Universidad de Málaga
Master en Biotecnología Avanzada
Bioinformática y Tratamiento de Datos
(BIF) 2018-2019

Bioinformática Estructural

Florencio Pazos (CNB-CSIC)

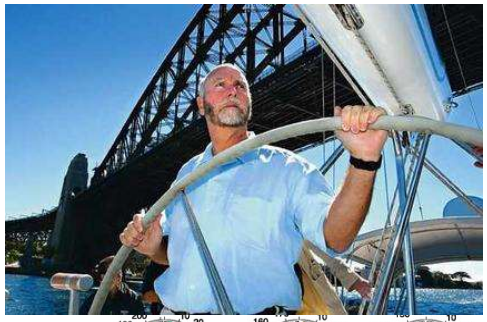
Florencio Pazos Cabaleiro
Computational Systems Biology Group (CNB-CSIC)
pazos@cnb.csic.es
http://csbg.cnb.csic.es

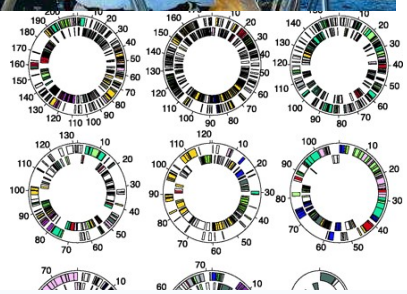
U. Málaga
November 2018

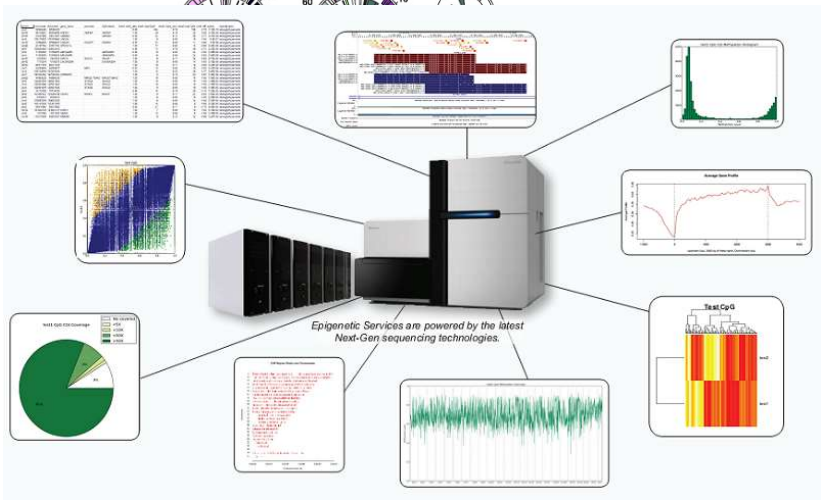
Structural Bioinformatics

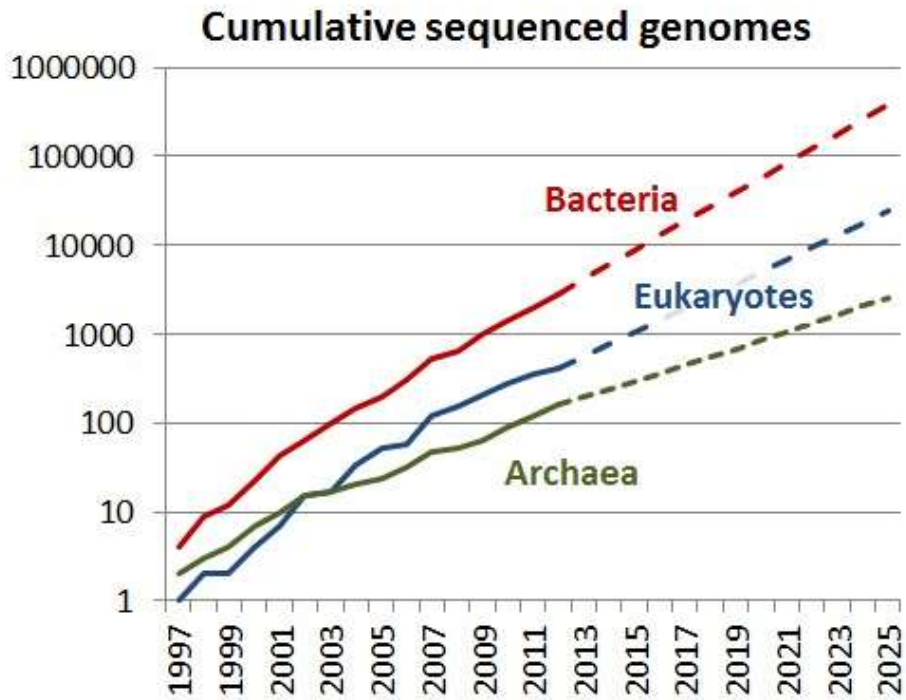
- Experimental knowledge on protein sequences and structures
- Combining experimental structure determination with prediction to cover the structural space
- Structure Visualization
- Structural alignments / Protein domains
- Characteristics of the space of structures. Relationship with that of sequences
- Protein homology
- Hierarchical classification of the protein universe
- Classification of protein structure prediction methods
- Prediction of 1D characteristics
    - Secondary structure and solvent accessibility
    - Transmembrane helices
    - Unstructured regions
- Prediction of 3D structure
    - Homology modeling
    - Threading
    - Combined and fragment-based approaches
- Model filtering
- Correlated mutations as distance constraints
- Assessment of prediction methods
- Bibliography

Obtaining protein sequences









August 2015: GenBank: **187 million** seqs (**DNA**) from **500.000 diff organisms** -> **50 million** seqs translated into **proteins** (UniProt/TrEMBL)

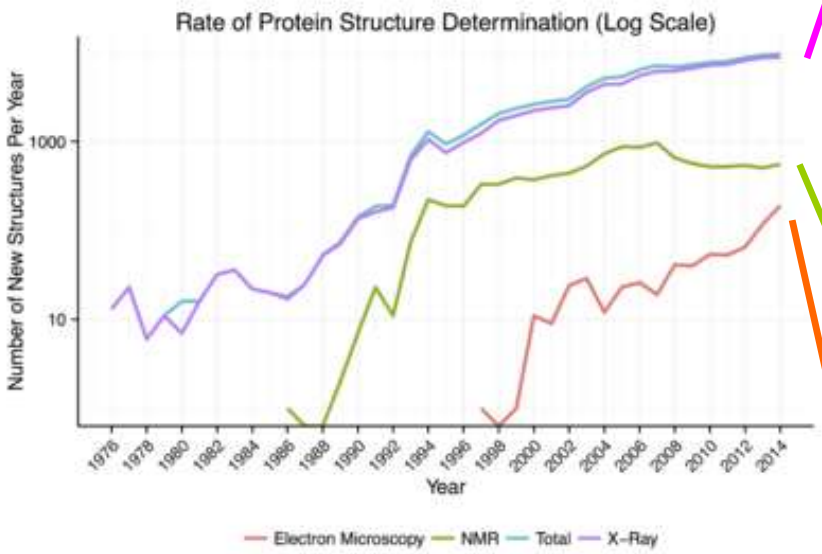
It is “easy” to obtain protein sequences

• van Dijk, E.L., Auger, H., Jaszczyszyn, Y. and Thermes, C. (2014) Ten years of next-generation sequencing technology., *Trends Genet*, **30**, 418-426.

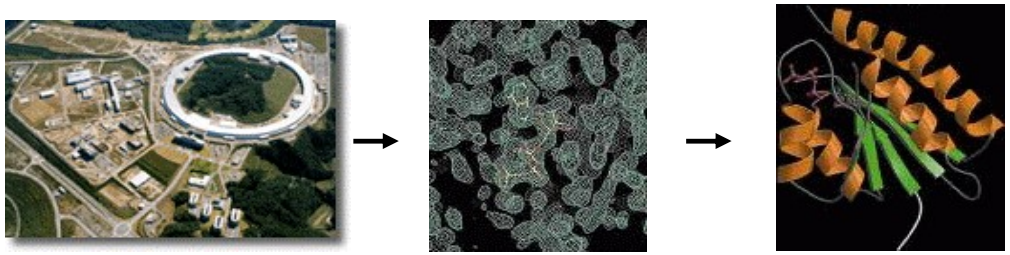
• Collins, F.S., Green, E.D., Guttmacher, A.E. & Guyer, M.S. (2003) A vision for the future of genomic research. *Nature*, **422**, 835-847.

• Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. (2004). *Science* **304,** 66-74.

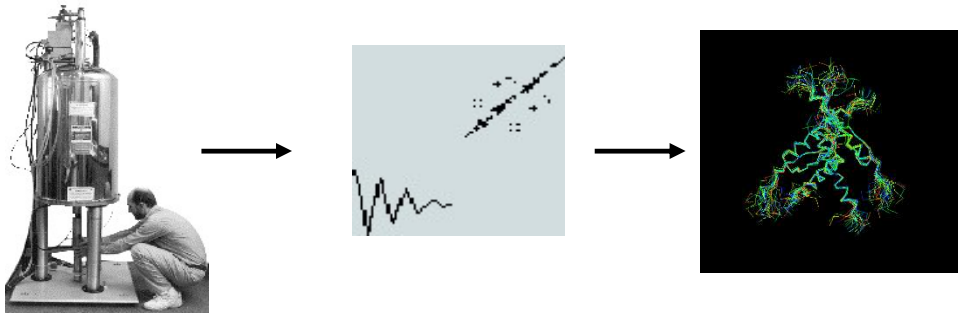
Experimental determination of protein 3D structures



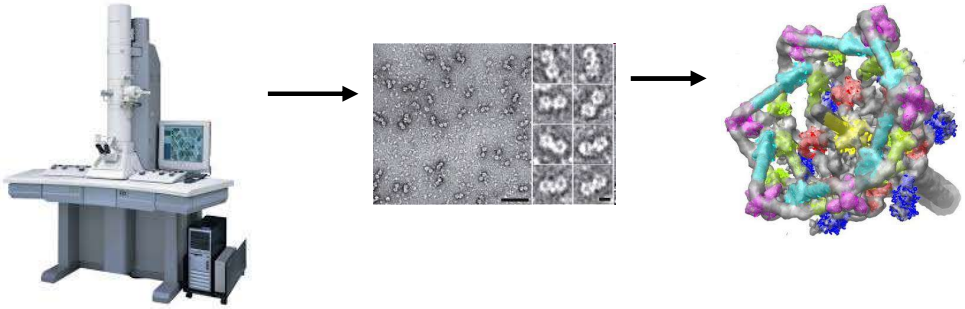
X-ray



NMR



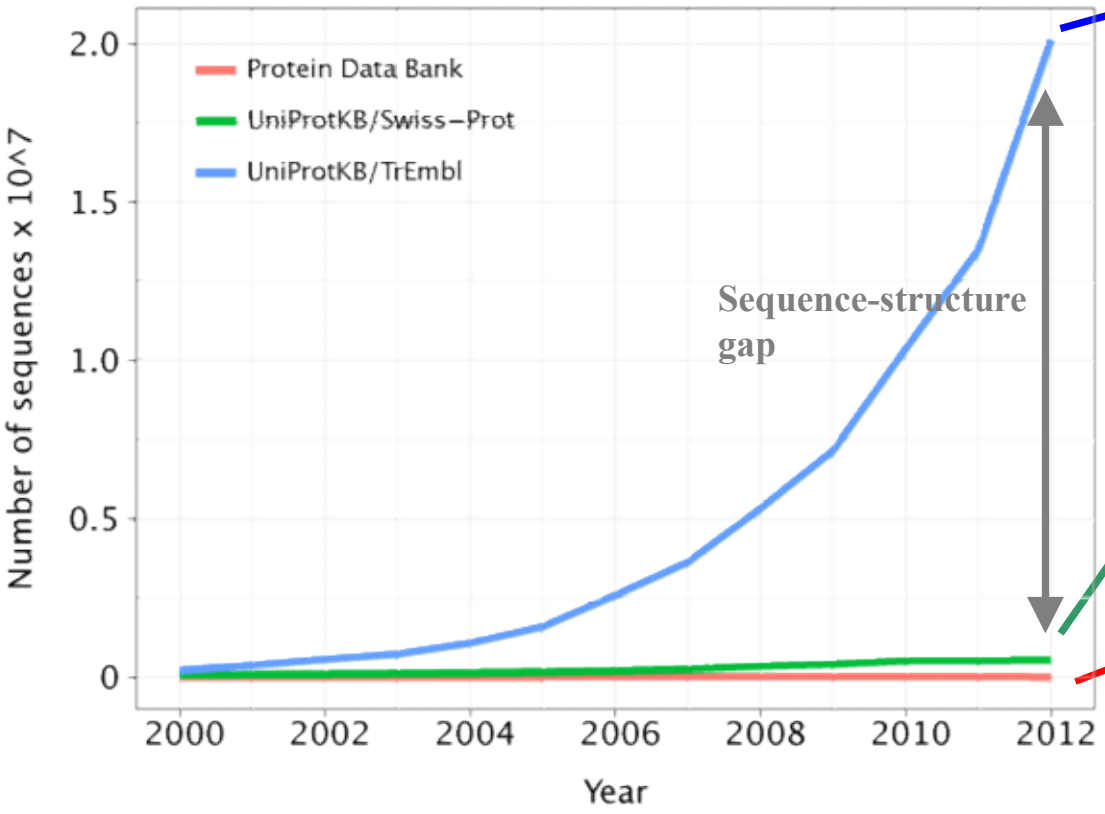
Electron microscopy



August 2015: **100.000 protein structures** experimentally solved (PDB)
=> **0.2%** of known protein sequences

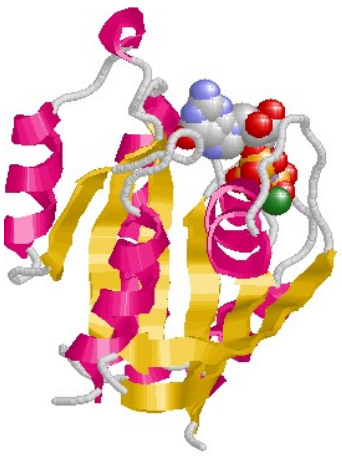
… and not so easy to obtain 3D structures. Methods keep on improving, although NMR is declining and e- microscopy increasingly used for HR.

Experimental knowledge on protein sequences, functions and 3D structures



MREYKLVVLGSGGVGKSALTVQFVQGIFVDEYDPTIEDSY
RKQVEVDCQQCMLEILDTAGTEQFTAMRDLYMKNGQGFA
VYSITAQSTFNDLQDLREQILRVKDTEDVPMILVGNKCDL
EDERVVGKEQGQNLARQWCNCAFLESSAKSKINVNEIFYD
LVRQINR

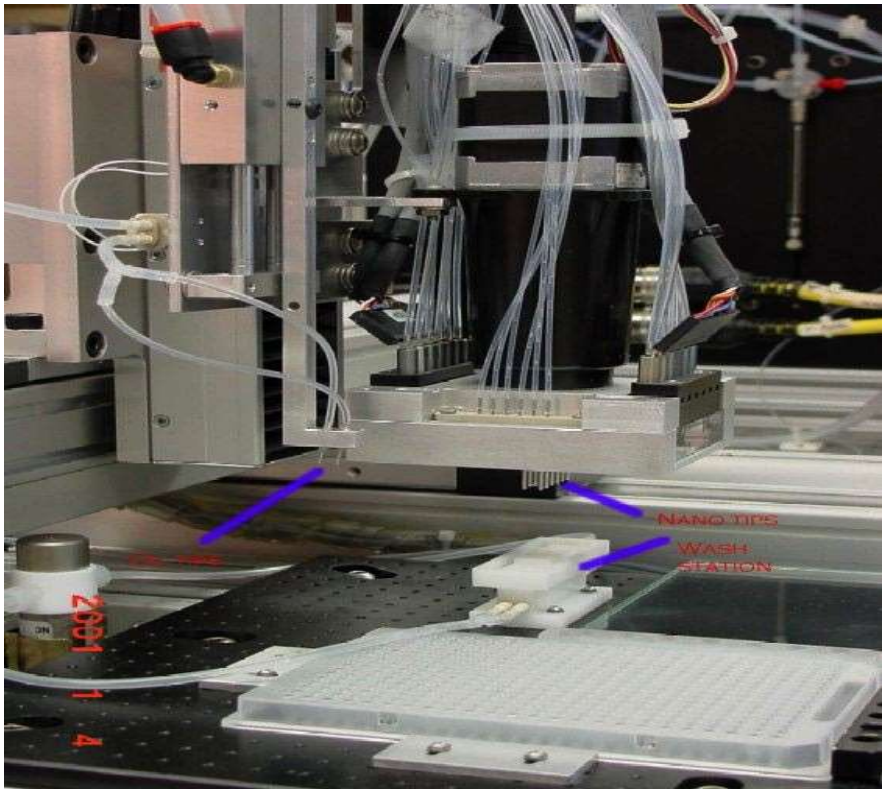
MLEILDTAGTEQFTAMRDLYMKNGQGFAL
VYSITAQSTFNDLQDLREQILRVKDTEDVPMIL
VGNKCDLEDERV

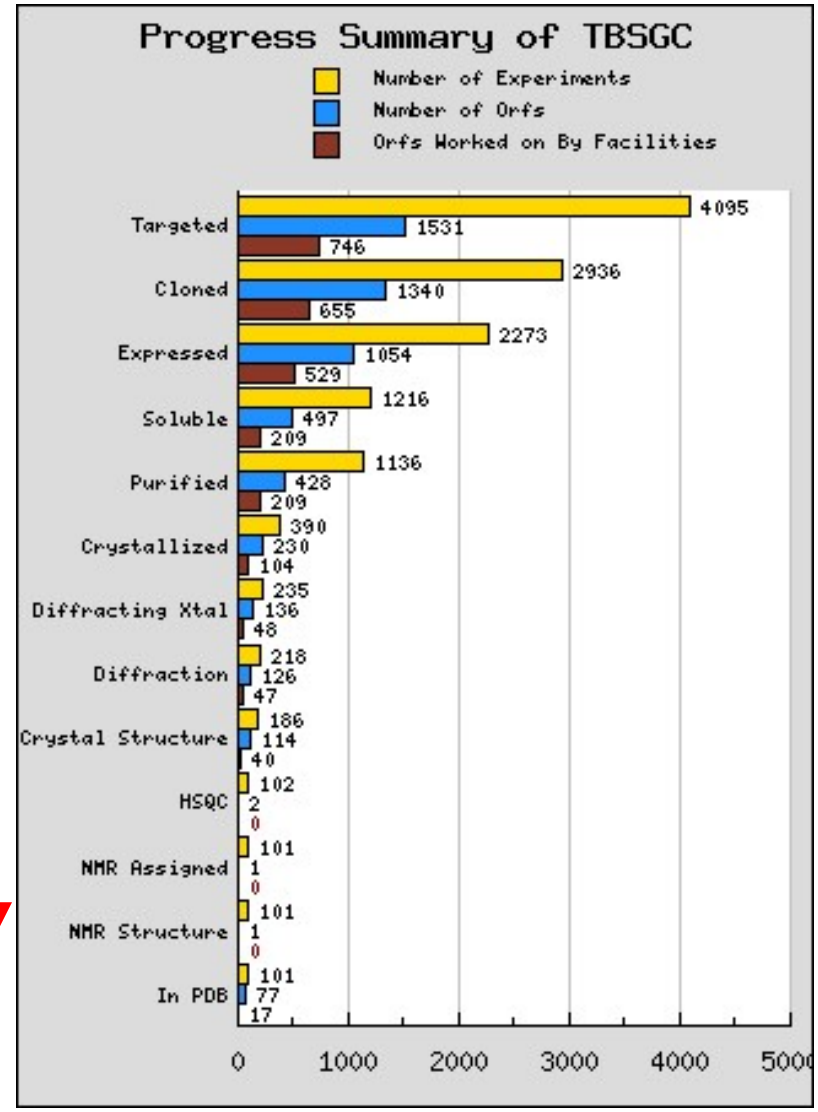


Consequently… The number of known 3D structures is orders of magnitude lower than the number of known sequences: “sequence/structure gap”

Structural Genomics

Aim: massive (high-throughput) determination of protein 3D structures

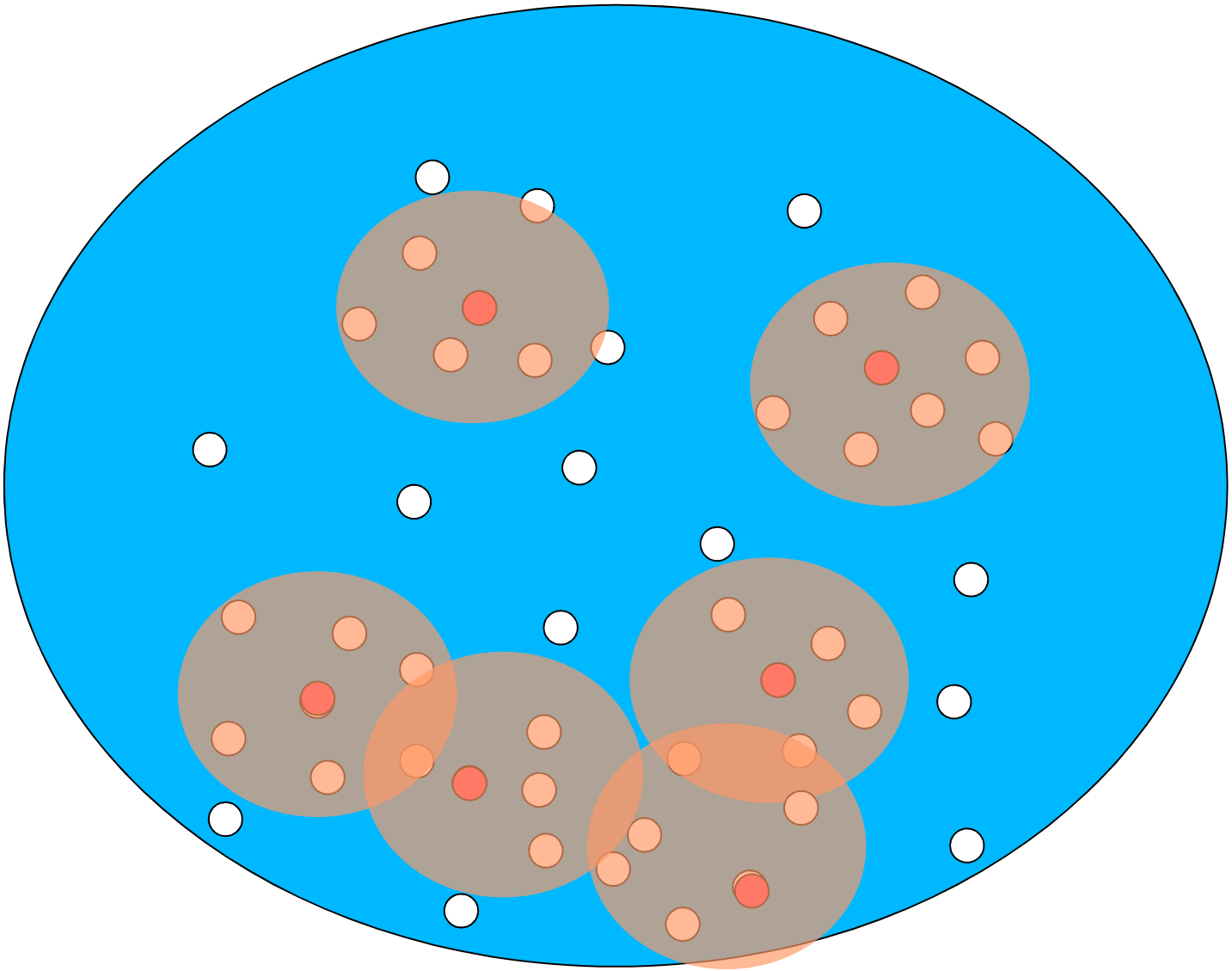




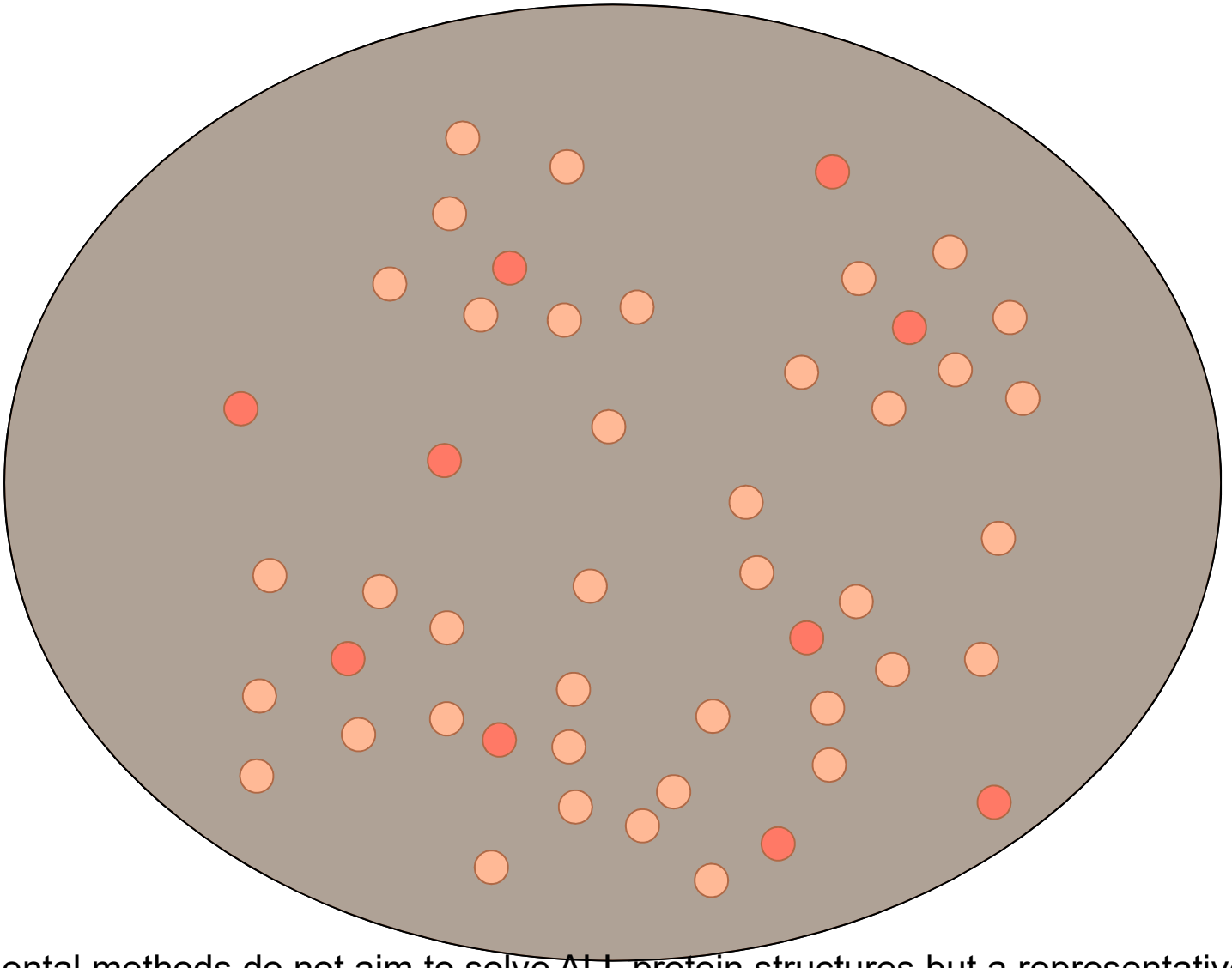
Structural genomics is not as “massive” as expected

