

Introduction and Basic Concepts

Laboratory of Bioinformatics I
Module 2

Emidio Capriotti

<http://biofold.org/>



Biomolecules
Folding and
Disease

Department of Pharmacy and
Biotechnology (FaBiT)
University of Bologna



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.

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

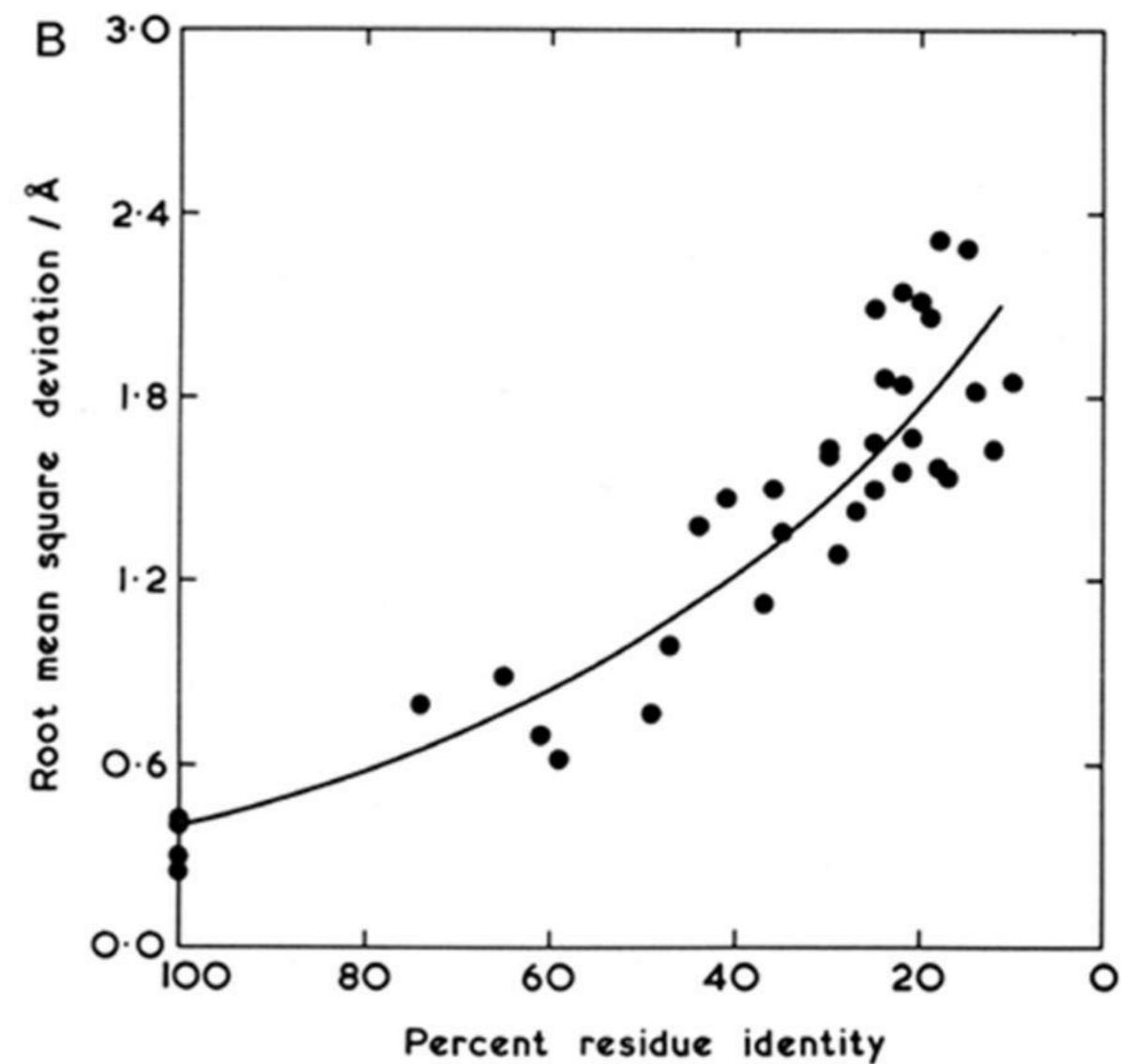
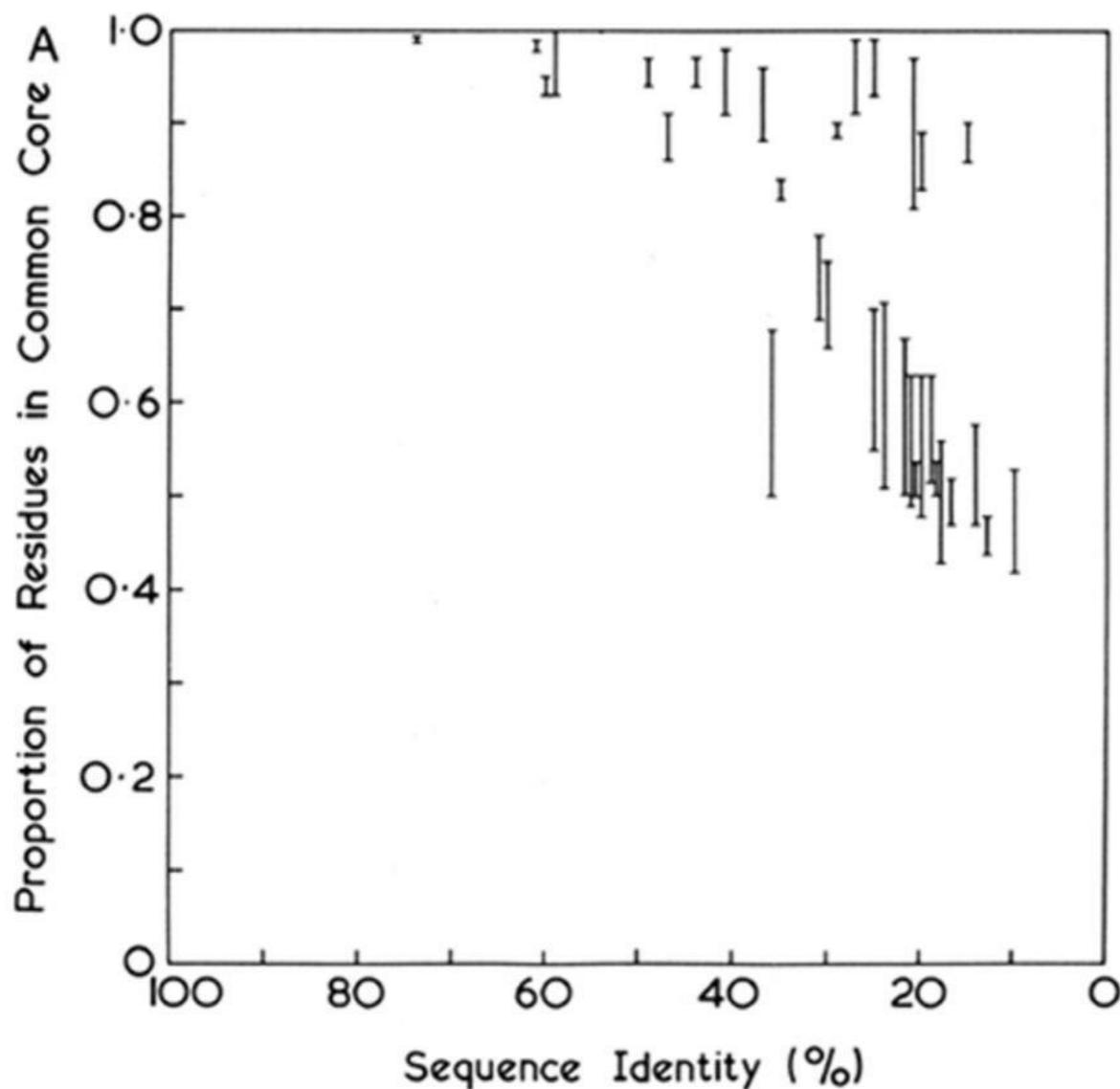
$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^n (a_i - b_i)^2}{n}}$$

