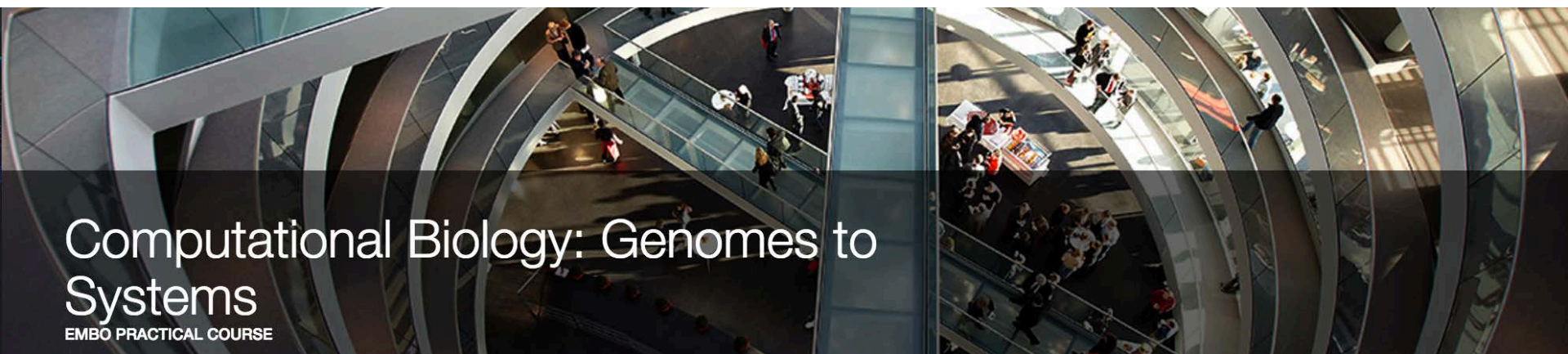


# Using protein structures to understand protein function

Rebecca Wade

Computational Biology: Genomes to Systems

EMBO PRACTICAL COURSE





# Using protein structures to understand protein function

**Rebecca C. Wade**

Molecular & Cellular Modeling Group,  
Heidelberg Institute for Theoretical Studies (HITS)

&

Zentrum für Molekulare Biologie Heidelberg (ZMBH)  
Heidelberg University

[rebecca.wade@h-its.org](mailto:rebecca.wade@h-its.org)

<http://www.h-its.org/mcm/>

# Heidelberg Institute for Theoretical Studies

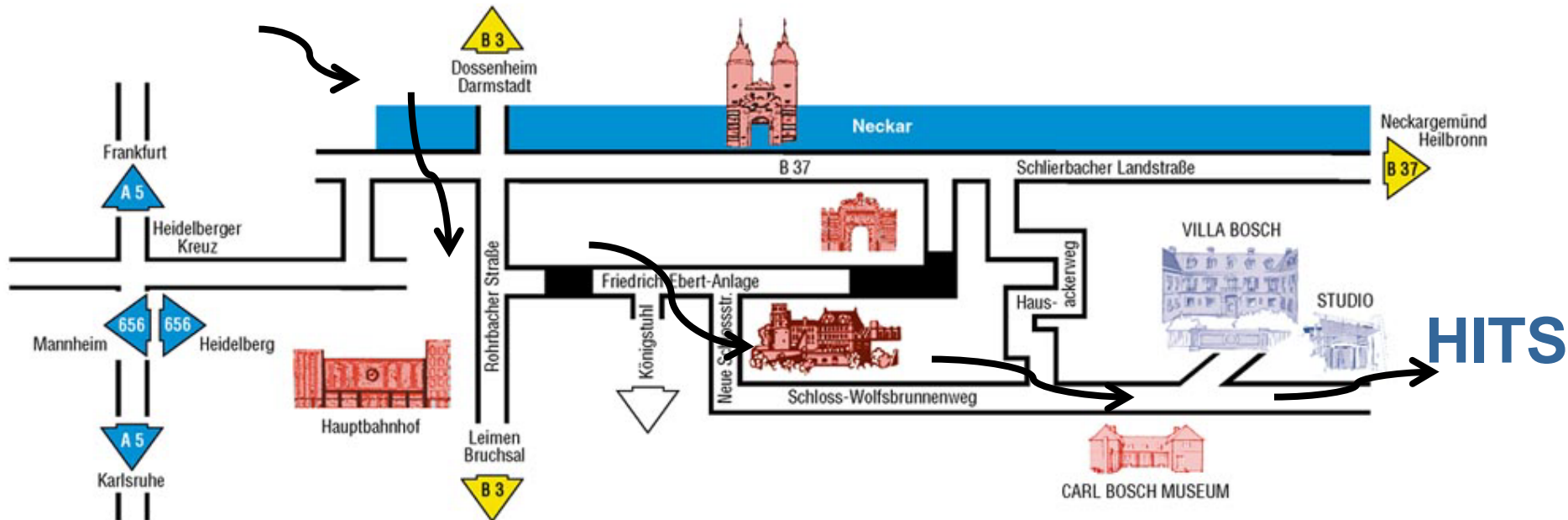
Heidelberg Institute for Theoretical Studies



Multidisciplinary computational sciences  
*Bioinformatics to computational linguistics to astrophysics to .....*



Klaus Tschira Foundation



# Heidelberg Institute for Theoretical Studies



## Multidisciplinary computational sciences:

- Molecular modeling and simulation ; Bioinformatics & databases ; Computational linguistics ; Theoretical astrophysics ; ...



## Klaus Tschira Stiftung:

Supports informatics, natural sciences, mathematics



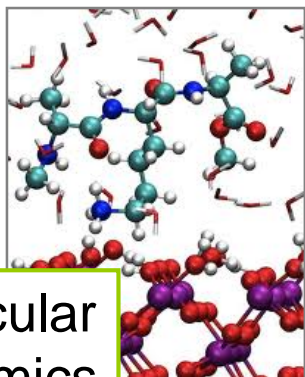
## Klaus Tschira (1940-2015)

1972 - Co-founded the German software giant SAP AG

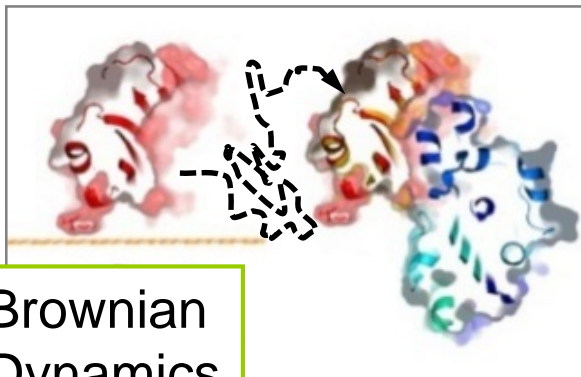
1995 – Bought **Villa Bosch**, founded Klaus Tschira Stiftung



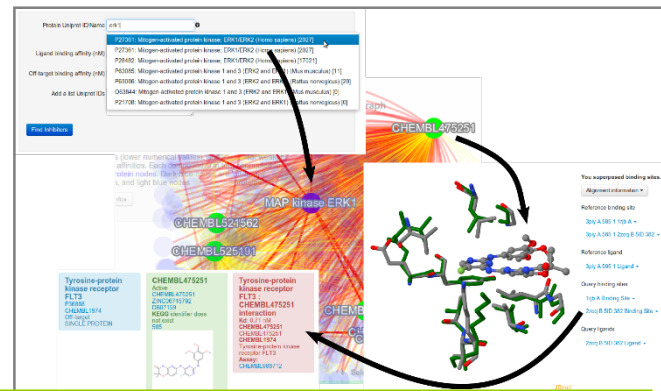
# Computing Protein Interactions: Methods ↔ Applications



Molecular  
Dynamics

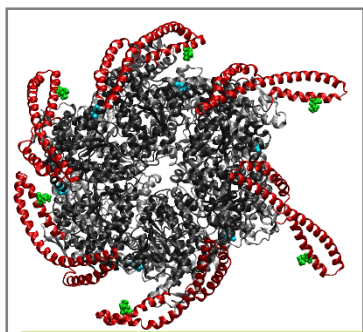


Brownian  
Dynamics

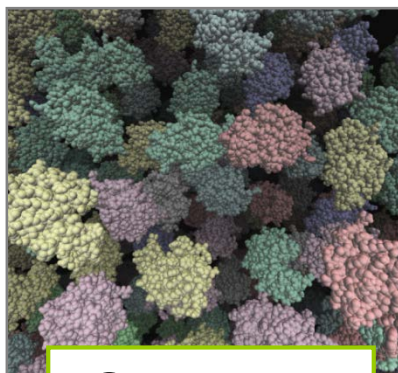


Bio/Chemo-informatics

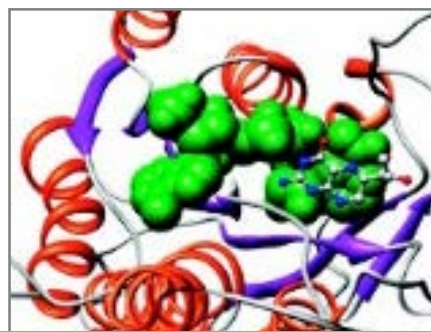
Multiscale Molecular Simulation



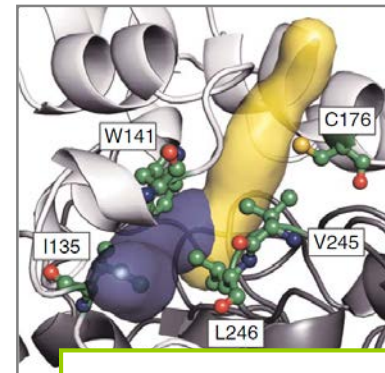
Molecular  
Biology



Systems  
Biology



Structure-based  
Drug Design



Protein  
Engineering

# Using protein structures to learn about protein function: Learning objectives

- Protein structure and function
- Modeling protein structure and dynamics
- Computing interaction properties

# Proteins:

DNA: gatccagctg taccattatg taatataata agacacggac gcac.....



PROTEIN: MFKPVDFSETSPVPPDIDLAPTQSPHHVAPSQDSSYDLLS.....  
.....SMLKNKSFLLHGKDYPNNADNNDNEDIRAKTMNRSQSHV

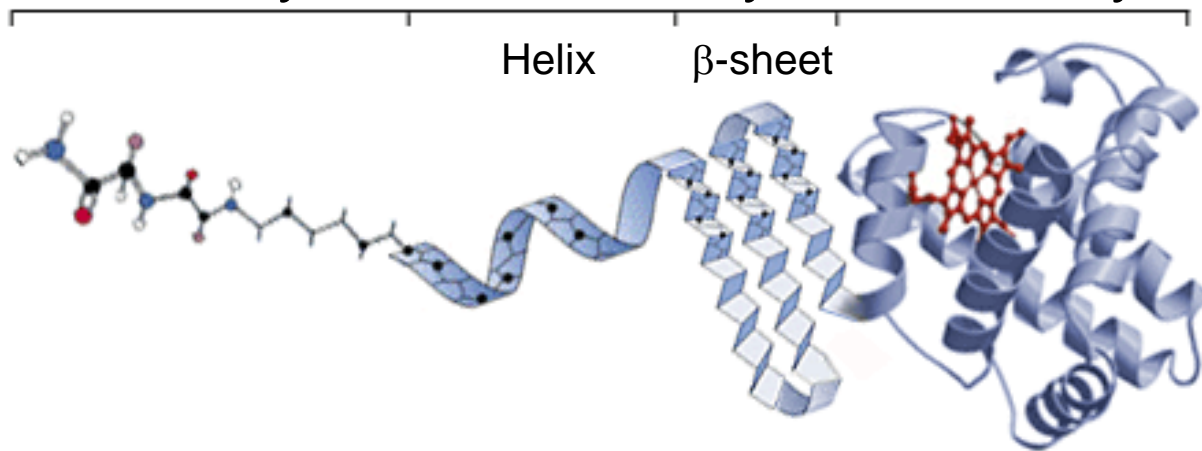


Structure:

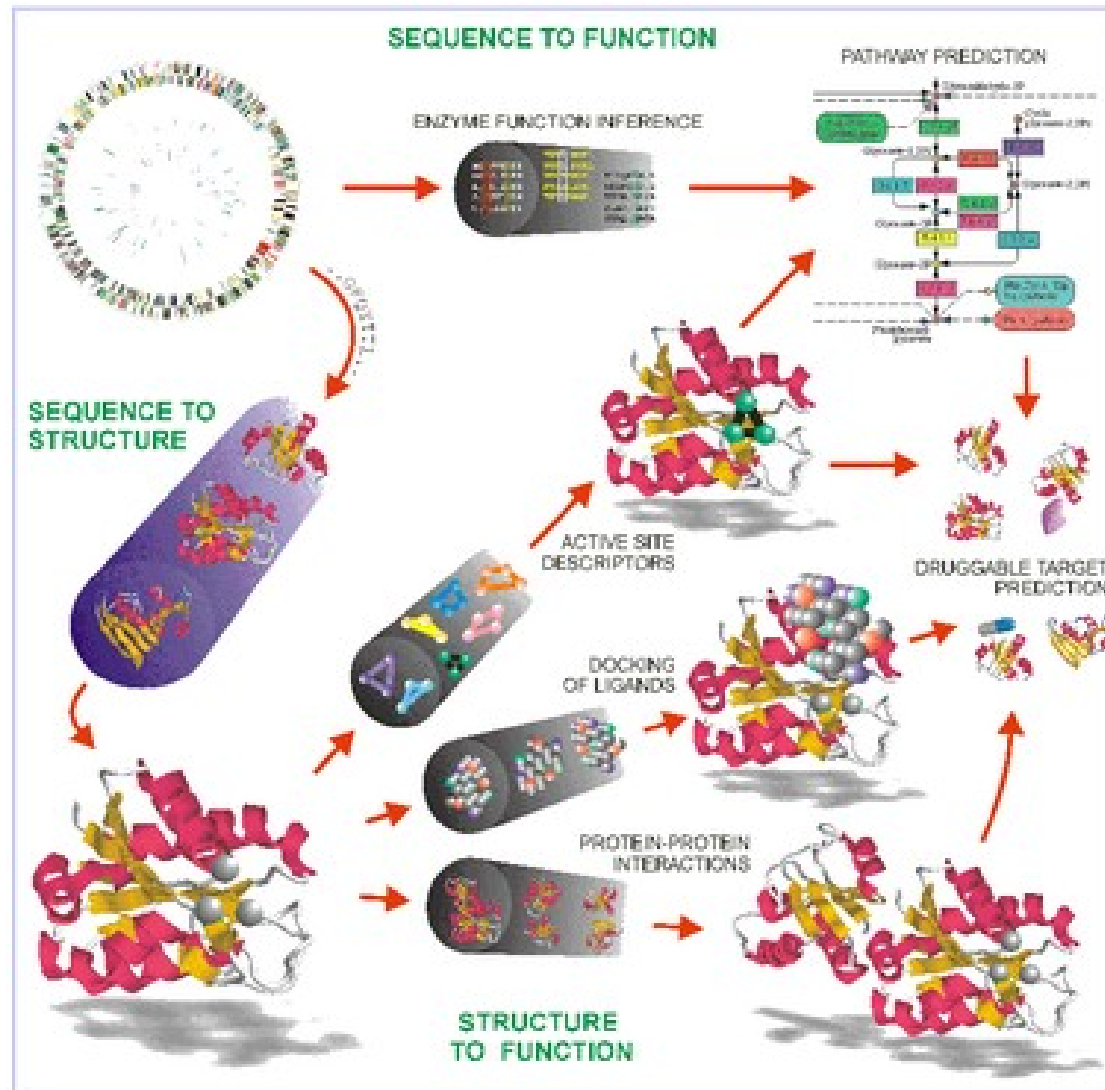
Primary

Secondary

Tertiary



# Sequence → Structure → Function





# The most important question:

- **What do I want to use my protein model(s) (experimental or predicted) for?**

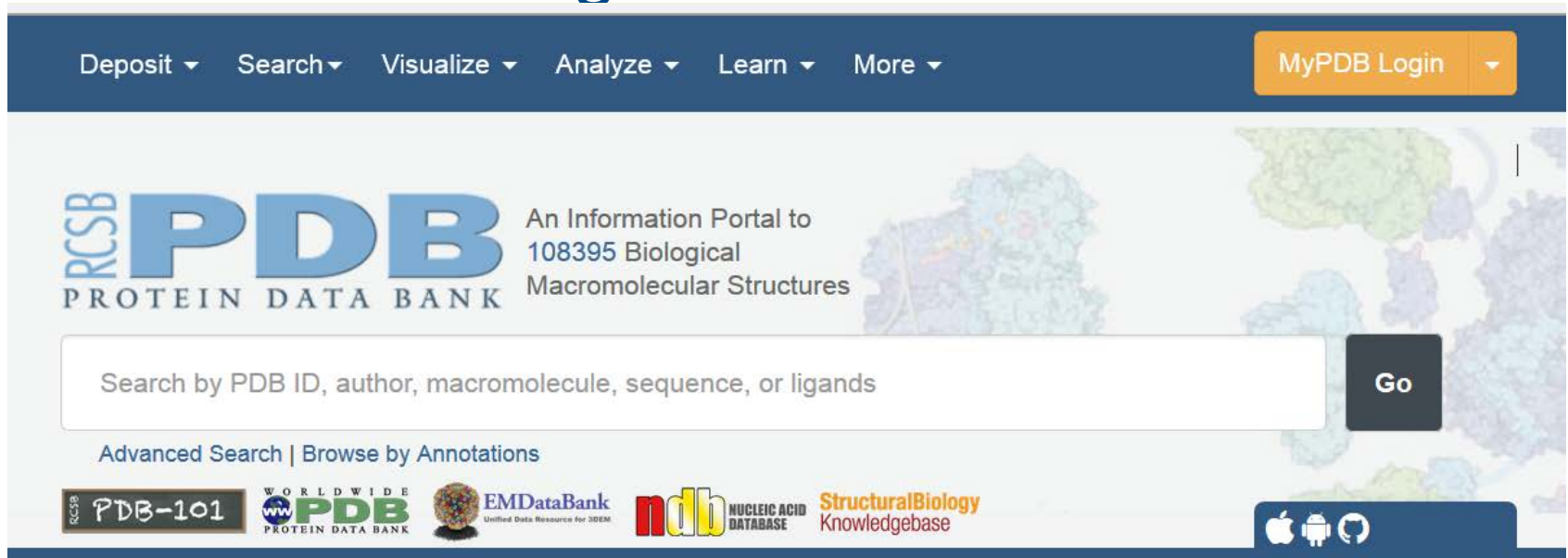
(What effect will errors in the models have?)

**Is the structure of the protein already known from experiments?**

**If so, does this provide me with a suitable 3D model for answering my question?**

# The PDB: Protein Data Bank

- [www.rcsb.org](http://www.rcsb.org)



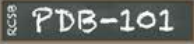





Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Learn ▾ More ▾ MyPDB Login ▾

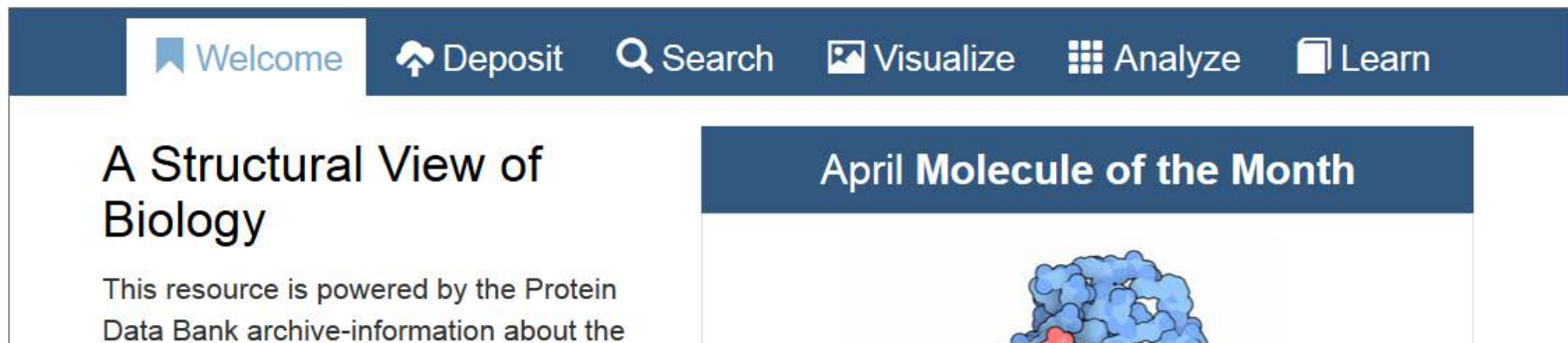
**RCSB PDB**  
PROTEIN DATA BANK

An Information Portal to  
108395 Biological  
Macromolecular Structures

Search by PDB ID, author, macromolecule, sequence, or ligands Go

[Advanced Search](#) | [Browse by Annotations](#)




Welcome Deposit Search Visualize Analyze Learn

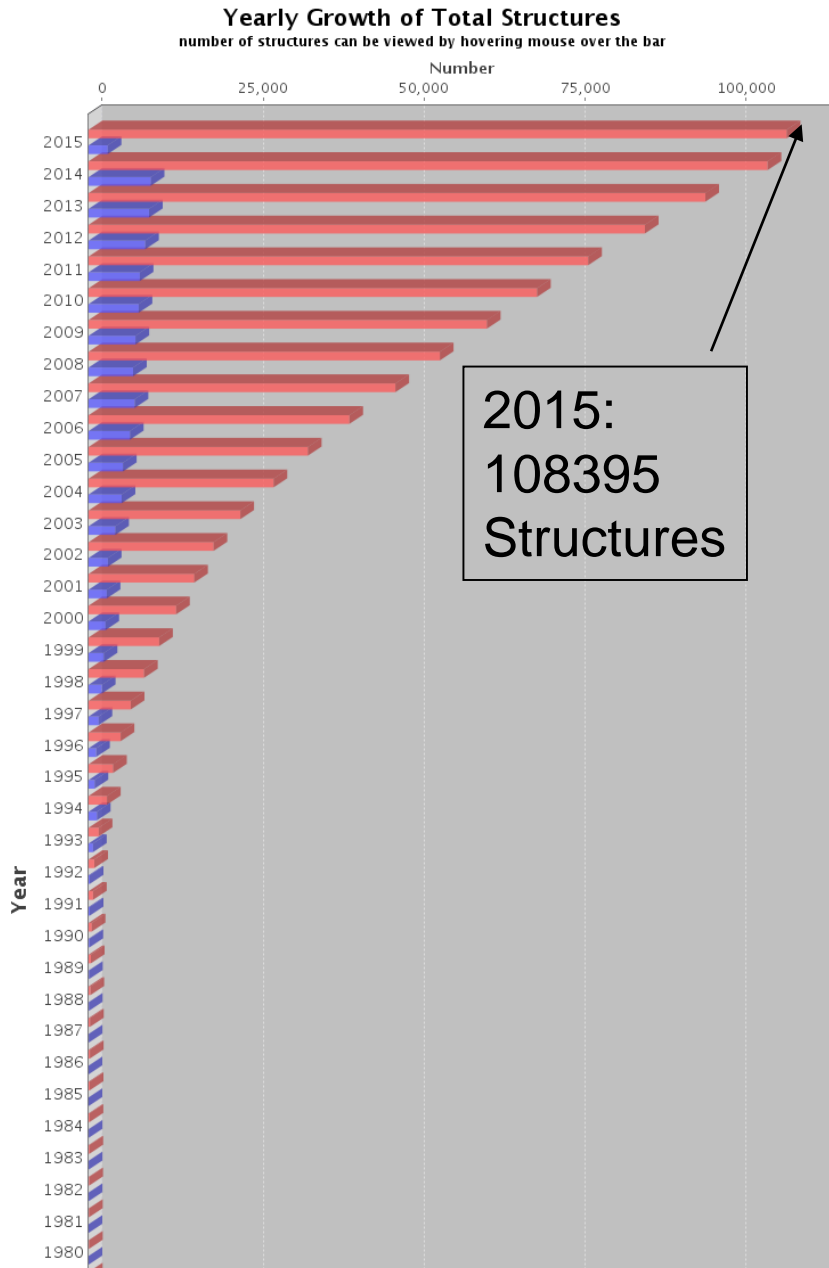
## A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the

## April Molecule of the Month



# Structures in the PDB:

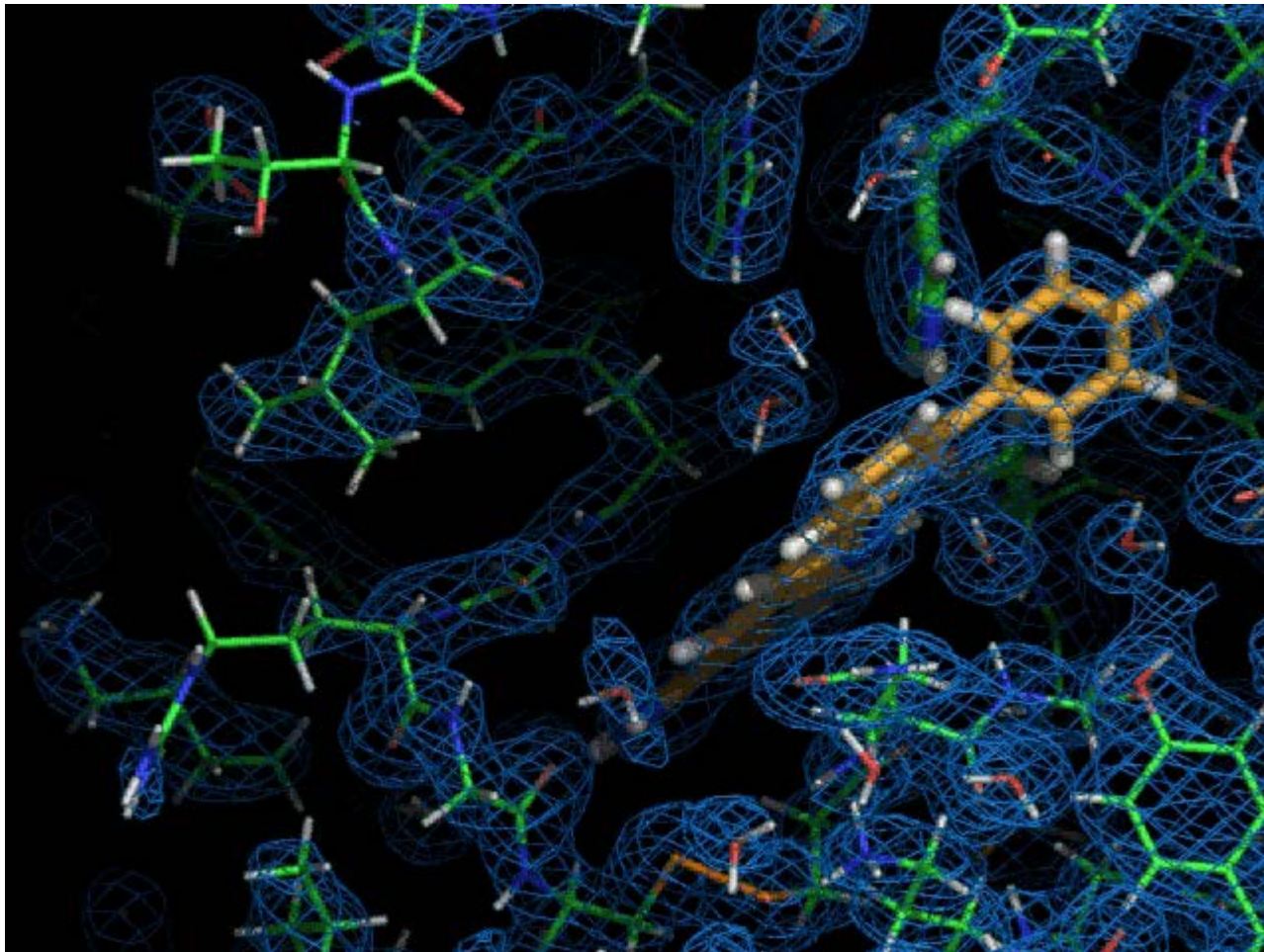


## Structures from:

- Xray crystallography
- Nuclear Magnetic Resonance
- Electron microscopy

# Crystal structures are models:

*Protein-ligand complex with electron density map*

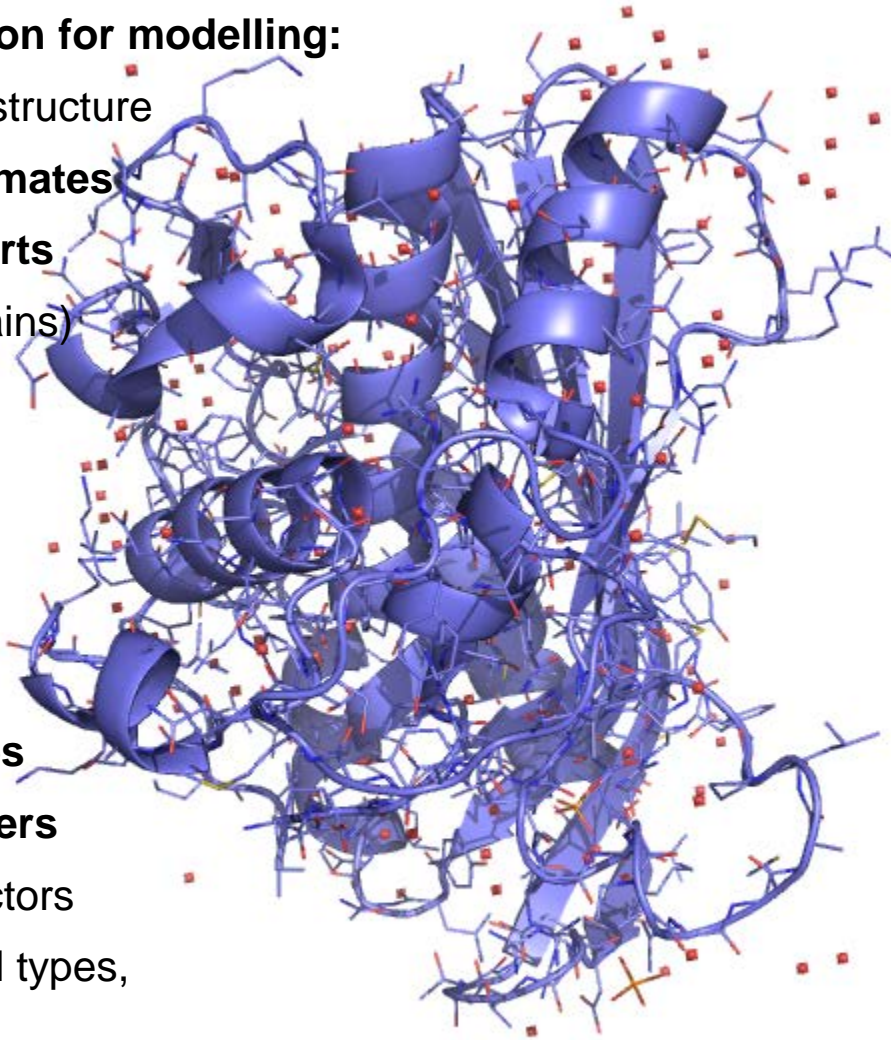


human thrombin at a resolution of 1.68 Å (PDB code: 1o5a)

