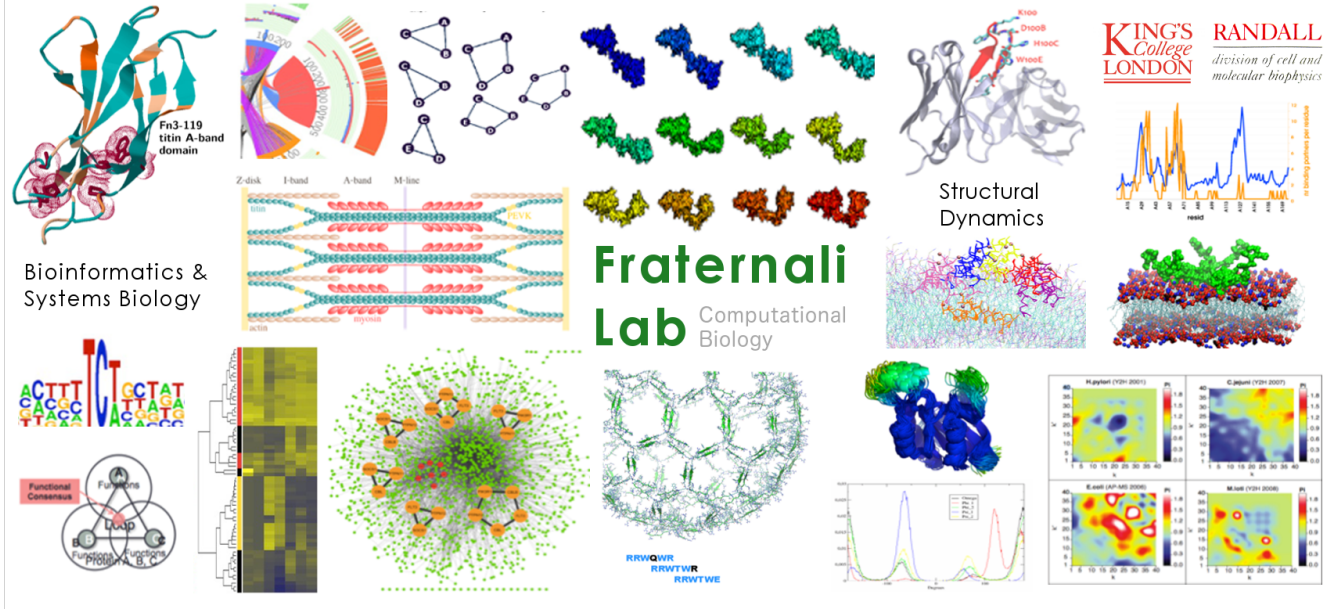
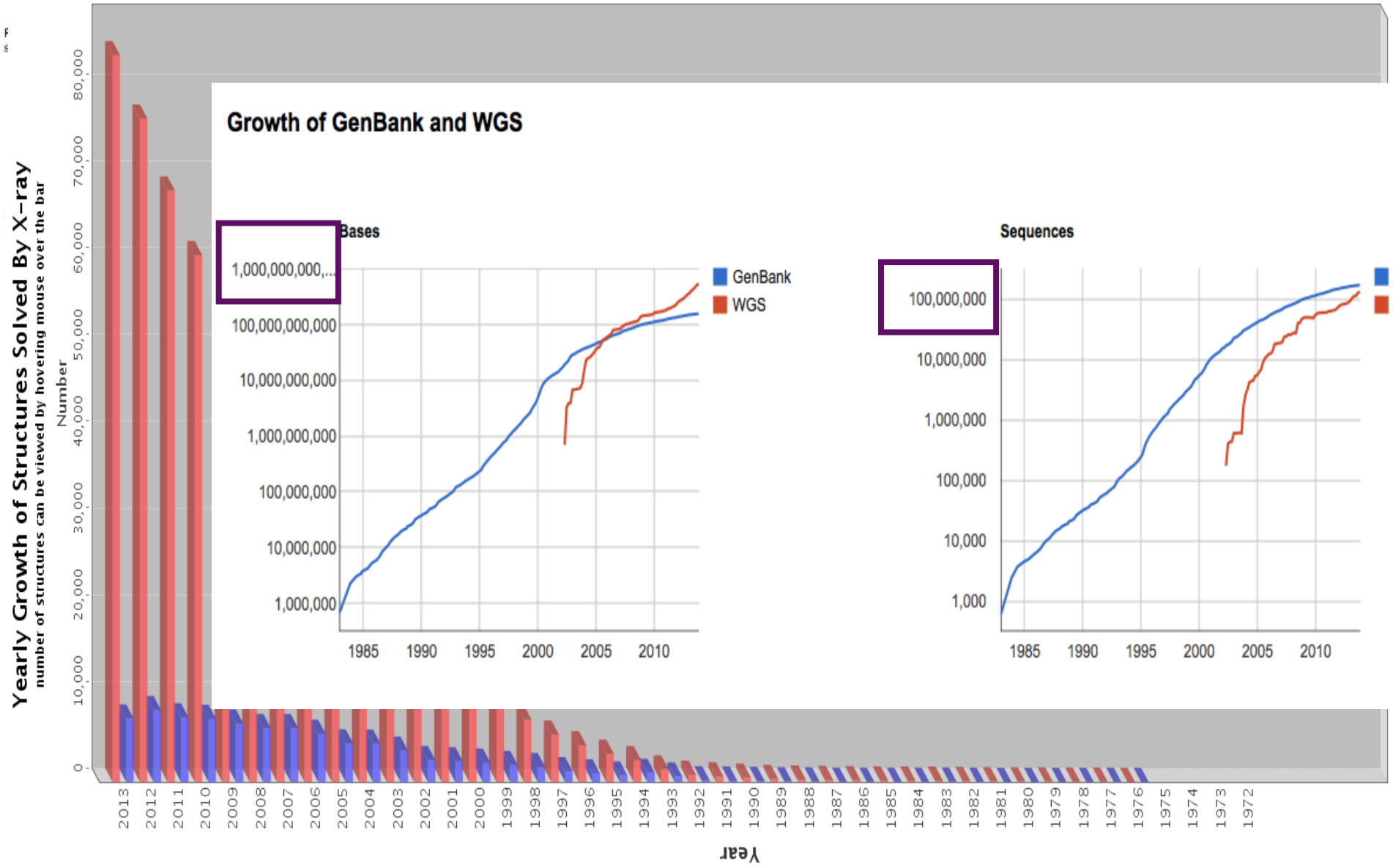


http://fraternalilab.kcl.ac.uk

80.000

Primary structure



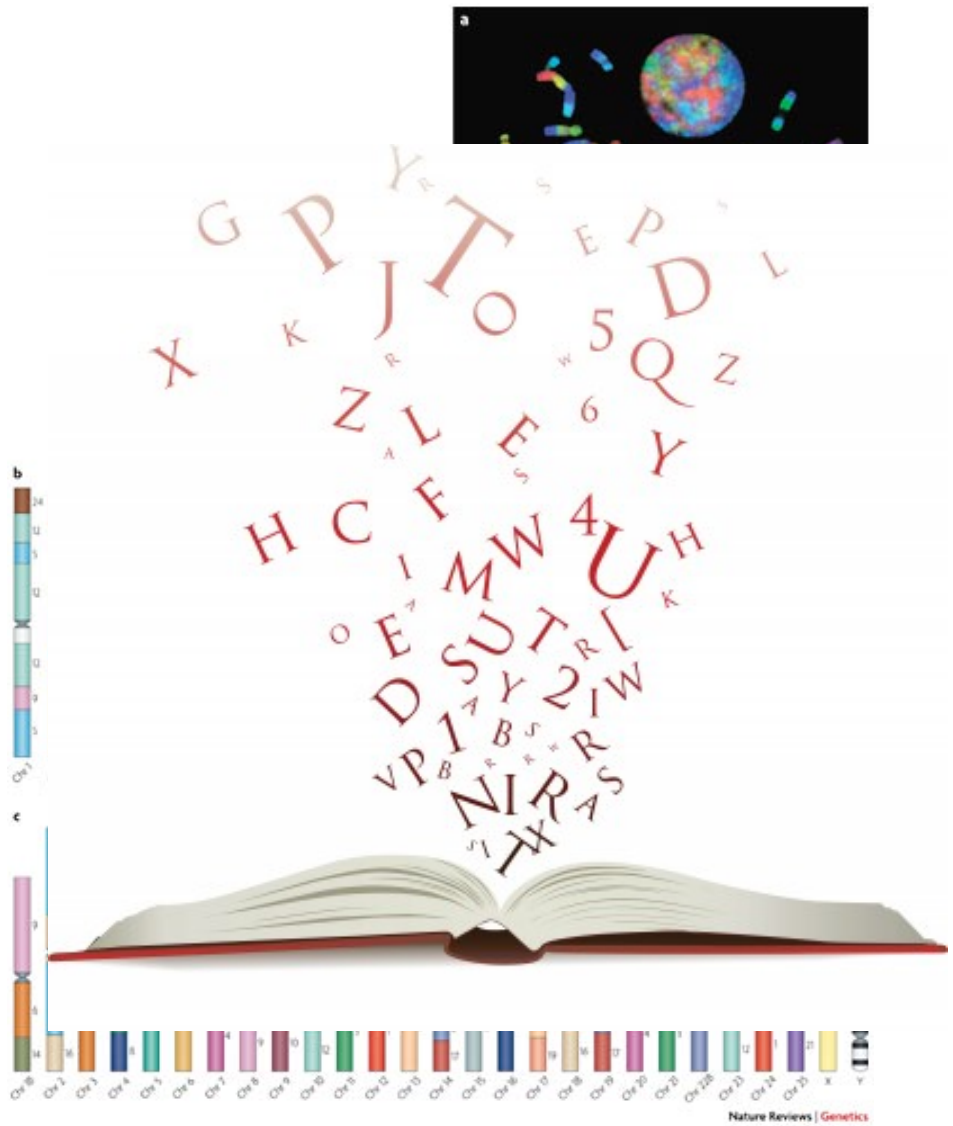
Sequence-structure gap

Despite the efforts from structural genomic projects the number of new structures per year has decreased

NEW sequencing techniques are becoming routinely available to scientists.....

Many genomes have been completely sequenced
During the last 250 years, 1.2 million eukaryotic species have been identified and taxonomically classified.

Number of species estimated to exist on Earth: bacterial and archaea species, from 100,000 to 10 million; eukaryotic species, approximately 8.7 million.


a
b
c
Nature Reviews | Genetics

-

sequence

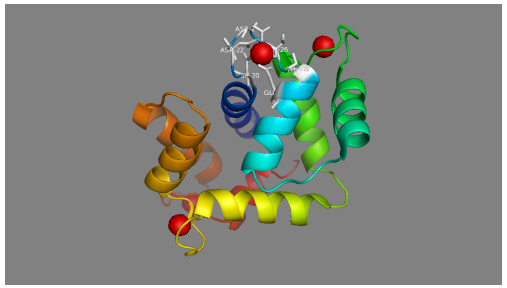
structure

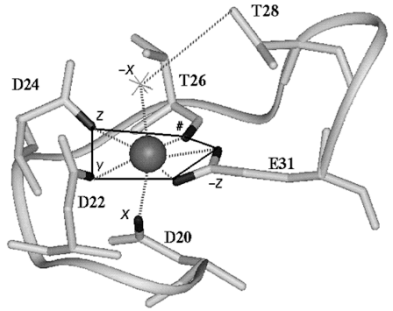
+

function

conservation

```
YT67_CAEEL/162-190      LAEKVFRELDVSANGHLSADQFATIVEDY
SCP_PERVT/129-157       MAPASFDAIDTNNDGLLSQEEFVTAGSDF
O96676_DROME/249-277    DAAVSFAVIDKNGDGQLSLKEFVHLGRQF
LBP_RENRE/138-166       KAITCFNTLDFNKNGQISRDEFLVTVNDF
```

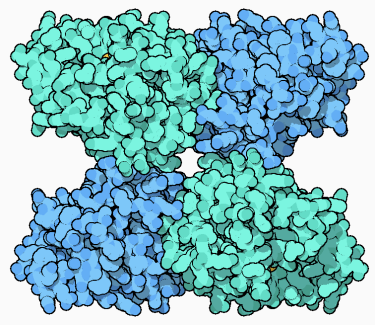




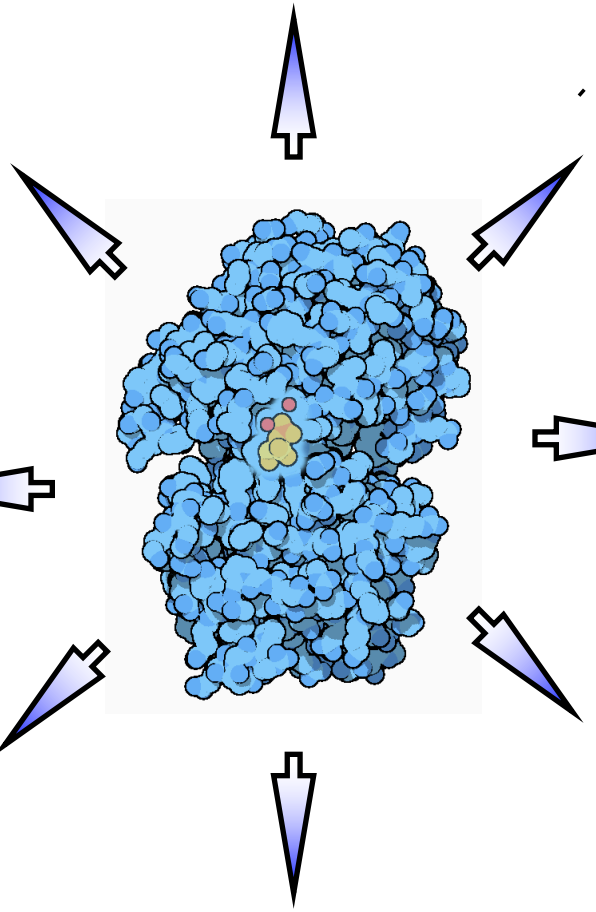
Function is attributed to very few atoms absolutely conserved during the evolutionary process

The Centrality of a 3D Structure

biological multimeric state



crystal packing: putative interaction sites

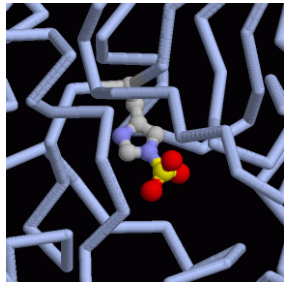


evolutionary relationships

```
>KPYM_HUMAN/409-529
TEATAVGAVEASFKCCSGAIIVL
>A6TVD2_ALKMQ/356-471
TDAISHATCSTASDLQASAILIA
>G9SRZ3_ENTFC/365-480
TETIGLSVARAAKNLGVKTIVAA
>E7C344_9BACT/358-465
--KMLRSAAMLAMRMKNAAVLVF
```

mutants (SNPs) and conservation

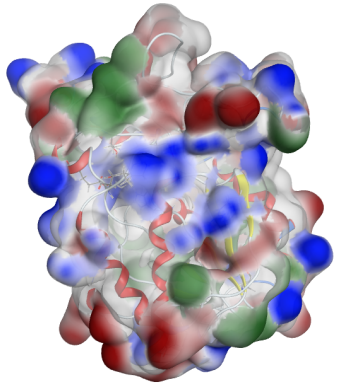
```
>KPYM_HUMAN/409-529
TEATAVGAVEASFKCCSGAIIVL
```



ligands and functional sites

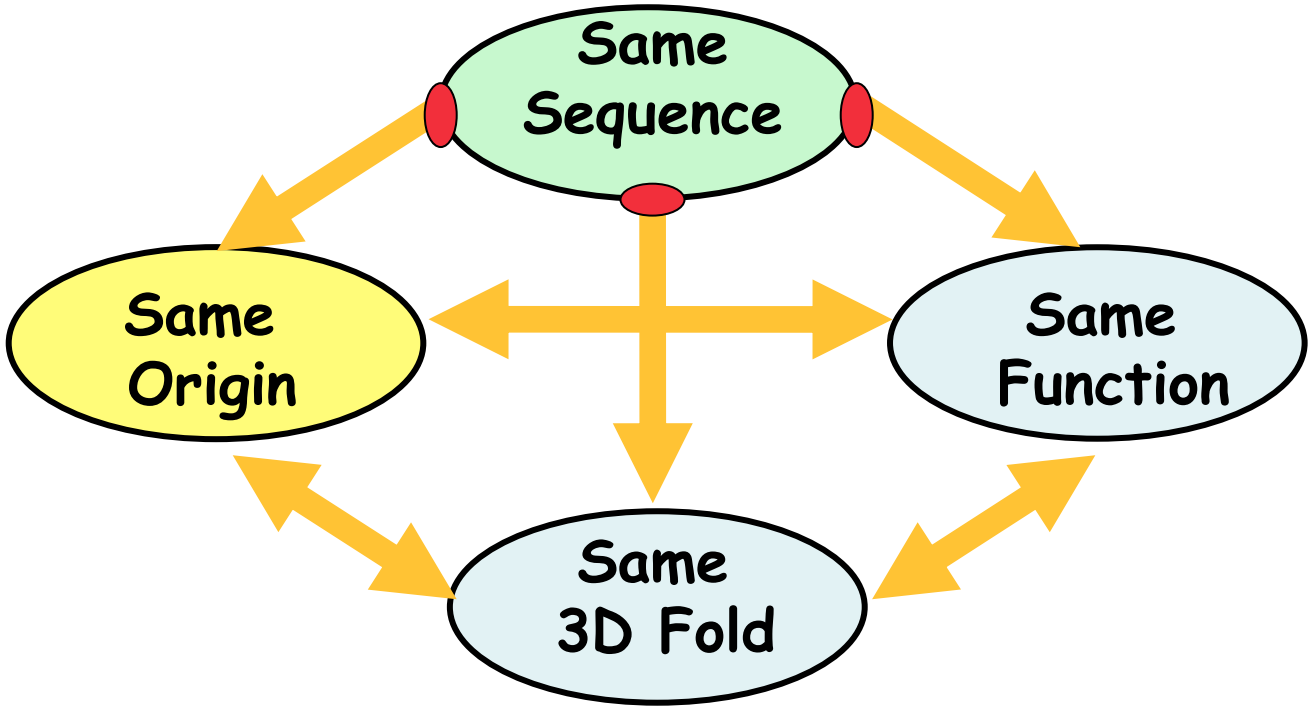
structural/functional motifs

interaction sites, catalytic sites



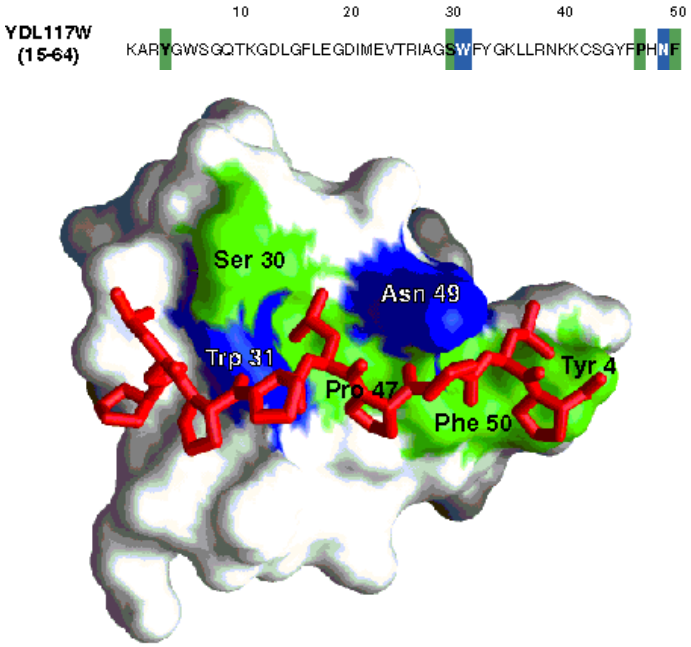
surface properties: electrostatics, clefts, patches

Why Does It Make Sense To Align Sequences ?



Protein Structure Prediction: sequence vs structure

- The biochemical function (activity) of a protein is defined by its interactions with other molecules.
- The biological function is in large part a consequence of these interactions.
- The 3D structure is more informative than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence.



As mentioned before structure is more conserved in evolution than sequence.

The net result is that patterns in space are frequently more recognizable than patterns in sequence.

Protein Structure Prediction: the gap in numbers

Why Protein Structure Prediction?

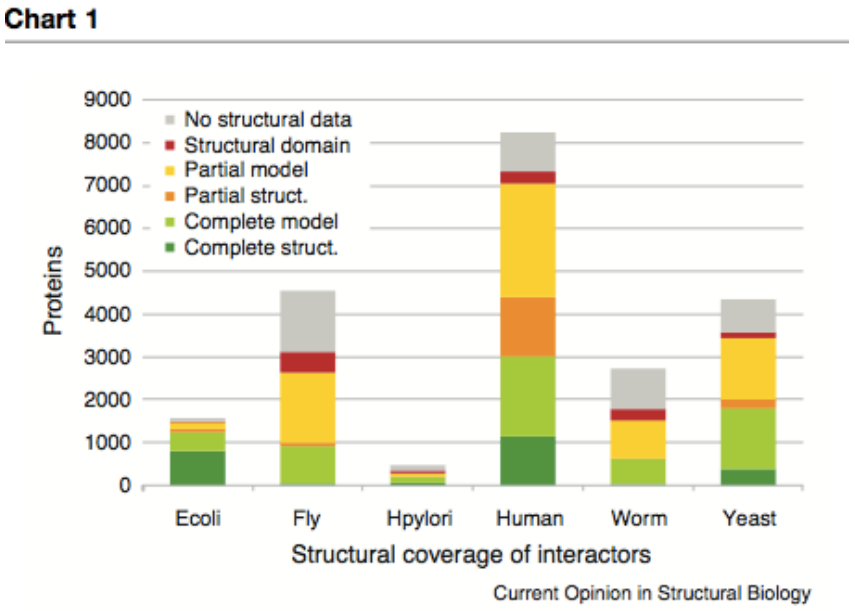
	2018
Sequences (UniProt)	~556,000
Structures (PDB)	~14,000

UniProt: repository of protein sequences (www.uniprot.org)

PDB: repository of protein structures (www.rcsb.org)

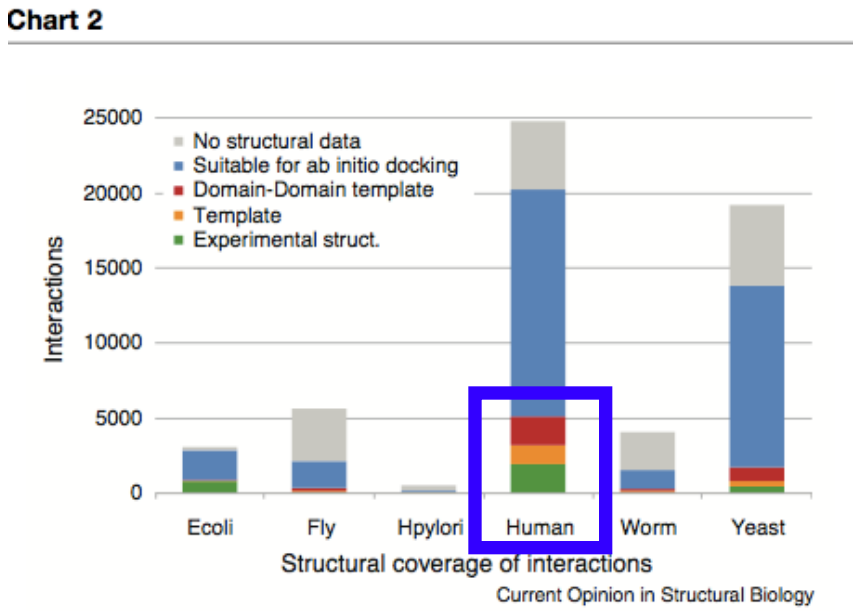
Protein Structure Prediction: the gap in numbers

Single protein structures



Structural coverage of interactors.

Interaction protein pair structures

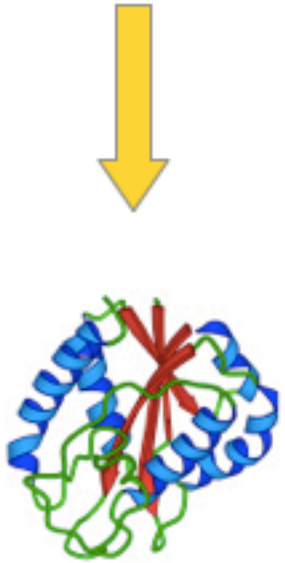


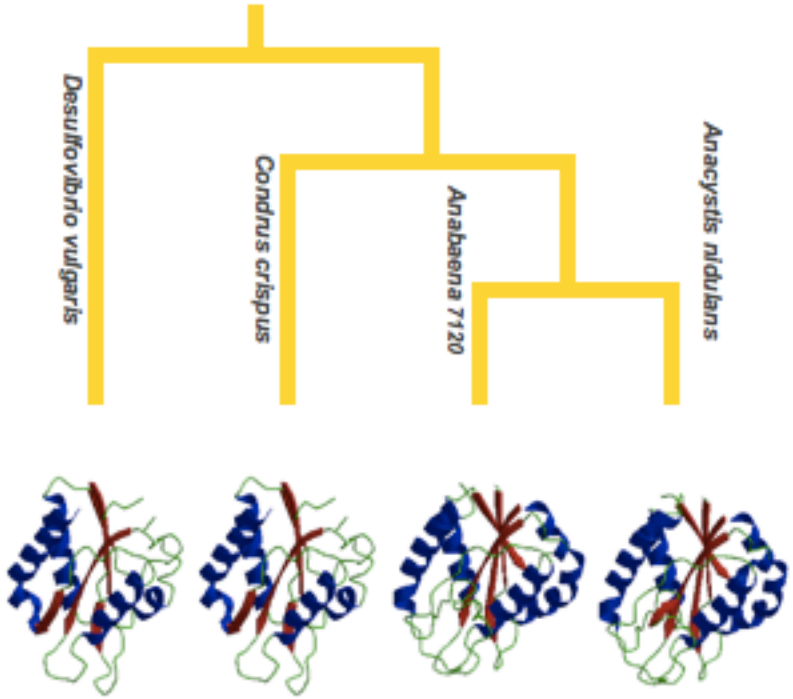
Structural coverage of interactions.

Stein, A., Mosca, R. & Aloy, P. Three-dimensional modeling of protein interactions and complexes is going ‘omics. *Curr. Opin. Struct. Biol.* **21,** 200–208 (2011).