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 $s(i,j)$

BLOSUM62

A: ALASVLIRLITRLYP
B: ASAVALNRLITRLYP

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

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

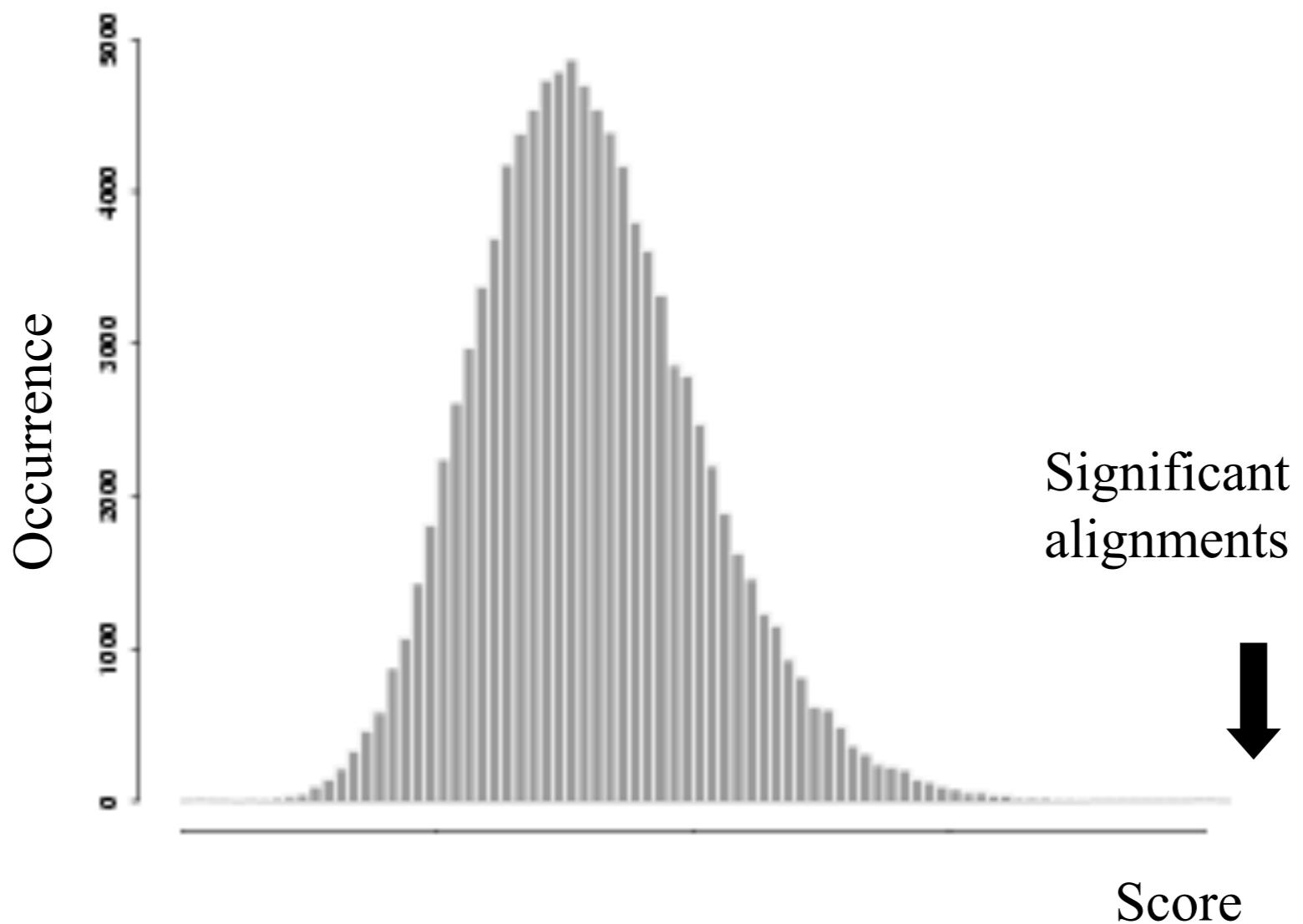
Algorithms for finding the **minimum distance** between two sequences

- **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

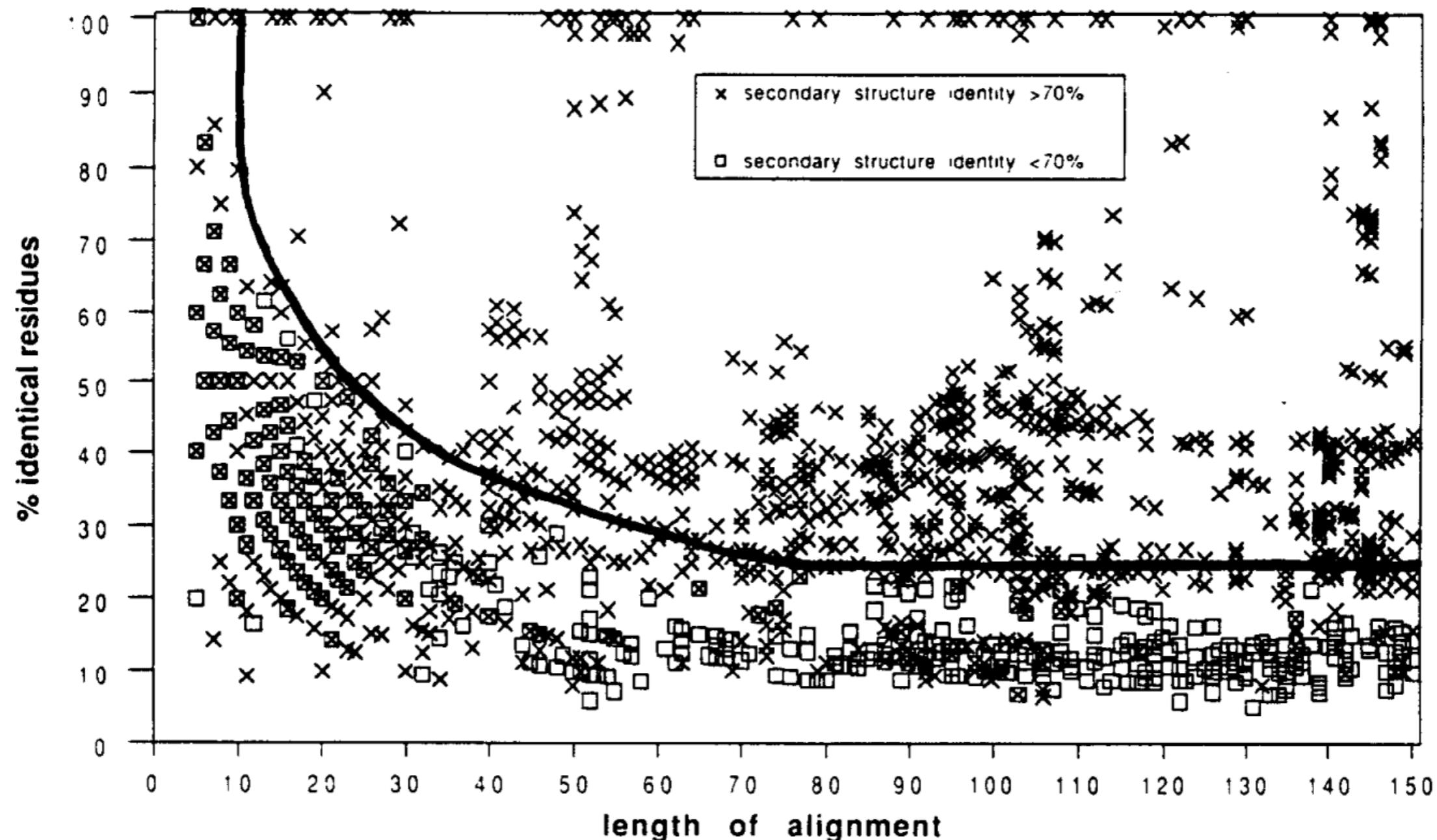
Given an alignment with score S , is it significant?