# Preparation of a structure:

## Structure preparation for modelling:

- downloading PDB structure
- adding **symmetry mates**
- adding **missing parts**  
(e.g., loops, side chains)
- adding **hydrogens**  
(protonation state!)
- re-orientation of  
side chains  
**(His, Asn, Gln)**  
and **water molecules**
- assigning **parameters**  
for ligands and cofactors  
(e.g., atom and bond types,  
angles, charges)
- for MD:  
adding water box

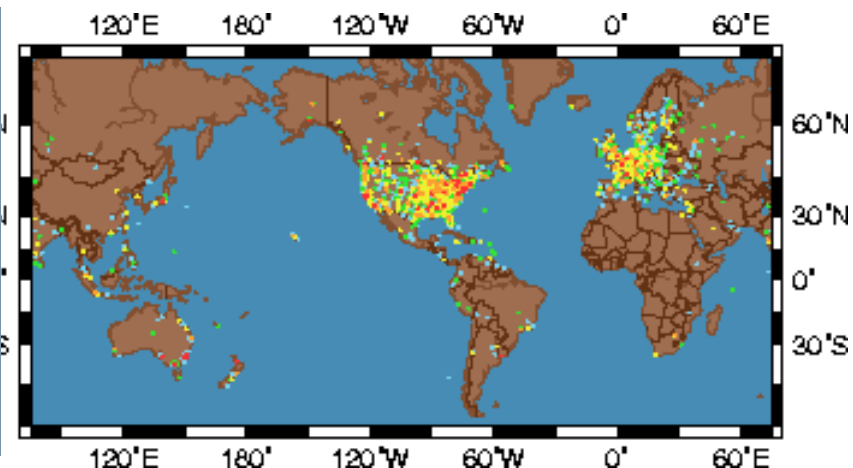


human thymidylate  
synthase (1ypv)  
space group: P3<sub>1</sub>21

**If I only know the protein sequence,  
can I fold the protein and predict the  
structure?**

# Folding from first principles (ab initio)

- Molecular dynamics simulation
  - CPU
  - Distributed computers
- Villin headpiece (36AA)
  - Fastest folding protein (microsec)
  - Duan&Kollmann (Science 1998)
  - +Many others
- Folding@Home
  - V. Pande, [Folding@Home](#)
  - Multiple folding simulations

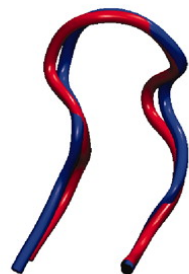




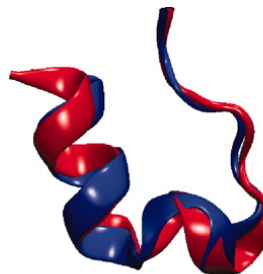
# Folding from first principles (ab initio)

- 12 fast-folding small diverse proteins

- DE Shaw et al, Science 2011
- MD simulations: 100  $\mu$ s - 1 ms



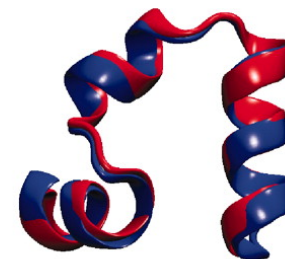
**Chignolin** 106  $\mu$ s  
cln025 1.0 Å 0.6  $\mu$ s



**Trp-cage** 208  $\mu$ s  
2JOF 1.4 Å 14  $\mu$ s



**BBA** 325  $\mu$ s  
1FME 1.6 Å 18  $\mu$ s



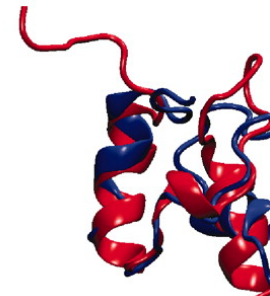
**Villin** 125  $\mu$ s  
2F4K 1.3 Å 2.8  $\mu$ s



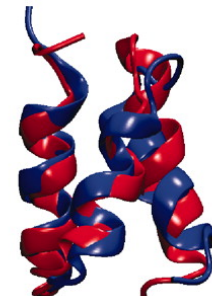
**WW domain** 1137  $\mu$ s  
2F21 1.2 Å 21  $\mu$ s



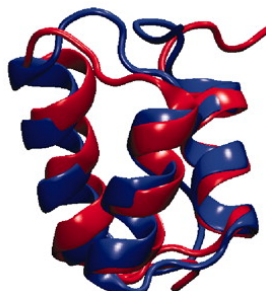
**NTL9** 2936  $\mu$ s  
2HBA 0.5 Å 29  $\mu$ s



**BBL** 429  $\mu$ s  
2WXC 4.8 Å 29  $\mu$ s



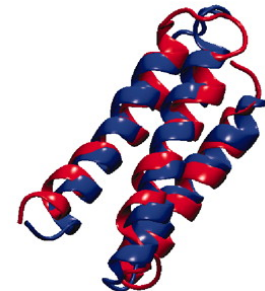
**Protein B** 104  $\mu$ s  
1PRB 3.3 Å 3.9  $\mu$ s



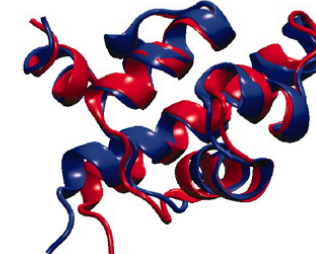
**Homeodomain** 327  $\mu$ s  
2P6J 3.6 Å 3.1  $\mu$ s



**Protein G** 1154  $\mu$ s  
1MIO 1.2 Å 65  $\mu$ s



**$\alpha$ 3D** 707  $\mu$ s  
2A3D 3.1 Å 27  $\mu$ s

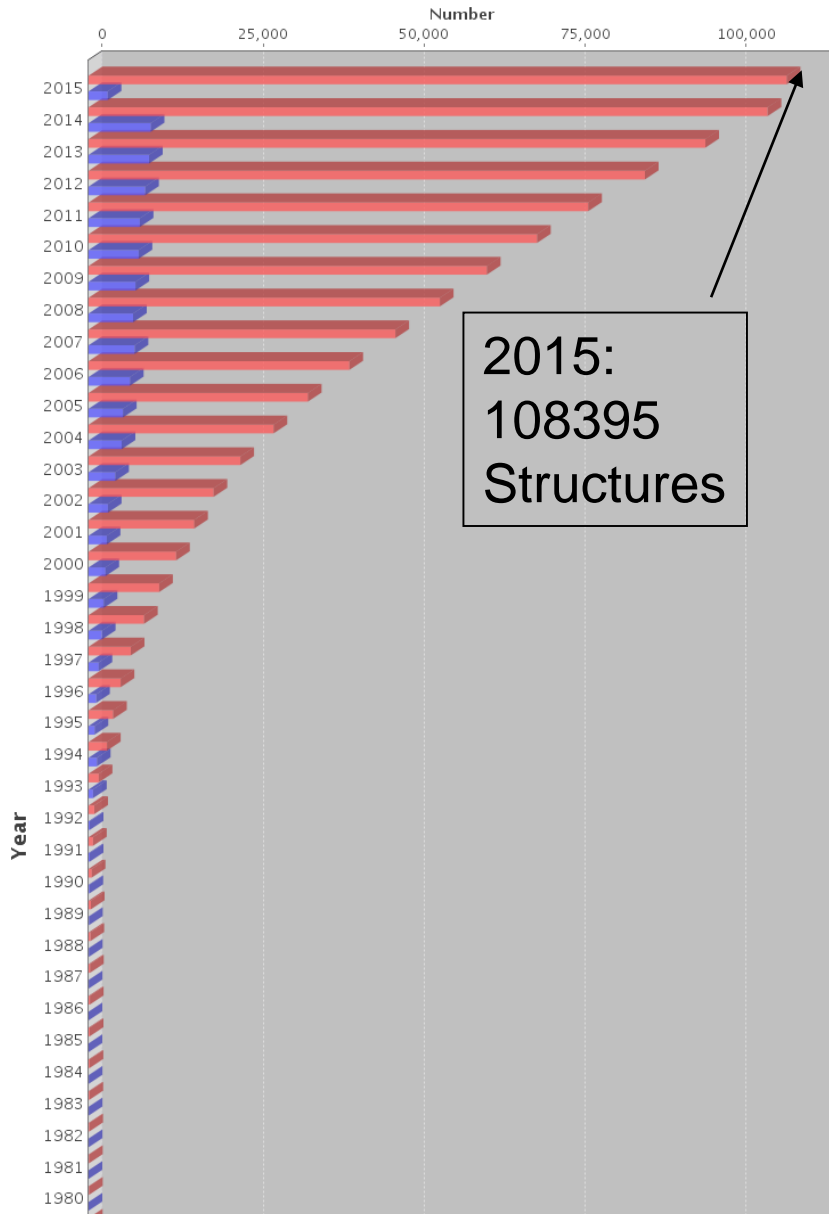


**$\lambda$ -repressor** 643  $\mu$ s  
1LMB 1.8 Å 49  $\mu$ s

# Protein 3D Structures in the PDB:

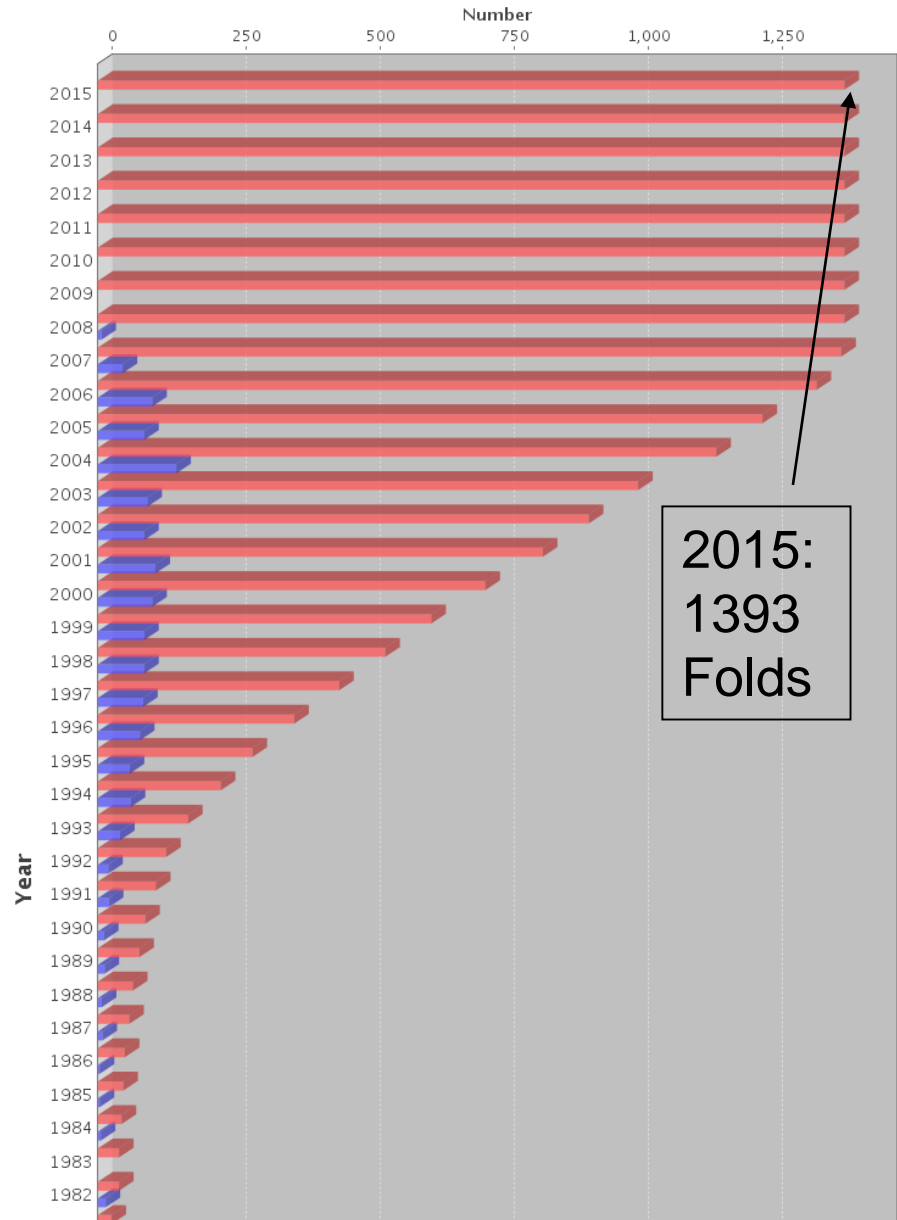
## Yearly Growth of Total Structures

number of structures can be viewed by hovering mouse over the bar

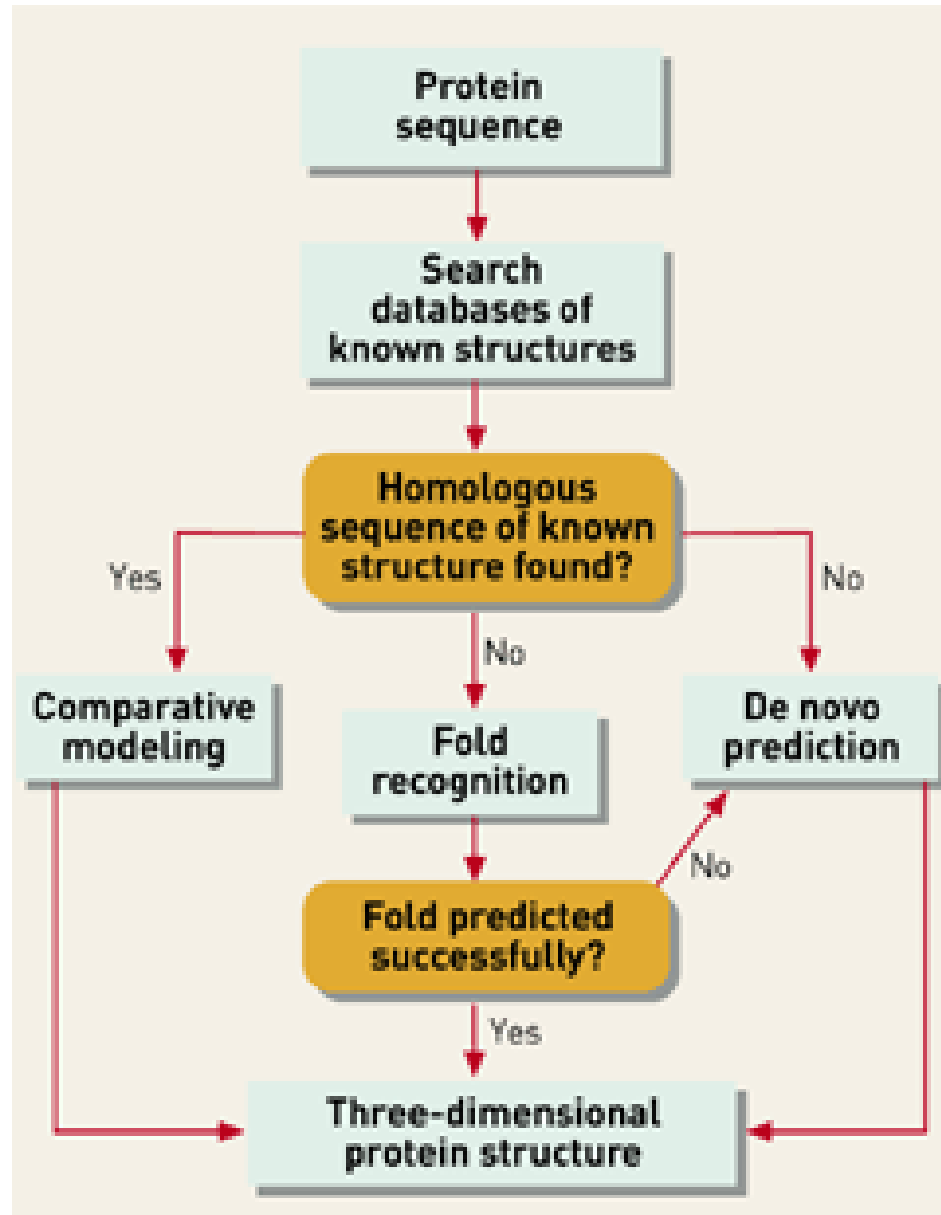


## Growth Of Unique Folds Per Year As Defined By SCOP (v1.75)

number of folds can be viewed by hovering mouse over the bar



# Can I make a 3D model of this protein?



From:  
C&EN, 04.08.2003  
(R. Russell)

# Can I find a 3D model of this protein?

- **Is my protein in a structural database?**
  - Experimentally determined structures:
    - PDB (RCSB)
    - <http://www.rcsb.org>
  - Comparative model databases, e.g.:
    - ModBase,
    - SwissModel Repository
  - Single point of entry for finding protein structures:
    - [Protein Structure Initiative Knowledgebase \(PSI KB\)](http://www.proteinmodelportal.org/)
    - <http://www.proteinmodelportal.org/>

# The Protein Model Portal (PMP)

– <http://www.proteinmodelportal.org/>

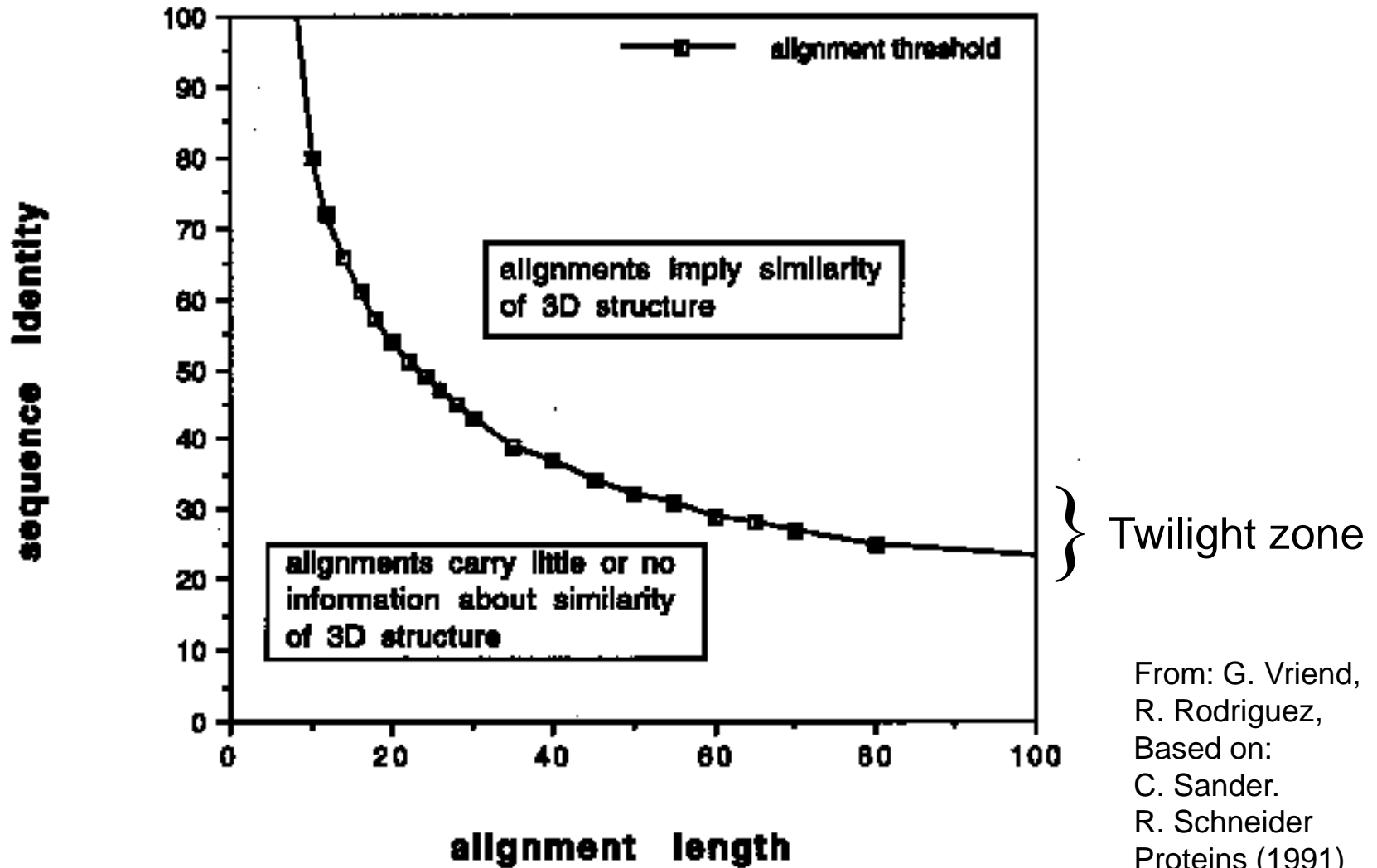
The screenshot shows a web browser window displaying the Protein Model Portal (PMP) homepage. The browser's address bar shows the URL [www.proteinmodelportal.org](http://www.proteinmodelportal.org/). The page features a navigation menu with options: Home, Interactive Modeling, Quality Estimation, Protein Modeling 101, and More. A search bar is present with the placeholder text "Please enter your query." Below the navigation, a welcome message reads "Welcome to the Protein Model Portal (PMP)". A descriptive paragraph states: "PMP gives access to various models computed by comparative modeling methods provided by different partner sites, and provides access to various interactive services for model building, and quality assessment." A search input field with the placeholder "Please enter your query." is located below the text. A search button is positioned to the left of the input field. To the right of the search button, examples of search terms are listed: [UniProt AC], [UniProt ID], [RefSeq], [PDBID], [Sequence], and [Free Text]. On the right side of the page, a section titled "Modeling Highlights" (with a "Show all" link) displays three protein structure models. The top row is labeled "FRAGFOLD with contacts" and the bottom row is labeled "PDB". The models are shown in red and green, with numerical values 0.58, 0.68, and 0.66 placed below each model. Below the models, the text reads: "De novo structure prediction of globular proteins aided by sequence variation-derived contacts." The browser's taskbar at the bottom shows various application icons and system tray icons, including the Windows logo, taskbar, and system tray icons for network, volume, and power. The system tray also displays the date and time: "Kosciolek T., Jones DT. PLoS One. (2014) 9(3):97" and "DEU".

# Can I make a 3D model of this protein?

- **Is it similar in sequence to proteins with known structure?**
  - Sequence alignment
    - Blast, FASTA, Modeller
    - Multiple-sequence alignment (PSI-BLAST)
    - Structure-based

# Can I make a 3D model of this protein?

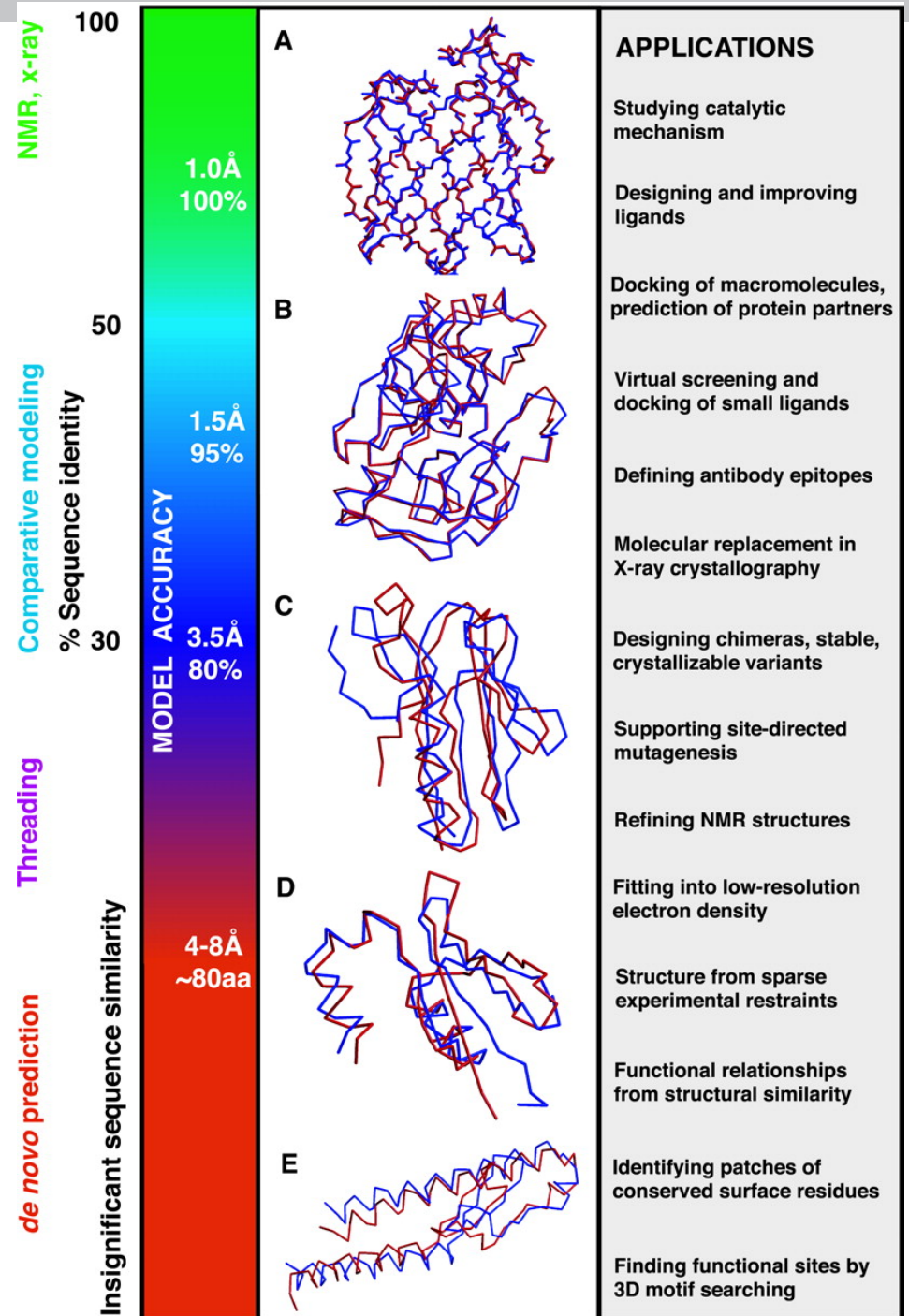
## Threshold for structural homology



From: G. Vriend,  
R. Rodriguez,  
Based on:  
C. Sander.  
R. Schneider  
Proteins (1991)  
9, 56-68.

# Can I make a 3D model of this protein?

Sali, Baker,  
Science 2001





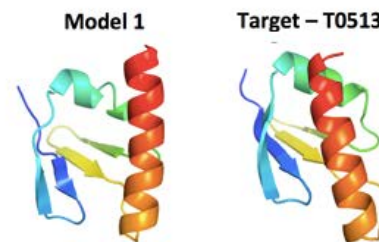
# Can I make a 3D model of this protein?

- What if my protein is not similar in sequence to any proteins with known structure?