Protein Structure Prediction: Principles

GFCHIKAYTRLIMVG…





Folding

(physics)

***Ab initio* prediction**

Evolution

("statistical" rules)

Threading
Comparative Modeling

Protein Structure Prediction: Principles

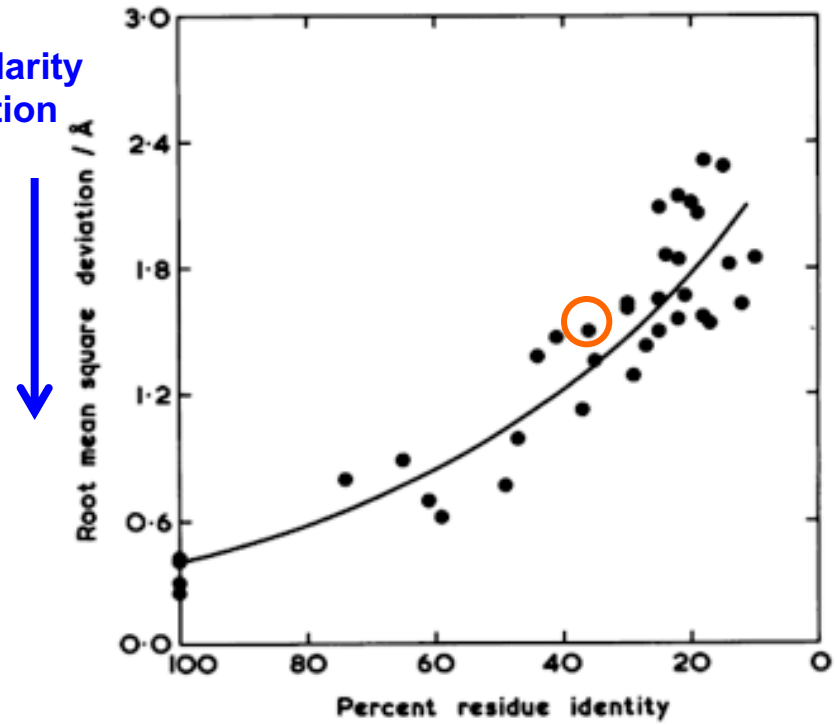
During evolution protein structure is more conserved than sequence

The EMBO Journal vol.5 no.4 pp.823–826, 1986

The relation between the divergence of sequence and structure in proteins

Cyrus Chothia[1] and Arthur M.Lesk[2]

Measure of structure similarity (Root Mean Square Deviation or RMSD)



Measure of sequence similarity (Sequence Identity)

Levitt conformational preferences of aa in globular proteins Biochemistry (1978) 17, 427

TABLE VI: Conformational Preferences[a] of Amino Acids for α Helix, β Sheet and Reverse Turns.

type of secondary struct	favoring (h)	indifferent (i)	breaking (b)
α helix	Ala, Leu, Met, His. Glu, Gln. Lys, (Cys)	Val. Ile. Phe, Trp. Asp, Asn, Arg	Tyr. Thr. Gly, Ser. Pro
β sheet	Val, Ile, Phe, (Trp), Tyr, Thr	Ala, Leu. Met, His. Gly, Ser. Arg	Glu. Gln. Lys, Asp, Asn, Pro. cys
reverse turn	Gly. Ser. Asp. Asn. Pro	Gly. Gln, Lys, Tyr. Thr. (Arg)	Ala. Leu. Met. His, Val, Ile. Phe, (Trp). (Cys), (Arg)

[a] These preferences are assigned with at least 75% confidence for the h and b classes (24 out of 40 are with at least 95% confidence). unless the amino acid is enclosed in parentheses when the confidence is as low as 56%. The confidence with which the i structure code is assigned is generally lower than for the h and b structure codes.

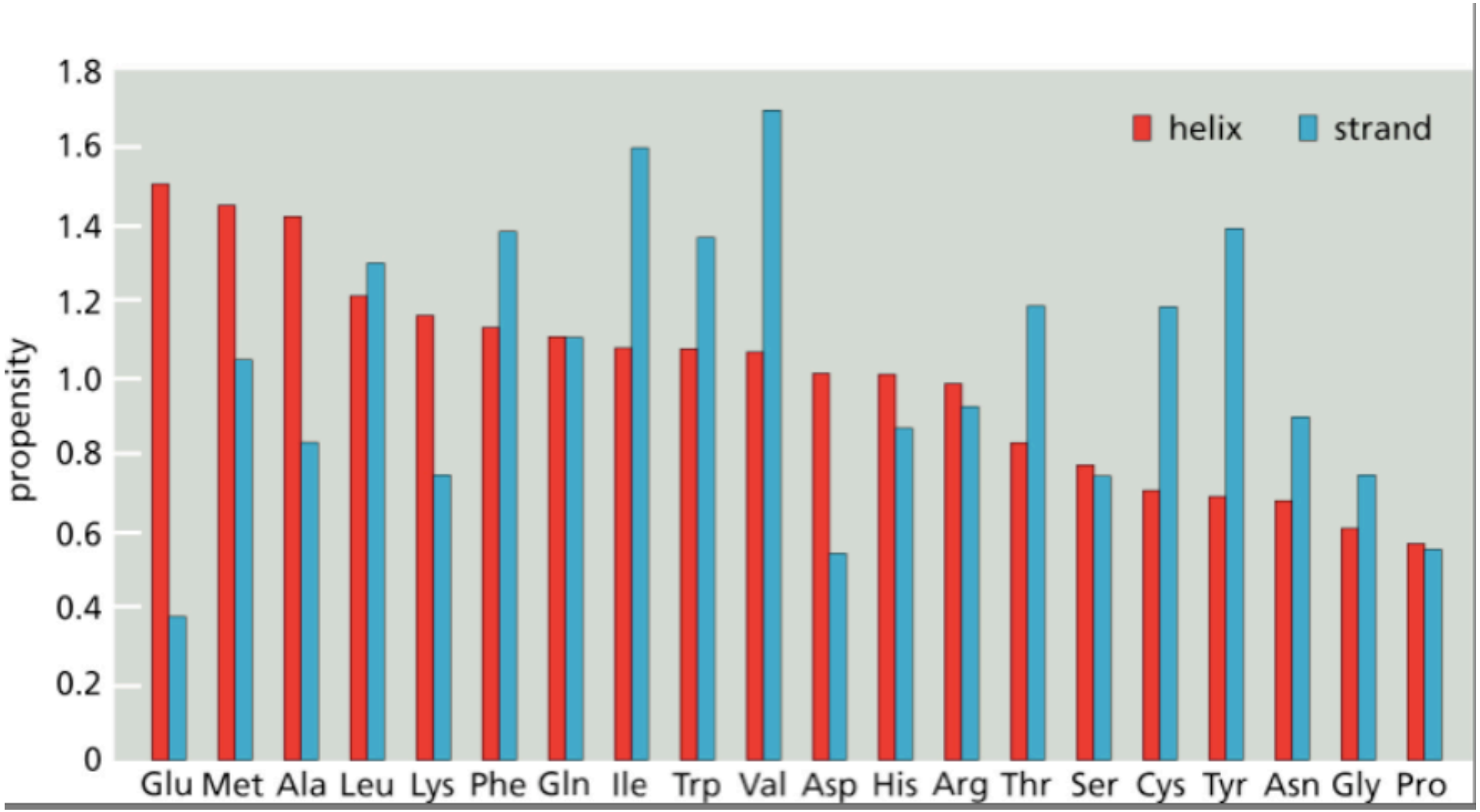
In 1974 there were insufficient data (less than 2500 residues) to accurately determine the values.

The propensities were recalculated several times as more data became available, In 1998 using a dataset of over 33,000 residues leading to some notable differences in the two sets of propensities.

By 2004 the proteins structure datasets were large enough to derive residue propensities at different positions within alpha-helices.

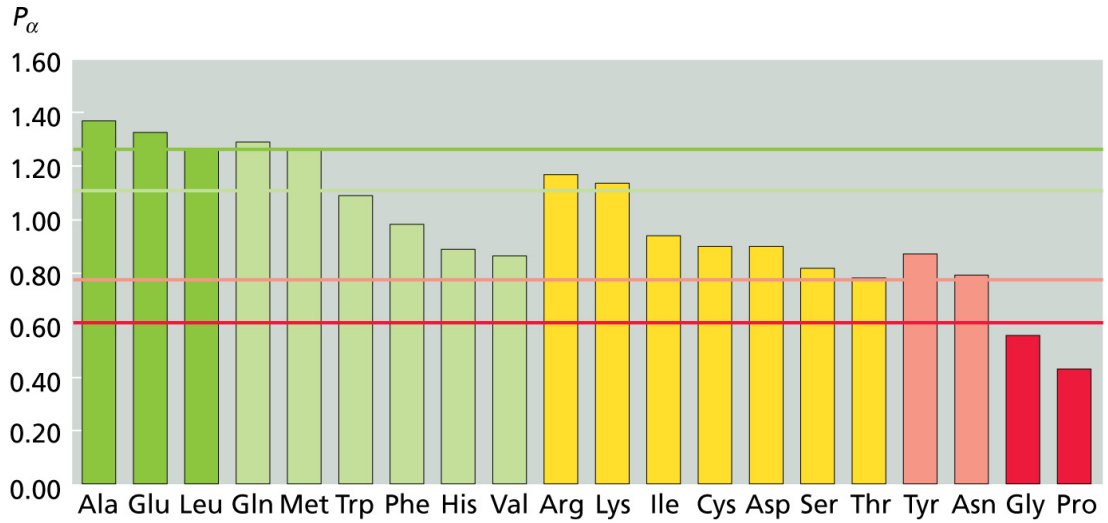
One of the most accurate determinations of beta-turns residue propensities was by the group of Janet Thornton in 1999 based on almost 4000 beta-turns.

Conformational propensities

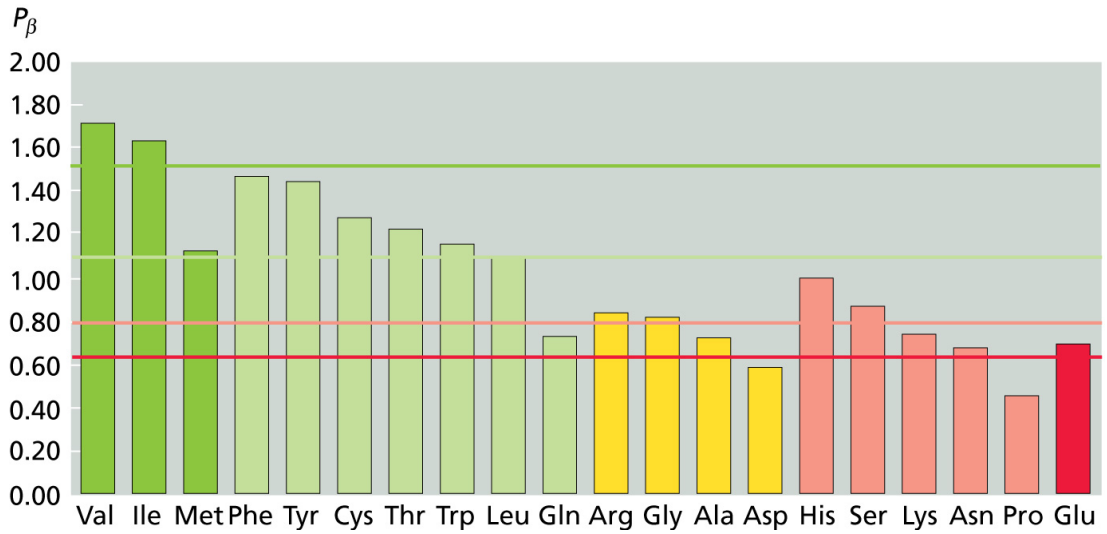


Conformational propensities

alpha-helix propensities

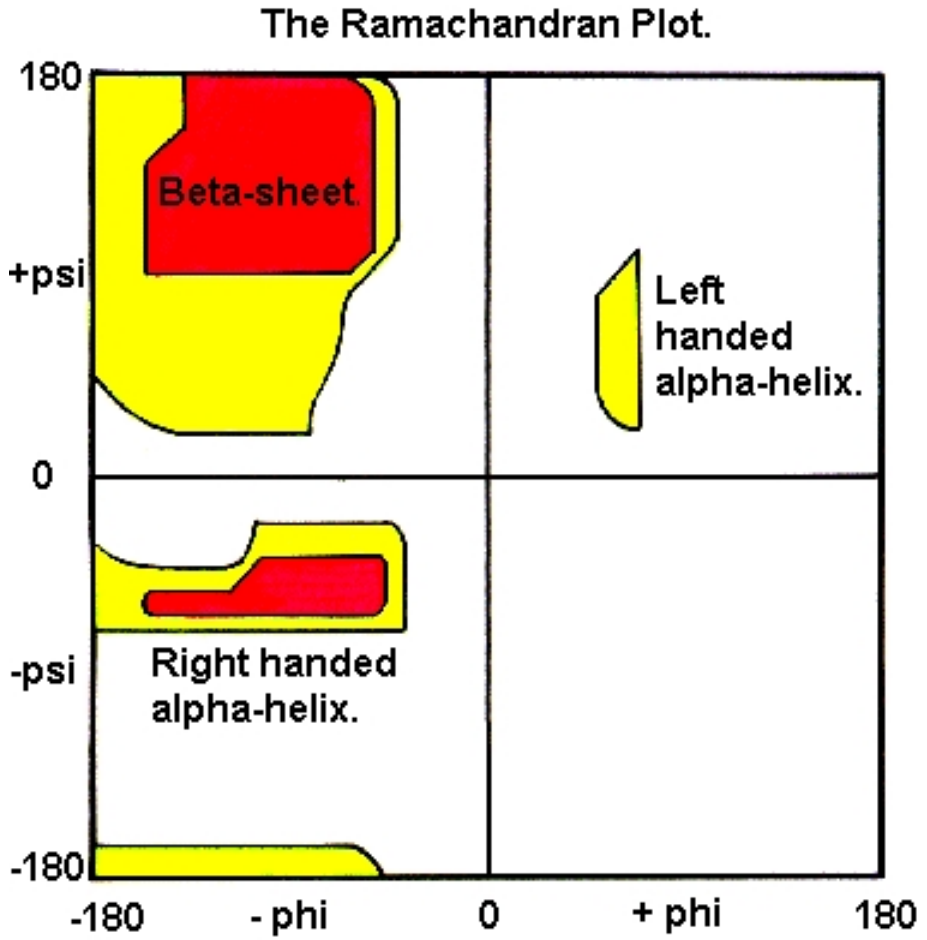


beta-sheet propensities



strong formers, formers, indifferent, breakers, strong breakers

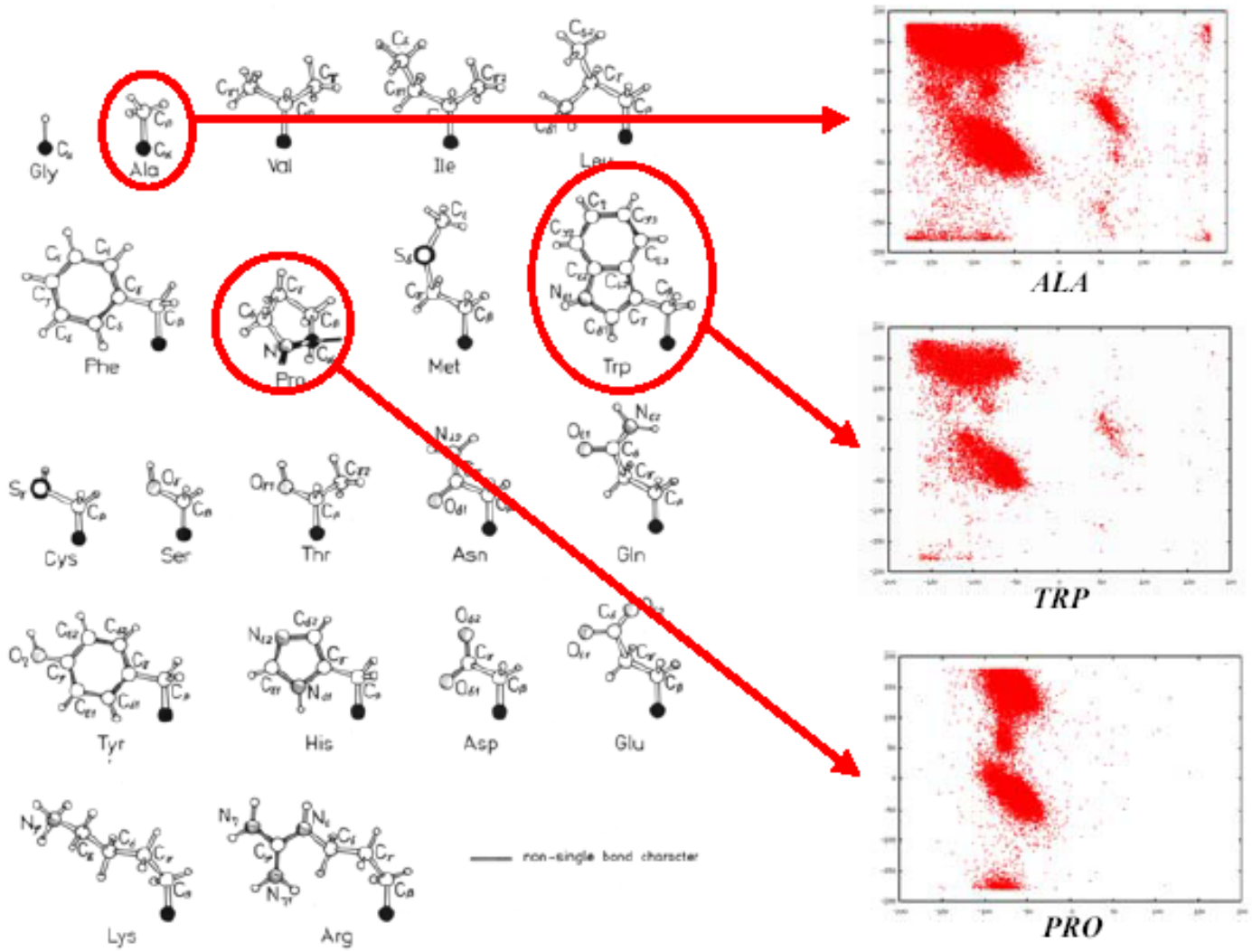
Ramachandran Plot



Projection of a torus onto a plane.

Note the clustering of low-energy states of single residues.

Conformational propensities



Secondary Structure Propensities

(Koehl-Levitt, 1999)

Alpha-helix propensity derived from designed sequences

A/L	R/K	N/M	D/F	C/P	Q/S	E/T	G/W	H/Y	I/V
-0.04	-0.30	0.25	0.27	0.57	-0.02	-0.33	1.24	-0.11	-0.26
-0.38	-0.18	-0.09	-0.01	0.	0.15	0.39	0.21	0.05	-0.06

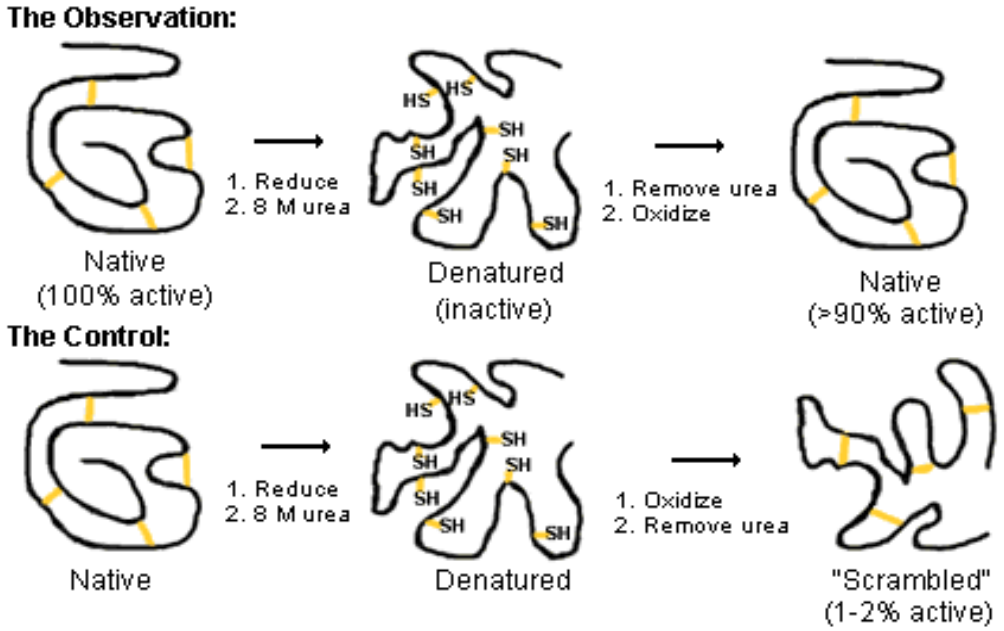
Beta-sheet propensity derived from designed sequences

A/L	R/K	N/M	D/F	C/P	Q/S	E/T	G/W	H/Y	I/V
-0.12	0.34	1.05	1.12	-0.63	1.67	0.91	0.76	1.34	-0.77
0.15	0.29	-0.71	-0.67	0.	1.45	-0.70	-0.14	-0.49	-0.70

http://www.genome.jp/aaindex/

Anfinsen "Thermodynamic Hypothesis" states that the native conformation of a protein is adopted spontaneously.
In other words, there is sufficient information contained in the protein sequence to guarantee correct folding from any of a large number of unfolded states.

The Anfinsen experiment

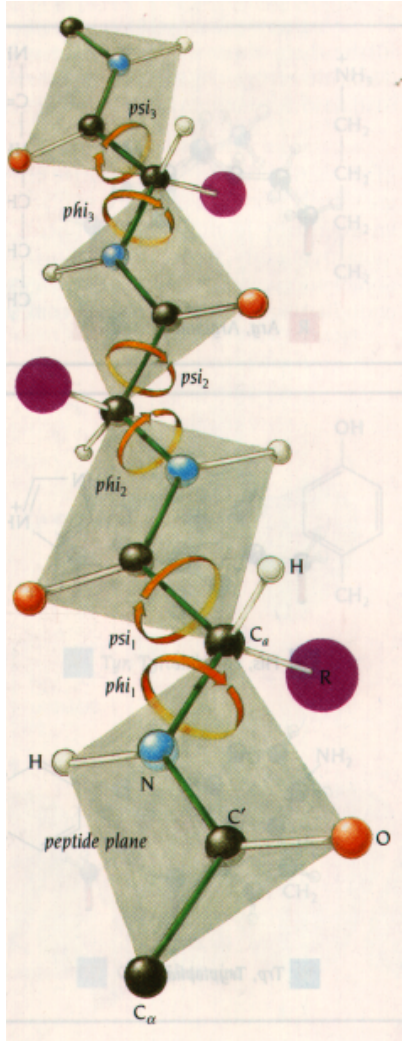


A reasonable objection can be raised to the above result by suggesting that perhaps RNaseA was not completely unfolded in 8 M urea.
To address this class of objections, RNAseA was first reduced and denatured as above.
But in the second phase, the enzyme was first oxidized to form S-S bonds, and then the urea was removed, i.e. the order of steps in the second phase of the experiment was reversed.
The resulting activity was only about 1-2% of the untreated enzyme.
Sequence analysis showed a random assortment of S-S

Protein Structure Prediction

- In theory, a protein structure can be solved computationally
- A protein folds into a 3D structure to minimizes its free potential energy
- The problem can be formulated as a search problem for minimum energy
  - the search space is enormous
  - the number of local minima increases exponentially

Computationally it is an exceedingly difficult problem



Levinthal paradox

In 1969 Cyrus Levinthal noted that, because of the very large number of degrees of freedom in an unfolded polypeptide chain, the molecule has an astronomical number of possible conformations.

The estimate 3^{300} or 10^{143} appears in the original article. If the protein is to attain its corrected folded configuration by sequentially sampling all the possible conformations, it would require a time longer than the age of universe to arrive at its correct native conformation.

Levinthal himself was aware that proteins fold spontaneously and on short timescales, and that a random conformational search is therefore impossible.

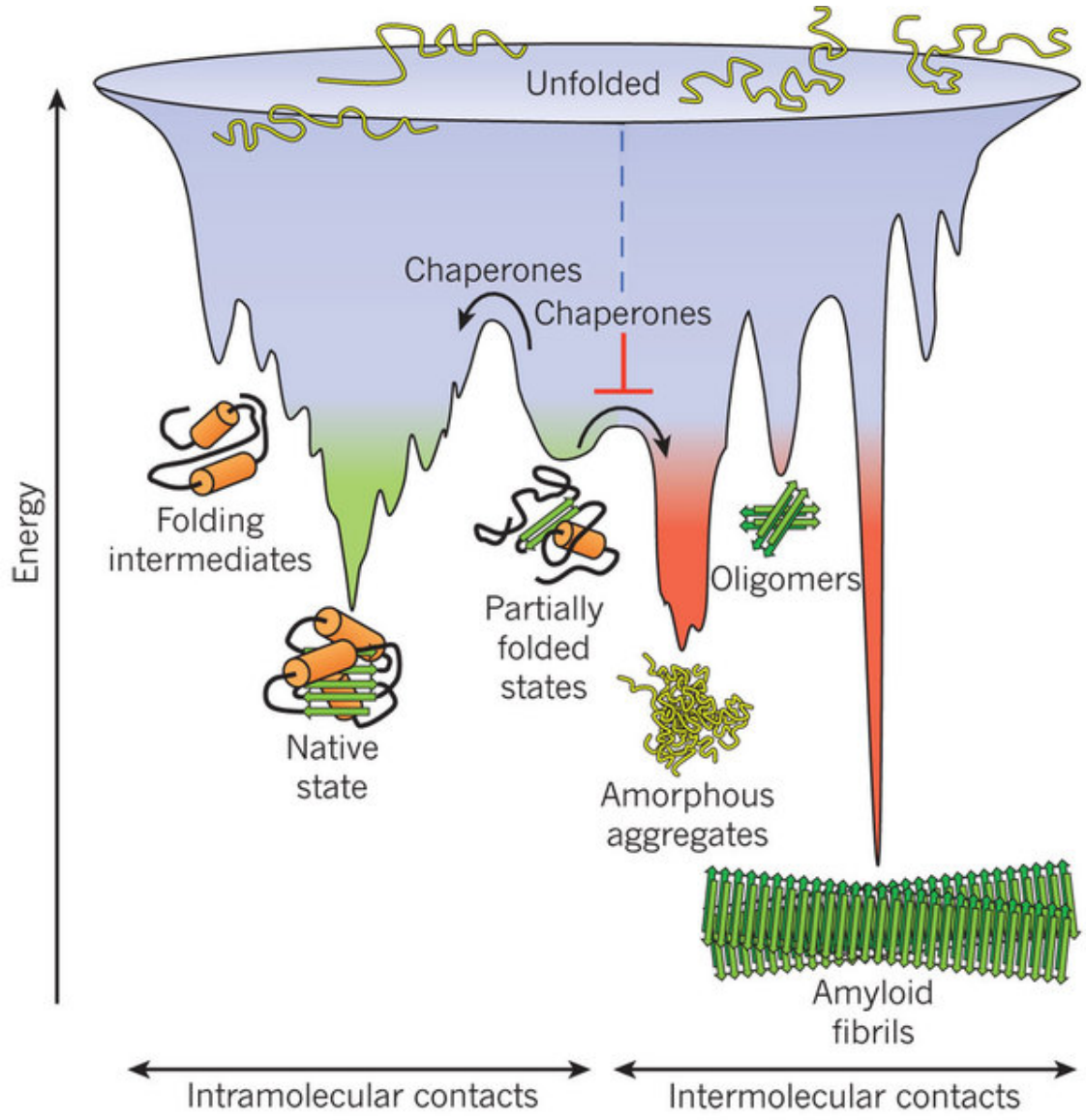
Christian B. Anfinsen's 1971 Nobel Prize lecture revisits some of the same themes.

