Introduction to protein structure analysis and prediction

Mónica Chagoyen monica.chagoyen@cnb.csic.es

Protein sequence analysis and prediction service

Centro Nacional de Biotecnologia (CNB-CSIC)

24-26 October 2011

Course organization and contents

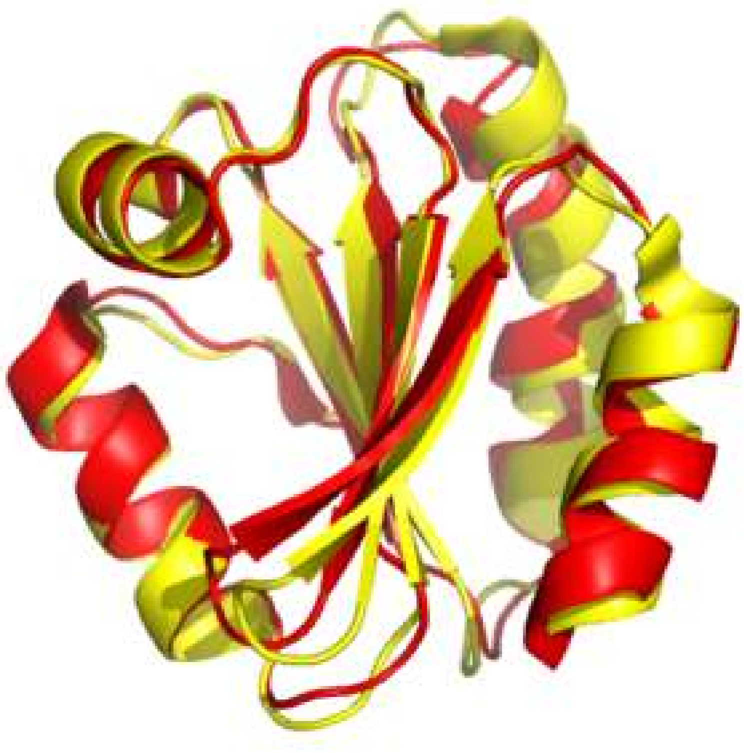
Day 1:
The protein structure universe, resources and visualization

Day 2:
Structural alignment, classification and 1D prediction

Day 3:
3D structure prediction

Structural alignment

Structural alignment



Establishing equivalences between amino acid residues based on the 3D structure of two or more protein folds

No prior knowledge of what amino acids are equivalent

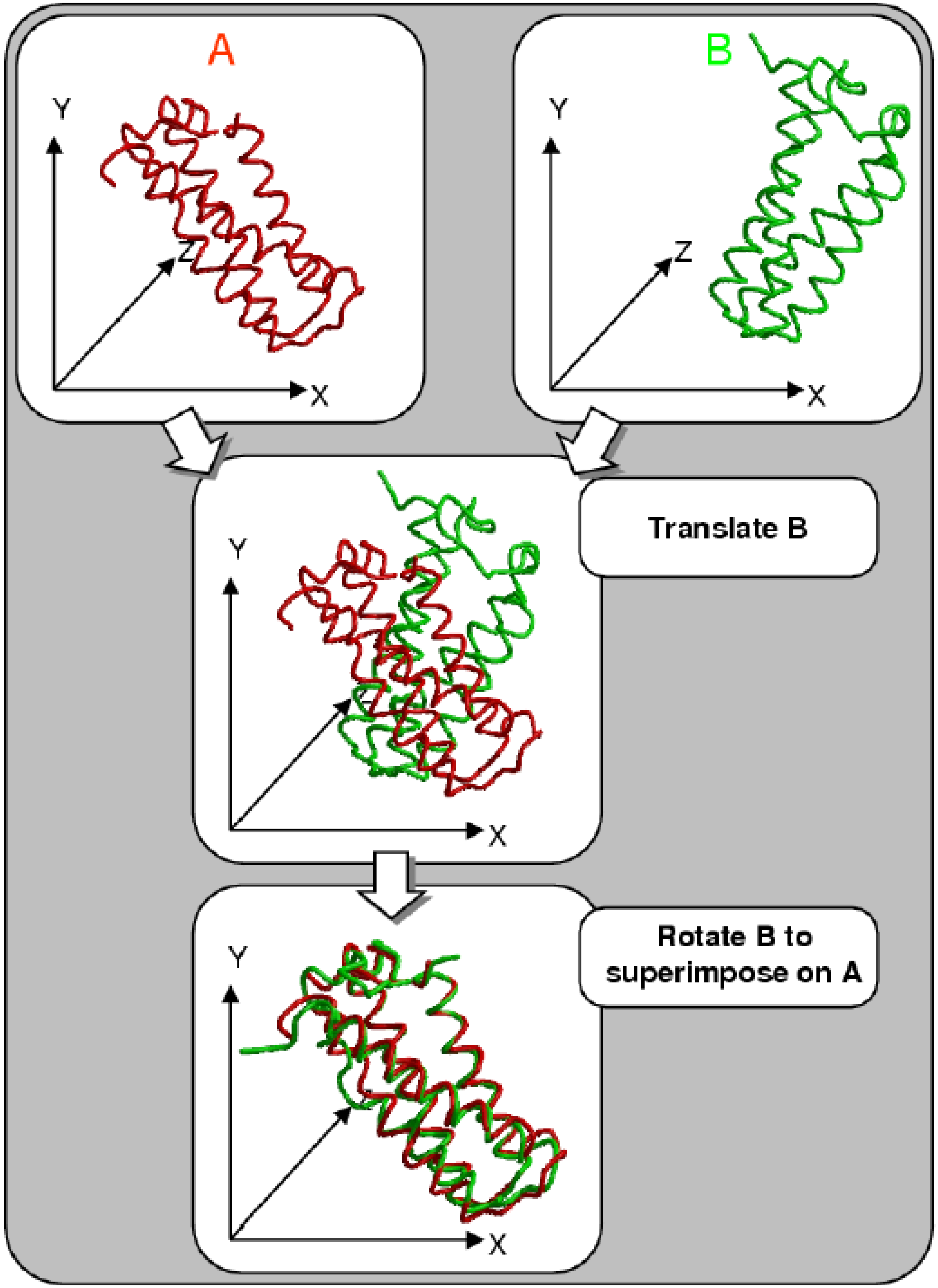
Rigid body superposition

Steps

1. Represent proteins A and B (backbone)

2. Rotate & translate B

3. Score the alignment



Translate B

Rotate B to superimpose on A

Image Dr. Alison Cuff
http://www.bioinf.org.uk/teaching/c40/pdf/ali_lecture.pdf

Scores

H Hasegawa and L Holm
Advances and pitfalls of protein structural alignment
Current Opinion in Structural Biology 2009, 19:341–348

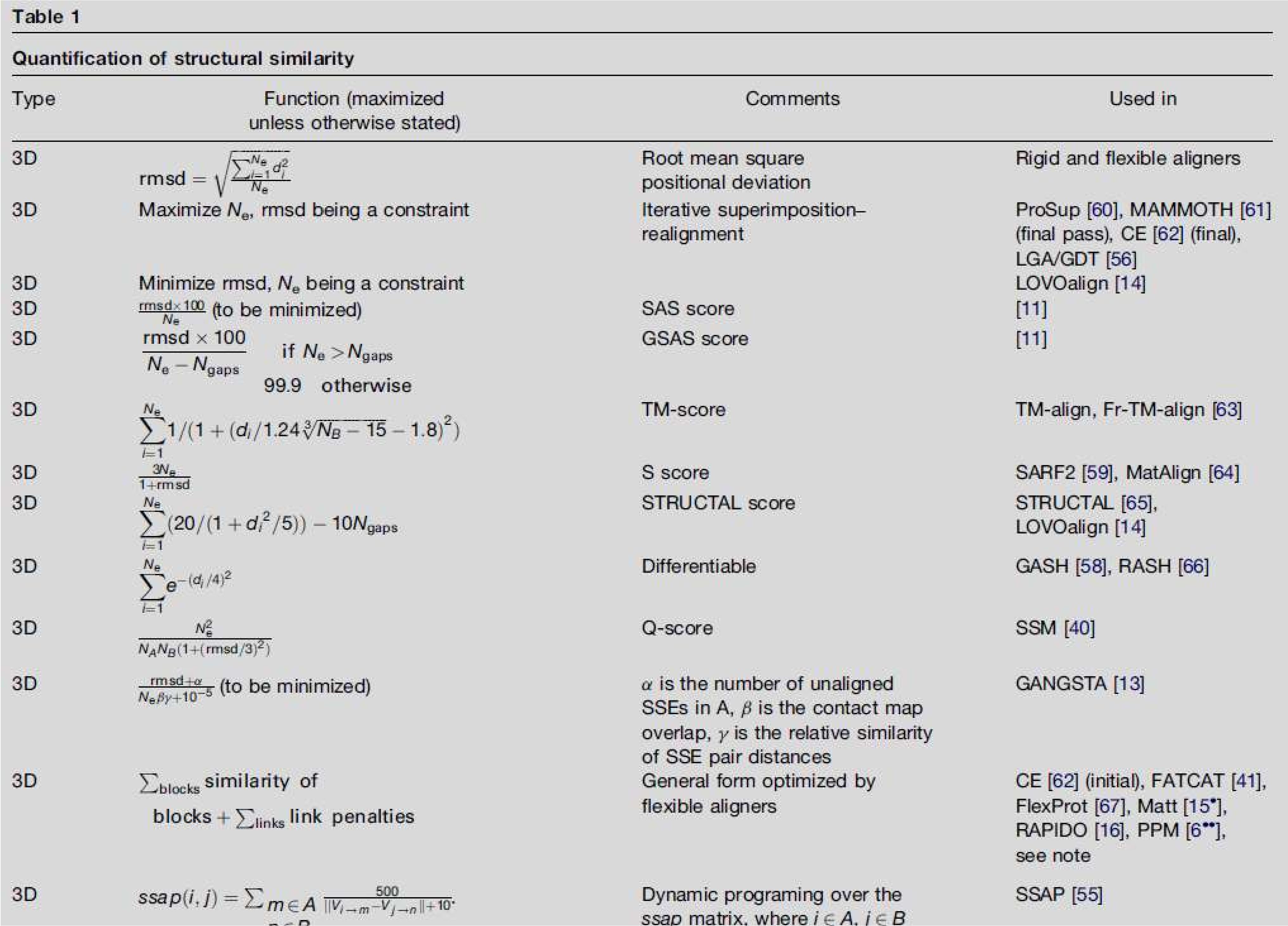
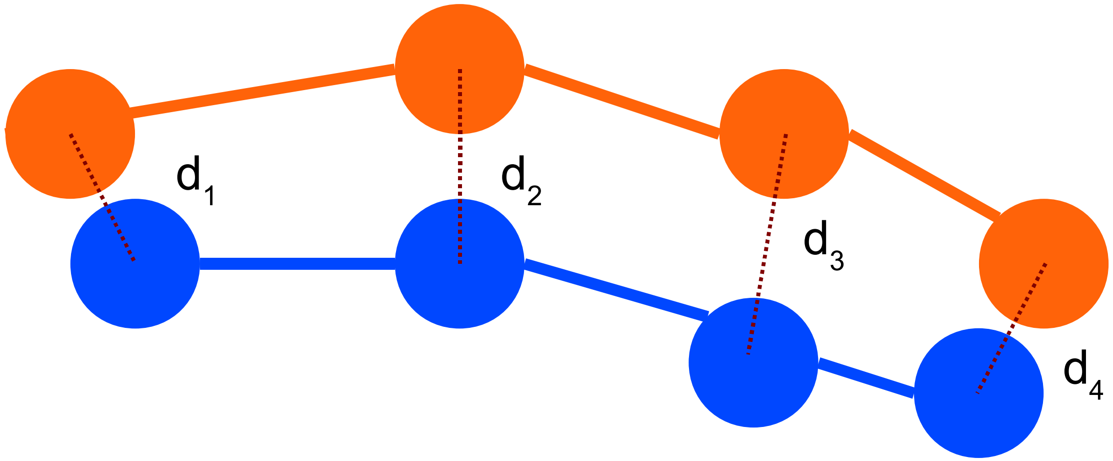


Table 1

Quantification of structural similarity

Type	Function (maximized unless otherwise stated)	Comments	Used in
3D	$\mathrm{rmsd} = \sqrt{\frac{\sum_{i=1}^{N_e} d_i^2}{N_e}}$	Root mean square positional deviation	Rigid and flexible aligners
3D	Maximize N_e, rmsd being a constraint	Iterative superimposition–realignment	ProSup [60], MAMMOTH [61] (final pass), CE [62] (final), LGA/GDT [56]
3D	Minimize rmsd, N_e being a constraint		LOVOalign [14]
3D	$\frac{\mathrm{rmsd}\times 100}{N_e}$ (to be minimized)	SAS score	[11]
3D	$\frac{\mathrm{rmsd}\times 100}{N_e - N_{\mathrm{gaps}}}$ if $N_e > N_{\mathrm{gaps}}$ 99.9 otherwise	GSAS score	[11]
3D	$\sum_{i=1}^{N_e} 1/(1+(d_i/1.24\sqrt[3]{N_B-15}-1.8)^2)$	TM-score	TM-align, Fr-TM-align [63]
3D	$\frac{3N_e}{1+\mathrm{rmsd}}$	S score	SARF2 [59], MatAlign [64]
3D	$\sum_{i=1}^{N_e}(20/(1+d_i^2/5)) - 10N_{\mathrm{gaps}}$	STRUCTAL score	STRUCTAL [65], LOVOalign [14]
3D	$\sum_{i=1}^{N_e} e^{-(d_i/4)^2}$	Differentiable	GASH [58], RASH [66]
3D	$\frac{N_e^2}{N_A N_B(1+(\mathrm{rmsd}/3)^2)}$	Q-score	SSM [40]
3D	$\frac{\mathrm{rmsd}+\alpha}{N_e\beta\gamma+10^{-5}}$ (to be minimized)	α is the number of unaligned SSEs in A, β is the contact map overlap, γ is the relative similarity of SSE pair distances	GANGSTA [13]
3D	$\sum_{\mathrm{blocks}}$ similarity of blocks $+ \sum_{\mathrm{links}}$ link penalties	General form optimized by flexible aligners	CE [62] (initial), FATCAT [41], FlexProt [67], Matt [15•], RAPIDO [16], PPM [6••], see note
3D	$ssap(i,j) = \sum_{m\in A, n\in B} \frac{500}{\lVert V_{i\to m} - V_{j\to n}\rVert + 10}$	Dynamic programing over the *ssap* matrix, where $i \in A$, $j \in B$	SSAP [55]