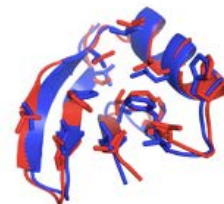
- Fold Recognition
  - Threading of sequence on structures in database
- De Novo Prediction
  - Knowledge-based
  - Simulation
  - *For small proteins*



www.bakerlab.org



2.66 Å over 62 residues



0.84 Å over 39 residues

de novo prediction by Robetta in CASP-8

## REGISTRATION

[ [Register](#) / [Update](#) ] [ [Login](#) ]

## DOCUMENTATION

[ [Docs](#) / [FAQs](#) ]

## SERVICES

Domain Parsing & 3-D Modeling  
[ [Queue](#) ] [ [Submit](#) ]

Interface Alanine Scanning  
[ [Queue](#) ] [ [Submit](#) ]

Fragment Libraries  
[ [Queue](#) ] [ [Submit](#) ]

DNA Interface Residue Scanning  
[ [Queue](#) ] [ [Submit](#) ]

## RELATED SITES

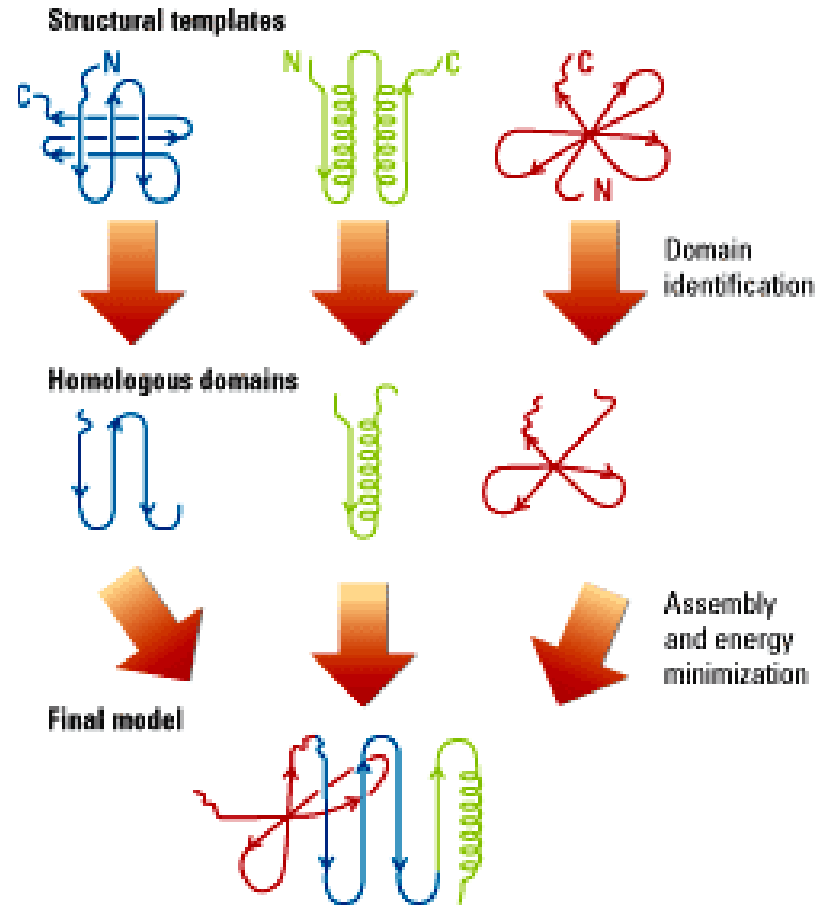
[Rosetta Commons](#)  
[Rosetta Commons ROSIE server \\*NEW\\*](#)  
[RosettaBackrub Server](#)  
[RosettaDesign Server](#)  
[FoldIt](#)  
[Rosetta@home](#)  
[Human Proteome Folding Project](#)  
[Rosetta@Cloud](#)

Robetta.bakerlab.org

# Can I make a 3D model of this protein?

- What if my protein is not similar in sequence to any proteins with known structure?

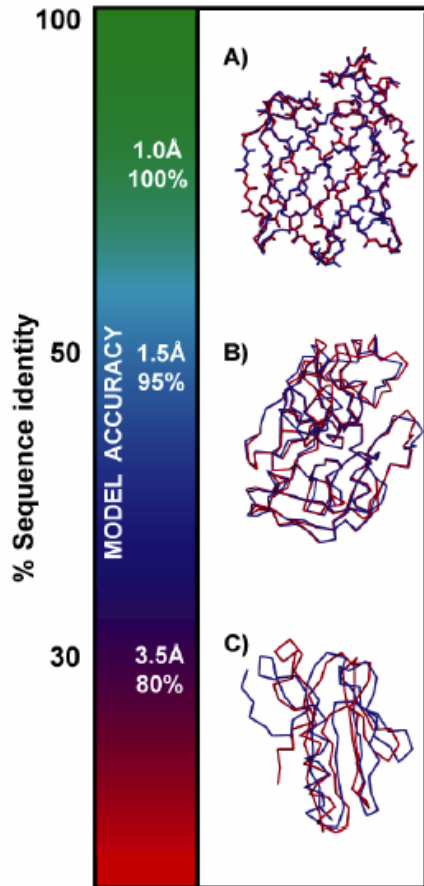
- Fold Recognition
  - Threading of sequence on structures in database
- De Novo Prediction
  - Knowledge-based
  - Simulation
  - *For small proteins*



Rosetta:

Simons, K. T., et al. *Proteins Suppl.* **1999**, 3, 171–176

# Can I make a 3D model of this protein?



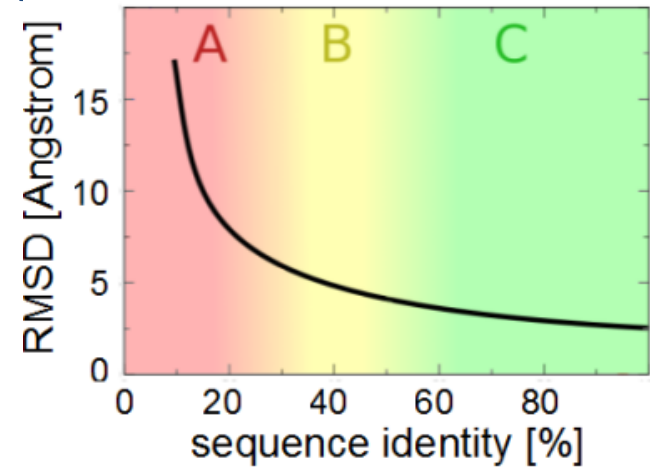
Marti-Renom,  
Yerkovitch,  
Sali *Current Prot.*  
*Protein Sci.*(2002)  
2.9.1

◆ How much sequence identity for comparative (homology) modeling?

- ☞ >50% : “reliable” model
- ☞ >30% : “useful” model
- ☞ <30% : “might be useful” model

◆ Other metrics for deciding on comparative modeling

- ◆ Multiple sequence alignment
- ◆ Structure-based alignment



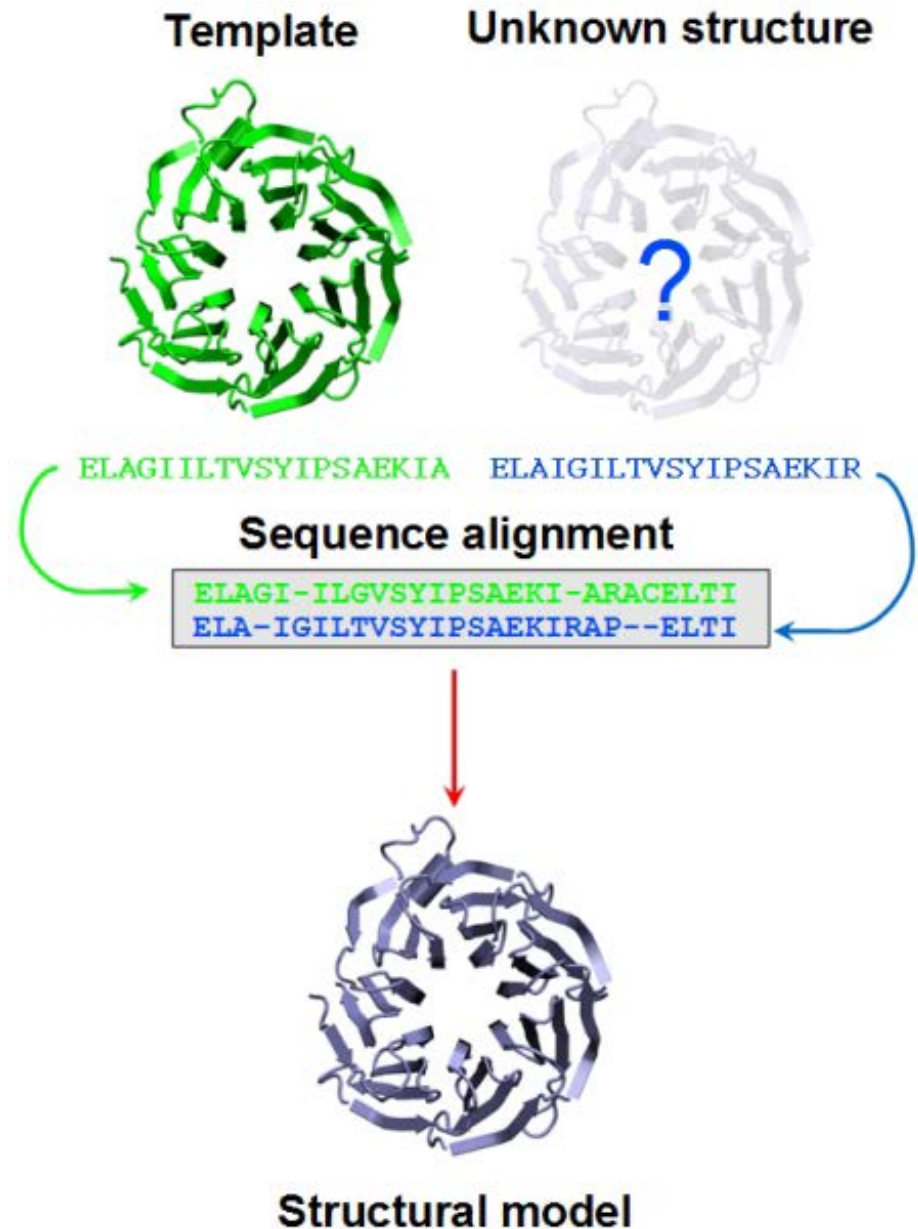
SWISS-MODEL  
webserver

# How can I make a 3D model of this protein?

- **Use a server!**

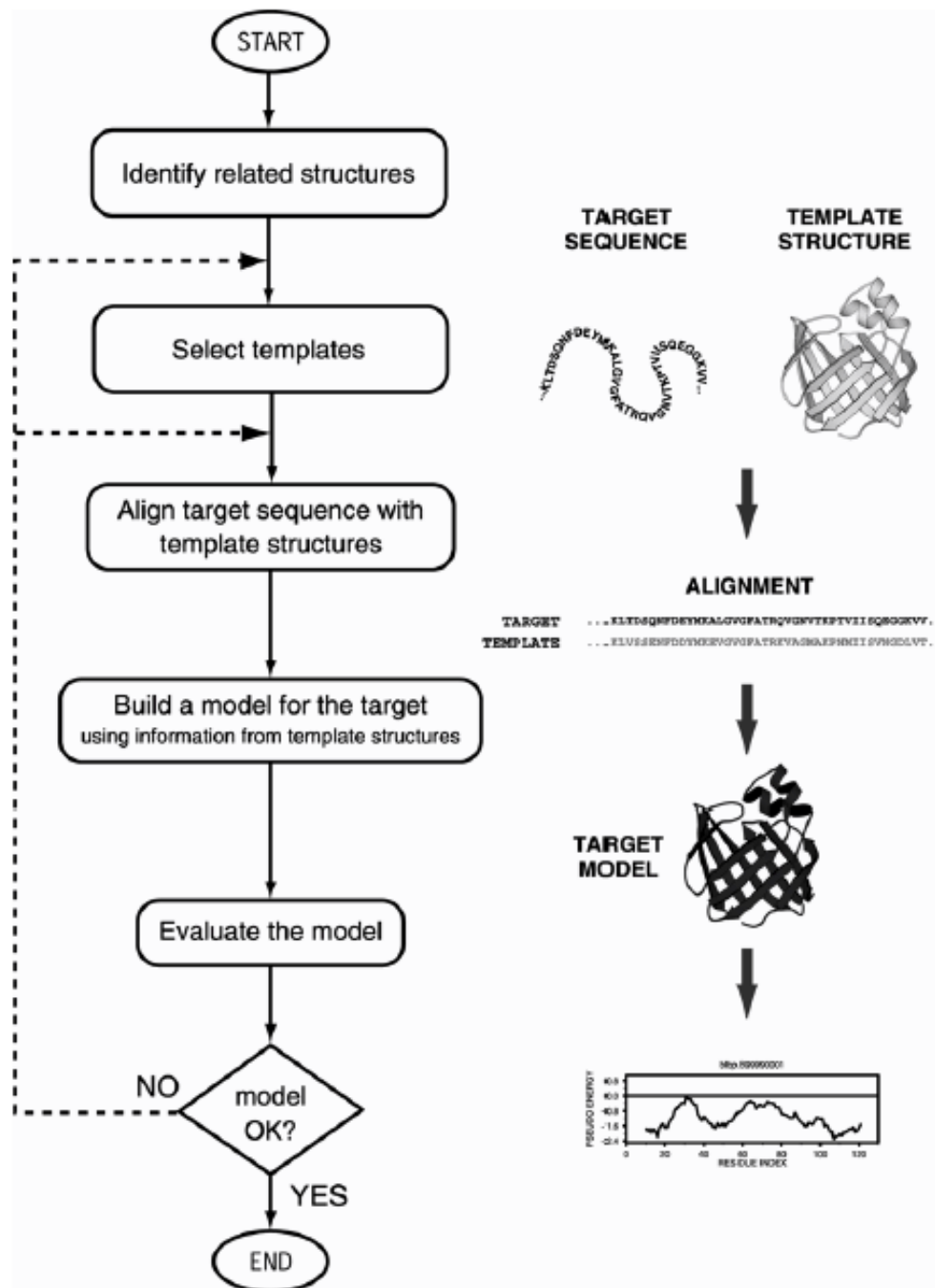
- [http://www.proteinmodelportal.org/?pid=modelling\\_interactive](http://www.proteinmodelportal.org/?pid=modelling_interactive)
- Enter your aminoacid sequence
- Run these tools (comparative modeling and/or threading)
  - ModWeb
  - M4T
  - SwissModel
  - I-TASSER
  - Hhpred
  - Phyre2
  - InfFOLD2
  - RaptorX

# How can I make a 3D comparative model of my protein?



<http://www.unil.ch/pmf/en/home/menueinst/technologies/homology-modeling.html>

# How can I make a 3D comparative model of my protein?

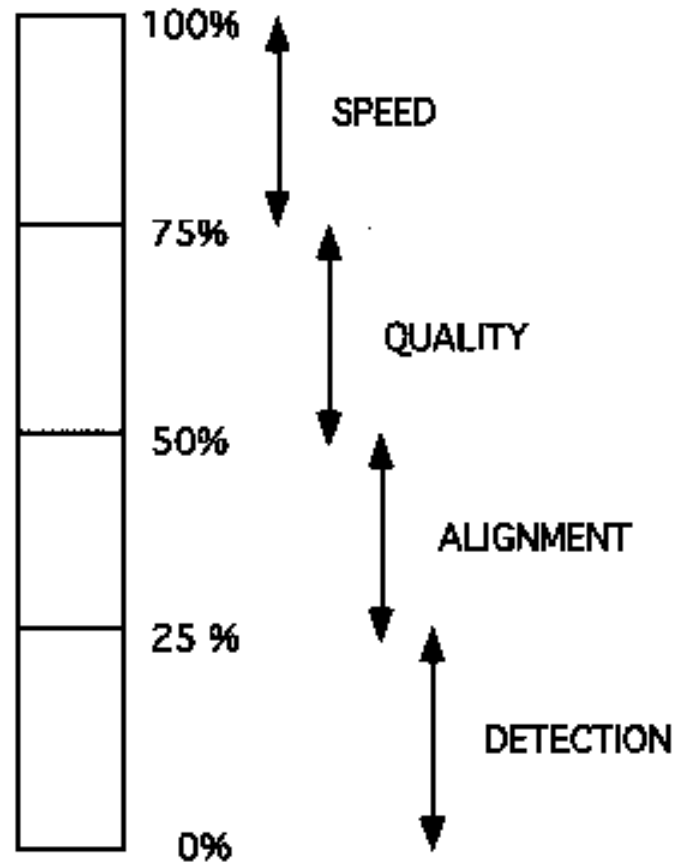


Marti-Renom,  
Yerkovitch,  
Sali Current Prot.  
Protein Sci.(2002)  
2.9.1

# Can I make a 3D comparative model of my protein?

range of sequence similarity in % identical residues

key limiting factor in model building by homology



From: G. Vriend

Figure 1. The main limiting steps for model building by homology as function of the percentage sequence identity between the structure and the model.

# Local sequence predictions

VIEWS

**Dashboard** >

DETAILED PREDICTIONS

Secondary Structure >

Transmembrane Regions >

Protein Disorder and Flexibility >

Disulphide and Metal Binding >

Binding Sites >

Subcellular Localization >

Transmembrane Beta-barrels >

FURTHER ANALYSIS

Functional Changes >

Literature Search >

HELP

Site Tutorial >

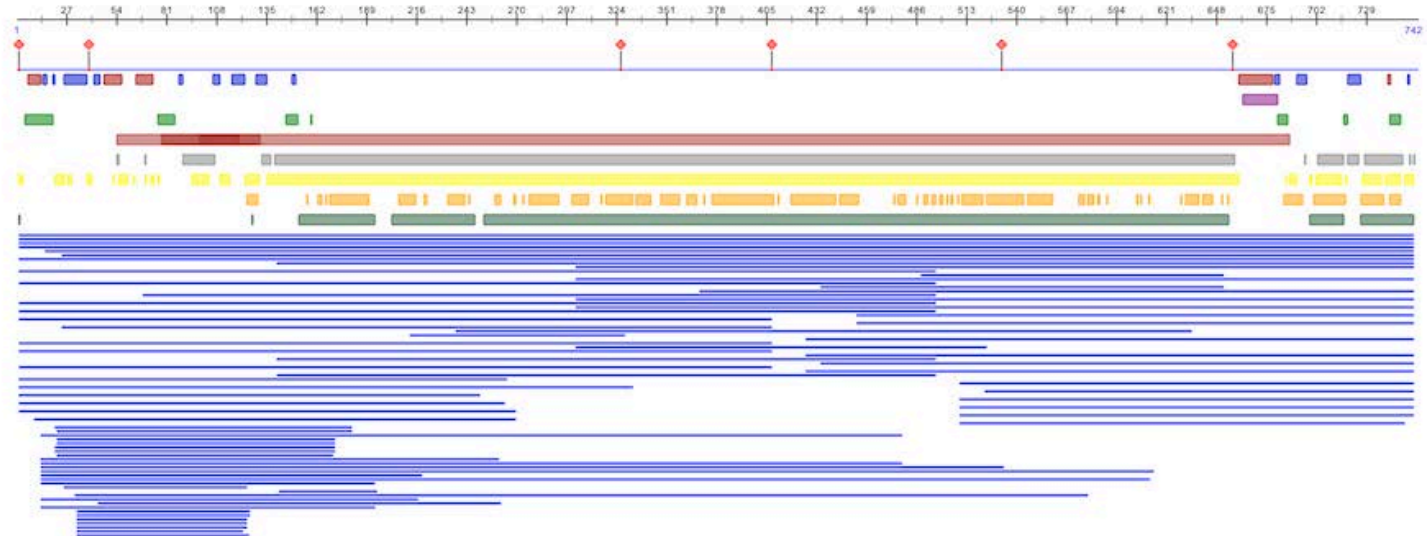
## Dashboard Overview for CD44\_HUMAN

Email Export

What am I seeing Here? This viewer lays out predicted features that correspond to regions within the queried sequence. Mouse over the different colored boxes to learn more about the annotations

Zoom - Start:1, End:742

Export to image



Summary

Amino Acid composition



# Local sequence predictions (with global information)

sequence information from protein family

profile derived from multiple alignment for a window of adjacent residues

two levels of neural network systems: PHDsec and PHDhtm

one level network: PHDacc

local alignment 13 adjacent residues

global statistics of whole protein

input local in sequence

A	C	L	I	G	S	V	ins	del	cons
100	0	0	0	0	0	0	0	0	1.17
100	0	0	0	0	0	0	33	0	0.42
0	0	100	0	0	0	0	0	33	0.92
0	0	33	66	0	0	0	0	0	0.74
66	0	0	0	33	0	0	0	0	1.17
0	66	0	0	0	33	0	0	0	0.74
0	0	0	33	0	0	66	0	0	0.48

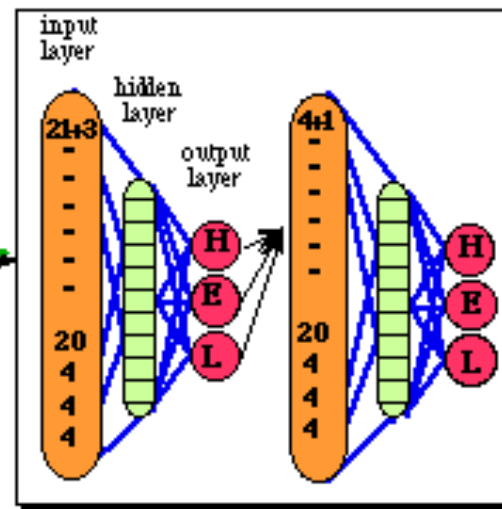
input global in sequence

percentage of each amino acid in protein

length of protein ( $\leq 60, \leq 120, \leq 240, > 240$ )

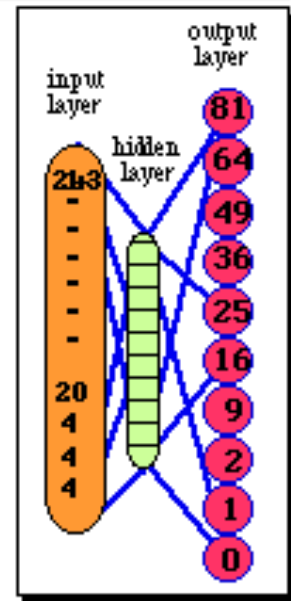
distance: centre, H-term ( $\leq 40, \leq 30, \leq 20, \leq 10$ )

distance: centre, C-term ( $\leq 40, \leq 30, \leq 20, \leq 10$ )



first level  
sequence-to-structure  
network

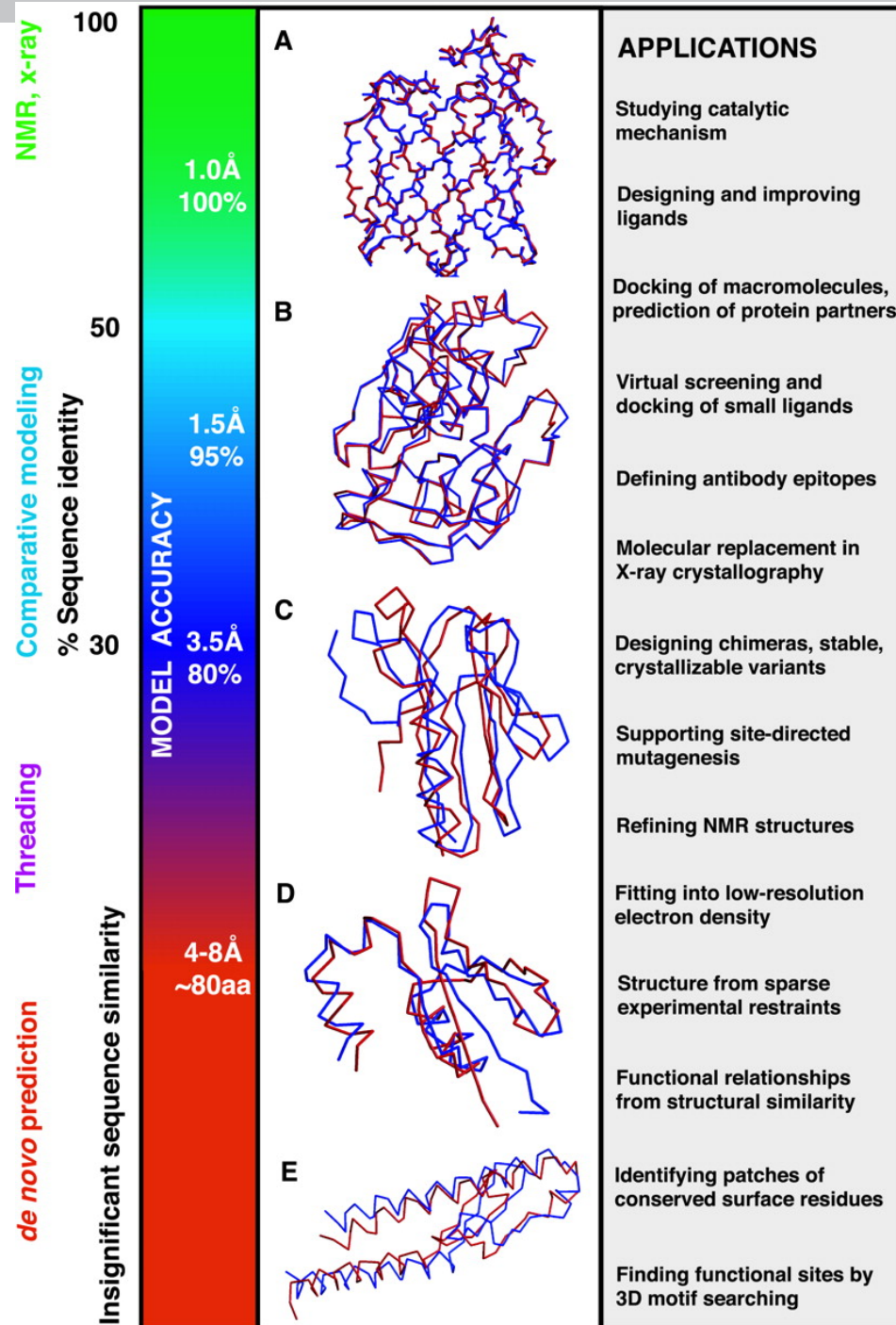
second level  
structure-to-structure  
network



first level only

# What can I do with my protein model?

Sali, Baker,  
Science 2001



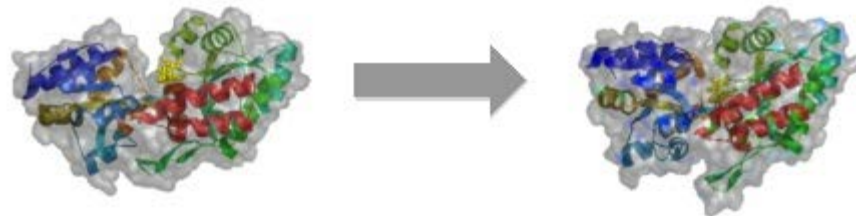
# Challenges for Protein Modeling:

- Sequence alignment
- Loop modeling
- Modeling complexes and oligomers
- Distinguishing the best model(s)
- **!! Proteins are dynamic !!**
- Predicting bound structure from unbound structure
- Predicting correct oligomeric state and structure
- Protein folding

# Protein dynamics:

- Morphing:
  - ◆ Between two known structures of one protein
  - ◆ Morph (Gerstein)
    - ☞ <http://www.molmovdb.org/molmovdb/morph/>
    - ☞ Interpolation and energy minimization

## The Yale Morph Server



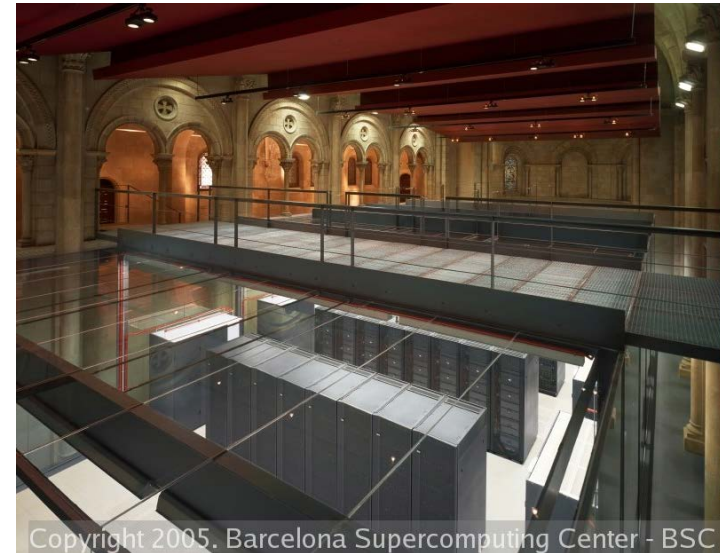
# Protein dynamics:

- Molecular dynamics simulations

- ◆ Databases

- ☞ MoDEL - Molecular Dynamics Extended Library (Orozco)

- <http://mmb.pcb.ub.es/MODEL/>
- Standard MD simulations
- Principal components available



Copyright 2005. Barcelona Supercomputing Center - BSC

- ☞ Dynameomics (Daggett)

- Protein unfolding using high temperature
- <http://www.dynameomics.org/>

- ☞ Etc

# Where can I find out about protein structure prediction tools?

- Collection:
  - <http://www.proteinmodelportal.org/?pid=101>
- CASP:
  - <http://predictioncenter.org/>
  - <http://www.forcasp.org/>
- Tertiary and Secondary structure prediction:
  - <http://predictioncenter.org/index.cgi?page=links>
  - **<http://www.jove.com/video/3259/a-protocol-for-computer-based-protein-structure-function>**
- Comparative modeling:
  - <http://www.salilab.org/modeller/modeller.html>
  - <http://swift.cmbi.ru.nl/teach/HOMMOD/>
- Folding@Home
  - <http://folding.stanford.edu/>

What do we already know about mutations in this protein or at this site in the protein ?

