Humans, aliens, and eHarmony
or
why there is no such thing as a free lunch in protein structure determination from sparse experimental data

Mark Berjanskii, July 27th, 2012

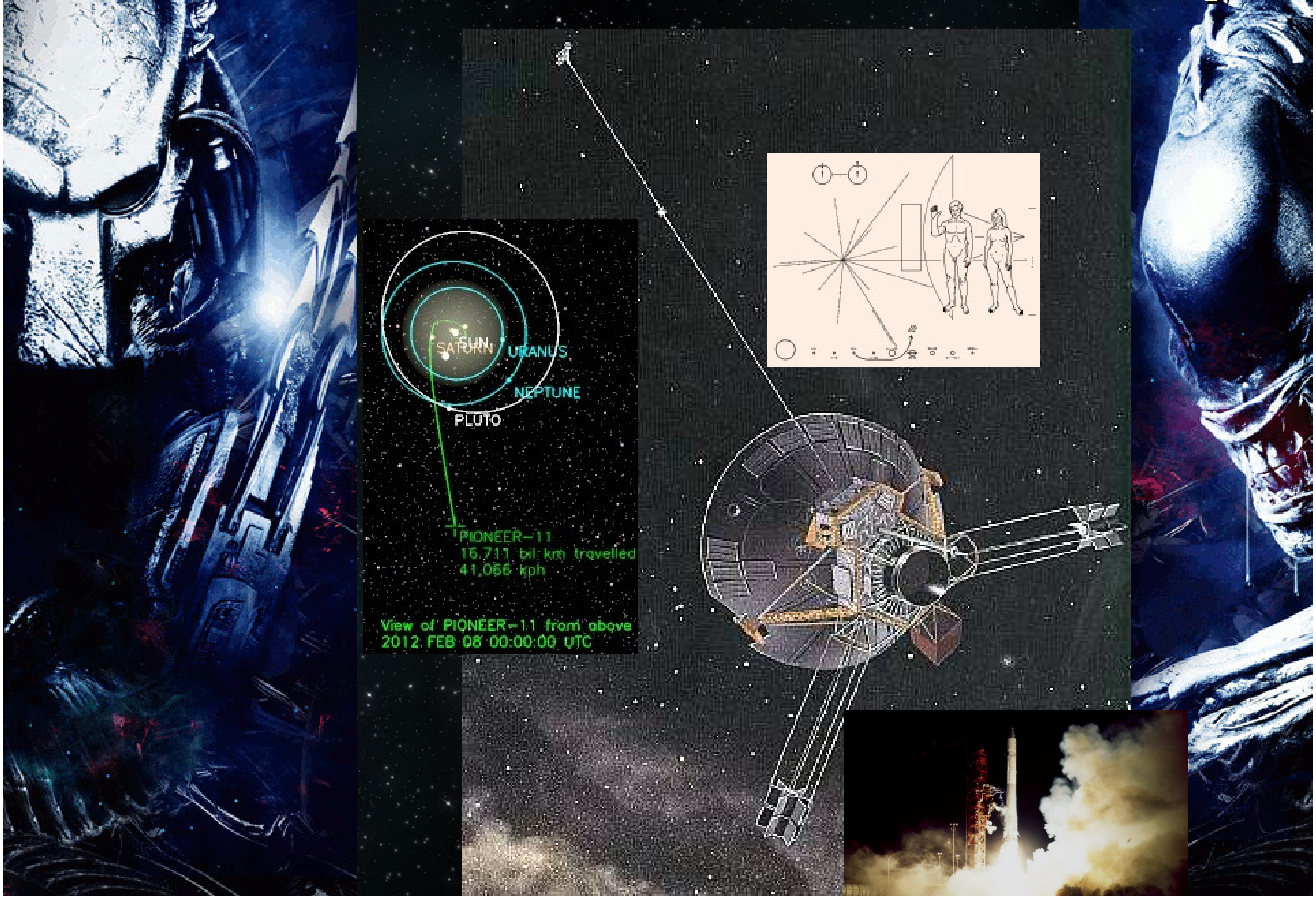
… dedicated to all experimentalists who lost their way

Outline

- Purposes and origins of protein structural models
- Theoretical models of protein structure
- Models from non-sparse experimental data
- Models from sparse experimental data
- Recent developments