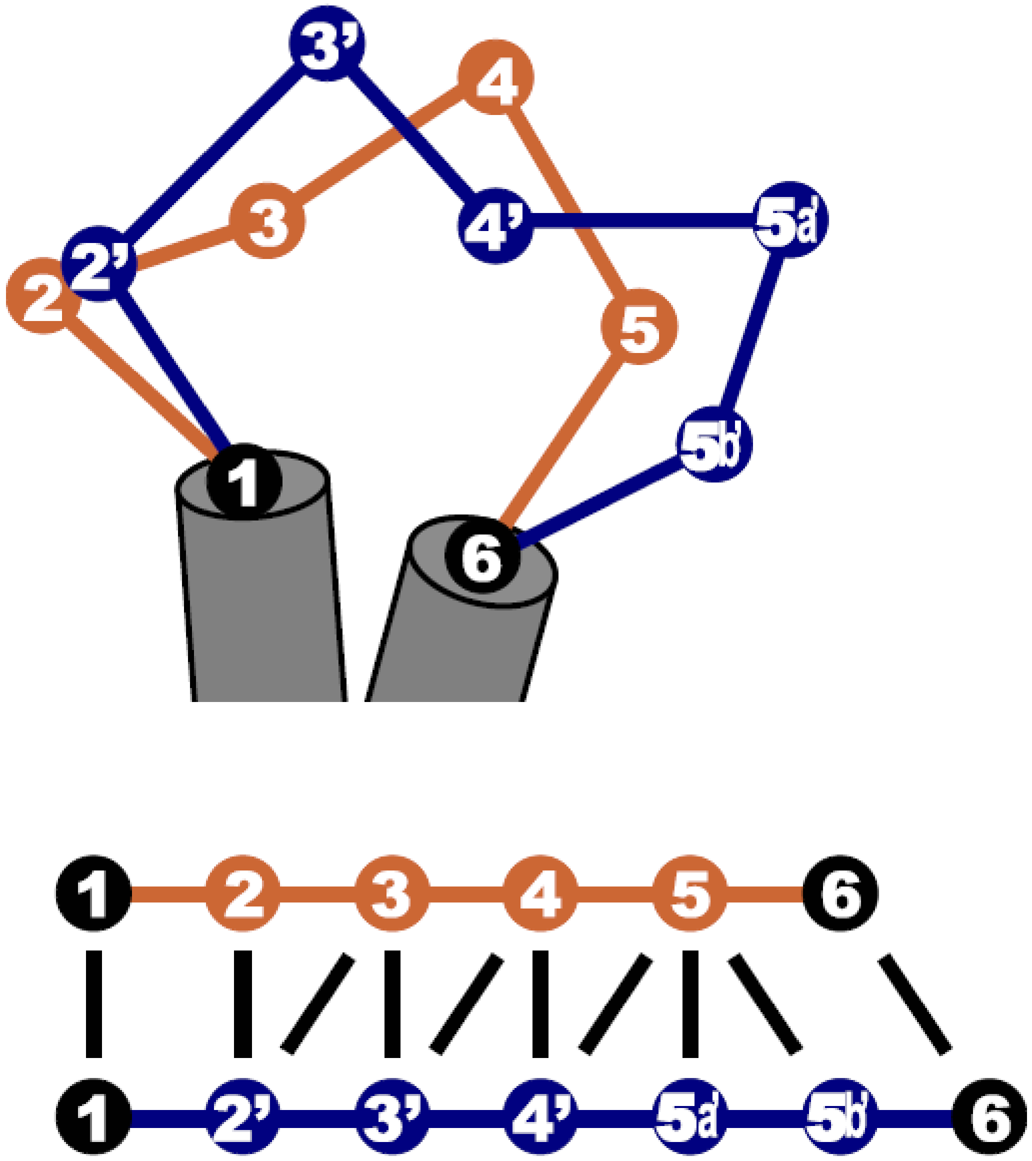
r.m.s.d (root mean square deviation)



Ambiguity

A single insertion (5') can lead to ambiguity in the pairwise residue alignment between the loops.

Therefore, a simple one-to-one functional equivalence between residues from different proteins may not exist.



W Pirovano, KA Feenstra, J Heringa
The meaning of alignment: lessons from structural diversity
BMC Bioinformatics 2008, 9:556

Software for structural alignment

Pair-wise and database searches

Dali
http://ekhidna.biocenter.helsinki.fi/dali_server

CE (Combinatorial Extension)
http://cl.sdsc.edu/

SSAP (CATH database)
http://www.cathdb.info (select Tools)

PDBefold
http://www.ebi.ac.uk/msd-srv/ssm

jFATCAT-rigid algorithm
PDB www.pdb.org (all-against-all PDB, 40% sequence similarity clustering)

Multiple structure alignment

Mammoth-Mult
http://ub.cbm.uam.es/software/online/mamothmult.php

MultiProt
http://bioinfo3d.cs.tau.ac.il/MultiProt/

SuperPose
http://wishart.biology.ualberta.ca/SuperPose/

MUSTANG
(for download)

Flexible alignments

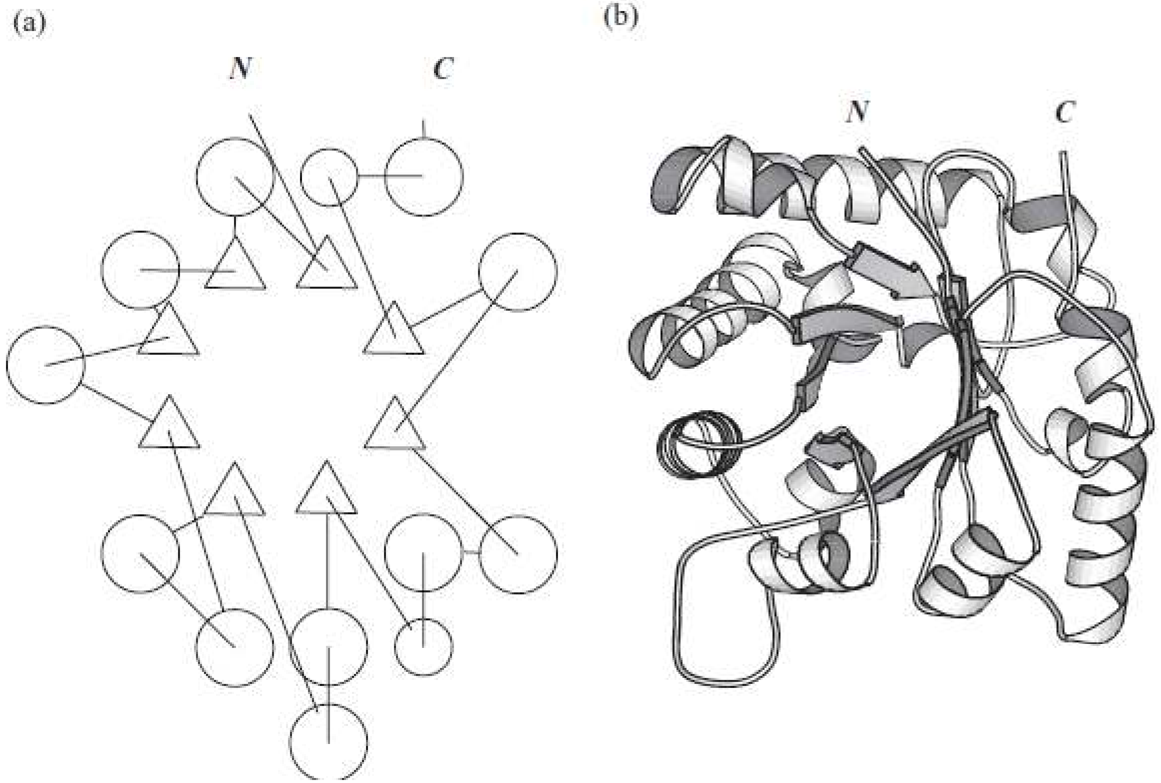
FATCAT
http://fatcat.burnham.org/

RAPIDO
http://webapps.embl-hamburg.de/rapido/

FlexProt
http://bioinfo3d.cs.tau.ac.il/FlexProt/ (only PDB ids)

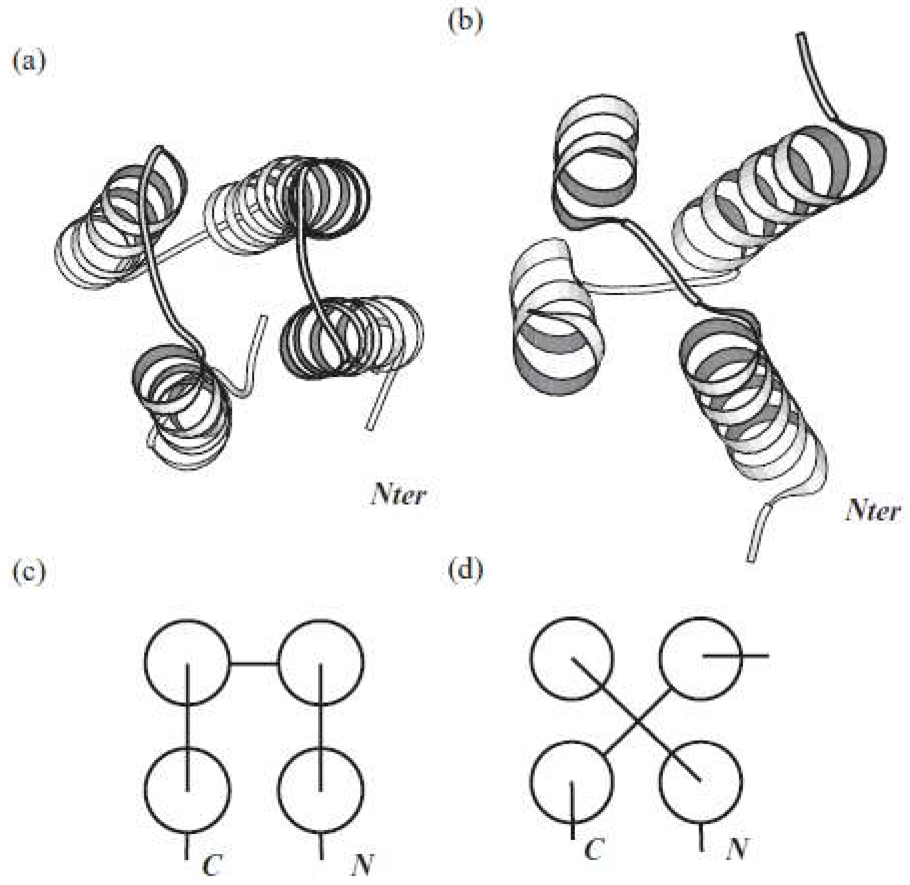
Structural classification

Topology and cartoon representation of the TIM barrel



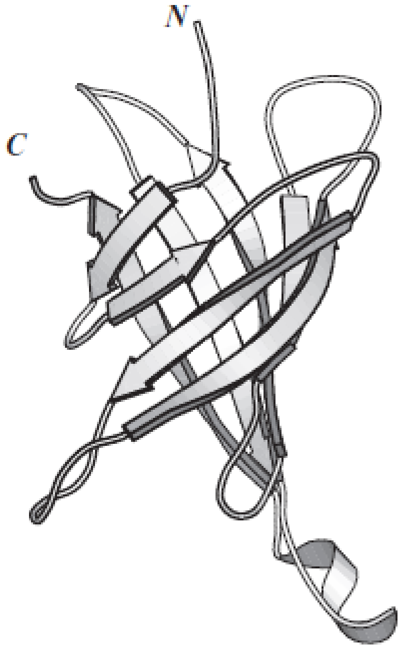
Reviews in Computational Chemistry, Volume 22
edited by Kenny B. Lipkowitz, Thomas R. Cundari, and Valerie J. Gillet
Copyright 2006 Wiley-VCH, John Wiley & Sons, Inc.

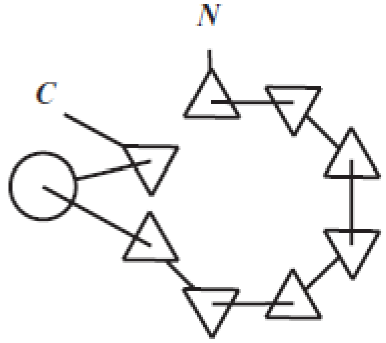
Two different topologies of four-helix bundles



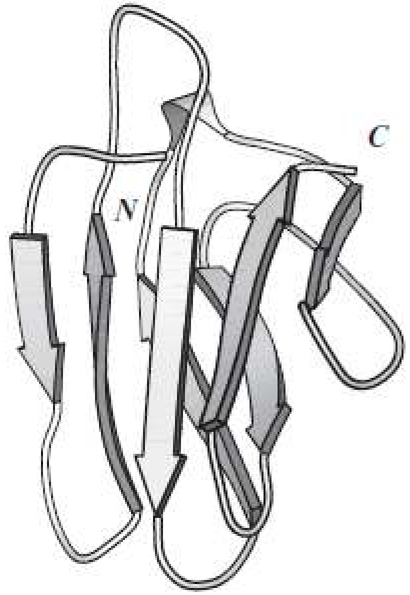
Reviews in Computational Chemistry, Volume 22
edited by Kenny B. Lipkowitz, Thomas R. Cundari, and Valerie J. Gillet
Copyright 2006 Wiley-VCH, John Wiley & Sons, Inc.