Vitkup, D., Melamud, E., Moult, J. and Sander, C. (2001) Completeness in structural genomics. *Nat Struct Biol*, **8**, 559-566.

Combining experimental structure determination
with prediction to cover the protein structural space

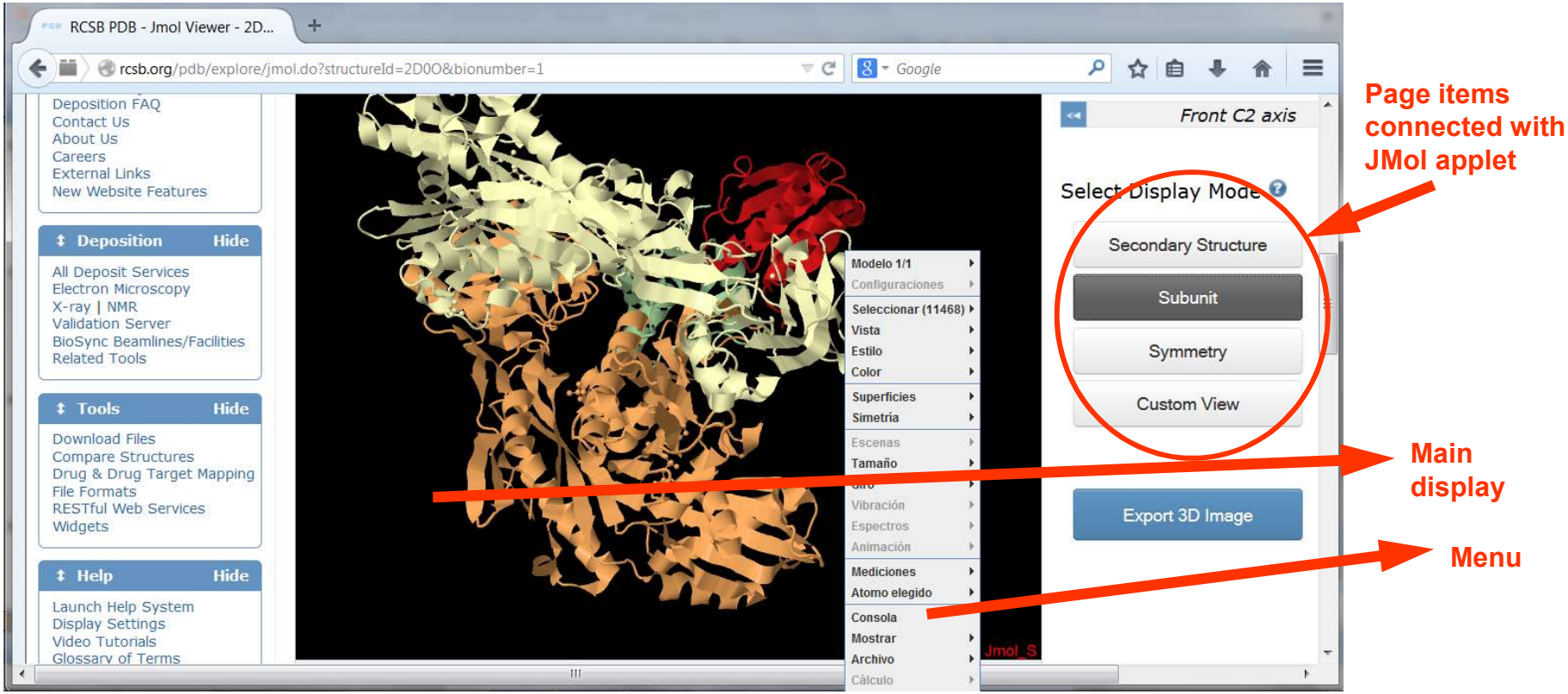


Combining experimental structure determination with prediction to cover the protein structural space



Experimental methods do not aim to solve ALL protein structures but a representative set so that the rest can be modeled (predicted) based on them.

Structure Visualization

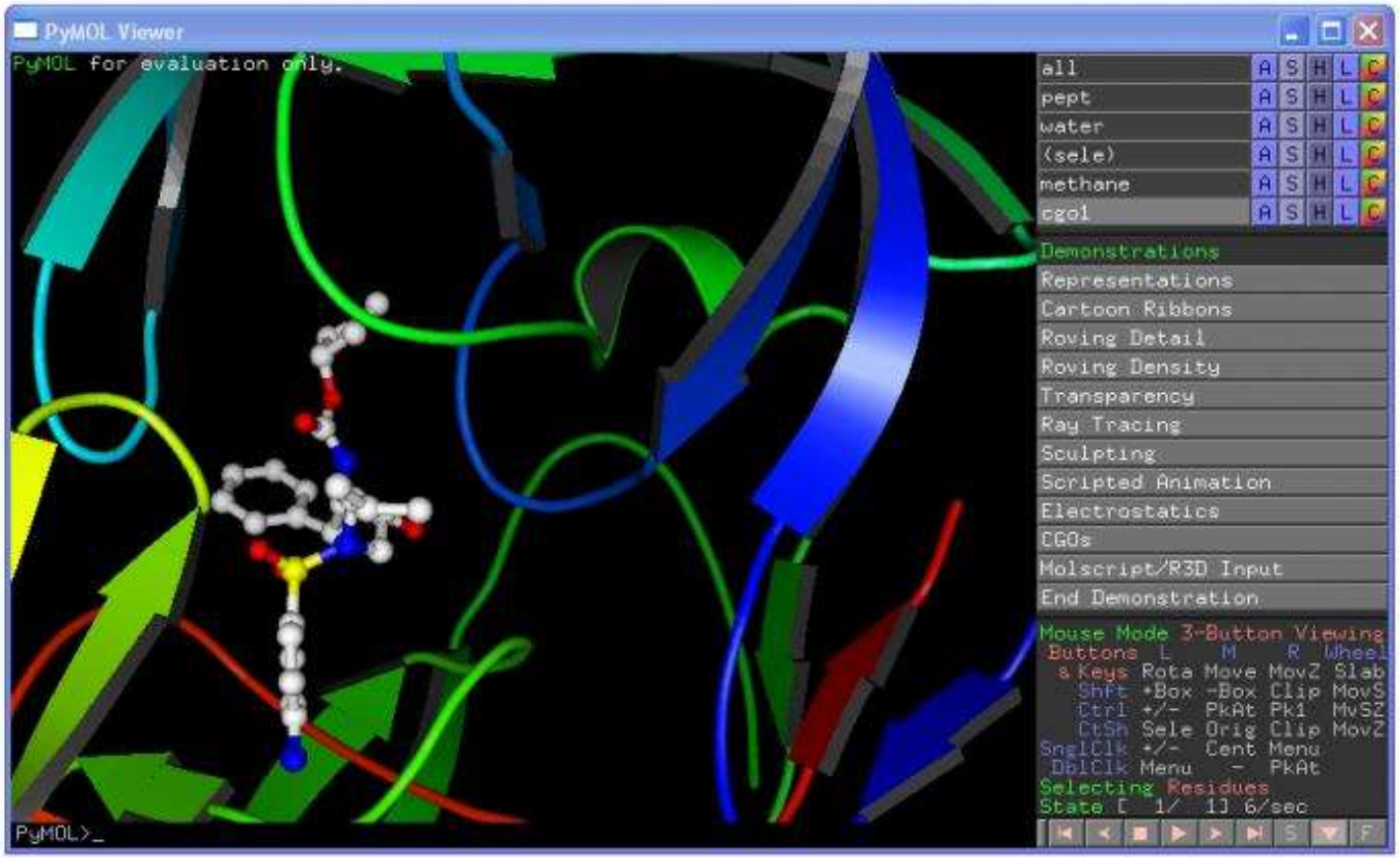


JMol (applet) /JSMol (javascript)

- Does not require installation (java applet (or JS) embedded in web pages) [Also available as standalone java program]
- Easy to customize and connect with page elements/controls
- Not many features: suitable for a quick view.

http://jmol.sourceforge.net/

Structure Visualization

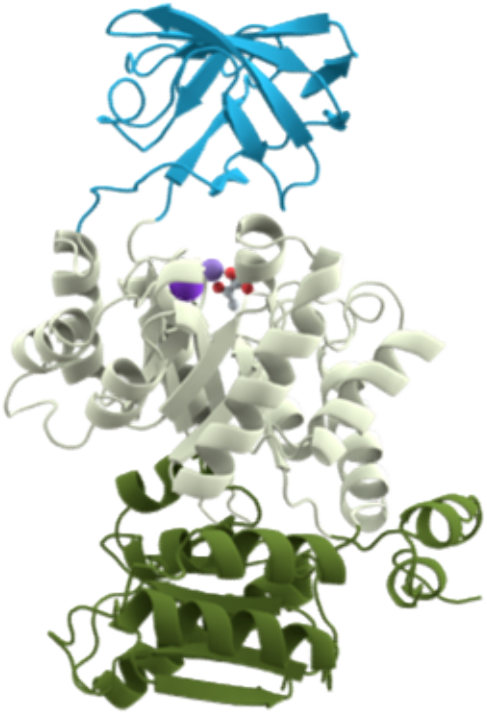


PyMol

- Local installation.
- Many features
- Easily expandable with modules/scripts in python.

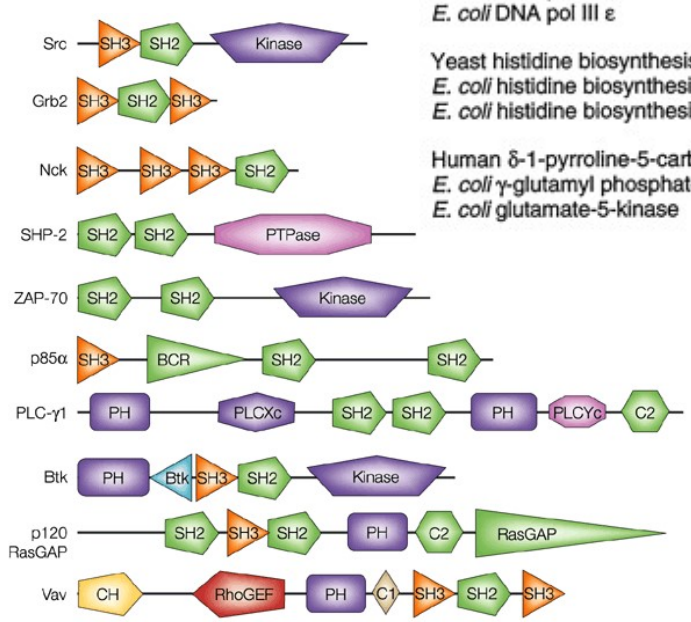
https://www.pymol.org/

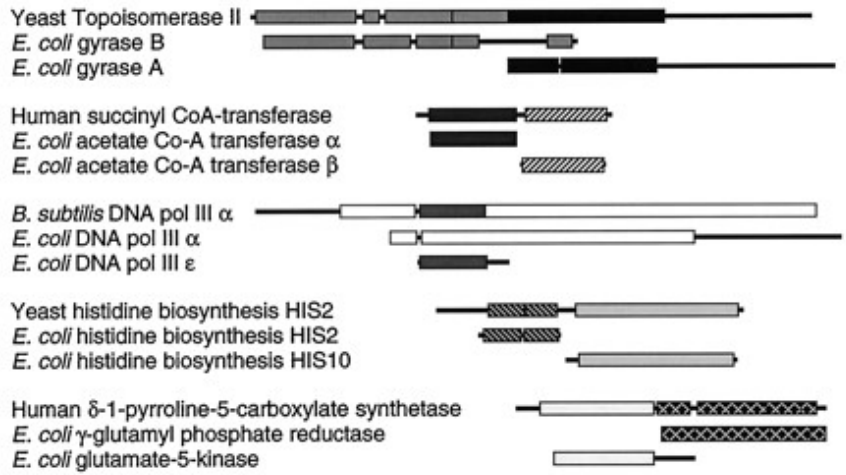
Protein domains



Rabbit pyruvate kinase

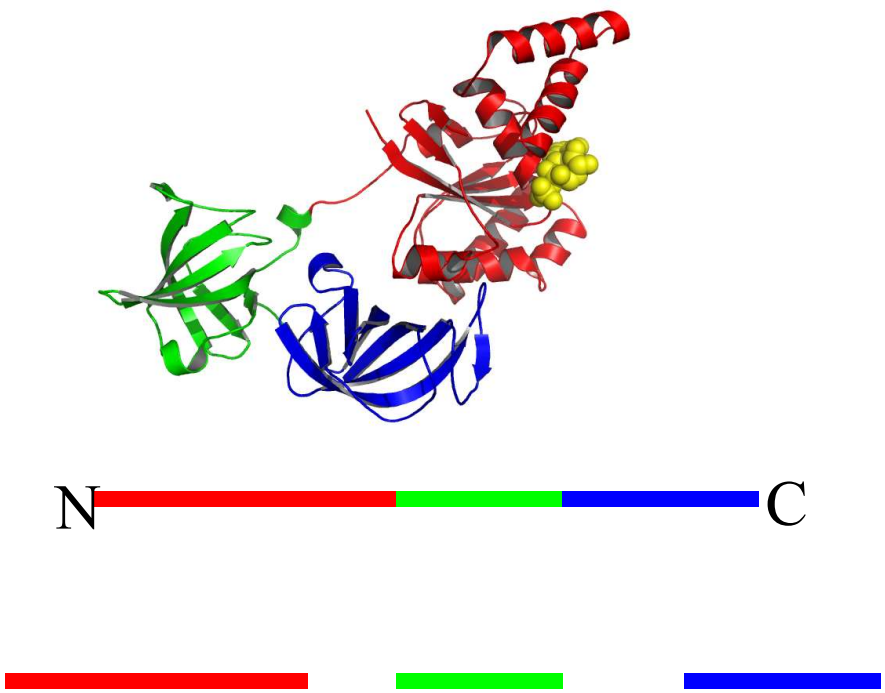
(Wikipedia)





Domains are the functional, structural and evolutionary units of proteins. They are quite independent in all these aspects. And as so they should be considered in protein studies (evolution, structure prediction, …)

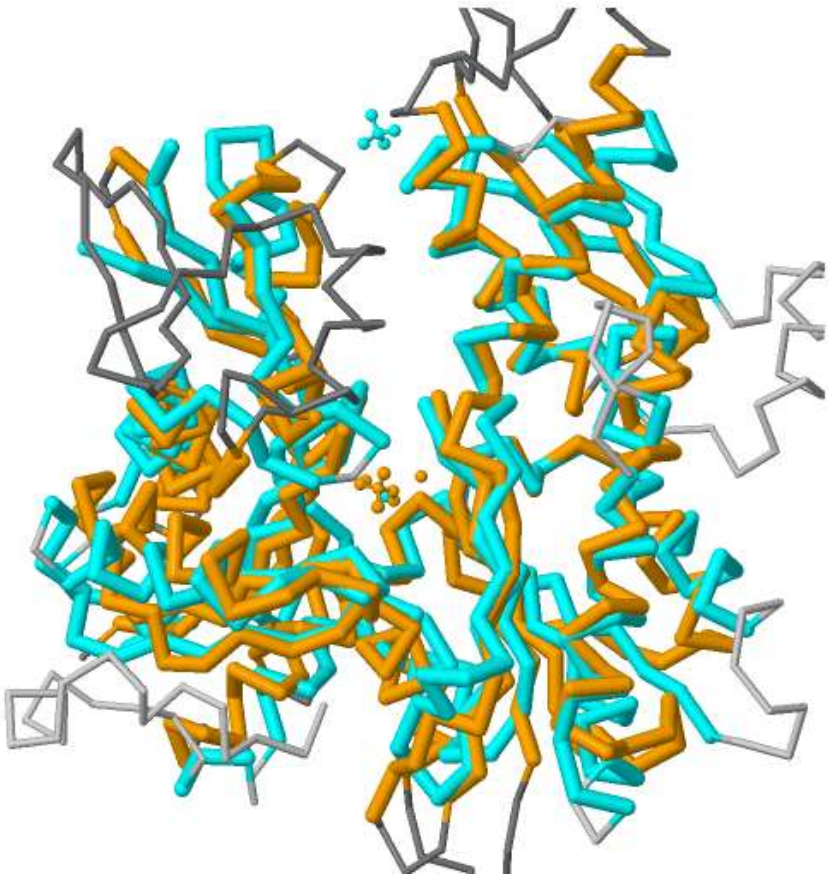
Multi-domain proteins

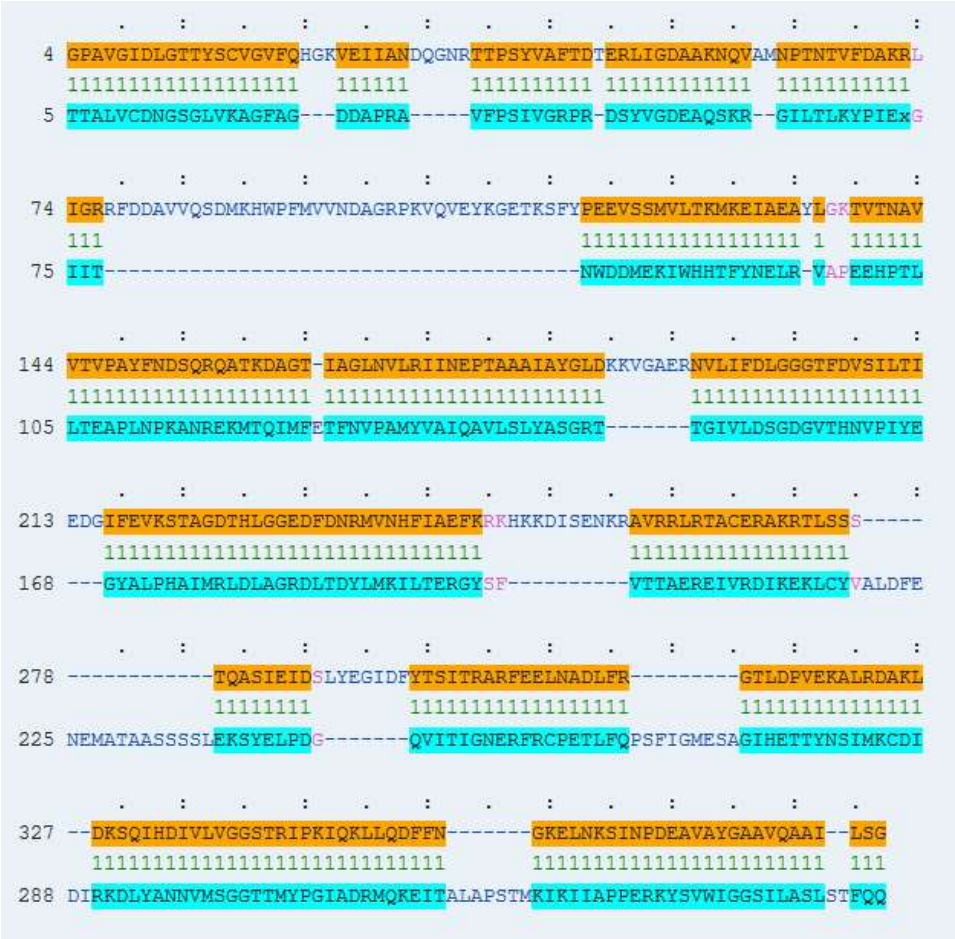


Many of the discussed approaches for structure prediction split into domains **implicitly**. E.g. different fragments/templates for the different domains.

But… If 1) the domain composition of the target sequence is known (or suspected) or 2) models present problems apparently due to domains => Model individual domains separately.

Structural alignments



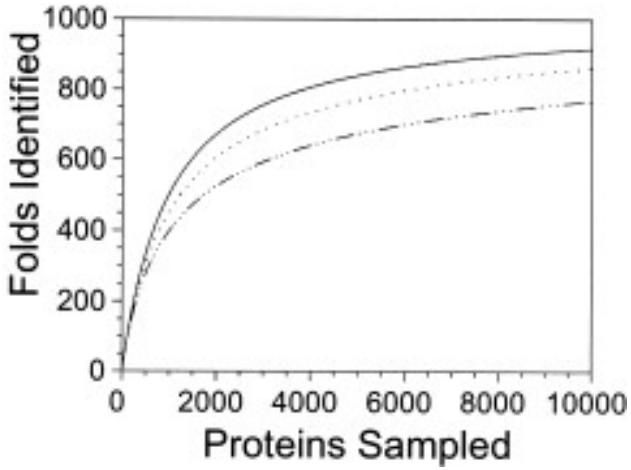


Based on structural/geometric criteria (not sequence matching)

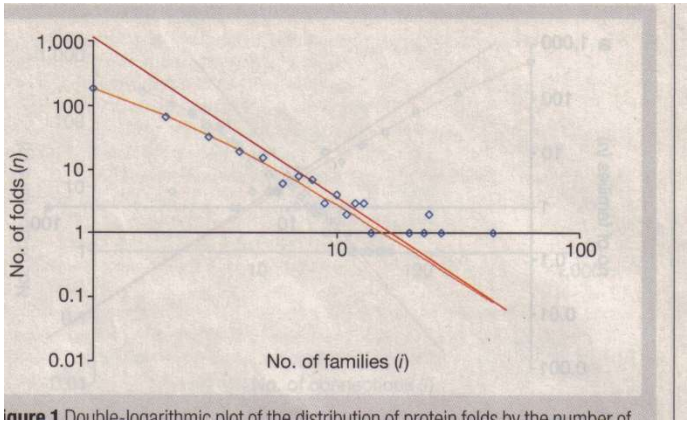
Only way to align distant homologs (e.g. to locate equivalent ("conserved"/functional) residues, etc.) and structural analogs (same structure but no homology)

Also used for constructing protein classifications (detect similar folds), evaluate models, ….