# ProSAT+: PROtein Structure Annotation Tool



## Find a protein structure

First Step: Find a protein structure to display annotations

You can get a protein structure to display sequence annotations via a PDB id, a UniProt id, or a sequence.

by PDB   by UniProt   by Sequence

Please enter a PDB code:

PDB code:

Check PDB



# ProSAT+:

PROtein Structure Annotation Tool



Display sequence annotations on

Chain A, Source P69905 (Homo sapiens)

3D structure

modified residue >

strand >

helix >

variant >

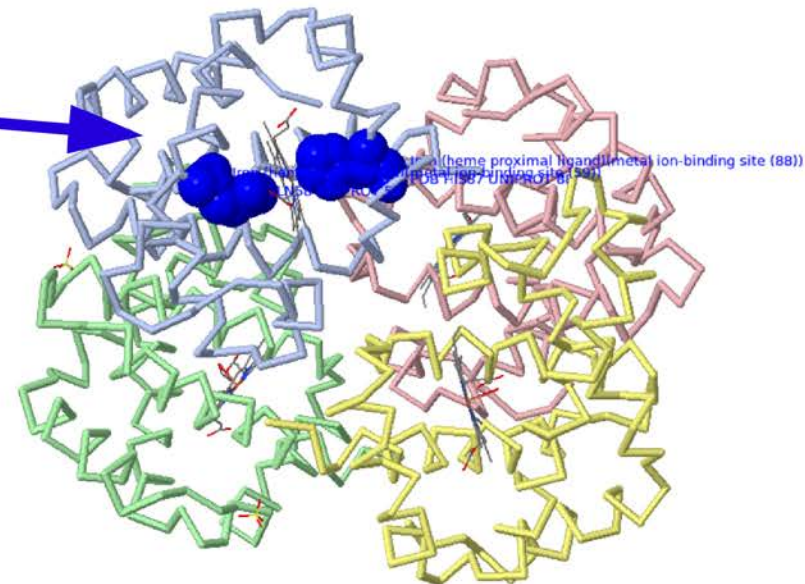
binding >

Iron (heme distal ligand) 58:A (59)

Iron (heme proximal ligand) 87:A (88)

turn >

kinetics >



<http://prosat.h-its.org>

Stank, Richter, Wade PEDS, 2016, in press

# ProSAT+: PROtein Structure Annotation Tool



Define sequence annotations  
and generate URL

ProSAT+ - Generate user URL

User defined sequence annotations

#	UniProt accession	Position(s)	Name	Description	
1	<a href="#">P69905</a>	<a href="#">1-3,5</a>	<a href="#">Name 1</a>	<a href="#">Sequence Description 1</a>	<input type="button" value="X"/>
2	<a href="#">P68871</a>	<a href="#">66</a>	<a href="#">Name 2</a>	<a href="#">Sequence Description 2</a>	<input type="button" value="X"/>
<input type="text"/>	:	<input type="text" value="position(s)"/>	<input type="text" value="name, do not use ;:  or &amp;"/>	<input type="text" value="description, do not use ;:  or &amp;"/>	<input type="button" value="Add"/>

Here you can enter sequence annotations in the following format

P60174:123-129,120,135-140:name1:ThisisOneFeatureDescription|P60174:250,255:name2:AnotherFeature

You can also transfer all your selected sequence features and amino acids into the table:

Your Url is:

```
http://localhost:9000/prosat/prosatexe?pdbcode=1j7y&submit=Check+PDB&
user=P69905:1-3,5:Name%20:Sequence%20Description%20|P68871:66:Name%20:Sequence%20Description%20|
```

# ProSAT+: PROtein Structure Annotation Tool



Use sequence annotations from evolutionarily related proteins

1J7Y Hemoglobin

Chain A Chain B Chain C Chain D

Chain A

Uniprot Accession: P69905  
Name: Hemoglobin subunit alpha  
Enzyme class:  
Taxon: Homo sapiens  
P69905 Sequence: M/- V/M L S P A D K T N V K A A W G

Similarity

- P01923 (99 %) Gorilla gorilla gorilla, 19
- Q9TS35 (99 %) Hylobates lar, 2
- P06635 (98 %) Pongo pygmaeus, 20
- P01924 (98 %) Semnopithecus entellus, 19
- P63107 (97 %) Macaca fuscata fuscata, 18
- P67817 (96 %) Ateles geoffroyi, 18
- Q9TS34 (96 %) Hylobates lar, 2
- P18972 (97 %) Callithrix argentata, 18
- P01926 (96 %) Chlorocebus aethiops, 18
- P21767 (97 %) Macaca fascicularis, 23

Sequence Features [Show features panel](#)

Q9TS35 Sequence (Hylobates lar)

P06635 Sequence (Pongo pygmaeus)

P01924 Sequence (Semnopithecus entellus)

P63107 Sequence (Macaca fuscata)

Chain A, Source P21767 (Macaca fascicularis)

- modified residue
- variant
- In alpha-R. (T8S) 8:A (8)
- In alpha-R. (V55I) 55:A (55)
- In alpha-Q. (G71D) 71:A (71)
- In alpha-R and alpha-T. (Q78H) 78:A (78)
- binding

# ProSAT+:

PROtein Structure Annotation Tool

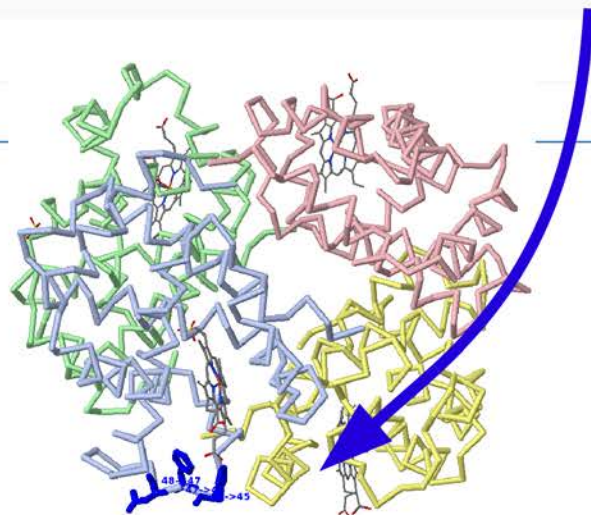


## Search for sequence motifs

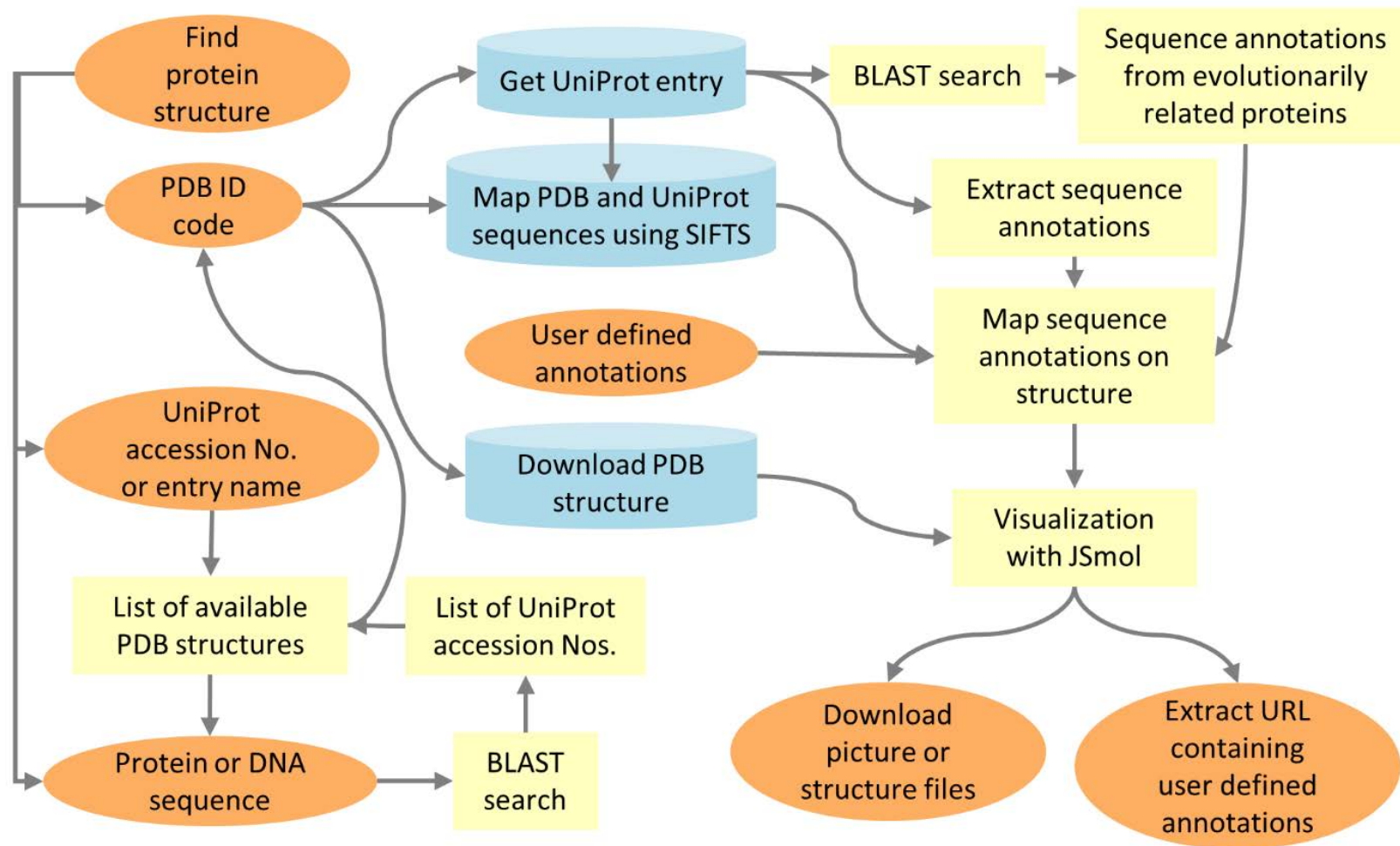
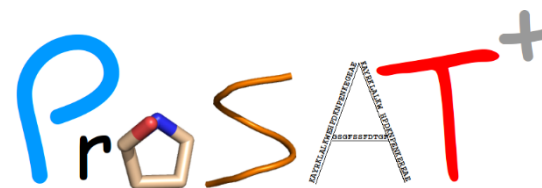
Motif search:

Motif found:

Chain	Match	Position(s)
<input checked="" type="checkbox"/> Chain A	HFD	<input checked="" type="checkbox"/> A 46-48



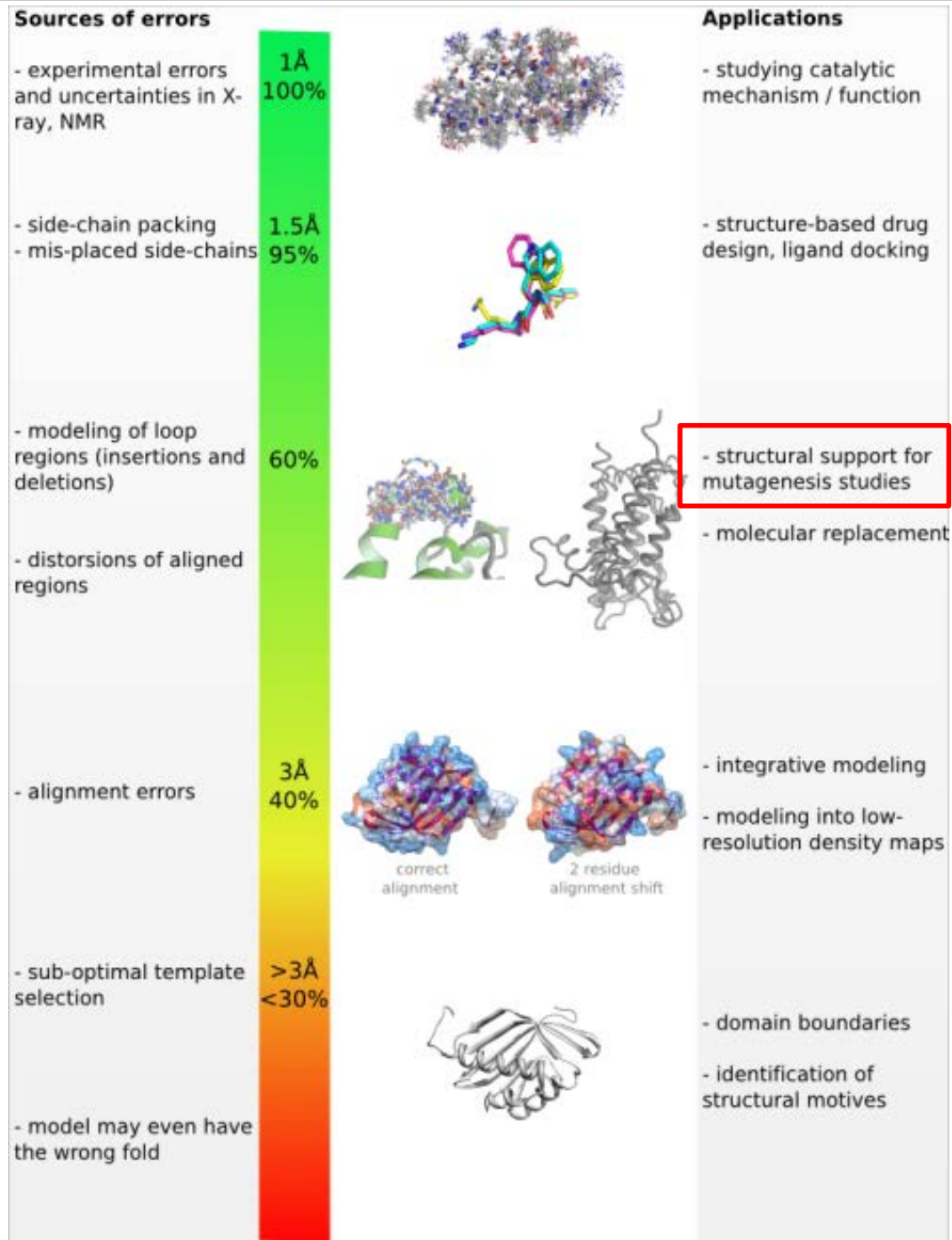
# ProSAT+: Workflow



What mechanistic effect does a particular mutation have?

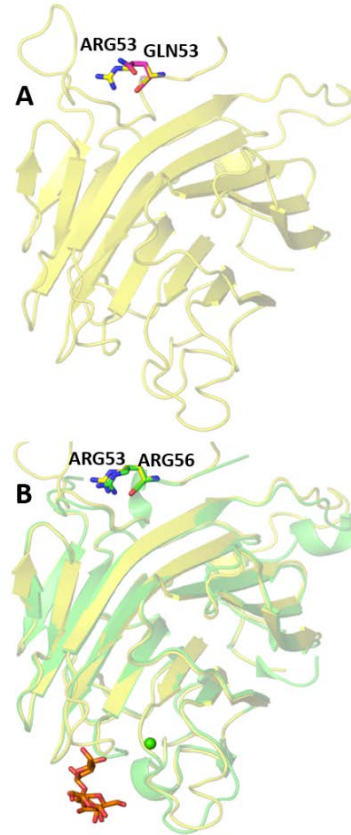
# Using protein models to predict the effects of mutations

<http://www.proteinmodelportal.org/?pid=documentation#modelquality>

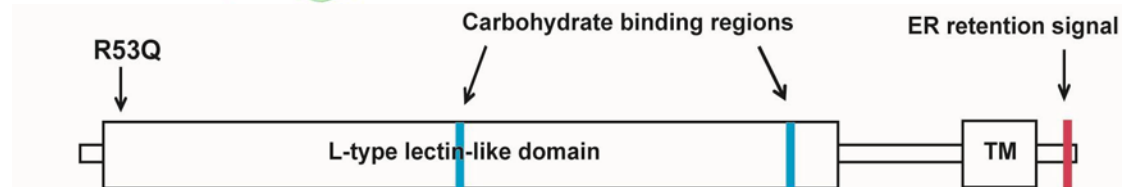


# Missense mutation : mechanism?

- Homozygous missense mutation in *lectin mannose-binding 2-like* (**LMAN2L**) gene (**R53Q**), identified in whole exome sequencing, segregates with severe **intellectual disability & epilepsy** in consanguineous Pakistani family
- **LMAN2L**: ER cargo receptor for glycoprotein transport & quality control, expressed in brain



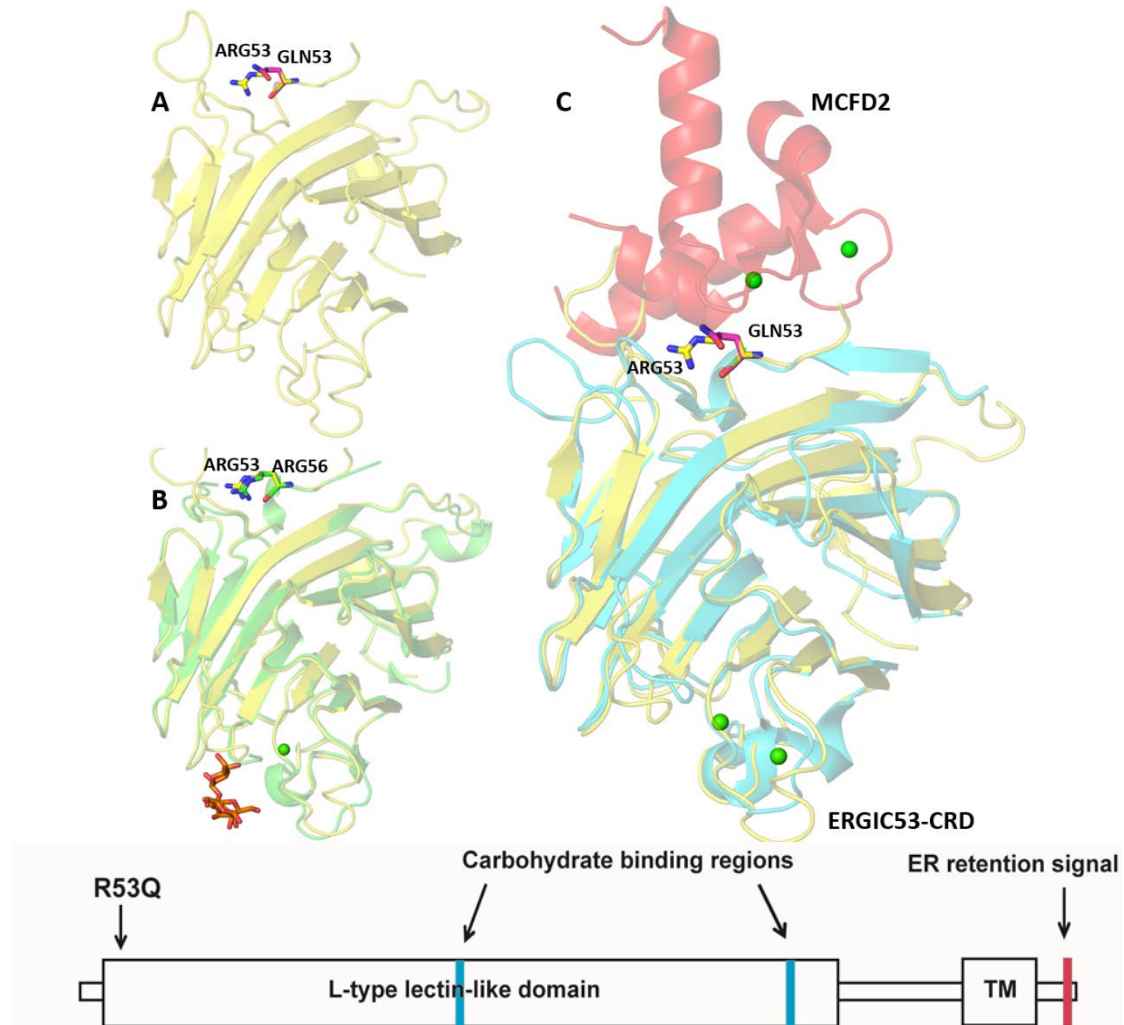
- **LMAN2L** (Swiss) modeled on **Vip36** (*lectin mannose-binding 2, dog*) template (63% SI)
- Mutation on opposite side from **Vip36**  $\text{Ca}^{2+}$  and **mannose** binding site





# Missense mutation : mechanism?

- Superimpose model on **ERGIC-53-CRD** (*ER Golgi intermediate compartment 53 carbohydrate recognition domain, man*) (35% SI to **LMAN2L** domain) bound to **MCFD2** (*multiple coagulation factor deficiency 2*) to form cargo receptor
- 3D model indicates **R53Q** impairs protein-protein interaction in **LMAN2L**
  - Multimer
  - Unknown protein
  - ERGIC-53



# Missense mutation : mechanism?

Chain A Chain B Chain C Chain D Chain E

Chain A

Uniprot Accession **P49256**  
Name Vesicular integral-membrane protein VIP36  
Enzyme class  
Taxon **Canis lupus**

P49256 Sequence

Positions are given in natural positions.

/	-	T	/	D	/	G	/	N	/	S	E	H	L	K	R	E	H	S	L	I	K	P	
16	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	

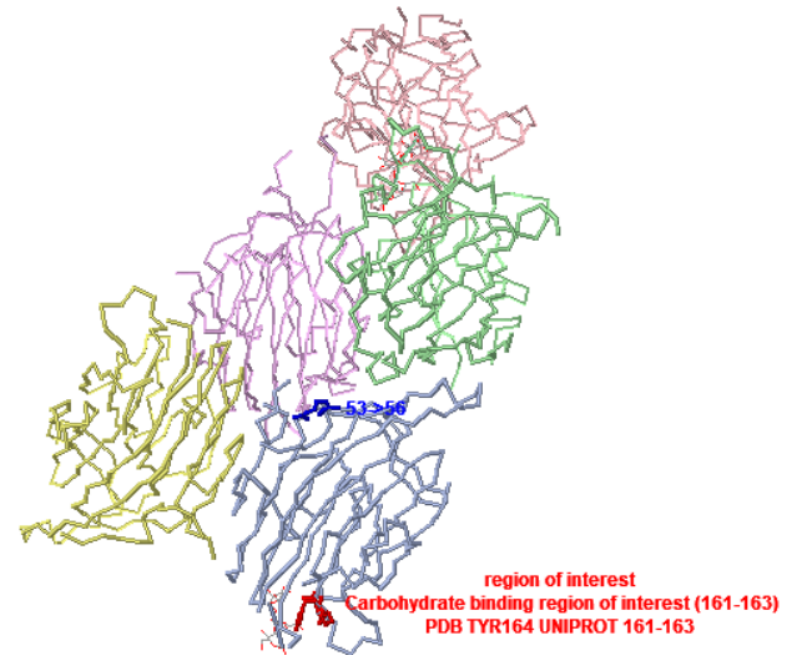
Caution: The transfer of annotations from similar sequences can be misleading. Lower sequence similarity will result in a lower reliability of the annotations.

Q9H0V9 Sequence (Homo sapiens)

T	/	S	P	/	E	Y	/	H	L	K	R	E	H	S	L	S	/	I	K	P	Y	Q	G	V	G	T	
47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	

Sequence Features [Show features panel](#)

Sequence features and sequences taken from [Uniprot](#).



JSmo

# Missense mutation : mechanism?

You searched for UniProt Accession P49257 (LMAN1\_HUMAN):

Check PDB

## 3A4U Protein ERGIC-53

Chain A Chain B

Chain A

Uniprot Accession P49257  
Name Protein ERGIC-53  
Enzyme class  
Taxon Homo sapiens

P49257 Sequence

Positions are given in natural positions.

A	V	A	L	P	H	R	R	F	E	Y	K	Y	S	F	K	G	P	H	I
									✓										
38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57

Similarity

Find similar sequences

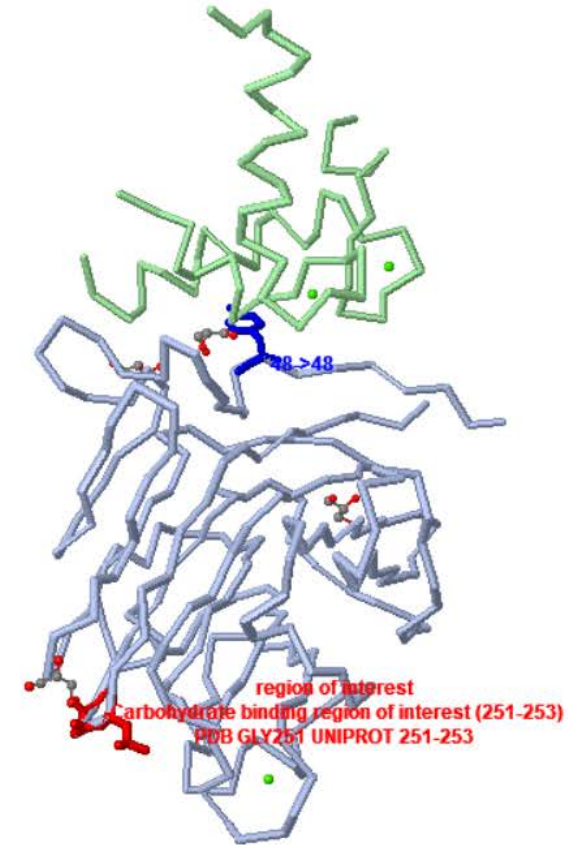
Sequence Features

Show features panel

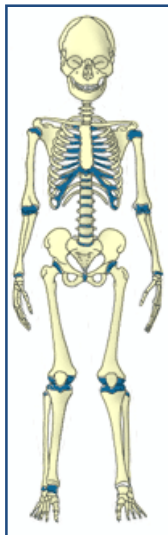
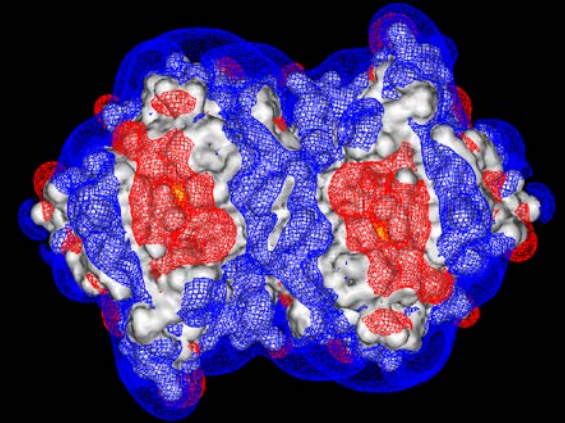
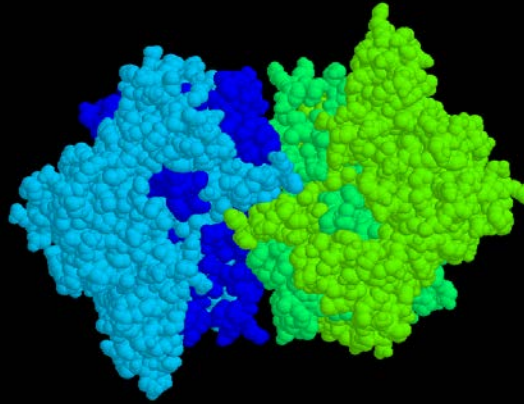
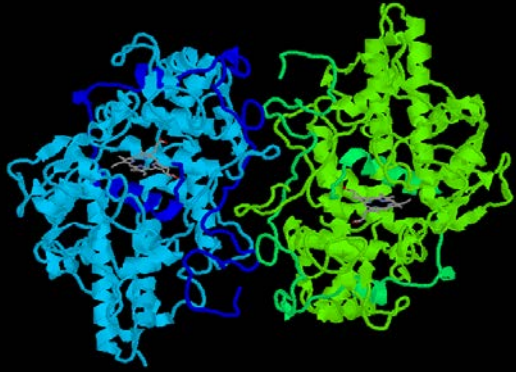
Sequence features and sequences taken from Uniprot.

Structural data retrieved from RCSB.

Mapping data obtained from SIFTS or pairwise alignments using BioJava, additional similar sequences added using Blast.