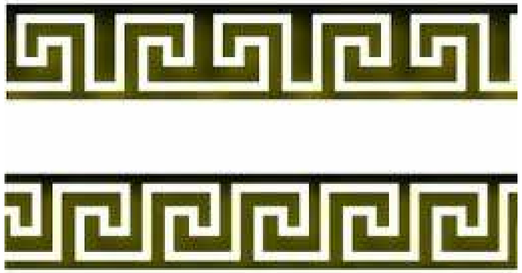
Barrel

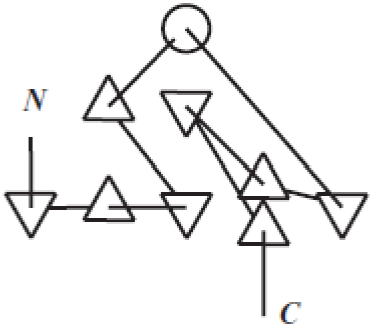




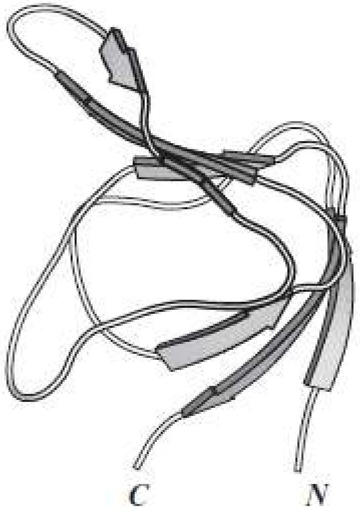
Greek key

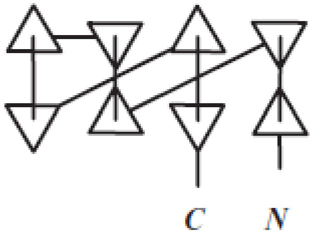




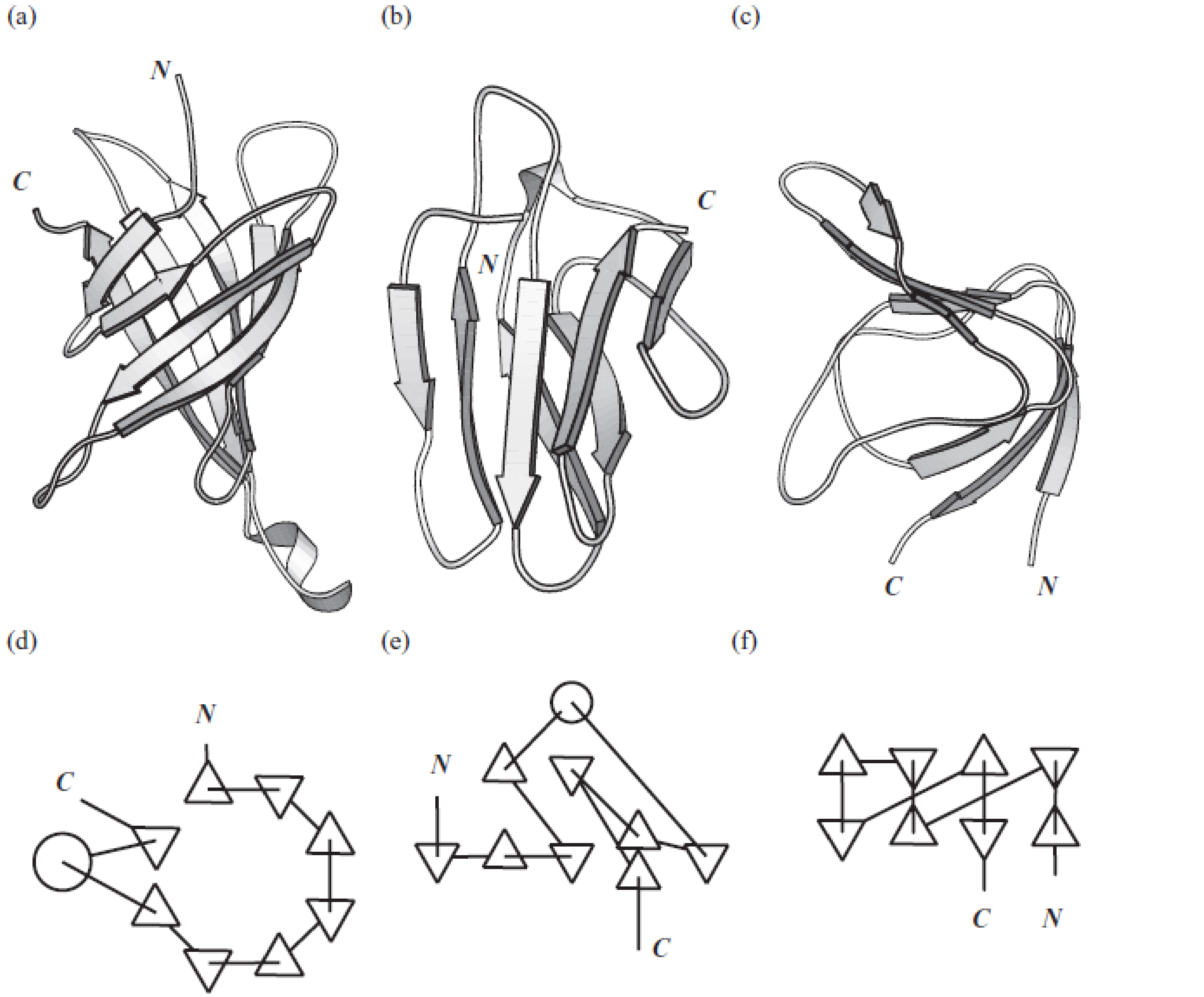


Jelly roll



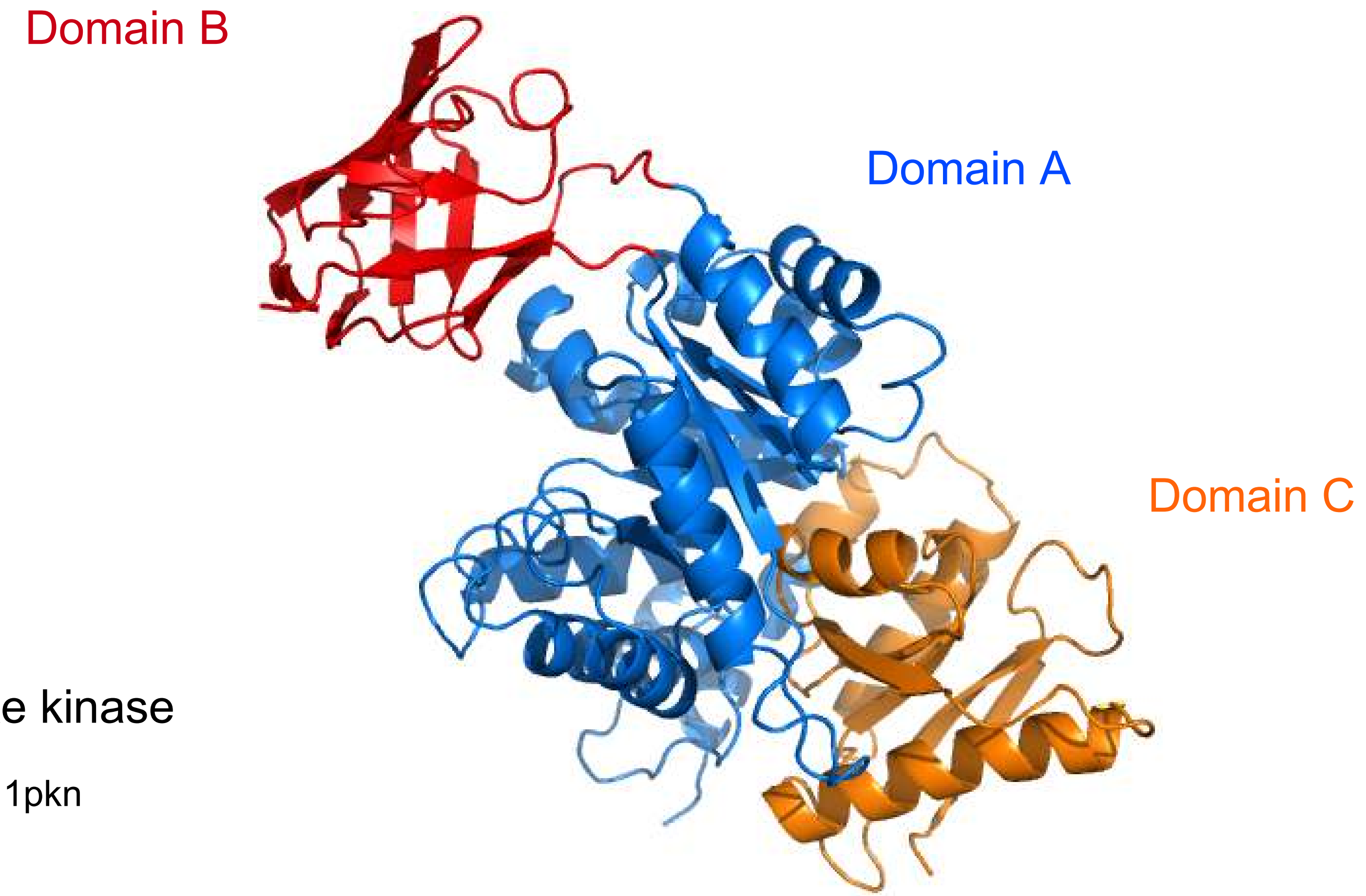


Three common sandwich topologies of beta proteins



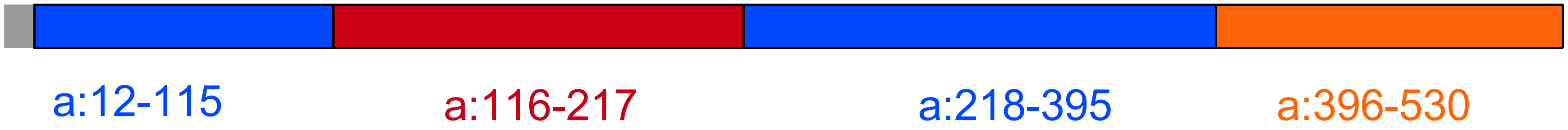
Reviews in Computational Chemistry, Volume 22
edited by Kenny B. Lipkowitz, Thomas R. Cundari, and Valerie J. Gillet