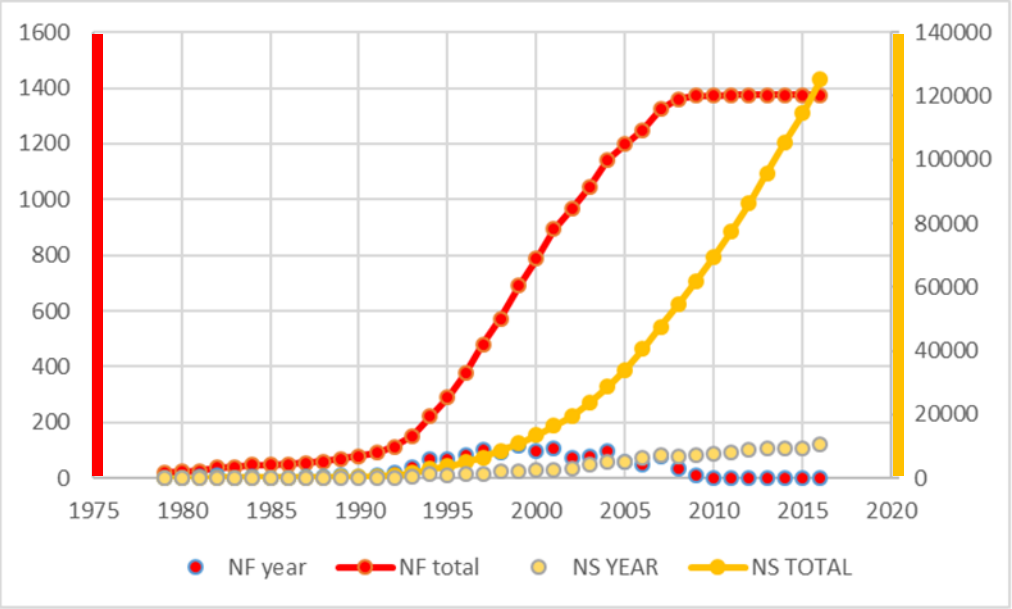
Characteristics of the structural space
Relationship with the sequence space



Leonov, H., Mitchell, J.S. & Arkin, I.T. (2003) Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions. *Proteins*, **51**, 352-359.



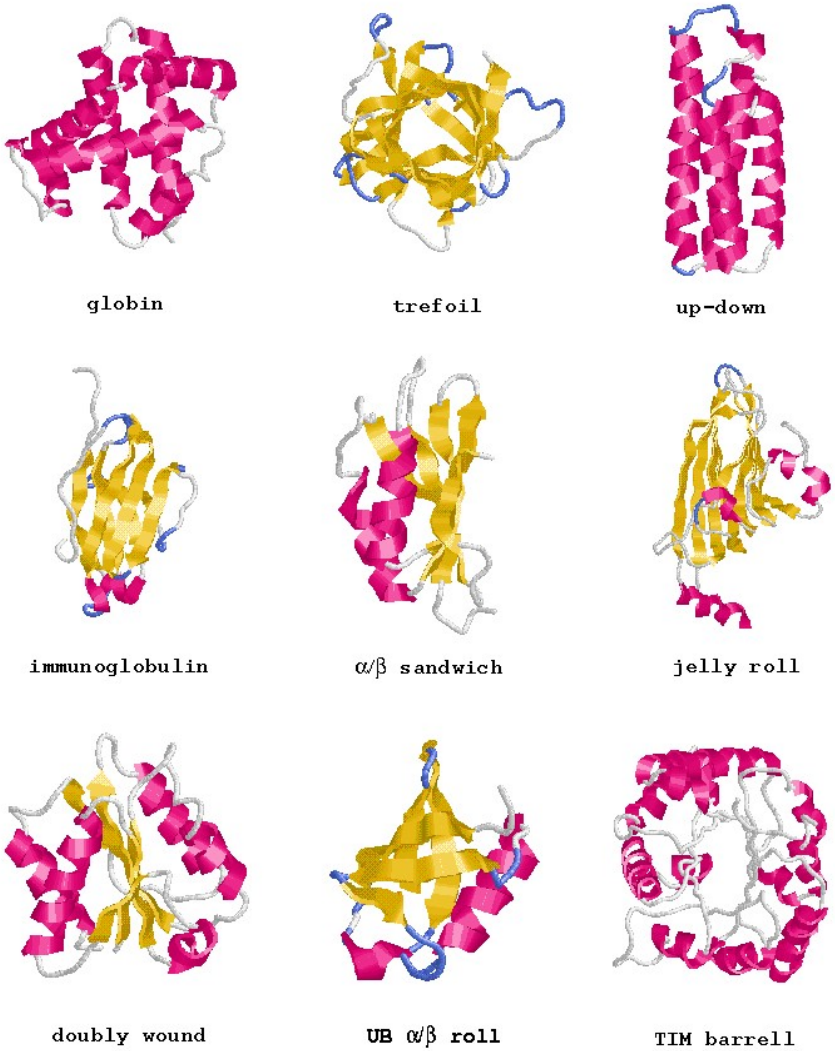
Koonin, E.V., Wolf, Y.I. & Karev, G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, **420**, 218-223.



There is a “small” number of different folds/topologies in nature and their sequence population is highly un-even

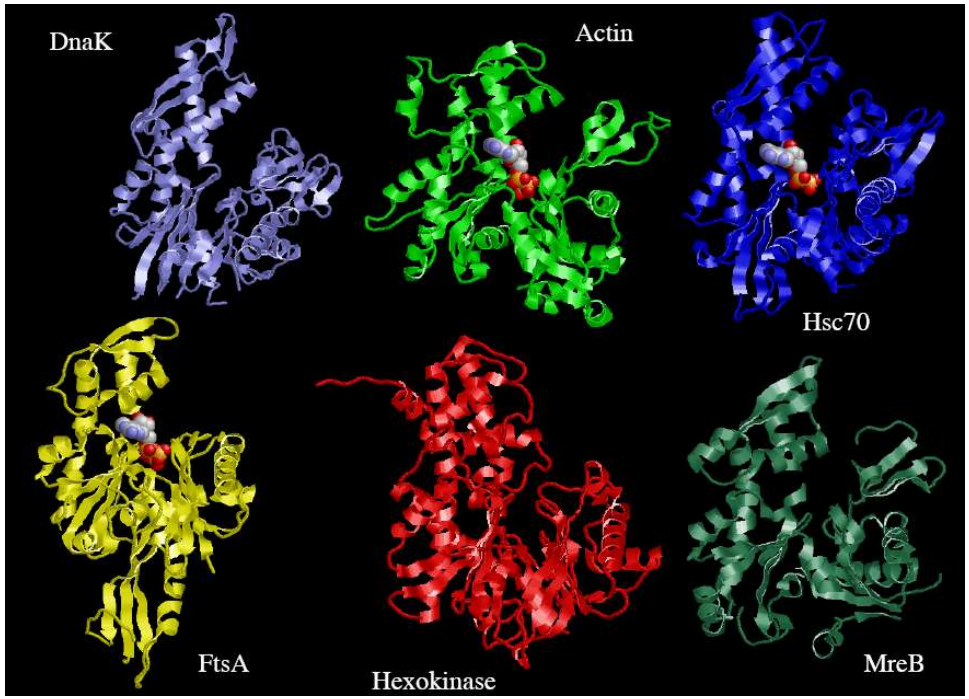
Characteristics of the structural space
Relationship with the sequence space

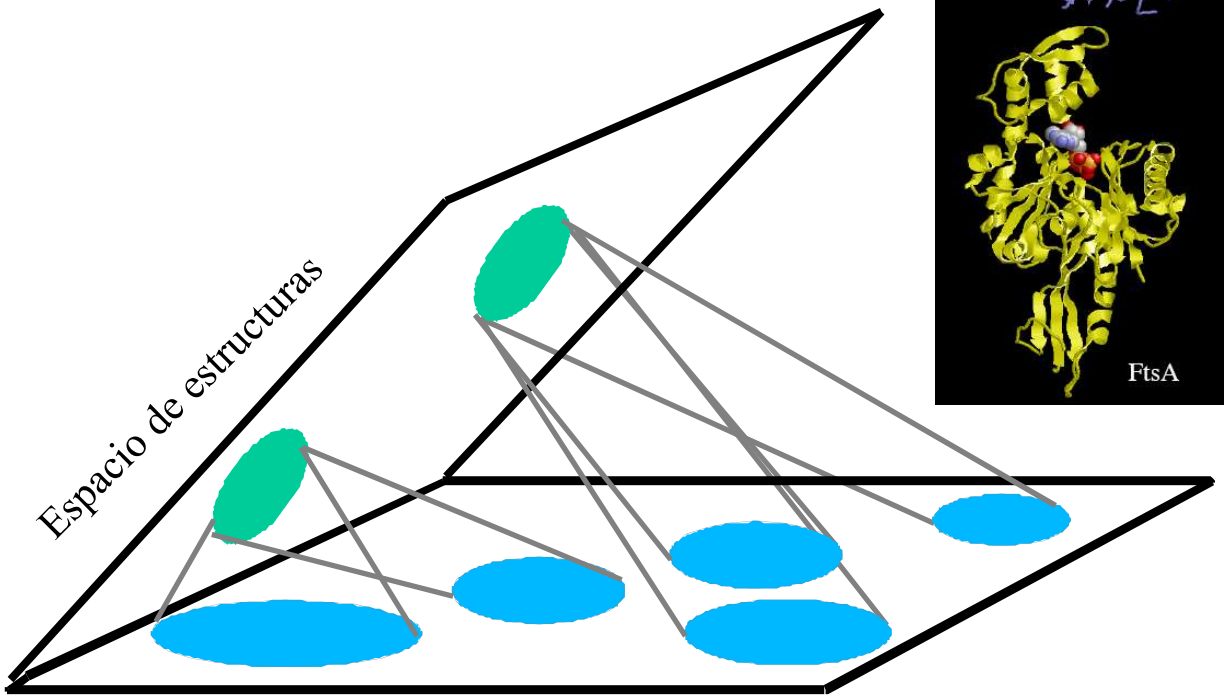
Highly populated folds
(*superfolds*)



Orengo, C.A., Jones, D.T. & Thornton, J.M. (1994) Protein superfamilies and domain superfolds. *Nature*, **372**, 631-634.

Characteristics of the structural space
Relationship with the sequence space

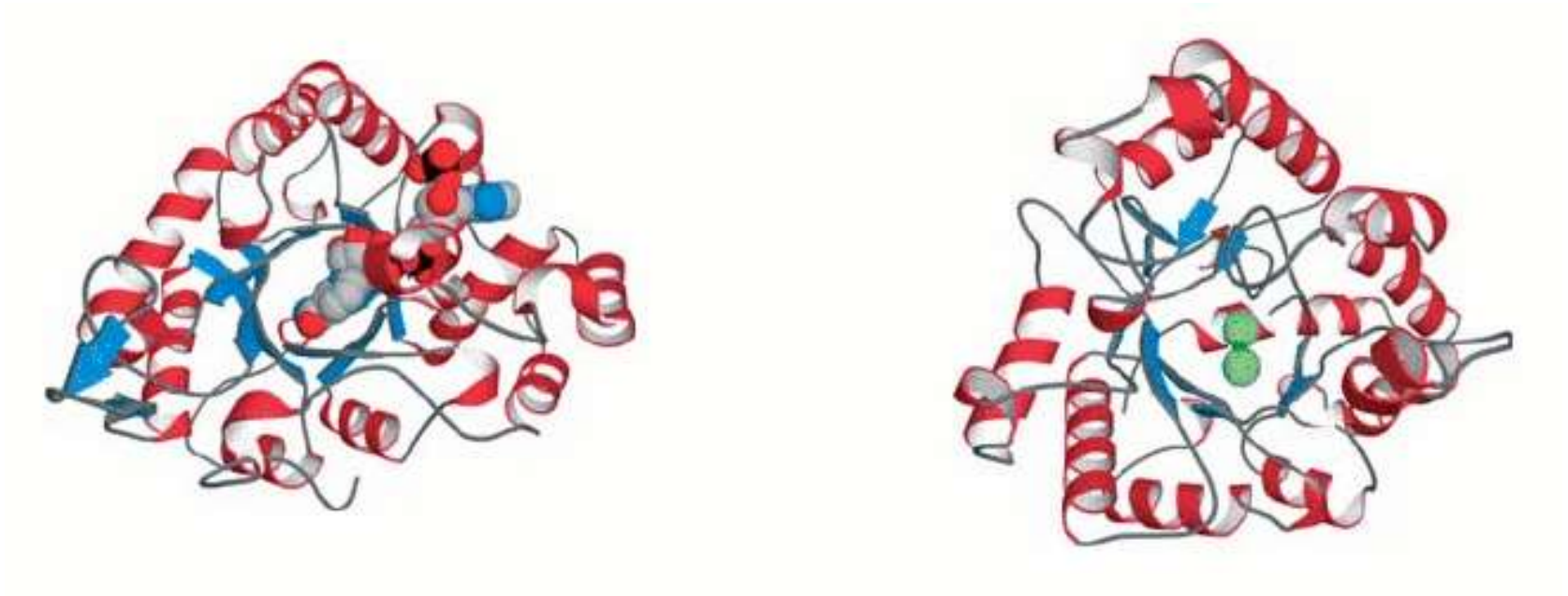




Very different sequences can fold into the same 3D structure…. Either having the same (distant) evolutionary origin ("distant homologs")….

Holmes, K., Sander, C., Valencia., A. (1992) A new ATP-binding fold in actin, hexokinase and Hsc70, *Trends in Cell Biol.*, **3**, 53-59.

Characteristics of the structural space
Relationship with the sequence space

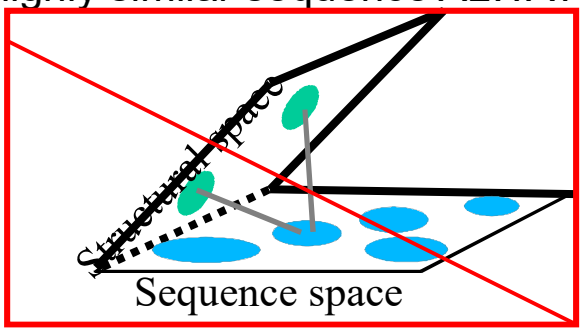


… Or without any traceable homology (common ancestry) => totally unrelated sequences (convergent evolution to the same structure)

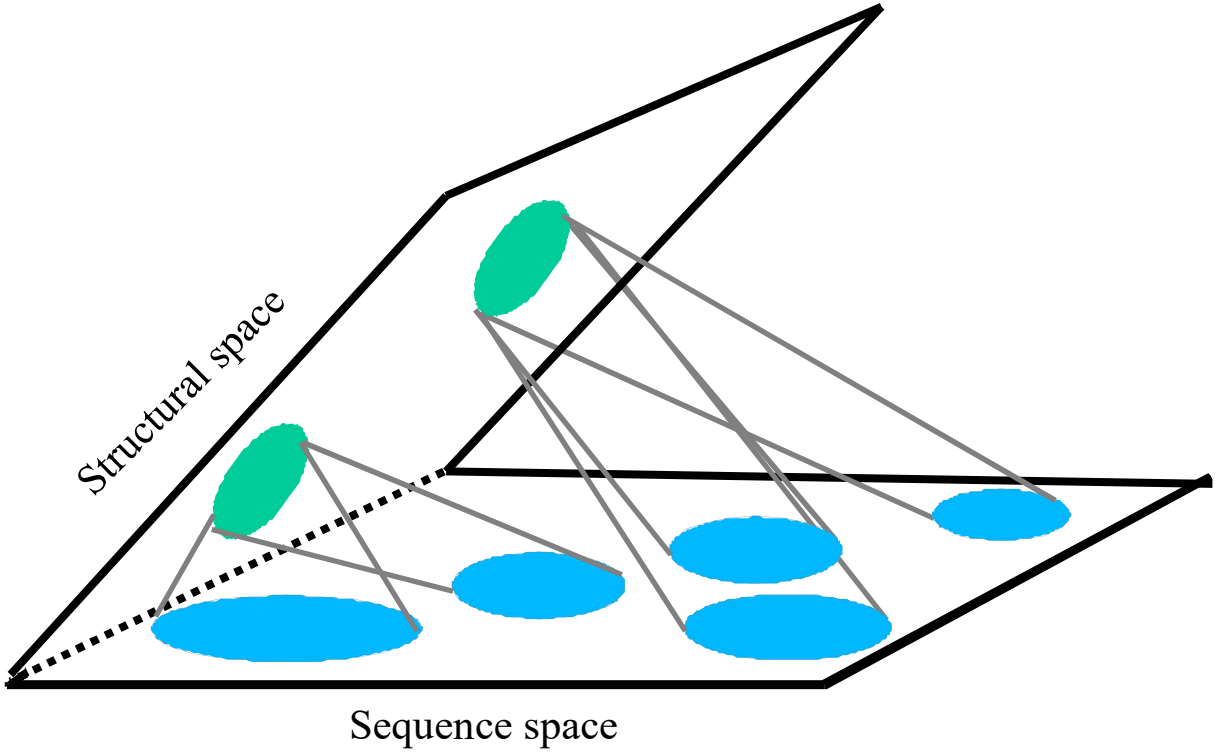
Petsko, GA, Ringe, D (2007) Protein Structure and Function.New Science Press.

Characteristics of the structural space
Relationship with the sequence space

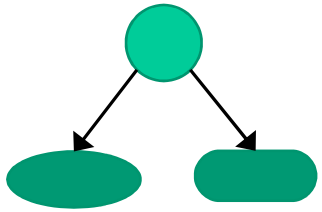
But the contrary is not true: highly similar sequence ALWAYS fold into the same 3D structure



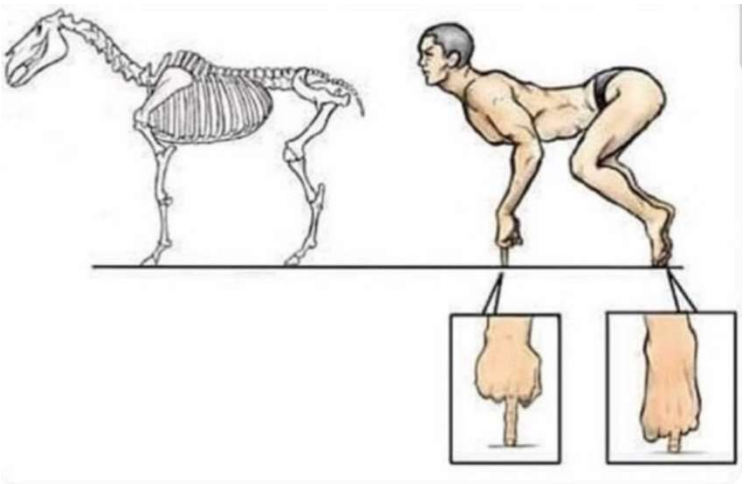
So the relationship between the sequence and structural spaces is “convergent”

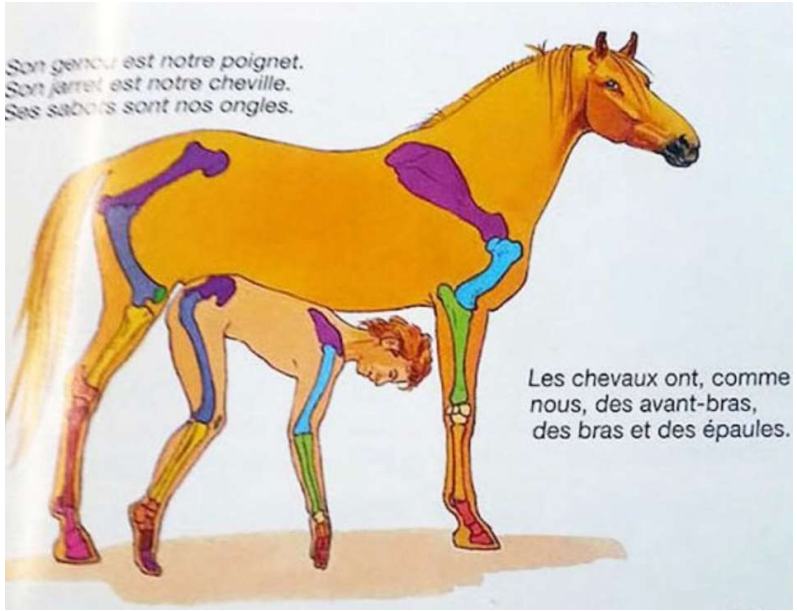


Homology

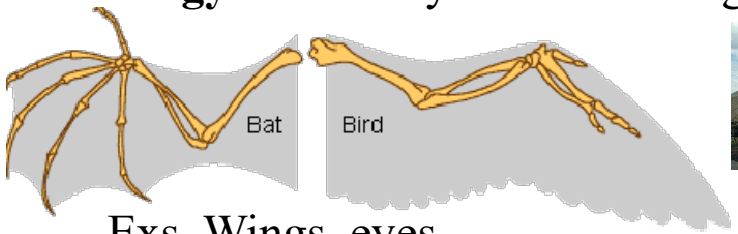


Common ancestry
Reflected in sequence, structure and function similarity
=> These features can be (with caution) transferred between homologs

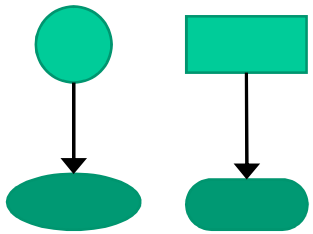




Analogy: Similarity due to convergent evolution from different origins



Exs. Wings, eyes, …



Pazos, F. and Sanchez-Pulido, L. (2014) Protein Superfamilies. In, *eLS*. John Wiley & Sons, Ltd, Chichester, DOI: 10.1002/9780470015902.a9780470025587.