# How do biomolecules recognize each other?

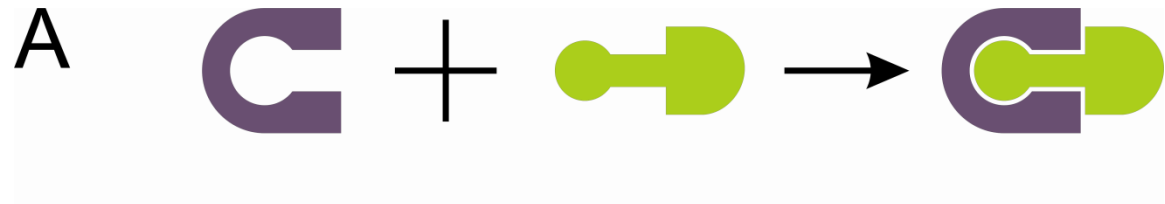


# Challenges for predicting receptor-ligand interactions computationally

- **Scoring**
  - Force field, energy function, etc
- **Sampling**
  - Simulation length, space, degrees of freedom, multigraining, etc

# Receptor-ligand binding paradigms

Lock and key

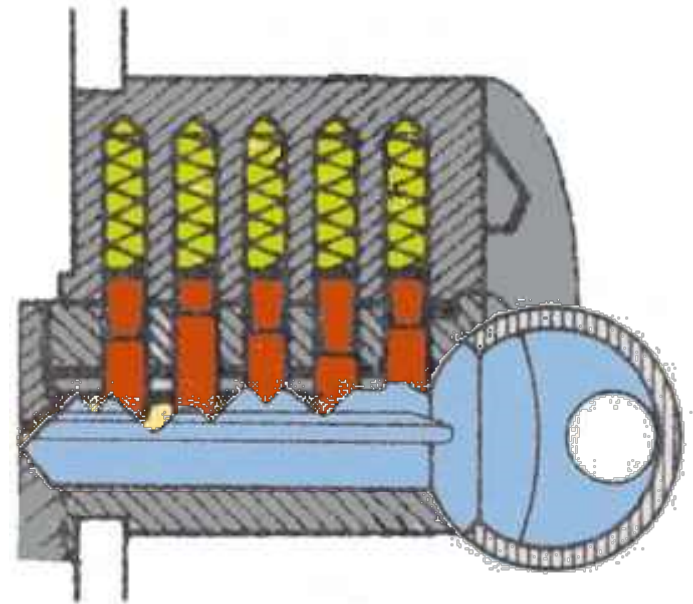
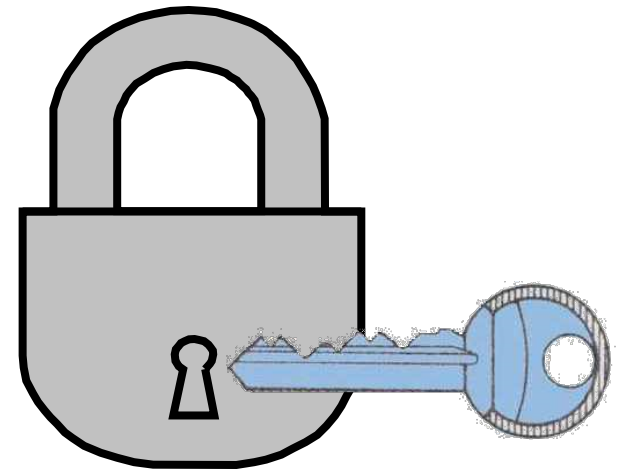


Receptor

Ligand

# Lock & Key Model

- Fischer, E. (1894) Einfluss der Configuration auf die Wirkung der Enzyme Ber. Deutsch Chem Ges., 27, 2985-2993.
- „.....dass Enzym und Glucosid wie Schloss und Schlüssel zu einander passen müssen....“
- Laskowski & Thornton (1995): SURFNET: detect 80% of enzyme active sites by looking for crevices on proteins



# MetaPocket

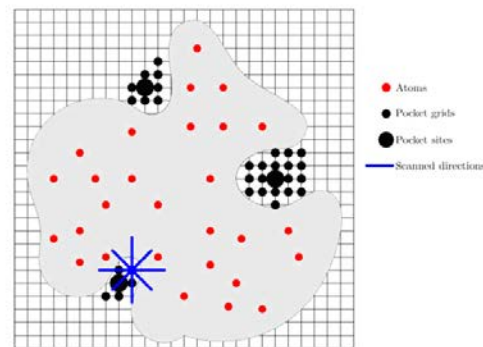
- MetaPocket 2.0 [projects.biotec.tu-dresden.de/metapocket/](http://projects.biotec.tu-dresden.de/metapocket/)

– Consensus method to predict binding sites on protein surfaces

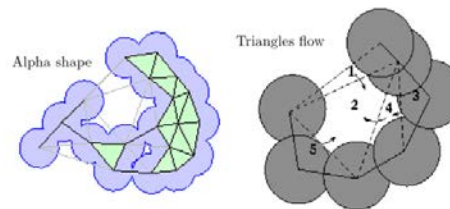
– 8 methods

- LIGSITE<sup>CS</sup>
- PASS
- Q-SiteFinder
- SURFNET
- Fpocket
- GHECOM
- ConCavity
- POCASA

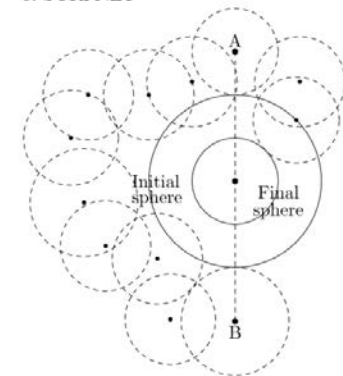
a. POCKET, LIGSITE, LIGSITE<sup>CS</sup>



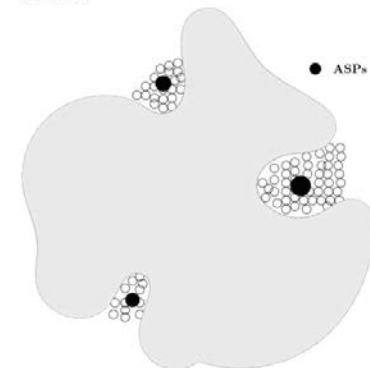
c. CAST



b. SURFNET



d. PASS



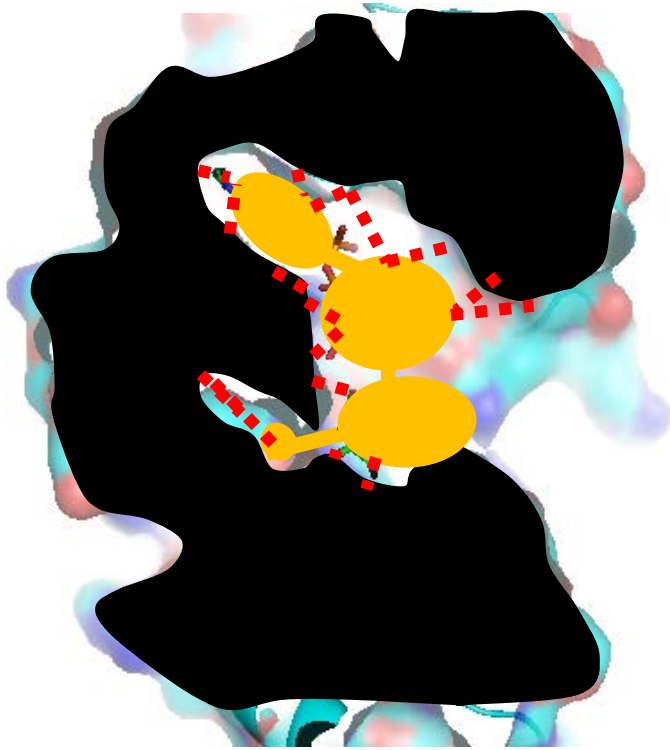


# MetaPocket

- **MetaPocket 2.0** <http://projects.biotec.tu-dresden.de/metapocket/>
  - Consensus method to predict binding sites on protein surfaces
  - 8 methods (MPK2) (MPK1: 4 methods)

Dataset	Version	Top1	Top2	Top3	Top4
48(bound)	MPK2	85	92	96	96
	MPK1	83	94	96	96
48(unbound)	MPK2	80	90	94	96
	MPK1	75	85	90	92
210(bound)	MPK2	81	91	95	96
	MPK1	76	89	94	96
Drug-Target	MPK2	61	70	74	76
198(bound)	MPK1	55	65	68	72

# Structure-based drug design: the lock & key paradigm

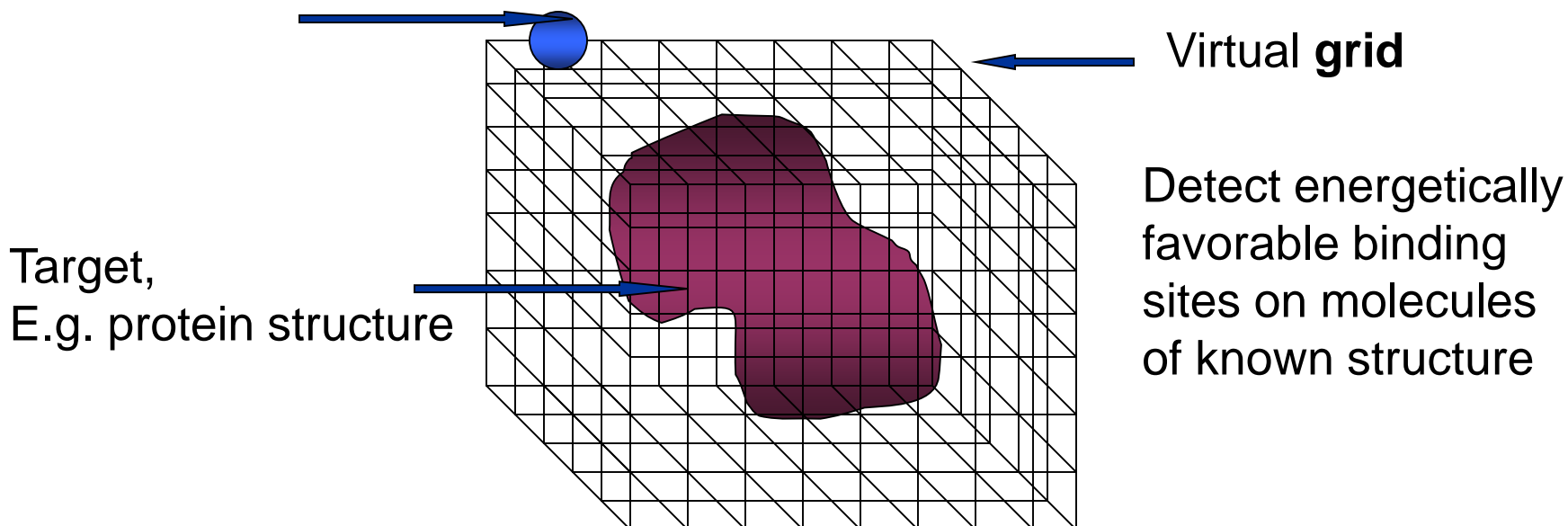


A good compound must:

- **Bind strongly**
  - many favourable contacts, shape and chemical complementarity
- **Bind selectively**
  - complex shape, etc

# GRID Molecular Interaction Fields

Chemical probe,  
E.g. water, amino group, proton



$$\Delta E = \sum_i E_{LJ} + \sum_i E_{EL} + \sum_i E_{HB} + S$$

Goodford, PJ *J. Med. Chem.* (1985) 28, 849-857.  
Boobbyer et al, *JMC.* 1989, Wade et al, *JMC.* 1993

# Structure-based drug design : Influenza

ARTICLES *Nature*. 1993 363:418-23.

## Rational design of potent sialidase-based inhibitors of influenza virus replication

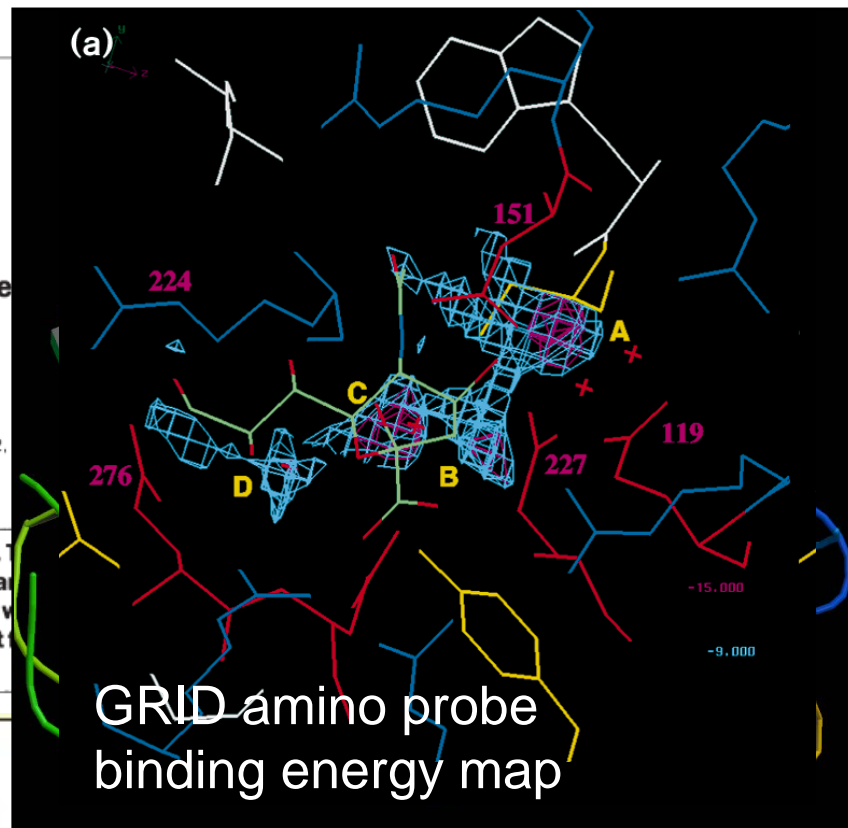
Mark von Itzstein\*, Wen-Yang Wu\*, Gaik B. Kok\*, Michael S. Pegg\*, Jeffrey C. Dyason\*, Betty Jin\*, Tho Van Phan \*, Mark L. Smythe\*, Hume F. White Stuart W. Oliver\*, Peter M. Colman‡, Joseph N. Varghese‡, D. Michael Ryan§, Jacqueline M. Woods§, Richard C. Bethell§, Vanessa J. Hotham§, Janet M. Cameron§ & Charles R. Penn§

\*Department of Pharmaceutical Chemistry, Victorian College of Pharmacy, Monash University, 381 Royal Parade, Parkville, Victoria 3052, Australia

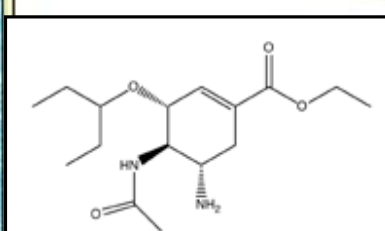
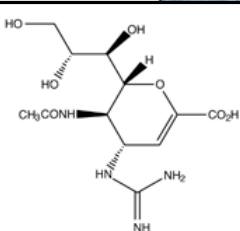
‡CSIRO Division of Biomolecular Engineering, Parkville, Victoria 3052, Australia

§Glaxo Group Research Ltd, Greenford, Middlesex UB6 OHE, UK

Two potent inhibitors based on the crystal structure of influenza virus sialidase have been designed. The compounds are effective inhibitors not only of the enzyme, but also of the virus in cell culture and in animal models. The results provide an example of the power of rational, computer-assisted drug design, as well as indicating significant progress in the development of a new therapeutic or prophylactic treatment for influenza infection.



GRID amino probe  
binding energy map



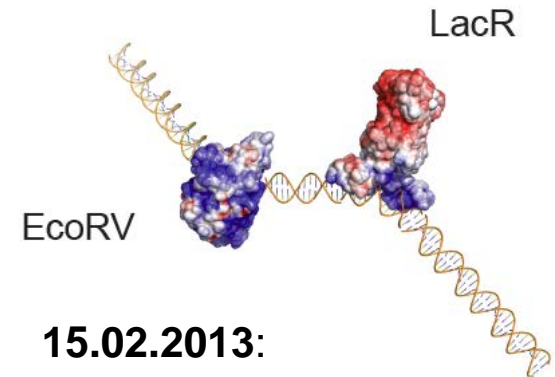
Review:

SBDD against influenza:  
Wade, *Structure*, 1997, 5,  
1139-1145

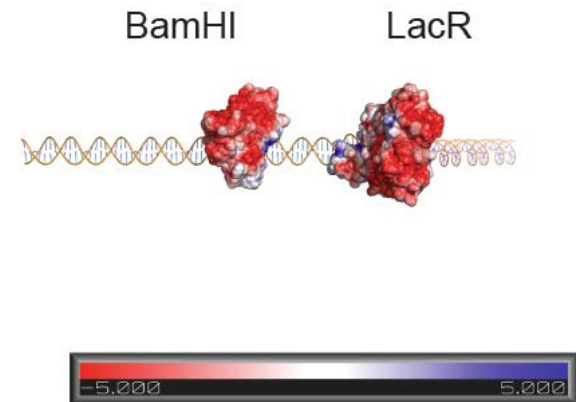
# **Protein electrostatic properties**

# ...in Science:

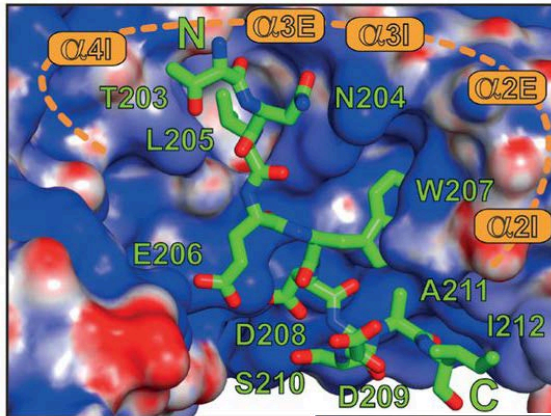
A



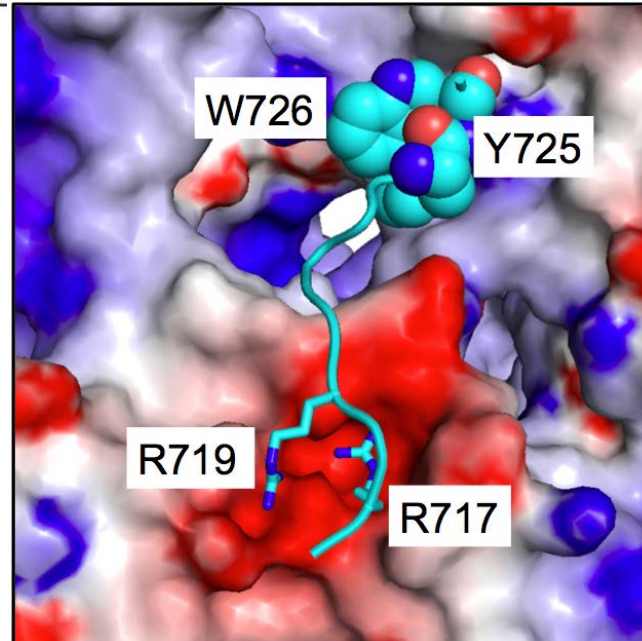
B



C



19.04.2013:  
Pernigo et al,  
(2013) 340, 356

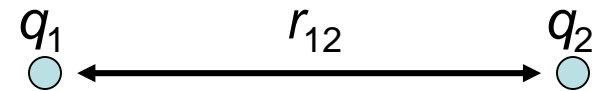


31.05.2013:  
Kato et al,  
(2013) 340, 1110

# Coulomb's Law - Energy

- Interaction energy of two point-charges *in vacuo*
- Solve Poisson equation
- In SI units:

$$U = \frac{q_1 q_2}{4\pi\epsilon_0 r_{12}}$$



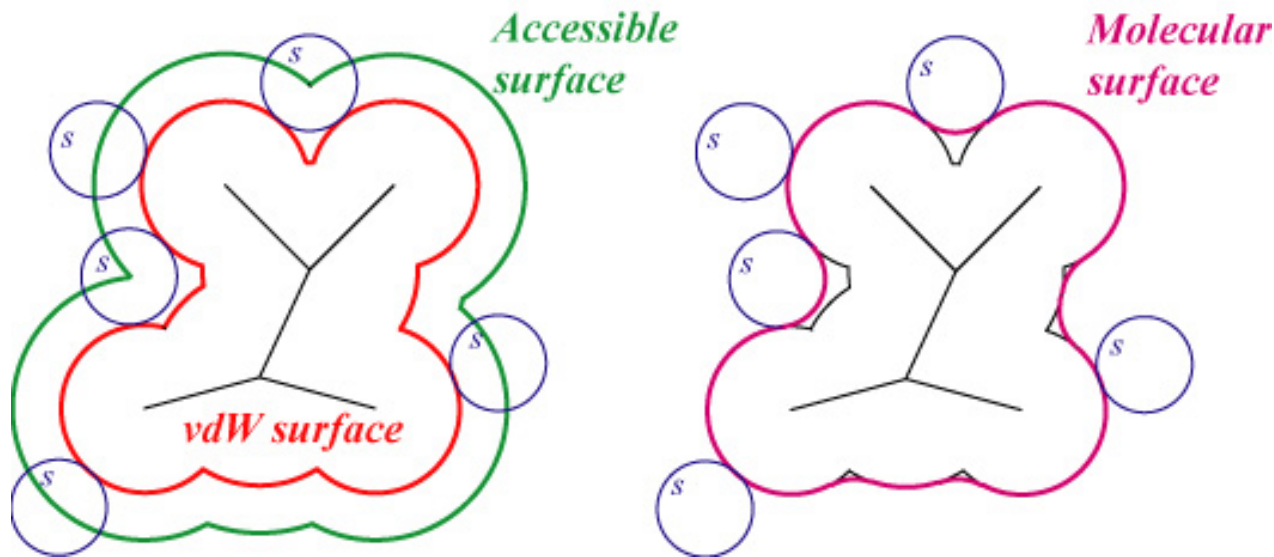
- In “biomolecular” units:

$$U = \frac{332q_1q_2}{r_{12}}$$

Energy,  $U$ : kcal/mol  
Charge,  $q$ : electron charges  
Distance,  $r$ : Angstroms

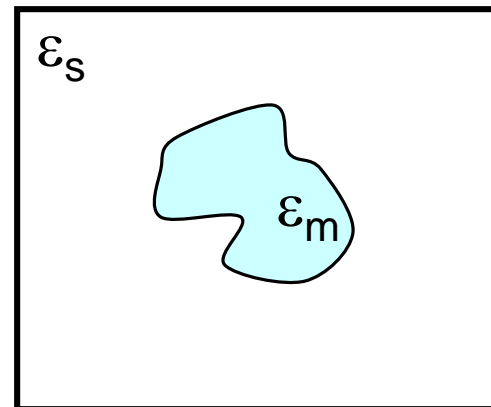
- Electrostatic potential:  $\phi(\mathbf{r}_2) = 332q_1/r_{12}$

# Continuum electrostatics for molecules



<http://csb.stanford.edu/koehl/ProShape/>

Solute Molecule:  $\epsilon_m \sim 2-4$   
Solvent (water) :  $\epsilon_s \sim 80$





# Mobile ions in the solvent

- Ionic solution with dissolved ions (electrolyte)
  - Ions redistribute in the presence of a molecule with charges to weaken/screen its electrostatic interactions
- Debye-Hueckel theory
  - Implicit model of the ions in the solvent
  - Ions assumed to distribute according to the local potential with a Boltzmann factor

$$c_{ion}(\mathbf{r}) = c_{ion,bulk} e^{-\beta\phi(\mathbf{r})q_{ion}}$$

# Continuum Electrostatics

- Poisson-Boltzmann equation

$$-\varepsilon_0 \nabla \cdot [\varepsilon_r(\mathbf{r}) \nabla \phi(\mathbf{r})] = \rho^f(r) + \sum_{i=1}^N q_i c_{i,bulk}(\mathbf{r}) e^{-\beta \phi(\mathbf{r}) q_i}$$

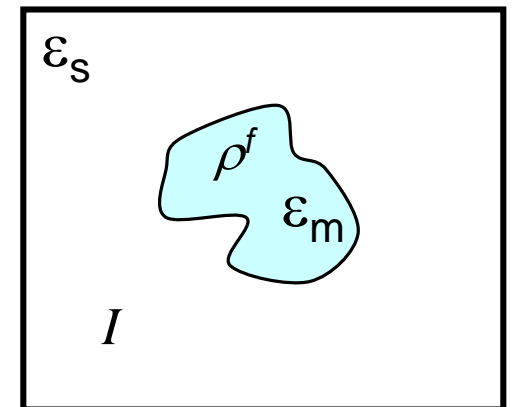
- Linearized Poisson-Boltzmann equation (weak  $\phi$ , low  $I$ )

$$-\varepsilon_0 \nabla \cdot [\varepsilon_r(\mathbf{r}) \nabla \phi(\mathbf{r})] = \rho^f(r) - \varepsilon_0 \varepsilon_r(r) \kappa^2(r) \phi(r)$$

$$\kappa^2(r) = \frac{\beta}{\varepsilon_0 \varepsilon_r} \sum_1^N c_{i,bulk} q_i^2 = \frac{2e^2 N_A I}{\varepsilon_0 \varepsilon_r kT}$$

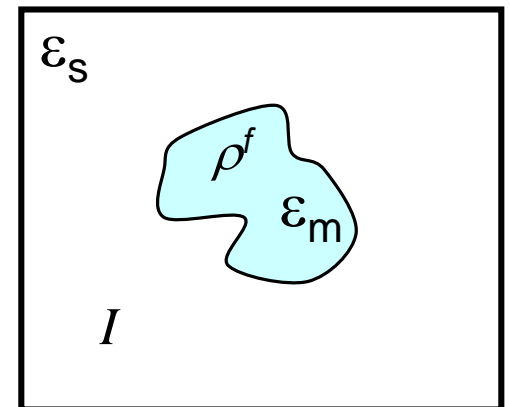
Debye length =  $1/\kappa$   
 =  $3.04/I^{1/2}$  Å at 300K  
 = 8 Å at  $I=150$ mM

- Analytical solution only for simple shapes
- Numerical solution
- Simple approximations, e.g.  $\varepsilon_r = K r_{ij}$

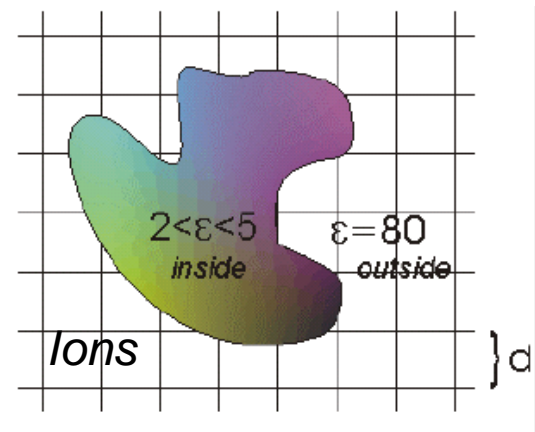
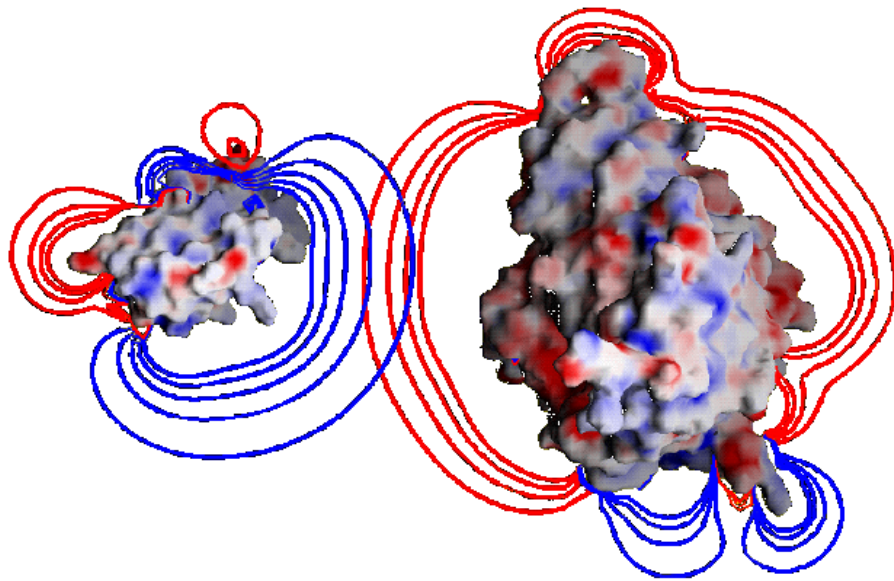
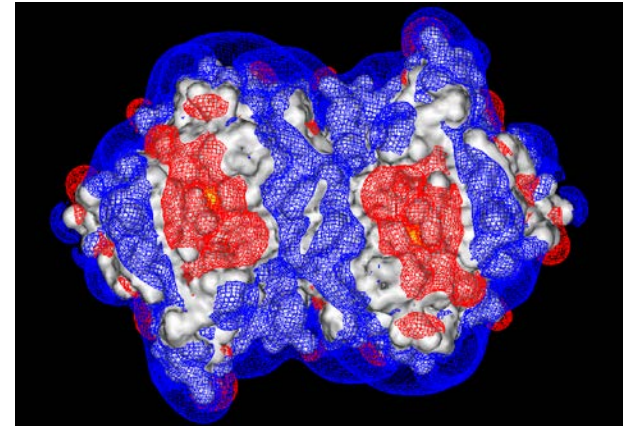
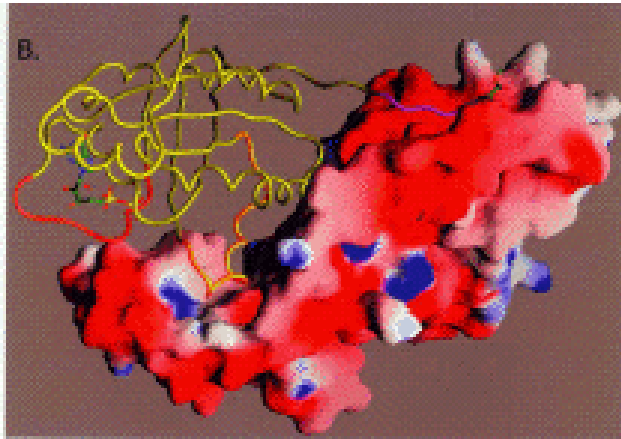