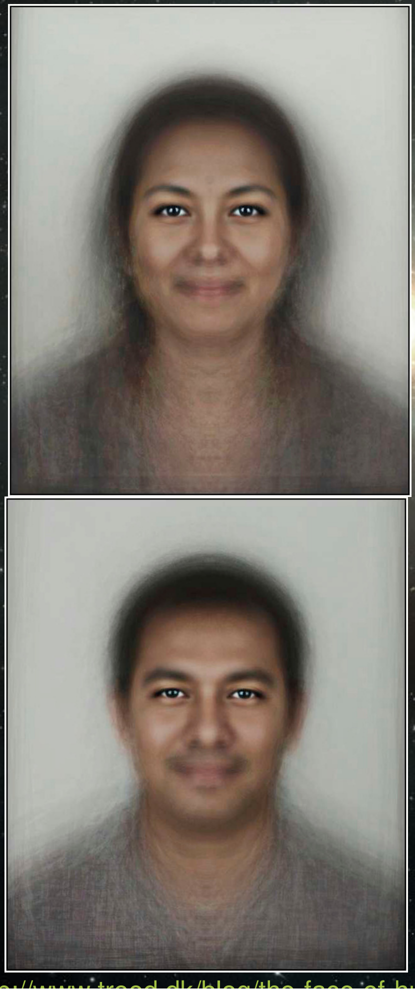
Humans, aliens, and eHarmony®

Humans, aliens, and eHarmony



Humans, aliens, and eHarmony

Typical human face
by David Tood



http://www.trood.dk/blog/the-face-of-humanity/

What is enough for aliens, not enough for eHarmony

Typical ≠ Accurate

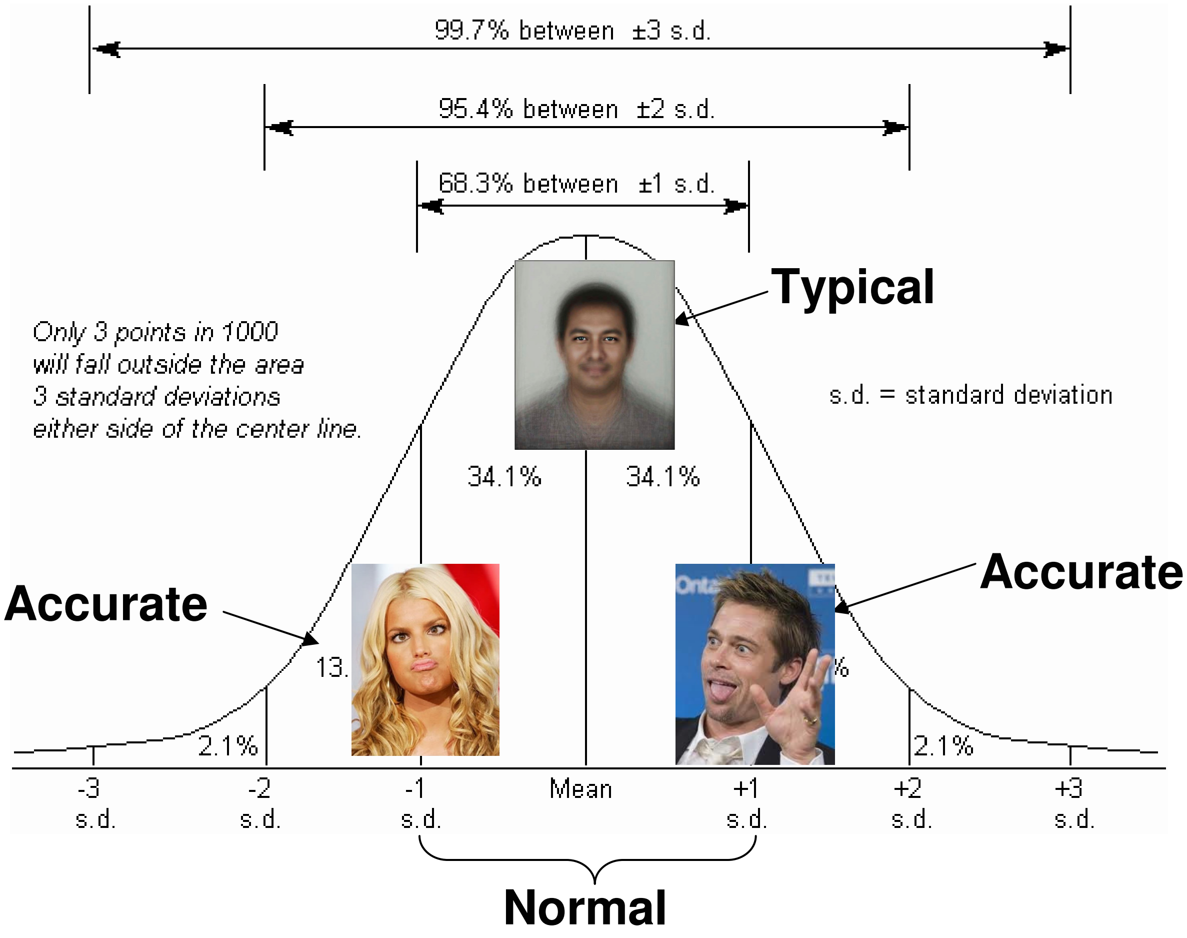
Typical human face
by David Tood

http://www.trood.dk/blog/the-face-of-humanity/

“Accurate” human faces

Take-home message

Typical ≠ Accurate



Typical

Accurate

Accurate

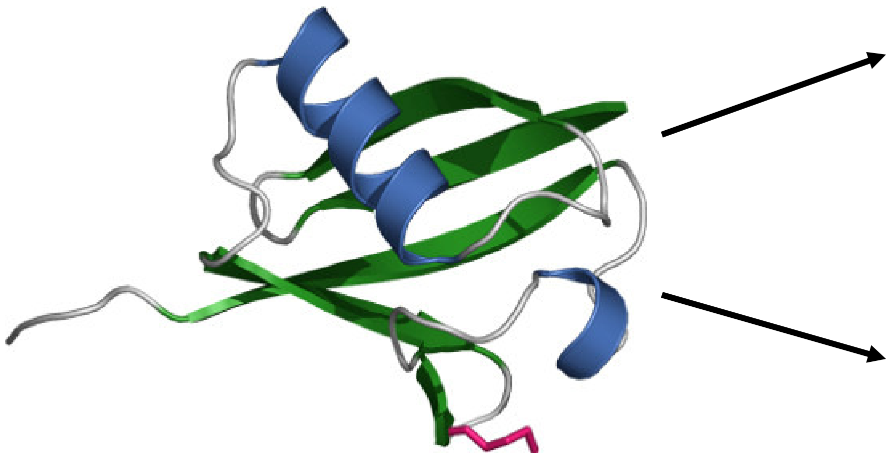
Normal

Warning!!! Most protein structure validation tools check how typical or normal your protein model is, not how accurate your protein model is.

The level of required protein structural accuracy also depends on its purpose

Why do we need to know protein structures?

1) Prediction of protein function from 3D structure (e.g. fold, motifs, active site prediction)



1) Ubiquitin

- degradation by the proteasome,

2) Ubiquitin-like modifiers

- function regulation by post-translation modification

2) Sequence-to-function prediction

3) Mechanism of protein function (e.g. enzyme catalysis, structural effect of known mutations).

4) Rational drug design and structure design

5) Design of novel proteins with novel function

When do we need to do a structural experiment?

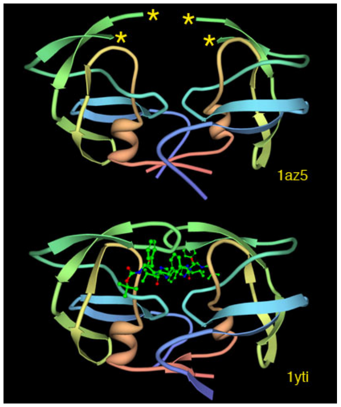
1) Structure is not known

2) Structure is known but can not be used to answer your scientific question

a) structure is incomplete

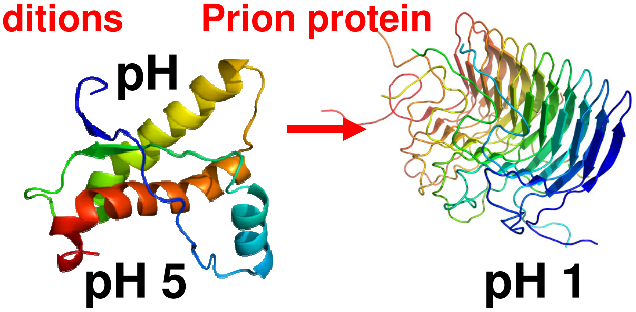
b) structure was determined at different conditions

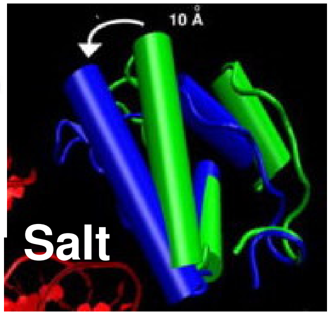
SIV protease



Tom Cruse under stress conditions

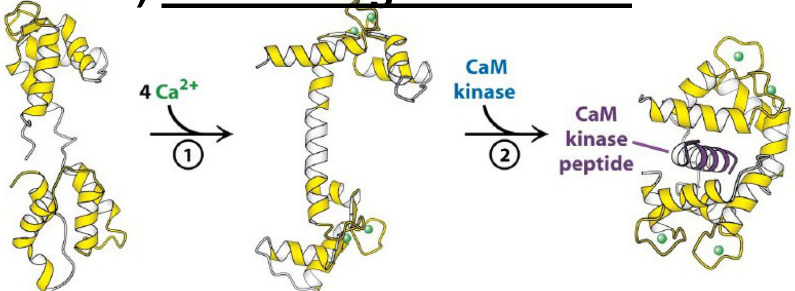
Prion protein



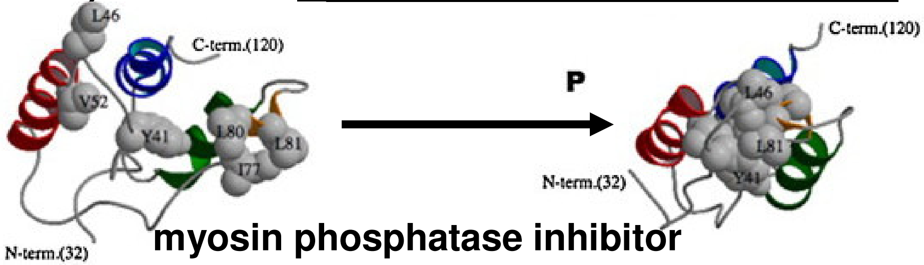


African swine fever virus DNA polymerase X
50mM salt vs 500 mM salt

c) different liganded state

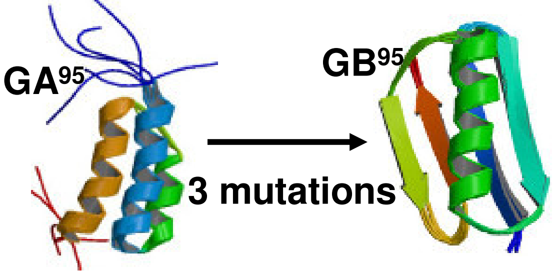


d) different post-tranlational modification

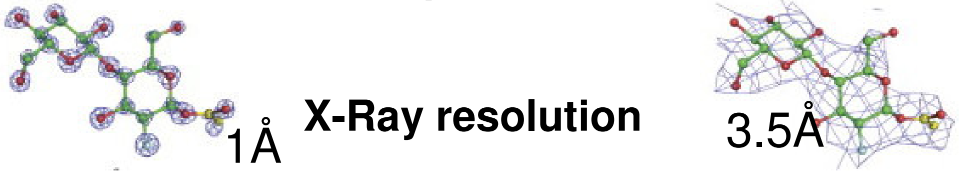


myosin phosphatase inhibitor

e) mutations



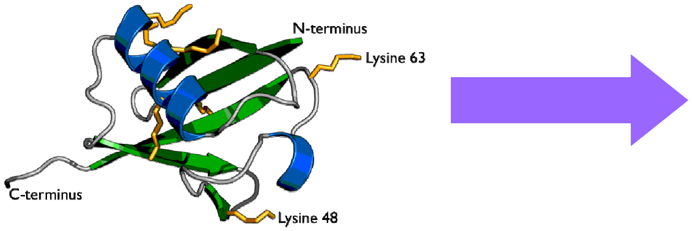
f) Insufficient experimental data



X-Ray resolution

Why use incomplete experimental data for protein structure determination?