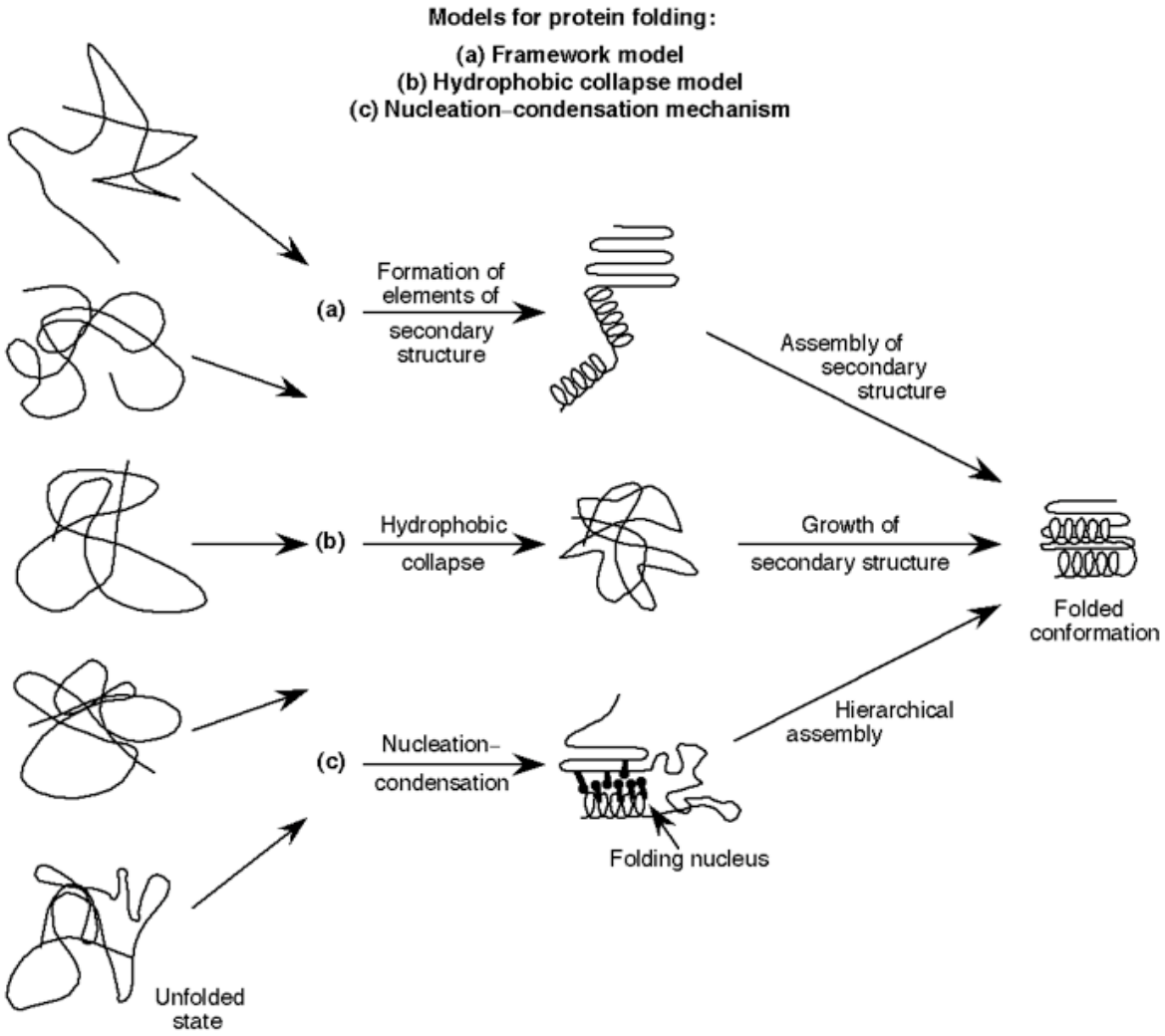
From Wikipedia

Protein folded states: ‘explored’ as contained in the PDB structures

- But how does a protein fold?

- According to Anfinsen and Levinthal a protein cannot visit all the possible ϕ and Ψ values before finding the native structure


Unfolded
Chaperones
Chaperones
Energy
Folding intermediates
Native state
Partially folded states
Oligomers
Amorphous aggregates
Amyloid fibrils
Intramolecular contacts
Intermolecular contacts



It is likely that folding mechanisms vary significantly according to protein size, stability and structure.

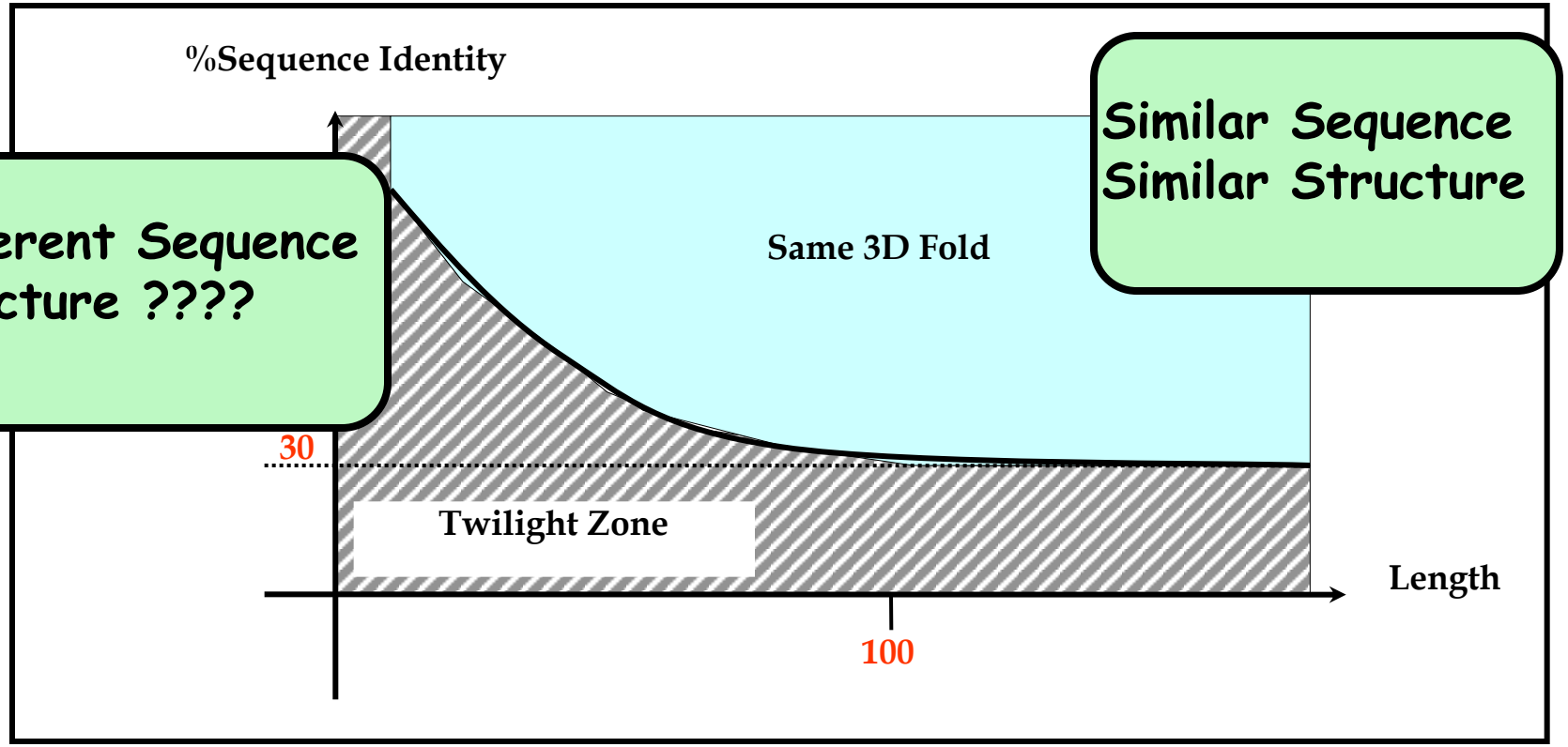
The nucleation-condensation model has been supported by experimental evidence from several small proteins including chymotrysin inhibitor-II and barstar.

Bychkova and Ptitsyn have studied more than 20 proteins and found that nearly all adopted a molten globule state under mild denaturing conditions.

This points to the hydrophobic collapse model, a model favoured by many for the case of larger proteins.

But what if …
we explore the ‘knowledge’ of the structurally determined protein folded states: the ‘explored’ ones, contained in the PDB structures

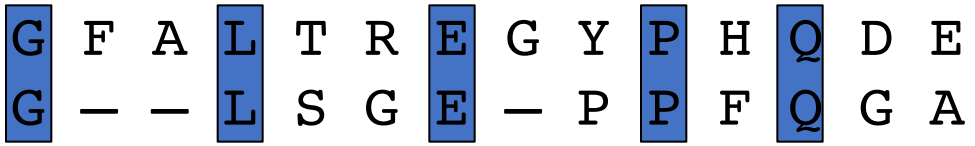
How Can We Compare Sequences ?
The Twilight Zone



Some Basic Principles: Sequence identity

How is sequence identity defined?

It is the **fraction** of identical amino acids correctly aligned



14 residues 5 of which identical in the aligned positions:
5/14 x100 = 35% identity

Structure Prediction: state of the art

Ab initio folding (force-field and simulation based)

1998 Duan and Kollman :36 residues, 1000 ns, 256 processors, 2 months

Recently examples of folding small proteins via computer simulations has been achieved V. PANDE (see movie)

Ab initio folding (knowledge-based scoring functions)

Rosetta (BAKER)

I-Tasser (ZHANG)

Deep Learning methods (from 2018)

Template-based (or knowledge-based) methods

- Homology modeling: sequence-sequence alignment, works if sequence identity > 30%
- Threading
- Protein threading: sequence-structure alignment, can go beyond the 25% limit

Comparative modeling overview

- Why build comparative models?
  - Many more sequences available than structures (millions vs. tens of thousands)
  - Many applications (e.g. determination of function) rely on structural information
  - Structure is often more conserved than sequencesince evolution tends to preserve function,

Comparative modeling overview

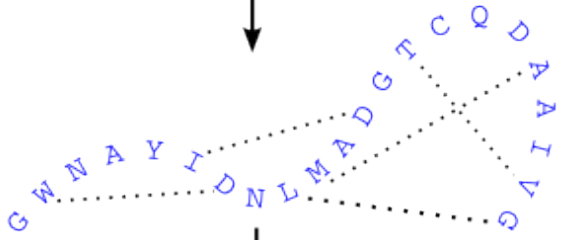
- How does it work?
  - Extract information from known structures (one or more templates), and use to build the structure for the ‘target’ sequence
  - Should also consider information from other sources: physical force fields, statistics (e.g. PDB mining)
- Classes of methods for comparative modeling
  - Assembly of rigid bodies (core, loops, sidechains)
  - Segment matching
  - Satisfaction of spatial restraints

Comparative modeling by satisfaction of spatial restraints - MODELLER

1. Align sequence with structures

Template structure(s) SWQTYVDTNLVGTGAVTQA--AI
Target sequence -GWNAYIDNLMADGTCQDAAIVG

2. Extract spatial restraints



3. Satisfy spatial restraints

A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

Some Basic Principles: Homology

Pragmatic Definition of Homology:

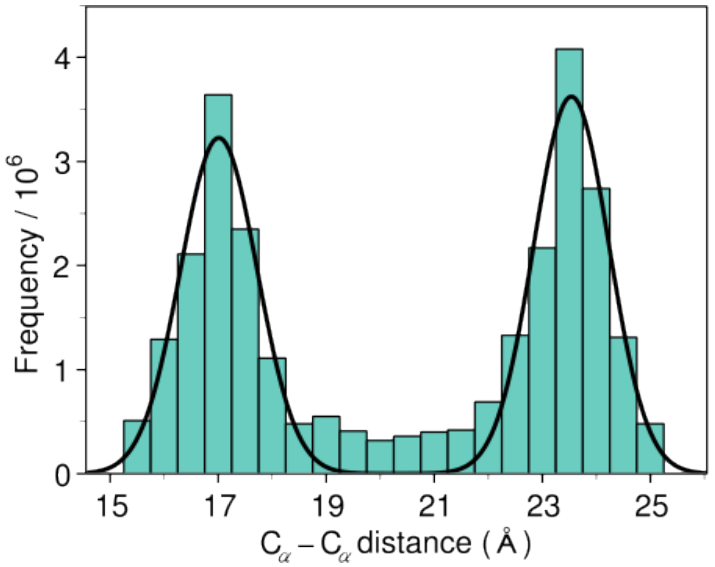
The probability of two sequences to share **more than 30% sequence identity by CHANCE is so low,** one can safely assume that they share **a common ancestor**.

1. Align sequence with structures

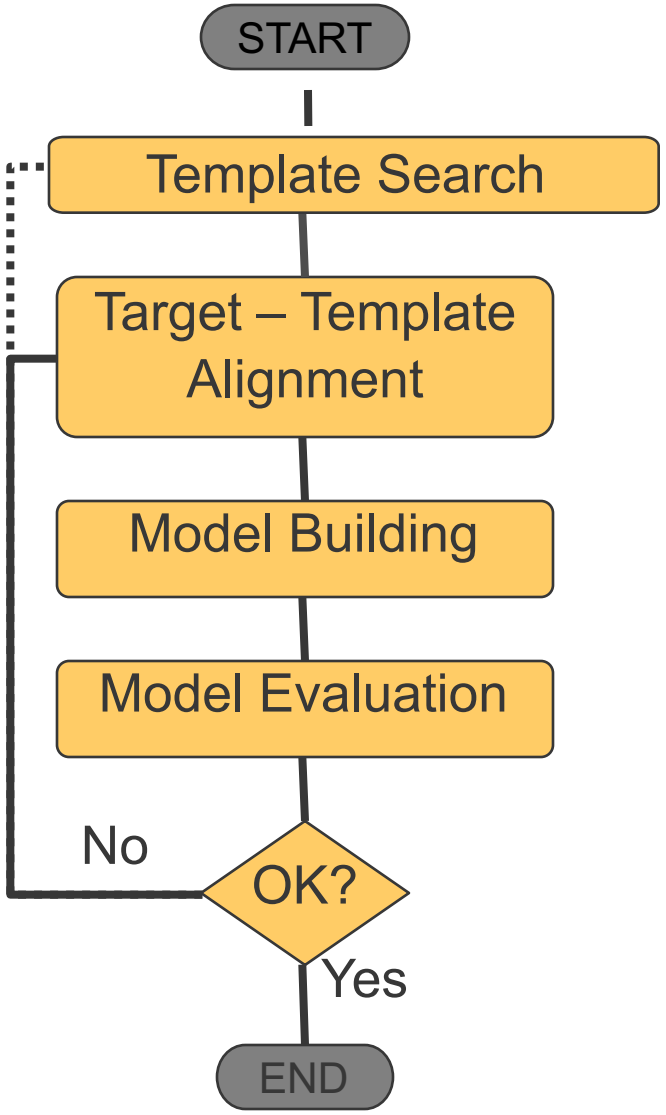
- First, must determine the template structures
  - Simplistically, try to align the target sequence against every known structure's sequence
  - In practice, this is too slow, so heuristics are used (e.g. BLAST)
  - Profile or HMM searches are generally more sensitive in difficult cases (e.g. Modeller's profile.build method, or PSI-BLAST)
  - Could also use threading or other web servers
- Alignment to templates generally uses global dynamic programming
  - Sequence-sequence: relies purely on a matrix of observed residue-residue mutation probabilities ('align')
  - Sequence-structure: gap insertion is penalized within secondary structure (helices etc.) ('align2d')
  - Other features and/or user-defined ('salign') or use an external program

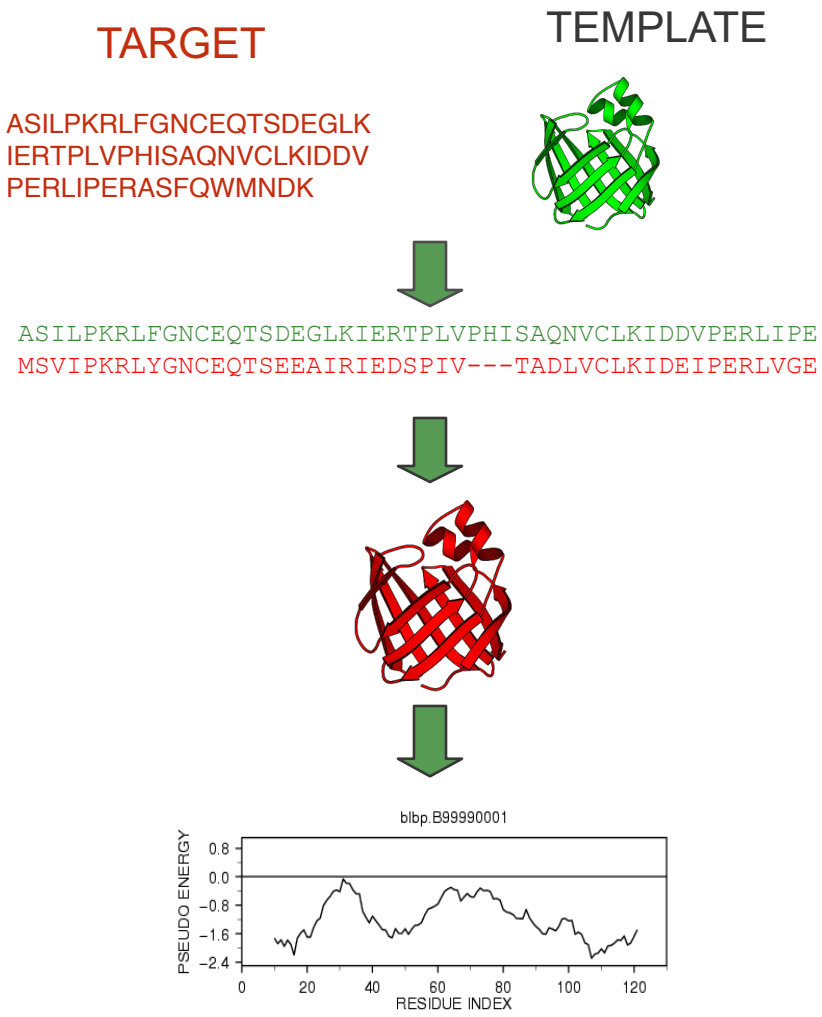
2. Extract spatial restraints

- Spatial restraints incorporate homology information, statistical preferences, and physical knowledge
  - Template Cα- Cα internal distances
  - Backbone dihedrals (ϕ/ψ)
  - Sidechain dihedrals given residue type of both target and template
  - Force field stereochemistry (bond, angle, dihedral)
  - Statistical potentials
  - Other experimental constraints
  - etc.



Comparative Modelling





A. Šali, Curr. Opin. Biotech. 6, 437, 1995.
R. Sánchez & A. Šali, Curr. Opin. Str. Biol. 7, 206, 1997.
M. Marti et al. Ann. Rev. Biophys. Biomolec. Struct., 29, 291, 2000.

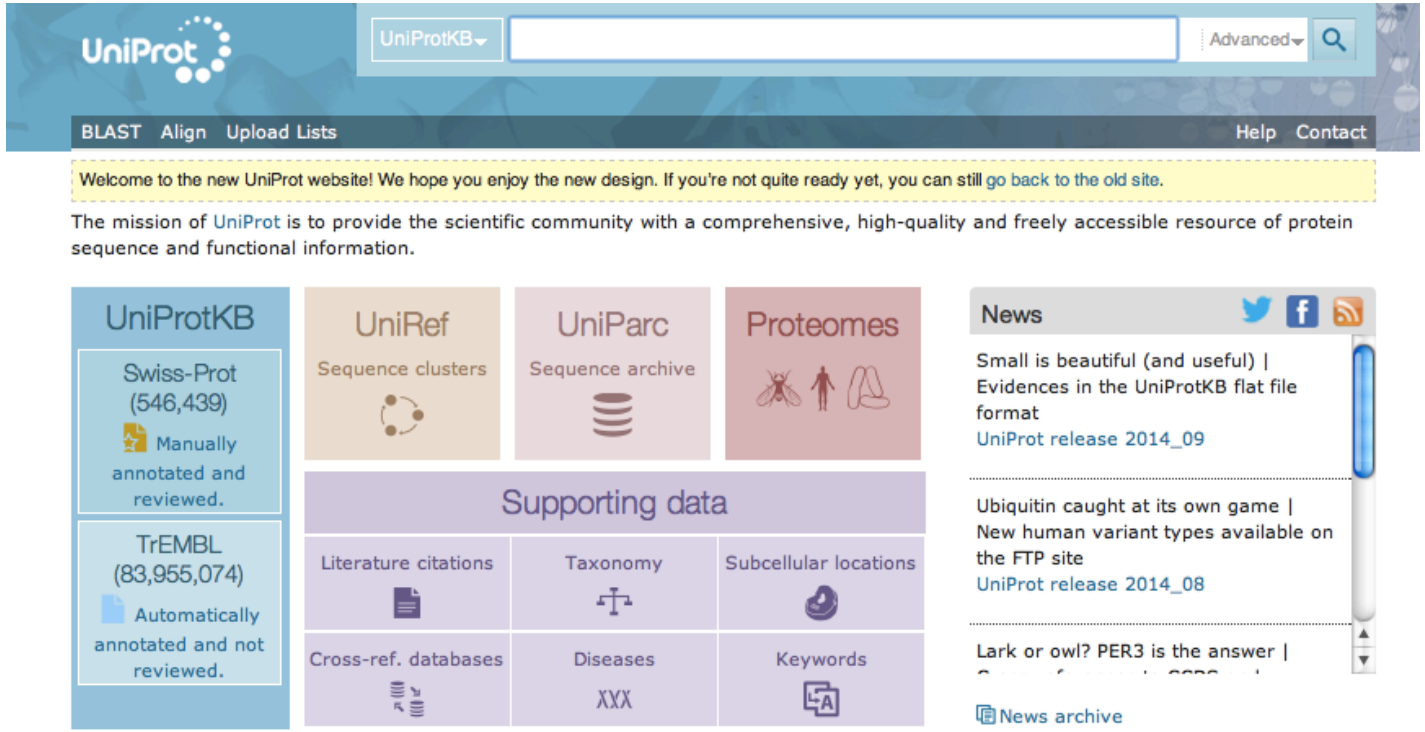
(Adapted from slides in https://salilab.org/modeller/london.zip)

3. Satisfy spatial restraints

- All information is combined into a single objective function
  - Restraints and statistics are converted to an “energy” by taking the negative log
  - Force field (CHARMM 22) simply added in
- Function is optimized by conjugate gradients and simulated annealing molecular dynamics, starting from the target sequence threaded onto template structure(s)
- Multiple models are generally recommended; ‘best’ model or cluster or models chosen by simply taking the lowest objective function score, or using a model assessment method such as Modeller’s own DOPE or GA341, fit to EM density, or external programs such as PROSA or DFIRE