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



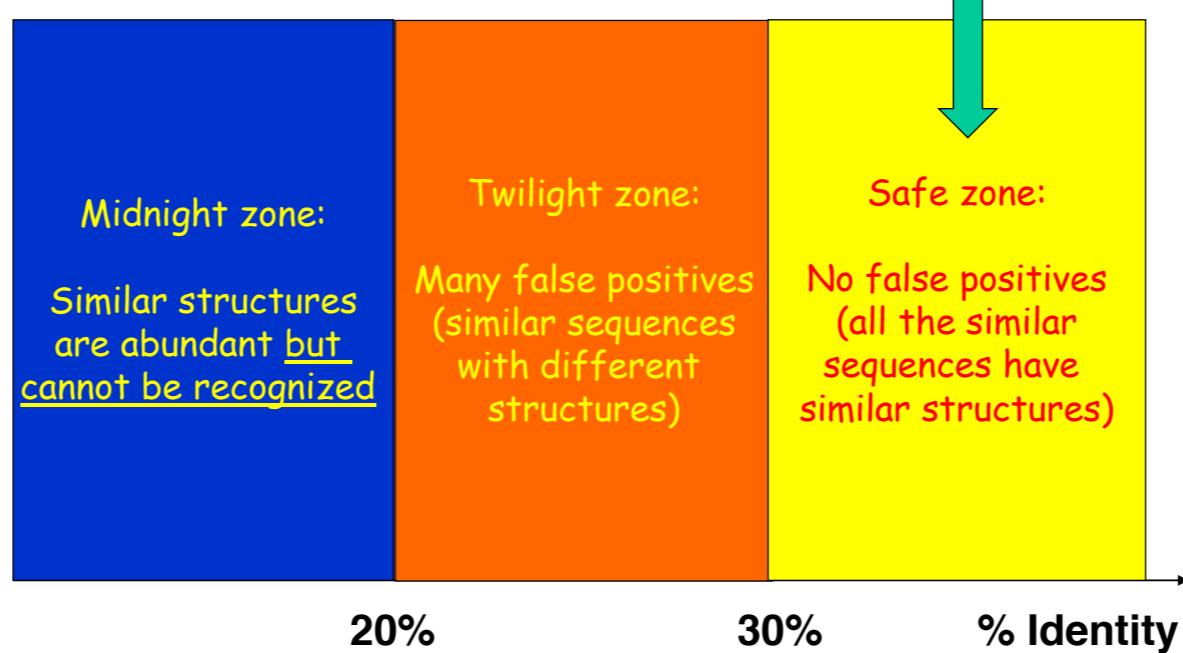
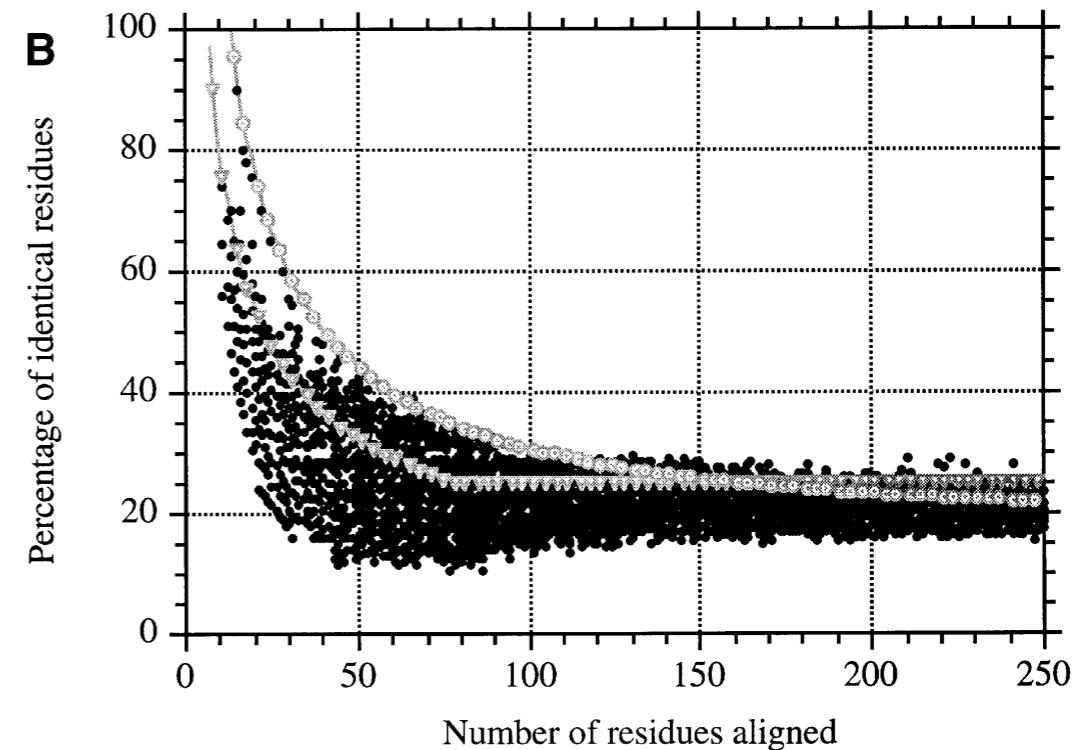
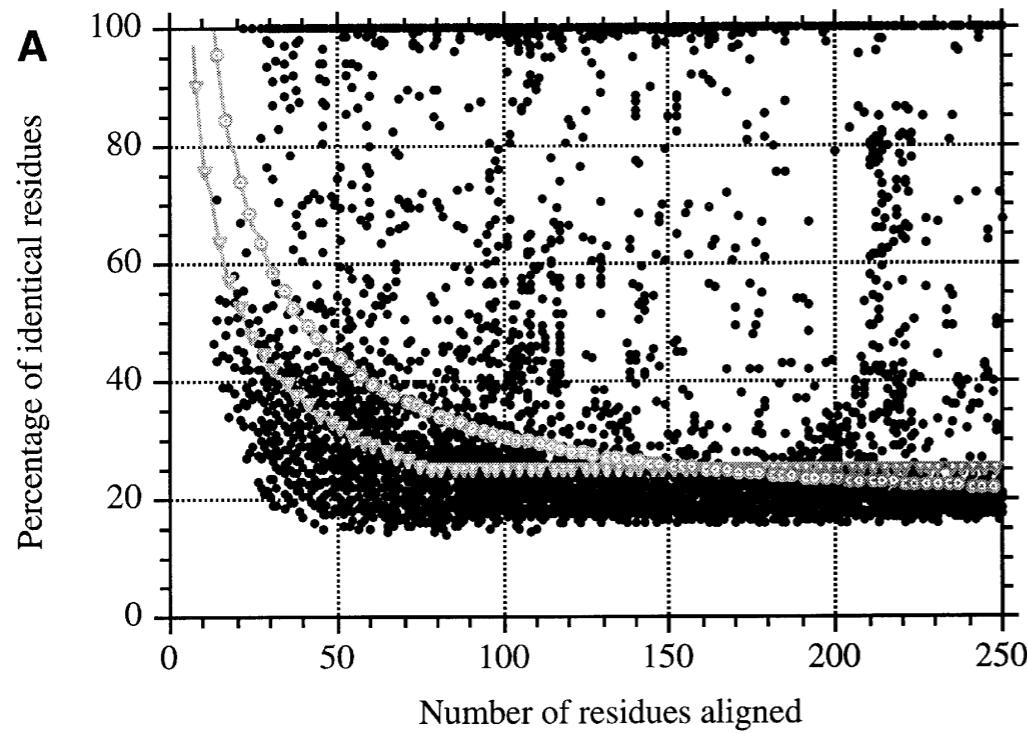
Structural Homology