1) Some biological questions (e.g. prediction of function from protein fold) may not require high structural accuracy



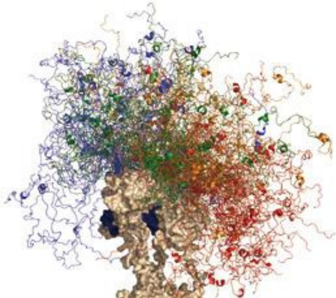
Ubiquitin

2) Some biologically interesting proteins are too difficult to study by any high-res method:

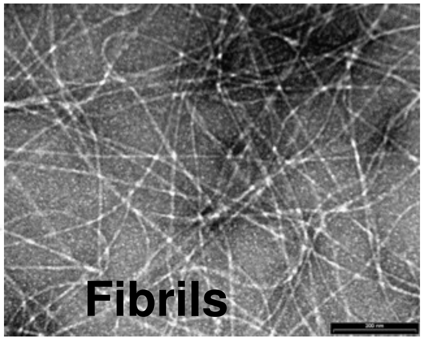
- proteins with extended flexible regions

-**large proteins**

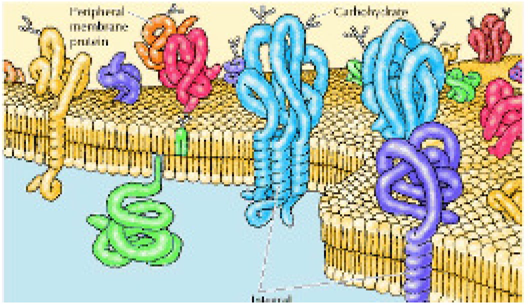
- fibrillar and membrane proteins



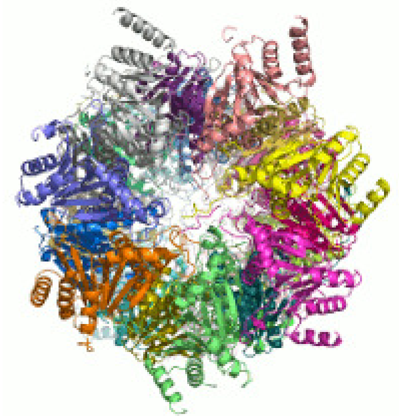
Flexible



Fibrils

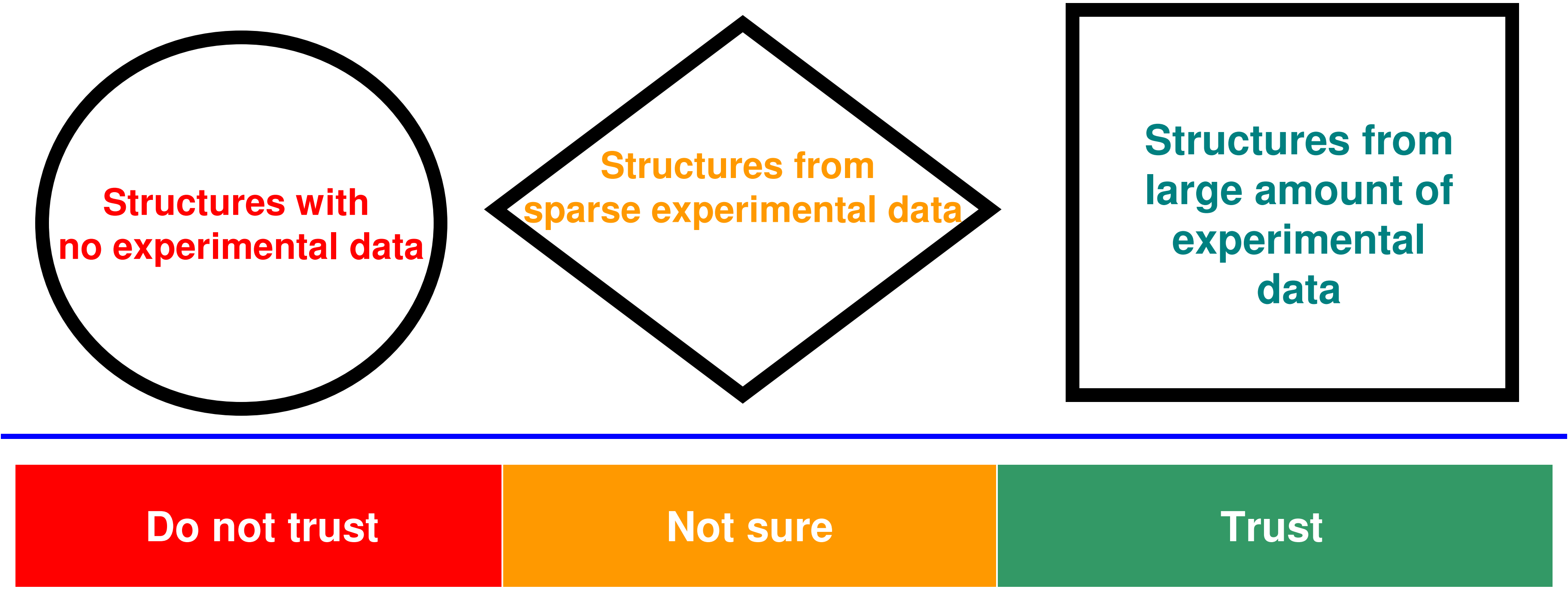


Membrane proteins



Large proteins

Protein structures from the point of view of an experimentalist



Structures with no experimental data

Structures from sparse experimental data

Structures from large amount of experimental data

Do not trust

Not sure

Trust

What is expert's opinion?

PDBsum	A database of
ProFunc	Analysis of a
EC-PDB	Enzyme Struc
SAS	Sequence An
DrugPort	Database of d
Arch Schema	Interactive gra
	Atlas of sidec
ENCODE	ENCODE protein analysis