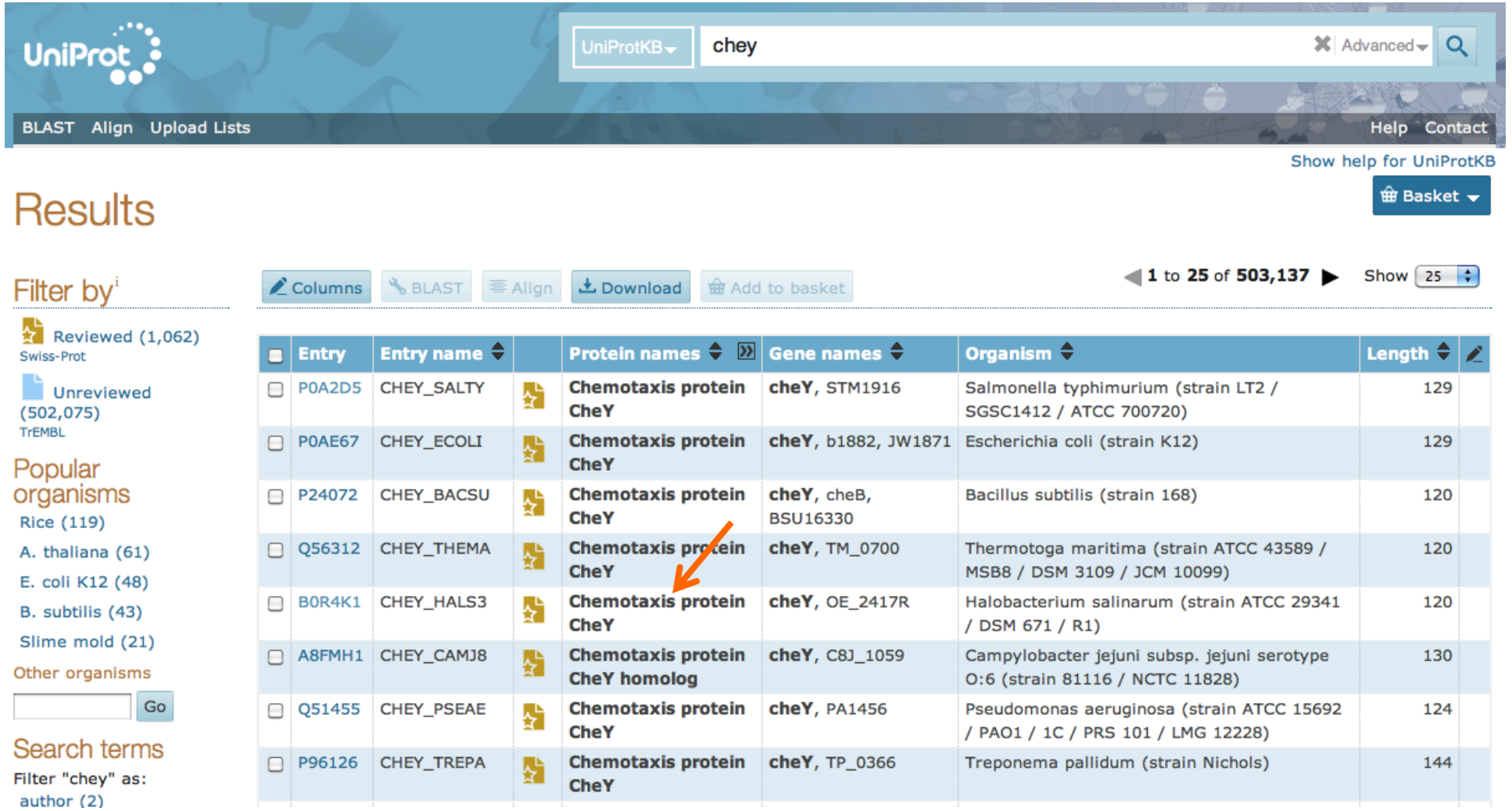
Comparative Modelling: Template Search

The UniProt database

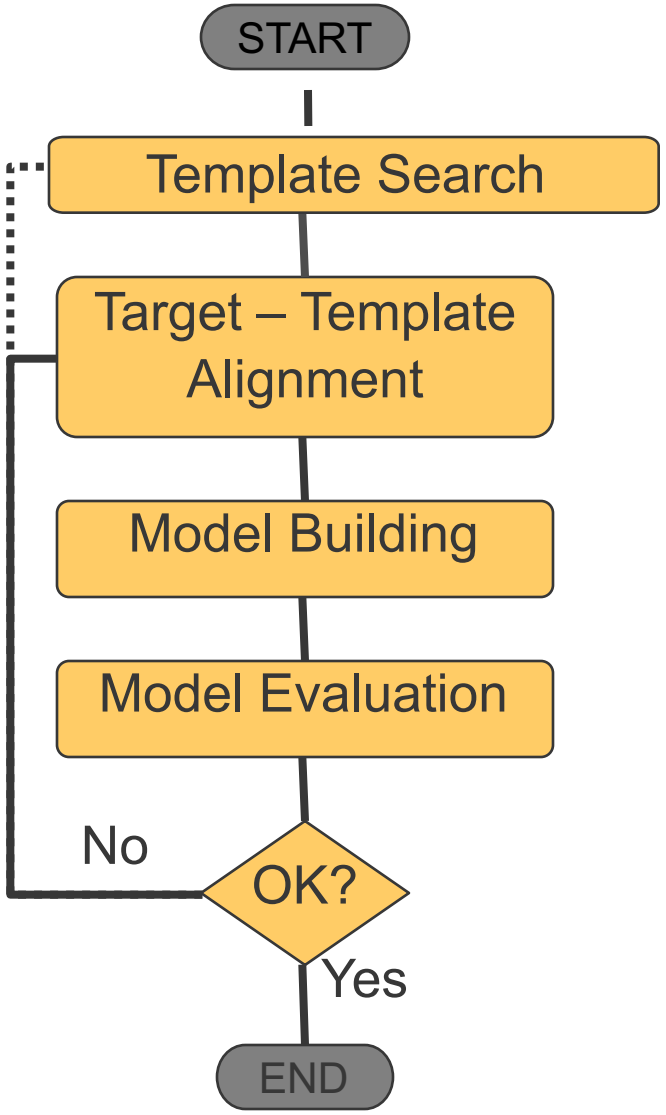


Comparative Modelling: Template Search

The UniProt database



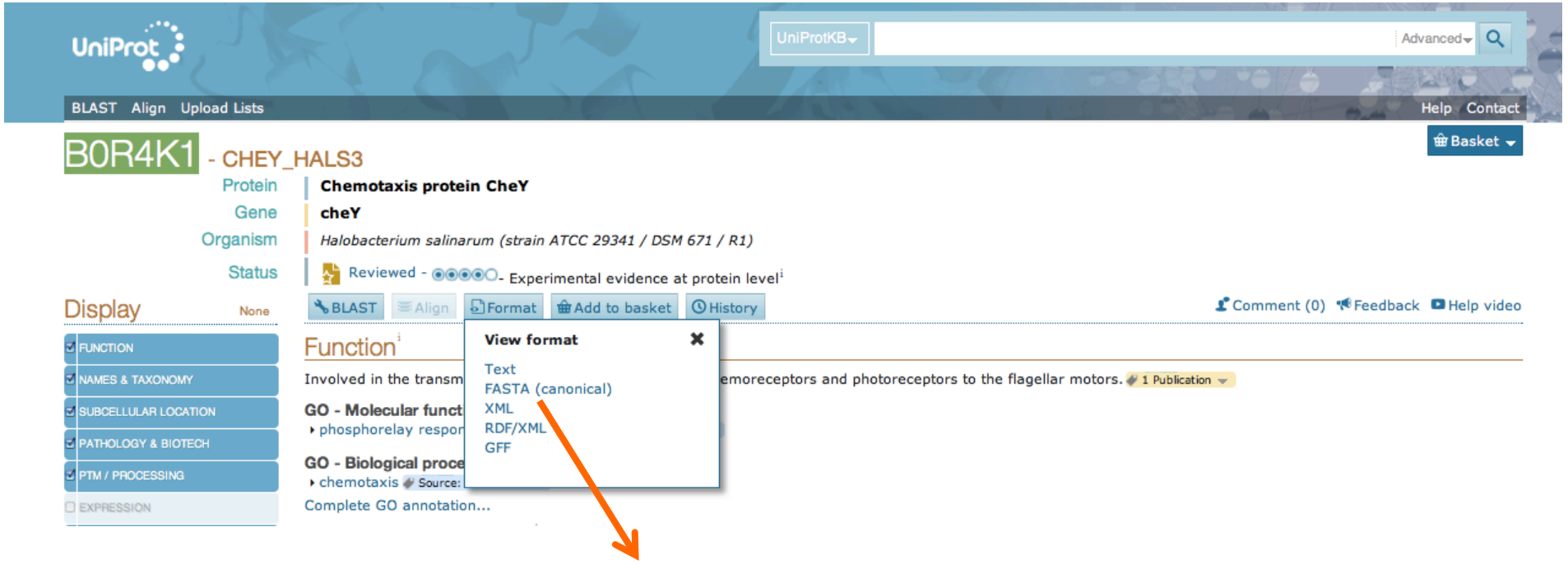
Comparative Modelling



← **Step 0 (retrieve the sequence)**

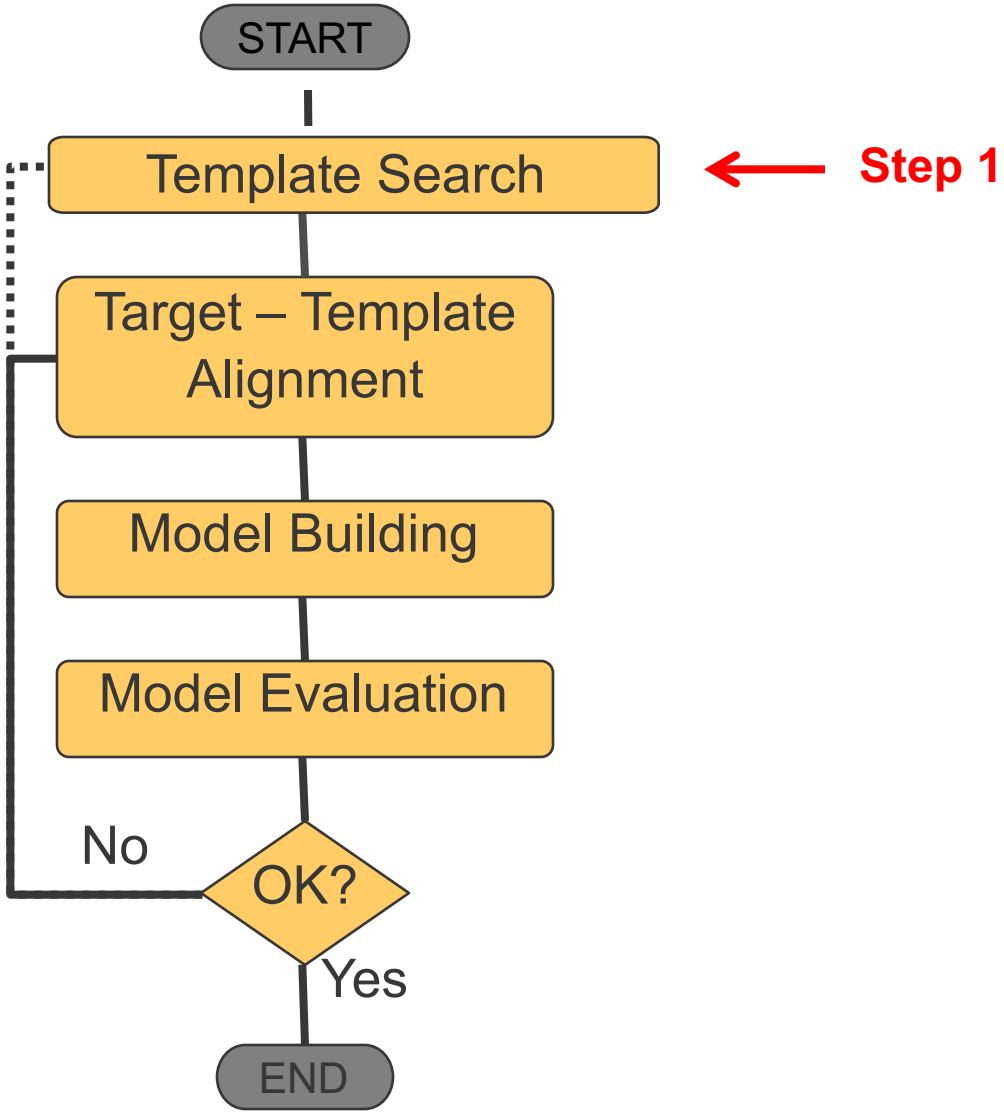
Comparative Modelling: Template Search

The FASTA format




```
>sp|B0R4K1|CHEY_HALS3 Chemotaxis protein CheY OS=Halobacterium salinarum (strain ATCC 29341 / DSM 671 / R1) GN=cheY PE=1 SV=1
MAKQVLLVDDSEFMRNLLREILEEEFEIADEAENGVEAVEMYKEYDPDLVMMDIVMPIRD
GIEATSEIKEFDAGAHIIMCTSIGQEEKMKKAVKAGADGYITKPFQKPSVMDAISDVLTA
```

Comparative Modelling



Comparative Modelling: Template Search

Task:

- Query: new sequence (300 aa)
- Database (searching space): very many sequences
- Goal: find sequences related to query

We want:

- fast tool
- primarily a filter: most sequences will be unrelated to the query
- fine-tune the alignment later

Comparative Modelling: Template Search

BLAST

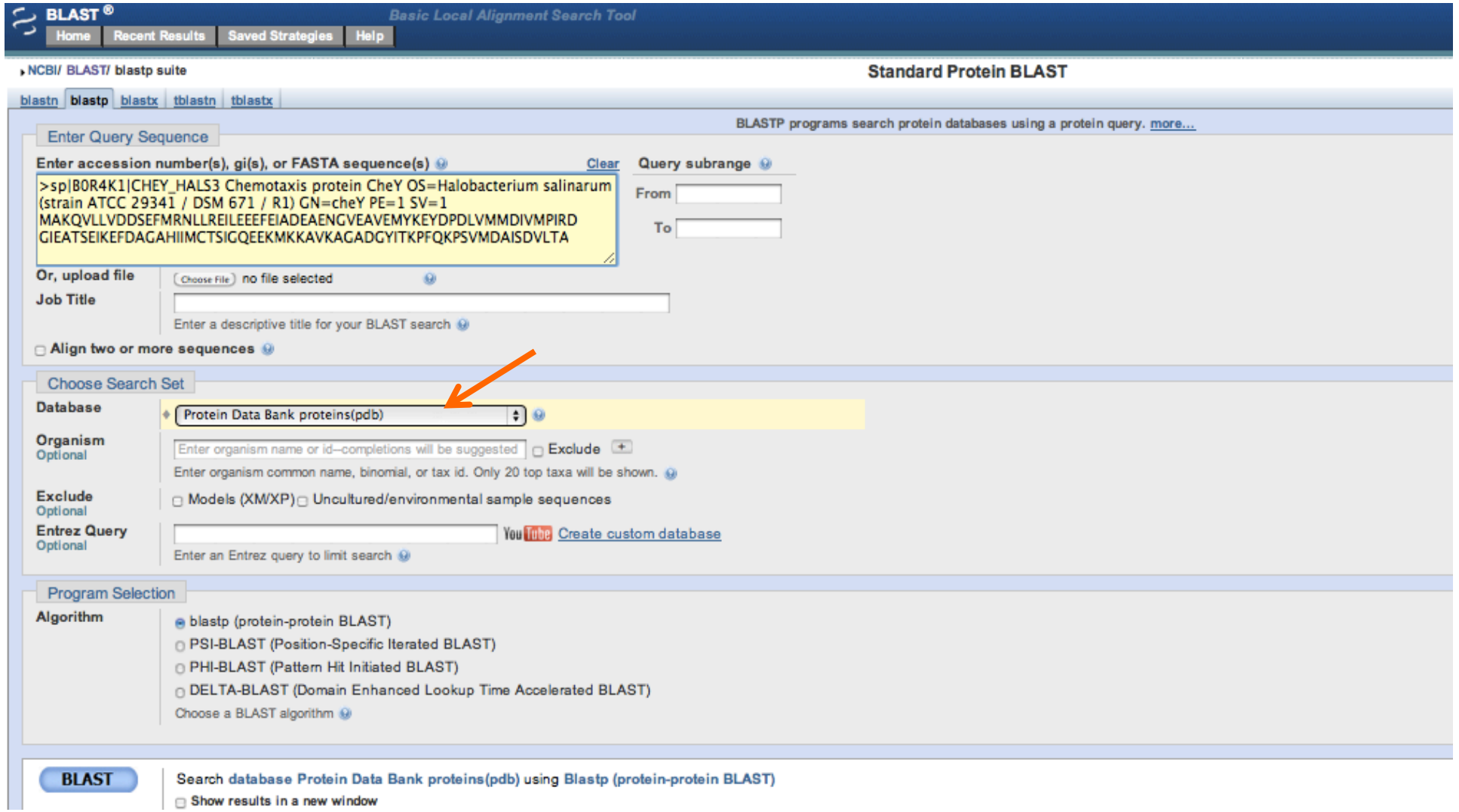
Basic Local Alignment Search Tool

BLAST is a program designed for rapidly comparing your sequence with every sequence in a database and report the most similar sequences

A good general reference is in wikipedia
http://en.wikipedia.org/wiki/BLAST

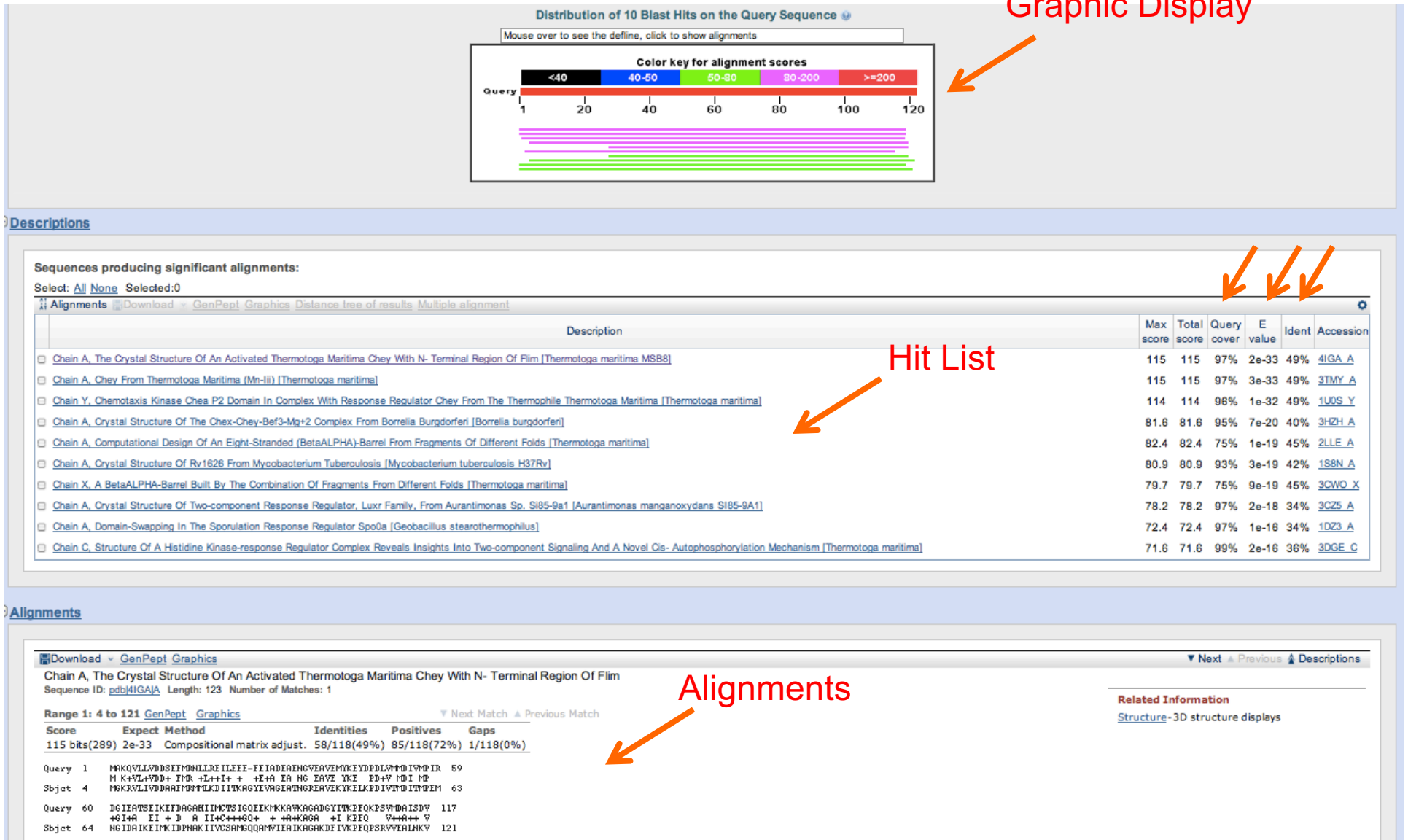
Comparative Modelling: Template Search

BLAST
(http://blast.ncbi.nlm.nih.gov/Blast.cgi)



Comparative Modelling: Template Search

Understanding the BLAST output



Comparative Modelling: Template Search

BLAST Scoring (E-value)

To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone.

In this context, "chance" can mean the comparison of (i) real but non-homologous sequences; (ii) real sequences that are shuffled to preserve compositional properties; or (iii) sequences that are generated randomly based upon a DNA or protein sequence model.

(http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html)

Useful rules of thumb:

The E-value indicates the probability of encountering a false positive (expected by chance) in the sequence hit list. Therefore if one expects many hits from the database (e.g. 100) the E-values should be low enough to exclude the chance of non related hit.

E-values between 10^{-4} and 10^{-6} should be sufficient to exclude this chance.

If one expects few hits from the database (e.g. 1 to 10) the E-values should be high enough to allow distant sequences to be included in the hit list.

E-values between 10^{-1} and 10^{-4} should be high enough for this purpose.

Comparative Modelling: Template Search

Understanding the BLAST output

Hit List

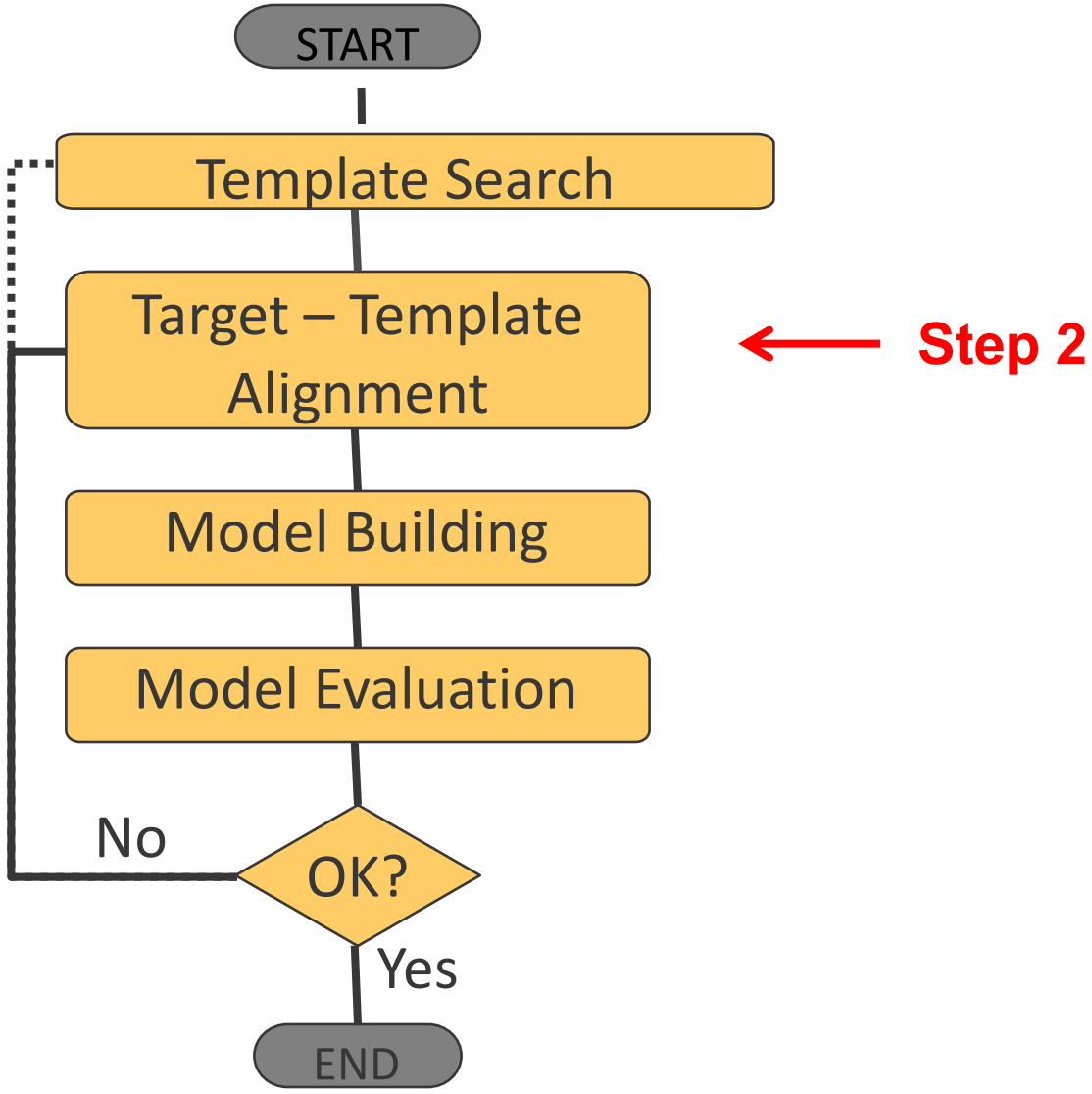
Sequences producing significant alignments:

Select: All None Selected:0

Alignments | Download | GenPept | Graphics | Distance tree of results | Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
Chain A, The Crystal Structure Of An Activated Thermotoga Maritima Chey With N- Terminal Region Of Flim [Thermotoga maritima MSB8]	115	115	97%	2e-33	49%	4IGA_A
Chain A, Chey From Thermotoga Maritima (Mn-Iii) [Thermotoga maritima]	115	115	97%	3e-33	49%	3TMY_A
Chain Y, Chemotaxis Kinase Chea P2 Domain In Complex With Response Regulator Chey From The Thermophile Thermotoga Maritima [Thermotoga mari	114	114	96%	1e-32	49%	1U0S_Y
Chain A, Crystal Structure Of The Chex-Chey-Bef3-Mg+2 Complex From Borrelia Burgdorferi [Borrelia burgdorferi]	81.6	81.6	95%	7e-20	40%	3HZH_A
Chain A, Computational Design Of An Eight-Stranded (BetaALPHA)-Barrel From Fragments Of Different Folds [Thermotoga maritima]	82.4	82.4	75%	1e-19	45%	2LLE_A
Chain A, Crystal Structure Of Rv1626 From Mycobacterium Tuberculosis [Mycobacterium tuberculosis H37Rv]	80.9	80.9	93%	3e-19	42%	1S8N_A
Chain X, A BetaALPHA-Barrel Built By The Combination Of Fragments From Different Folds [Thermotoga maritima]	79.7	79.7	75%	9e-19	45%	3CWO_X
Chain A, Crystal Structure Of Two-component Response Regulator, Luxr Family, From Aurantimonas Sp. Si85-9a1 [Aurantimonas manganoxydans SI85-9A	78.2	78.2	97%	2e-18	34%	3CZ5_A
Chain A, Domain-Swapping In The Sporulation Response Regulator Spo0a [Geobacillus stearothermophilus]	72.4	72.4	97%	1e-16	34%	1DZ3_A
Chain C, Structure Of A Histidine Kinase-response Regulator Complex Reveals Insights Into Two-component Signaling And A Novel Cis- Autophosphorylat	71.6	71.6	99%	2e-16	36%	3DGE_C