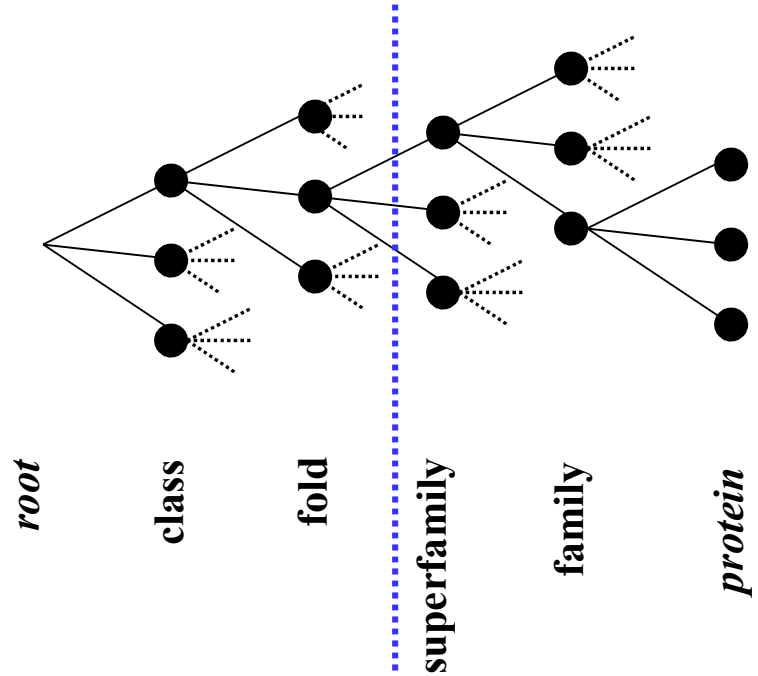
https://evolution.berkeley.edu/evolibrary/article/evo_09

Hierarchical classification of the structural (and sequence) space
SCOP

Root: scop

Classes:

1. All alpha proteins [46456] (218)
2. All beta proteins [48724] (144)
3. Alpha and beta proteins (a/b) [51349] (136)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) [53931] (279)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) [56572] (46)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides [56835] (47)
   Does not include proteins in the immune system
7. Small proteins [56992] (75)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins [57942] (6)
   Not a true class
9. Low resolution protein structures [58117] (24)
   Not a true class
10. Peptides [58231] (116)
    Peptides and fragments. Not a true class
11. Designed proteins [58788] (42)
    Experimental structures of proteins with essentially non-natural sequences. Not a true class

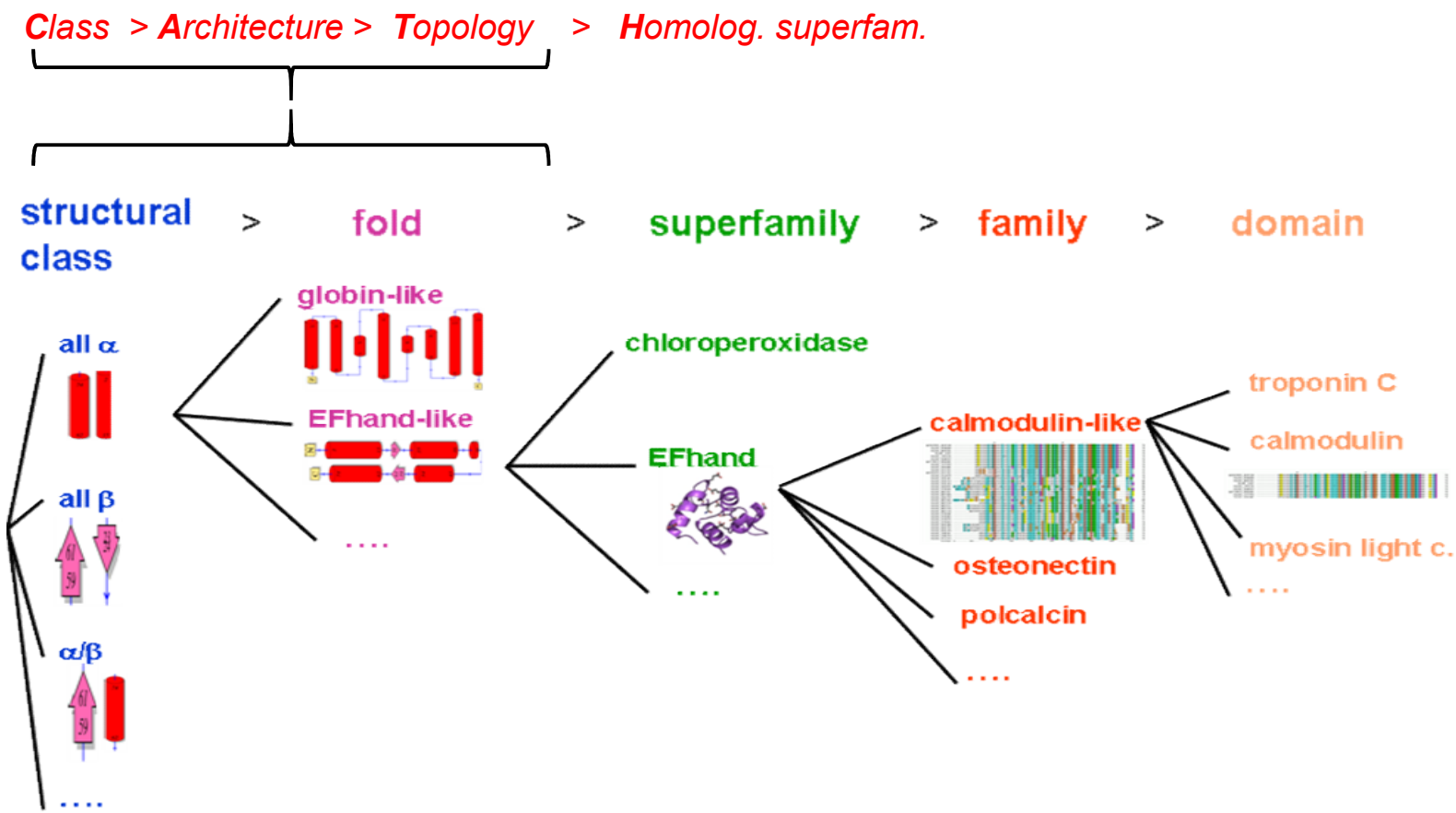


Class: Proteins with similar secondary structure content
Fold: " .. and with similar arrangement of sec. str. elements
Superfamily: " … and with the same evolutionary origin (homologs)
Family: " …. and with clear sequence similarity

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226-229.

http://scop.mrc-lmb.cam.ac.uk/scop/index.html

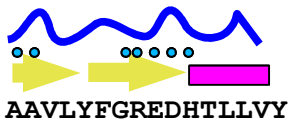
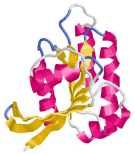
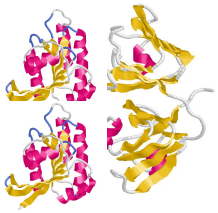
Hierarchical classification of the structural (and sequence) space
SCOP / *CATH*



Pearl F, Todd A, Sillitoe I *et al.* (2005) The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Research* **33:** D247–D251. **http://www.cathdb.info/**

Pazos, F. and Sanchez-Pulido, L. (2014) Protein Superfamilies. In, *eLS*. John Wiley & Sons, Ltd, Chichester, DOI: 10.1002/9780470015902.a9780470025587.

Protein Structure prediction
(Historical) classification of methods

Protein structural "level"	secondary	--------	tertiary	quaternary
Protein representation / Use aa sequence alone?	1D 	2D 	3D 	4D 
Ab Initio	Secondary structure prediction	Contact prediction	- Molecular dynamics - Energy minimisation	*docking*
No Ab-Initio	Secondary structure prediction		- Homology modeling - *Threading*	*docking with restraints*

Protein structure prediction
1D characteristics

1D sequence characteristics: Characteristics that can be represented by a single value associated with each amino acid (B. Rost).

These values often take the form of status labels, such as secondary structure (H-> helix, E> sheet, T-> turn). They can also take continuous values (% accessible surface ...)

Some 1D features:

- secondary structure
- Solvent accessibility
- Post-translational modifications
- signal peptides
- Coiled-coils
- disordered regions
- etc.

Why to predict secondary structure and other 1D characteristics, instead of 3D directly?

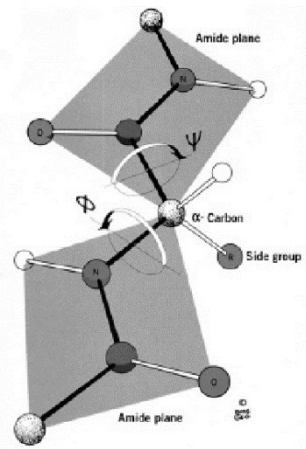
It is not always possible to generate a 3D model (reliable).

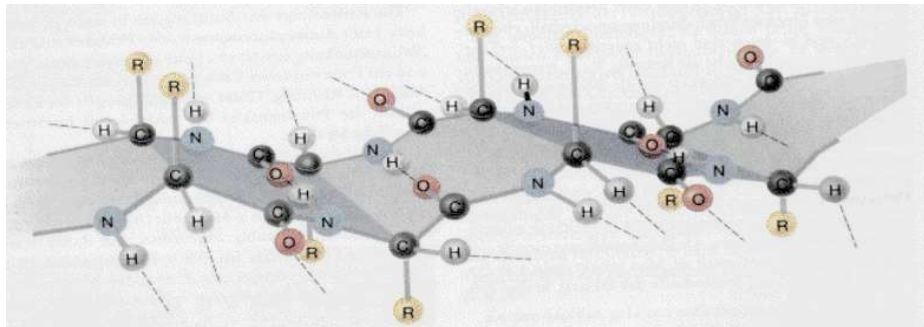
Help predicting 3D folding (restricts possible folds/models)

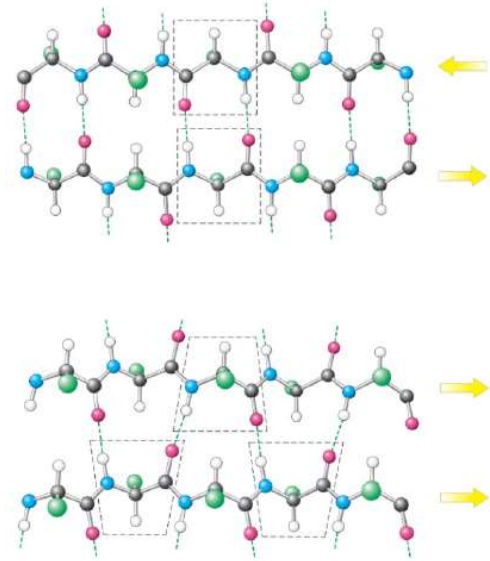
Function Prediction: e.g. particular secondary structure motifs associated to certain functions, disordered regions involved in binding

The mapping of all the 1D predictions along a sequence gives much information about possible structural and functional domains, active sites, distinct areas ....

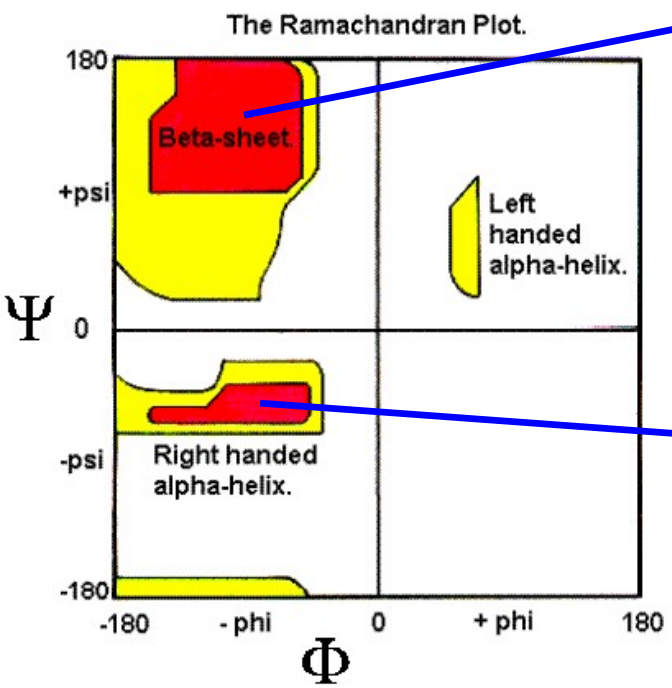
1D characteristics
Secondary structure



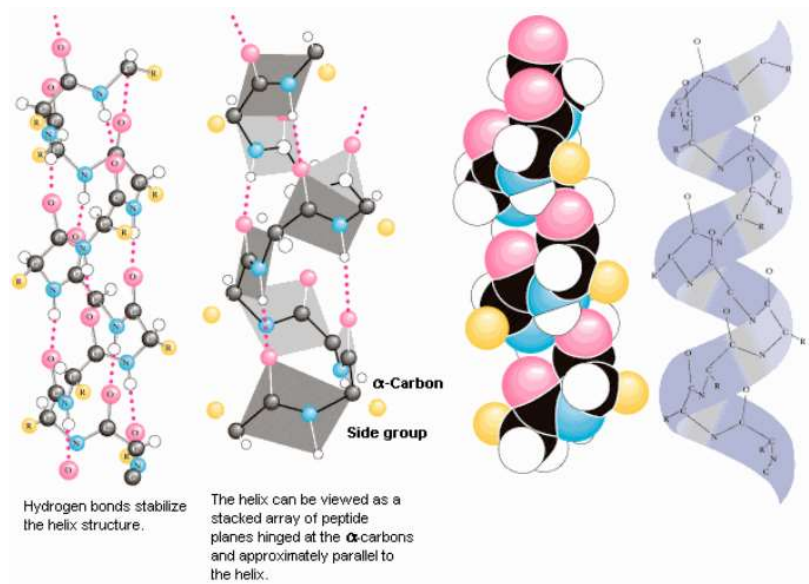




β-strand



α-helix



1D characteristics
Secondary structure

```
  1 ASKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTT
      TTGGGGSSEEEEEEEEEEEETTEEEEEEEEEEEETTTTEEEEEEEETT

 51 GKLPVPWPTLVTTFSYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIFF
    SS SS GGGGHHHHSSS GGG B  GGGGGG HHHHTTTT EEEEEEEEE

101 KDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYNYNSHNV
    TTS EEEEEEEEEEETTEEEEEEEEEE   TTSTTTTT B S    EEE

151 YIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHY
    EEEEEGGGTEEEEEEEEEEETTS EEEEEEEEEEESSSS      SEE

201 LSTQSALSKDPNEKRDHMVLLEFVTAAGIT HGMDELYK
    EEEEEEEE  TT  SSEEEEEEEEEES
```

Usually different "vocabularies" of secondary structure states used for…

<u>Definition:</u> T=hydrogen bond turn, H=helix, G=310 helix, I=phi helix, B=residue in isolated beta bridge, E=strand, and S=bend

<u>Prediction:</u> H/E/T (3 states only)

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Secondary structure prediction
First generation methods

Based on simple statistics: propensities of each aminoacid to form each type of secondary structure.

- Chou and Fasman in 1974, proposed the first of these methods. They used statistics from the 15 structures solved by X-ray crystallography at the time. These probabilities were calculated separately for each residue. Later this method showed an accuracy of 57% (3-states) on 62 proteins.

- Garnier (1978). Similar but statistics based on pairs of residues (accuracy: ~ 60%)

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Aspartic Acid	101	54	146	0.147	0.110	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic Acid	151	037	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

Glu, Met Ala y Leu : tend to form hélices.
Val, Ile y Tyr: tend to be in beta strands.
Gly, Pro, ...: turns.

Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222-244/225.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97-120.

Secondary structure prediction
Second generation methods

- Input: longer windows of adjacent residues (=> context information). Coupled to more advanced machine learning and statistical methods: neural networks, graph theory, rule-based systems, multivar statistics.
- ~ 70% accuracy (3 states).
- Limitations:
  - Lower accuracies for β strands.
  - Tend to predict too short segments
- Due to…
  - Still low number of structures for training (and biased, e.g. more α than β).
- Long-range effects (3D contacts) not taken into account (only local)

Garnier, J. and Robson, B. (1989) The GOR method for predicting secondary structure in proteins. In D., F.G. (ed.), *Prediction of protein structure and the principles of protein conformation*. Plenum Press, New York, pp. 417-465

Secondary structure prediction
Third generation methods

Initiated by Levin (~69%) and Rost y Sander around 1994 (PHD 72%)

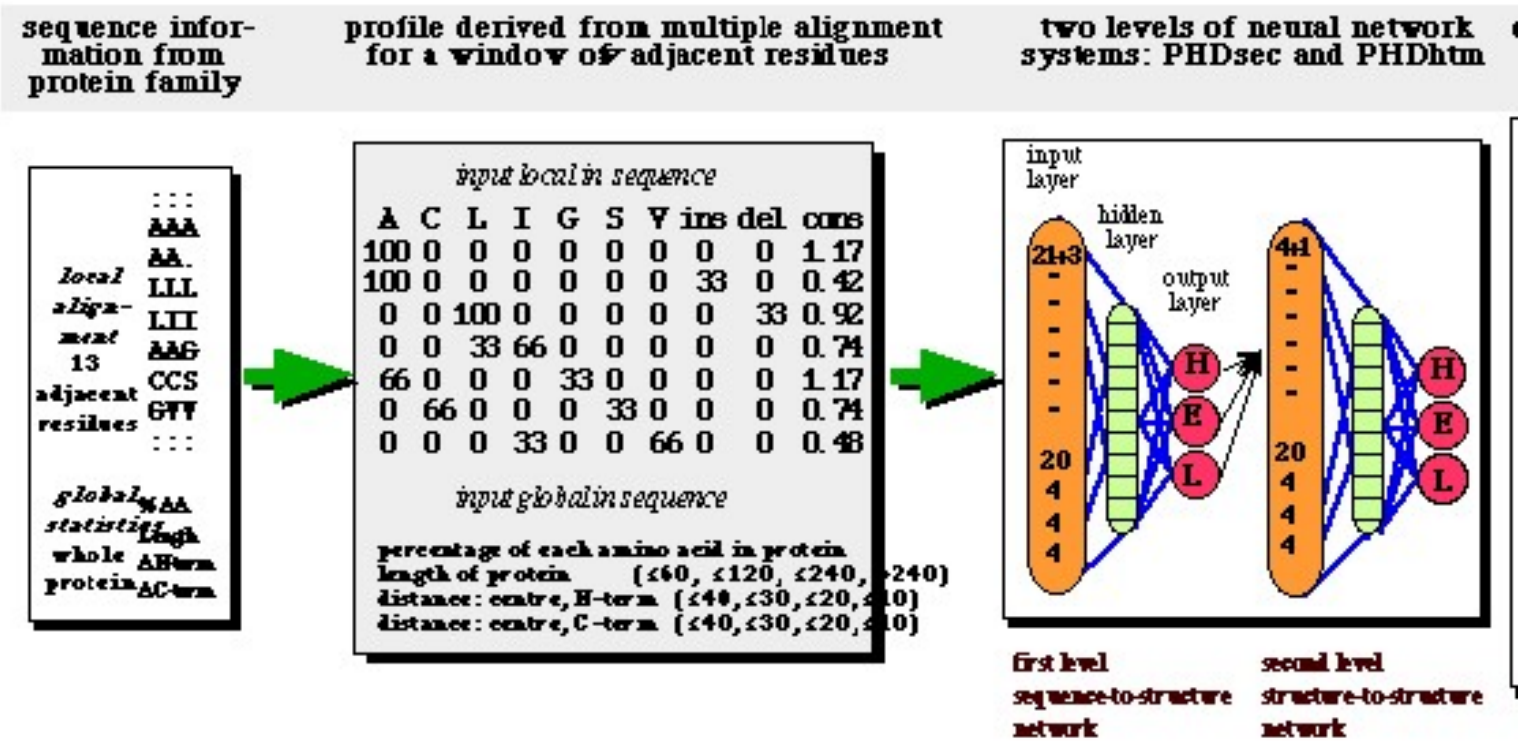
- Main innovation: include evolutionary information in the input: multiple sequence alignments and profiles.
- Solve the problem with the β strands by balancing the training set (richer in α)
- The prediction of a 1st NN is fed to a second one to "soft" the predictions: e.g. avoid too short elements, etc.
- All this breaks the 70% accuracy barrier.

Levin JM, Pascarella S, Argos P, Garnier J. (1993). Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* **6(8):**849-54.

Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

Secondary structure prediction
Example: PHD



Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

Secondary structure prediction
Current methods

- Same methods (NN) but fed with better alignments: e.g. including remote homologs detected by psi-blast (introduced by David Jones with PSIPRED (1999)), or by HMMs (Kevin Karplus *et al.* in SAMT99sec (1999)).

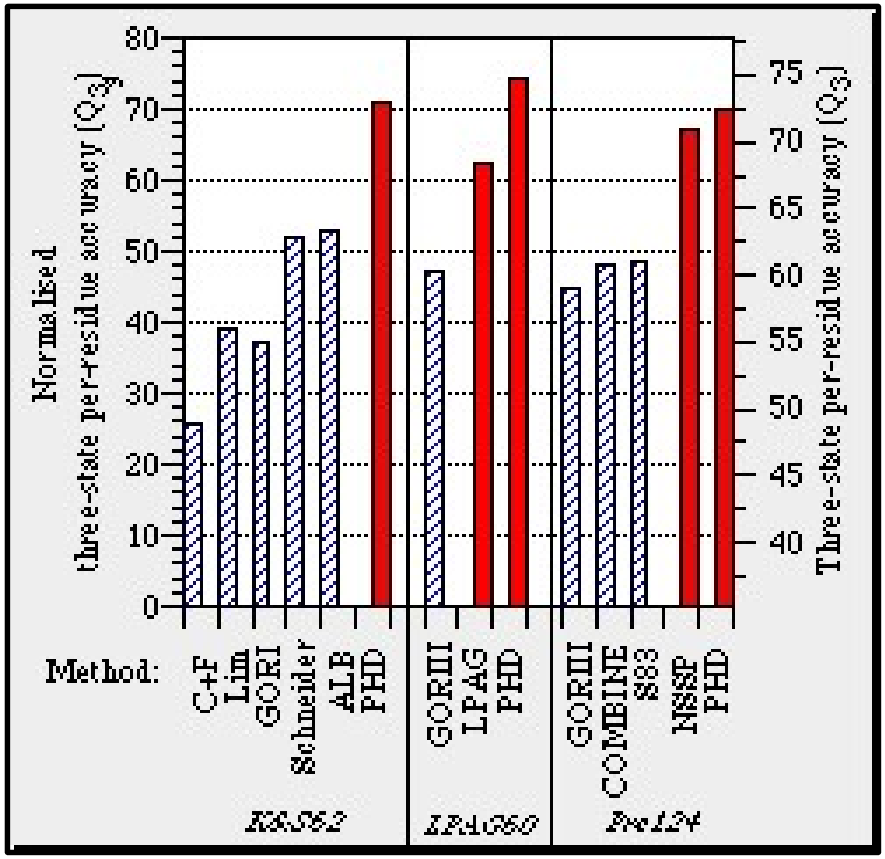
- Consensus methods: run different predictors and combine the predictions. E.g Jpred2 (Cuff y Barton, 2000).

Accuracies ~76-78%

Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195-202.

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. (1998). JPred: a consensus secondary structure prediction server. *Bioinformatics*. **14(10):**892-3.

Secondary Structure Prediction



Métodos de Primera generación: Chou & Fasman, Lim, GORI

Métodos de Segunda generación : Schneider, ALB, GORIII

Métodos de Tercera generación: LPAG, COMBINE, S83, NSSP, PHD

Accuracy limit?