# Solving the Poisson-Boltzmann equation for biomolecules

- Need to assign:
  - Atomic partial charges
  - Atomic radii
  - Dielectric constants (solute and solvent)
  - Ionic strength
  - Ion exclusion layer etc
  
- Linear vs non-linear
- Solution method (FD, FE etc)
- Convergence



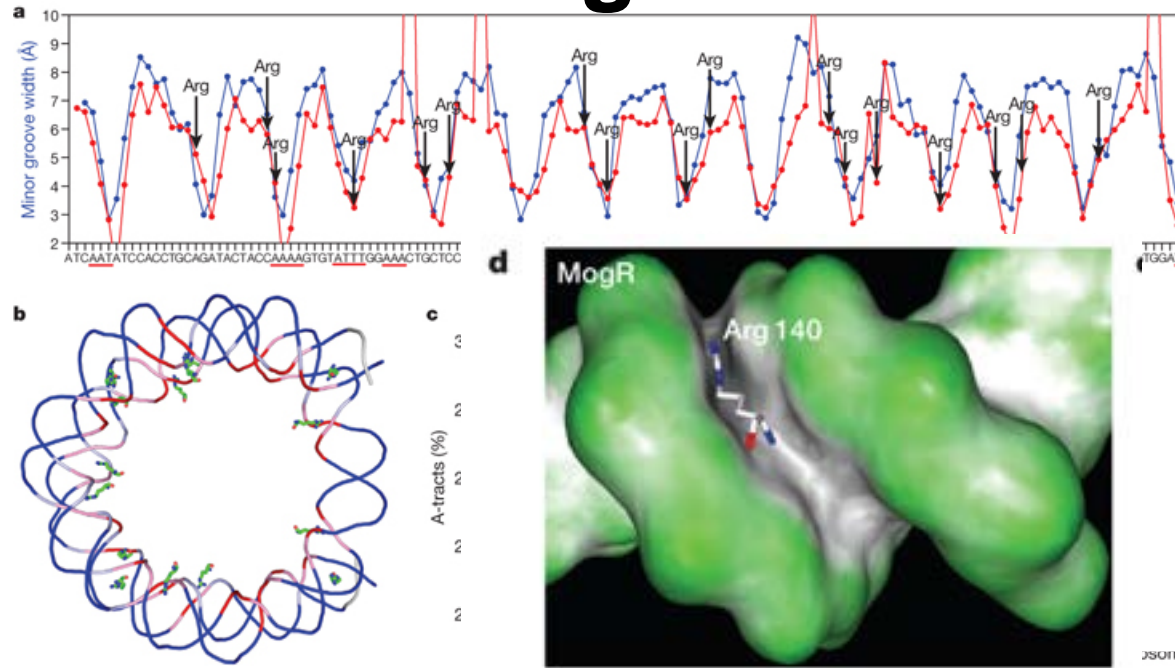
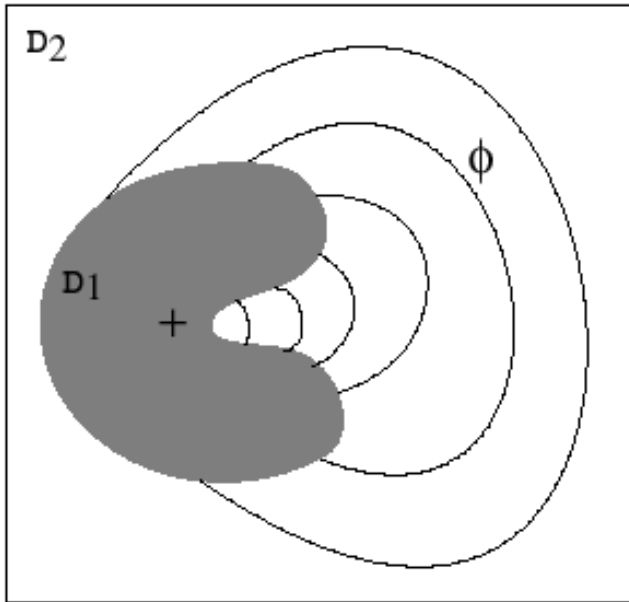
# Protein electrostatic potentials



$$\Phi(r) = \frac{q}{4\pi\epsilon_0\epsilon r}$$

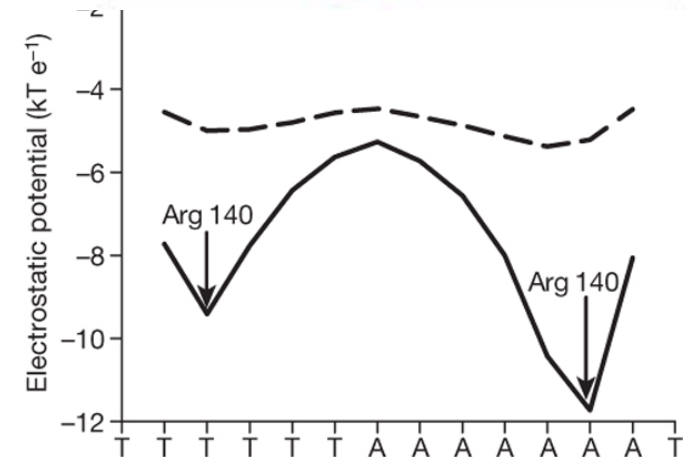
# Electrostatic focusing

- Due to total dielectric environment
- Non-spherical isocontours around charges



$$\epsilon_m = \epsilon_s = 80$$

$$\epsilon_m = 2; \epsilon_s = 80$$



Minor-groove shape recognition by arginines.

# Electrostatic solvation

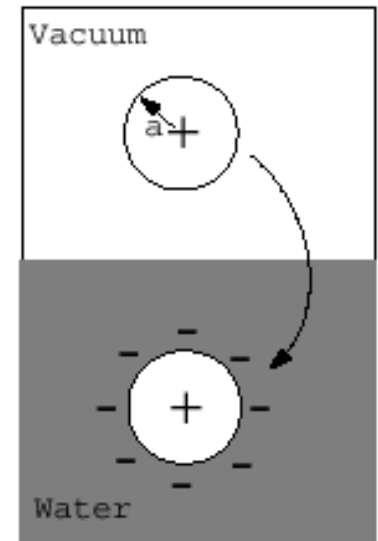
- In a high dielectric solvent (e.g. water), charges will tend to be repelled from low dielectric solutes
- Charge polarizes solvent, which produces *reaction field* at charge with which charge interacts

- **Born ion solvation**

- Work done to transfer Born ion between 2 dielectrics
- Born ion is point charge in spherical cavity

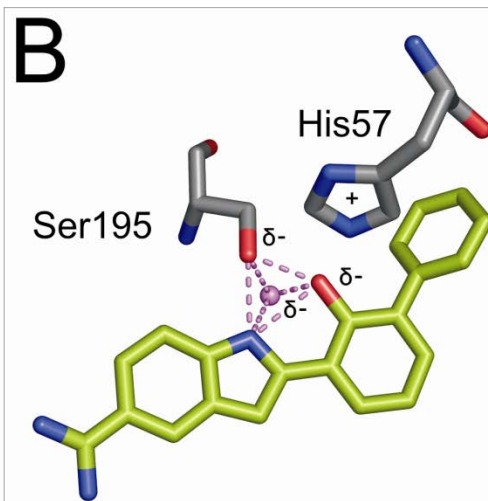
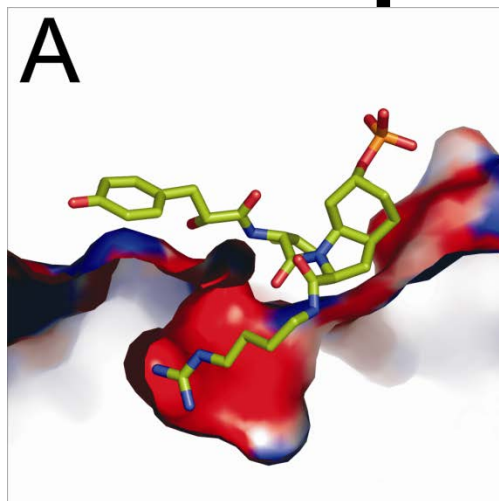
$$U_{Born} = \frac{332q^2}{2a} \left( \frac{1}{\epsilon_2} - \frac{1}{\epsilon_1} \right)$$

- E.g. Na<sup>+</sup>, K<sup>+</sup>, Cl<sup>-</sup> in water: a ~ 1.5-2.5 Å
- Free energy of hydration ~ -100- -50 kcal/mol



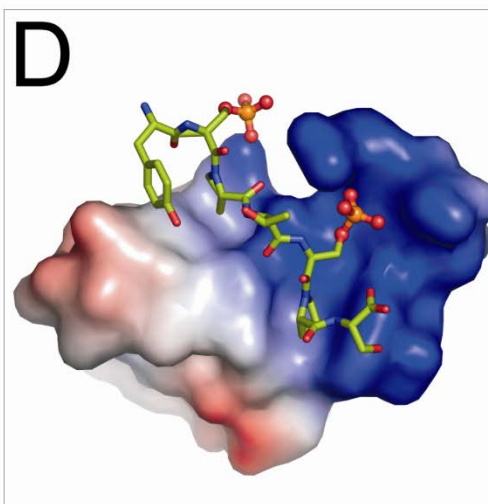
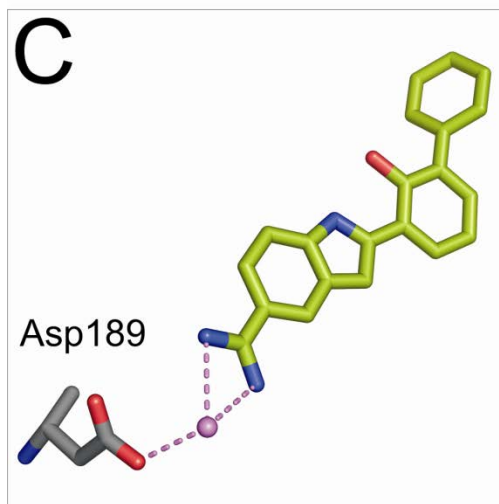
# Electrostatics in ligand-receptor complexes

Inhibitor Aeruginosin 98-B bound to negatively charged pocket of trypsin



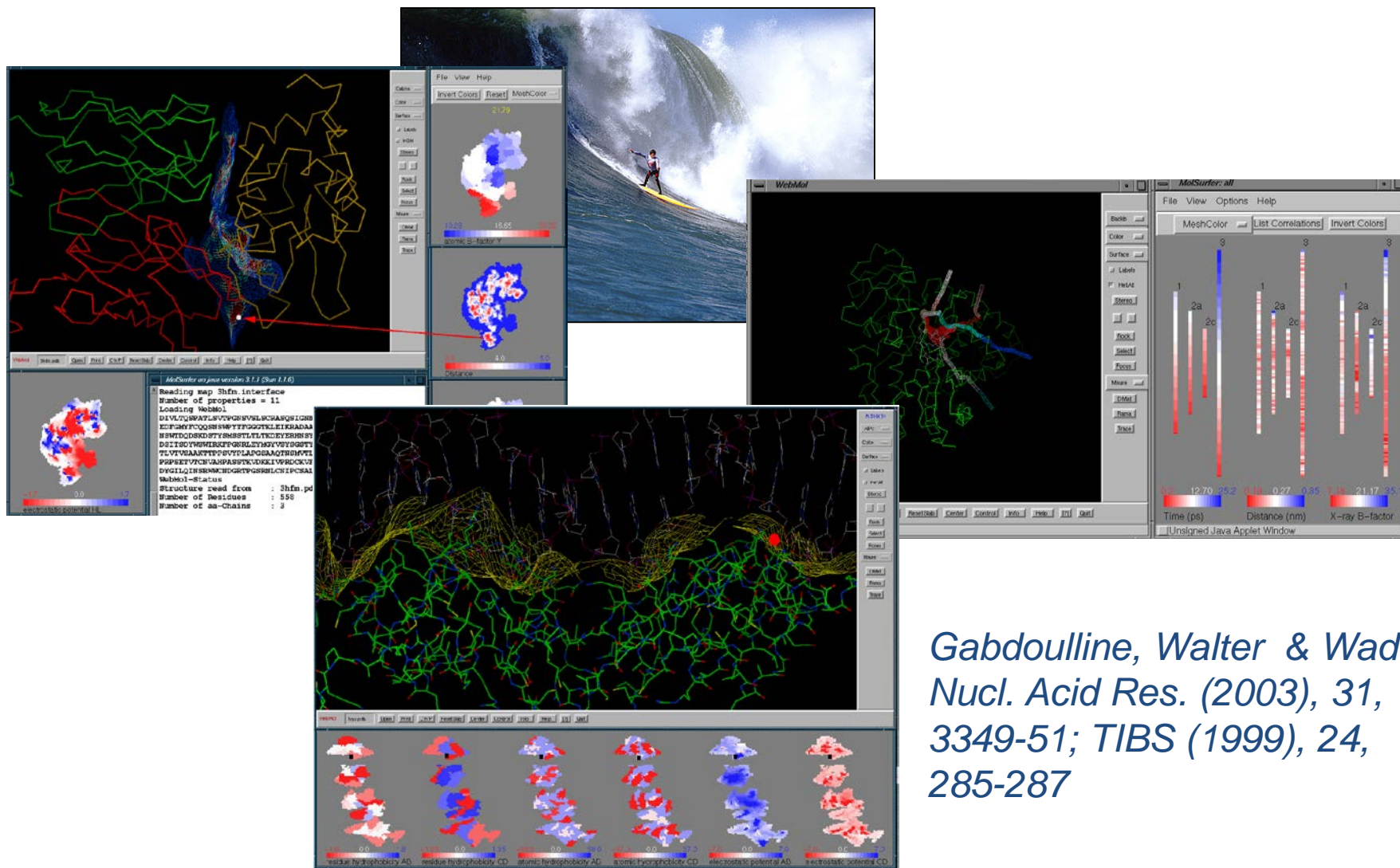
Water mediated short H-bond network in Urokinase-Type Plasminogen activator-inhibitor complex

Water mediated salt-bridge between Urokinase-Type Plasminogen activator Asp189 and inhibitor



Phosphoserine-proline containing peptide bound to group IV WW domain area with positive electrostatic potential

# Molsurfer: a Macromolecular Interface Navigator



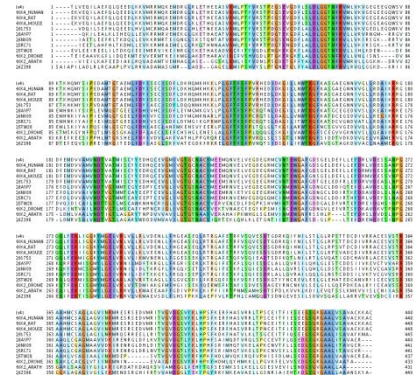
*Gabdoulline, Walter & Wade,  
Nucl. Acid Res. (2003), 31,  
3349-51; TIBS (1999), 24,  
285-287*

<http://projects.h-its.org/dbase/molsurfer/index.html>



# Levels of Protein Comparison

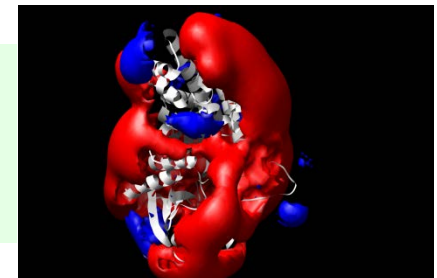
- Amino Acid Sequence Identity



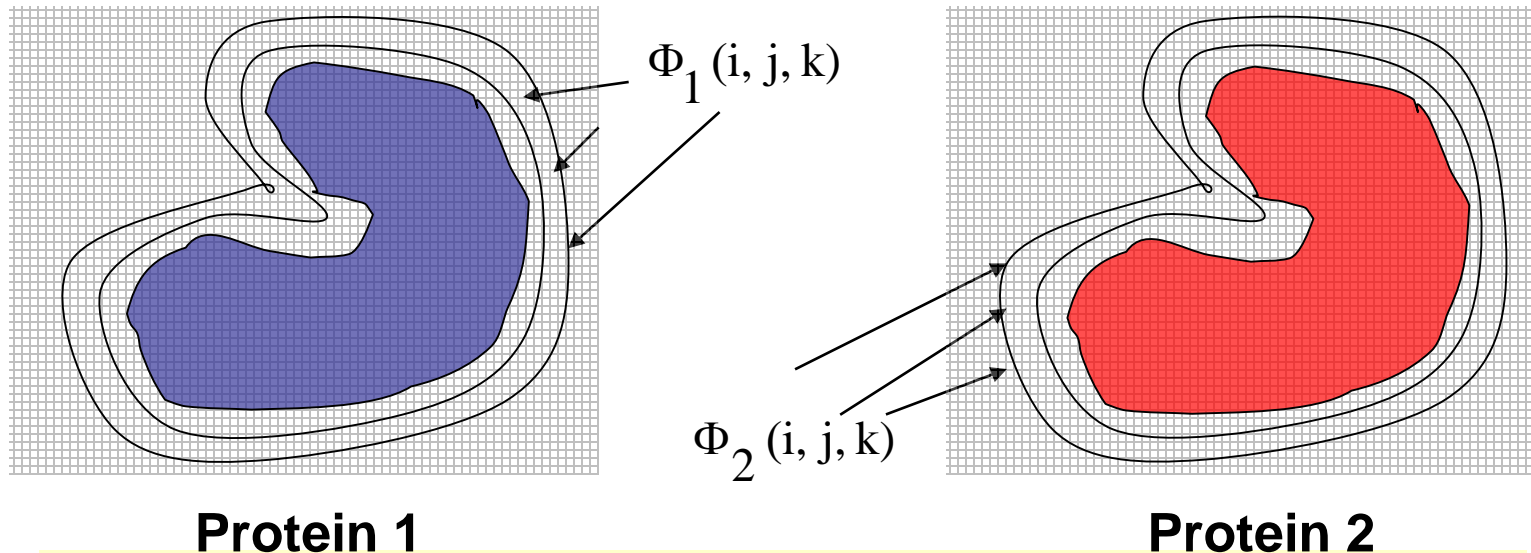
- Protein Structure (NMR, X-Ray)



- Protein Structure/Function Relationship - MIF

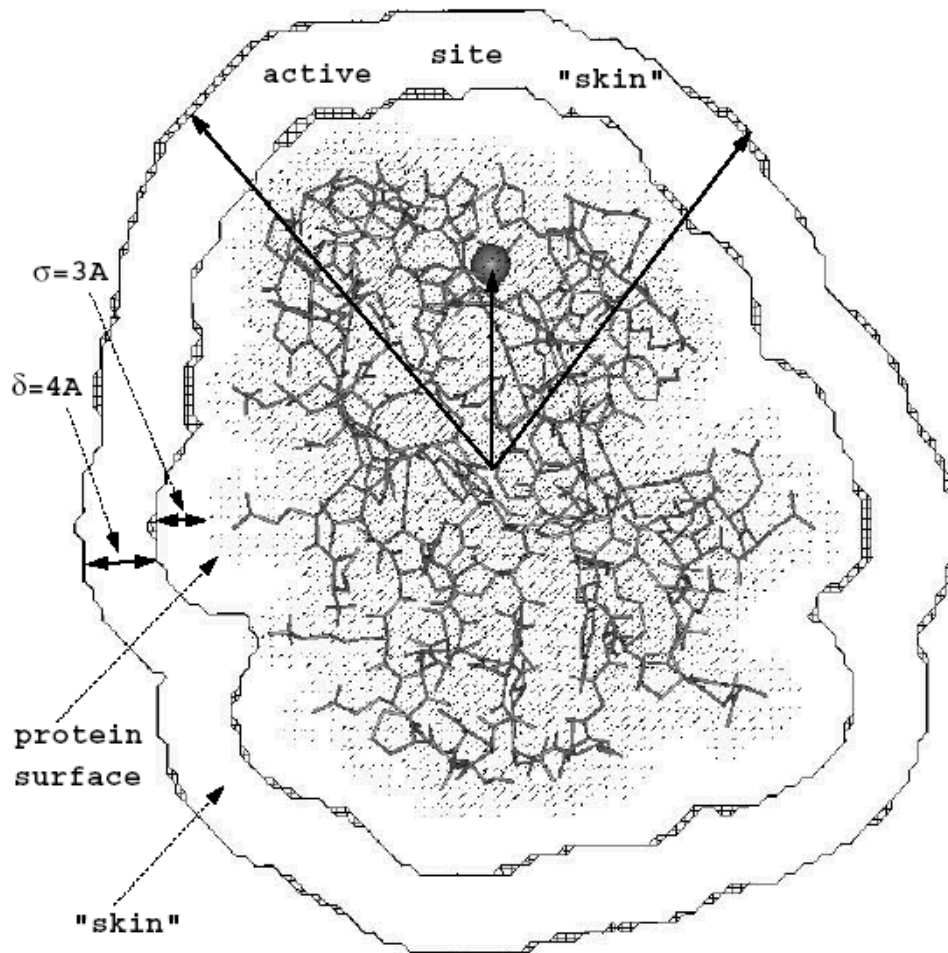


# PIPSA: Protein Interaction Property Similarity Analysis



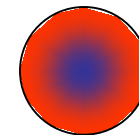
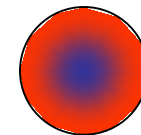
- Interaction fields are calculated on a set of points
- Field values on corresponding points are compared
- $\Phi$  = electrostatic potential, shape, probe interaction field, ...

# PIPSA: Protein Interaction Property Similarity Analysis

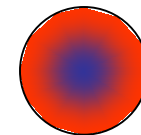


$$SI_{12} = \frac{2(\mathbf{p}_1, \mathbf{p}_2)}{(\mathbf{p}_1, \mathbf{p}_1) + (\mathbf{p}_2, \mathbf{p}_2)}$$

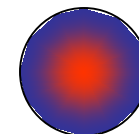
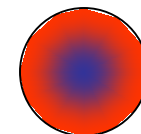
$$(\mathbf{p}_1, \mathbf{p}_2) = \sum_{i,j,k} \phi_1(i, j, k) \phi_2(i, j, k)$$



S = +1



S = 0



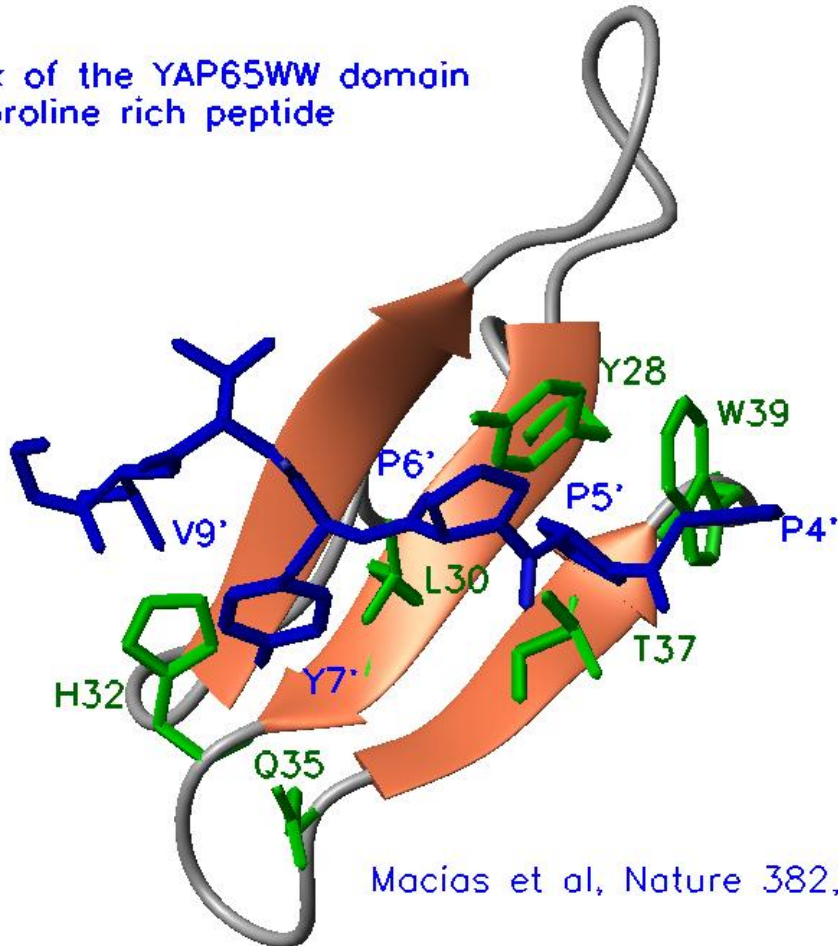
S = -1

*Wade et al., PNAS, 1998; Blomberg et al. Proteins 1999; De Rienzo et al. Protein Sci. 2000; Wade et al. Intl. J. Quant. Chem. 2001*

# WW domain/peptide complexes

Binding specificity and affinity determinants?