Software

PROCHECK
Program to check stereochemical quality of protein structures

LigPlot+
GUI version of LIGPLOT, including superposition of related plots

LIGPLOT
Program to plot schematic diagrams of protein-ligand interactions

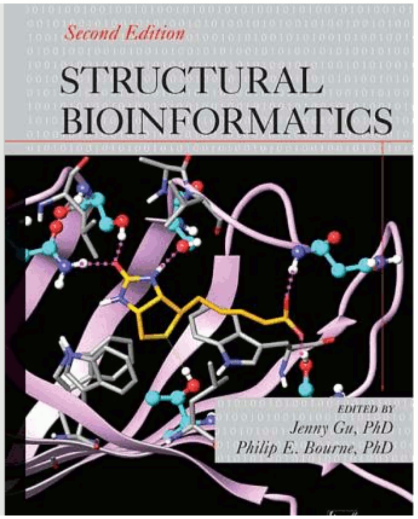
Roman Laskowski

Research Scientist at European Bioinformatics Institute

STRUCTURAL QUALITY ASSURANCE

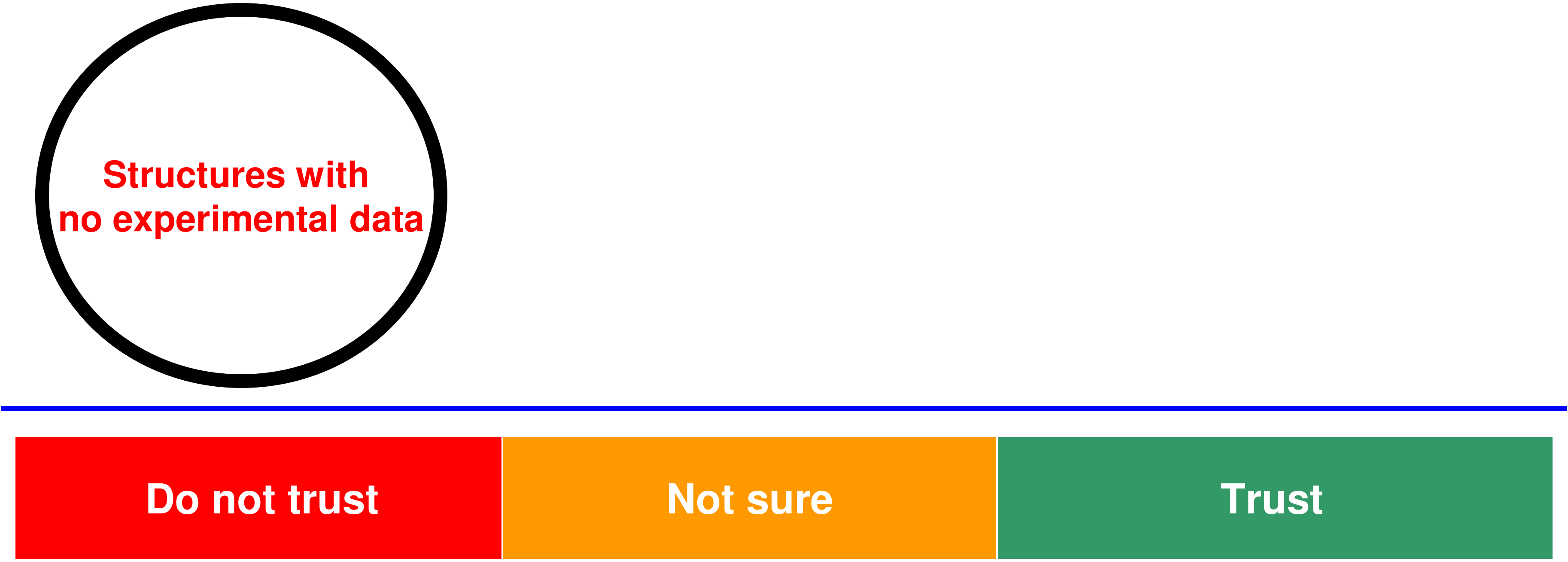
Roman A. Laskowski

Jenny Gu (Editor),
Philip E. Bourne (Editor)
ISBN: 978-0-470-18105-8
Hardcover
1067 pages
John Wiley & Sons, Inc.



Non-experimental structures

Protein structures from the point of view of an experimentalist



Structures with no experimental data

Do not trust

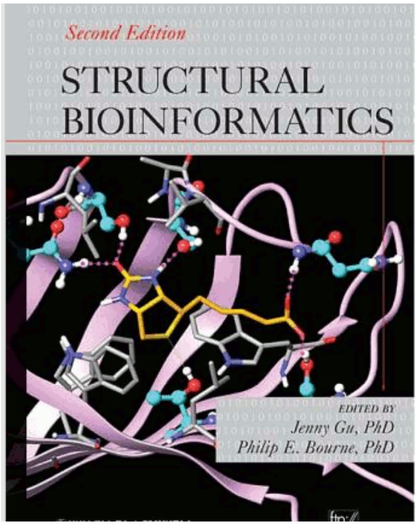
Not sure

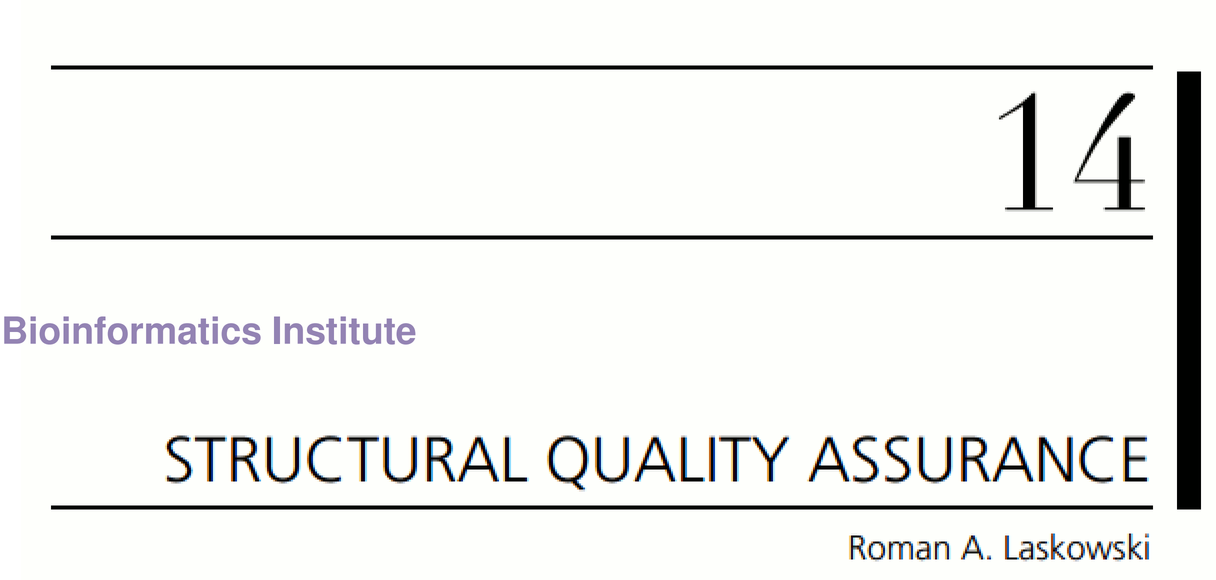
Trust

What does Roman Laskowski write about theoretical structures?

Roman Laskowski

Research Scientist at European Bioinformatics Institute

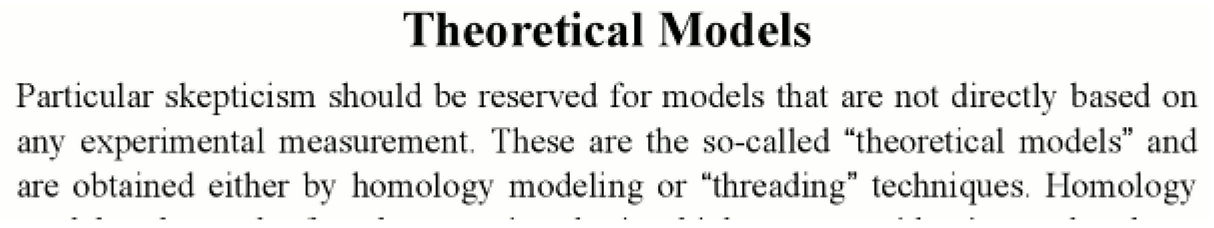




14

STRUCTURAL QUALITY ASSURANCE

Roman A. Laskowski



Theoretical Models

Particular skepticism should be reserved for models that are not directly based on any experimental measurement. These are the so-called "theoretical models" and are obtained either by homology modeling or "threading" techniques. Homology

What are the criteria of success in protein structure determination?