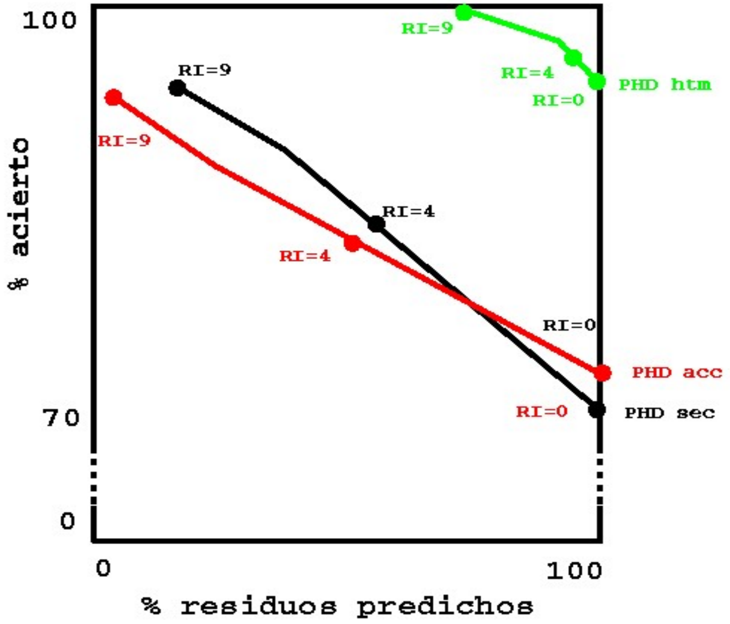
- Intrinsic limit due to the deffinition of secondary structure elements (DSSP vs. others)
- Local information limited

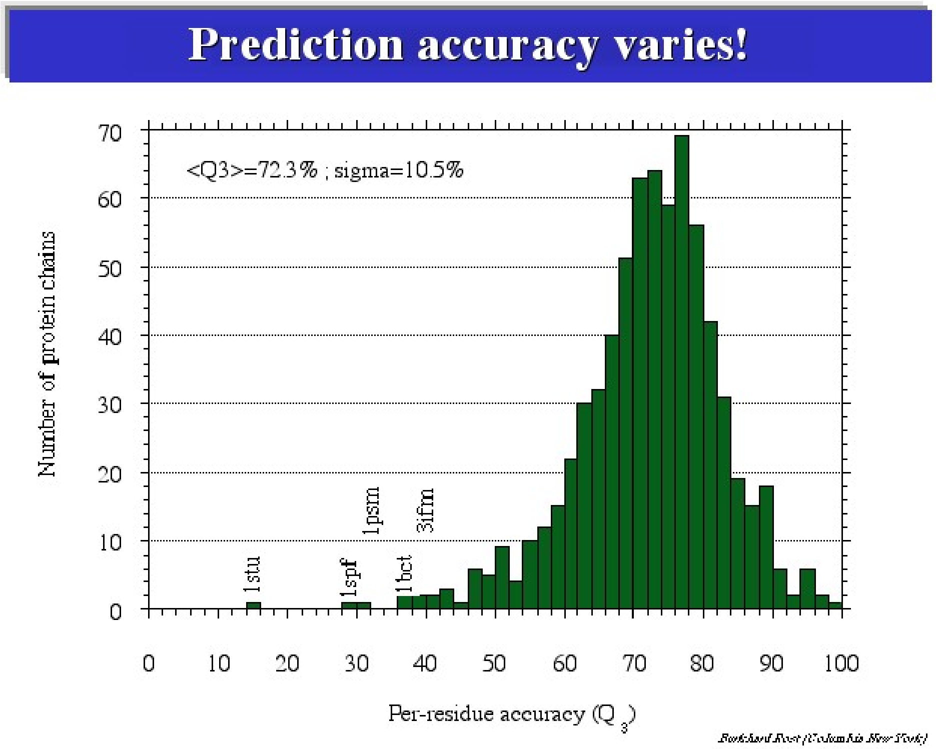
Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Secondary structure prediction
Factors to take into account

1) Equilibrium accuracy/coverage
(through "reliability")

2) Results vary depending on protein

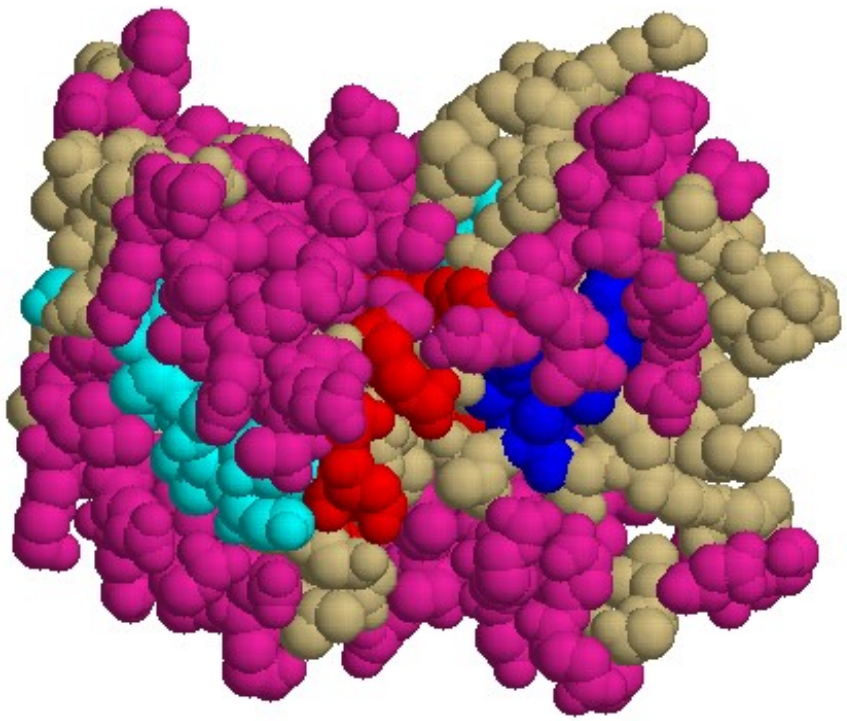




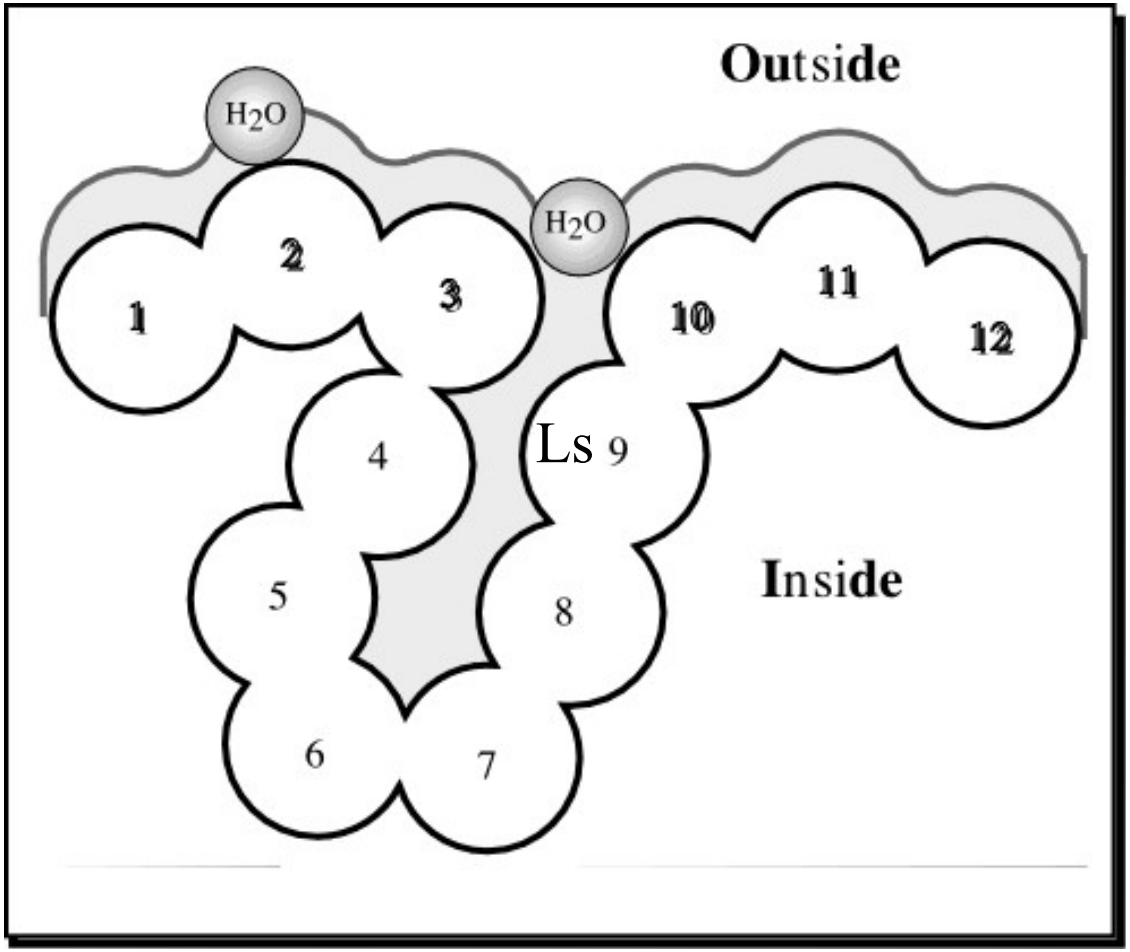
1D Methods
Solvent accessibility prediction

Useful for:

- Discriminating among alternative structural models
- Functional and interaction sites
- Design of mutants, labels for proteins, etc.



Solvent accessibility



Programs for defining solvent accessibility (from a 3D structure) report for each residue the accessible surface, in Å^2.

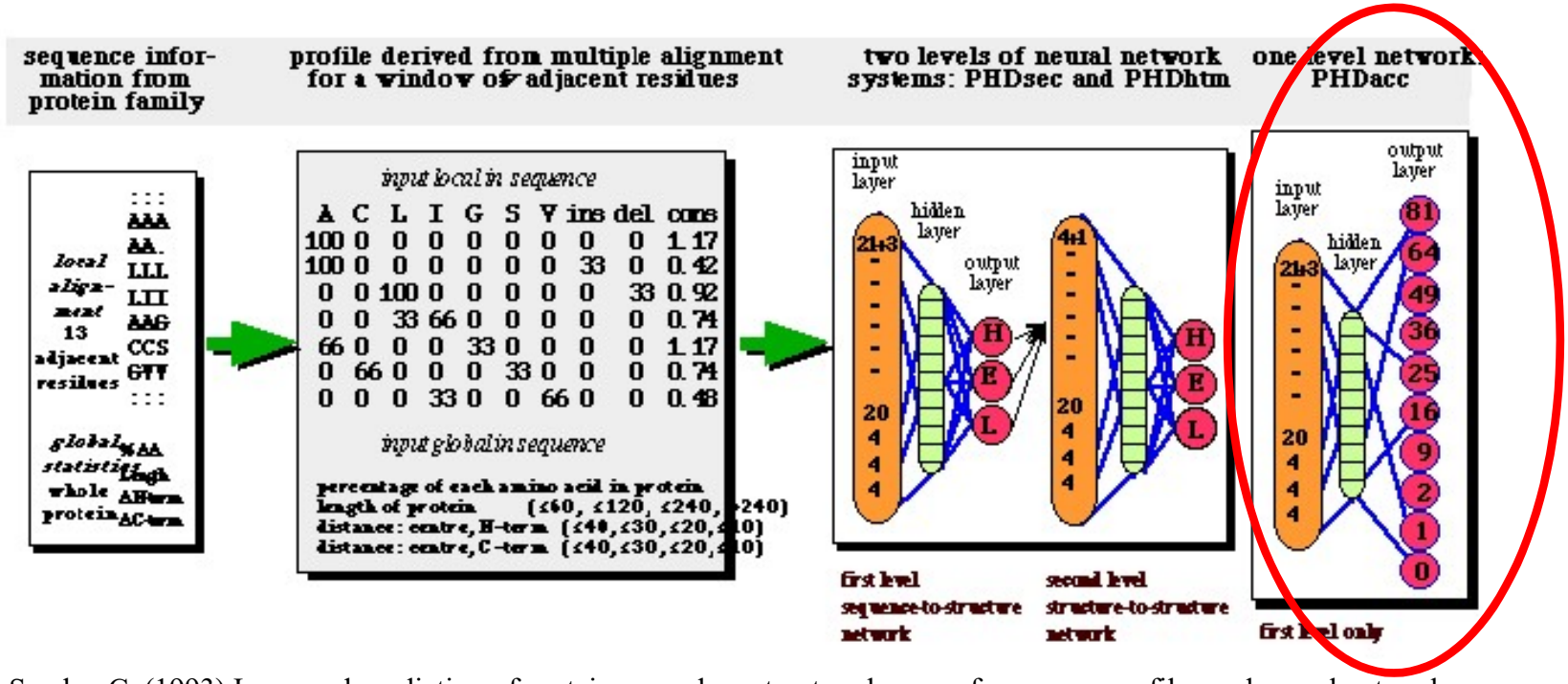
Most prediction methods reduce this to two a number of discrete states: E.g. 2: buried (accs. relative. <16%, abs <50 Å2) vs. exposed; or 10: different levels of accessibility

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.

Solvent accessibility prediction

- Same history as for secondary structure: propensities -> windows -> neural networks -> alignments -> better alignments & consensus methods

- Indeed, the programs are the same as for secondary structure, with minimal adaptations of the neural net to represent the accessibility states. So are the datasets used for training/testing, etc.

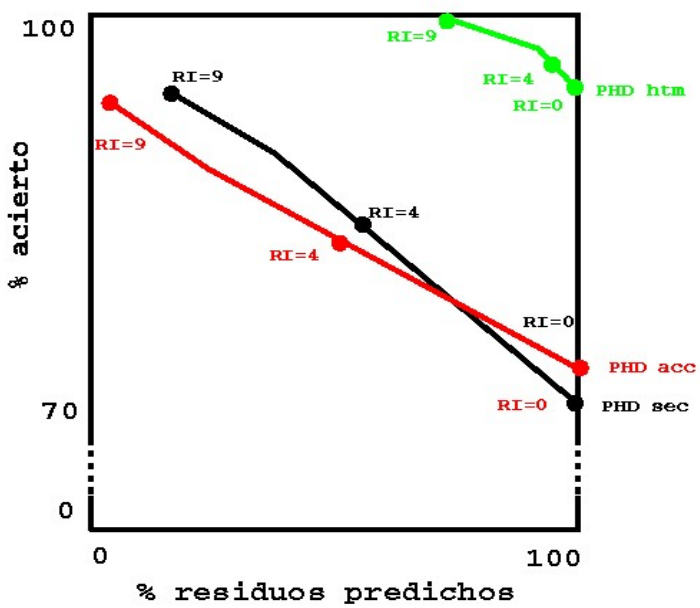


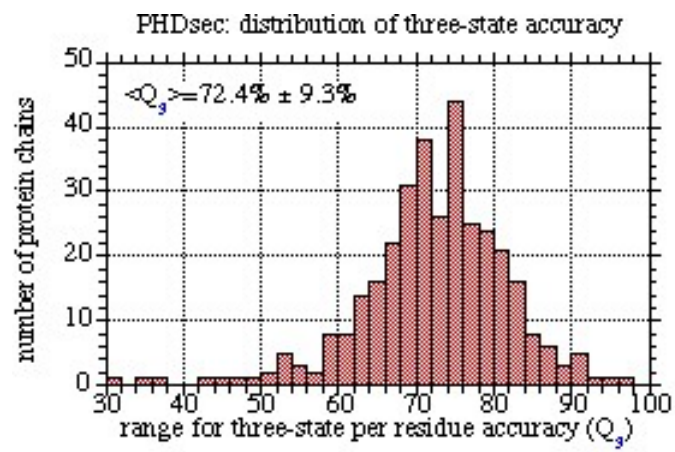
Rost, B. and Sander, C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci U S A*, **90**, 7558-7562.

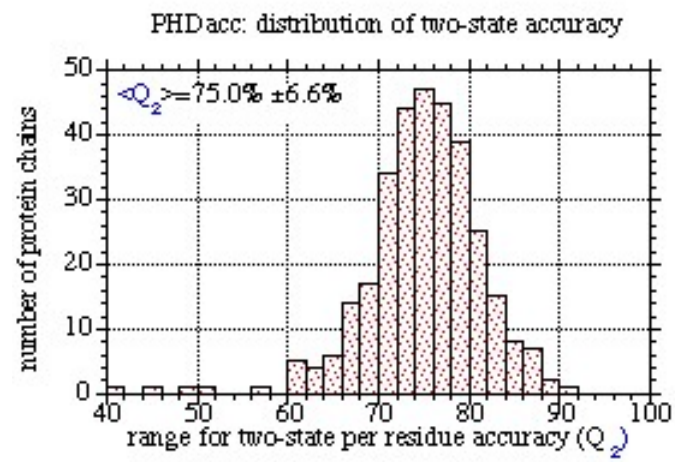
Rost, B., Sander, C. and Schneider, R. (1994) PHD - A mail server for protein secondary structure prediction. *Comp. Applic. Biosci.*, **10**, 53-60.

Solvent accessibility prediction

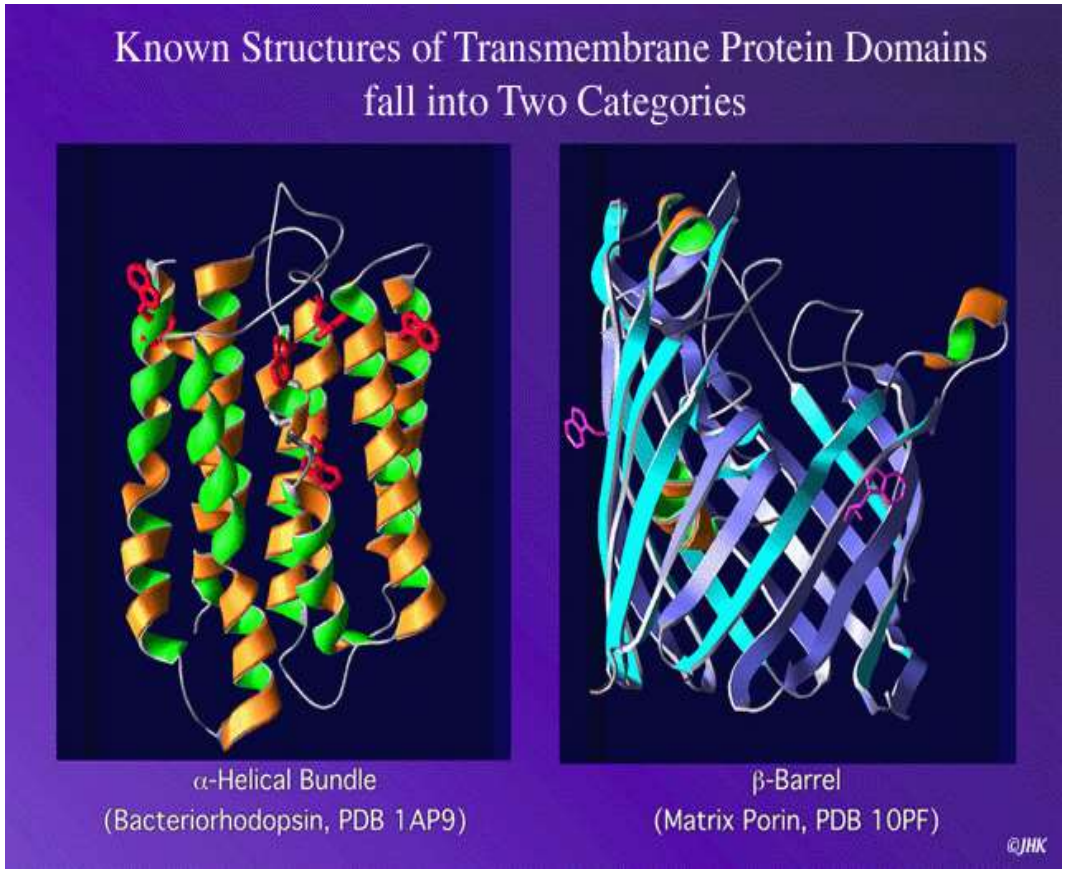
Same average accuracy
Same factors to take into account







1D Methods
Transmembrane segments



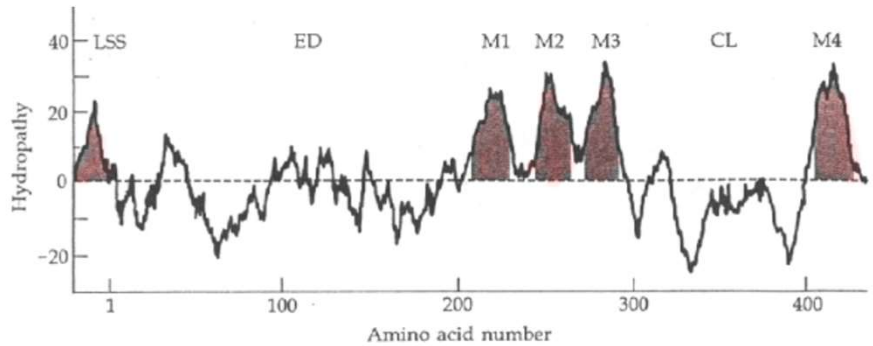
- Difficult to crystallize: very low number of known structures

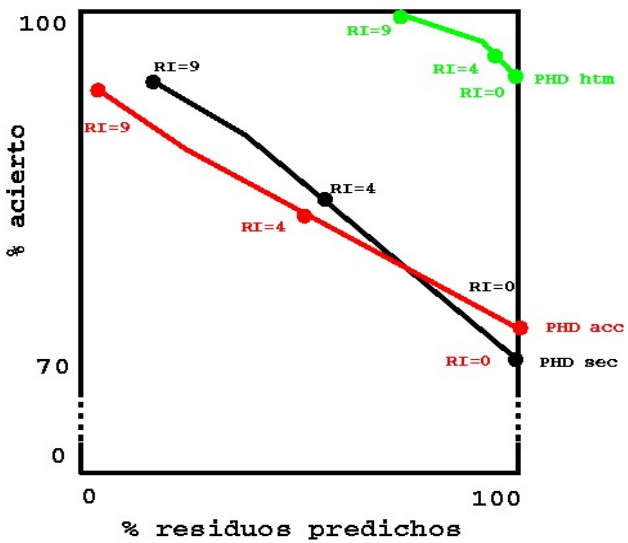
1D Methods
Transmembrane helices

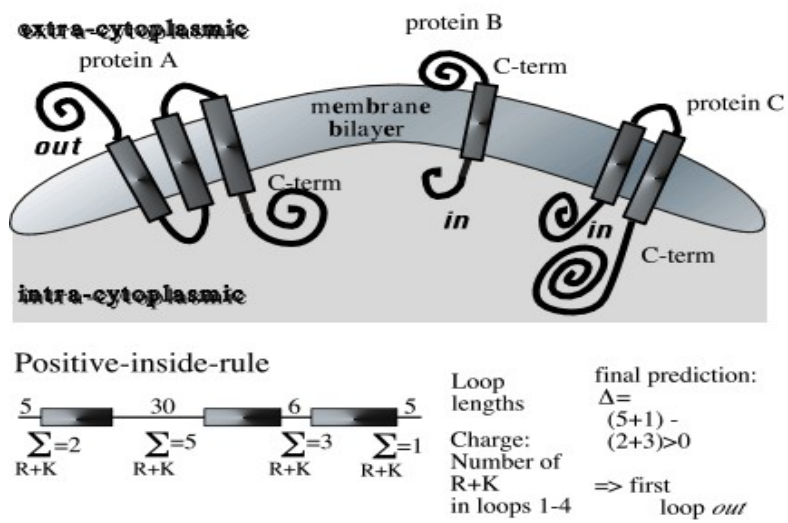
Same methods as for sec. str. and solvent accessibility.

Much higher accuracies due to the peculiarities of these elements:

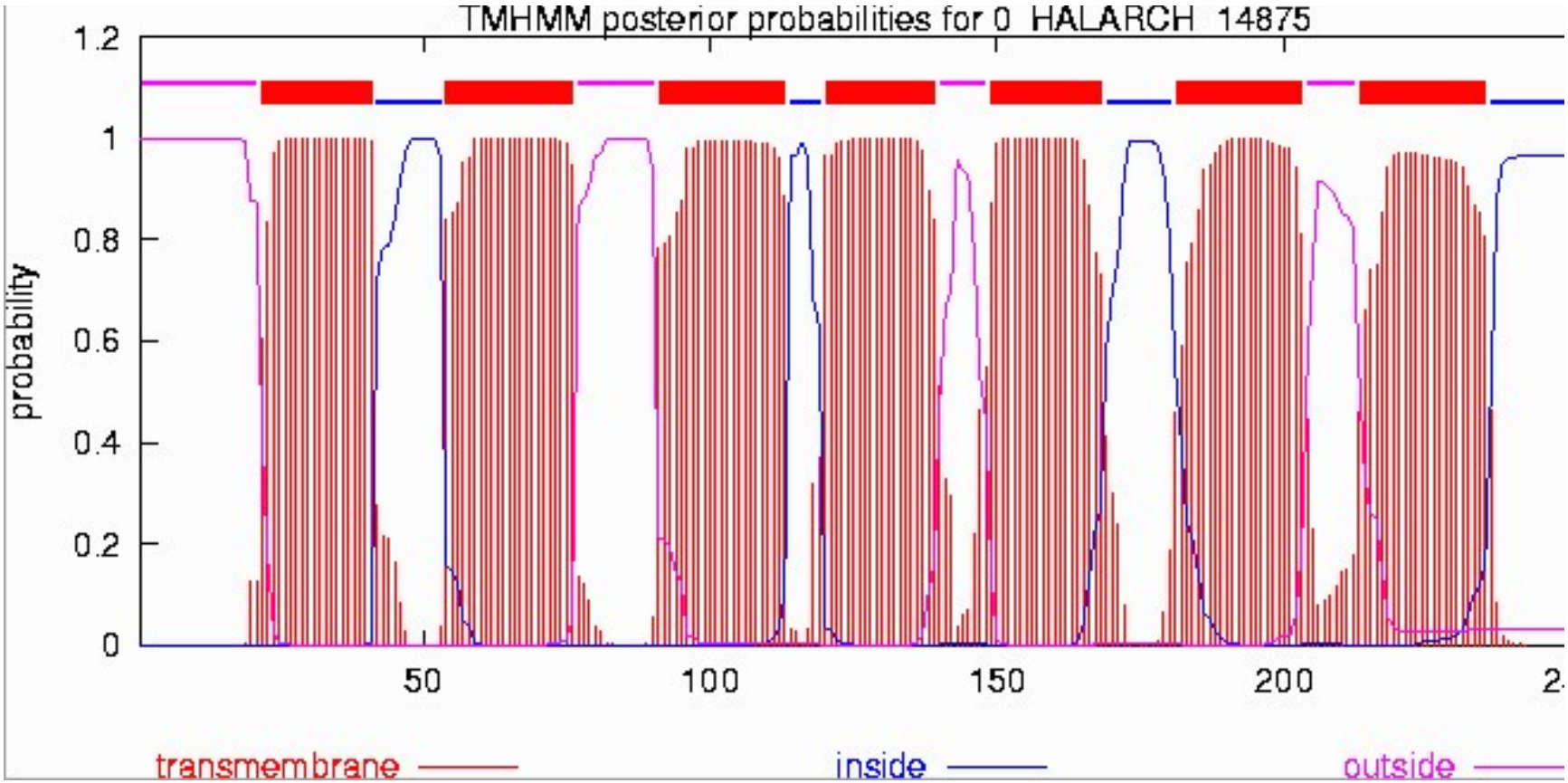
- Fixed length (20-30 res.)
- Rich in hydrophobic residues
- Loops connecting helices in the cytoplasm use to have positive charge.