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



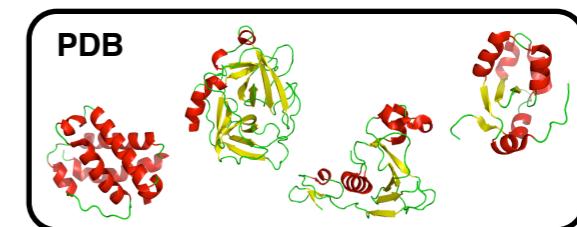
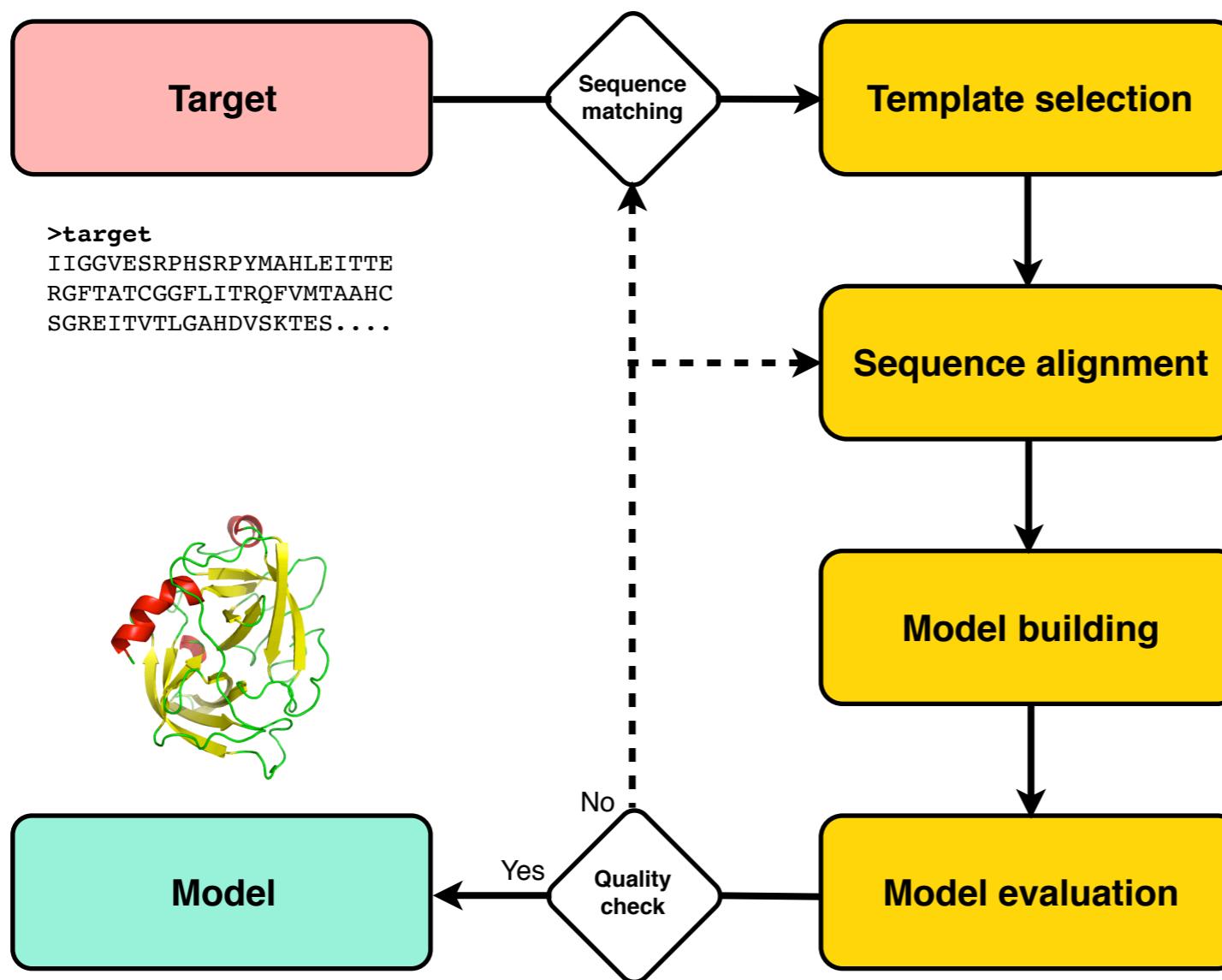
Twilight Zone

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

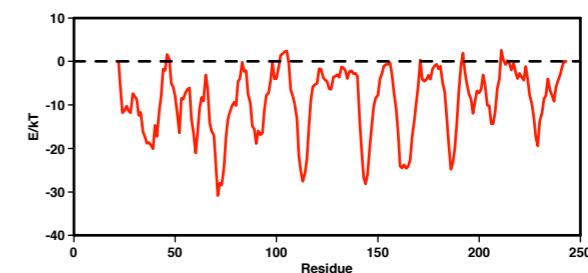
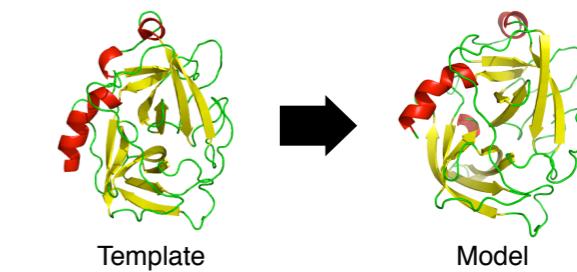


Comparative Modeling

Flow chart of Comparative Modeling



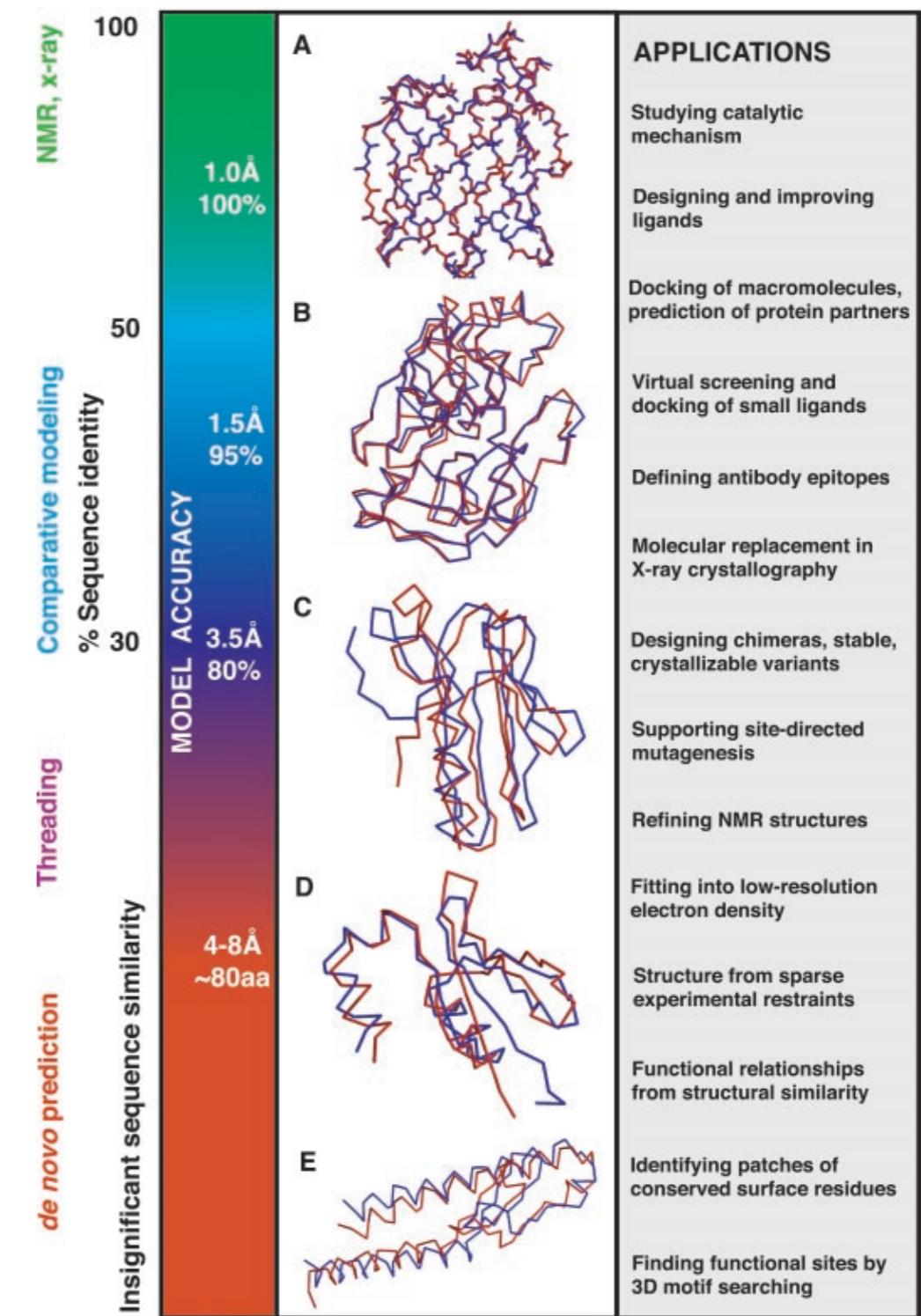
target IIGGVESRPHSRPYMAHLEI
3RP2A IIGGVESIPHSPYMAHLDI
target TTERGFTATCGGFLITRQ..
3RP2A VTEKGLRVICGGFLISRQ..