Structural domains



Pyruvate kinase

PDB:1pkn



SCOP classification

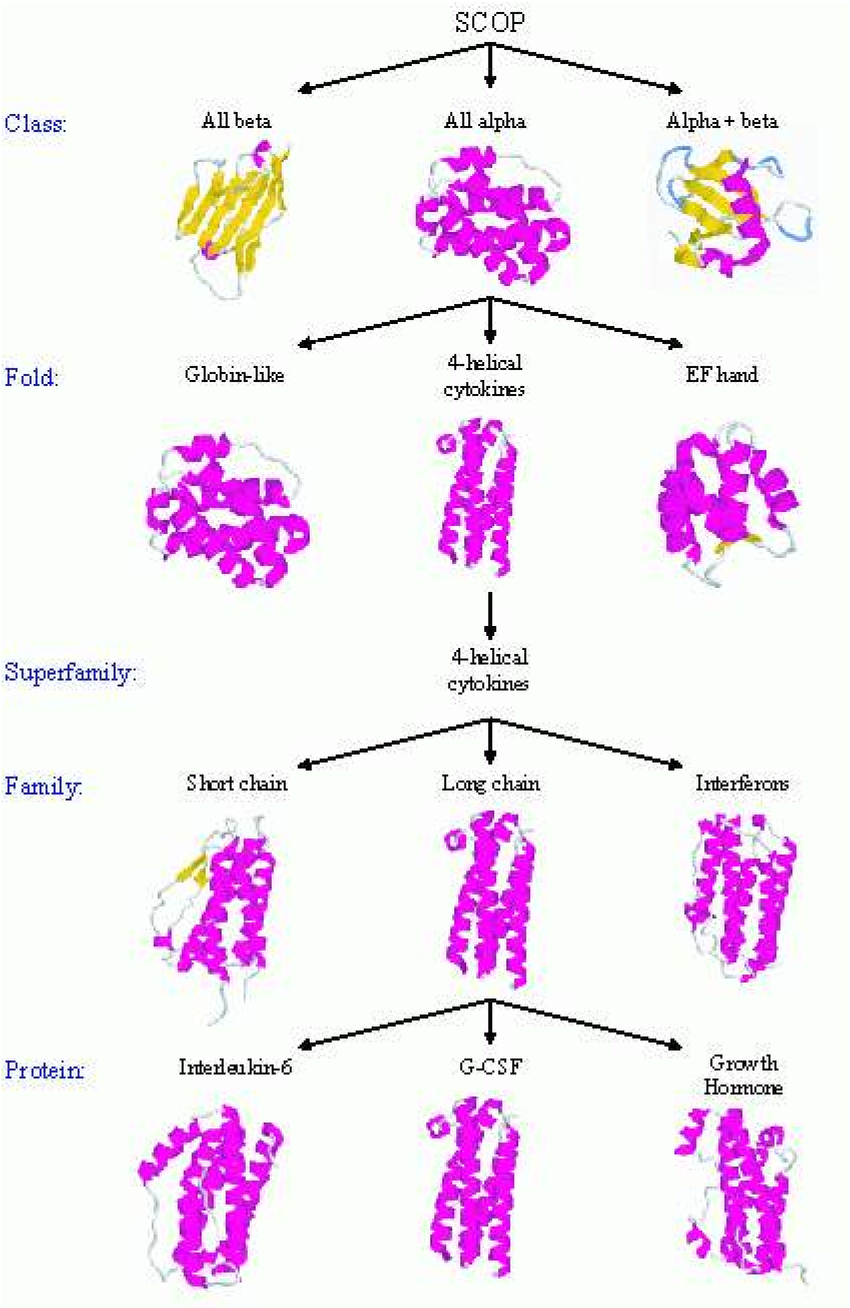


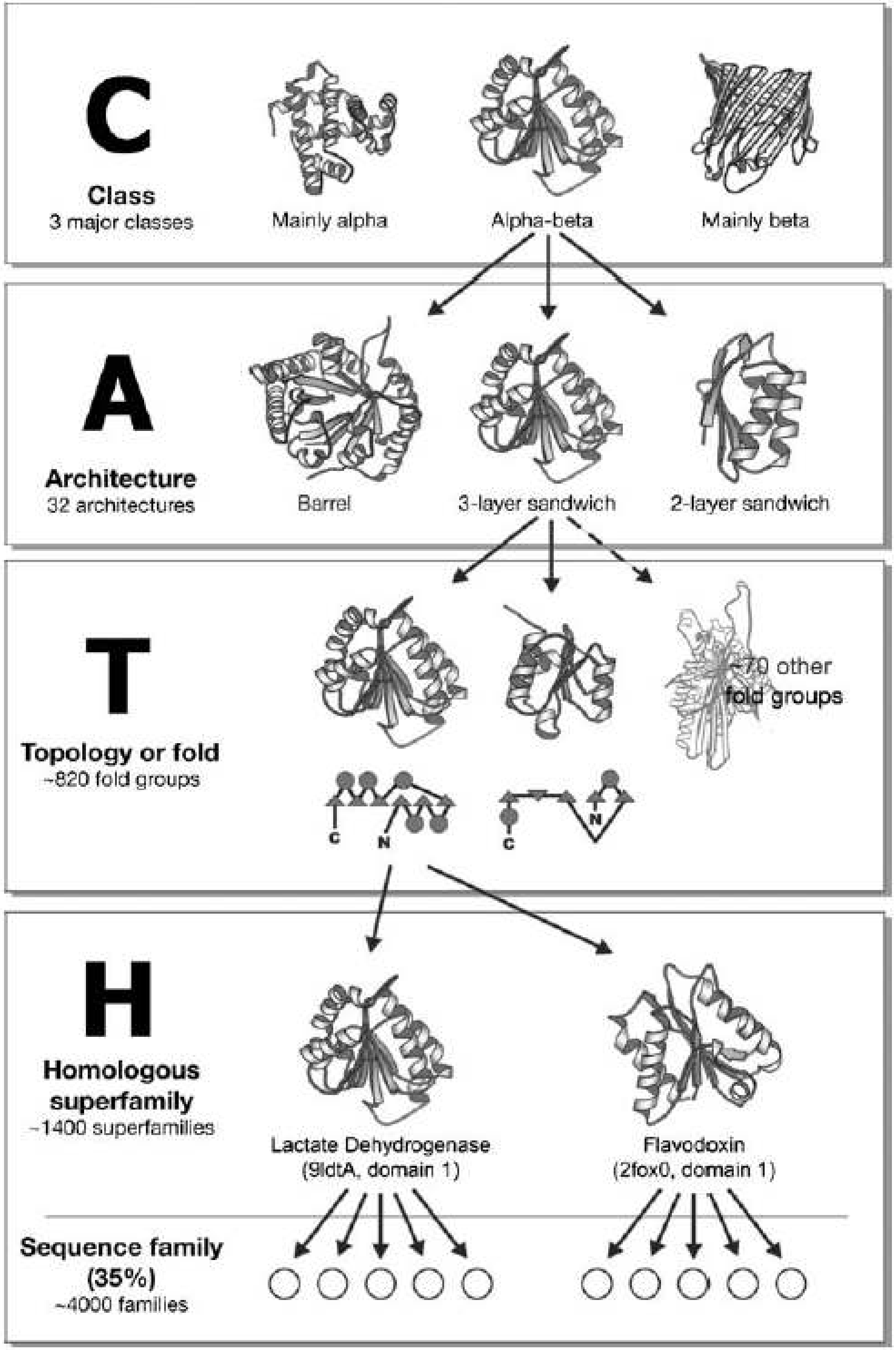
Image from: http://compbio.berkeley.edu/people/emma/scop_work.html

Fold: common structure (same SSEs in the same arrangement and topology)

Superfamily: probable common evolutionary origin (common structure and function despite low sequence identities)

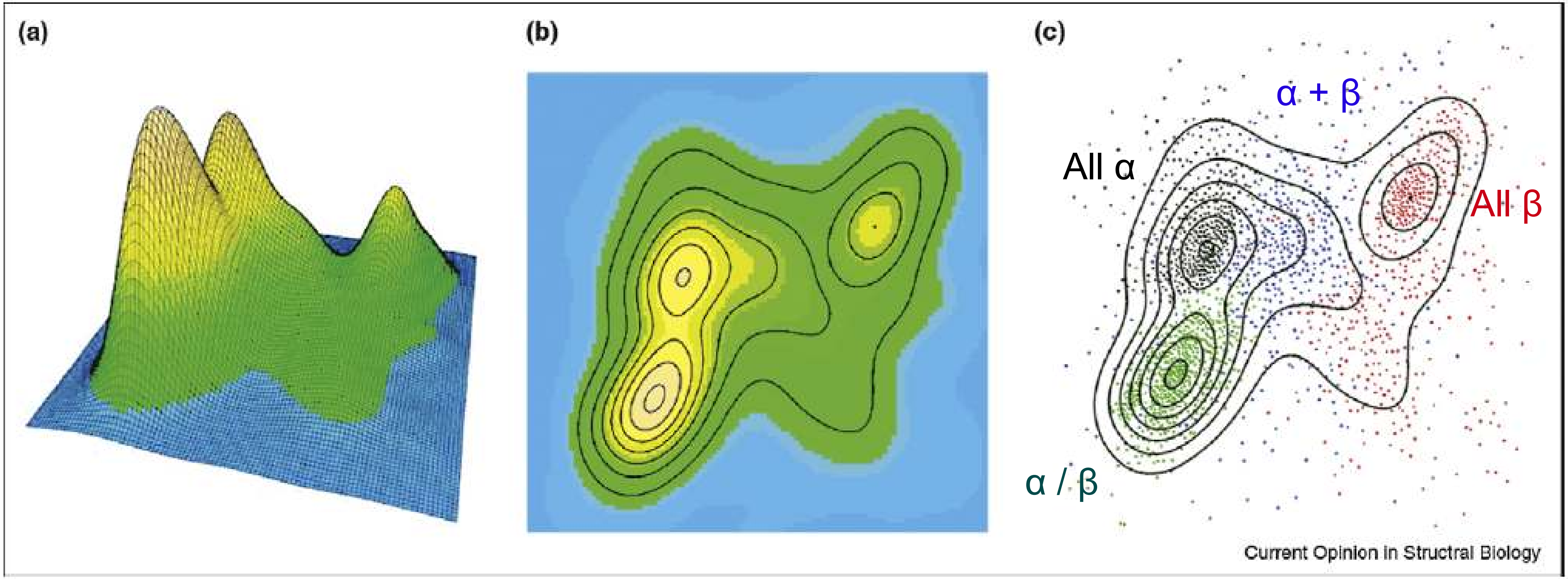
Family: clear common evolutionary origin (by sequence identity or extremely similar structure and function)

CATH classification



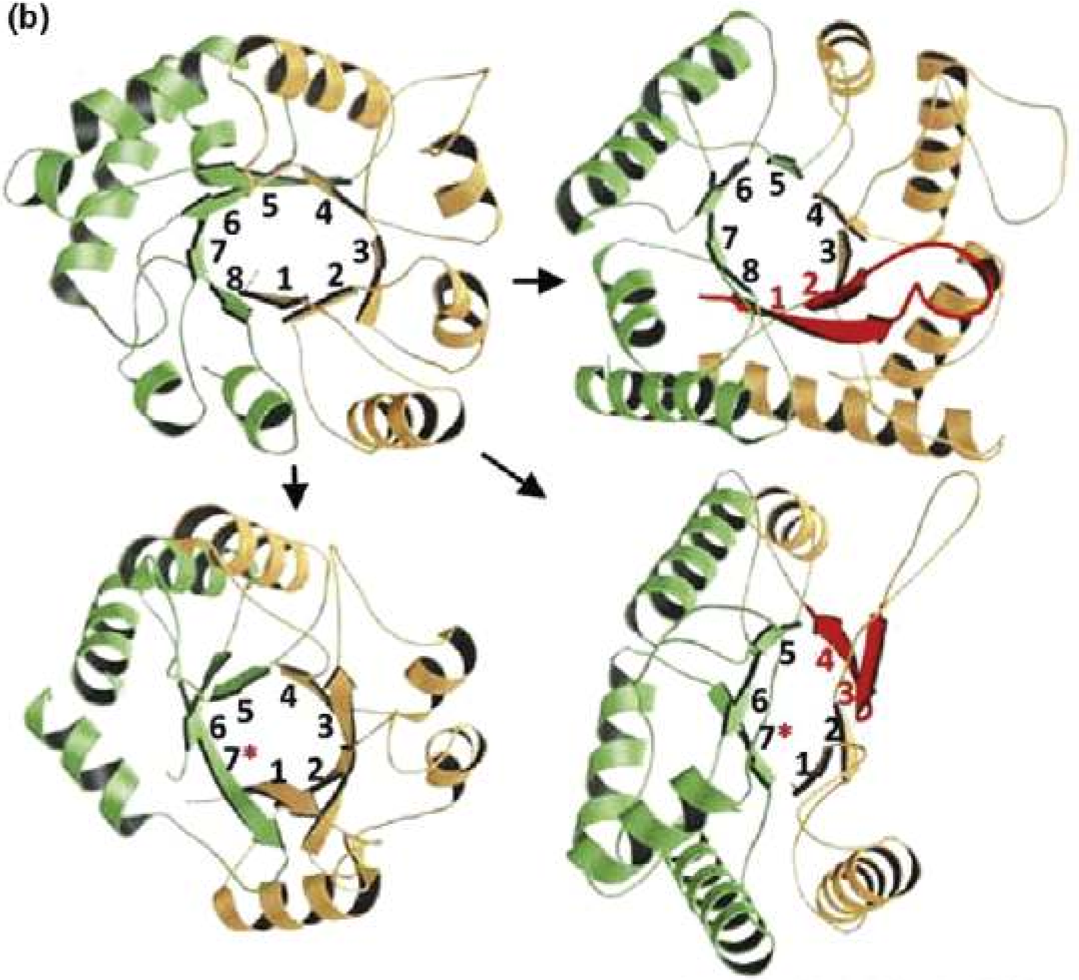
Annu. Rev. Biochem. 2005. 74:867–900

Discrete or continuous?