1D Methods
Transmembrane helices

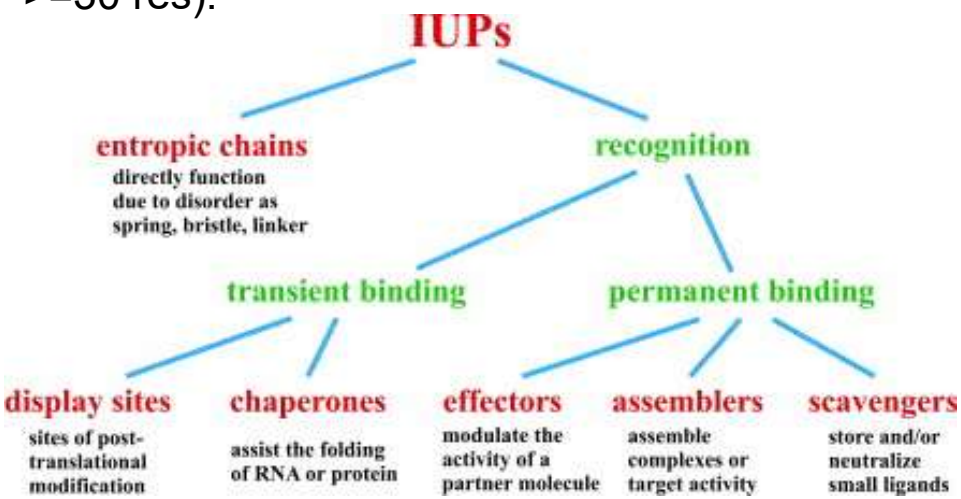


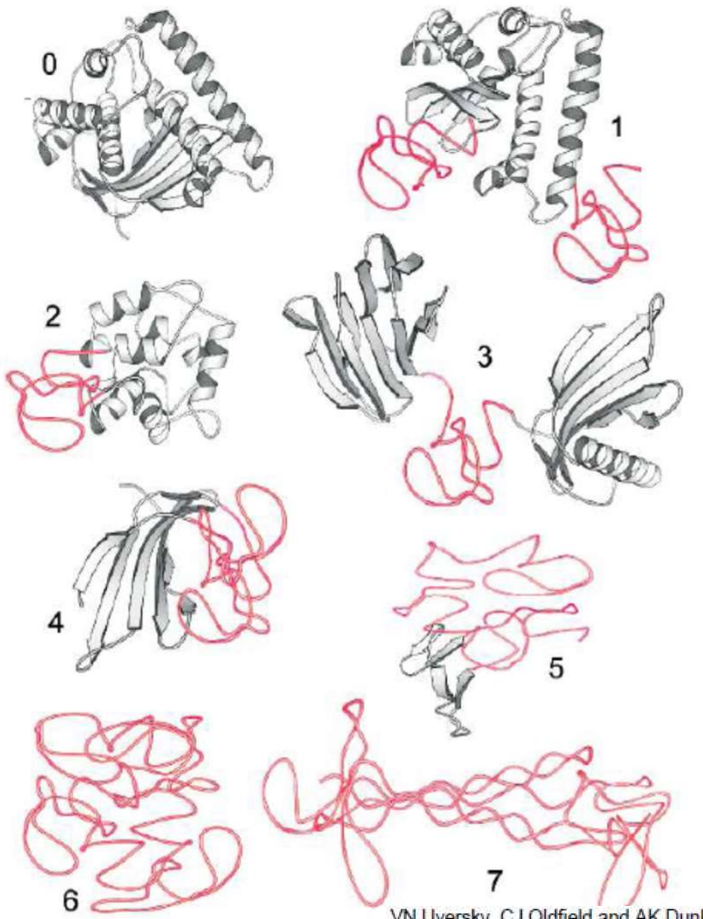
1D Methods
Unstructured regions

A.k.a. disordered regions, intrinsically unstructured regions (IUR), …

Proteins totally or partially unstructure in their native (functional) state.

Importance increasingly being recognized. Involved in central processes. ~70% human proteins predicted to have 1 or more IUR of >=30 res).

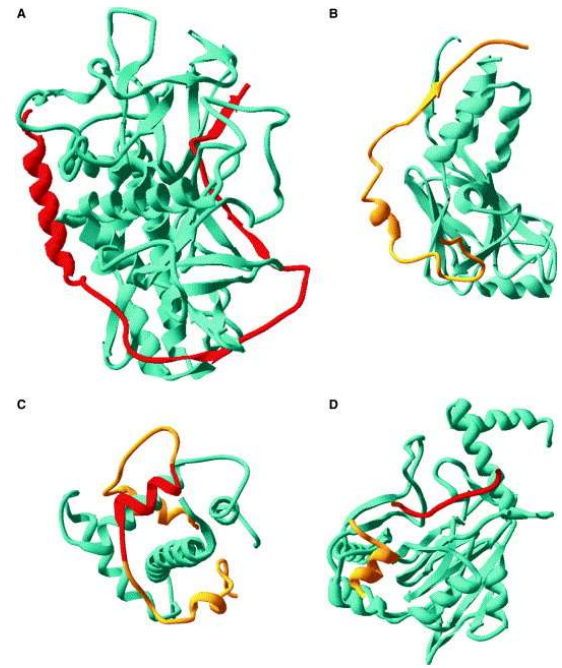




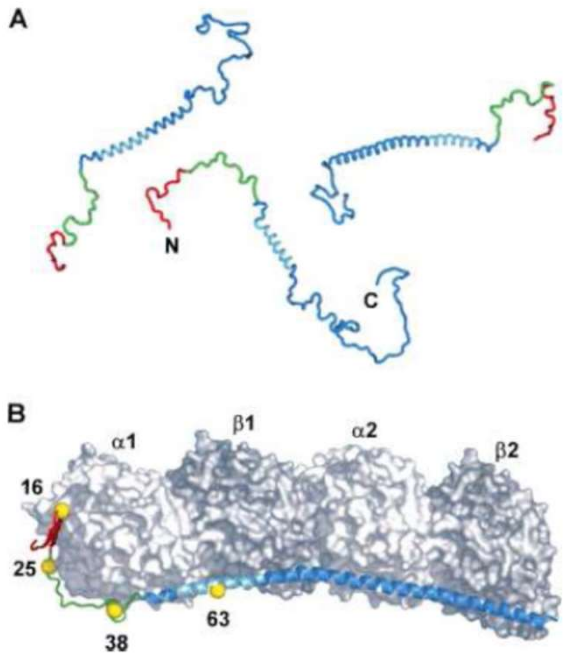
VN Uversky, CJ Oldfield and AK Dunker
Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling
J. Mol. Recognit. (2005) 18: 343–384

•Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*, **579**, 3346-3354.
•Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. Flavors of protein disorder. *Proteins* **52**, 573-84. (2003).
•Pazos, F., Pietrosemoli, N., García-Martín, J.A. and Solano, R. (2013) Protein intrinsic disorder in plants, *Front Plant Sci*, **4,** 363.

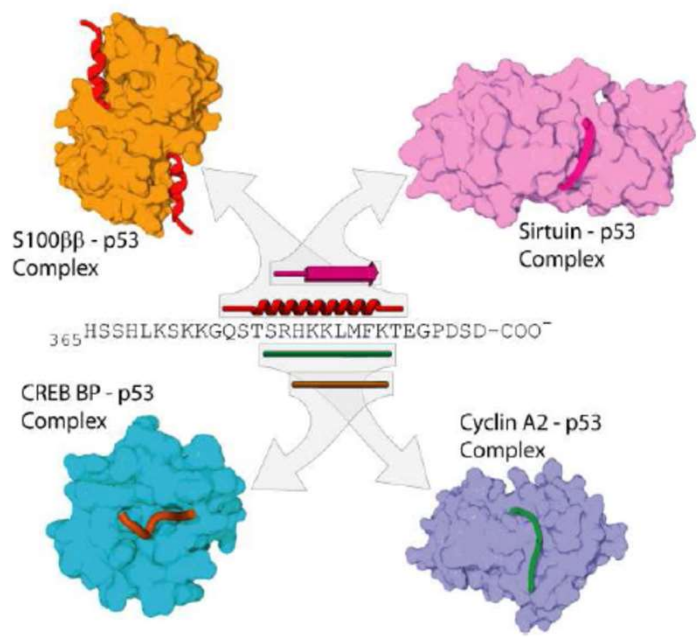
Unstructured regions



Frequently Involved in protein-protein interactions
In some cases they get structured upon binding



Honnappa S et al. J. Biol. Chem. 2006;281:16078-16083

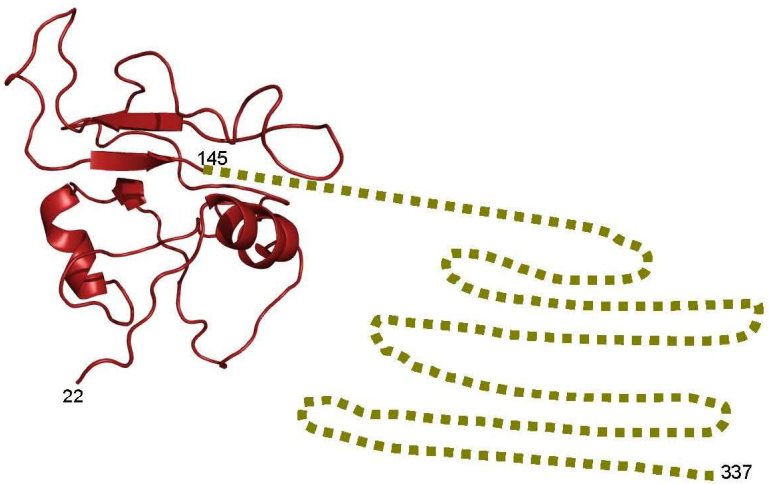


CJ Oldfield et al. BMC Genomics 2008 9(Suppl 1):S1

Prediction of unstructured regions

Why to predict them?

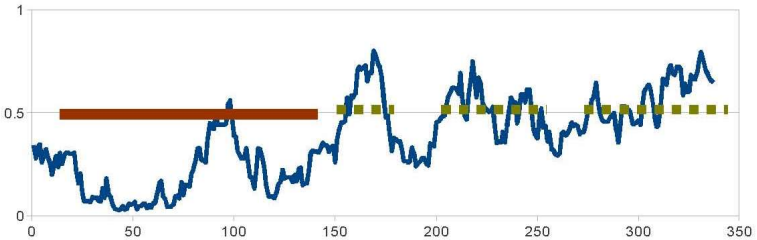
- Remove for crystallizing
- Might cause problems in sequence searches
- Map regions involved in transient interactions



<u>Compositionally biased regions</u>. *SEG*

<u>Specific for disorder.</u>

- *DISOPRED*
- *IUPRED*
- *ANCHOR* (for disorder involved in binding)
- …



• Wootton, J.C. and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Meth in Enzym*, **266**, 554-571

• Ward, J. J., McGuffin, L. J., Bryson K., Buxton, B. F. & Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20:**2138-2139.

• Dosztanyi, Z., Csizmok, V., Tompa, P. and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics*, **21**, 3433-3434.

• Dosztanyi, Z., Meszaros, B. and Simon, I. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins, *Bioinformatics*, **25**, 2745-2746.

1D Predictions
Other…

ExPASy Proteomics tools http://www.expasy.ch/tools/

COIL – Coiled-coil regions.
PSORT - prediction of signal proteins and localisation sites
SignalP - prediction of signal peptides

ChloroP - prediction of chloroplast peptides
NetOGlyc - prediction of O-glycosilation sites in mammalian proteins
Big-PI - prediction of glycosil -phosphatidyl inositol modification sites
DGPI - prediction of anchor and breakage sites for GPI

NetPhos - prediction of phosphorylation sites (Ser, Thr, Tyr) in eukaryotes
NetPicoRNA - prediction of cleavage sites for proteases in the picornavirus
NMT - prediction of N-miristoilation of N-terminals
Sulfinator - predicts sulphattation sites in tyrosines

$[\mathbf{a}bc\mathbf{d}efg]_n$

Lupas, A., Dyke, M.v. and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162-1164.

Protein Structure Prediction
3D Methods

- *Ab initio*

- Homology modelling/Comparative modelling.

- Fold recognition/
  Remote homology modelling/*threading*

3D Methods
Pure Ab Initio

Based on physico-chemical principles only (atom interaction energies, …)

Amino-acid sequence as only input (Anfinsen).

Interesting since they provide knowledge on the folding mechanism.

Purely ab-initio no usable for 3D prediction in general because:
• Empirical or semi-empirical interaction potentials with small inaccuracies that accumulate for large proteins and or long similations
• Require a lot of CPU power.

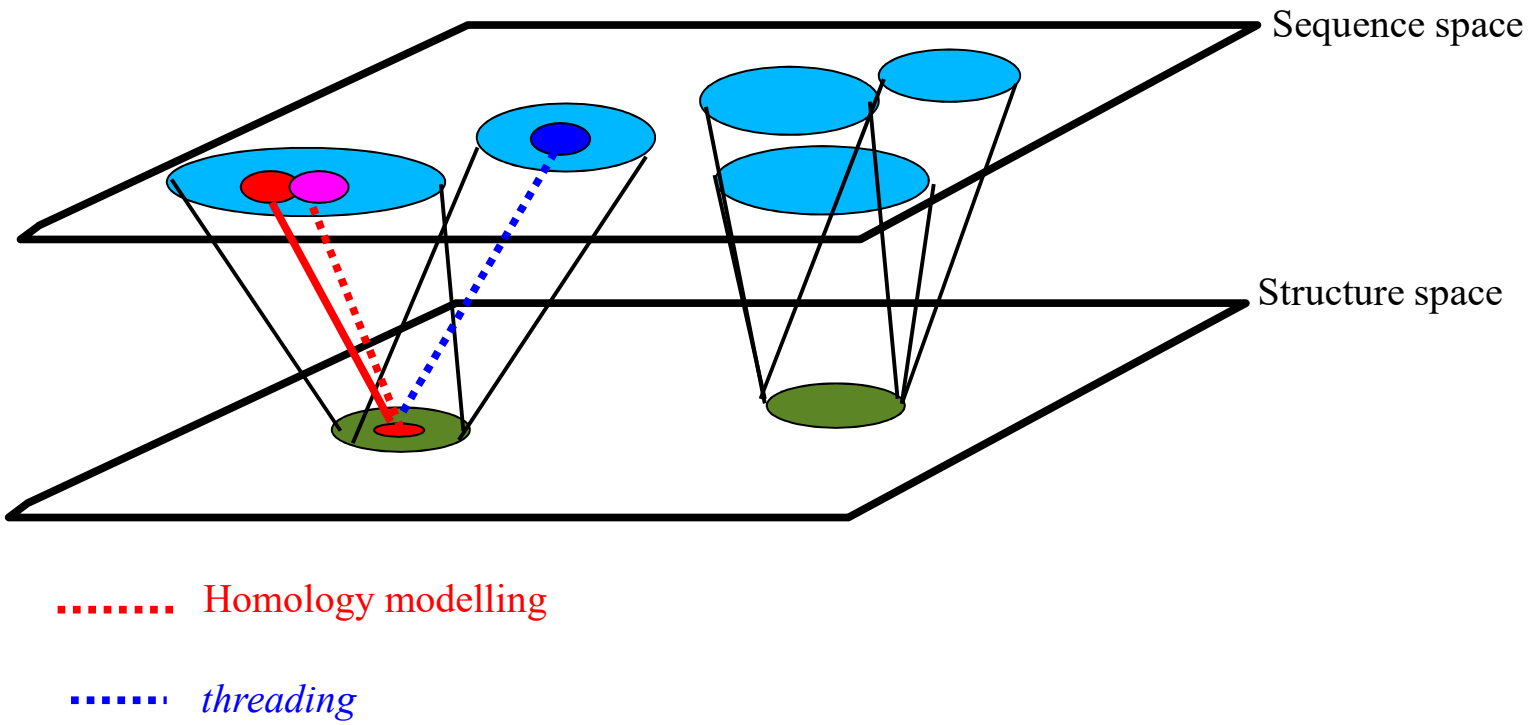
=> Practical utility for peptides or very short proteins

---

http://folding.stanford.edu/

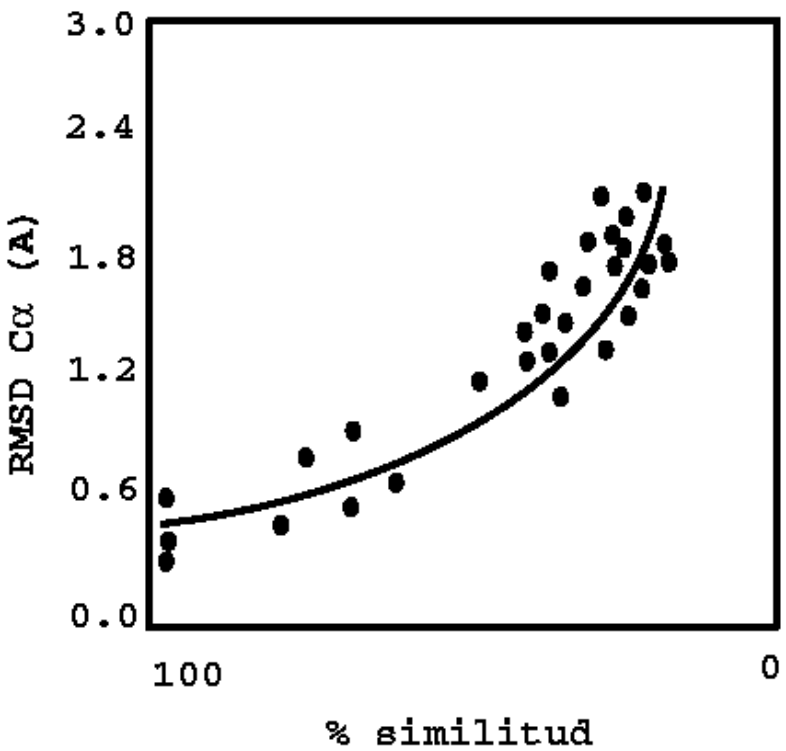
3D Methods

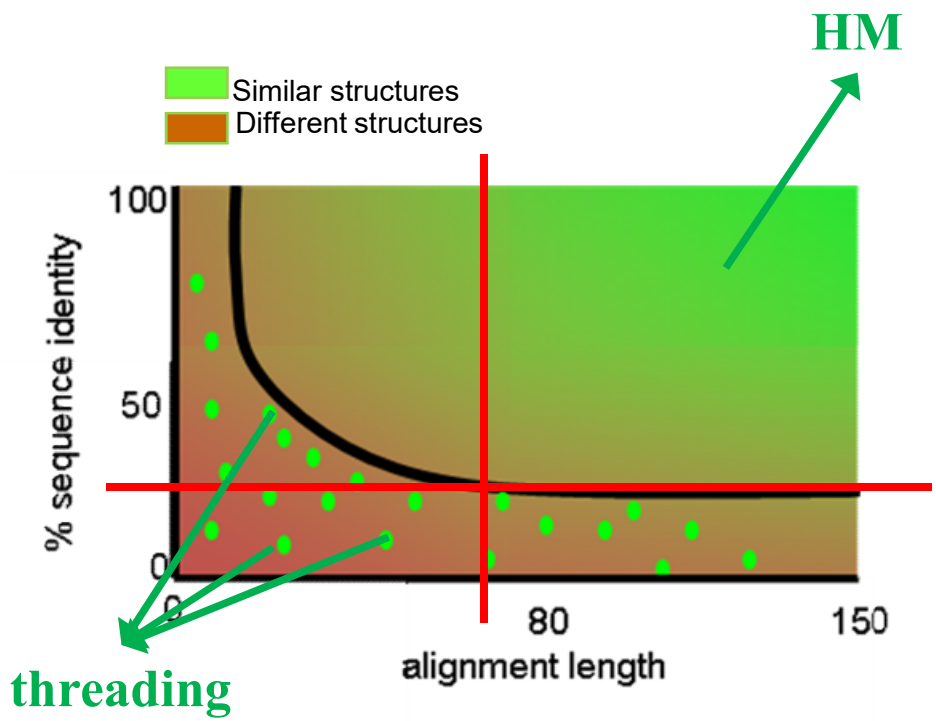
Homology modelling *vs. Threading*



Homology modeling

Based on the observation that similar sequences fold into the same (overall) structure





For a medium-length protein, a fair threshold for HM is around 25% seqid.

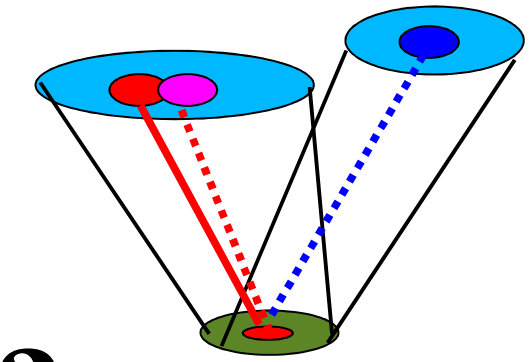
Chothia, C. & Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823-826.

Sander, C. & Schneider, R. (1993) The HSSP data base of protein structure-sequence alignments. *Nucleic Acids Res.*, **21**, 3105-3109.