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).

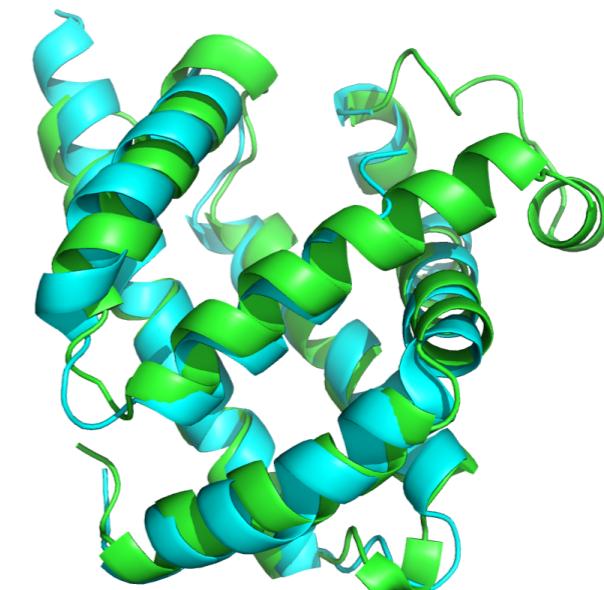
This **DO NOT** exclude that sequences sharing lower identity have similar structures.

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”

Sequence Identity Inference

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

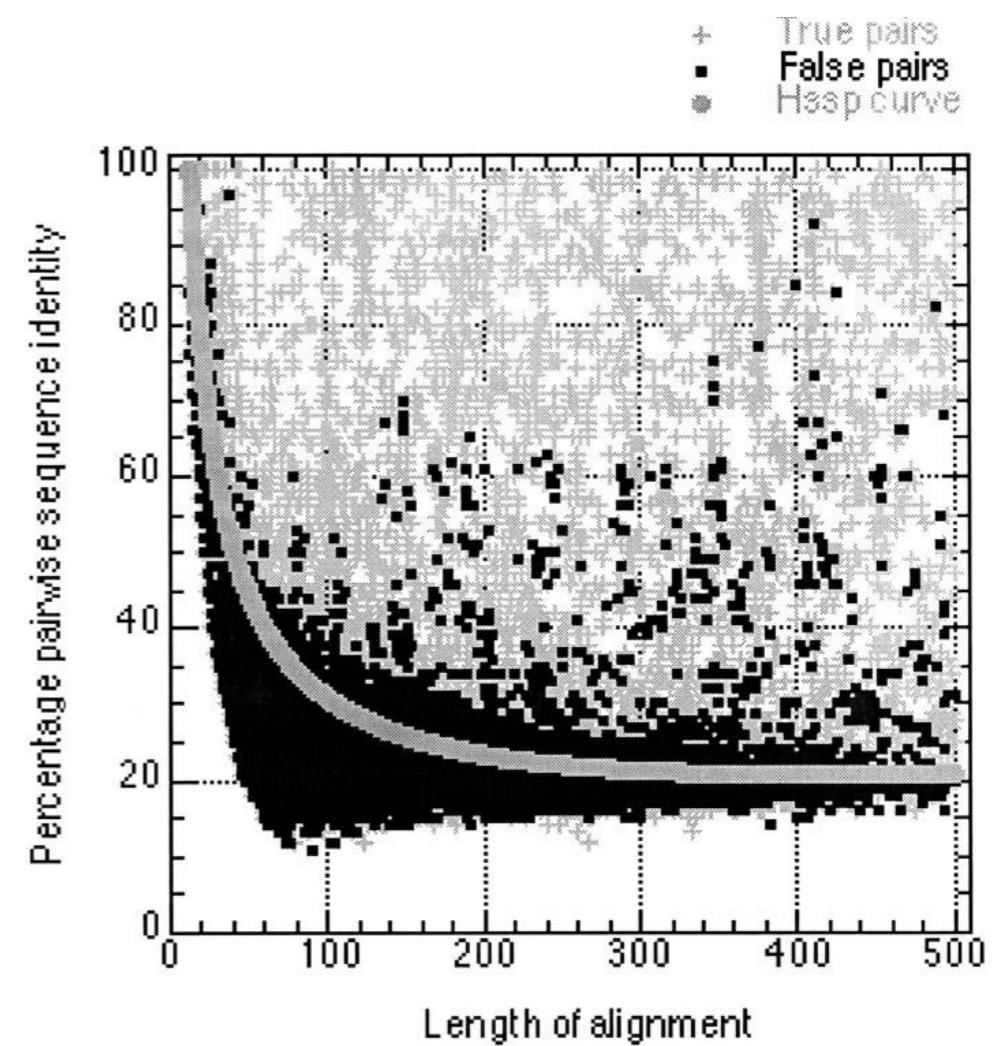
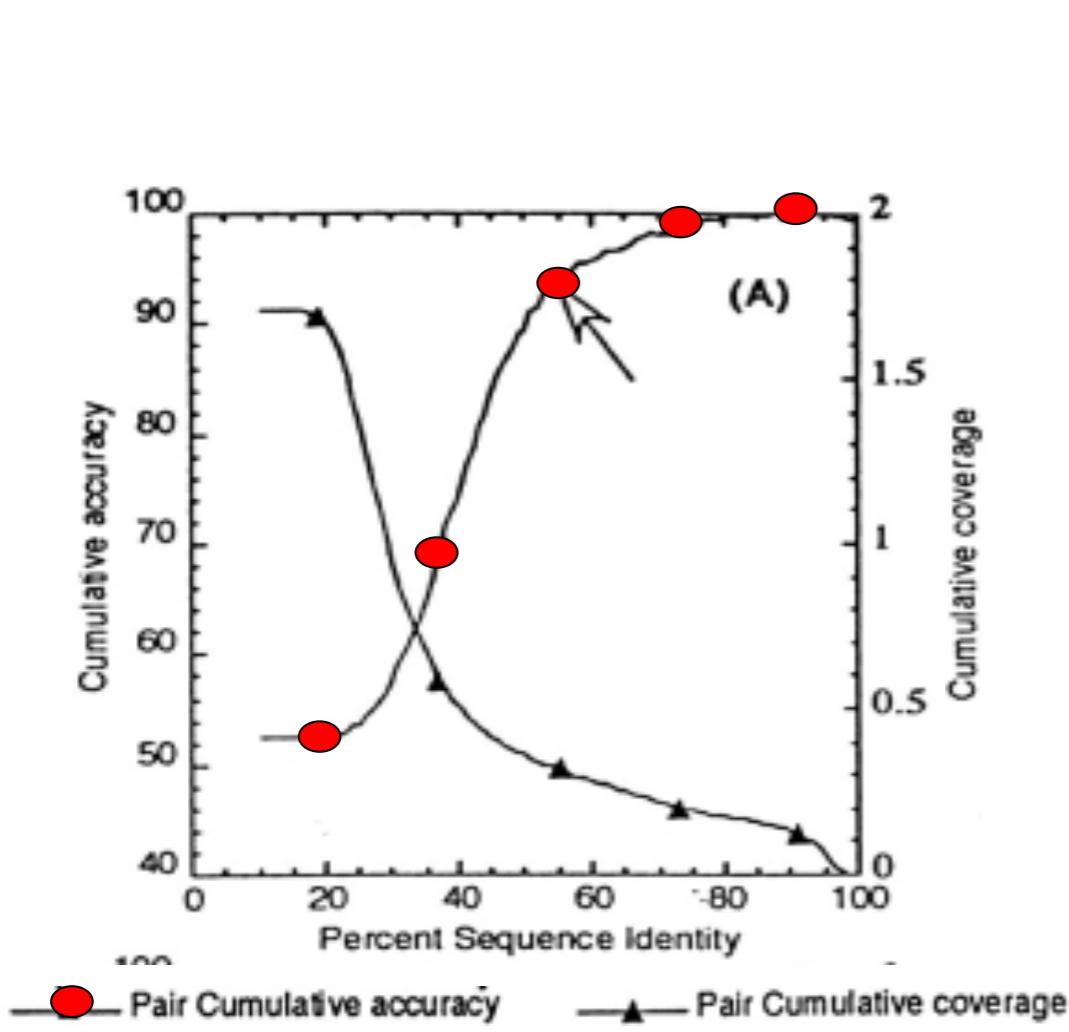
Solution: Define a the sequence similarity threshold that allow a reliable transfer of annotation features.