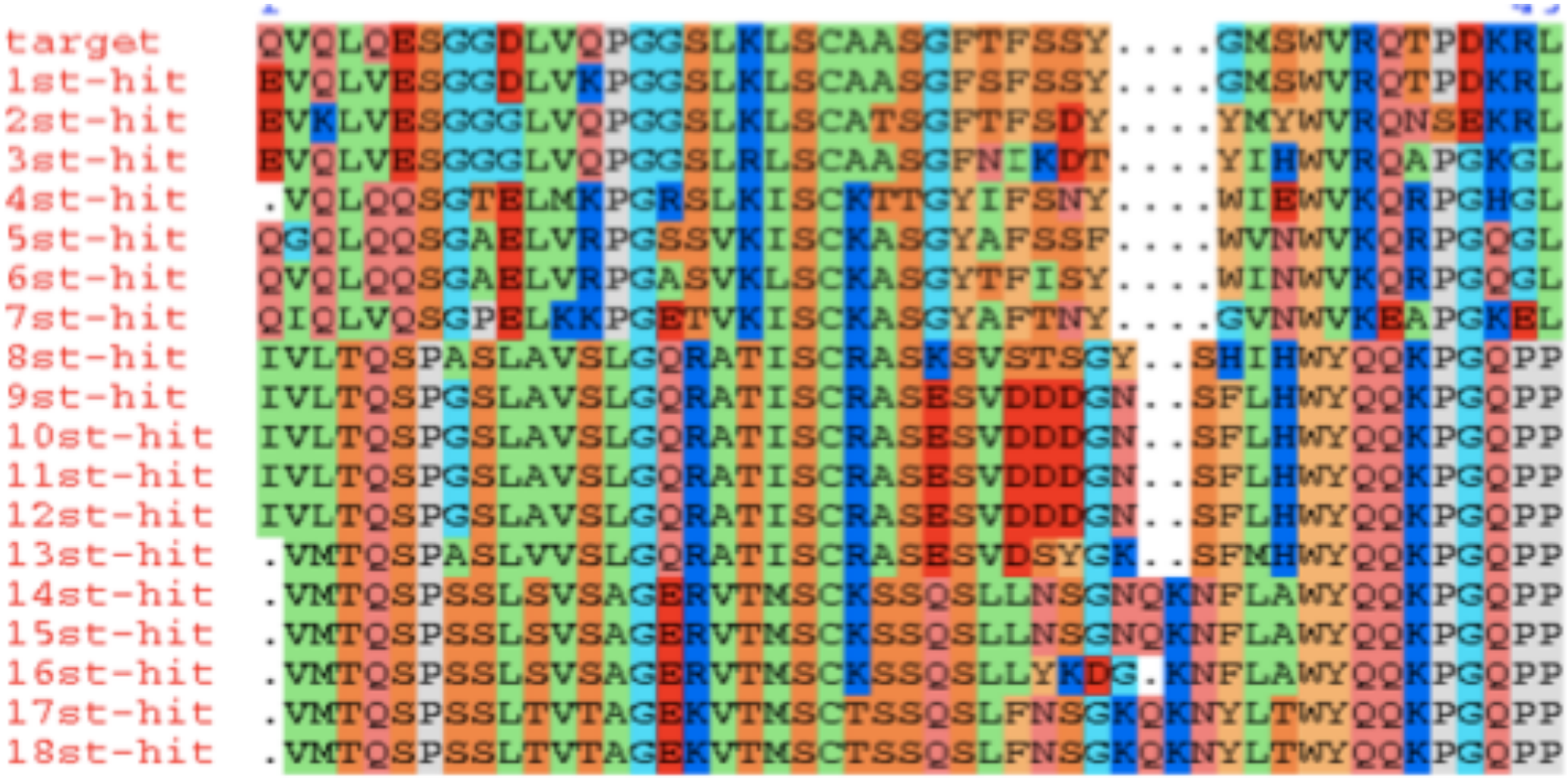
Comparative Modelling



Step 2

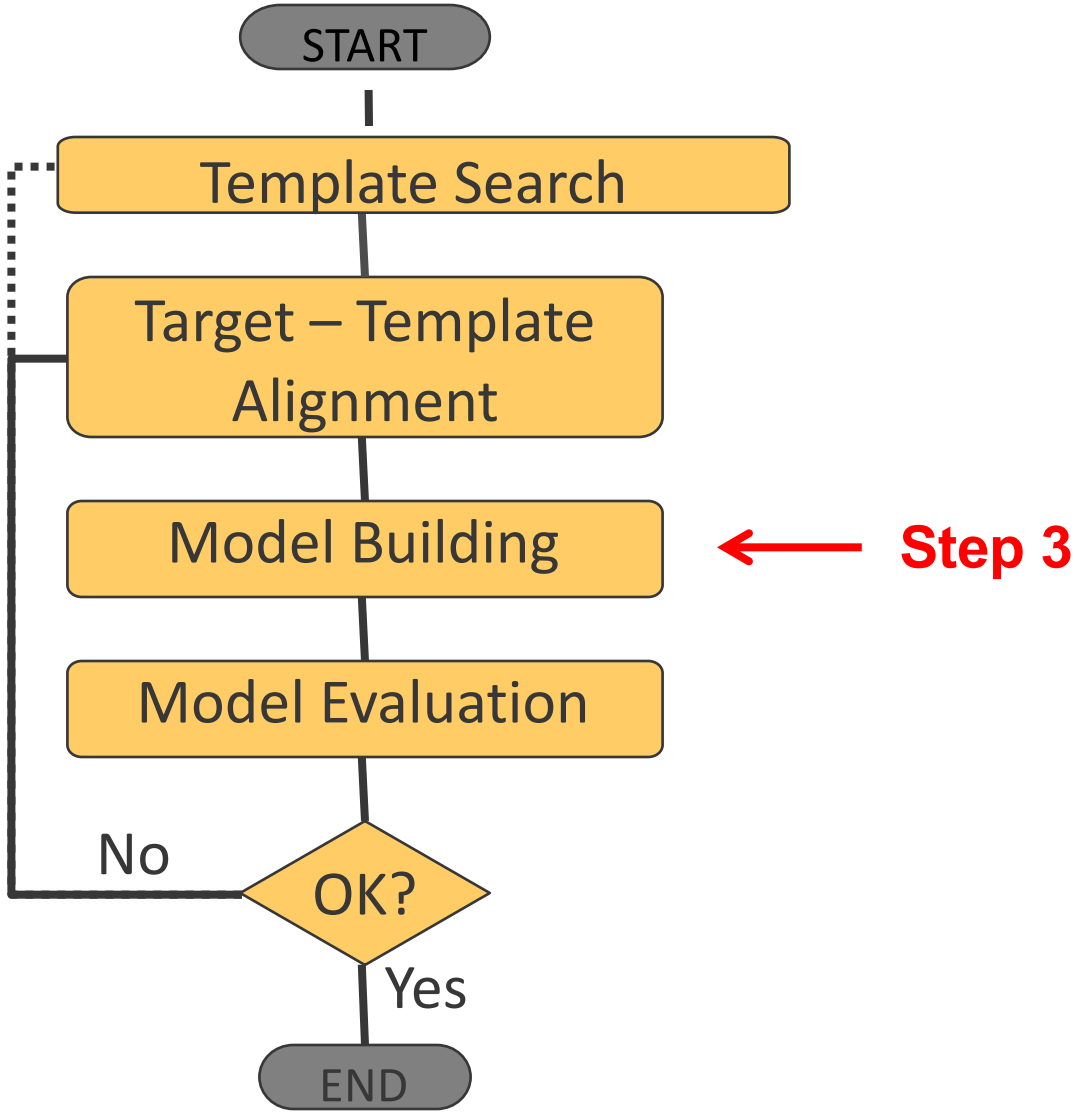
Comparative Modelling: Template Alignment

Multiple Sequence Alignment



A widely used software for sequence alignments is
T-Coffee http://tcoffee.vital-it.ch/apps/tcoffee/do:regular

Comparative Modelling



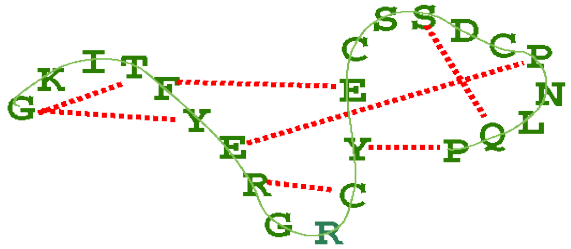
Step 3

Comparative Modelling: Model Building

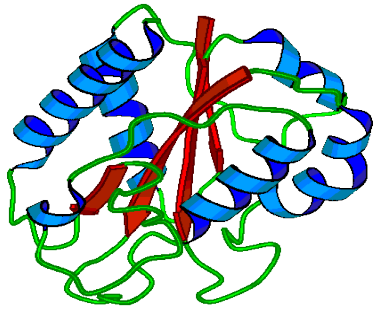
Comparative Modeling by Satisfaction of Spatial Restraints
MODELLER MODBASE Swissmodel

3D GKITFYERGFQGHCYESDC-NLQP...
SE GKITFYERG---RCYESDCPNLQP...

1. Extract spatial restraints



2. Satisfy spatial restraints



http://www.salilab.org/modeller

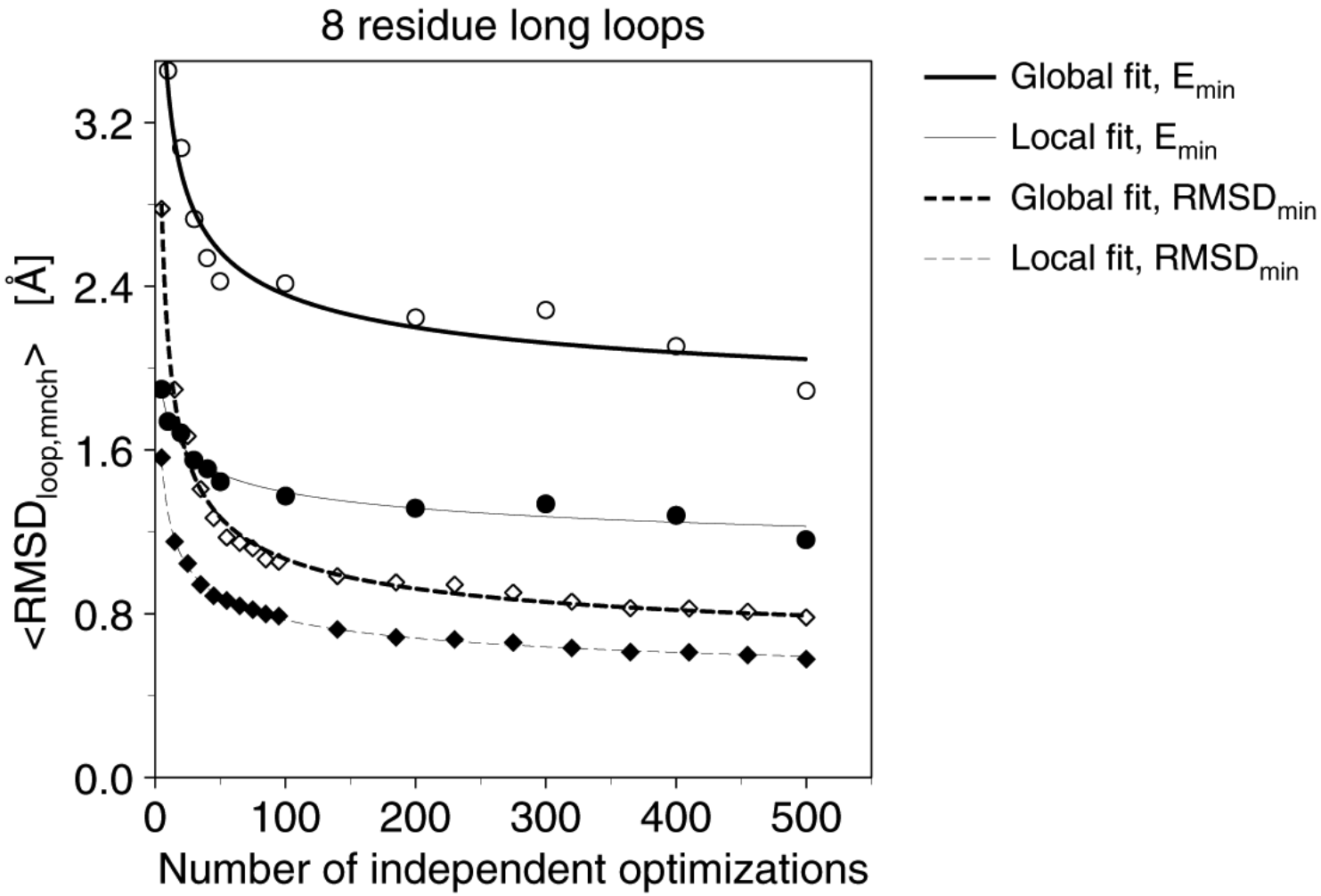
(Adapted from slides in https://salilab.org/modeller/london.zip)

A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.

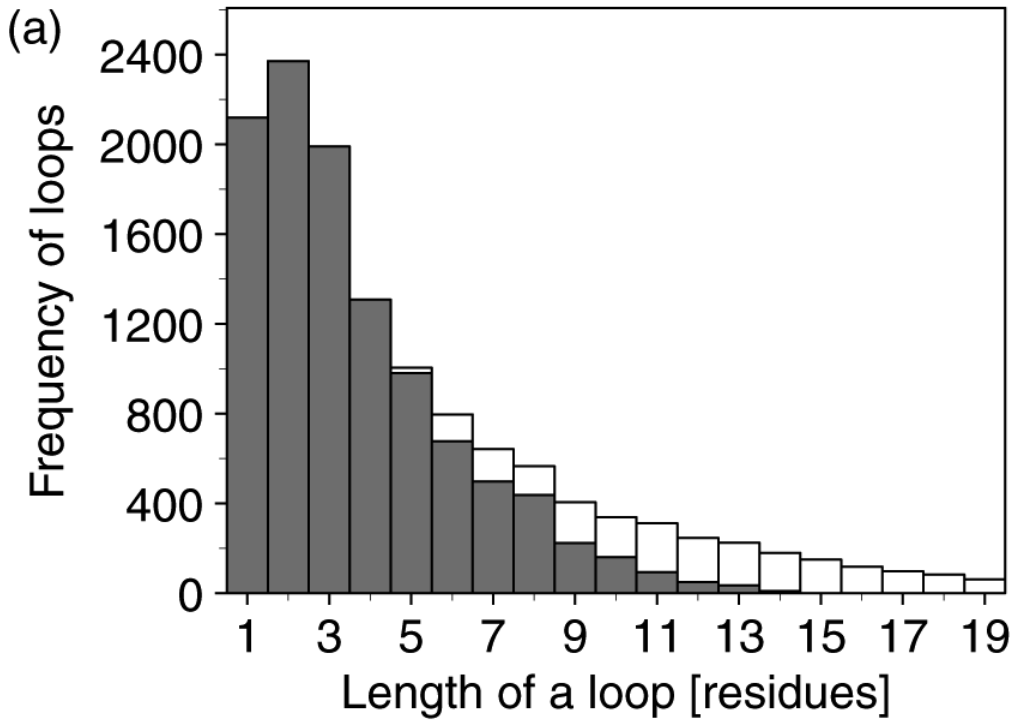
Loop modeling

- Often, there are parts of the sequence which have no detectable templates (usually loops)
- “Mini folding problem” – these loops must be sampled to get improved conformations
- Database searches only complete for 4-6 residue loops
- Modeller uses conformational search with a custom energy function optimized for loop modeling (statistical potential derived from PDB)
  - Fiser/Melo protocol (‘loopmodel’)
  - Newer DOPE + GB/SA protocol (‘dope_loopmodel’)

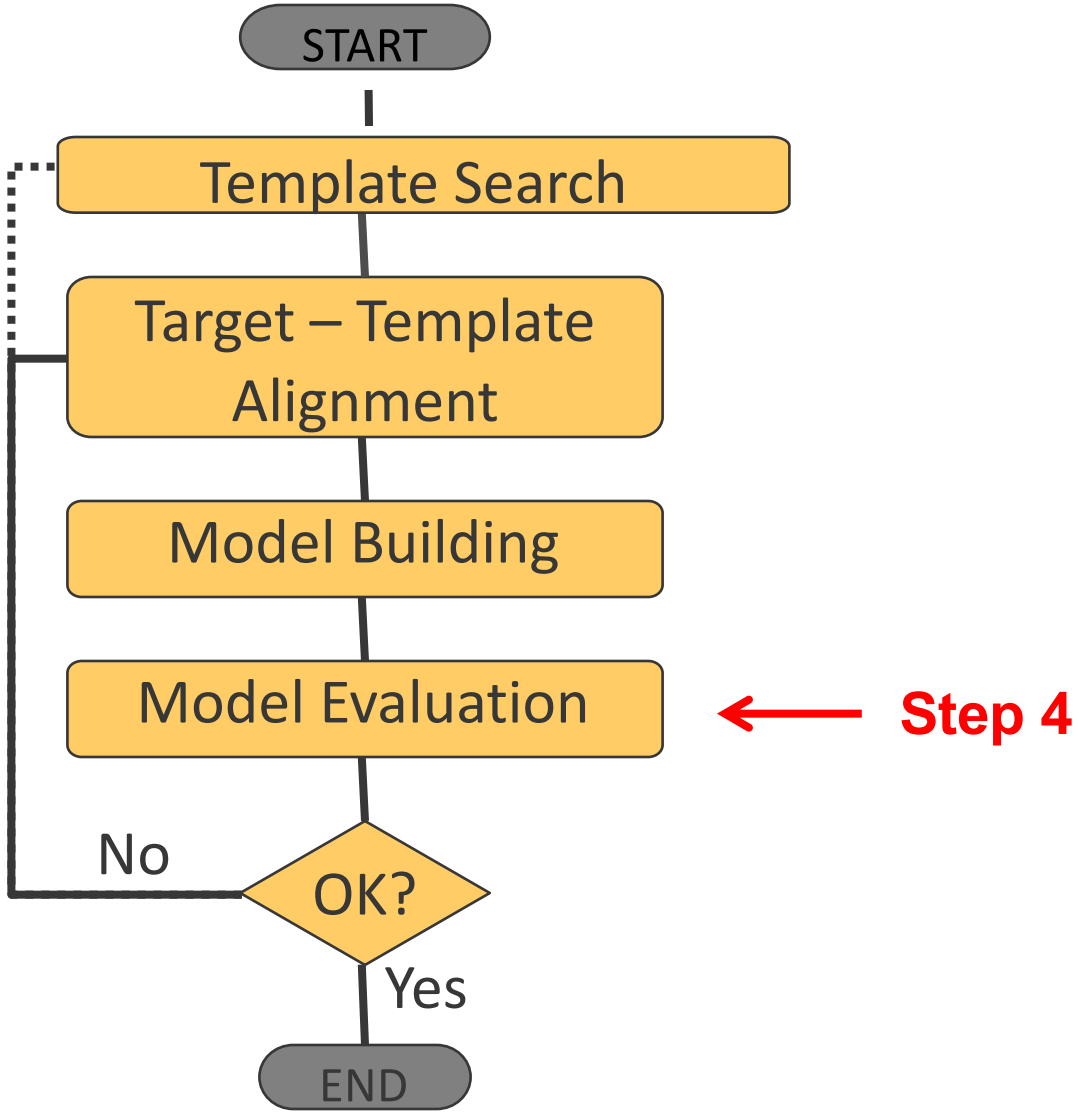
Accuracy of loop models as a function of amount of optimization



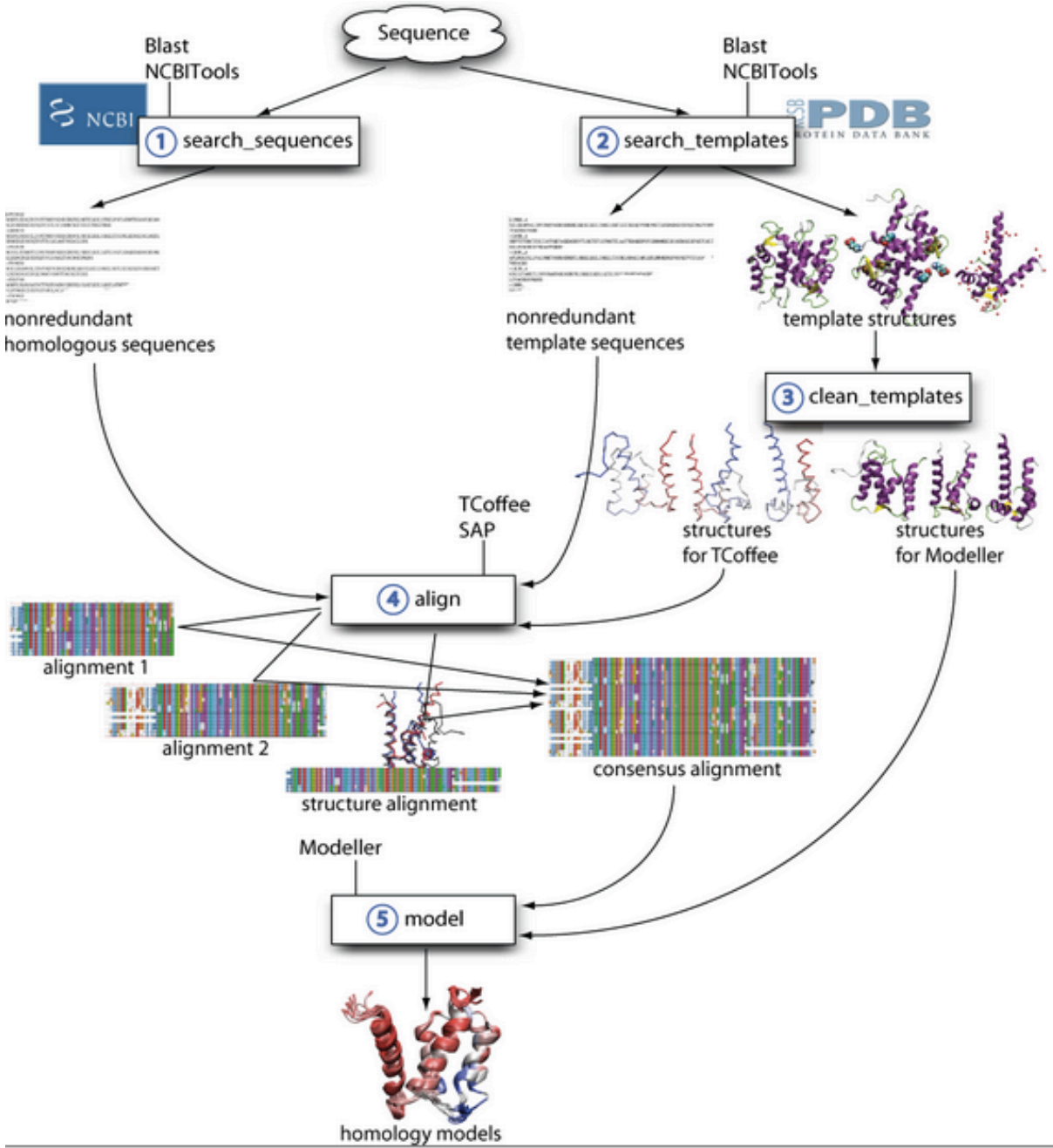
Fraction of loops modeled with medium accuracy (<2Å)



Comparative Modelling



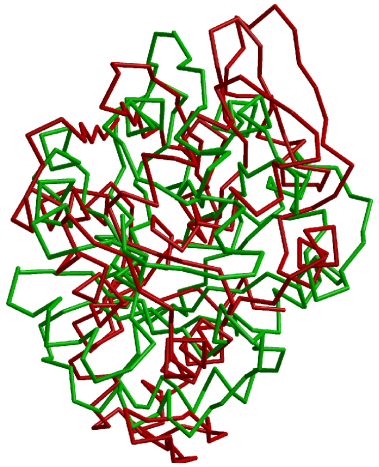
Comparative Modelling



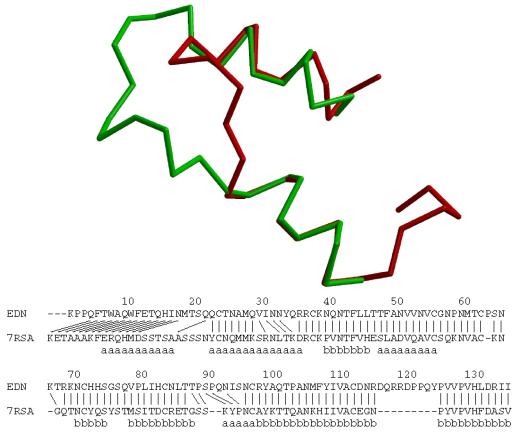
Typical errors in comparative models

MODEL
X-RAY
TEMPLATE

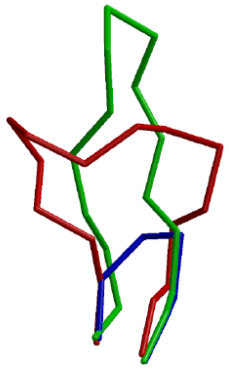
Incorrect template



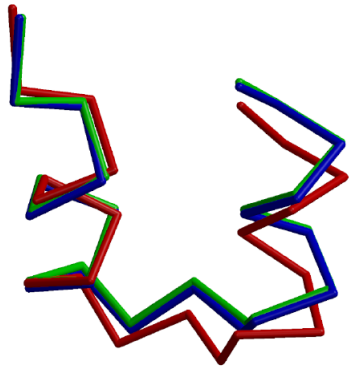
Misalignment



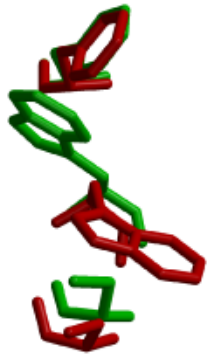
Region without a template



Distortion/shifts in aligned regions

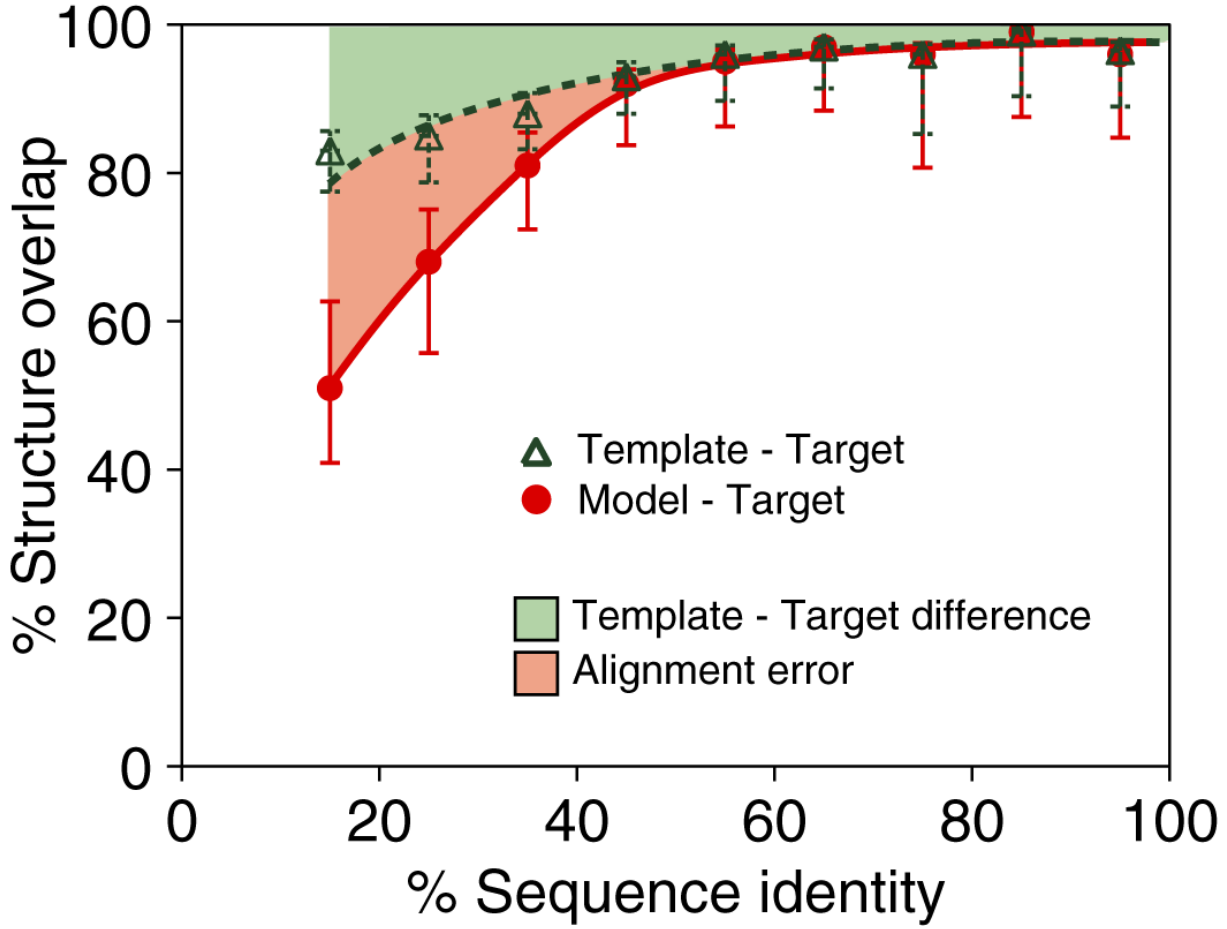


Sidechain packing



Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Model Accuracy as a Function of Target-Template Sequence Identity

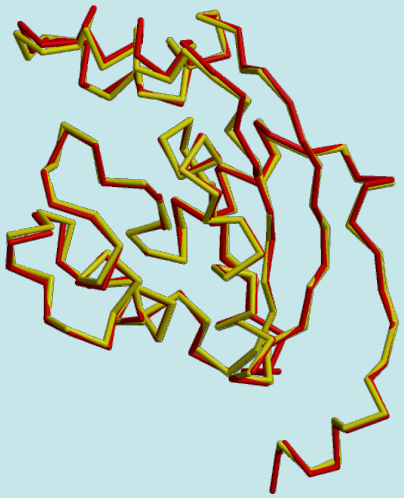


Sánchez, R., Šali, A. Proc Natl Acad Sci U S A. 95 pp13597-602. (1998).