RI Sadreyev, BH Kim and NV Grishin
Discrete–continuous duality of protein structure space
Current Opinion in Structural Biology 2009, 19:321–328

TIM-barrel homologs with deviations from canonical fold

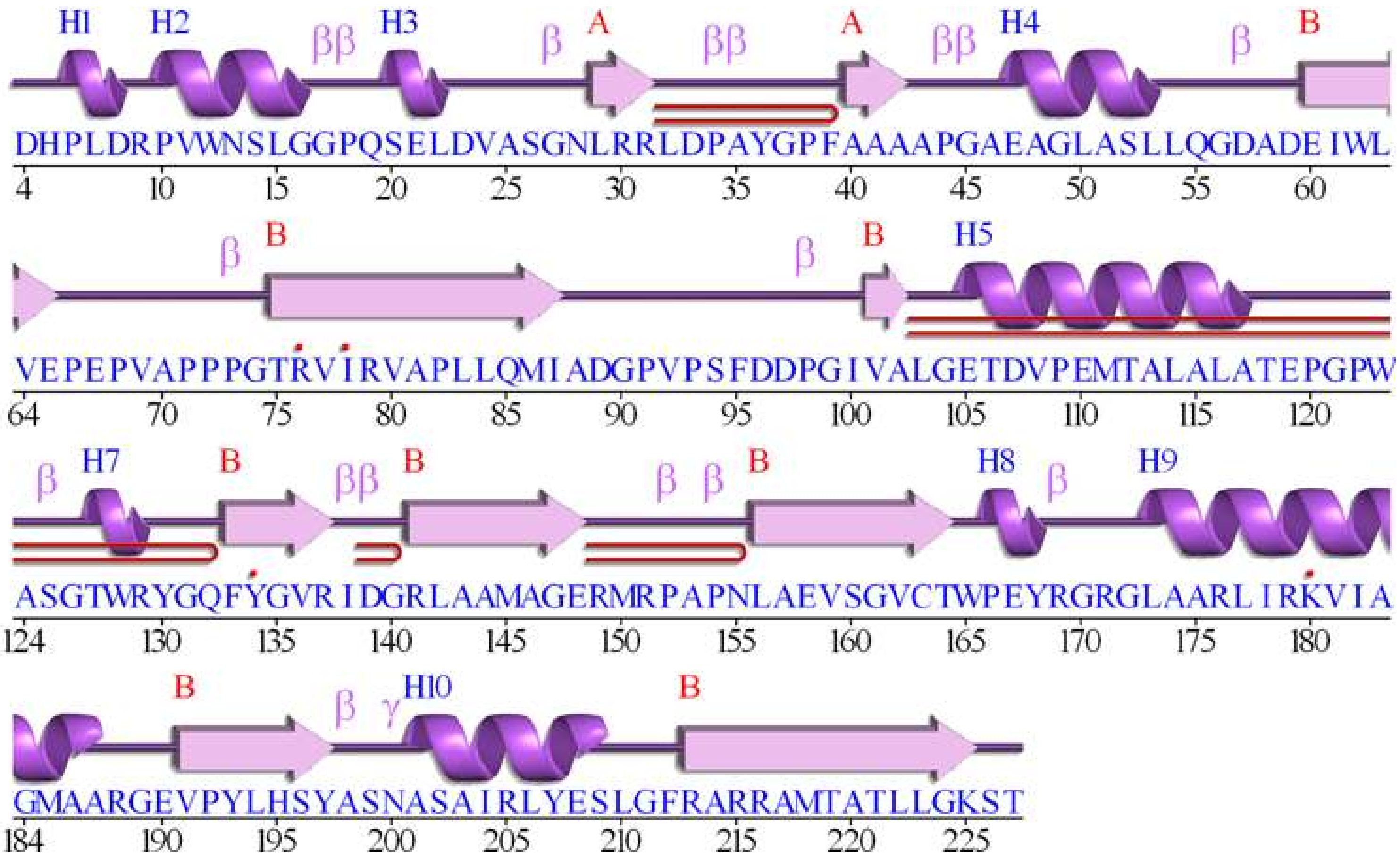


Current Opinion in Structural Biology

RI Sadreyev, BH Kim and NV Grishin
Discrete–continuous duality of protein structure space
Current Opinion in Structural Biology 2009, 19:321–328

Prediction 1D

Prediction of secondary structure


H1
H2
ββ
H3
β
A
ββ
A
ββ
H4
β
B
DHPLDRPVWNSLGGPQSELDVASGNLRRLDPAYGPFAAAAPGAEAGLASLLQGDADEIWL
4 10 15 20 25 30 35 40 45 50 55 60
β
B
β
B
H5
VEPEPVAPPPGTRVIRVAPLLQMIADGPVPSFDDPGIVALGETDVPEMTALALATEPGPW
64 70 75 80 85 90 95 100 105 110 115 120
β
H7
B
ββ
B
β β
B
H8
β
H9
ASGTWRYGQFYGVRIDGRLAAMAGERMRPAPNLAEVSGVCTWPEYRGRGLAARLIRKVIA
124 130 135 140 145 150 155 160 165 170 175 180
B
β γ
H10
B
GMAARGEVPYLHSYASNASAIRLYESLGFRARRAMTATLLGKST
184 190 195 200 205 210 215 220 225

DSSP* secondary structure elements

H = alpha helix
B = residue in isolated beta-bridge
E = extended strand, participates in beta ladder
G = 3-helix (3/10 helix)
I = 5 helix (pi helix)
T = hydrogen bonded turn
S = bend

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

Most secondary structure predictors

H = helix	(DSSP's H + G + I classes)
E = strand	(DSSP's E + B classes)
C = the rest	(DSSP's T + S + the rest)

First methods (70s) were based on single amino acid propensities

~ 60% accuracy

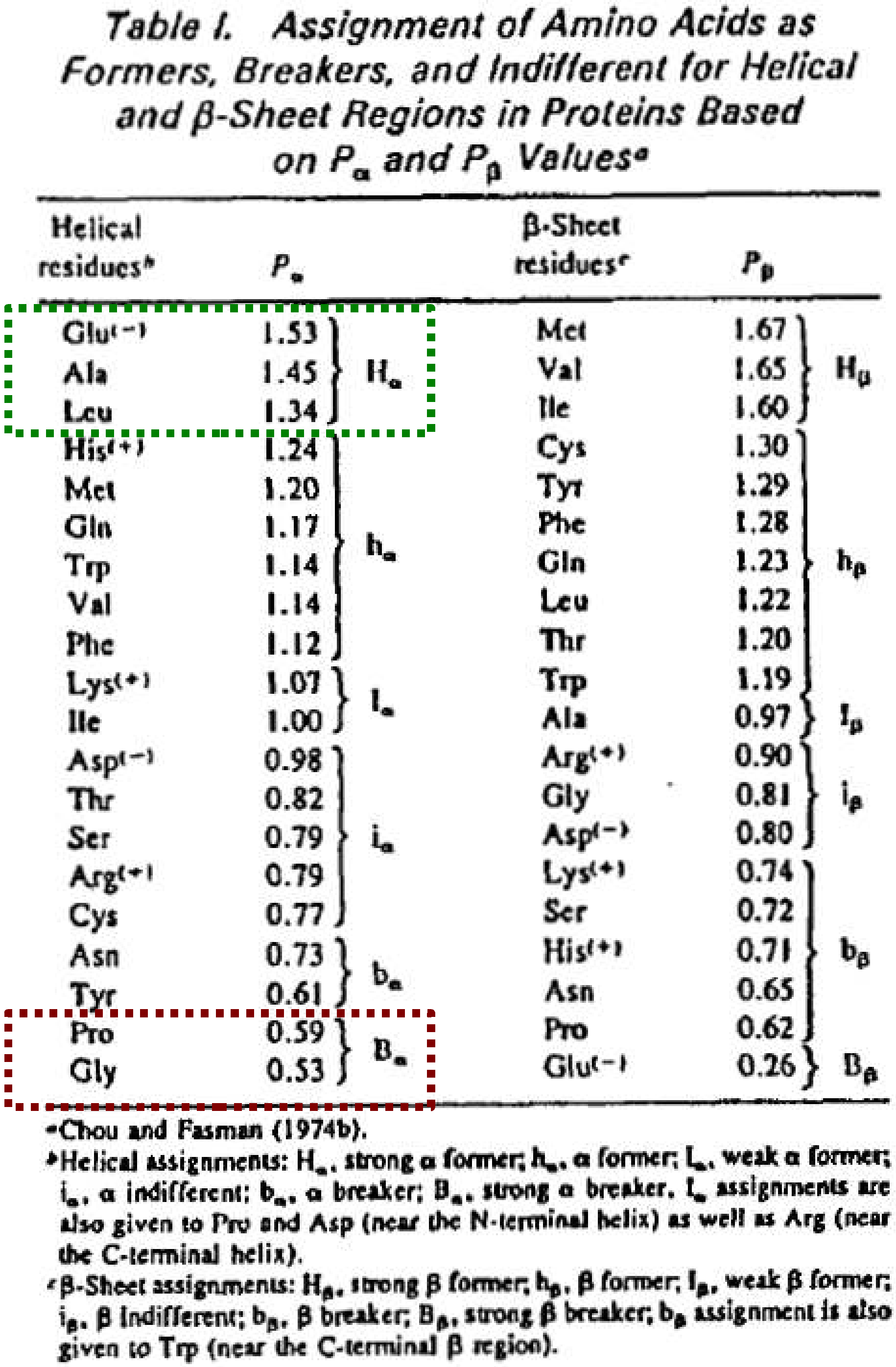


Table I. Assignment of Amino Acids as Formers, Breakers, and Indifferent for Helical and β-Sheet Regions in Proteins Based on P_α and P_β Values[a]

Helical residues[b]	P_α		β-Sheet residues[c]	P_β	
$Glu^{(-)}$	1.53	H_α	Met	1.67	H_β
Ala	1.45		Val	1.65	
Leu	1.34		Ile	1.60	
$His^{(+)}$	1.24	h_α	Cys	1.30	h_β
Met	1.20		Tyr	1.29	
Gln	1.17		Phe	1.28	
Trp	1.14		Gln	1.23	
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
$Lys^{(+)}$	1.07	I_α	Trp	1.19	
Ile	1.00		Ala	0.97	I_β
$Asp^{(-)}$	0.98	i_α	$Arg^{(+)}$	0.90	i_β
Thr	0.82		Gly	0.81	
Ser	0.79		$Asp^{(-)}$	0.80	
$Arg^{(+)}$	0.79		$Lys^{(+)}$	0.74	b_β
Cys	0.77		Ser	0.72	
Asn	0.73	b_α	$His^{(+)}$	0.71	
Tyr	0.61		Asn	0.65	
Pro	0.59	B_α	Pro	0.62	
Gly	0.53		$Glu^{(-)}$	0.26	B_β

[a] Chou and Fasman (1974b).

[b] Helical assignments: H_α, strong α former; h_α, α former; I_α, weak α former; i_α, α indifferent; b_α, α breaker; B_α, strong α breaker. I_α assignments are also given to Pro and Asp (near the N-terminal helix) as well as Arg (near the C-terminal helix).

[c] β-Sheet assignments: H_β, strong β former; h_β, β former; I_β, weak β former; i_β, β indifferent; b_β, β breaker; B_β, strong β breaker; b_β assignment is also given to Trp (near the C-terminal β region).

Chou PY, Fasman GD (1974). "Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins".
Biochemistry 13 (2): 211–222

Second generation methods (until early 90s)

<70% accuracy

Compiled propensities for segments of adjacent residues (3-51 residues)

But, secondary structure formation is partially determined by nonlocal interactions (e.g. sheets). Local information was estimated to account for roughly 65% of the secondary structure information.

Third generation methods

>70% accuracy

Multiple sequence alignments

Larger databases

More advanced algorithms

Secondary structure prediction software

PSIPRED	http://bioinf.cs.ucl.ac.uk/psipred/
PROF	http://www.aber.ac.uk/~phiwww/prof/
SSpro	http://scratch.proteomics.ics.uci.edu/
Porter	http://distill.ucd.ie/porter/
APSSP2	http://www.imtech.res.in/raghava/apssp2/
SAM	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html
YASPIN	http://www.ibi.vu.nl/programs/yaspinwww/
Jpred	http://www.compbio.dundee.ac.uk/jpred/

Trasmembrane segments

M Punta, *et al.*
Membrane protein prediction methods
Methods (2007) 41:460–474

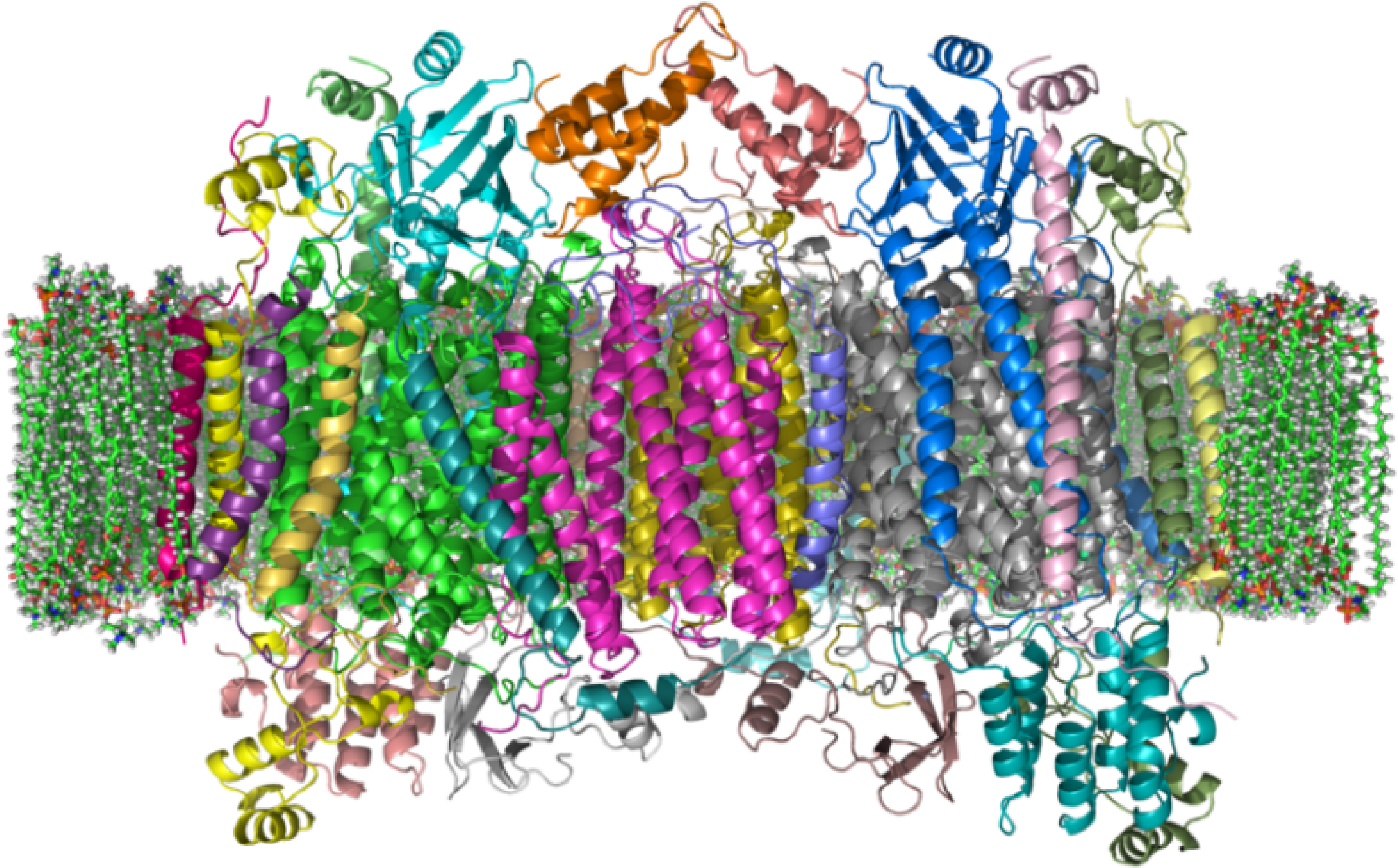
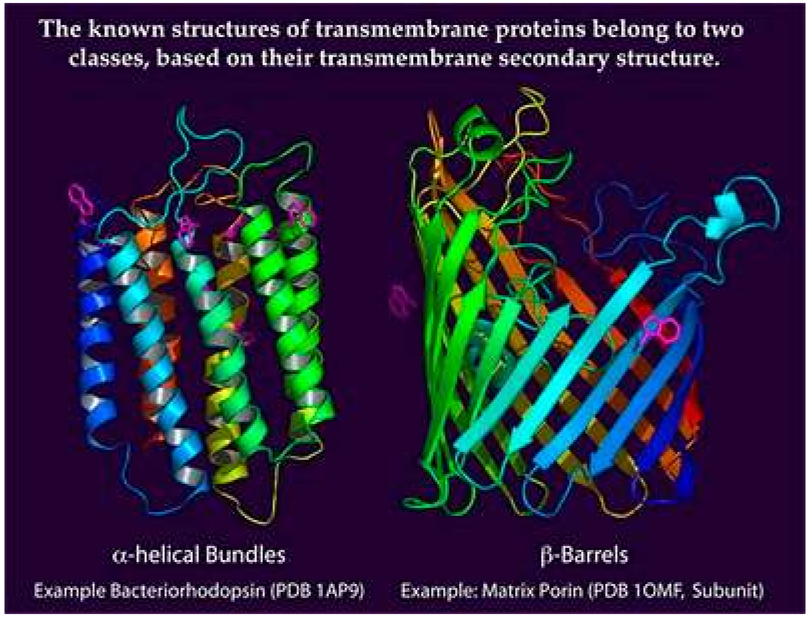