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 ::	VVEMDGDEM ::	TRVIWKFI :::	KDKLIFPF :::	LELDI :::	KYFDLGLP :::	NRDFTDD :::	KVTI ::
10	20	30	40	50	60			
sp Q9S	MAFEKIKVAN ::	PIVEMDGDEM ::	TRVIWKS :::	IKDKLITP :::	FVELDI :::	KYFDLGLP :::	HRRDATDD :::	KVTI ::
10	20	30	40	50	60			
sp Q9S	ETAEATLKYN ::	VAIKCATITP ::	DEARVREFGL :::	KMMWRSP :::	PNGTIRN :::	ILNGTVFRE :::	PIICRNIP ::	
70	80	90	100	110	120			
sp Q9S	ESAEATKKYN ::	VAIKCATITP ::	DEGRVTEFGL :::	KQMWRSP :::	PNGTIRN :::	ILNGTVFRE :::	PIICKNVP ::	
70	80	90	100	110	120			
sp Q9S	RLVPGWT :::	KPKICIGR :::	HAFGDQYR :::	ATDLIVNE :::	PGKLKL :::	VFEPGS :::	SQKTEFEV :::	FNFTG-GGV ::
130	140	150	160	170				
sp Q9S	KLVPGWT :::	KPKICIGR :::	HAFGDQYR :::	ATDAVIKG :::	PGKLTMTF --	GKDGTETEV :::	FTFTGEGGV ::	
130	140	150	160	170				
sp Q9S	180	190	200	210	220	230		
sp Q9S	ALAMYNT :::	DESIRAF :::	AESSMYT :::	AYQKKW :::	PLYLST :::	KNTILKI :::	DGRFKD :::	IFQE :::
180	190	200	210	220	230			
sp Q9S	240	250	260	270	280	290		
sp Q9S	YEAA :::	AGI :::	WYE :::	HLI :::	DDMV :::	AYAM :::	KSEGGY :::	VWACK :::
sp Q9S	240	250	260	270	280	290		
sp Q9S	300	310	320	330	340	350		
sp Q9S	DGKTIE :::	EA :::	AA :::	HGT :::	TVTR :::	RHQ :::	KG :::	GET :::
300	310	320	330	340	350			
sp Q9S	360	370	380	390	400	410		
sp Q9S	LEAACM :::	GTV :::	ESGK :::	MKT :::	KDL :::	ALLI :::	HGA :::	VRRD :::
360	370	380	390	400	410			
sp Q9S	LEAACV :::	GT :::	ESGK :::	MKT :::	KDL :::	ALLI :::	HGSK :::	LRTYLN :::
360	370	380	390	400	410			

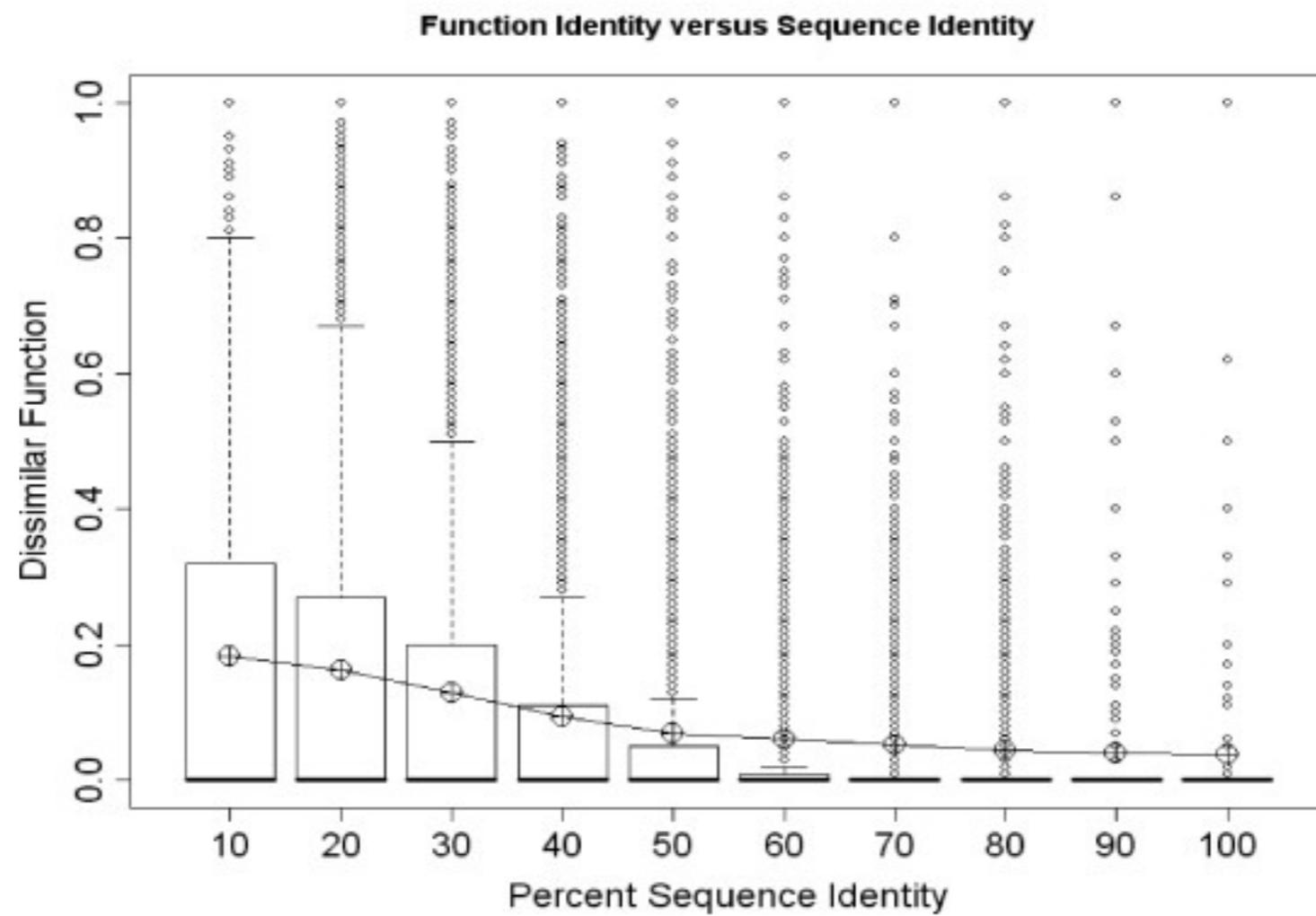
Q9SLK0 (ICDHX_ARATH):
Peroxisomal isocitrate dehydrogenase