1) In method development tests:

- Global accuracy: RMSD, TM-score

- Agreement with experimental data (when available)
- Agreement with protein quality metrics

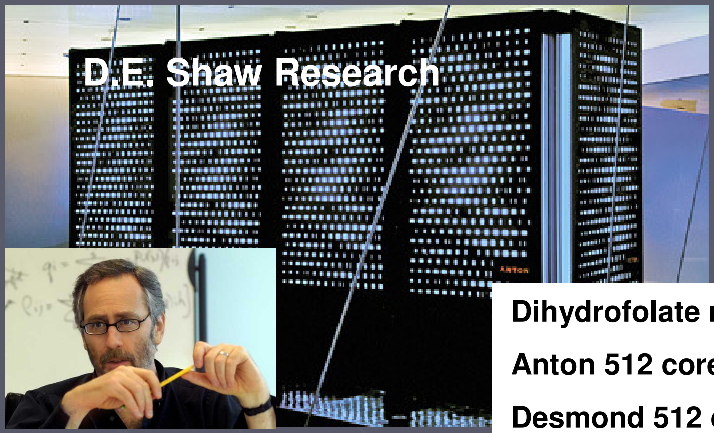
2) In real-life research:

- Global accuracy

- Agreement with experimental data (when available)
- Agreement with protein quality metrics

Ab-Initio structures by molecular dynamics

SuperComputer “Anton” for MD simulations



Dihydrofolate reductase:

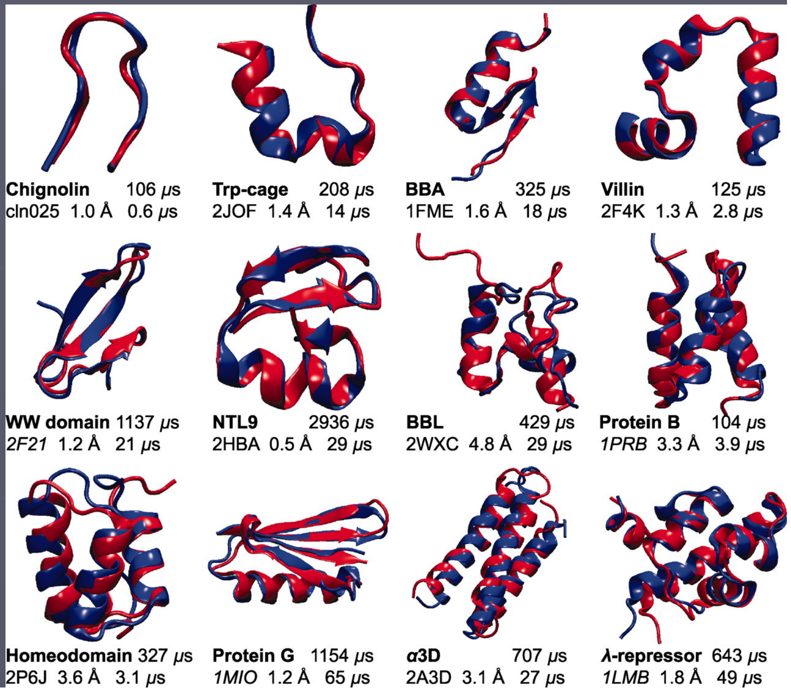
Anton 512 cores: 15 μs/day

Desmond 512 cores: 0.5 μs/day

Amber-GPU 64 cores: 80 ns/day

Amber 48 cores: 20ns/day

Gromacs 8 cores: ~5-10ns/day



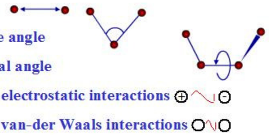
How Fast-Folding Proteins Fold
Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, David E. Shaw;
Science 28 October 2011: Vol. 334 no. 6055 pp. 517-520

MD force-field

$$
\begin{aligned}
V(\bar{\mathbf{r}}) &= \sum_{\text{bonds}} \overset{(1)}{K_b(b-b_0)^2} + \sum_{\text{angles}} \overset{(2)}{K_\theta(\theta-\theta_o)^2} \\
&+ \sum_{\text{dihedrals}} \overset{(3)}{K_\chi(1+\cos(n\chi-\delta))} \\
&+ \sum_{\text{nonbonded-pairs},i,j} \left[\overset{(4)}{\frac{q_i q_j}{4\pi e_o r_{ij}}} - \varepsilon_{ij} \left\{ \left(\frac{R_{\min ij}}{r_{ij}} \right.^{12} \overset{(5)}{-} 2 \left(\frac{R_{\min ij}}{r_{ij}} \right.^{6} \right\} \right]
\end{aligned}
$$

Energy dependencies on:

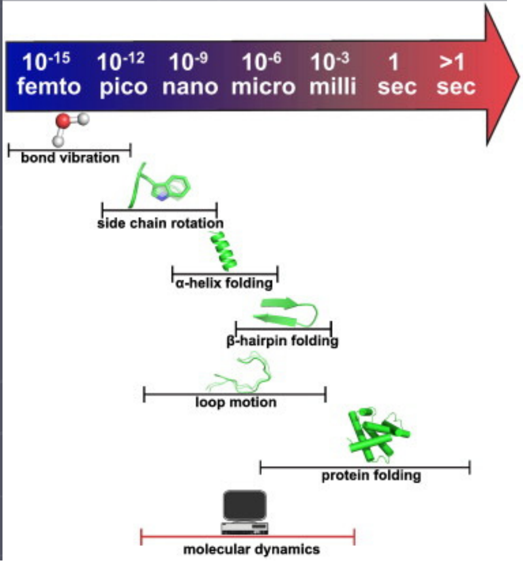
1. Bond length
2. Bond valence angle
3. Bond dihedral angle
4. Non-bonded electrostatic interactions
5. Non-bonded van-der Waals interactions



Newton equation of motion

$$\vec{f}_i = m_i \vec{a}_i$$

Folding time-scales:



Limitations:

1) Requires powerful hardware or computing time

2) Limited to small/simple proteins

3) Can not take into account chaperone action

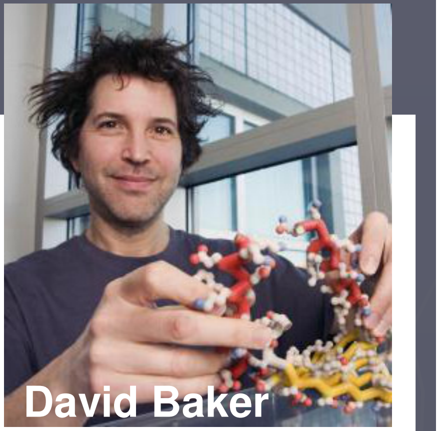
4) Criteria for success???

Fragment-based *ab initio* structures

(Non-experimental structures)