Image: Wikimedia Commons

In comparison to water-soluble proteins,

IMP chains are able to sample only a limited number of folds



The known structures of transmembrane proteins belong to two classes, based on their transmembrane secondary structure.

α-helical Bundles
Example Bacteriorhodopsin (PDB 1AP9)

β-Barrels
Example: Matrix Porin (PDB 1OMF, Subunit)

SCOP Class: Membrane and cell surface proteins and peptides

See http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.g.html

The number, location and cross-membrane direction of TM segments can be predicted rather accurately

Strong compositional biases imposed by the bilayer

TMHs > 15 residues, predominantly hydrophobic amino acids.

TMBs > 10 residues , alternating hydrophobic and polar amino acids

'Positive-inside rule'

regions connecting TM segments that are not translocated across the bilayer ("inside" or cytoplasmic regions) are enriched in positively charged amino acids

Early predictions of TM helical segments

4-step procedure:

(1) Derive a ‘transmembrane propensity scale’,
(2) Generate a plot of propensity values along the query sequence.
(3) Smooth the plot by taking the average propensity value in a window of N residues and plot the average at the center of the window (i.e. a sliding-window average).
(4) Identify TM stretches on the smoothed plot using some propensity threshold.

Current predictions

Machine learning approaches
- Neural networks (NN)
- Hidden Markov Models (HMM)
- Support Vector Machines (SVM)

Larger databases

Trasmembrane prediction software

Transmembrane helices

MEMSAT	http://bioinf.cs.ucl.ac.uk/psipred/
TopPred	http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html
HMMTOP	http://www.enzim.hu/hmmtop/
DAS	http://www.enzim.hu/DAS/DAS.html
TMHMM	http://www.cbs.dtu.dk/services/TMHMM-2.0/
Tmpred	http://www.ch.embnet.org/software/TMPRED_form.html
MINNOU	http://minnou.cchmc.org/
Phobius	http://phobius.sbc.su.se/

Trasmembrane prediction software

Transmembrane barrels

PRED-TMBB	http://biophysics.biol.uoa.gr/PRED-TMBB/
BOMP	http://services.cbu.uib.no/tools/bomp
TMB-HUNT	http://www.bioinformatics.leeds.ac.uk
B2TMR, HMM-B2TMR PROFtmb	
ConBBPRED	http://bioinformatics.biol.uoa.gr/ConBBPRED/

Trasmembrane proteins databases

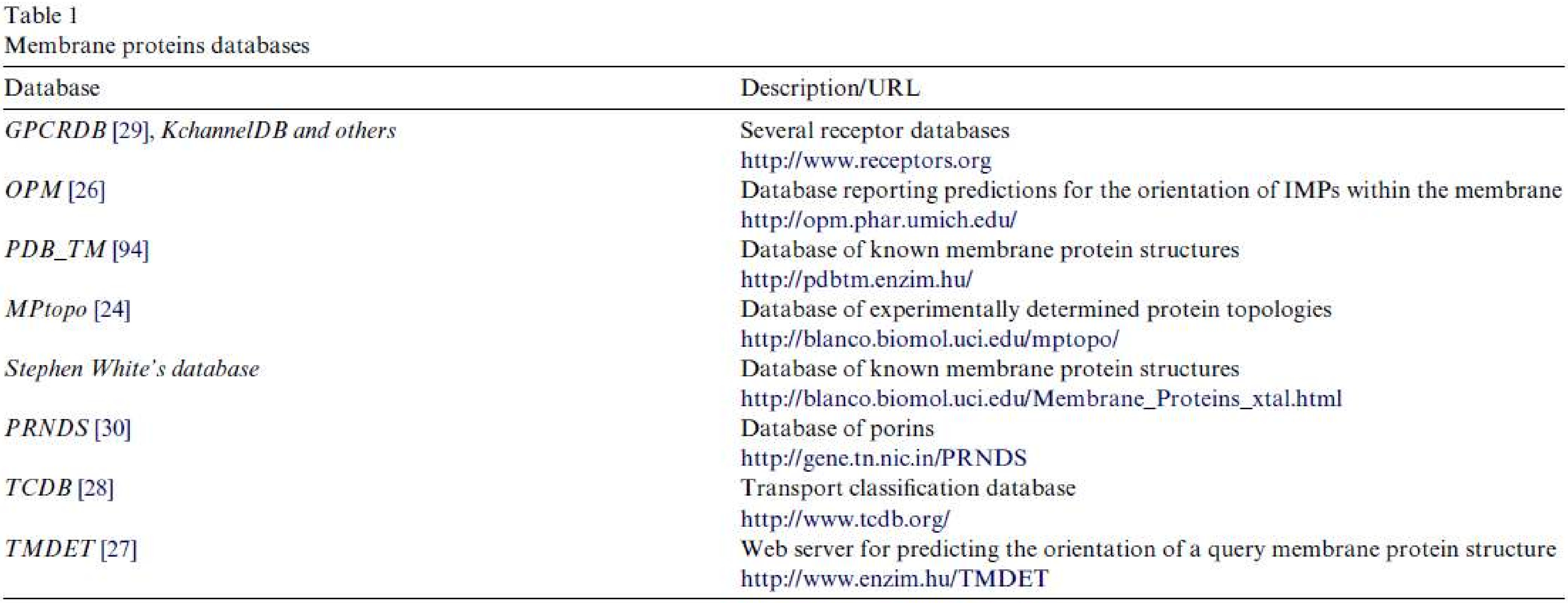
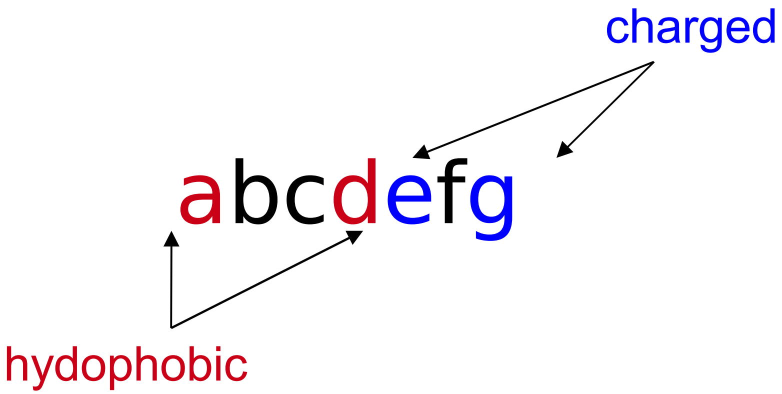


Table 1
Membrane proteins databases

Database	Description/URL
GPCRDB [29], *KchannelDB and others*	Several receptor databases http://www.receptors.org
OPM [26]	Database reporting predictions for the orientation of IMPs within the membrane http://opm.phar.umich.edu/
PDB_TM [94]	Database of known membrane protein structures http://pdbtm.enzim.hu/
MPtopo [24]	Database of experimentally determined protein topologies http://blanco.biomol.uci.edu/mptopo/
Stephen White's database	Database of known membrane protein structures http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html
PRNDS [30]	Database of porins http://gene.tn.nic.in/PRNDS
TCDB [28]	Transport classification database http://www.tcdb.org/
TMDET [27]	Web server for predicting the orientation of a query membrane protein structure http://www.enzim.hu/TMDET