Model accuracy

HIGH ACCURACY

NM23
Seq id 77%
Cα equiv 147/148
RMSD 0.41Å

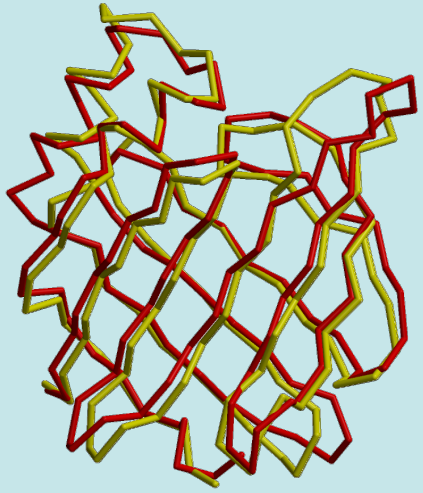


Scope for improvement:
Sidechains

X-RAY / MODEL

MEDIUM ACCURACY

CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops

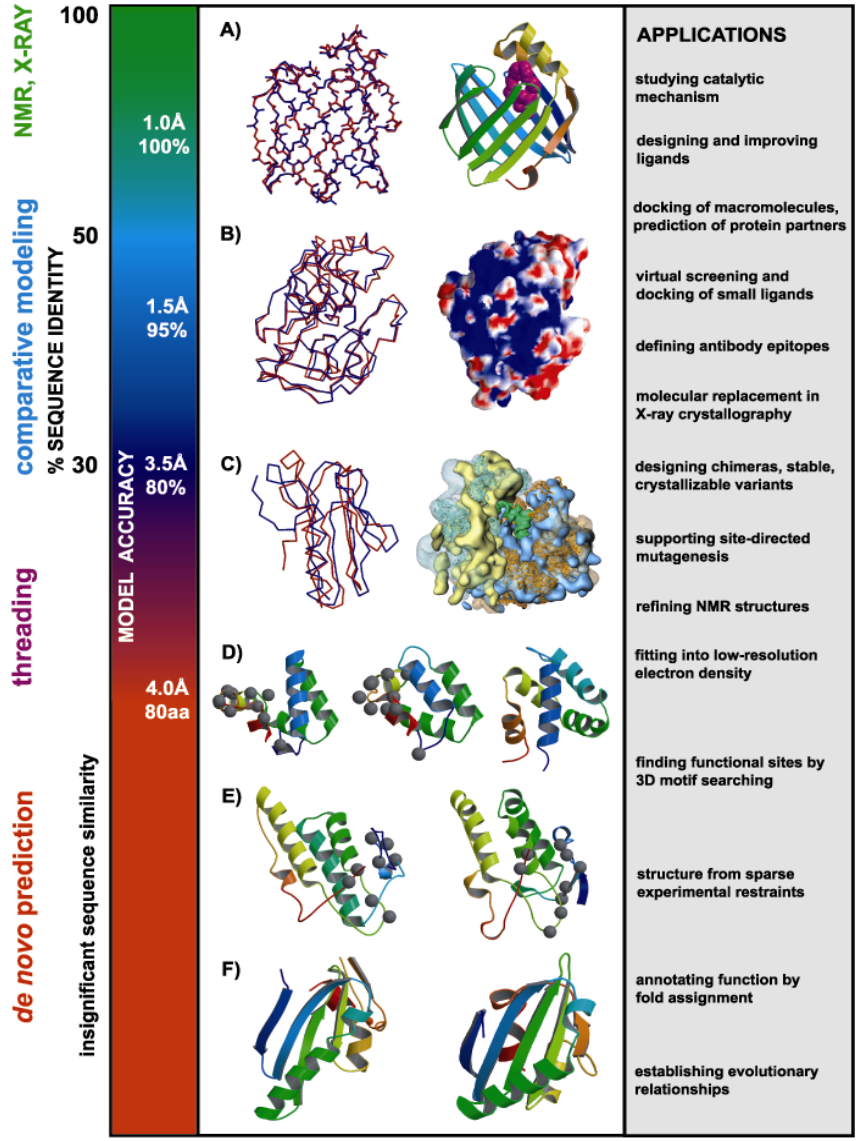
LOW ACCURACY

EDN
Seq id 33%
Cα equiv 90/134
RMSD 1.17Å

Sidechains
Core backbone, Loops
Alignment,
Fold assignment

Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Applications of protein structure models

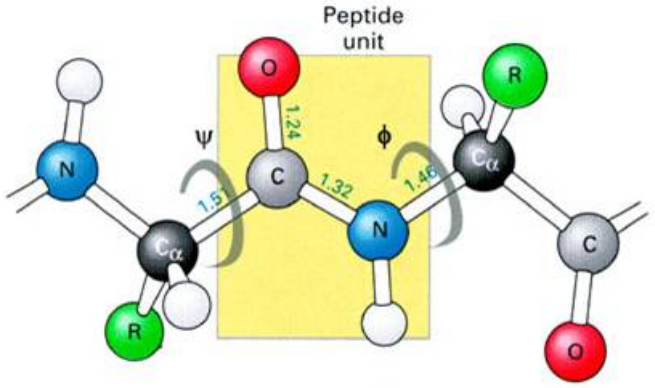


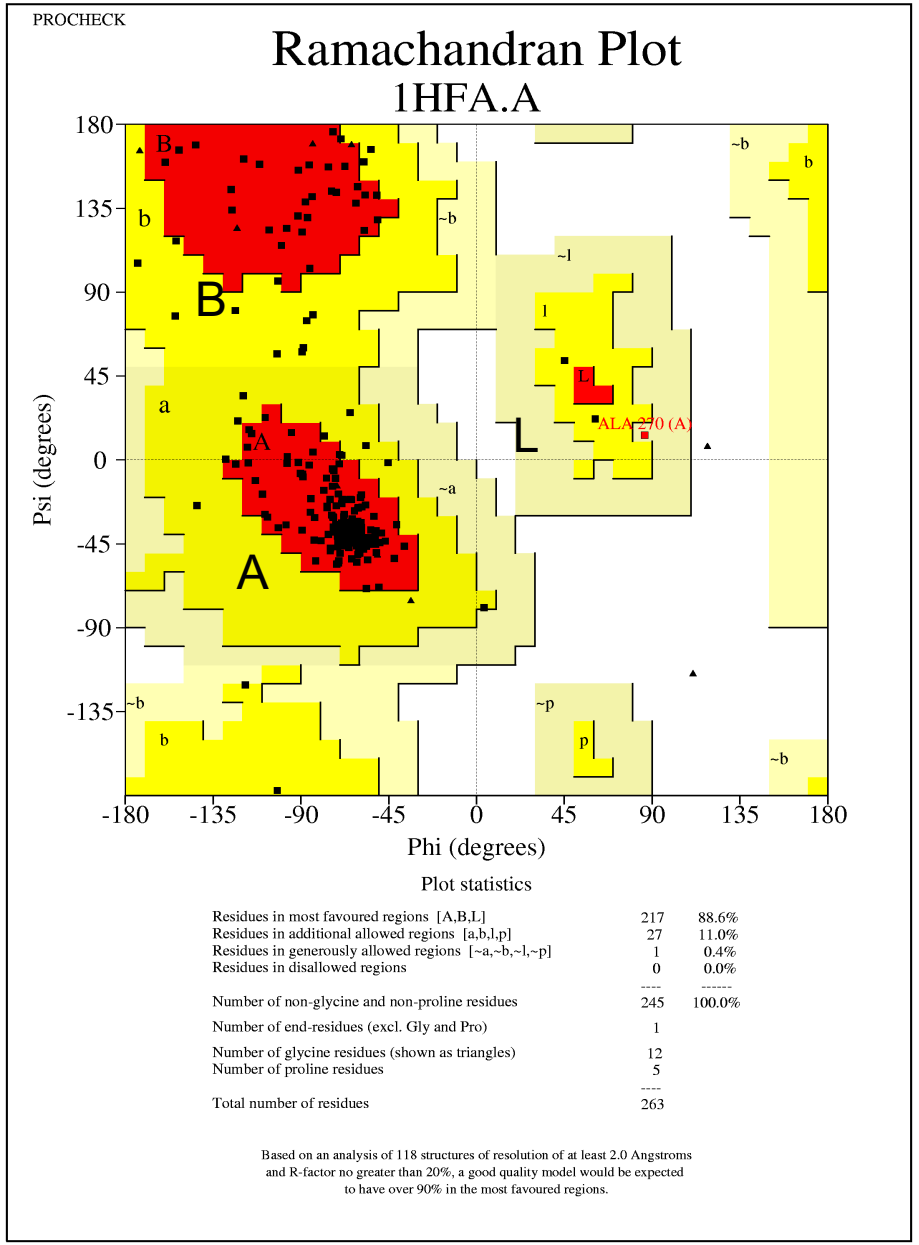
D. Baker & A. Sali.
Science 294, 93, 2001.

Comparative Modelling: Model Evaluation

Ramachandran plot

- B. Beta strand
- A. Right handed helix
- L. Left handed helix
- Color coding
  - White. Disallowed
  - Red. Most favorable
  - Yellow. Allowed region
- Glycine t

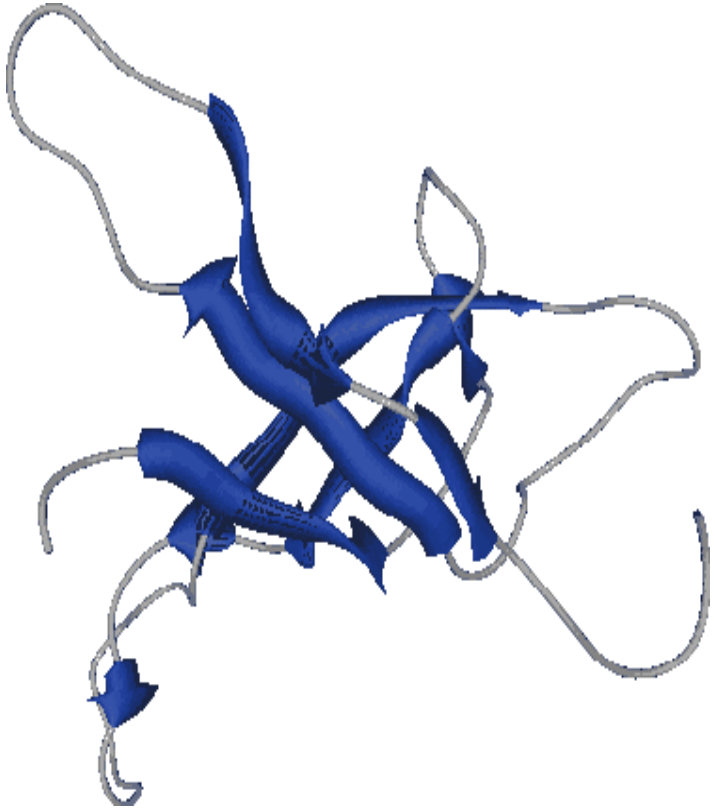




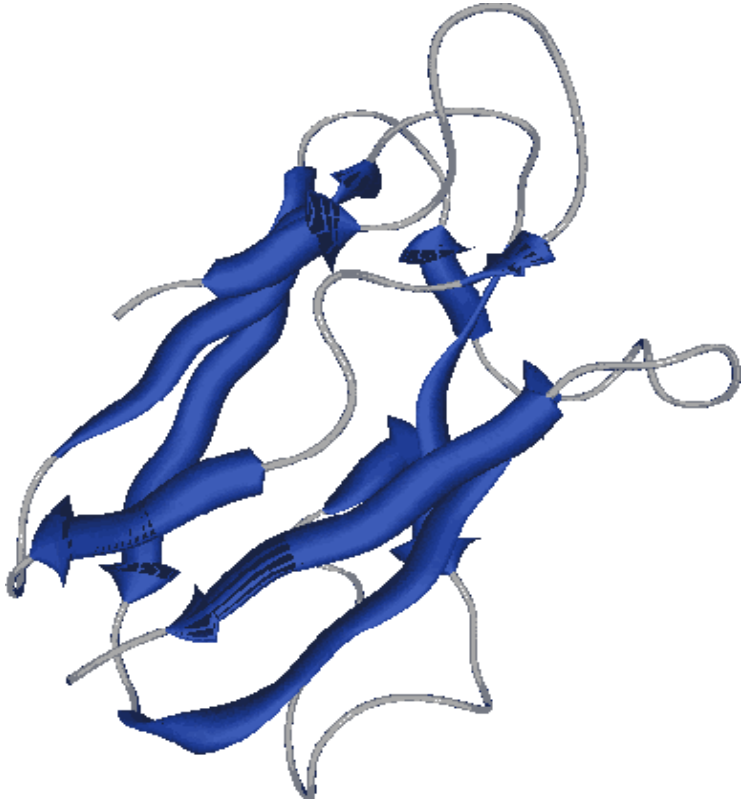
Comparative Modelling: Model Evaluation

Let's find the wrong structure!

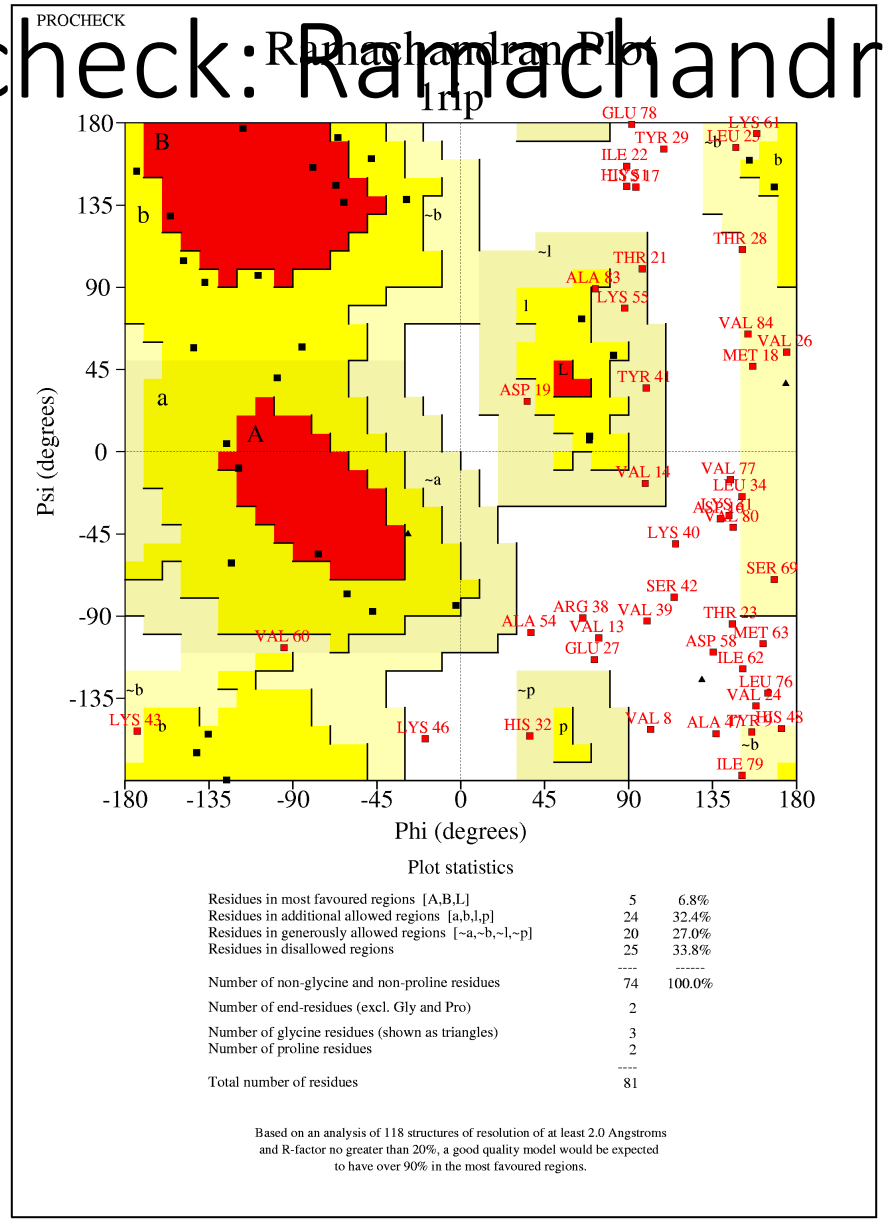
- 1RIP Ribosomal protein

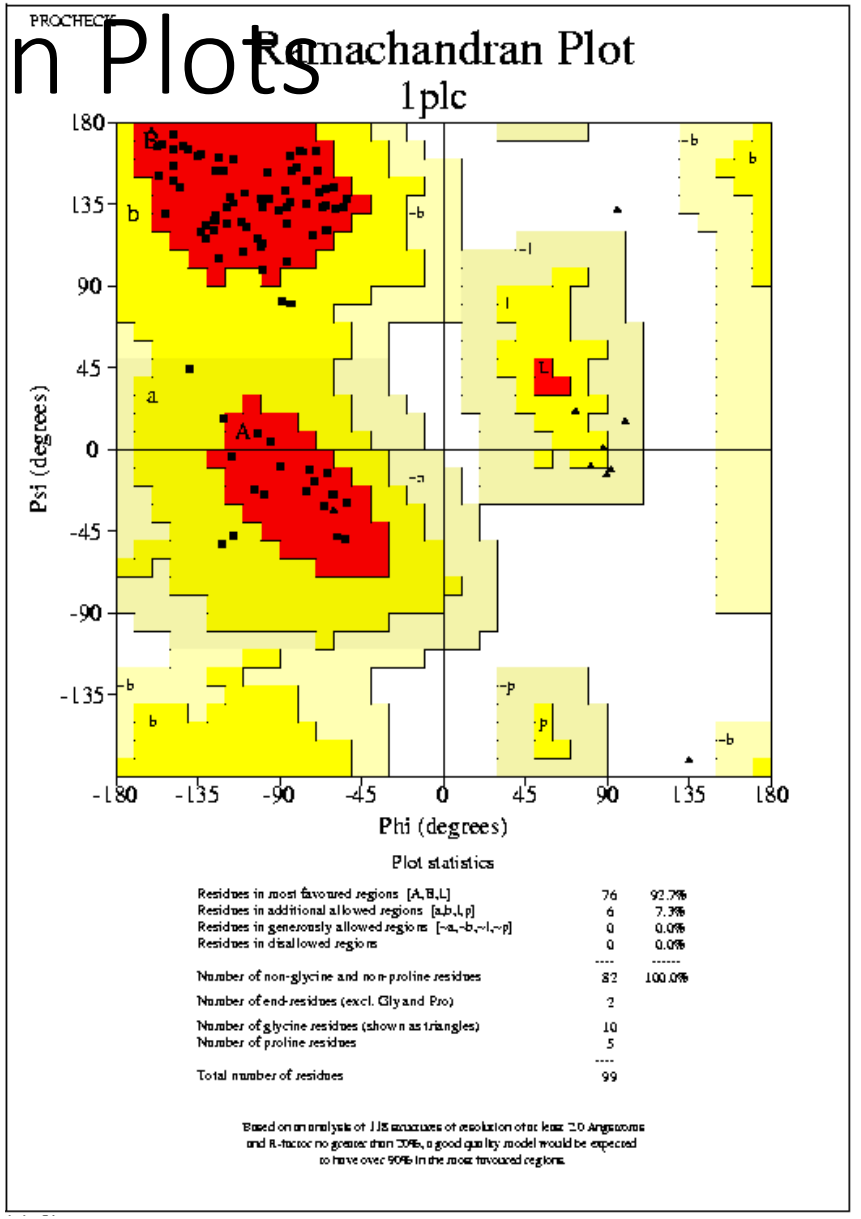


1PLC Electron transport protein

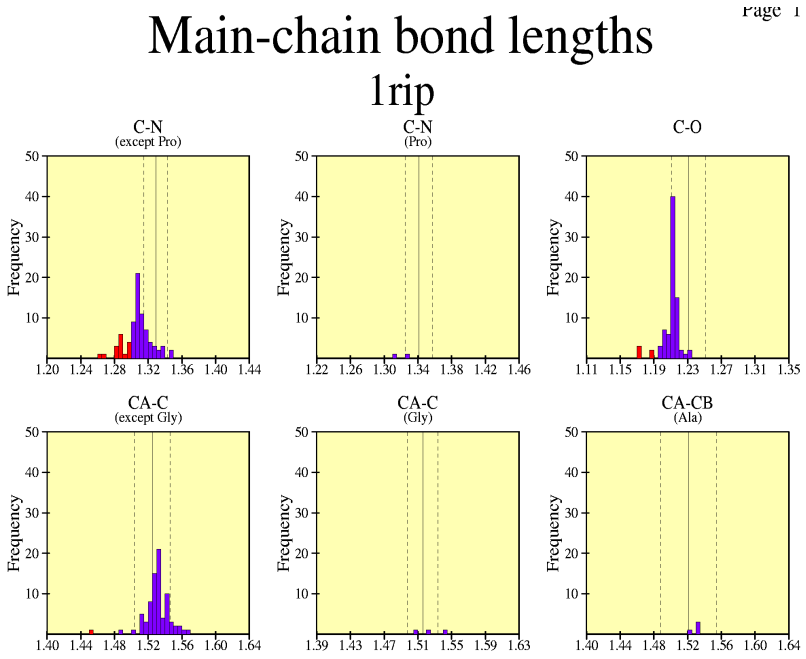


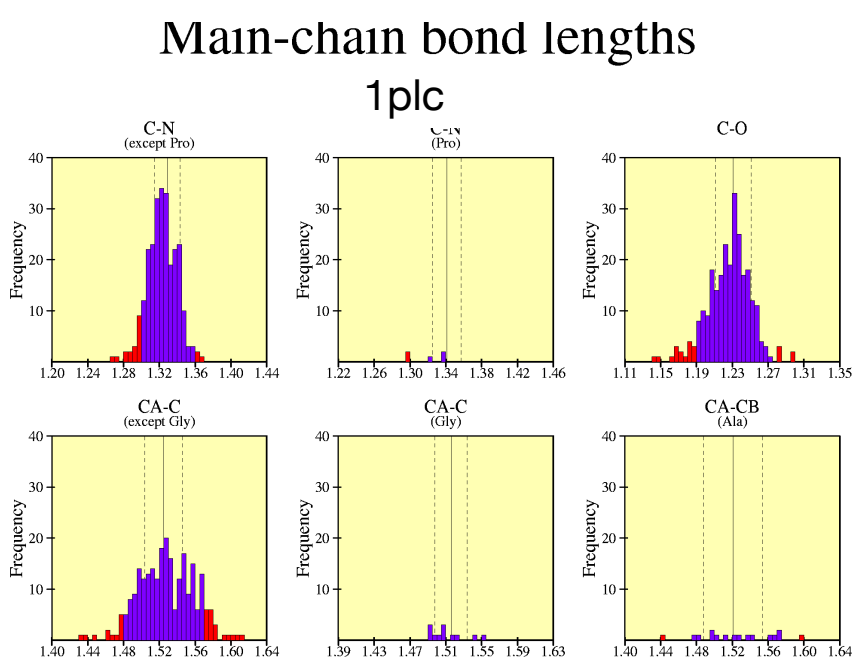
Procheck: Ramachandran Plots





Procheck: Bond Lenghts

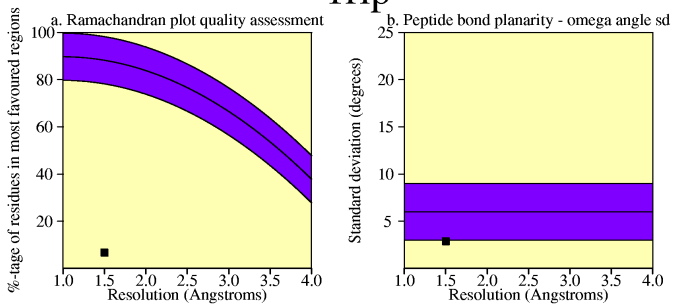


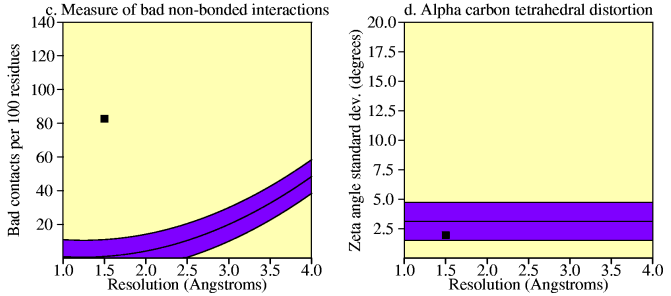


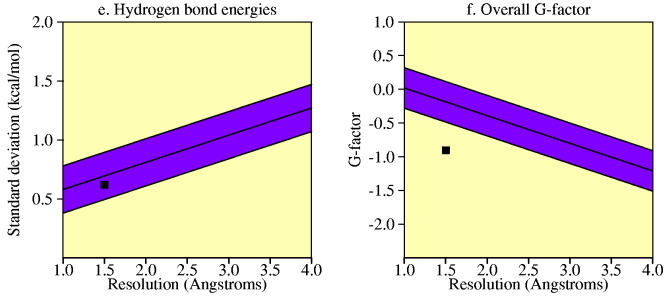
PROCHECK

Main-chain parameters

1rip







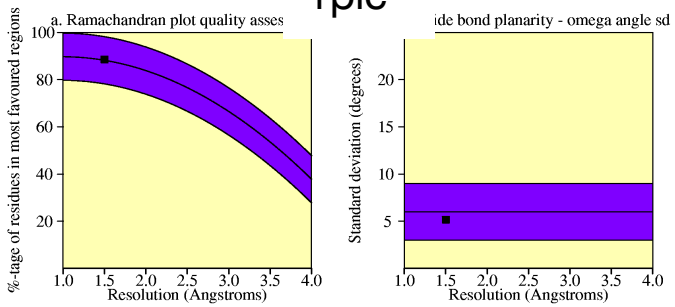
Plot statistics

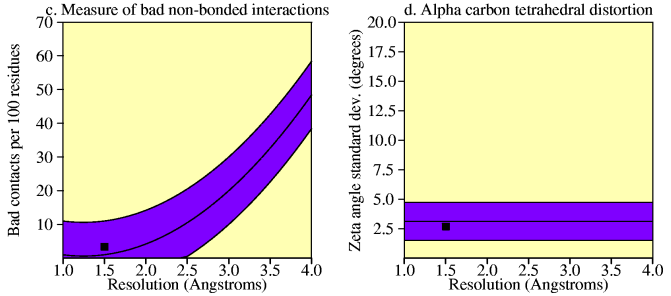
Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	74	6.8	88.2	10.0	-8.1	WORSE
b. Omega angle st dev	80	2.9	6.0	3.0	-1.0	BETTER
c. Bad contacts / 100 residues	67	82.7	1.0	10.0	8.2	WORSE
d. Zeta angle st dev	78	2.0	3.1	1.6	-0.7	Inside
e. H-bond energy st dev	24	0.6	0.7	0.2	-0.4	Inside
f. Overall G-factor	81	-0.9	-0.2	0.3	-2.4	WORSE

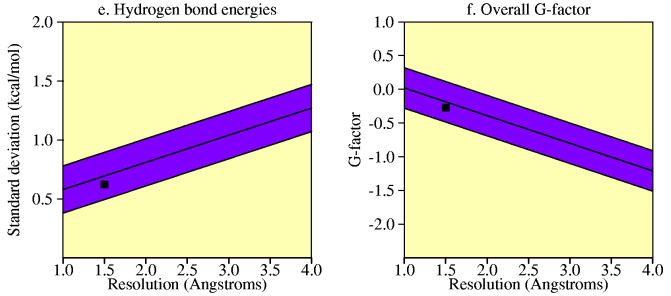
PROCHECK

Main-chain parameters

1plc





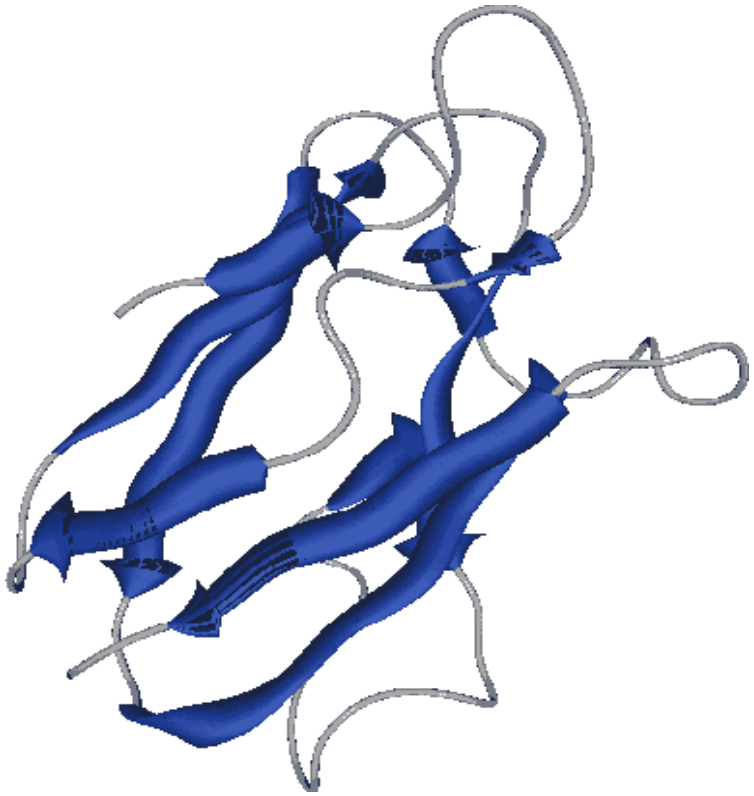


Plot statistics

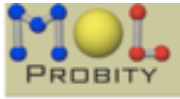
Stereochemical parameter	No. of data pts	Parameter value	Comparison values Typical value	Band width	No. of band widths from mean	
a. %-tage residues in A, B, L	245	88.6	88.2	10.0	0.0	Inside
b. Omega angle st dev	262	5.2	6.0	3.0	-0.3	Inside
c. Bad contacts / 100 residues	9	3.4	1.0	10.0	0.2	Inside
d. Zeta angle st dev	251	2.7	3.1	1.6	-0.3	Inside
e. H-bond energy st dev	181	0.6	0.7	0.2	-0.4	Inside
f. Overall G-factor	263	-0.3	-0.2	0.3	-0.3	Inside

Comparative Modelling: Model Evaluation

Let's find the wrong structure!