Complex of the YAP65WW domain  
and a proline rich peptide



Macias et al, Nature 382,646–649 (1996)

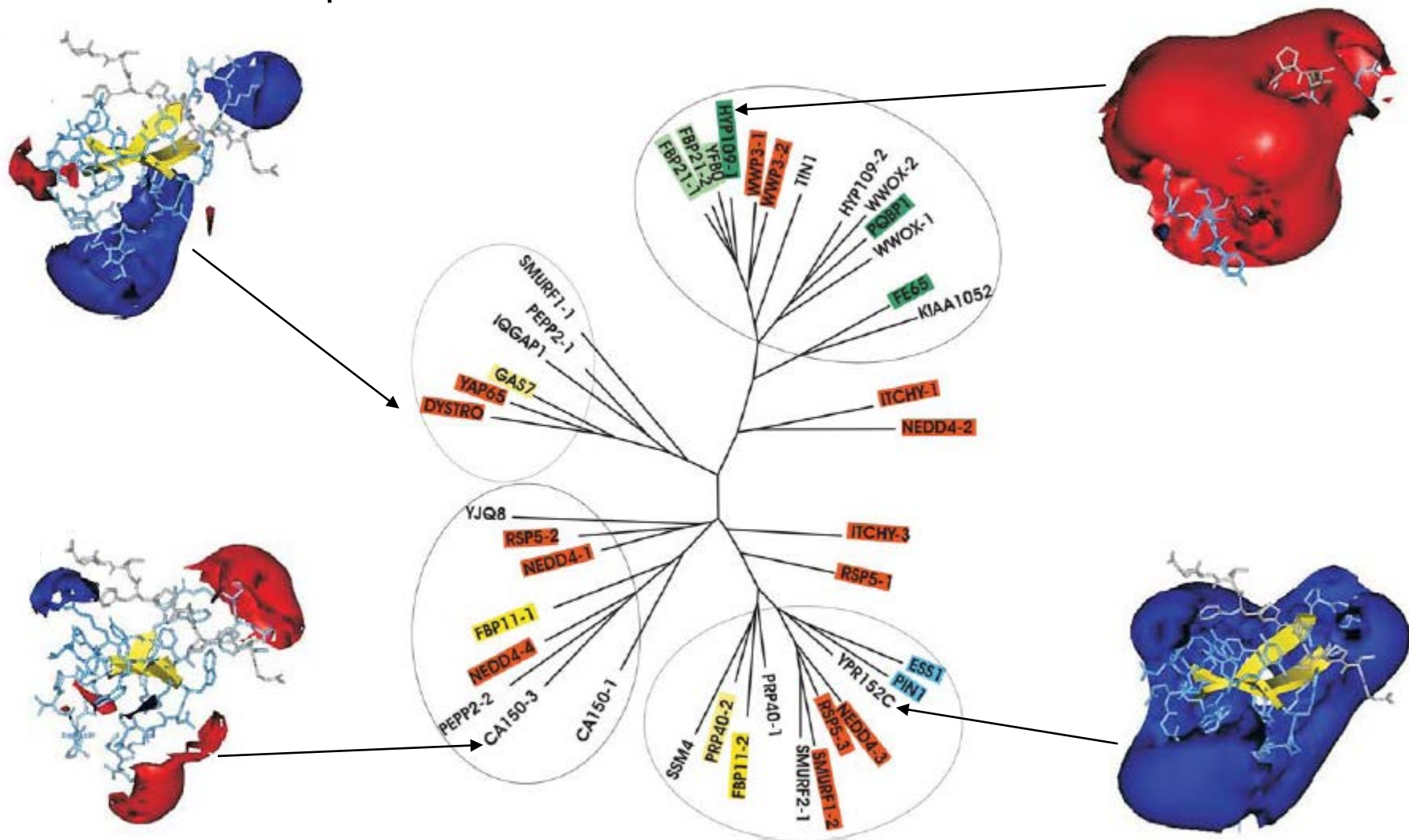
WW domain peptide  
binding preferences:

- xPPx(Y/poY)
- (p/Φ)P(p,g)PpR
- (p/Φ)PPRgpPp
- PPLPp
- (p/Ψ)PPPPP
- (poS/poT)P

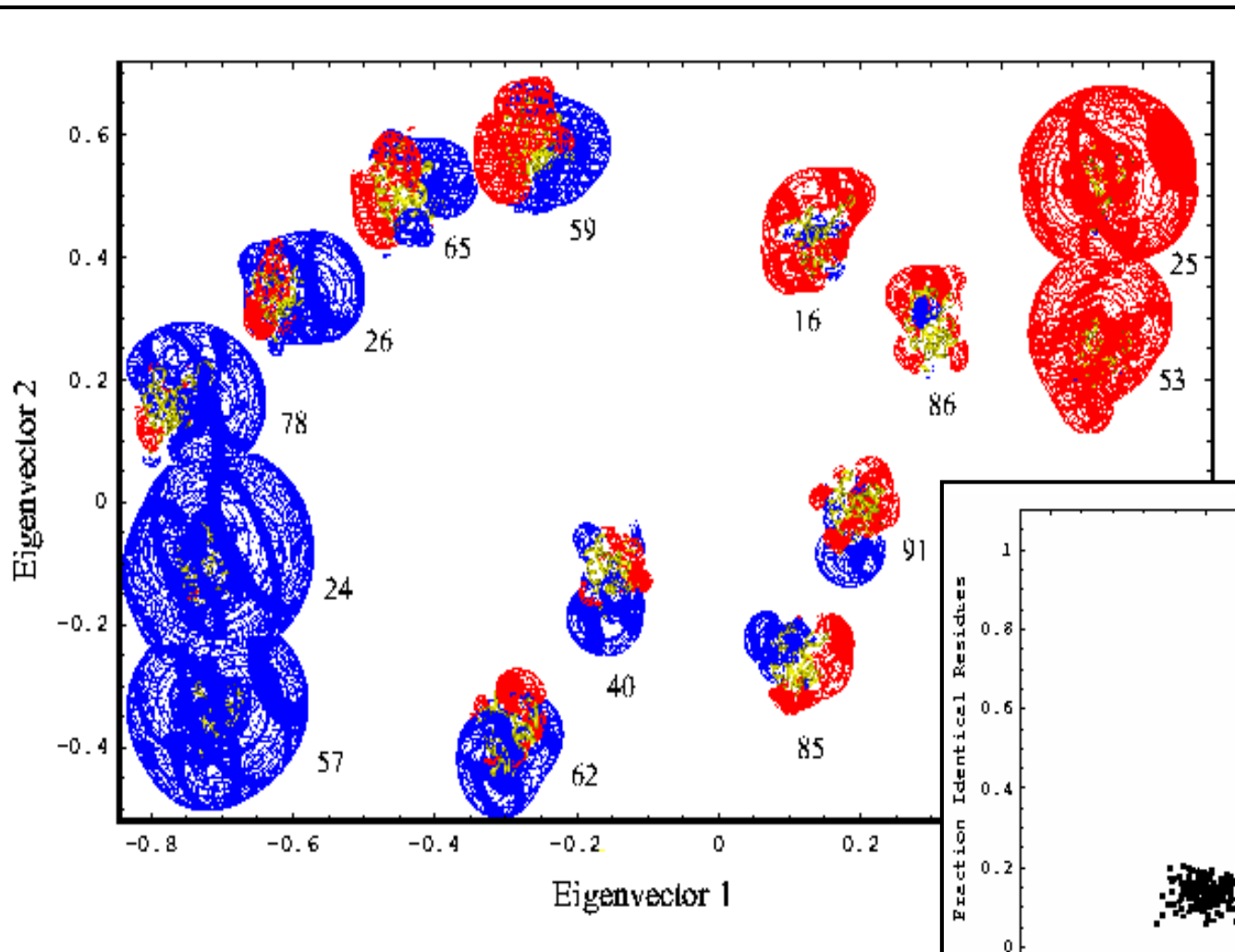
Otte et al. (2003) *Protein Sci.* 12, 491

# 42 WW Domains: PIPSA epogram for Molecular electrostatic potential

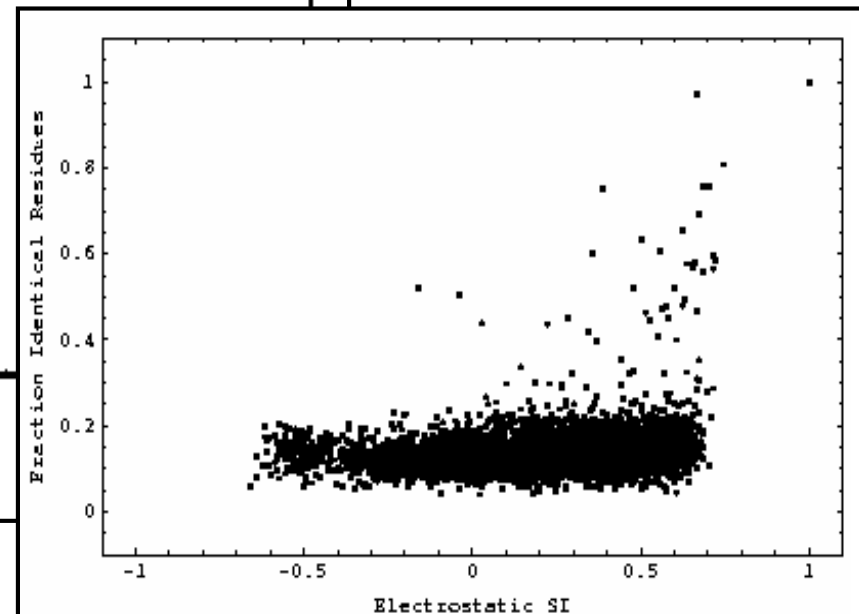
isopotential contours: -0.4 / +0.4 kcal/mol/e



# PH domains - distribution in electrostatic potential similarity space



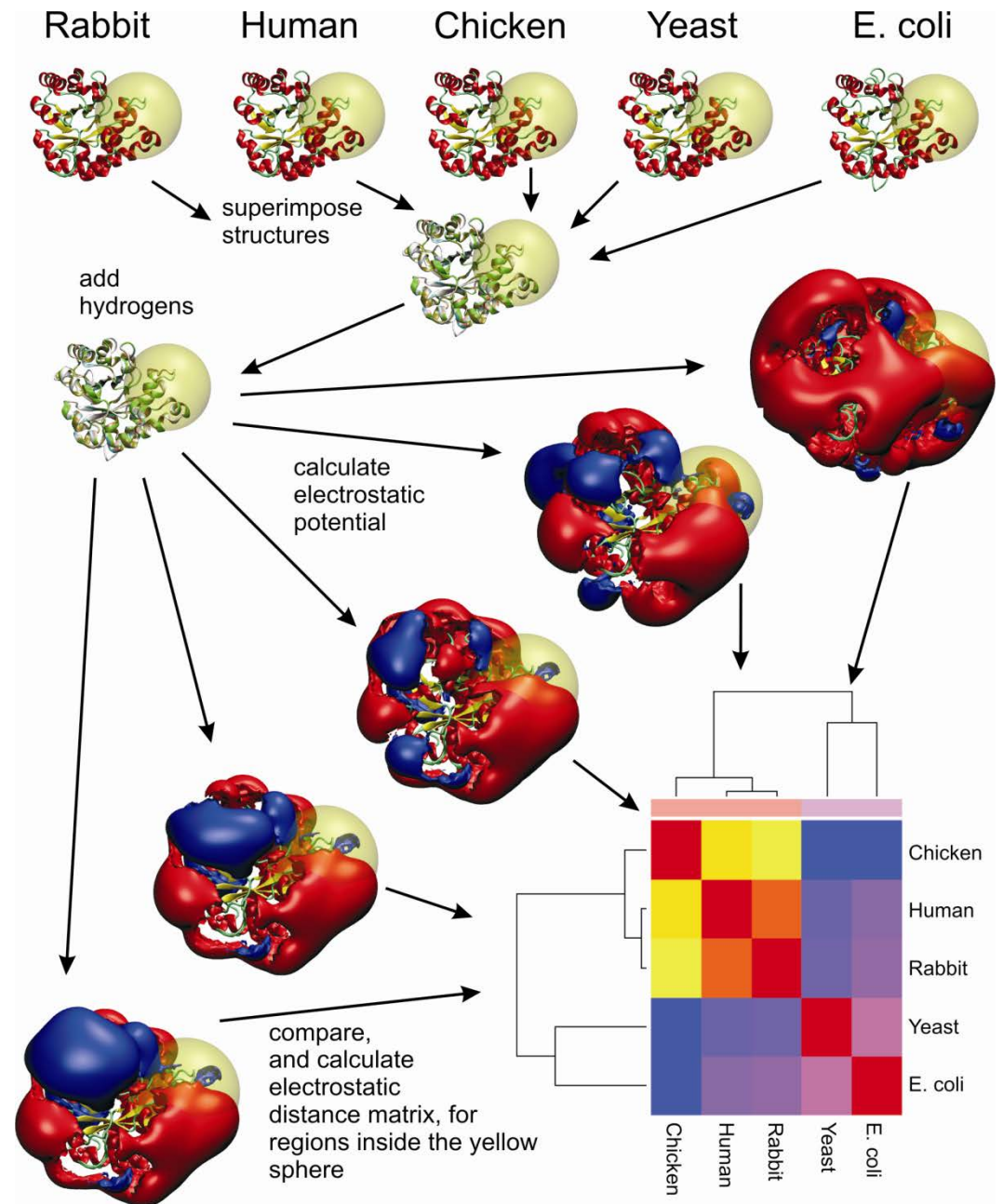
Distribution for DH-linked and internal PH repeat domains



# webPIPSA

webPIPSA:  
pipsa.h-its.org

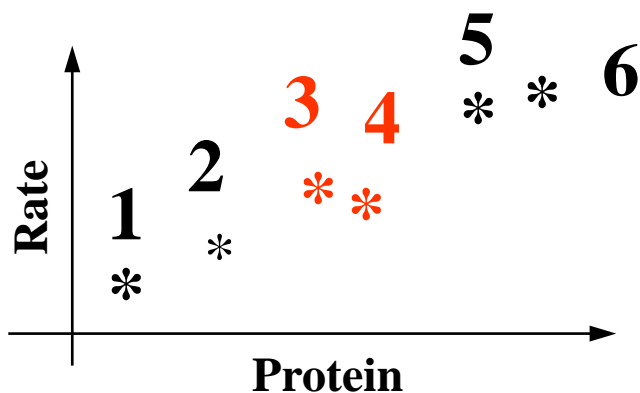
triosephosphate  
isomerase



Richter et al.,  
*Nucleic Acids Research*, 2008

# quantitative PIPSA (qPIPSA)

- Compare Molecular Interaction Fields
- Quantify similarities and differences
- Training set required with experimental information
- Predict relative ordering and trends

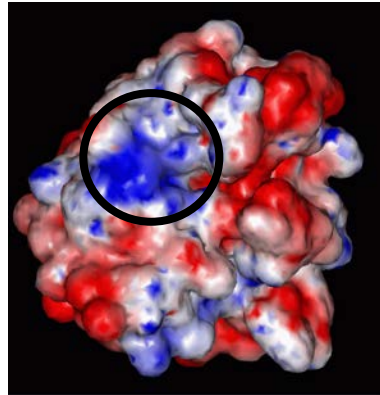
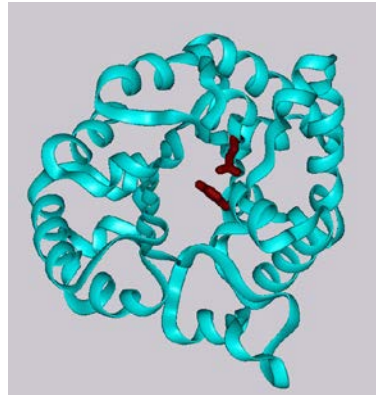
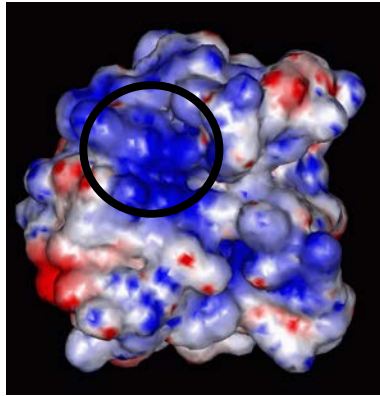


4



# Triose Phosphate Isomerase

*T. brucei*



*V. marinus*

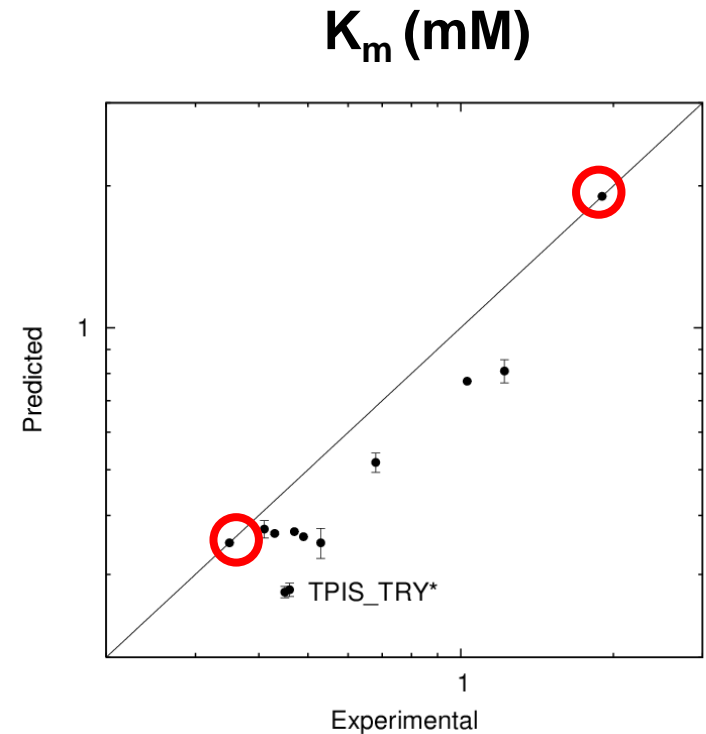
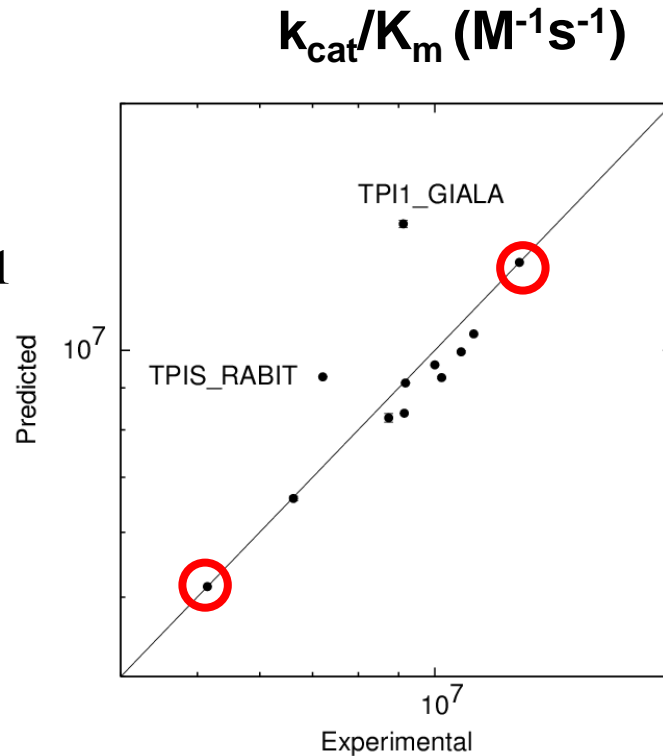
- 40/55% sequence identity/homology
- same fold
- very similar active site
- factor of 3 difference in  $k_{cat}/K_m$

## 12 species:

Giardia lamblia  
Spinach  
Chicken  
E. coli  
Human  
L. mexicana  
P. falciparum  
Rabbit  
T. brucei  
T. cruzi  
V. marinus  
Yeast

# Triose Phosphate Isomerase

$$\ln(k_a / k_b) \sim \alpha \cdot \sum_{\mathbf{R}} (\Phi_a - \Phi_b) / \sum_{\mathbf{R}} 1$$



**Predictions for 10 TPISs for the substrate glyceraldehyde-3-phosphate based on experimental measurements for the two TPISs from *V. marinus* (TPIS\_VIBMA) and *P. falciparum* (TPIS\_PLAFA)**

1 In unit increase is related to ca. 1.59

1 In unit decrease is related to ca. 0.85

kcal/mol/e increase of av. elec. pot.

kcal/mol/e increase of av. elec. pot.

*Gabdoulline, Stein, Wade, (2007) BMC Bioinformatics 8, 373*



# SYCAMORE: Systems biology's Computational Analysis and Modeling Research Environment sycamore.h-its.org

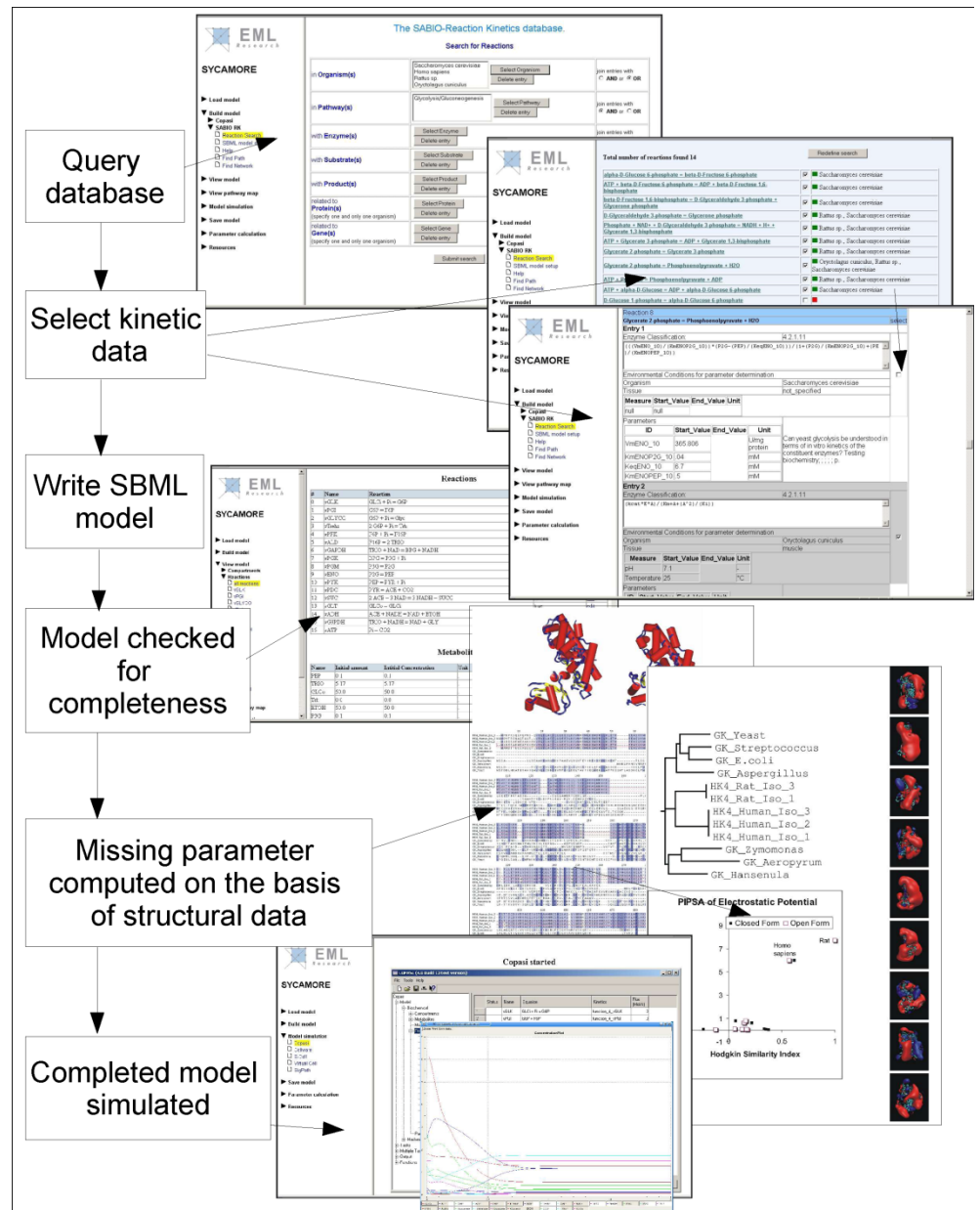


Fig 1: Example of using the available features in SYCAMORE. The case study is modeling and simulating glycolysis in hepatocytes. First, the database is queried and the relevant kinetic data selected. Then the SBML model file is created. This is checked for completeness (this is not implemented for automatic use yet). A missing parameter (here we assume Km for glucokinase to be missing) is then computed using structural data. Finally, the completed model is simulated.

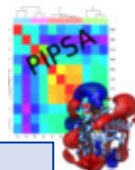
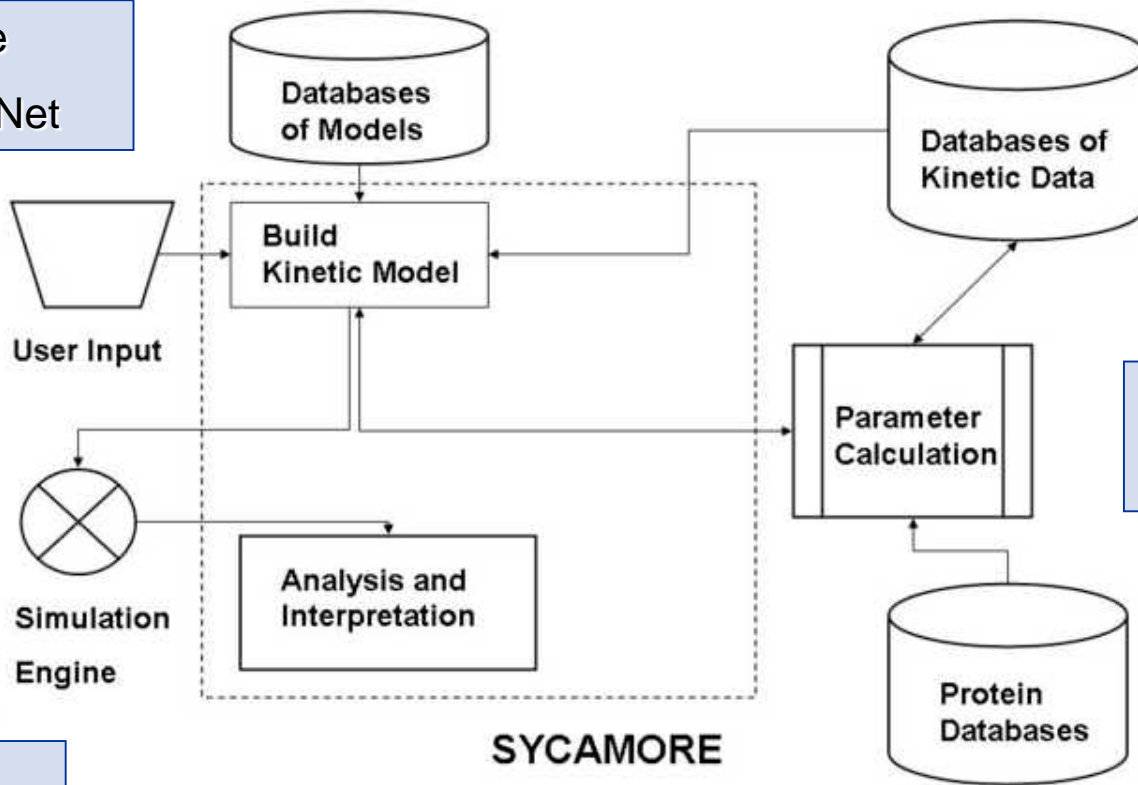


# SYCAMORE –Systems Biology’s Computational and Analysis Modelling Research Environment



JWS Online  
Biomodels.Net

BRENDA  
SABIO-RK



qPIPSA  
etc.

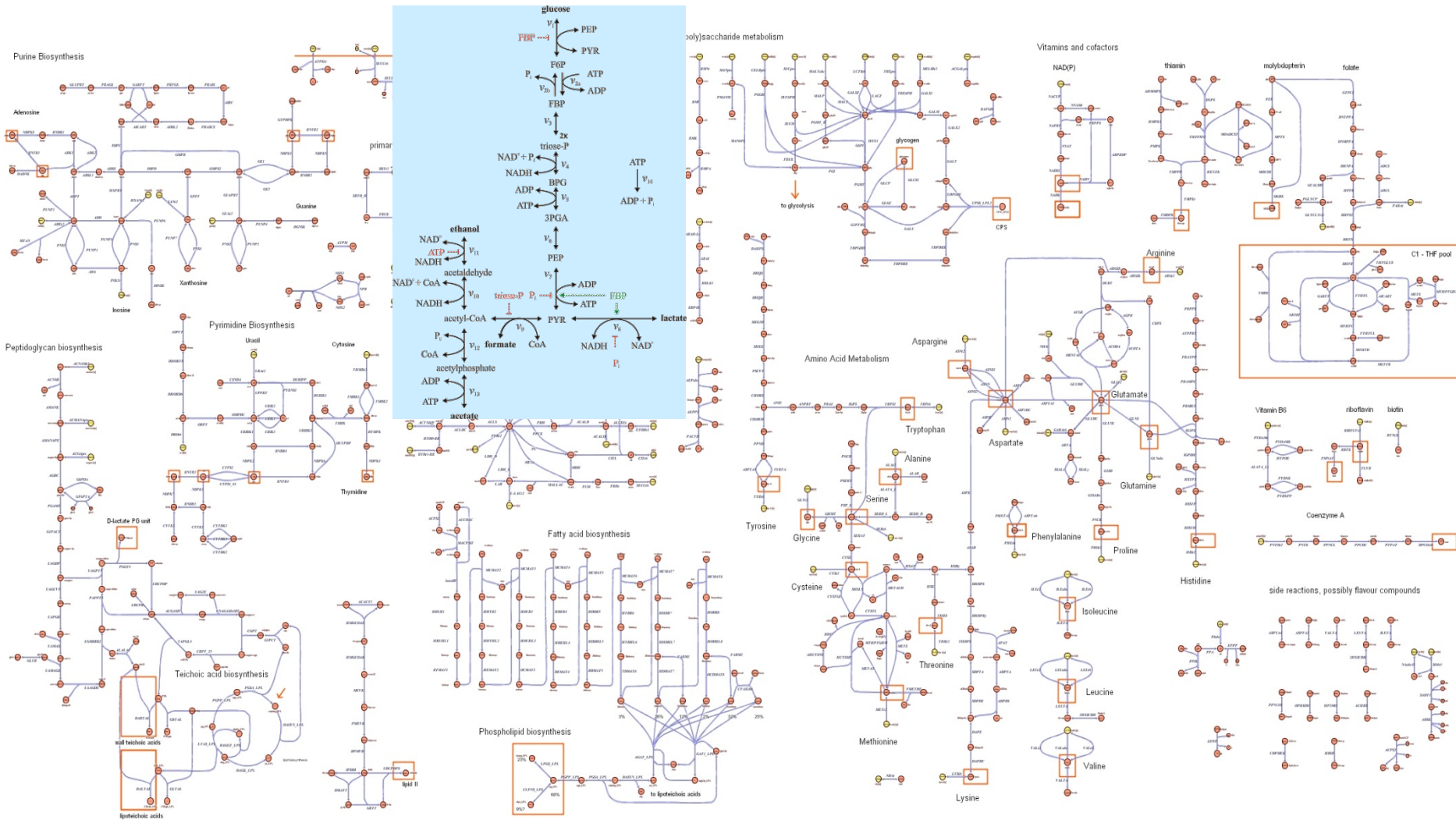


COPASI  
etc.