Homology modeling – General strategy

- Locate template
- Generate alignment between sequences of target and template
- For backbone atoms take the coordinates of the corresponding template atoms
- For conserved residues between target and template take the atom coordinates for the side chains also.
- Side chains of other aa.
  - Use rotamer libraries
  - Take coordinates of as many as possible equivalent atoms (Cβ -> Cγ, ->…)
- Model loops (insertion/deletions)
- Optimize final structure (MD, ...)
- **Evaluate model**

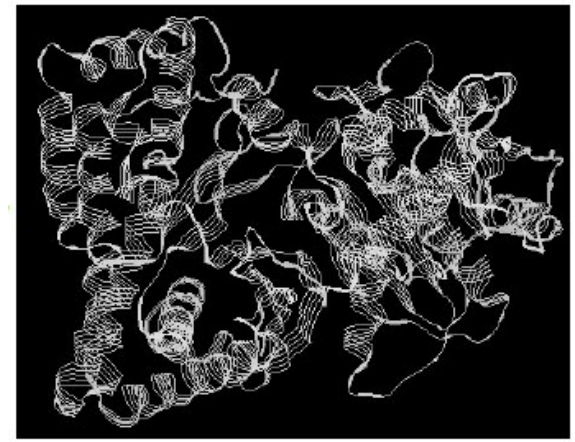
Homology Modelling
target-template alignment



Look for **solved structures** with sequences similar to our **target** (e.g. BLAST against PDB)

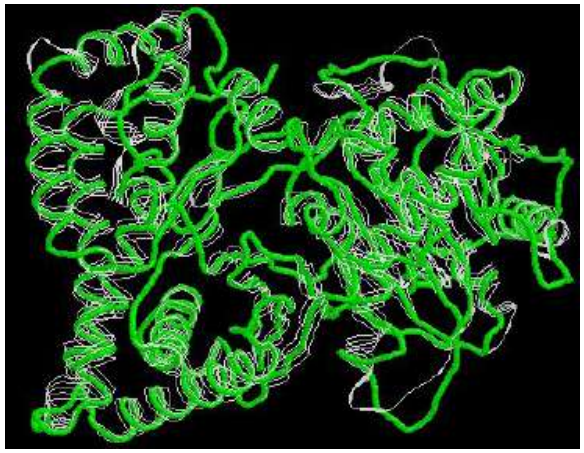
```
AHPLTSDFGGHTERDLHA   target
||  |||:  |||||  |:||
AHTLTSEGGGHTEADVHA   template
```

?



template (1ndb)

Homology modelling



model

Alignment: crucial step.

The %ID target-template is the most important *a-priori* indicator of the expected quality of your model

Homology Modelling
Public servers and model repositories

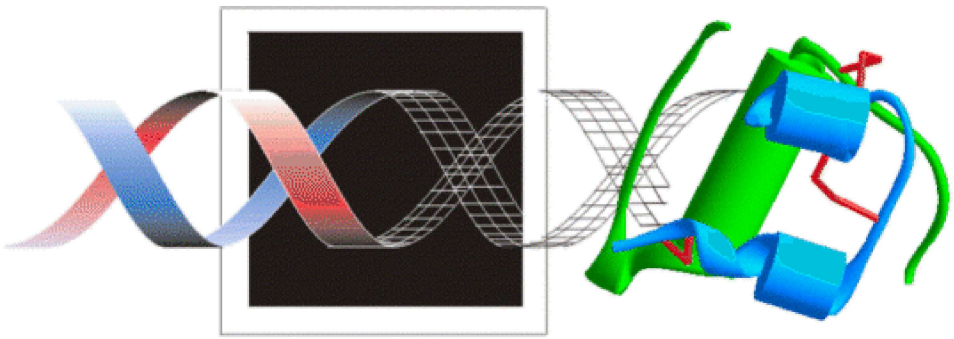
SWISS-MODEL - *www.expasy.ch/swissmod/SWISS-MODEL.html*
An automated comparative modelling server (ExPASy, CH)

CPHmodels - *www.cbs.dtu.dk/services/CPHmodels/*
Server using homology modelling (BioCentrum, Denmark)

SDSC1 - *cl.sdsc.edu/hm.html*
Protein structure homology modeling server (San Diego, USA)

3D-JIGSAW - *www.bmm.icnet.uk/servers/3djigsaw/*
Automated system for 3D models for proteins (Cancer Research UK)

There are public web servers for modeling by homology a given sequence, as well as repositories of pre-generated models



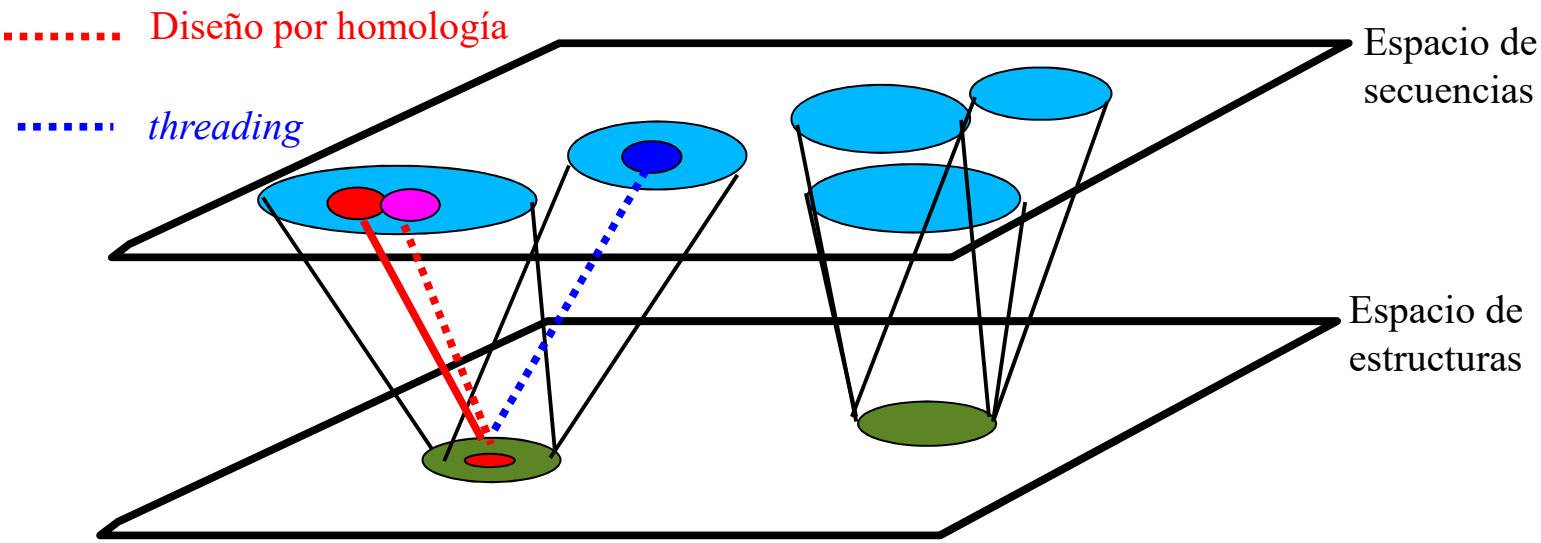
http://www.expasy.ch/swissmod/SM_3DCrunch.html

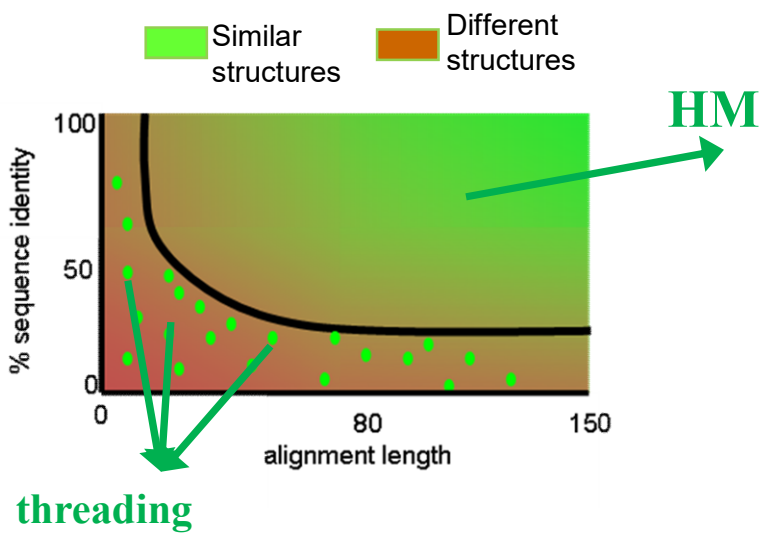
MODBASE

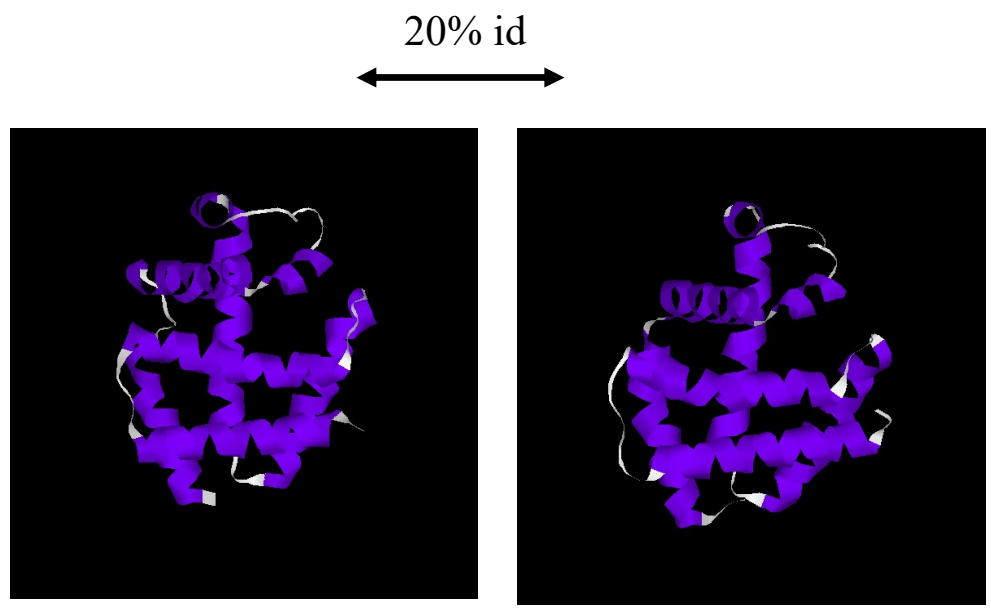
Database of Comparative Protein Structure Models

http://pipe.rockefeller.edu/modbase

3D Methods
Homology modeling *vs. Threading*

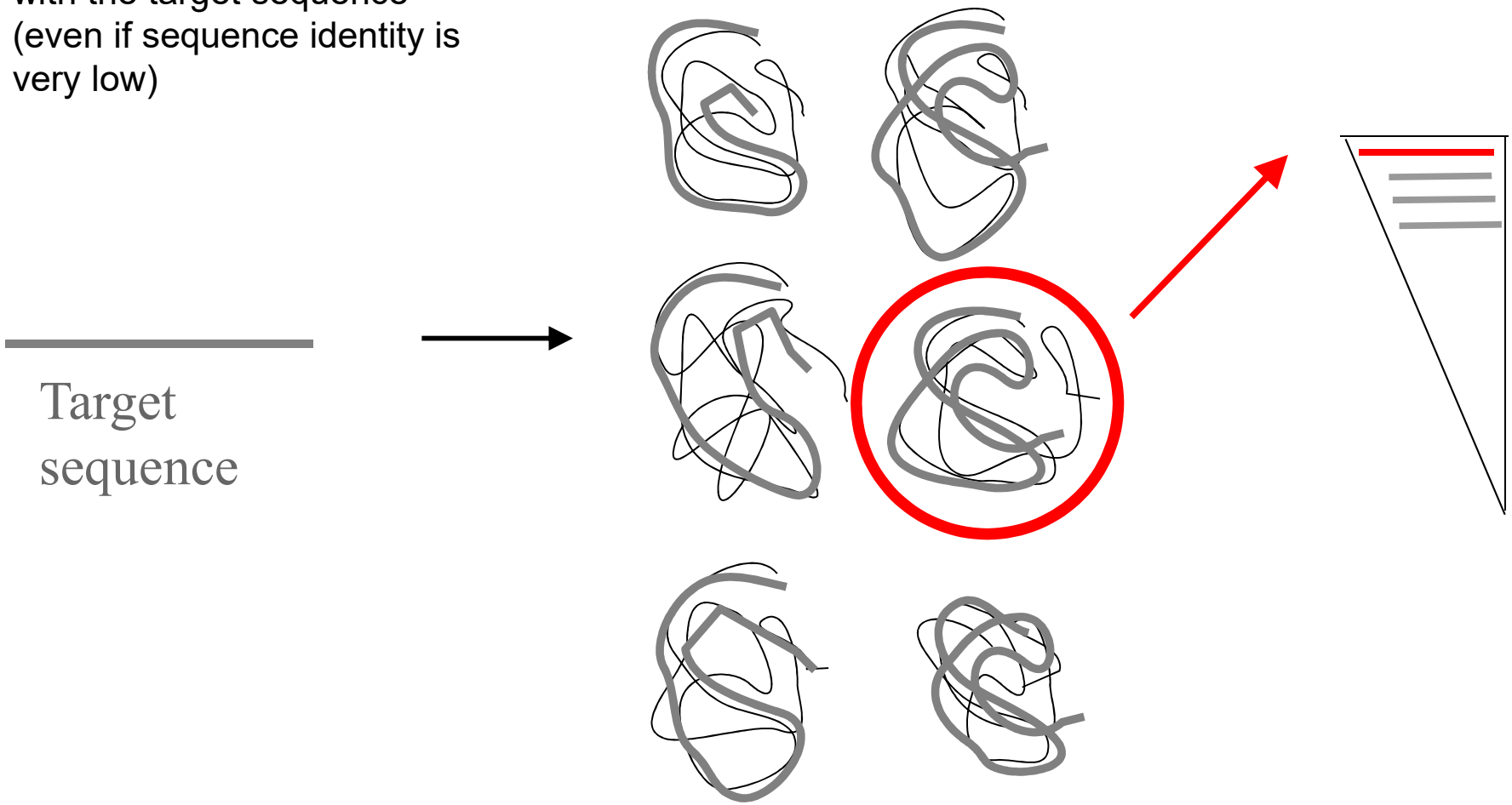






Threading. General Strategy

Look for folds “compatible” with the target sequence (even if sequence identity is very low)



Put (*thread*) the target sequence in the structures contained in a fold library (possible templates) and evaluate in which one it “fits best”, scoring these target-template matches by….

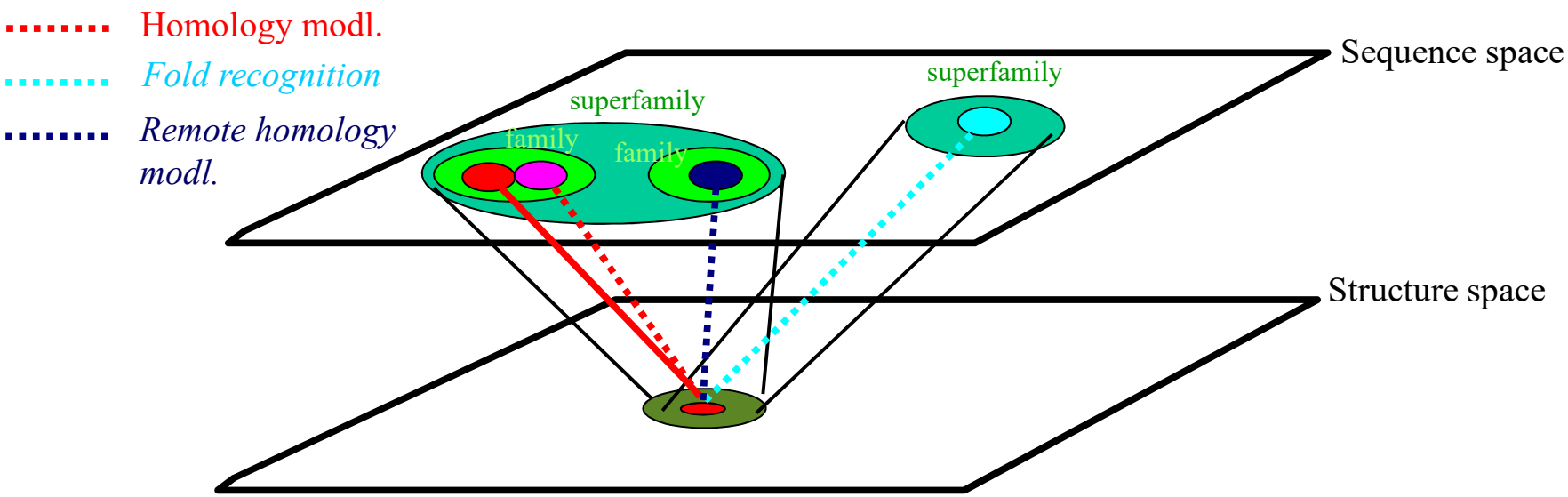
Threading
Template search/scoring

- Amino-acids in similar environments as they are in known structures (*pair potentials*)

- Solvatation potentials.

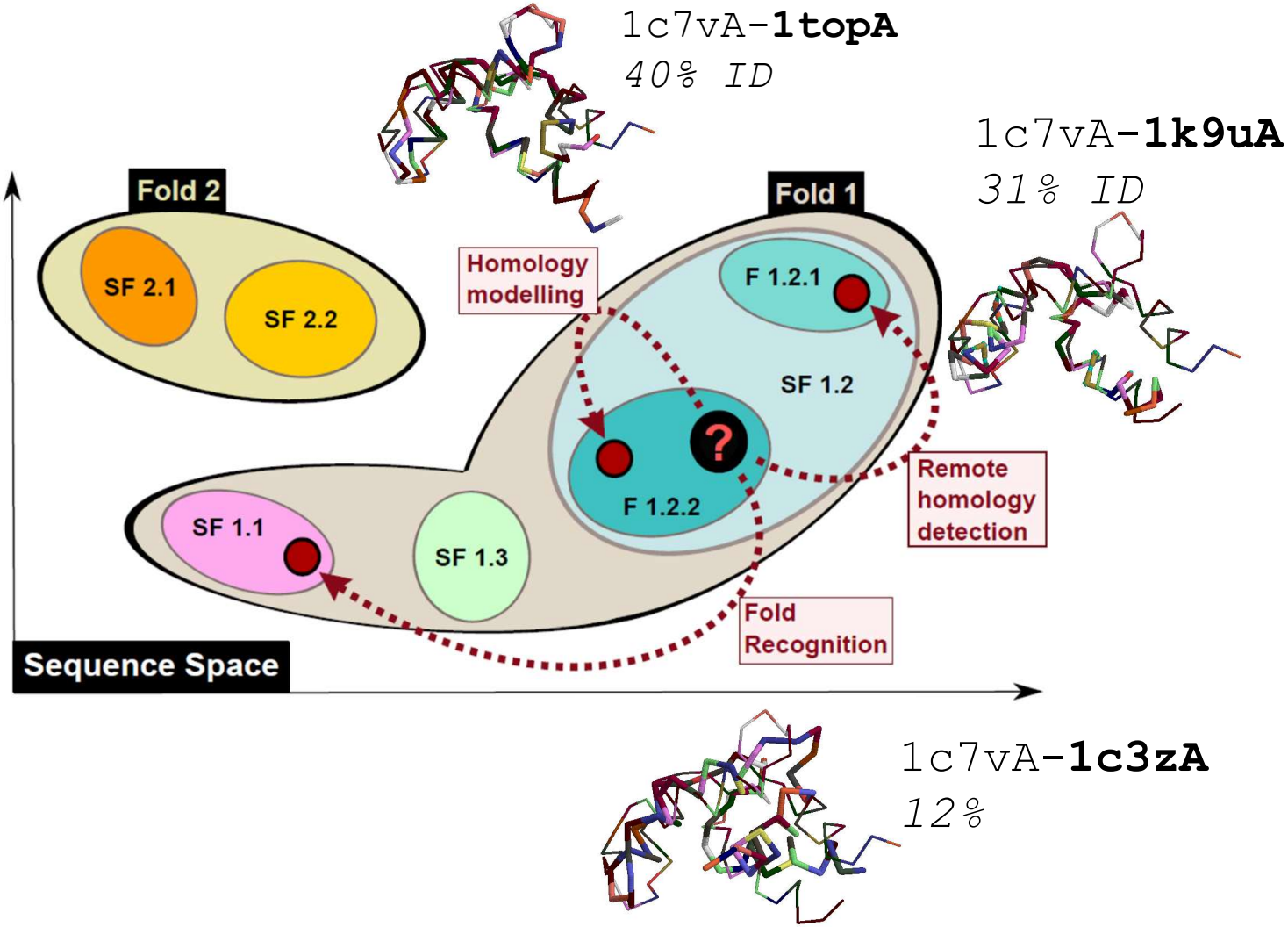
- Matching of secondary structures (predicted – real (template)) and accessibilities

- Remote homolog detection. Using profiles, HMMs (HHPRED)



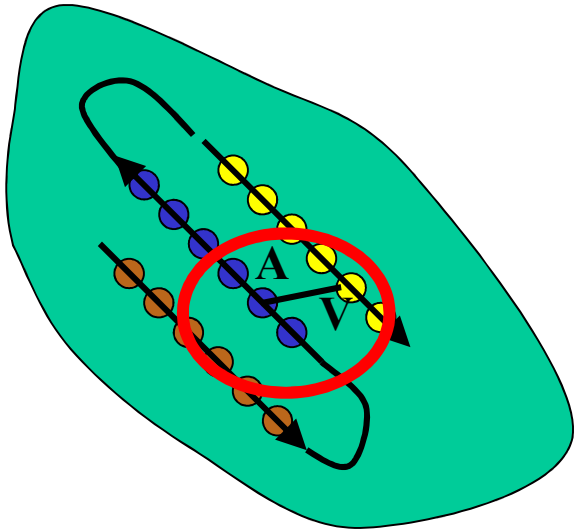
Söding J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21**, 951-960.

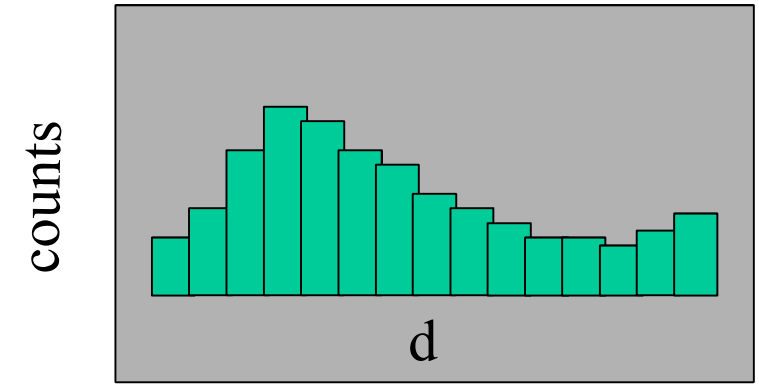
Threading
Template search/scoring



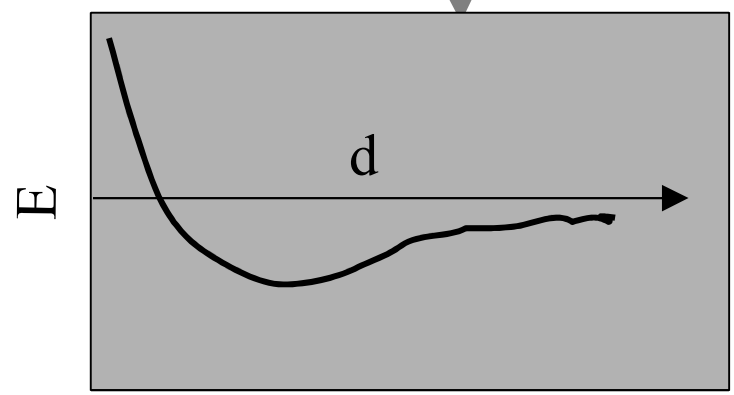
Pazos, F. and Sanchez-Pulido, L. (2014) Protein Superfamilies. In, *eLS*. John Wiley & Sons, Ltd, Chichester, DOI: 10.1002/9780470015902.a9780470025587.

