | 70                  | 80                  | 90                  | 100                 | 110 |
|          | <u>KKFLTDKGKADQAV</u>                                  |                     |                     |                     |                     |     |
|          | 100                                                    |                     |                     |                     |                     |     |
| sp   P99 | DLIAYLKKATNE                                           |                     |                     |                     |                     |     |
|          | . . : . . :                                            |                     |                     |                     |                     |     |
| sp   P00 | <u>GVTKMTFKLANEQQRKDVVAYLATLK</u>                      |                     |                     |                     |                     |     |
|          | 120                                                    | 130                 |                     |                     |                     |     |

# 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



# Sequence vs Structure (III)

In this case the sequence alignment is significantly different from 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    | <u>MGDVEKGKKIFIMKCSQCHTVEKGGKHKTG</u>                                        |    |    |
|         | : . . : . : : . : . : . : .                                                  |    |    |
| A. Thal | DFILLKKLAPPLTAVLLAVSPICFPPE <u>SLGQTLDIQRGATLFNRACIGCHDT-GGNIIQPG</u>        |    |    |
|         | 50            60            70            80            90            100    |    |    |
|         | 40            50            60            70            80            90     |    |    |
| sp P99  | <u>PNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLENPKKYIPG</u> <b>T</b> KMIFVGIKKKE |    |    |
|         | . :        . . :        : . . . . . . . :        : . : . : . : . . .         |    |    |
| sp Q93  | <u>ATLFTKDLERNGVD-----TEEEIYRVTYFGKGRMPGFGE---KCTPRGQCTF-GPR</u> LQD         |    |    |
|         | 110            120            130            140            150              |    |    |
|         | 100                                                                          |    |    |
| ;       |                                                                              |    |    |
| sp P99  | ERADLIAYLKKATNE                                                              |    |    |
|         | . . . . : . .                                                                |    |    |
| sp Q93  | EEIKLLAEFVKFQADQGWPTVSTD                                                     |    |    |
|         | 160            170                                                           |    |    |

# 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.