Comparative Modelling: Model Evaluation



About

Citation

Please cite the following paper when referring to MolProbity in print or during a presentation:

Simon C. Lovell, Ian W. Davis, W. Bryan Arendall III, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, David C. Richardson (2003) Structure validation by C-alpha geometry: phi, psi, and C-beta deviation. Proteins: Structure, Function, and Genetics. 50: 437-450.

The following paper is also relevant when using MolProbity with nucleic acids:

Ian W. Davis, Laura Weston Murray, Jane S. Richardson, David C. Richardson (2004) MolProbity: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Research. 32: W615-W619.

http://molprobity.biochem.duke.edu/

1sddb

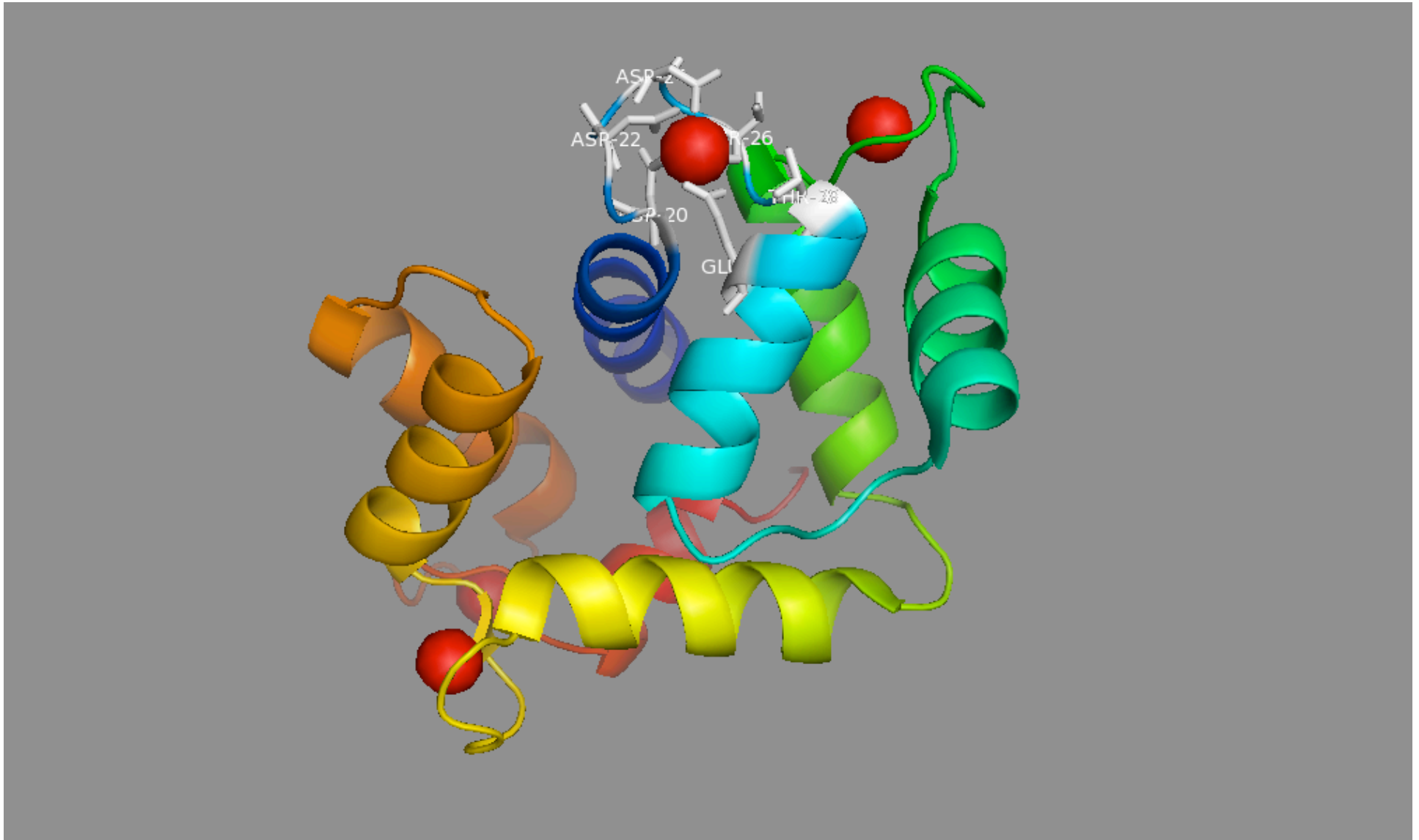
Protein Geometry	Rotamer outliers	2.90%	Goal: <1%
	Ramachandran outliers	6.00%	Goal: <0.2%
	Ramachandran favored	77.33%	Goal: >98%
	Cβ deviations >0.25Å	5	Goal: 0
	Residues with bad bonds:	0.00%	Goal: <1%
	Residues with bad angles:	2.63%	Goal: <0.5%

1czta

Protein Geometry	Rotamer outliers	1.46%	Goal: <1%
	Ramachandran outliers	2.01%	Goal: <0.2%
	Ramachandran favored	89.26%	Goal: >98%
	Cβ deviations >0.25Å	0	Goal: 0
	Residues with bad bonds:	0.66%	Goal: <1%
	Residues with bad angles:	1.99%	Goal: <0.5%

Calmodulin

How



Secondary Structure Assignment

(given a 3D structure, assign secondary structural elements)

The DSSP program defines 7 secondary structure states

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

The secondary structure assignment with DSSP over a database of structures can be used as 'standard of truth' for secondary structure prediction methods.

Kabsch & Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.
Biopolymers, 22(12), 2577-2637 (1983). http:/swift.cmbi.ru.nl/gv/dssp/
http://www.cmbi.ru.nl/hsspsoap/

web-server

Secondary Structure Assignment of a PDB file

Output

The output from DSSP contains secondary structure assignments and other information, one line per residue. Extract from 1est.dssp (simplified):

```
HEADER    HYDROLASE   (SERINE PROTEINASE)         17-MAY-76   1EST
...
  240  1   4   4   0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS,
                     NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)                .
 10891.0    ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
  162 67.5    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  ; PER 100 RESIDUES
    0  0.0    TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES; PER 100 RESIDUES
   84 35.0    TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES; PER 100 RESIDUES
...
   26 10.8    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2)
   30 12.5    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3)
   10  4.2    TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4)
...
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC    N-H-->O  O-->H-N  N-H-->O  O-->H-N
    2   17   V  B 3   +A  182   0A   8  180,-2.5 180,-1.9   1,-0.2 134,-0.1
                                     ...Next two lines wrapped as a pair...
                                      TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
                                    -0.776 360.0   8.1 -84.5 125.5  -14.7   34.4   34.8
....;....1....;....2....;....3....;....4....;....5....;....6....;....7..
    .-- sequential resnumber, including chain breaks as extra residues
    |     .-- original PDB resname, not nec. sequential, may contain letters
    |     |    .-- amino acid sequence in one letter code
    |     |    |  .-- secondary structure summary based on columns 19-38
    |     |    |  | xxxxxxxxxxxxxxxxxx recommend columns for secstruc details
    |     |    |  .-- 3-turns/helix
    |     |    |  |.-- 4-turns/helix
    |     |    |  ||.-- 5-turns/helix
    |     |    |  |||.-- geometrical bend
    |     |    |  ||||.-- chirality
    |     |    |  |||||.-- beta bridge label
    |     |    |  ||||||.-- beta bridge label
    |     |    |  |||||||    .-- beta bridge partner resnum
    |     |    |  |||||||    |    .-- beta bridge partner resnum
    |     |    |  |||||||    |    |.-- beta sheet label
    |     |    |  |||||||    |    ||    .-- solvent accessibility
    |     |    |  |||||||    |    ||    |
  #  RESIDUE AA STRUCTURE BP1 BP2   ACC
    |     |    |  |||||||    |    ||    |
   35    47    I  E     +     0    0    2
   36    48    R  E >  S- K   0   39C  97
   37    49    Q  T 3  S+     0    0   86
   38    50    N  T 3  S+     0    0   34
   39    51    W  E <   -KL  36   98C   6
```

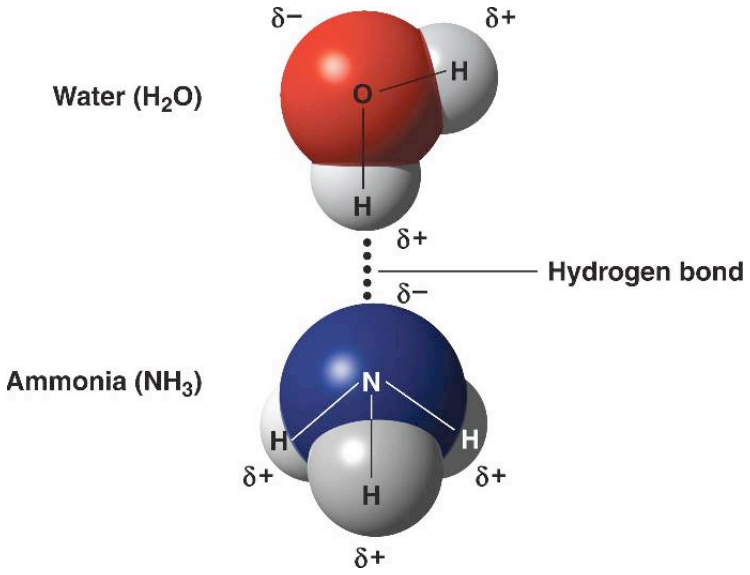
The secondary structure assignment with DSSP over a database of structures can be used as 'standard of truth' for secondary structure prediction methods.

DSSP uses mainly hydrogen bond assessments to discriminate secondary structure elements

What is a Hydrogen Bond?

A hydrogen bond is formed when a proton (H) covalently attached to one electronegative donor atom (D) is shared with another electronegative acceptor atom (A).

One of the widely used schemes was proposed by Morokuma (1977) in which ab-initio calculations describe the interaction energy of a hydrogen bond in terms of electrostatic, charge transfer, polarization, exchange repulsion and coupling.



DSSP: H-bond energy cutoff

Hydrogen bonds in proteins have little wave-function overlap and are well described by an electrostatic model.[13] We calculate the electrostatic interaction energy between two H-bonding groups by placing partial charges on the C,O $(+q_1, -q_1)$ and N,H $(-q_2, +q_2)$ atoms, i.e.,

$$E = q_1 q_2 (1/r(\mathrm{ON}) + 1/r(\mathrm{CH}) - 1/r(\mathrm{OH}) - 1/r(\mathrm{CN}))*f$$

with $q_1 = 0.42e$ and $q_2 = 0.20e$, e being the unit electron charge and r(AB) the interatomic distance from A to B. In chemical units, r is in angstroms, the dimensional factor $f = 332$, and E is in kcal/mol. A good H bond has about −3 kcal/mol binding energy. We choose a generous cutoff to allow for bifurcated H bonds and errors in coordinates and assign an H bond between C═O of residue i and N-H of residue j if E is less than the cutoff, i.e., "Hbond(i,j)=: [E < −0.5kcal/mole]."

Kabsch W, Sander C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers (1983) 22:2577-637.

DSSP: H-bond energy cutoff

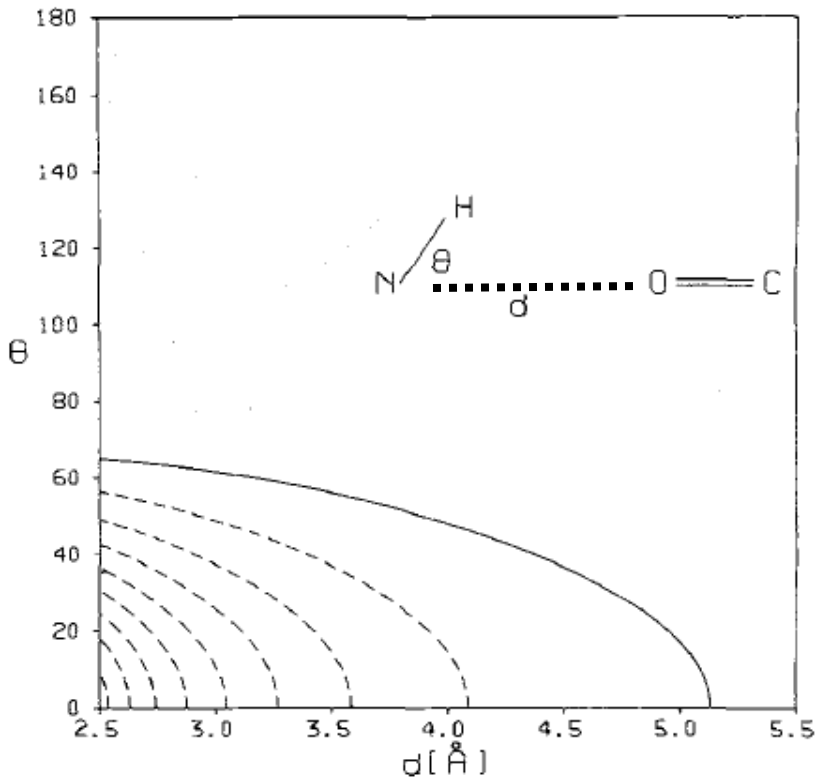


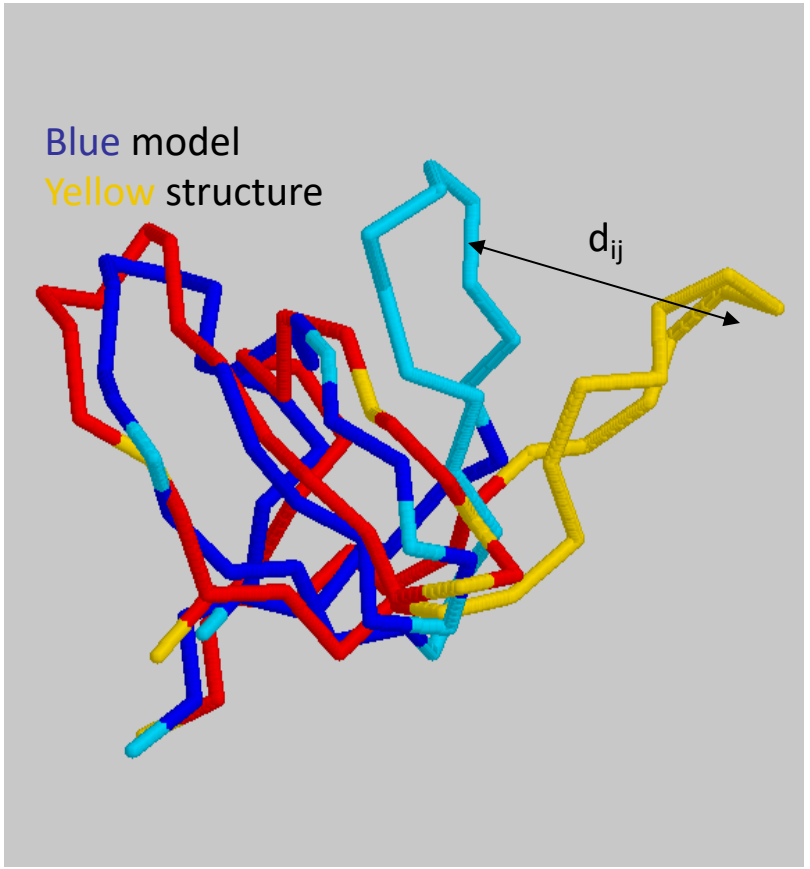
Fig. 1. H bond between peptide units is described here by the dominant electrostatic part E (see text) of the H-bond energy, drawn in contours of constant E at 0.5 kcal/mol intervals as a function of the distance, d, and the alignment angle θ. Dotted lines, E positive or zero; broken lines, E negative. An ideal H bond has $d = 2.9$ Å, $\theta = 0$, and $E = -3.0$ kcal/mol. We assume an H bond for E up to -0.5 kcal/mol (solid line). Thus, misalignment of up to 63° is allowed at the ideal length; an N-O distance of up to $d = 5.2$ Å is allowed for perfect alignment. This definition of H bonds is particularly simple and physically meaningful. It is more general than the historical definition of hydrogen "bond" and could be called polar interaction.

Model accuracy

If we know the answer

$$RMSD = \sqrt{\frac{1}{N} \bullet \sum_{i,j} d_{i,j}^2}$$

Fraction correct = N_c/N
N_c = number correct (dij<4Å)



Databases of multiple alignments ----> Domains

Very early in the days of protein sequence analysis, it was observed that some protein sequences contained long segments that were very similar to other proteins, while the rest of the sequence in that protein had no detectable similarity.

Today, we take more or less for granted that proteins are composed of domains, segments of sequence which have been joined together by genetic events during evolution so that the new protein has a function that is based on the activities of the domains it contains.

Often the domains detectable by sequence analysis correspond to structural domains in the 3D structure as well. There are now many well-documented cases where it has been shown that domains can exists perfectly well in isolation, when excised from the original protein. Surprisingly often, a domain can be expressed and folded all on its own.

There are today several databases that keep track of which domains have been discovered, which proteins are involved, and that store the multiple sequence alignments of the relevant segments of the protein sequences. We will discuss one such of databases, Pfam.

Also, several of the primary sequence databases now contain information about the domains in the sequence entries.

The idea behind Pfam is twofold:

1. Create and maintain good-quality multiple sequence alignments of well-defined protein sequence domains from proteins in SWISS-PROT.
2. Use these multiple alignments for creating so-called HMMs(Hidden Markov Models) machine learning algorithms, that can be used in profile searches of sequence databases.
https://en.wikipedia.org/wiki/Hidden_Markov_model

The multiple alignment used to define a domain (protein family) in Pfam are called the seed alignment. It is created by a curator, or taken from the literature. It is used to generate a profile HMM for identifying other sequences in the databases (SWISS-PROT and TREMBL) that contain the domain. The search results are inspected to decide which cutoff should be used for that particular Pfam entry. The search hits are then aligned automatically into a so-called full alignment.

There are today several databases that keep track of which domains have been discovered, which proteins are involved, and that store the multiple sequence alignments of the relevant segments of the protein sequences. We will discuss one such of databases, Pfam.
Also, several of the primary sequence databases now contain information about the domains in the sequence entries.

Pfam

http://pfam.sanger.ac.uk/

The alignments can be converted into hidden Markov models (HMM), which can be used to search for domains in a query protein sequence. The software HMMER (by Sean Eddy) is the computational foundation for Pfam.

http://nar.oxfordjournals.org/content/early/2015/05/04/nar.gkv397.full
http://hmmer.org/

HMMER is often used together with a profile database, such as Pfam or many of the databases that participate in Interpro. But HMMER can also work with query *sequences*, not just profiles, just like BLAST. For example, you can search a protein query sequence against a database with **phmmer**, or do an iterative search with **jackhmmer**.

The domain structure of protein sequences in SWISS-PROT and TrEMBL are available directly from the Pfam web sites, and it is also possible to search for domains in other sequences using servers at the web sites.

Prosite

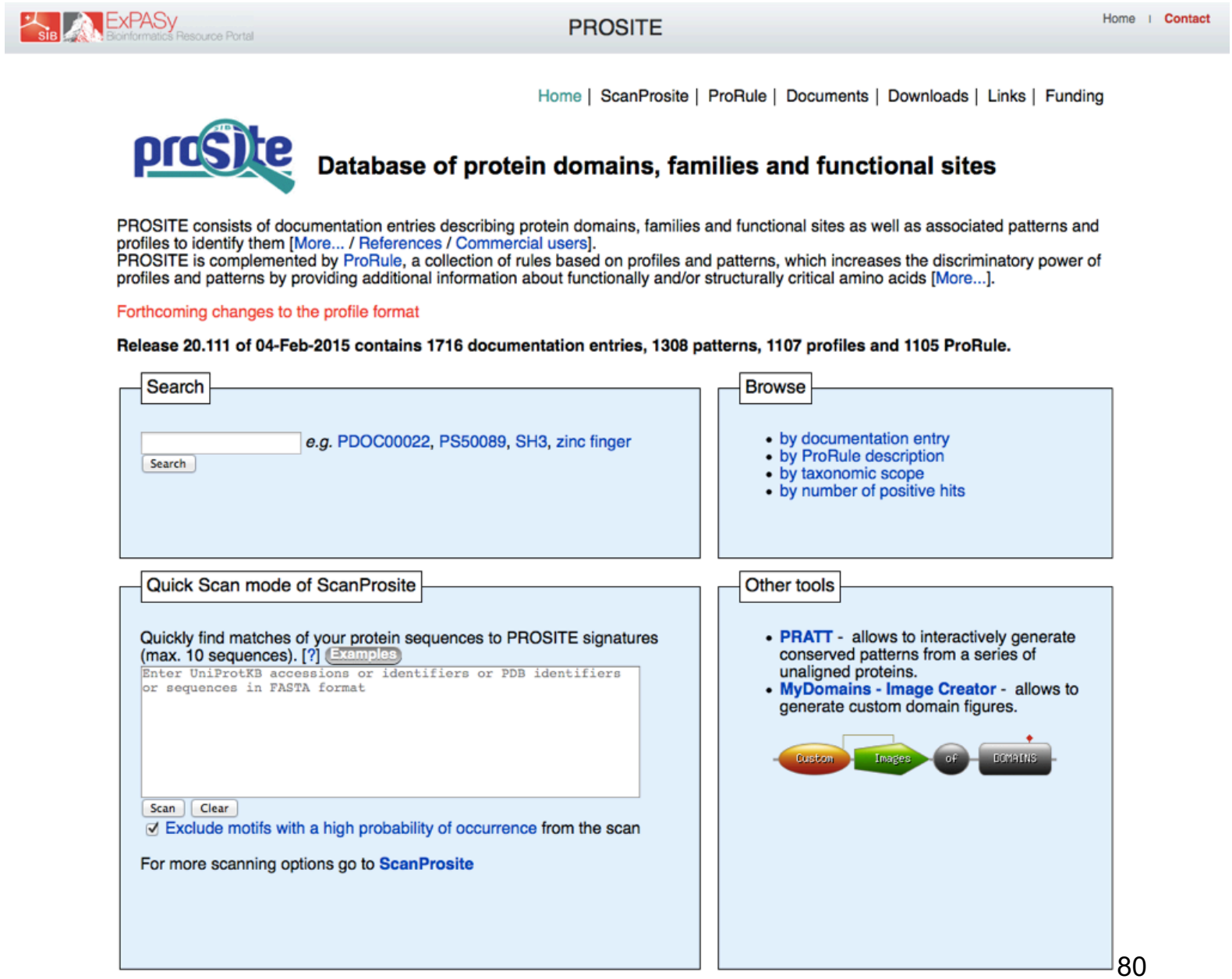
PROSITE www.expasy.ch/prosite/

PROSITE is a database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.

It was started by Amos Bairoch, is part of SWISS-PROT and is maintained in the same way as SWISS-PROT. The basis of it are regular expressions describing characteristic subsequences of specific protein families or domains. PROSITE has been extended to contain also some profiles, which can be described as probability patterns for specific protein sequence families.