Coiled-coils

Heptad repeat

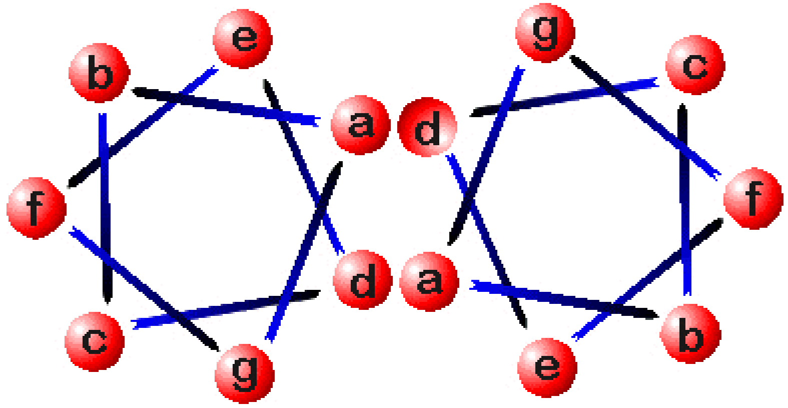


charged

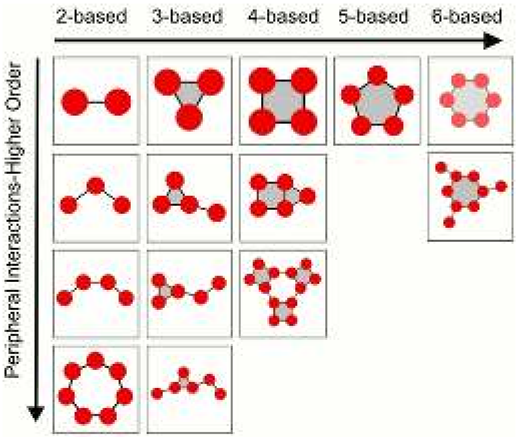
abcdefg

hydophobic

Amphipathic helix

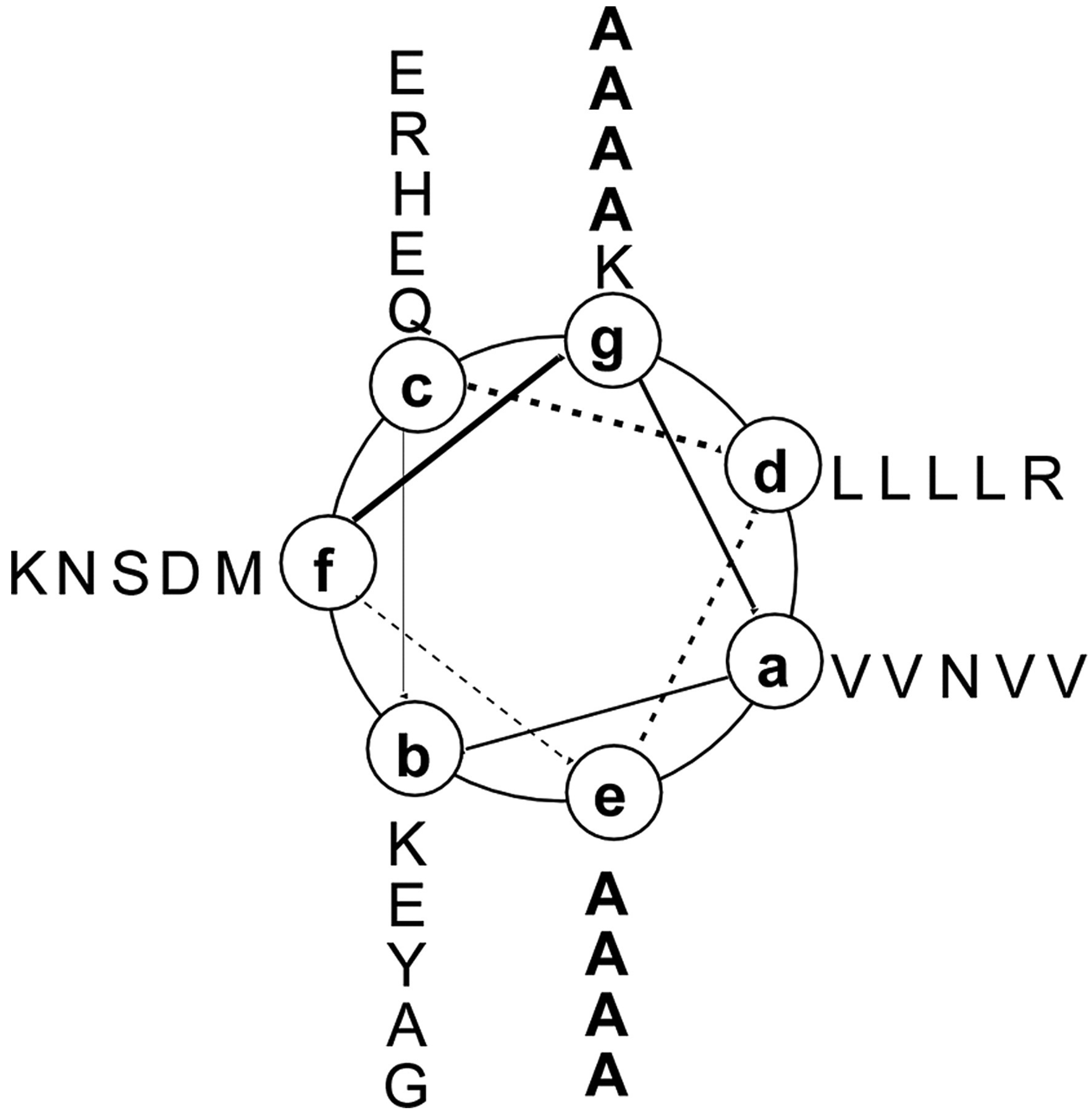


Coiled-coil architectures



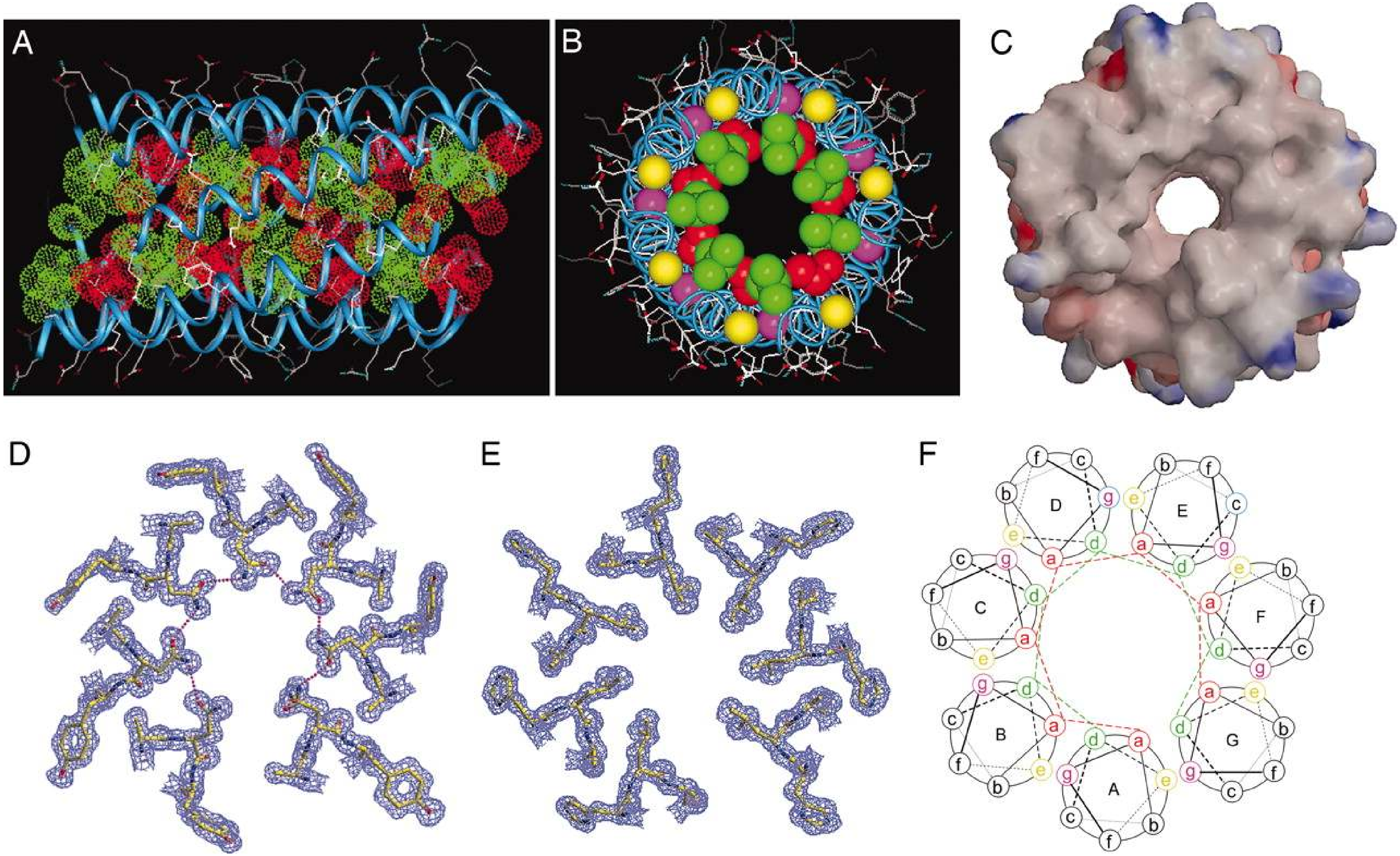
Images: CC+ database
coiledcoils.chm.bris.ac.uk/ccplus

Helical wheel projection of residues Met-1 to Arg-34 of the GCN4-pAA sequence.



Liu J et al. PNAS 2006;103:15457-15462

Crystal structure of GCN4-pAA.



Liu J et al. PNAS 2006;103:15457-15462

Coiled coils prediction software

COILS	http://www.ch.embnet.org/software/COILS_form.html
Paircoil2	http://groups.csail.mit.edu/cb/paircoil2/
bCIPA	http://www.molbiotech.uni-freiburg.de/bCIPA/
PrOCoil	http://www.bioinf.jku.at/software/procoil/

Database

CC+	http://coiledcoils.chm.bris.ac.uk/ccplus/

Disorder

A large number of naturally occurring proteins do not require a well-folded structure to fulfill their biological role

These intrinsically unstructured/disordered proteins (IUPs/IDPs) exist as ensembles of rapidly interconverting conformations, even under physiological conditions.

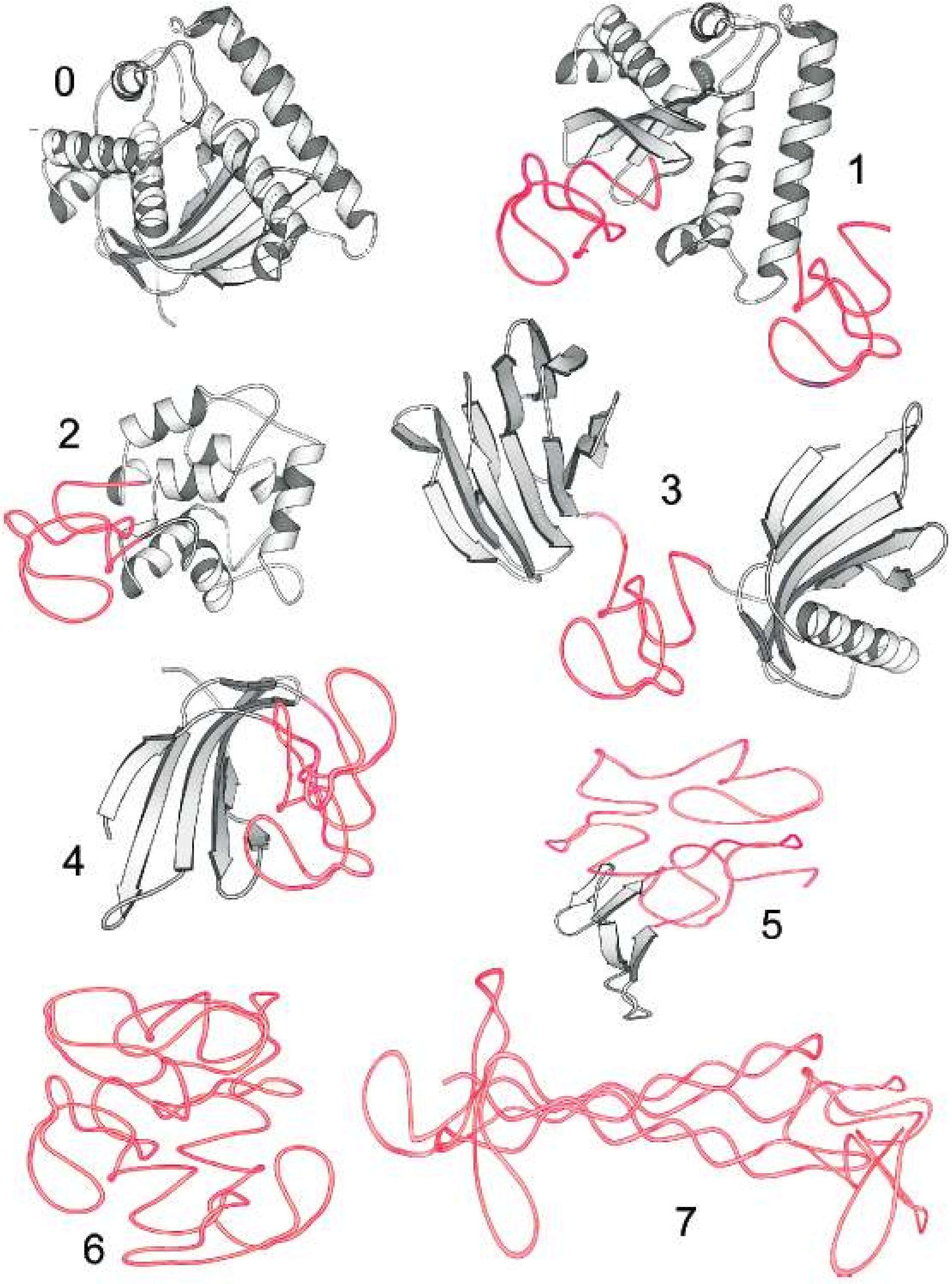
While some proteins appear fully disordered, many proteins are composed of both ordered and disordered regions of various lengths

It is estimated that 30–50% of eukaryotic proteins contain at least one long disordered segment

Disorder is related important regulatory functions in the cell including transcription, translation and cell signaling

Z Dosztanyi, B Meszaros and I Simon
Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins
Briefings in bioinformatics (2010) 11: 225-243

Different levels of **order** and **disorder**



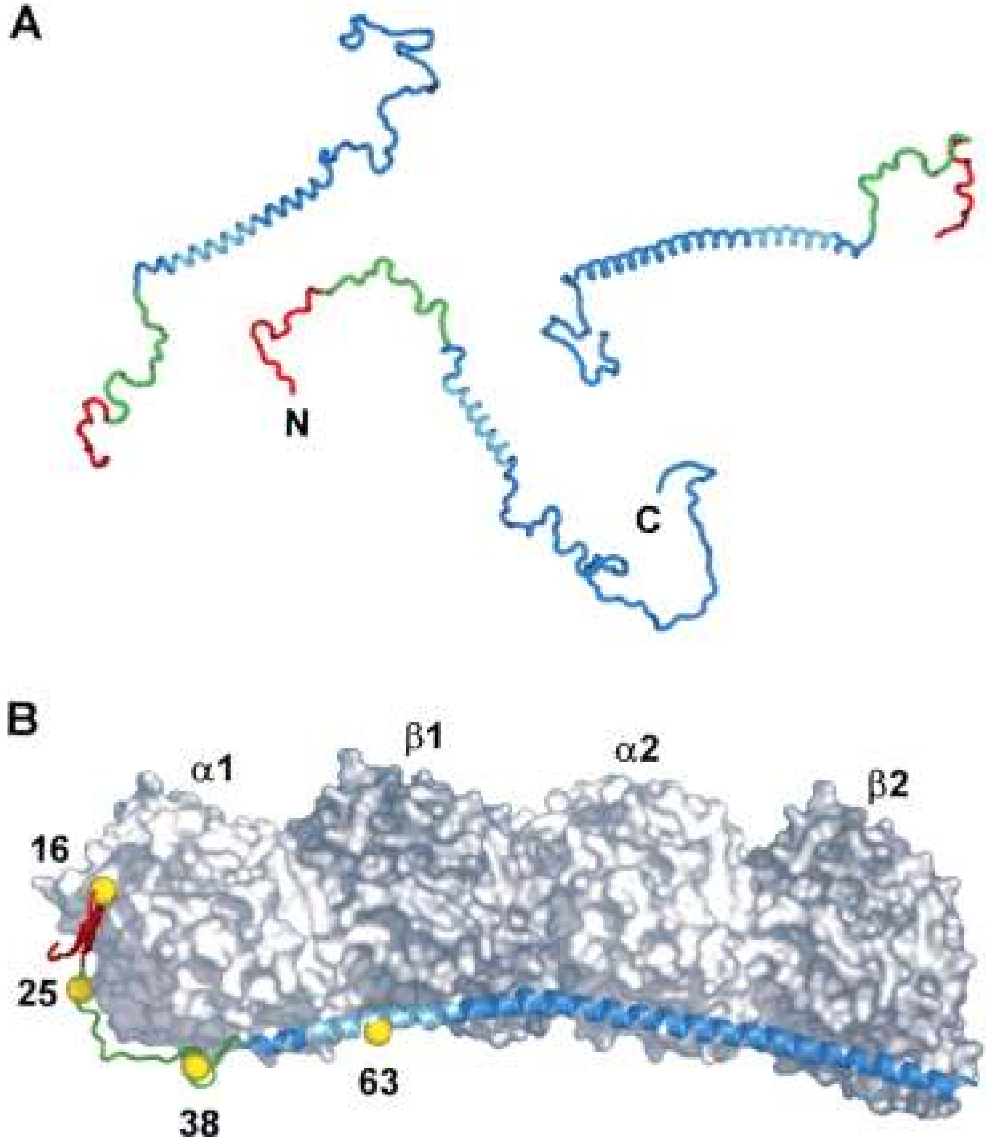
VN Uversky, CJ Oldfield and AK Dunker
Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling
J. Mol. Recognit. (2005) 18: 343–384