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

sp P04	MTKSHSEEVIVPEFVNSSAKELPRPLAECPSI KKFISAYDAKPDFVARSPGRVNLIGEH	10	20	30	40	50	60
sp P13	MNTN-----VPIFSSPVRLP RSFEQKHLAVVDAFFQTYHVKPDFIARSPGRVNLIGEH	10	20	30	40	50	
sp P04	IDYCDFSVLPLAIDFDMLCAV KVLNEKNPSITLINADPKFAQRKFDLPLDGSYVTIDPSV	70	80	90	100	110	120
sp P13	IDYCDFSVLPLAIDVDMLCAV KILDEKNPSITLTNA DPKFAQRKFDLPLDGSYMAIDPSV	60	70	80	90	100	110
sp P04	SDWSNYFKCGLHVAHSFLKKLAPERFASAPLAGLQVFCEGDVPTGSGLSSAA FIACVAL	130	140	150	160	170	180
sp P13	SEWSNYFKCGLHVAHSY LKKIAPERFNN TPLVGAQIFCQSDIPTGGGLSS--AFTCAAAL	120	130	140	150	160	170
sp P04	AVVKANMGP GYHMSKQNLMRITVV AHEYVGVNNGMDQAASVC GEE H ALYVE FPQLKA	190	200	210	220	230	240
sp P13	ATIRANMGK NFDISK KKDLTRITAVA EHEYVGVNNGMDQ ATSVY GEE H ALYVE FRPKLKA	180	190	200	210	220	230
sp P04	TPFKFPQLKNHEISFVIANTLV VSNKFETAP TNYNLRV VEVT TAANVLA TYGV VLLSGK	250	260	270	280	290	300
sp P13	TPFKFPQLKNHEISFVIANTLV KSNKFETAP TNYNLRV IEVT VAA NALATRY SVALPSHK	240	250	260	270	280	290
sp P04	EGSSTNKG NLRDFMN VYYARYHNISTPW NGDIES GERLTKMLVLVEESLANKKQGFSVD	310	320	330	340	350	360
sp P13	DNSNSERGN LRDFMDAY YARYENQAQP WNGDIGT GIERLLKMLQLV EESFSRK KSGFTVH	300	310	320	330	340	350
sp P04	DVAQSLNC SREEFTRDYL TTSPVRFQVLKLYQRAKH VYSES LRVLKAVKLM TTASFTA	370	380	390	400	410	420
sp P13	EASTALNC SREEFTRDYL TTFPVRFQVLKLYQRAKH VYSES LRVLKALKMM TSATFHT DE	360	370	380	390	400	410
sp P04	DFFKQFGALMNESQASC DKLYEC SCPEIDKIC SIAL NSNGSYGSRLTGAGWG GCTV HLVPG	430	440	450	460	470	480
sp P13	DFFTDFGRLMNESQASC DKLYEC CSIETNQIC SIAL ANGSG FSR LTGAGWG GCTI HLVPS	420	430	440	450	460	470
sp P04	GPN GNIEKVKEALANE FYKV KYP KITDAELEN AIIVSKP ALGSCLYEL	490	500	510	520		
sp P13	GANGN VEQVRKALIEKF YNVR YPD LTDEELKD AIIVSKP ALGT CLYE Q	480	490	500	510	520	

P04385 (GAL1_YEAST) Galactokinase

Catalytic activity

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

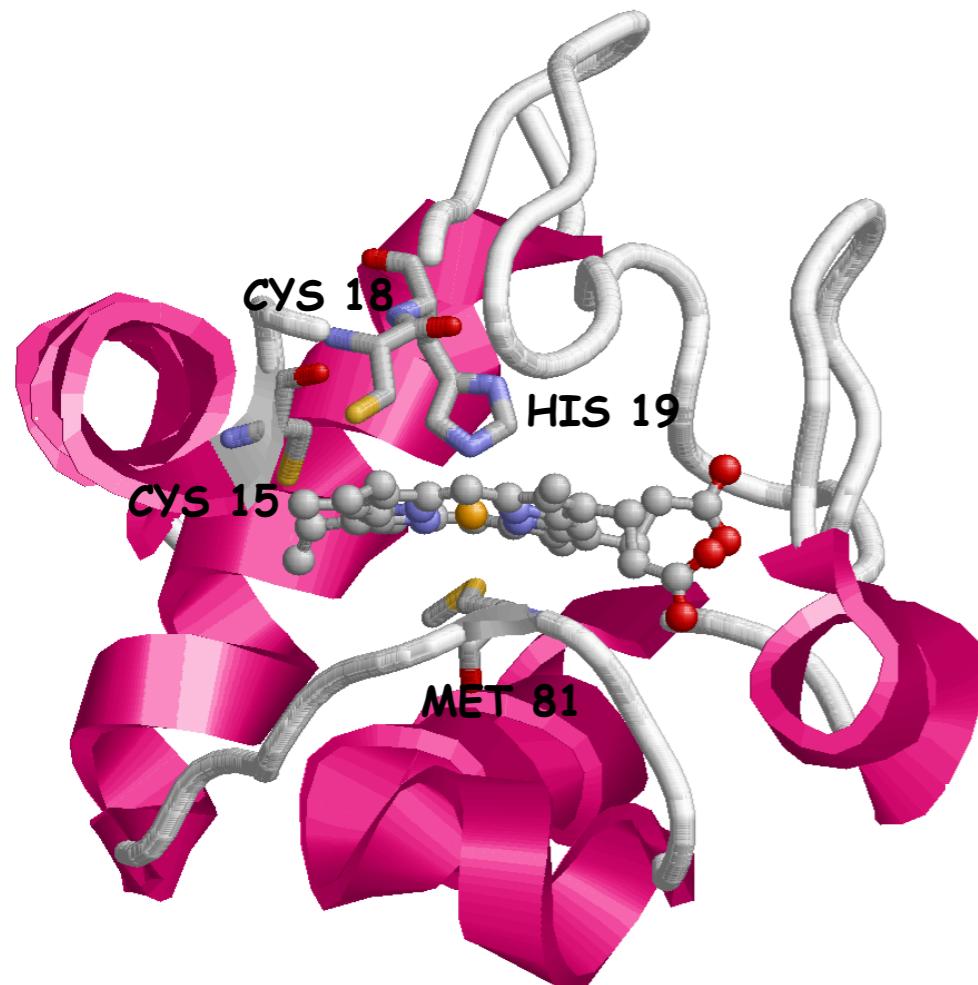
P13045 (GAL3_YEAST) Protein GAL3

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 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)

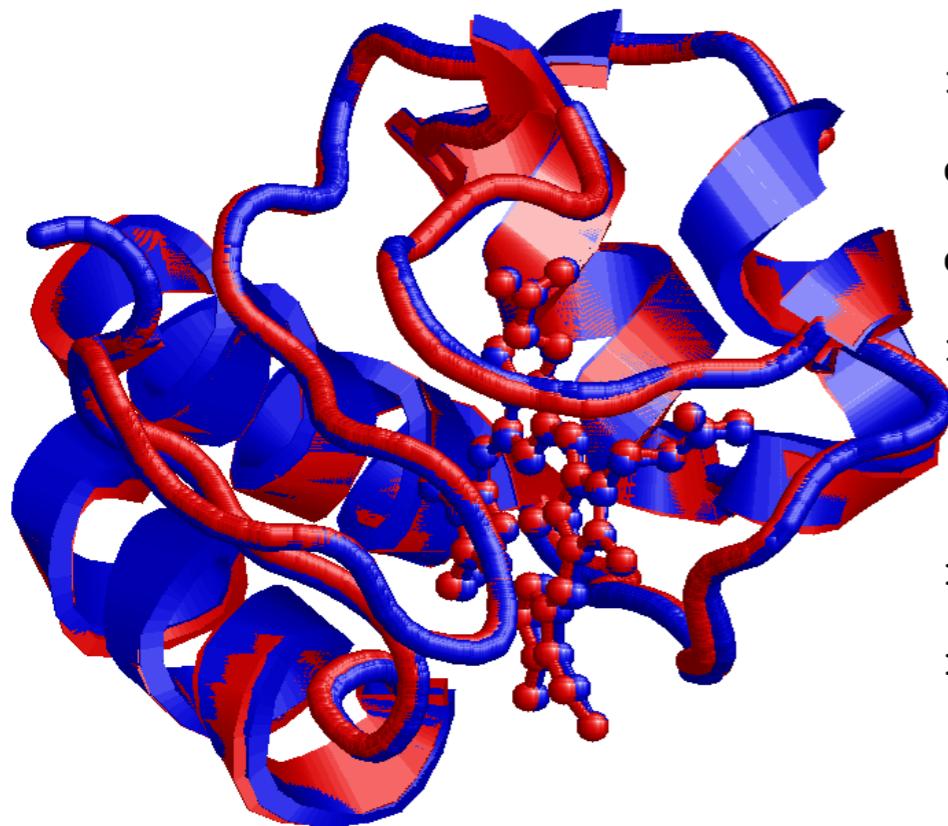
PDB: 3zcf:A

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 CSQCH TVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIIWGEDTLMEYLEN			
: : .	: .	: .	: .
GDVEKGKKIFVQK CAQCH TVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITWKEETLMEYLEN			
	.	.	.
1 : A	20 : A	40 : A	60 : A
80 : A	100 : A		
.	.		
PKKYIPGT KM IFVGIGKKEERADLIAYLKKATNE			
: : .	: .	: .	: .
PKKYIPGT KM IFAGIGKKTEREDLIAYLKKATNE			
.	.		
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)