The site above can be used to search by keyword or other text in the entries,
to search for a pattern in a sequence, or to search for proteins in SWISS-PROT
that match a pattern.

http://prosite.expasy.org/



Domain prediction by Smart

Schultz et al. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 5857-5864
Letunic et al. (2012) *Nucleic Acids Res* , doi:10.1093/nar/gkr931

SMART MODE:
NORMAL
GENOMIC

Simple
Modular
Architecture
Research
Tool

keywords...
Search SMART

HOME SETUP FAQ ABOUT GLOSSARY WHAT'S NEW FEEDBACK

Select your default SMART mode

You can use SMART in two different modes: **normal** or **genomic**.The main difference is in the underlying protein database used. In **Normal SMART**, the database contains Swiss-Prot, SP-TrEMBL and stable Ensembl proteomes. In **Genomic SMART** only the proteomes of completely sequenced genomes are used; Ensembl for metazoans and Swiss-Prot for the rest. The complete list of genomes in Genomic SMART is available here.

The protein database in Normal SMART has significant redundancy, even though identical proteins are removed. If you use SMART to explore domain architectures, or want to find exact domain counts in various genomes, consider switching to **Genomic** mode. The numbers in the domain annotation pages will be more accurate, and there will not be many protein fragments corresponding to the same gene in the architecture query results. Remember you are exploring a limited set of genomes, though.

Different color schemes are used to easily identify the mode you're in.

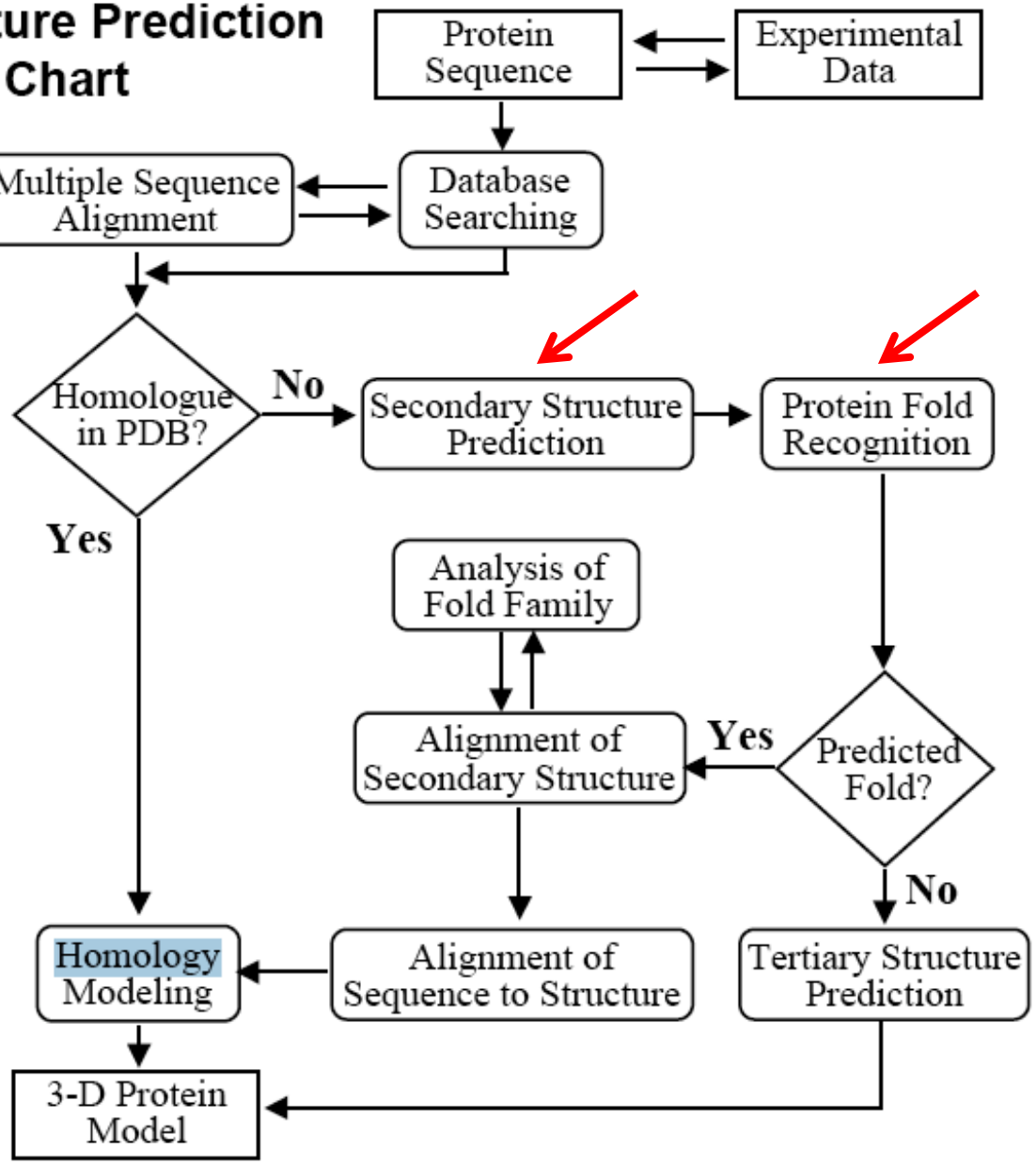
Normal mode	Genomic mode
SMART MODE: NORMAL GENOMIC Simple Modular Architecture Research Tool	SMART MODE: NORMAL GENOMIC Simple Modular Architecture Research Tool

Click on the images above to select your default mode.

Information about your selected mode is stored in a browser cookie. If you for whatever reason don't want/can't use cookies, access SMART through this page.

You can easily change modes later, by clicking on the links in the 'SMART MODE' header box, or in your personal preference settings ('SETUP' link in the menu):

Protein Structure Prediction Flow Chart



http://www.cs.wright.edu/~mraymer/cs790/Homology_Modeling.ppt

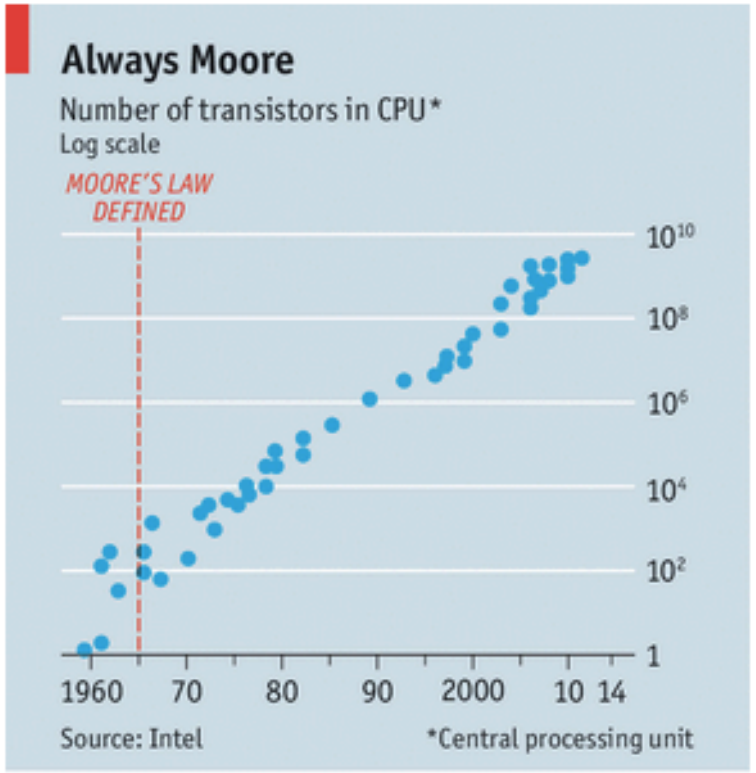
Moore's law turns 50

Ever more from Moore

A microchip pioneer's prediction has a bit more life left in it

Apr 18th 2015 | From the print edition

NEWS of the death of Moore's law has always been greatly exaggerated. People started to pronounce it deceased not long after Gordon Moore, co-founder of Intel, a chipmaker, published on April 19th 1965 a paper arguing that the number of transistors that can be etched on a given surface area of silicon would double every year. In a later paper he corrected his forecast to every two years, which has come to be stated as his "law". Regularly proving sceptics wrong, however, the exponential growth kept going (see chart), driving the digital revolution.



Economist.com

ulations of
n based on
years over

e simula-

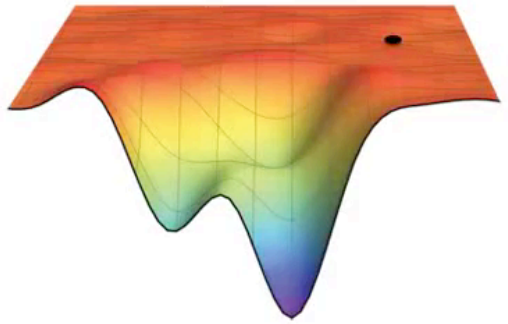
064-4092

Molecular Mechanics and Force Fields

Energy landscapes of proteins

3D Structure

Energy landscape



http://www.youtube.com/watch?v=YANAso8Jxrk

Molecular Mechanics Force Fields provide a 'recipe' (equations and parameters) to calculate the potential energy of a protein from its atomic coordinates

Molecular Mechanics and Force Fields

MM Force Fields

Levitt, M.et al *Computer Physics Communications* **91,** 215(1995)

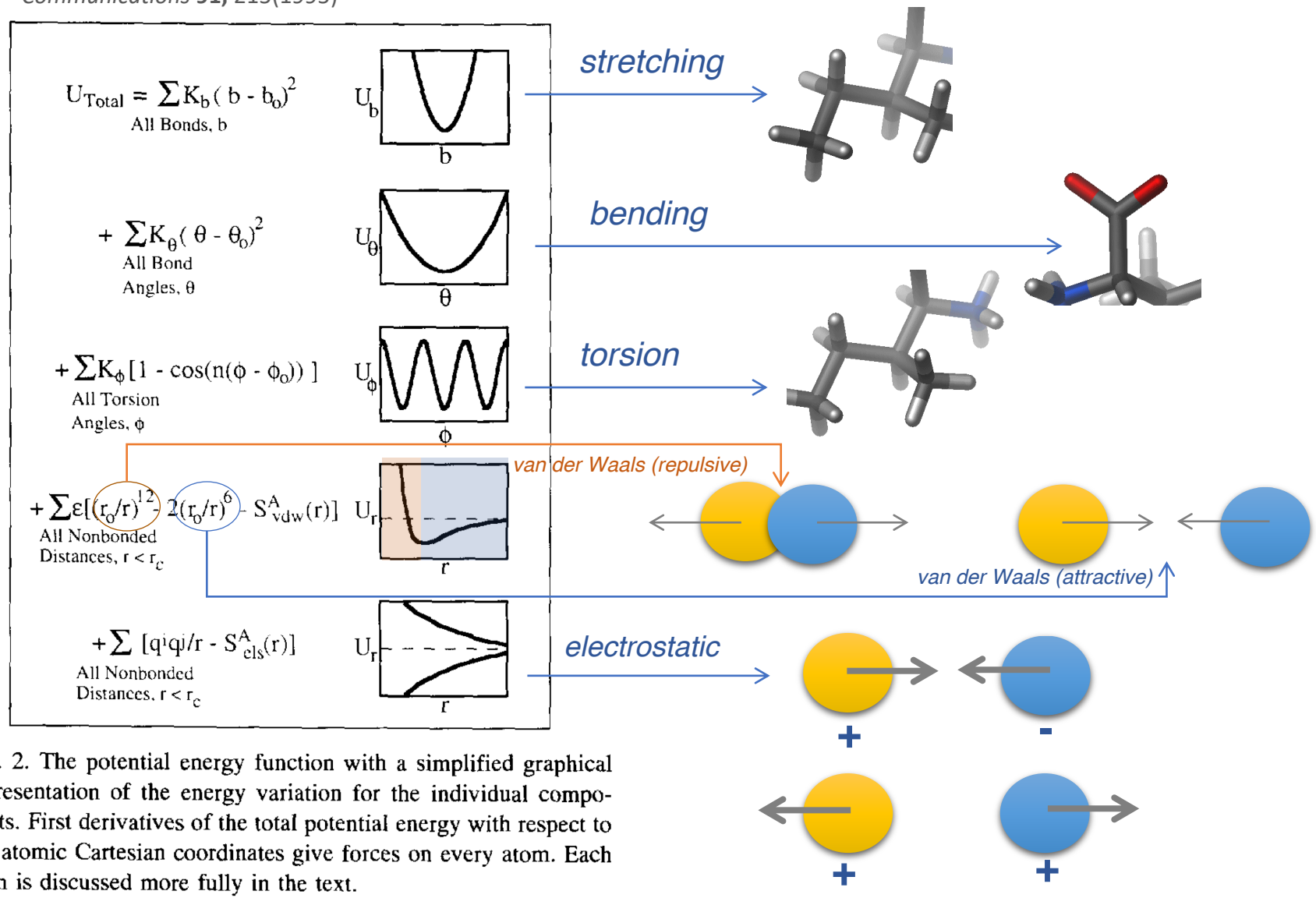


Fig. 2. The potential energy function with a simplified graphical representation of the energy variation for the individual components. First derivatives of the total potential energy with respect to the atomic Cartesian coordinates give forces on every atom. Each term is discussed more fully in the text.

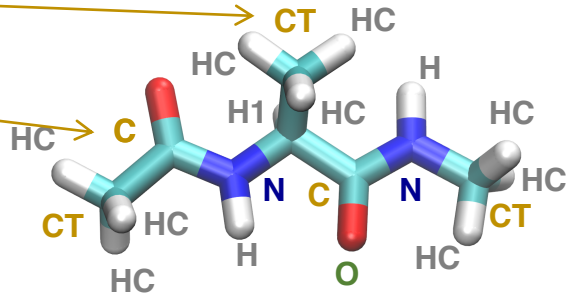
Molecular Mechanics and Force Fields

MM Force Fields

Assignment of Atom Types (based on the chemical environment)

Table 1. List of Atom Types[a]

atom	type	description
carbon	CT	any sp^3 carbon
	C	any carbonyl sp^2 carbon
	CA	any aromatic sp^2 carbon and (Cε of Arg)
	CM	any sp^2 carbon, double bonded
	CC	sp^2 aromatic in 5-membered ring with one substituent + next to nitrogen (Cγ in His)
	CV	sp^2 aromatic in 5-membered ring next to carbon and lone pair nitrogen (e.g. Cδ in His (δ))
	CW	sp^2 aromatic in 5-membered ring next to carbon and NH (e.g. Cδ in His (ε) and in Trp)
	CR	sp^2 aromatic in 5-membered ring next to two nitrogens (Cγ and Cε in His)
	CB	sp^2 aromatic at junction of 5- and 6-membered rings (Cδ in Trp) and both junction atoms in Ade and Gua
	C*	sp^2 aromatic in 5-membered ring next to two carbons (e.g. Cγ in Trp)
	CN	sp^2 junction between 5- and 6-membered rings and bonded to CH and NH (Cε in Trp)
	CK	sp^2 carbon in 5-membered aromatic between N and N-R (C8 in purines)
	CQ	sp^2 carbon in 6-membered ring between lone pair nitrogens (e.g. C2 in purines)
nitrogen	N	sp^2 nitrogen in amides
	NA	sp^2 nitrogen in aromatic rings with hydrogen attached (e.g. protonated His, Gua, Trp)
	NB	sp^2 nitrogen in 5-membered ring with lone pair (e.g. N7 in purines)
	NC	sp^2 nitrogen in 6-membered ring with lone pair (e.g. N3 in purines)
	N*	sp^2 nitrogen in 5-membered ring with carbon substituent (in purine nucleosides)
	N2	sp^2 nitrogen of aromatic amines and guanidinium ions
	N3	sp^3 nitrogen
oxygen	OW	sp^3 oxygen in TIP3P water
	OH	sp^3 oxygen in alcohols, tyrosine, and protonated carboxylic acids
	OS	sp^3 oxygen in ethers
	O	sp^2 oxygen in amides
	O2	sp^2 oxygen in anionic acids
sulfur	S	sulfur in methionine and cysteine
	SH	sulfur in cysteine
phosphorus	P	phosphorus in phosphates
hydrogen	H	H attached to N
	HW	H in TIP3P water
	HO	H in alcohols and acids
	HS	H attached to sulfur
	HA	H attached to aromatic carbon
	HC	H attached to aliphatic carbon with no electron-withdrawing substituents
	H1	H attached to aliphatic carbon with one electron-withdrawing substituent
	H2	H attached to aliphatic carbon with two electron-withdrawing substituents
	H3	H attached to aliphatic carbon with three electron-withdrawing substituents
	HP	H attached to carbon directly bonded to formally positive atoms (e.g. C next to NH_3^+ of lysine)
	H4	H attached to aromatic carbon with one electronegative neighbor (e.g. hydrogen on C5 of Trp, C6 of Thy)
	H5	H attached to aromatic carbon with two electronegative neighbors (e.g. H8 of Ade and Gua and H2 of Ade)

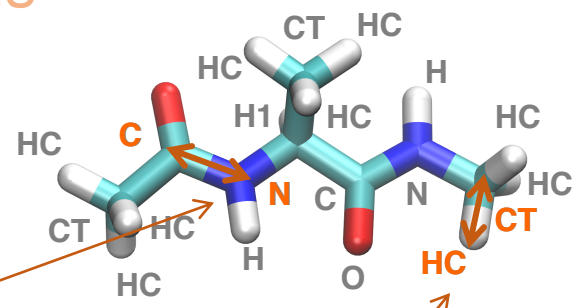


Molecular Mechanics and Force Fields

MM Force Fields

Assignment of parameters

$$U_{stretch}^{X-Y} = K_r^{X-Y} \times \left(r - r_{eq}^{X-Y}\right)^2$$



Bond Parameters

bond	K_r^b	r_{eq}^c	bond	K_r^b	r_{eq}^c	bond	K_r^b	r_{eq}^c	bond	K_r^b	r_{eq}^c
C−CA	469.0	1.409	CA−HA	367.0	1.080	CM−HA	367.0	1.080	CT−S	227.0	1.810
C−CB	447.0	1.419	CA−N2	481.0	1.340	CM−N*	448.0	1.365	CT−SH	237.0	1.810
C−CM	410.0	1.444	CA−NA	427.0	1.381	CN−NA	428.0	1.380	CV−H4	367.0	1.080
C−CT	317.0	1.522	CA−NC	483.0	1.339	CQ−H5	367.0	1.080	CV−NB	410.0	1.394
C−N	490.0	1.335	CB−CB	520.0	1.370	CQ−NC	502.0	1.324	CW−H4	367.0	1.080
C−N*	424.0	1.383	CB−CN	447.0	1.419	CR−H5	367.0	1.080	CW−NA	427.0	1.381
C−NA	418.0	1.388	CB−N*	436.0	1.374	CR−NA	477.0	1.343	H−N	434.0	1.010
C−NC	457.0	1.358	CB−NB	414.0	1.391	CR−NB	488.0	1.335	H−N*	434.0	1.010
C−O	570.0	1.229	CB−NC	461.0	1.354	CT−CT	310.0	1.526	H−N2	434.0	1.010
C−O2	656.0	1.250	CC−CT	317.0	1.504	CT−F	367.0	1.380	H−N3	434.0	1.010
C−OH	450.0	1.364	CC−CV	512.0	1.375	CT−H1	340.0	1.090	H−NA	434.0	1.010
C*−CB	388.0	1.459	CC−CW	518.0	1.371	CT−H2	340.0	1.090	HO−OH	553.0	0.960
C*−CT	317.0	1.495	CC−NA	422.0	1.385	CT−H3	340.0	1.090	HO−OS	553.0	0.960
C*−CW	546.0	1.352	CC−NB	410.0	1.394	CT−HC	340.0	1.090	HS−SH	274.0	1.336
C*−HC	367.0	1.080	CK−H5	367.0	1.080	CT−HP	340.0	1.090	O2−P	525.0	1.480
CA−CA	469.0	1.400	CK−N*	440.0	1.371	CT−N	337.0	1.449	OH−P	230.0	1.610
CA−CB	469.0	1.404	CK−NB	529.0	1.304	CT−N*	337.0	1.475	OS−P	230.0	1.610
CA−CM	427.0	1.433	CM−CM	549.0	1.350	CT−N2	337.0	1.463	OW−HW	553.0	0.9572
CA−CN	469.0	1.400	CM−CT	317.0	1.510	CT−N3	367.0	1.471	S−S	166.0	2.038
CA−CT	317.0	1.510	CM−H4	367.0	1.080	CT−OH	320.0	1.410			
CA−H4	367.0	1.080	CM−H5	367.0	1.080	CT−OS	320.0	1.410			

Molecular Mechanics and Force Fields

Force Field Parametrisation

Force Fields parameters are derived from experimental data or calculations performed at a higher level of theory (Quantum Mechanics)

Equilibrium bond distances and angles: X-ray crystallography

Bond and angle force constants: vibrational spectra, normal mode calculations with Quantum Mechanics (QM)

Dihedral angle parameters: difficult to measure directly with experiments; fit to QM calculations for rotations around a bond with other motions fixed

Atom charges: fit to experimental liquid properties, ESP charge fitting to reproduce electrostatic potentials of high level QM, X-ray crystallographic electron density

van der Waals parameters: often most difficult to determine, fit to experimental liquid properties, intermolecular energy fitting

Molecular Mechanics and Force Fields

Force Fields for Biomolecules

FFs commonly used for biomolecules:

AMBER
CHARMM
GROMOS
OPLS

Improved over time and validated against experimental data, including:

- experimental structures
- secondary structure propensities
- NMR data (e.g. order parameters, chemical shifts, NOEs etc…)

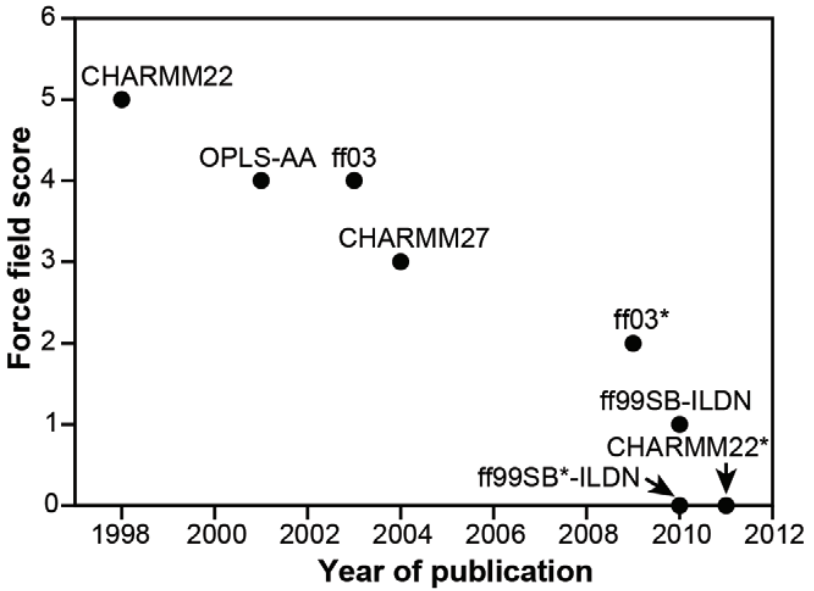


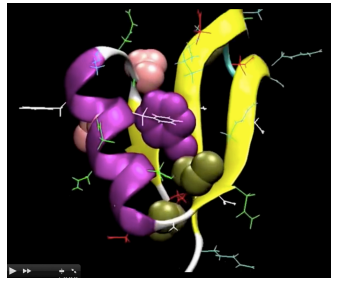
Figure 3. Improvement of force fields over time. For each force field, we assigned a score depending on the agreement with experiments in the tests presented here. Low scores indicate good agreement with experiments. These scores are plotted against the year in which the force field was published. For the force fields that involve multiple corrections (e.g., ff99SB*-ILDN), we use the year of the most recently published correction.

Lindorff-Larsen, K. *et al. PLoS ONE* **7,** e32131 (2012).

Ab-initio folding from first principles Force Fields and MD

Study of folding mechanisms

final structure



http://www.youtube.com/watch?v=gFcp2Xpd29I

Voelz, V. A. et al. *J. Am. Chem. Soc.* **132,** 1526–1528 (2010).

Simulation by Computer of a protein fast folder

Ab initio folding (force-field and simulation based)

http://www.youtube.com/watch?v=gFcp2Xpd29I

https://www.youtube.com/watch?v=sD6vyfTtE4U

Learning outcomes of this lecture:

Why do we need to predict protein structures
Importance of 3D structure Knowledge
Folding: Anfinsen theorem Levinthal paradox
Structure Prediction: Comparative Modelling: which are the main steps involved?
Main problems encountered in Comparative Modelling
What is DSSP and which information gives?
Assessing protein structures
Analysing Structures
Analysing Dynamics of Structures

Suggested readings

Secondary structure predictions:

1) DSSP: Kabsch W, Sander C.Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 22(12):2577-637.
2) Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Adv Enzymol Relat Areas Mol Biol. 1978;47:45-148.
3) Chen H, Gu F, Huang Z. Improved Chou-Fasman method for protein secondary structure prediction. BMC Bioinformatics. 2006 12:7 Suppl 4:S14

Tertiary structure predictions:

1) Forster MJ (2002) Molecular Modelling in structural biology. Micron 33, 365-384
2) Sutcliffe MJ, Hayes FR, Blundell TL. Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. Protein Eng. 1987 1(5):385-92.
3) Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng. 1987 1(5):377-84.

Very recent methods

Protein structure determination using metagenome sequence data. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. Science. 2017 Jan 20;355(6322):294-298. doi: 10.1126/science.aah4043.

De novo structure prediction with deeplearning based scoring R.Evans *,1 , J.Jumper *,1 , J.Kirkpatrick *,1 , L.Sifre *,1 , T.F.G.Green 1 , C.Qin ,1 , A.Zidek 1 , A.Nelson 1 , A.Bridgland 1 , H.Penedones 1 , S.Petersen 1 , K.Simonyan 1 , S.Crossan 1 , D.T.Jones 2 , D.Silver 1 , K.Kavukcuoglu 1 , D.Hassabis 1 , A.W.Senior (Goodle Deepmind)

Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Greener JG, Kandathil SM, Jones DT. Nat Commun. 2019 Sep 4;10(1):3977. doi: 10.1038/s41467-019-11994-0.

Web resources

Introductory MD simulation tutorials and How-tos

GROMACS (freely available)
http://www.gromacs.org/Documentation/Tutorials#General_GROMACS_Use
http://www.gromacs.org/Documentation/How-tos

NAMD (freely available)
http://www.ks.uiuc.edu/Training/Tutorials/index-all.html#namd

AMBER
http://ambermd.org/tutorials/#basic_tut

VMD (visualisation of MD trajectories, freely available)
http://www.ks.uiuc.edu/Training/Tutorials/vmd/tutorial-html/

YASARA
http://www.yasara.org/movies.htm

Servers

MDWeb
http://mmb.irbbarcelona.org/MDWeb/ (System set-up)
We-NMR
https://www.wenmr.eu/wenmr/molecular-dynamics-software (run simulations on the GRID, requires registration)

Databases

MolMovDB (database of macromolecular motions, contains morphs between different conformations)
http://www.molmovdb.org/