Rosetta

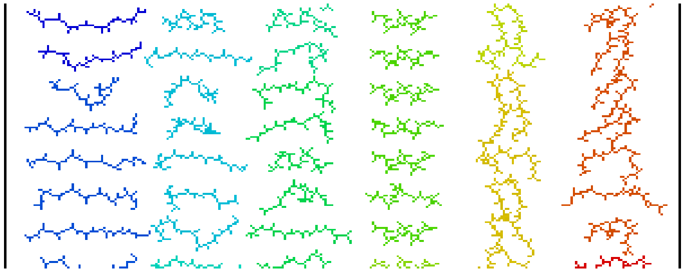
Developed by 12 labs
50 people



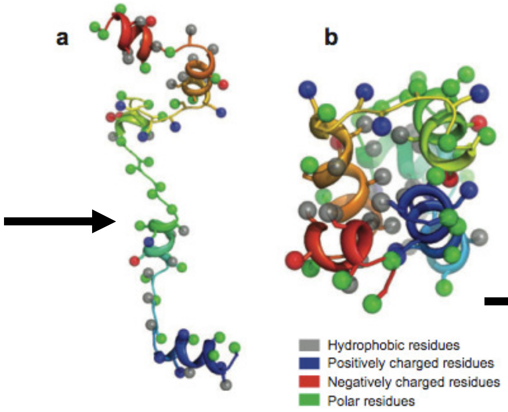
Sequence

Secondary Structure

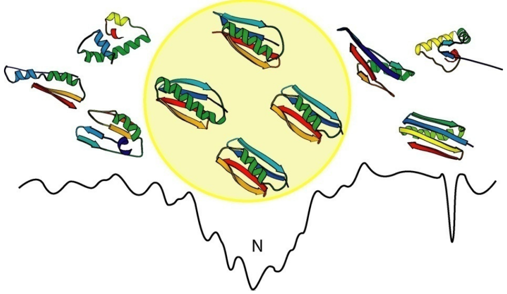
3- and 9-residue fragments



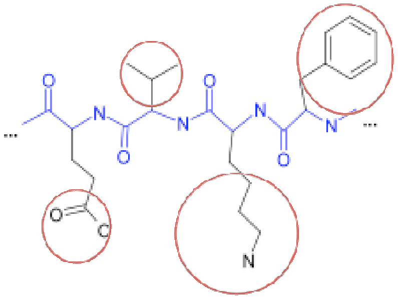
Low-resolution folding



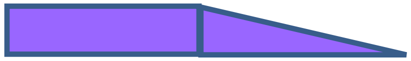
Best low-res decoy selection



Full-atom side-chain restoration

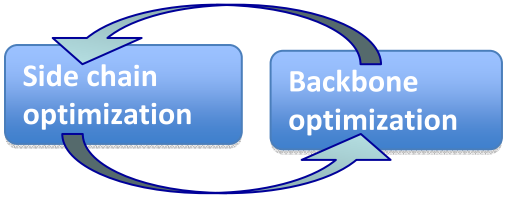


Full-atom refinement

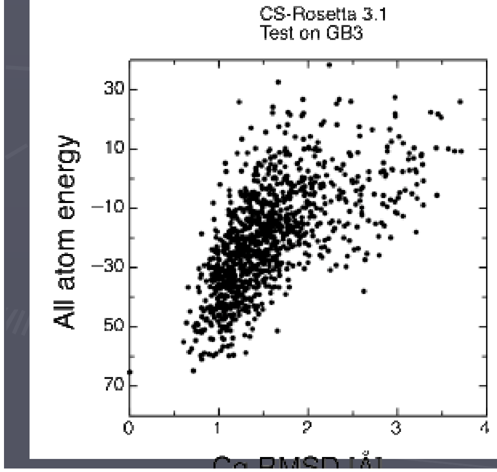


vdW repulsive

Small backbone moves



Model quality evaluation

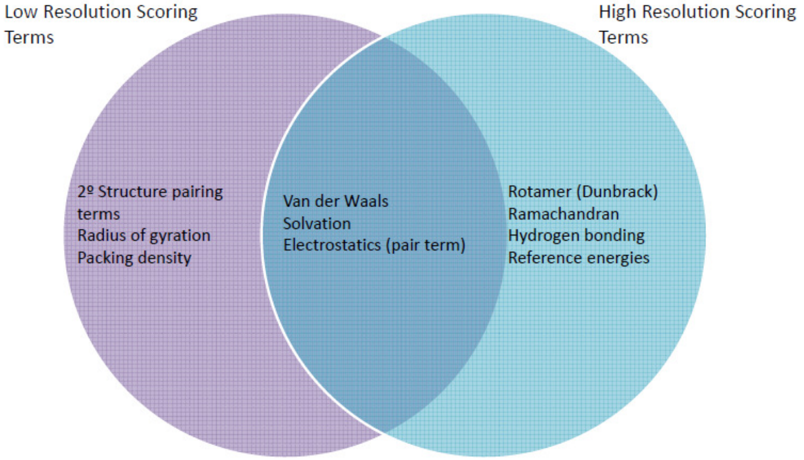


Fragment-based *ab initio* structures

(Non-experimental structures)

Rosetta

Scoring function



Low Resolution Scoring Terms

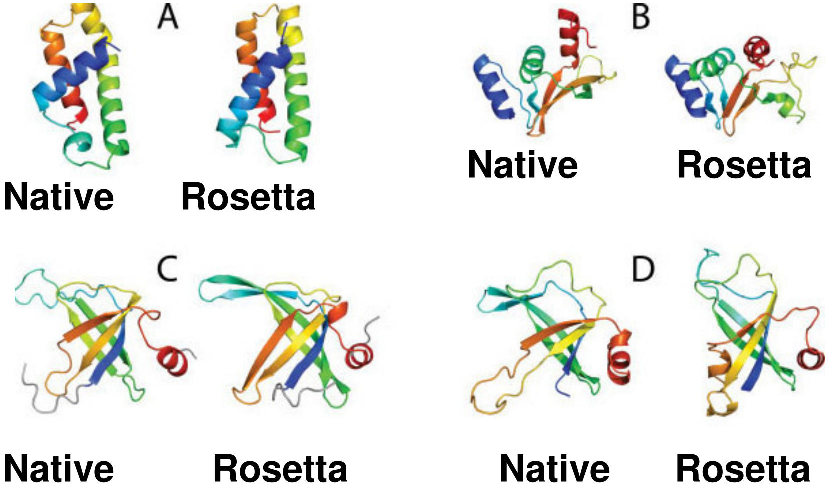
High Resolution Scoring Terms

2º Structure pairing terms
Radius of gyration
Packing density

Van der Waals
Solvation
Electrostatics (pair term)

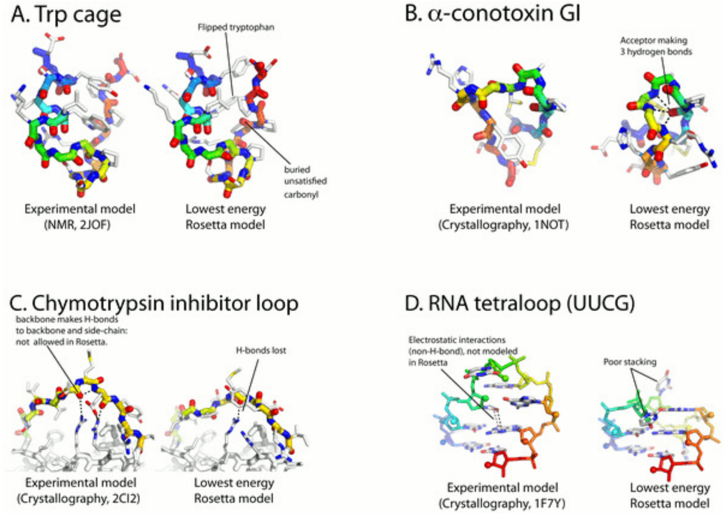
Rotamer (Dunbrack)
Ramachandran
Hydrogen bonding
Reference energies

Rosetta can fold small proteins (<100 residues)



Rosetta limitations

1) Does not fold well proteins above 100 residues (sampling problem)
2) Biased by fragment structure
3) Implicit solvation score is too simplistic and only weakly disfavors buried unsatisfied polar groups.
4) Hydrogen bond potential neglects the effects of charged atoms, (anti-) cooperativity within H-bond networks .
5) Ignores electrostatic interactions (besides H-bonds) and their screening,
6) Does not permit rigorous estimation of a model's free energy.
7) Does not fold properly some very small proteins and RNA


A. Trp cage
Flipped tryptophan
buried unsatisfied carbonyl
Experimental model (NMR, 2JOF)
Lowest energy Rosetta model
B. α-conotoxin GI
Acceptor making 3 hydrogen bonds
Experimental model (Crystallography, 1NOT)
Lowest energy Rosetta model
C. Chymotrypsin inhibitor loop
backbone makes H-bonds to backbone and side-chain: not allowed in Rosetta.
H-bonds lost
Experimental model (Crystallography, 2CI2)
Lowest energy Rosetta model
D. RNA tetraloop (UUCG)
Electrostatic interactions (non-H-bond), not modeled in Rosetta
Poor stacking
Experimental model (Crystallography, 1F7Y)
Lowest energy Rosetta model