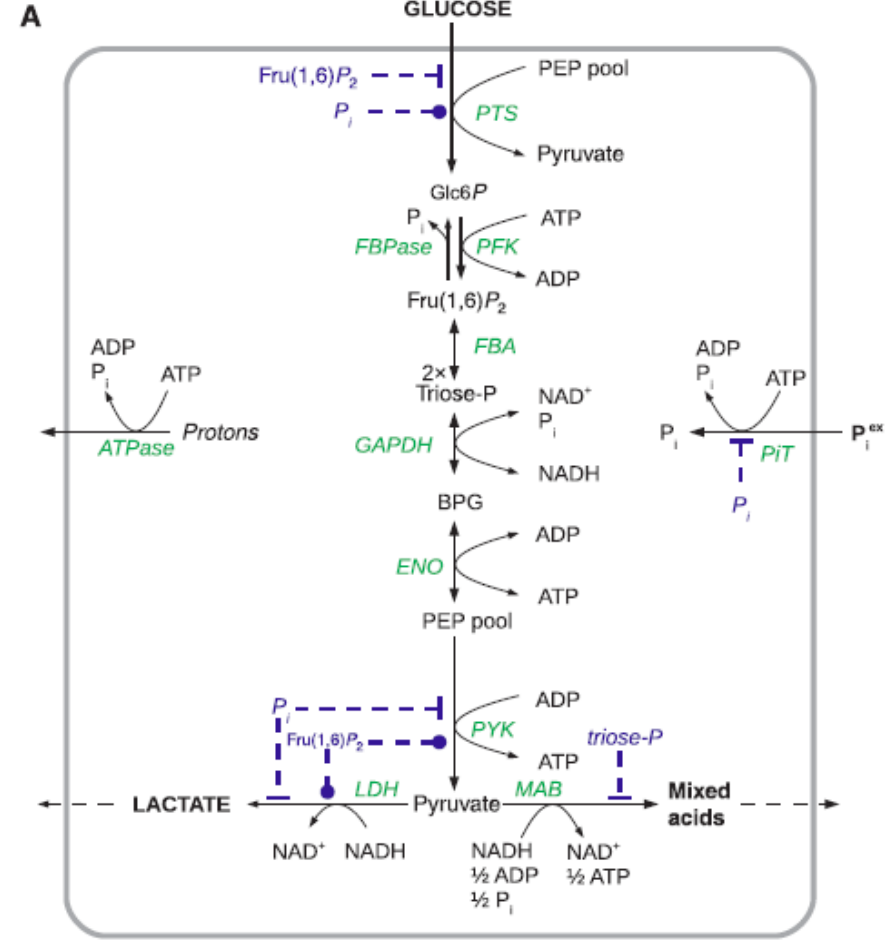
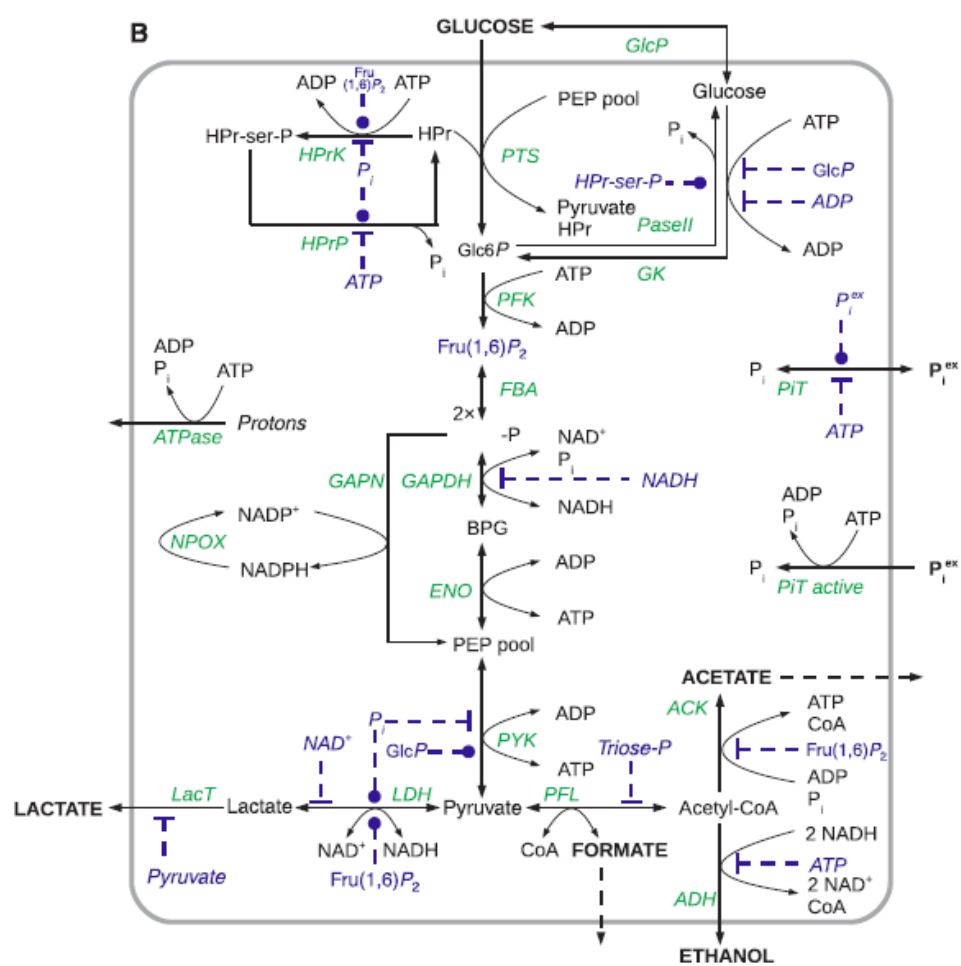
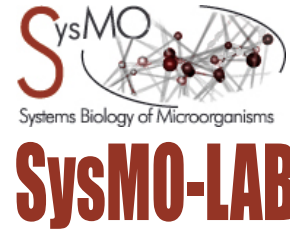
PDB  
MODBASE  
SwissModel

<http://sycamore.h-its.org>

# From protein structures to biochemical networks: regulation & cross-talk



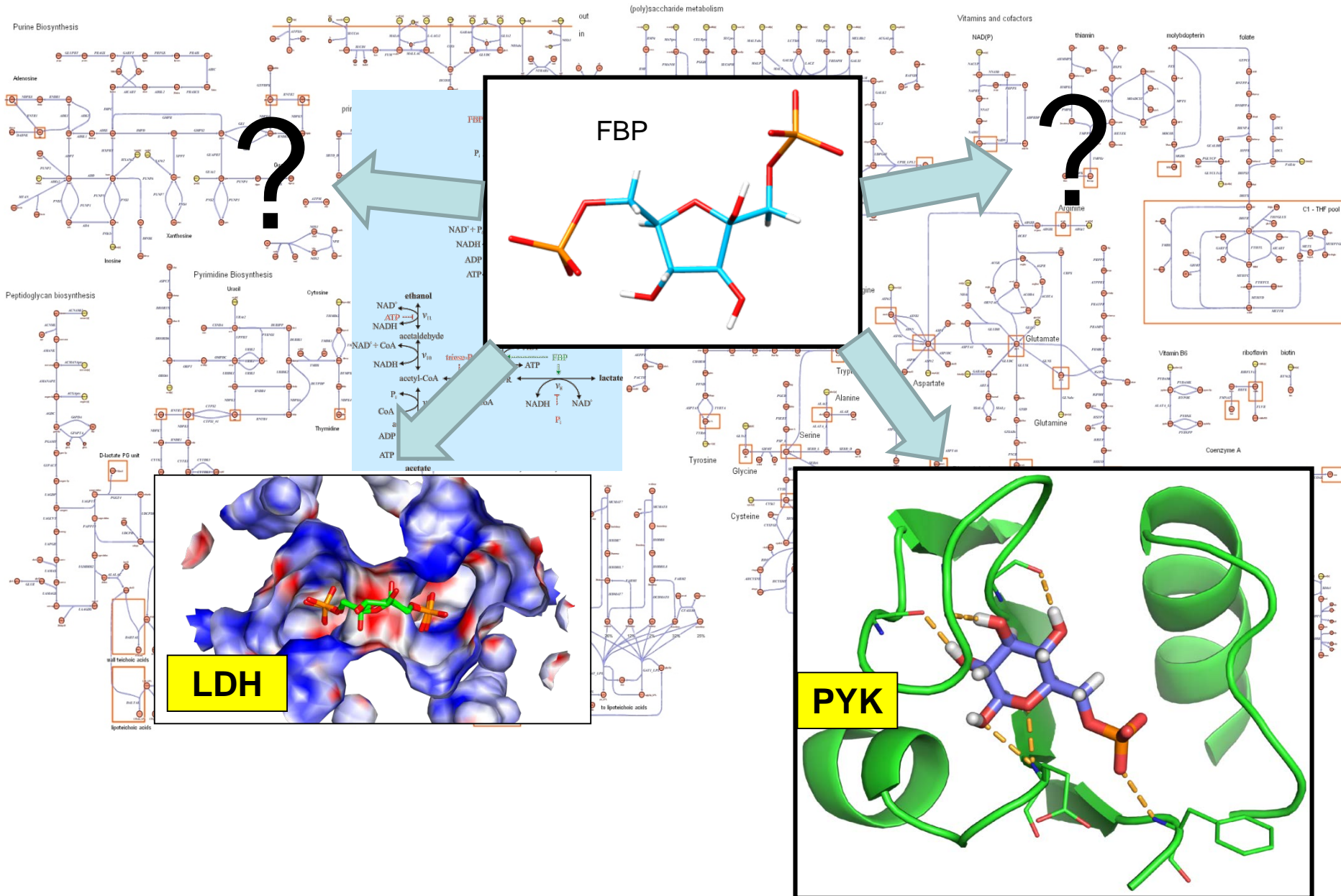
# Central metabolism of 2 lactic acid bacteria: regulation by phosphate



*S. pyogenes* central metabolism model

*L. Lactis* central metabolism kinetic model, Levering et al FEBS J. (2012) 279, 1274

# Finding cross-talk between reactions



# LigDig: a web server to answer ligand-based queries

Heidelberg Institute for  
Theoretical Studies



## This is LigDig

a web application for investigating ligand-protein interactions. LigDig can be used to query structural and functional properties. [Learn more](#)

Home

### SEARCH TOOLS

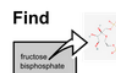
- [Find compound by name](#)
- [Find inhibitors](#)
- [Find ligand function](#)
- [Batch search for ligand ID](#)

### STRUCTURE TOOLS

- [Find protein structures](#)
- [Superpose ligand binding sites](#)
- [Structure preparation](#)


### LIGDIG INFO

- [Session info](#)
- [Workflow schema](#)
- [Links](#)
- [Use cases](#)

**Find**  
  
compound  
by name


[More info»](#)

[Find Compound](#)

**Find**  
  
Inhibitor


[More info»](#)

[Find an inhibitor](#)

**Ligand**  
  
functional  
annotation

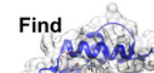
[More info»](#)

[Find Ligand Function](#)

**Batch**  
Search  


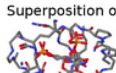
[More info»](#)

[Batch Search For Ligands](#)

**Find**  
  
Protein  
Structures

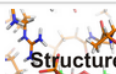
[More info»](#)

[Find protein structures](#)

**Superposition of**  
  
ligand  
binding sites


[More info»](#)

[Superpose ligand binding sites](#)

**Structure**  
Preparation  


[More info»](#)

[Structure Preparation](#)

**Review**  
  
Searches

[More info»](#)

[Session Info](#)

<http://mcm.h-its.org/ligdig>

Fuller et al, *Bioinformatics*, 2015, 31, 1147-49



# LigDig: Compound name disambiguation

## Find compound by name Search PubChem

You searched for: **fructose 1,6-bisphosphate**

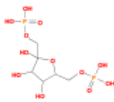

9 possible compounds were found (listed below).

For each compound found, the last column indicates the number of protein 3D structures that are available in the PDB.

Validate the corresponding checkbox(es) and click on the Submit to SearchPDB button at the end of the page. This will display all PDB files that contain your chosen ligand.

If no checkboxes are selected, or if you select only entries with no protein structures, you will also remain on this page when you click the Submit to SearchPDB button.

Submit to SearchPDB

CID	First Synonym	2D Structure	Additional compound information	# of PDB Structures	Select
172313	D-Fructose 1,6-bisphosphate <a href="#">Find ligand function</a>		More details for <b>D-Fructose 1,6-bisphosphate</b> <a href="#">Find binding partners for CHEMBL1089962 in ChEMBL</a>	40	<input type="checkbox"/>
10267	fructose-1,6-diphosphate <a href="#">Find ligand function</a>		<a href="#">Find binding partners for CHEMBL97893 in ChEMBL</a>		
445557	CHEMBL1089962 <a href="#">Find ligand function</a>		<a href="#">More info»</a>		
718	1,6-di-o-phosphonohex-2-ulofuranose <a href="#">Find ligand function</a>				
2734398	NCGC00166321-01 <a href="#">Find ligand function</a>		More details for <b>NCGC00166321-01</b> <a href="#">Find binding partners for CHEMBL2146112 in ChEMBL</a>	0	<input type="checkbox"/>
16219367	F6803_SIGMA				

RESTful →  
PubChem

# LigDig: Example application to kinase inhibitors

The screenshot illustrates the LigDig workflow for identifying kinase inhibitors. It starts with a search for the protein ERK1, leading to a network graph where nodes represent proteins and edges represent interactions. A specific node, MAP kinase ERK1, is highlighted, and a network of associated ligands is shown. One ligand, CHEMBL475251, is selected and shown in a 3D molecular model superposed with reference binding sites and ligands.

**Search Results:**

- Protein Uniprot ID/Name: erk1
- Ligand binding affinity (nM):
  - P27361: Mitogen-activated protein kinase; ERK1/ERK2 (Homo sapiens) [2027]
  - P28482: Mitogen-activated protein kinase; ERK1/ERK2 (Homo sapiens) [17021]
- Off-target binding affinity (nM):
  - P63085: Mitogen-activated protein kinase 1 and 3 (ERK2 and ERK1) (Mus musculus) [11]
  - P63086: Mitogen-activated protein kinase 1 and 3 (ERK2 and ERK1) (Rattus norvegicus) [29]
  - Q63844: Mitogen-activated protein kinase 1 and 3 (ERK2 and ERK1) (Mus musculus) [0]
  - P21708: Mitogen-activated protein kinase 1 and 3 (ERK2 and ERK1) (Rattus norvegicus) [0]

**Network Graph:**

- Central node: MAP kinase ERK1
- Other nodes: CHEMBL521562, CHEMBL525191, CHEMBL475251

**3D Molecular Model:**

- Reference binding site: 3piy A 585 1 1rjb A
- Reference ligand: 3piy A 585 1 Ligand
- Query binding sites: 1rjb A Binding Site, 2zoq B 5ID 382 Binding Site
- Query ligands: 2zoq B 5ID 382 Ligand

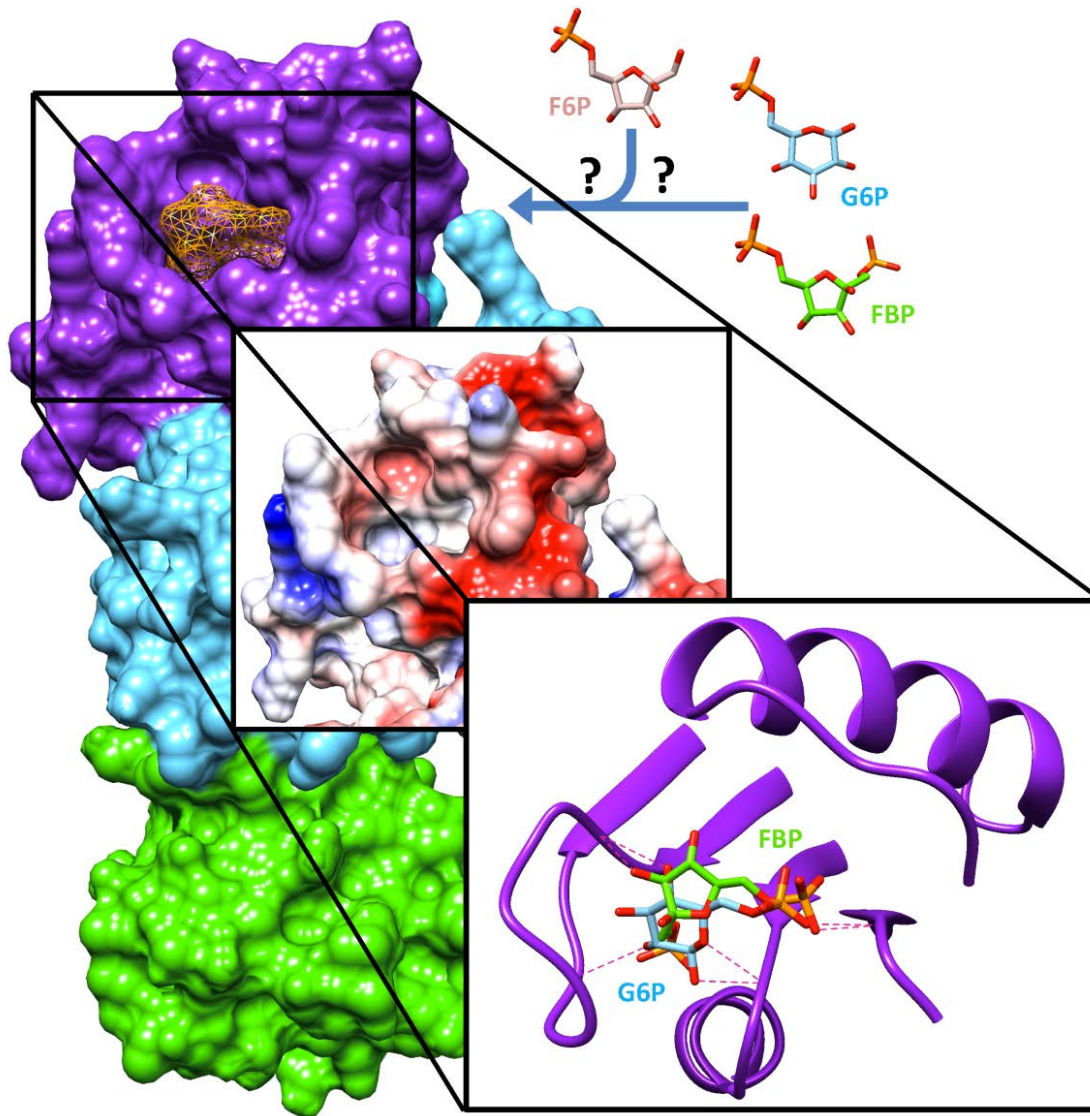
**Chemical Information:**

- Tyrosine-protein kinase receptor FLT3:** P36888, CHEMBL1974, Off-target SINGLE PROTEIN
- CHEMBL475251:** Active, ZINC06745792, DB07159, KEGG identifier does not exist, 585
- Tyrosine-protein kinase receptor FLT3 interaction:** Kd: 0.71 nM, CHEMBL475251, CHEMBL475251, CHEMBL1974, Tyrosine-protein kinase receptor FLT3, Assay: CHEMBL988712

<http://mcm.h-its.org/ligdig>

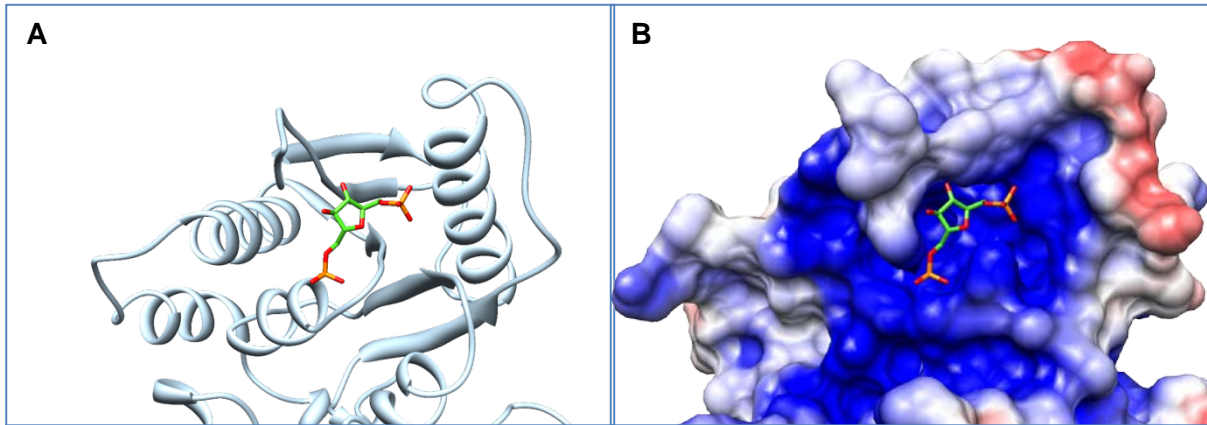
Fuller et al, *Bioinformatics*, 2015, 31, 1147-49

# Pyruvate kinase in lactic acid bacteria: Which are the activators?

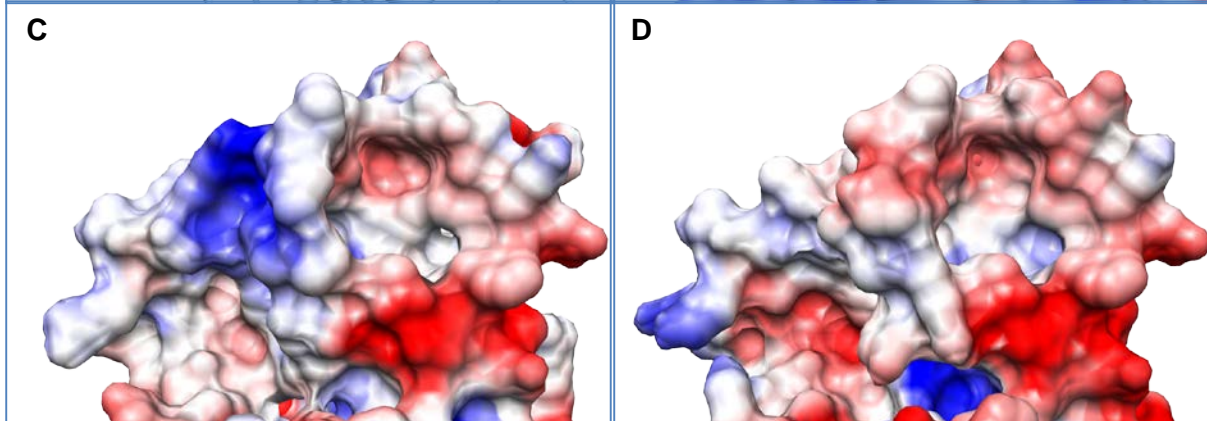


# Pyruvate kinase – allosteric site: different electrostatic potentials

**Crystal  
Structure of  
*S. cerevisiae*  
PYK with FBP  
bound**



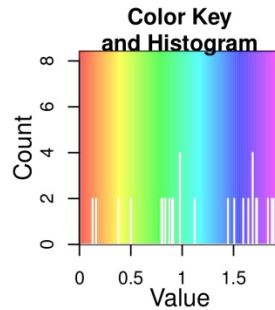
**Modelled  
Structure of  
*S. pyogenes*  
PYK with open  
allosteric site**



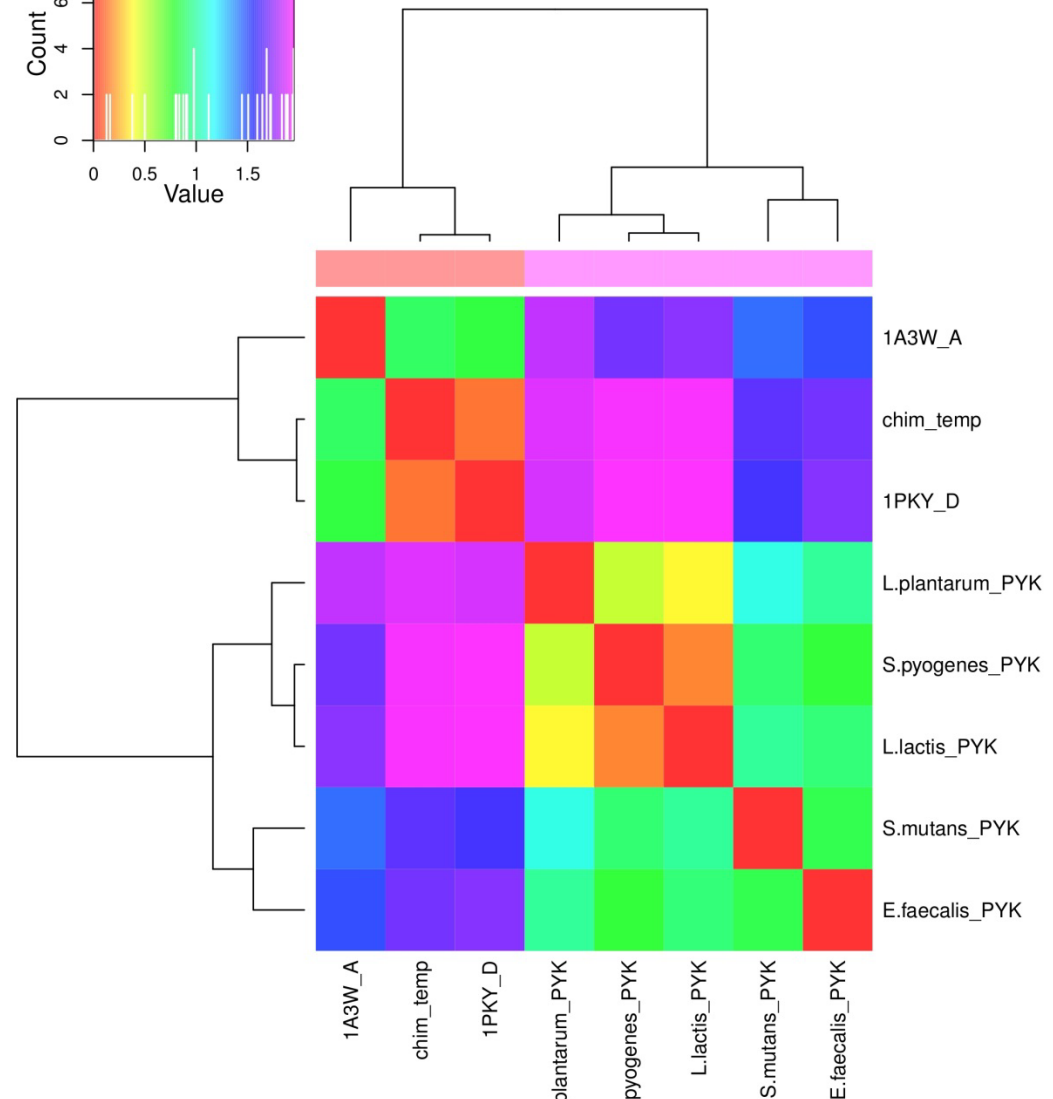
**Modelled  
Structure of  
*L. plantarum*  
PYK with open  
allosteric site**

Colored from -2 (red) to +2 (blue) kT/e

# Pyruvate kinase – allosteric site: electrostatic similarity



$$\text{Electrostatic Distance } D_{a,b} = \sqrt{2 - 2S_{a,b}}$$



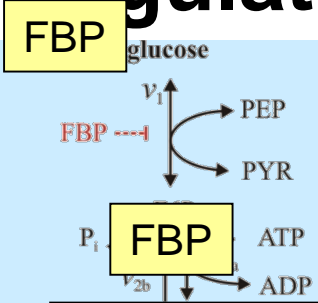
Cluster: yeast,  
chimeric template,  
E. coli

Cluster : L lactis, S. pyogenes,  
L. plantarum,

S. mutans, E. faecalis

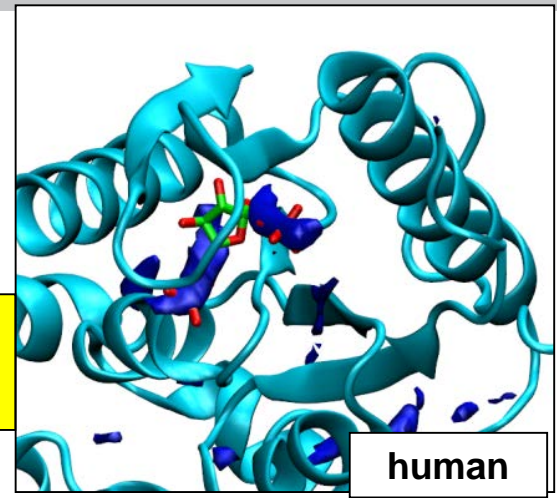
**PIPSA analysis**

# Allosteric regulation: PYK



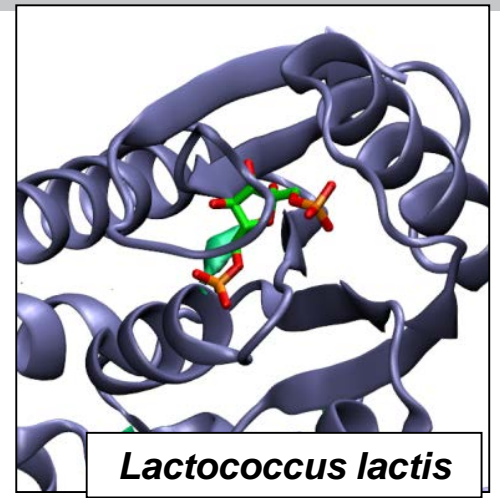
**FBP-regulation**

**pyruvate kinase (PYK)**



human

2 phosphate sites



*Lactococcus lactis*

1 phosphate sites

<i>E. coli</i>	FBP
<i>B. stearoth.</i>	R5P + AMP (G6P)
<i>S. mutans</i>	G6P (R5P)
<i>L. lactis</i>	G6P, R5P, F6P, FBP, Gal6P
<i>S. pyogenes</i>	G6P, Gal6P, R5P, F6P, FBP
<i>E. faecalis</i>	FBP
<i>L. plantarum</i>	FBP, G6P, Gal6P

**binding and allosteric regulation of phosphorylated sugars**

- Modifier used may depend on environment
- Different LABs adapted to different environments
- Some must survive in more environments

# Using protein structures to learn about protein function: Learning objectives

- Protein structure and function
- Modeling protein structure and dynamics
- Computing interaction properties

# Thank you for your attention!



Computational Biology: Genomes to  
Systems

EMBO PRACTICAL COURSE