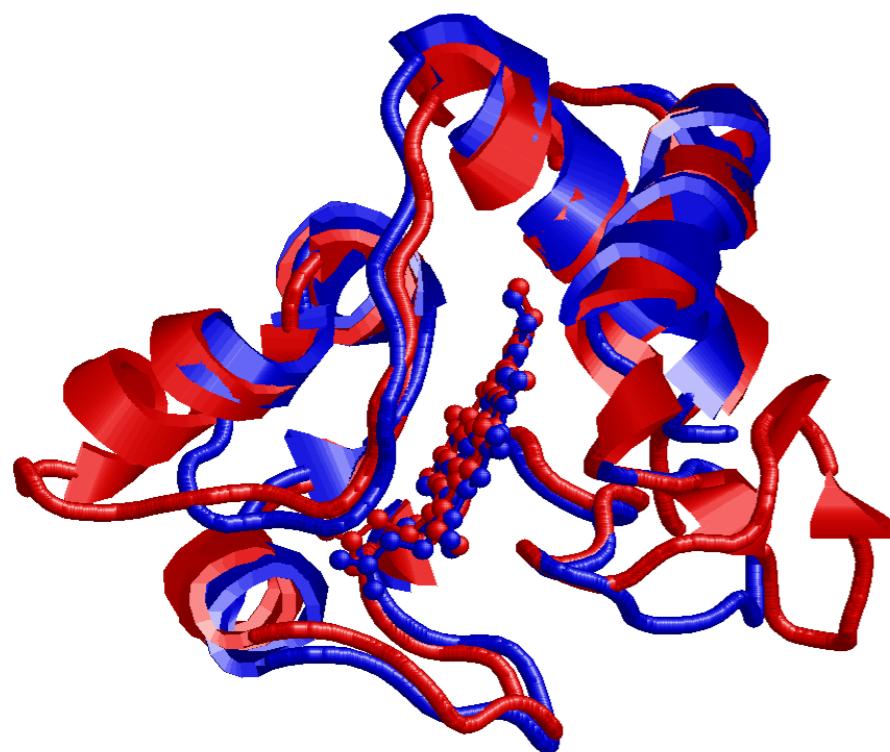
	10	20	30	40	50	60
Homo	MGDVEKGKKIFIMK <u>CSQCH</u> TVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAANKNKGIW					
	:::::::	:::::::	:::::::	:::::::	:::::::	:::::::
Horse	MGDVEKGKKIFVQK <u>CAQCH</u> TVEKGGKHKTGPNLHGLFGRKTGQAPGFTYTDANKNKGITW					
	10	20	30	40	50	60
	70	80	90	100		
Homo	GEDTLMEYLENPKKYIPGTM <u>I</u> FVGIKKKEERADLIAYLKKATNE					
	:::::::	:::::::	:::::::	:::::::	:::::::	:::::::
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 can be used for homology modeling after a refinement of the alignment because one binding site is not conserved.

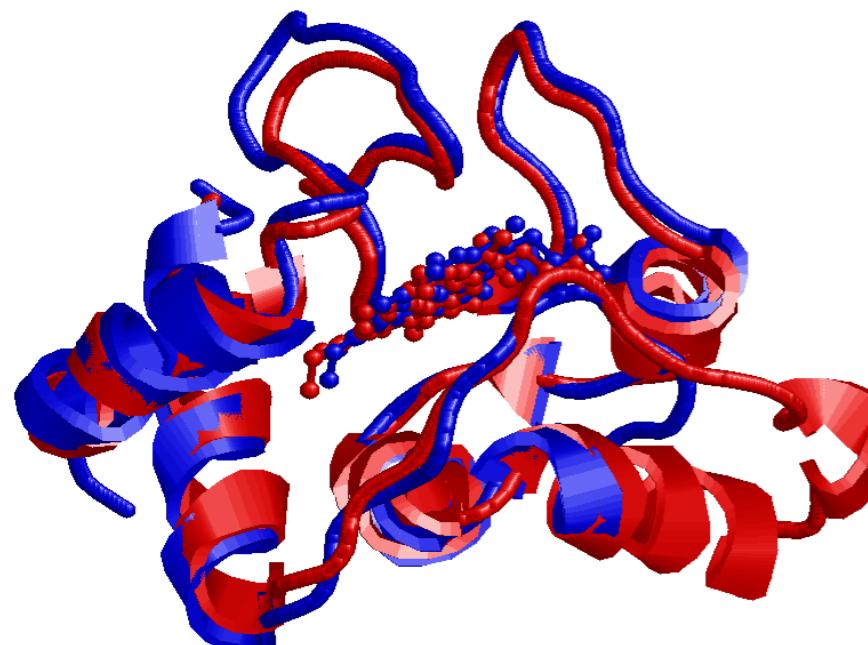
Structural alignment:
RMSD= 0,18 nm
28% sequence identity

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

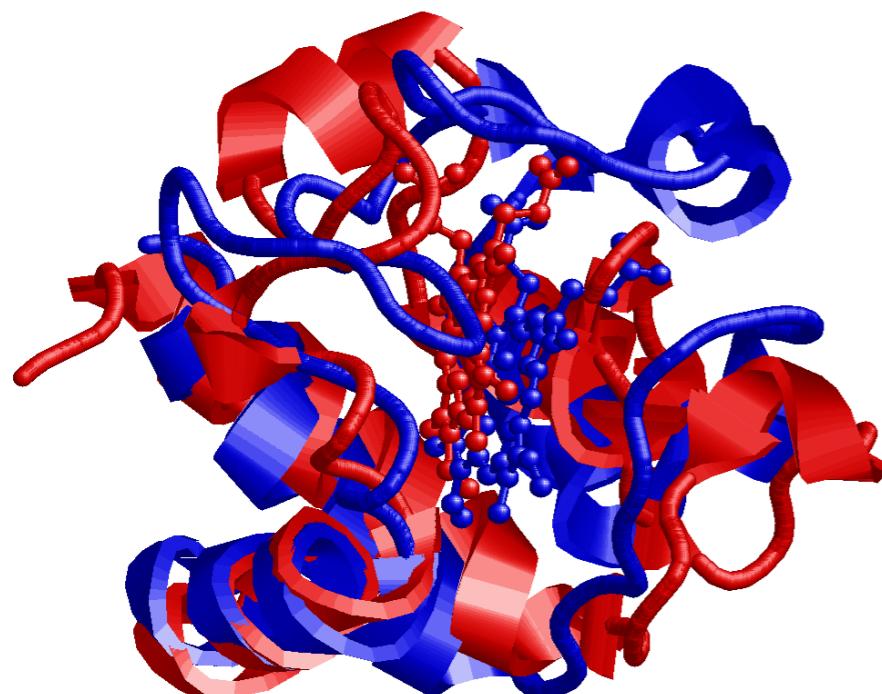
10	20	30		
sp P99	MGDVEKGKKIFIMK <u>CSQCH</u> TVEKGGKHKTGPNLHGL			
	: . . : . . : . . : : . . . : . : . . : . .			
sp P00	MVKLLTILSIAATAGSLSIGTASA <u>QDAKAGEAVF</u> ----KQCMT <u>CHRADKNMVGPA</u> LGGV			
	10 20 30 40 50			
40	50	60 70 80 90		
sp P99	<u>FGRKTGQAPGYSYTAANKNKG</u> --- <u>IIWGEDTLMEYLENPKKYIPGTKM</u> IFVGIKKKEERA			
	: . . : . . : . . : . . : . . : . . : . . : . . : . .			
sp P00	<u>VGRKAGTAAGFTYSPLNHNSGEAGLVWTADNIINYLNDPNAFL</u> --KKFLTDKGKADQAV			
	60 70 80 90 100 110			
100				
sp P99	DLIAYLKKATNE			
	. . : . .			
sp P00	GVTK <u>M</u> TFKLANEQQRKDVVAYLATLK			
	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

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.