Threading
Example – pair potentials





Energy of interaction =
-KT ln (frequency of interactions)
Boltzmann principle



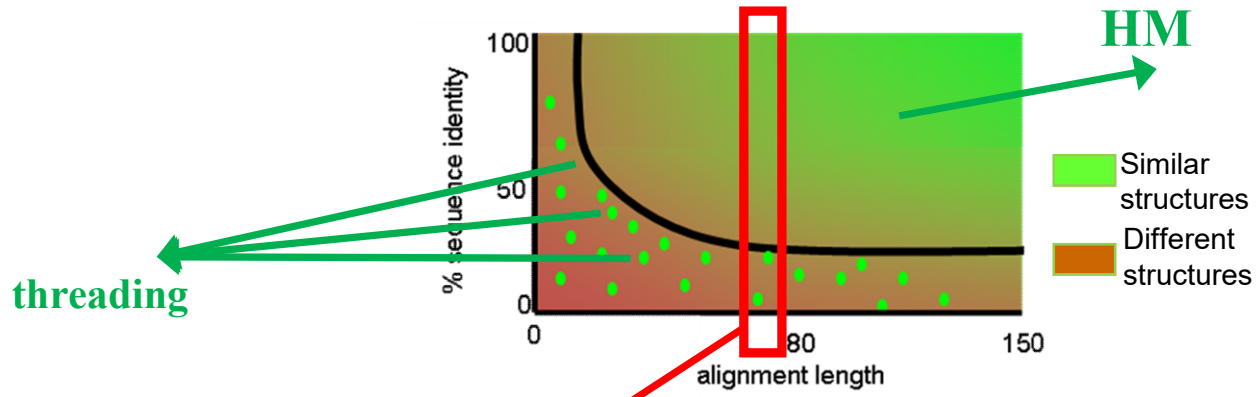
For a given residue pair, count instances at different separations

---

Jones, D., Taylor, W. and Thornton, J. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86-89.

Sippl, M.J. (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol*, **5**, 229-235.

Template-based modelling
Range of applicability and expected model quality



% id seq. con alguna estructura

0 30 50 70 100

threading
Remote homology modelling

modelado por homología

Fold ok

nivel atómico

calidad modelos

Detalles atomicos mal

RMSD ~1.0A
- **cadenas laterales**
- **loops**

RMSD ~3.5A
- **diferencia en loops y gaps**
- **movimientos de dominios**
- **cambios en el backbone**

Protein Structure prediction
Current (CASP) classification of methods

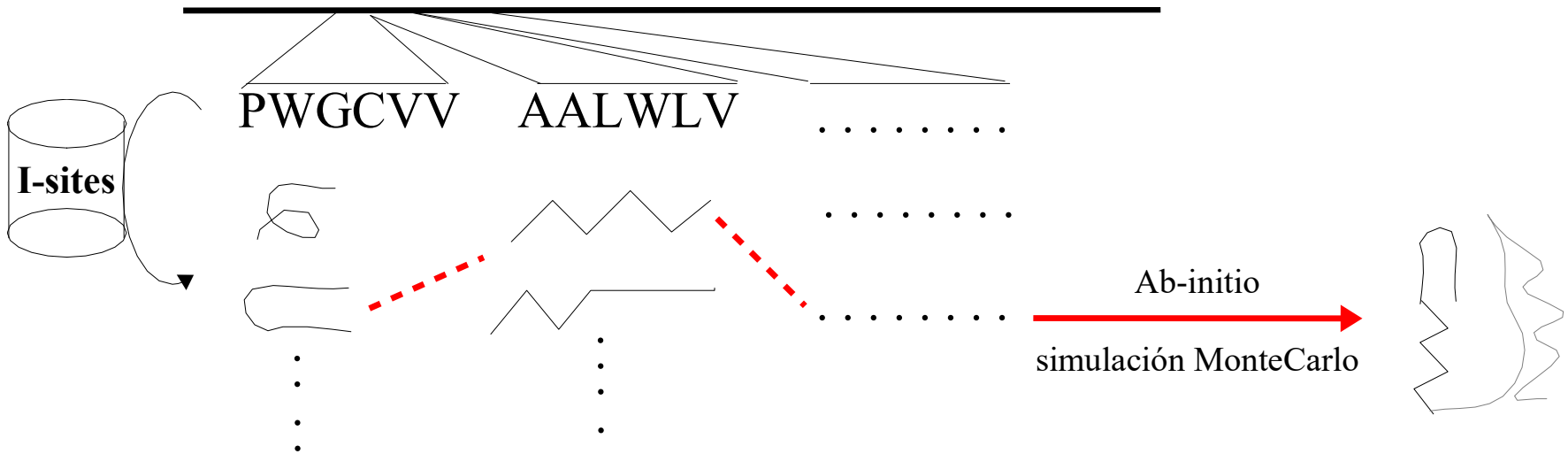
Protein structural "level"	secondary	--------	tertiary	quaternary
Protein representation	1D	2D	3D	4D
				
Use aa sequence alone?				
Ab Initio	Secondary structure prediction	Contact prediction	- Molecular dynamics - Energy minimisation	*docking*
No Ab-Initio	Secondary structure prediction		- Homology modeling - *Threading*	*docking with restraints*

Newer methods are hybrid approaches, taking concepts and methods from all these

TBM (whole-template based modelling)	Vs.	FM (free modelling)

CASP classification. For targets more than for methods

3D Methods
Fragment-based
mini-threading + Ab-initio
Rosetta



Mini-threading of small fragments of the target sequence (~10aa). Generation of the final model by exploring combinations of these small fragment models.

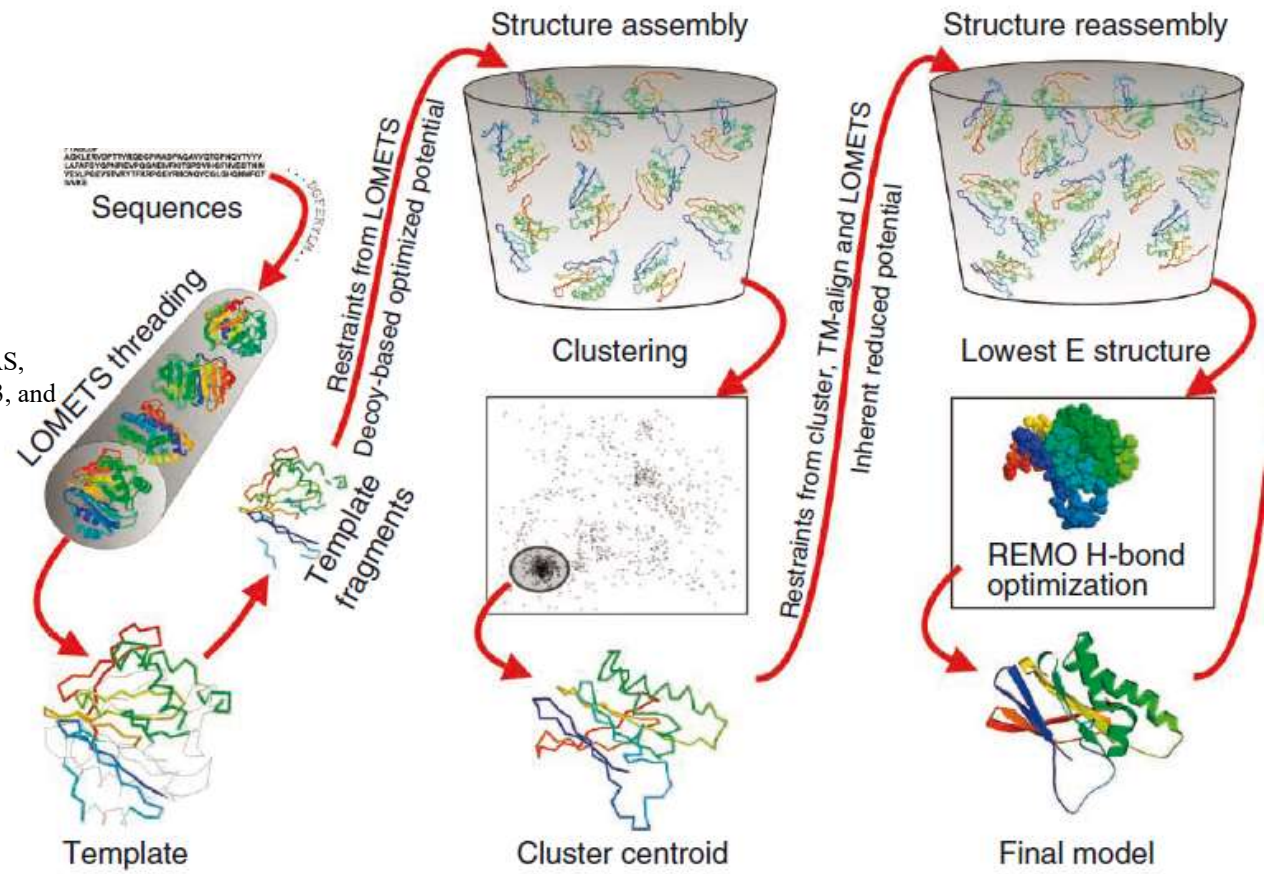
http://robetta.bakerlab.org/

Simons, K.T., Kooperberg, C., Huang, E. & Baker, D. (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, **268**, 209-225.

3D Methods - Fragment-based - Meta-method + Ab-initio

I-TASSER

FFAS-3D, HHsearch, MUSTER, pGenTHREADER, PPAS, PRC, PROSPECT2, SP3, and SPARKS-X



Combination of threading fragments of (overlapping) segments of the target sequence of different lengths. Clustering of models to look for overrepresented folds. Filtering by different constraints and optimization (energy minimization) of the final model(s).

http://zhanglab.ccmb.med.umich.edu/I-TASSER/

J Yang, R Yan, A Roy, D Xu, J Poisson, Y Zhang. (2015). The I-TASSER Suite: Protein structure and function prediction. *Nature Methods*, **12:** 7-8

3D Methods
Additional model filtering

Many approaches produce more than one alternative model. Filter then using any available information you might have on your particular protein.

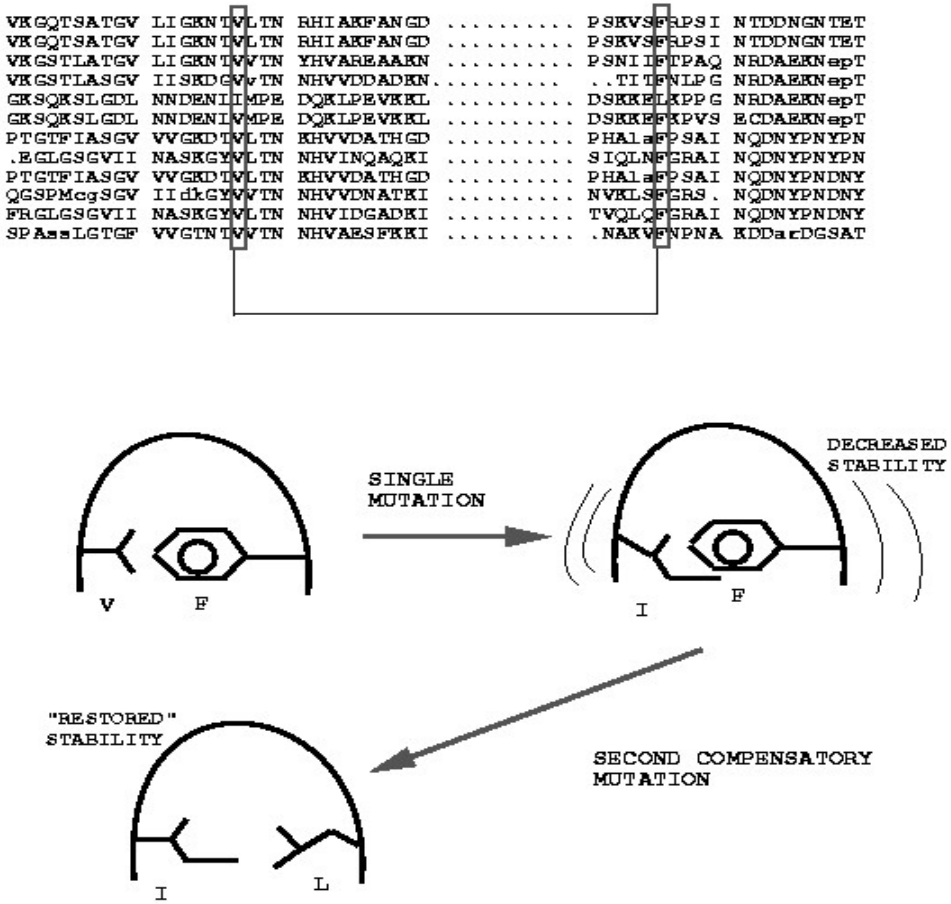
E.g.

- Distance constraints (experimental or predicted): NMR data, crosslinking, residue co-evolution (incorporated in some methods), …
- Functional information: function of the proposed fold, position of functional residues in the model, …

Proteins biannual issues on CASP (Critical assessment of protein structure prediction methods)

Juan, D., Pazos, F. and Valencia, A. Emerging methods in protein co-evolution. *Nat Rev Genet 2013;**14(4)**:249-261.*

Co-evolving positions as distance constraints



Until now… very low reliability. Useful (selecting among models or docking poses…) but not enough to predict 3D structure.

Göbel, U., Sander, C., Schneider, R. and Valencia, A. (1994) Correlated mutations and residue contacts in proteins, *Proteins*, **18**, 309-317.

Olmea, O. and Valencia, A. (1997) Improving contact predictions by the combination of correlated mutations and other sources of sequence information., *Fold Des*, **2**, S25-S32.

Florencio Pazos, Manuela Helmer Citterich, Gabriele Ausiello and Alfonso Valencia (1997). Correlated Mutations Contain Information About Protein-Protein Interaction. *Journal of Molecular Biology*. **271(4):**511-523.

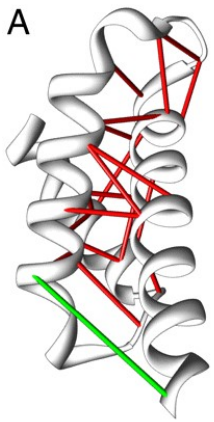
Correlated mutations - “New wave” methods

Evocouplings (C. Sander), *DCA* (M. Weigt), *PSICOV* (D. Jones)

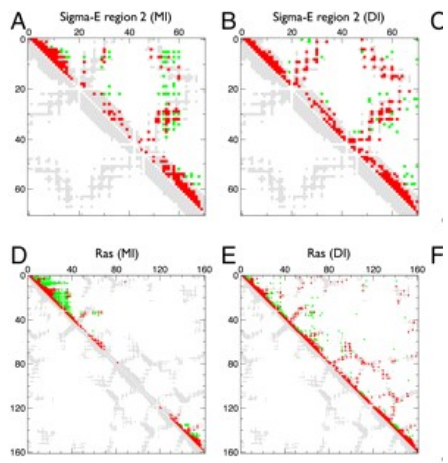
- Methodologies able to disentangle indirect correlations
- **Many more SEQUENCES**

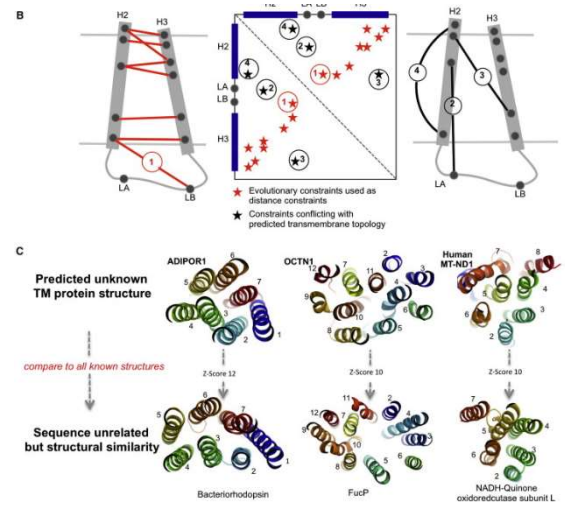
Accuracy of predicted contacts increased orders of magnitude!

“protein folding problem solved’, according with authors. A llimited number of reliable correlated pairs used as constraints for MD.
Models <= 2A RMSD



top DI pairs





Morcos, F. *et al.* (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl Acad. Sci. USA* 108, E1293–E1301.

Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28, 184–190.

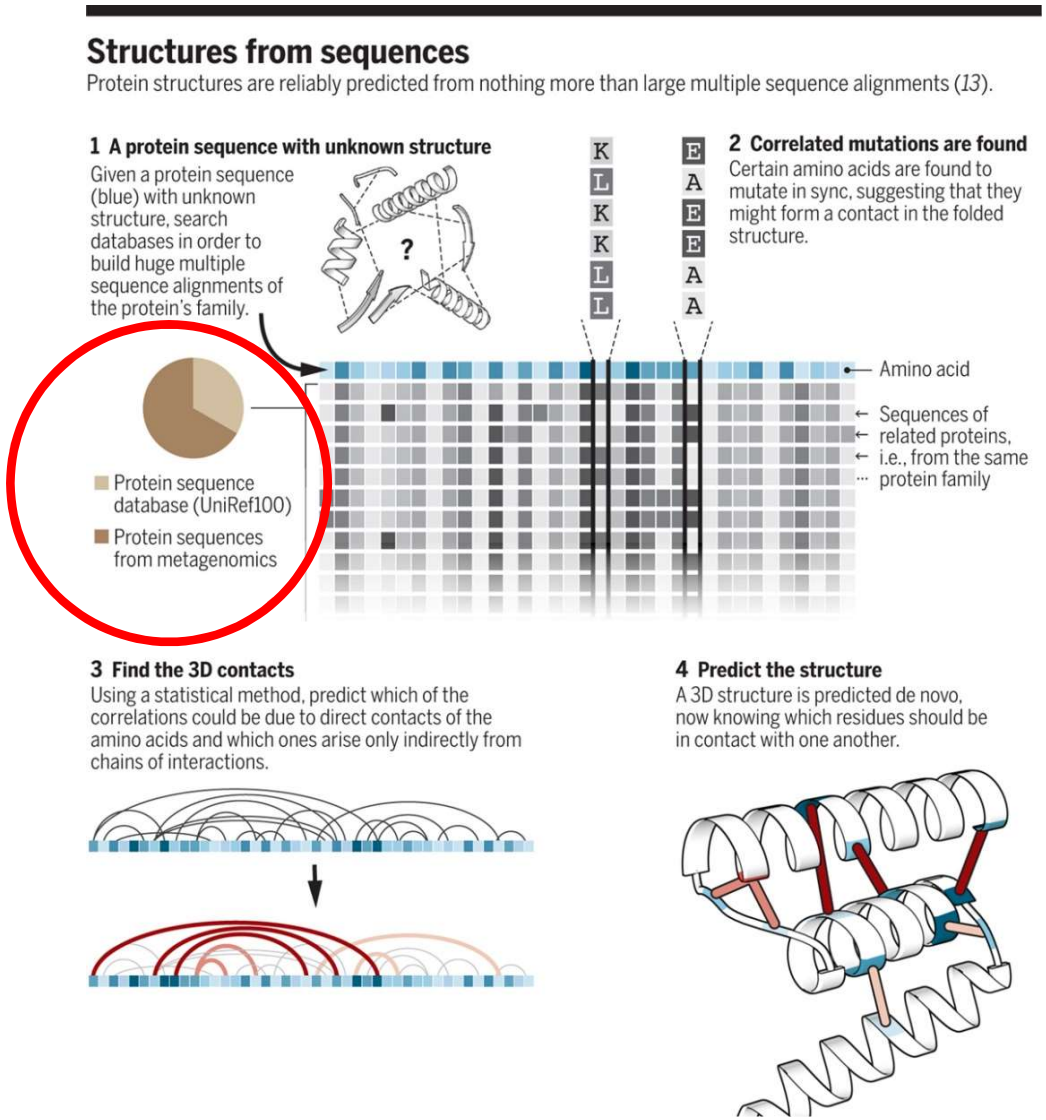
Hopf *et al.* (2012). Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell*. **149(7):**1607–1621

Correlated mutations - “New wave” methods

Problem: They need MANY homologous sequences. In the order of thousands. Available only for a limited number of protein families

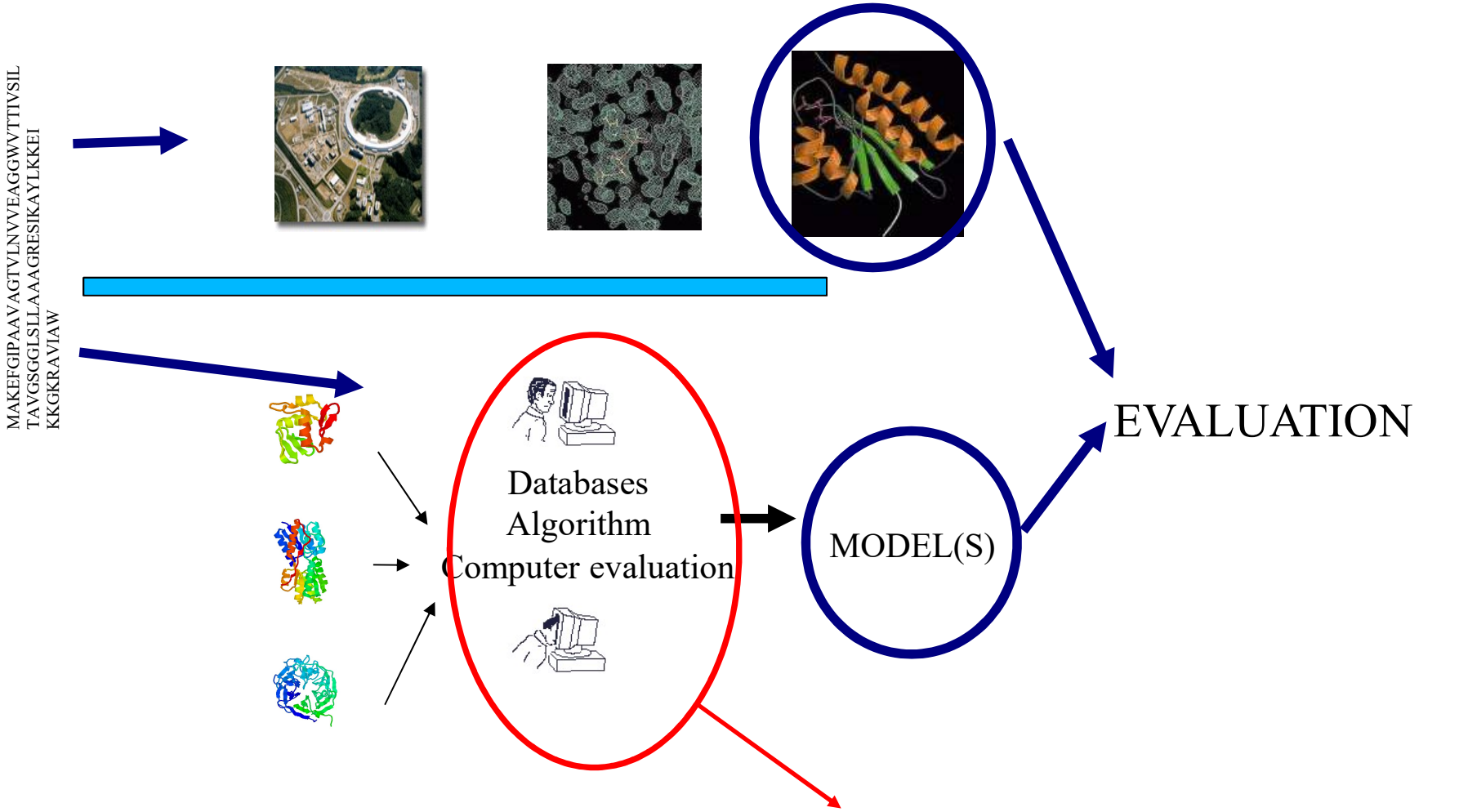
Possible solutions:

- Metagenomic sequences
- “A la carte” re-sequencing



Johannes Söding (2017). Big-data approaches to protein structure prediction. *Science.* 355(6322):248-249.

Assessment of prediction methods
CASP (bianual 94-)



Also section for fully automatic servers
(CAFASP / ROLL)

---

http://predictioncenter.org/

Proteins biannual issues on CASP/CAFASP

General Bibliography

• Thomas E. Creighton. Proteins: Structures and Molecular Properties. W. H. Freeman (ed); 2 Sub edition (August 15, 1992). ISBN-10: 071677030X

• Gregory A. Petsko & Dagmar Ringe. Protein Structure and Function. Sinauer Associates (eds) (January 2004). ISBN-10: 0878936637

• Florencio Pazos & Mónica Chagoyen. Practical Protein Bioinformatics. Springer. (January 2015). ISBN 978-3-319-12727-9

Structural Bioinformatics

Florencio Pazos

Computational Systems Biology Group (CNB-CSIC)
pazos@cnb.csic.es
http://csbg.cnb.csic.es

http://csbg.cnb.csic.es/Courses/UMA_BIF_2018/