Two types of disorder

Permanent disordered state

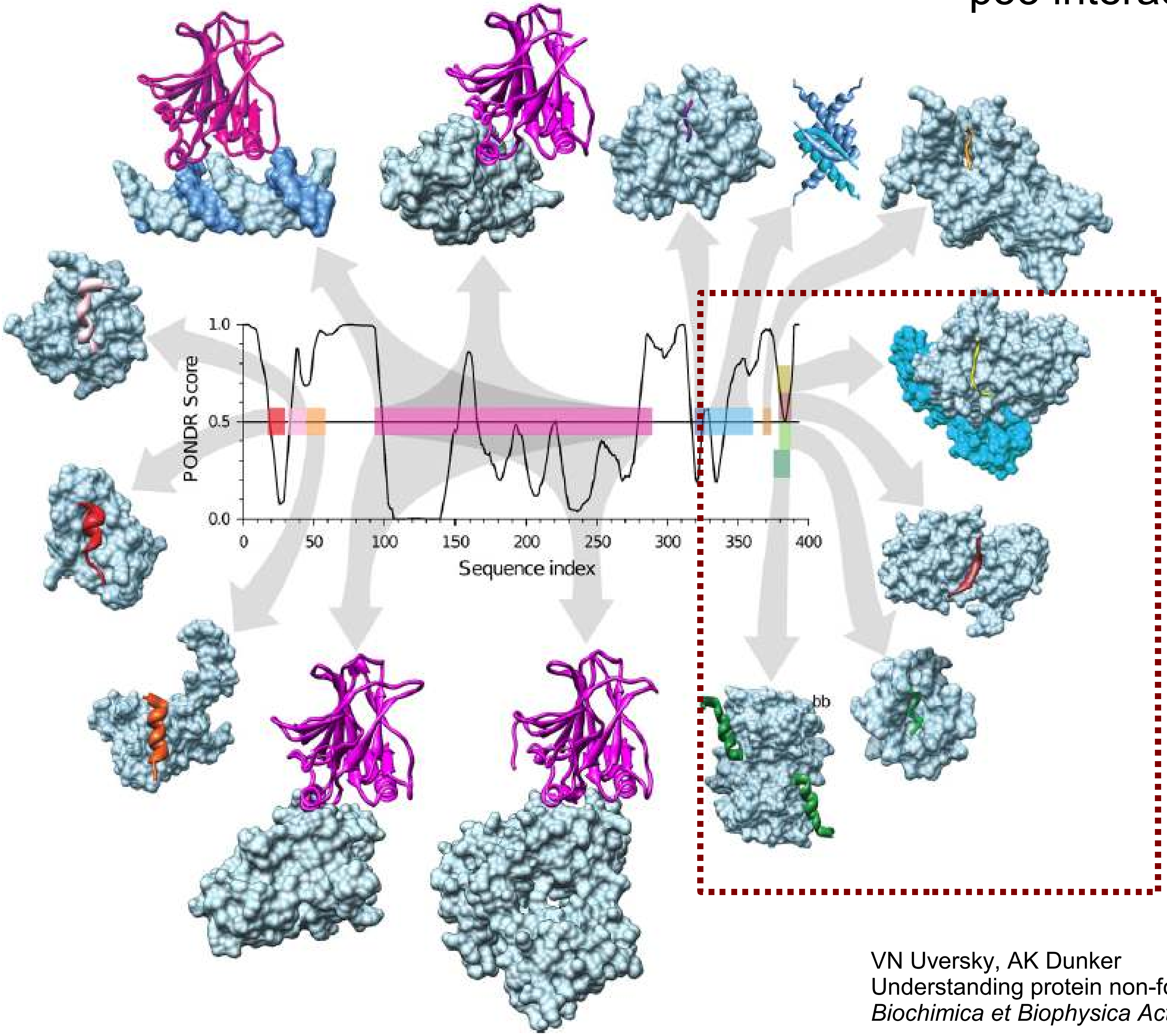
Disorder-to-order transition upon binding to other macromolecules

Stathmin is an intrinsically disordered protein that forms a ternary complex with tubulin.



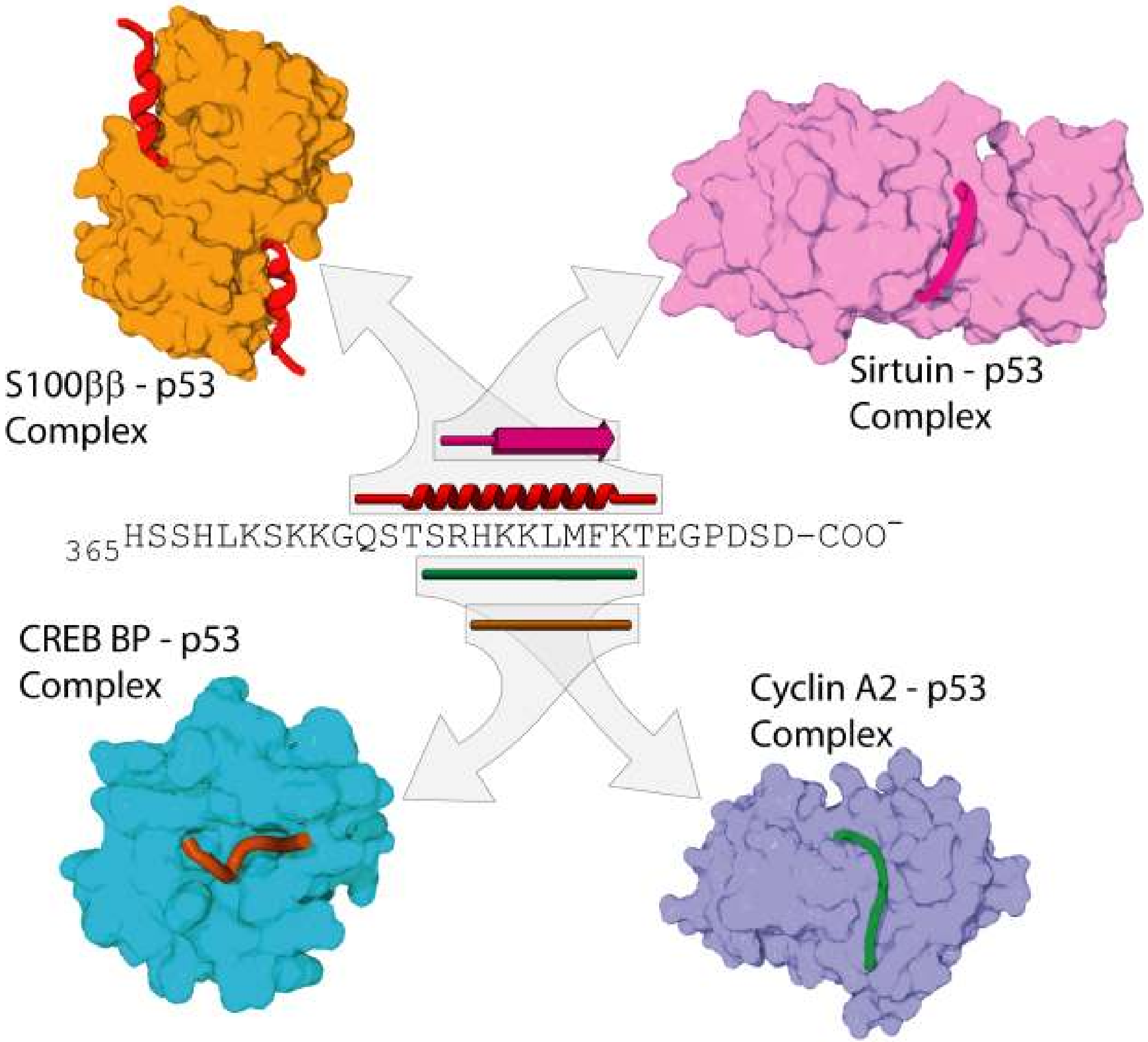
Honnappa S et al. J. Biol. Chem. 2006;281:16078-16083

p53 interactors



VN Uversky, AK Dunker
Understanding protein non-folding
Biochimica et Biophysica Acta (2010) 1231–1264

Structure comparison for the four overlapping complexes in the C-terminus of p53

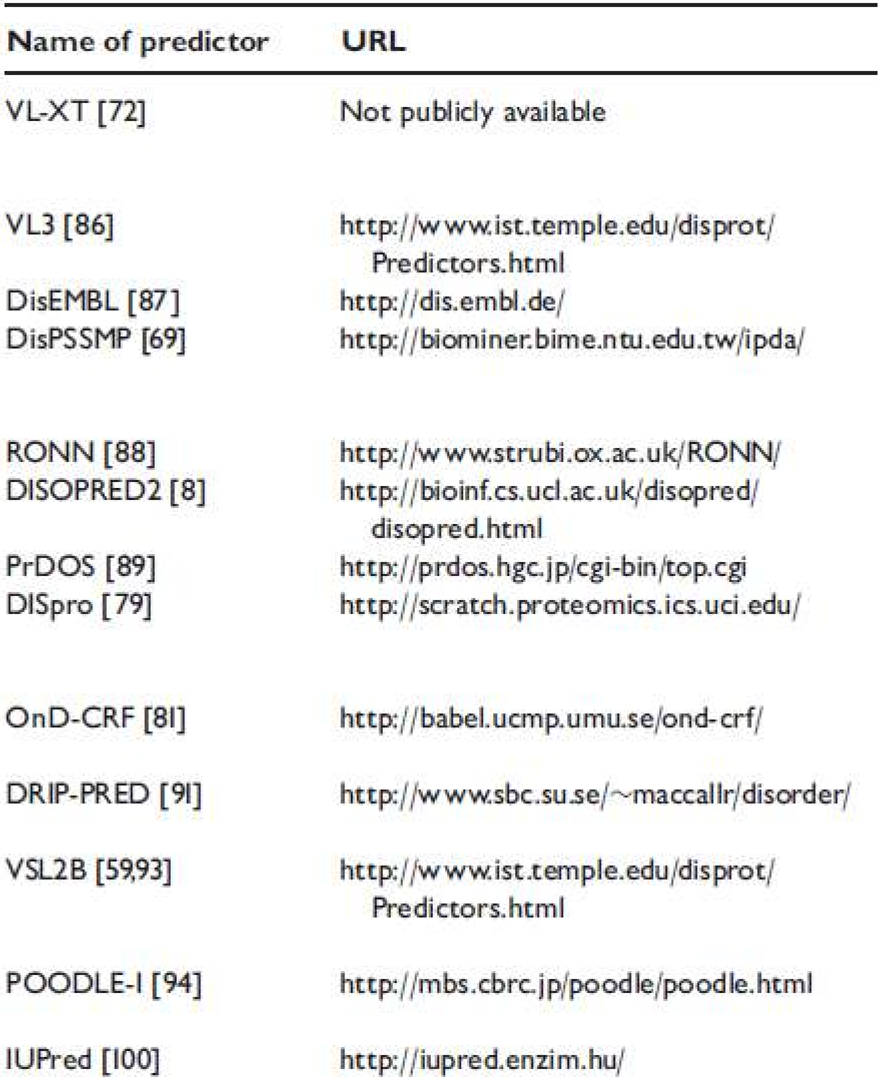


CJ Oldfield et al. BMC Genomics 2008 9(Suppl 1):S1

Disorder prediction is based on

- Basic sequence properties
    - Amino acid composition
    - Hydrophobicity
- Low complexity segments
- Evolutionary information
- Secondary structure
- Solvent accesibility

Disorder prediction software



Name of predictor	URL
VL-XT [72]	Not publicly available
VL3 [86]	http://www.ist.temple.edu/disprot/Predictors.html
DisEMBL [87]	http://dis.embl.de/
DisPSSMP [69]	http://biominer.bime.ntu.edu.tw/ipda/
RONN [88]	http://www.strubi.ox.ac.uk/RONN/
DISOPRED2 [8]	http://bioinf.cs.ucl.ac.uk/disopred/disopred.html
PrDOS [89]	http://prdos.hgc.jp/cgi-bin/top.cgi
DISpro [79]	http://scratch.proteomics.ics.uci.edu/
OnD-CRF [81]	http://babel.ucmp.umu.se/ond-crf/
DRIP-PRED [91]	http://www.sbc.su.se/~maccallr/disorder/
VSL2B [59,93]	http://www.ist.temple.edu/disprot/Predictors.html
POODLE-I [94]	http://mbs.cbrc.jp/poodle/poodle.html
IUPred [100]	http://iupred.enzim.hu/

Databases

DisProt
http://www.disprot.org/

Z Dosztanyi, B Meszaros and I Simon
Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins
Briefings in bioinformatics (2010) 11: 225-243

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

CBS Prediction Servers http://www.cbs.dtu.dk/services/

SIB ExPASy
Bioinformatics Resource Portal

Expasy proteomic tools http://www.expasy.org/proteomics http://csbg.cnb.csic.es/Courses/Struct_2011/

Practicals