Rosetta forces protein normality

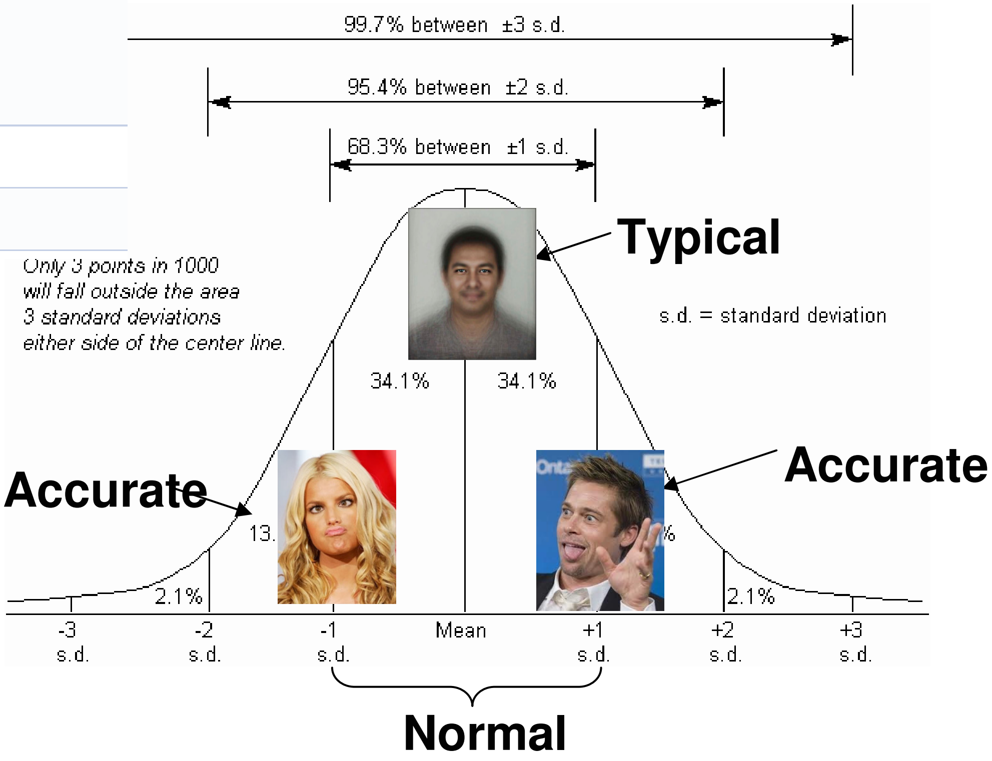
Rosetta scoring function

```
lennard-jones attractive
lennard-jones repulsive
lazaridis-jarplus solvation energy
lennard-jones repulsive between atoms in the same residue
statistics based pair term, favors salt bridges
pi-pi interaction between aromatic groups, by default = 0
internal energy of sidechain rotamers as derived from Dunbrack's statistics
reference energy for each amino acid
backbone-backbone hbonds distant in primary sequence
backbone-backbone hbonds close in primary sequence
sidechain-backbone hydrogen bond energy
sidechain-sidechain hydrogen bond energy
Probability of amino acid at phipsi
distance score in current disulfide
csangles score in current disulfide
dihedral score in current disulfide
ca dihedral score in current disulfide
proline ring closure energy
```

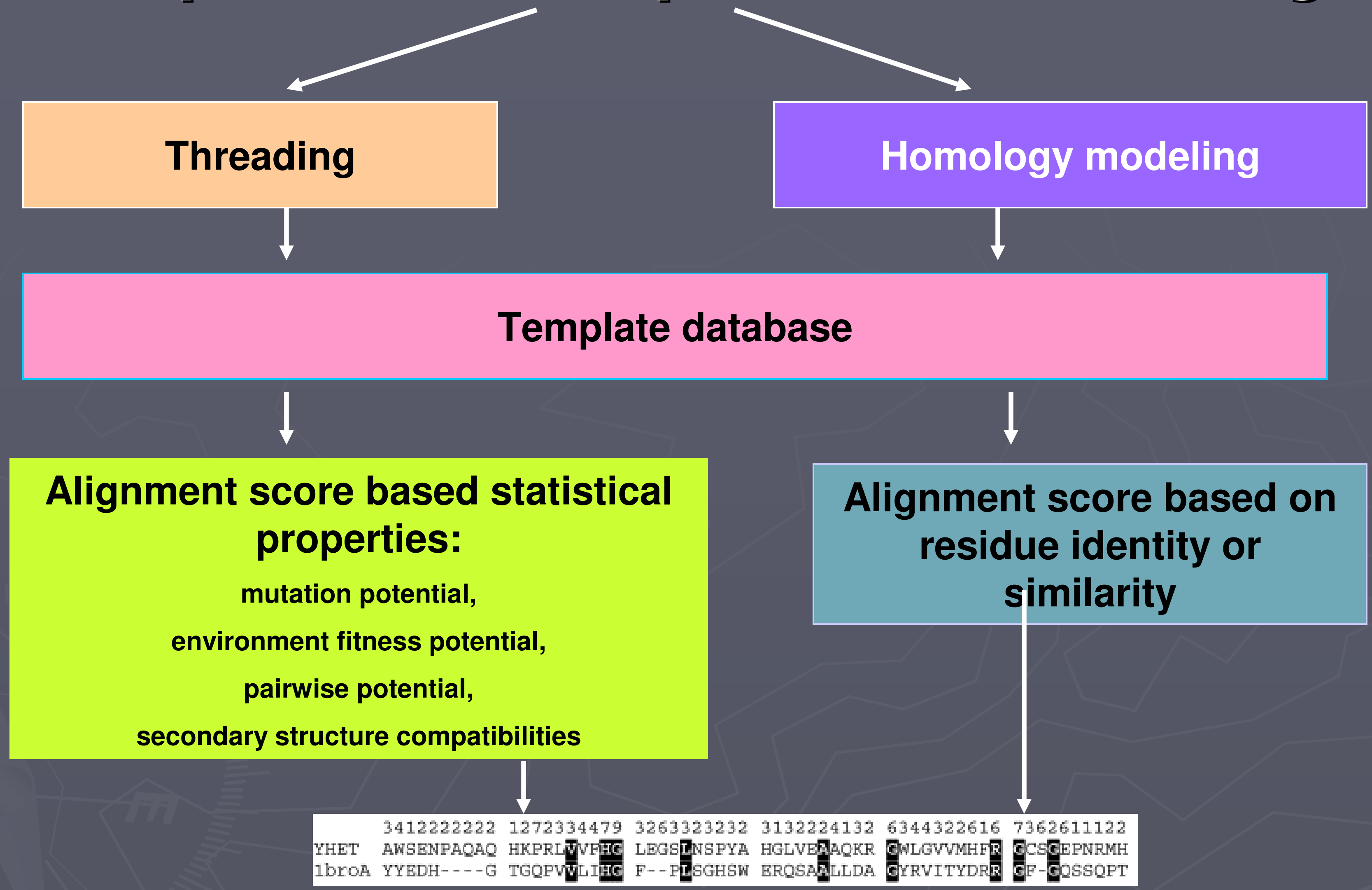
```
ramachandran preferences
omega dihedral in the backbone
```

Fragment idealization

Typical ≠ Accurate



Comparative or template-based modeling



Threading

Homology modeling

Template database

Alignment score based statistical properties:

mutation potential,

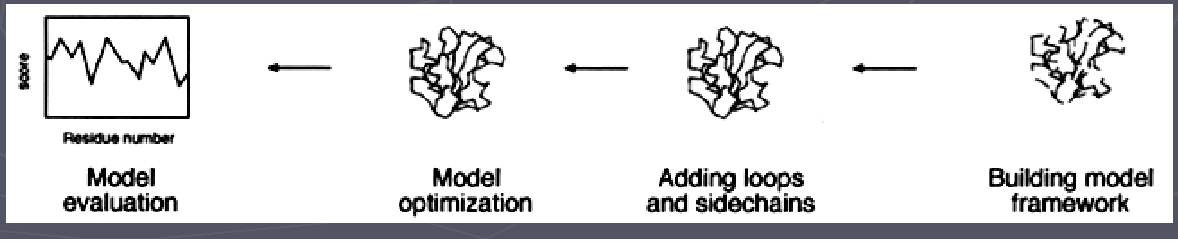
environment fitness potential,

pairwise potential,

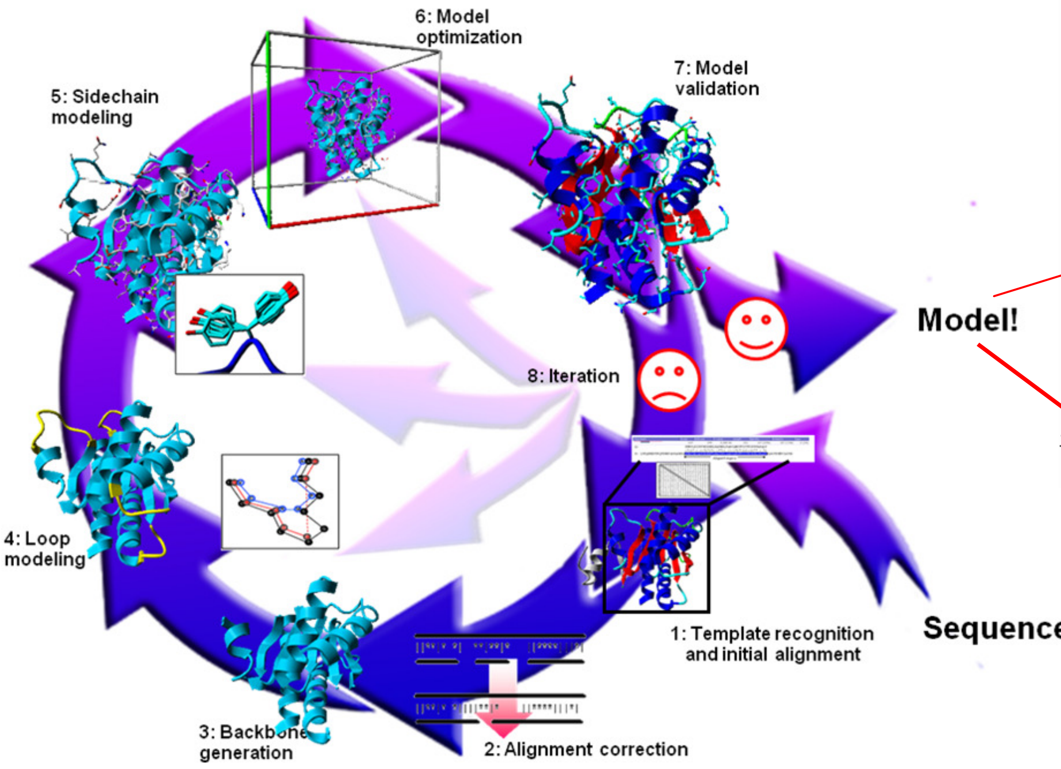
secondary structure compatibilities

Alignment score based on residue identity or similarity

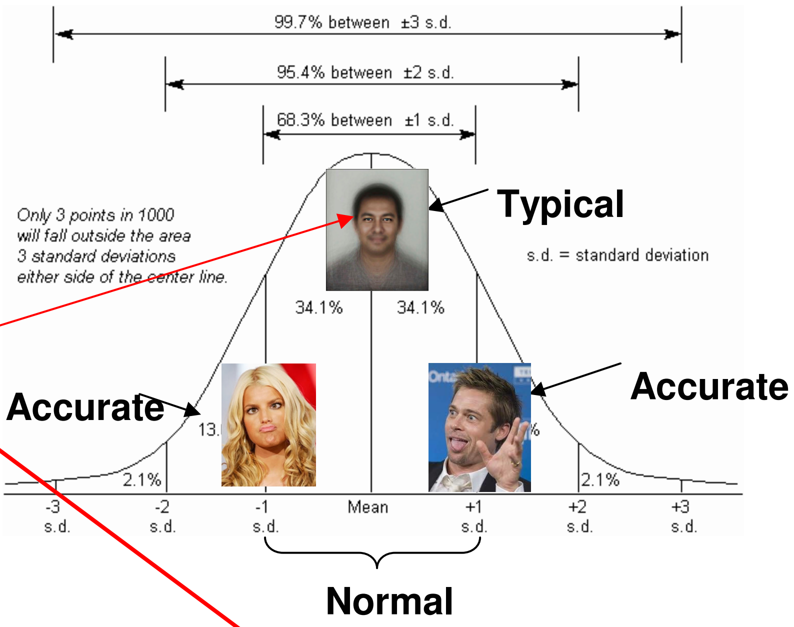
```
      3412222222 1272334479 3263323232 3132224132 6344322616 7362611122
YHET  AWSENPAQAQ HKPRLVVFHG LEGSLNSPYA HGLVEAAQKR GWLGVVMHFR GCSGEPNRMH
1broA YYEDH----G TGQPVVLIHG F--PLSGHSW ERQSAALLDA GYRVITYDRR GF-GQSSQPT
```

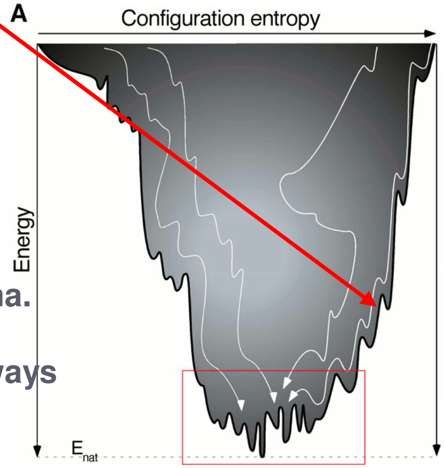


In a real-life scenario, success of homology modeling is judged based on model normality, not model accuracy.



Typical ≠ Accurate

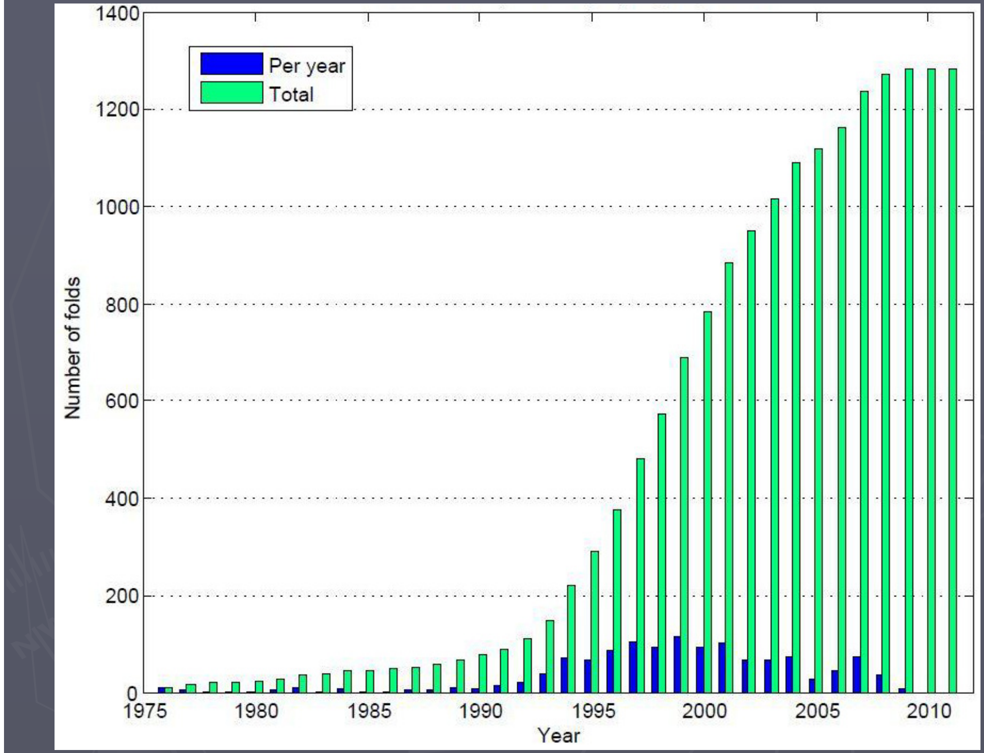




Theoretical scores for protein quality assessment may have wrong energetic minima.

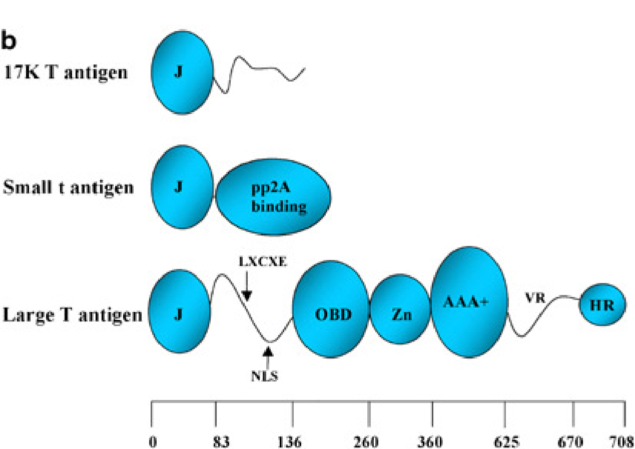
Limited conformational sampling may not always yield the native conformation

Some people may think that any structure can be determined via homology modeling because "all" folds of NMRable and XRAYable proteins are "known"



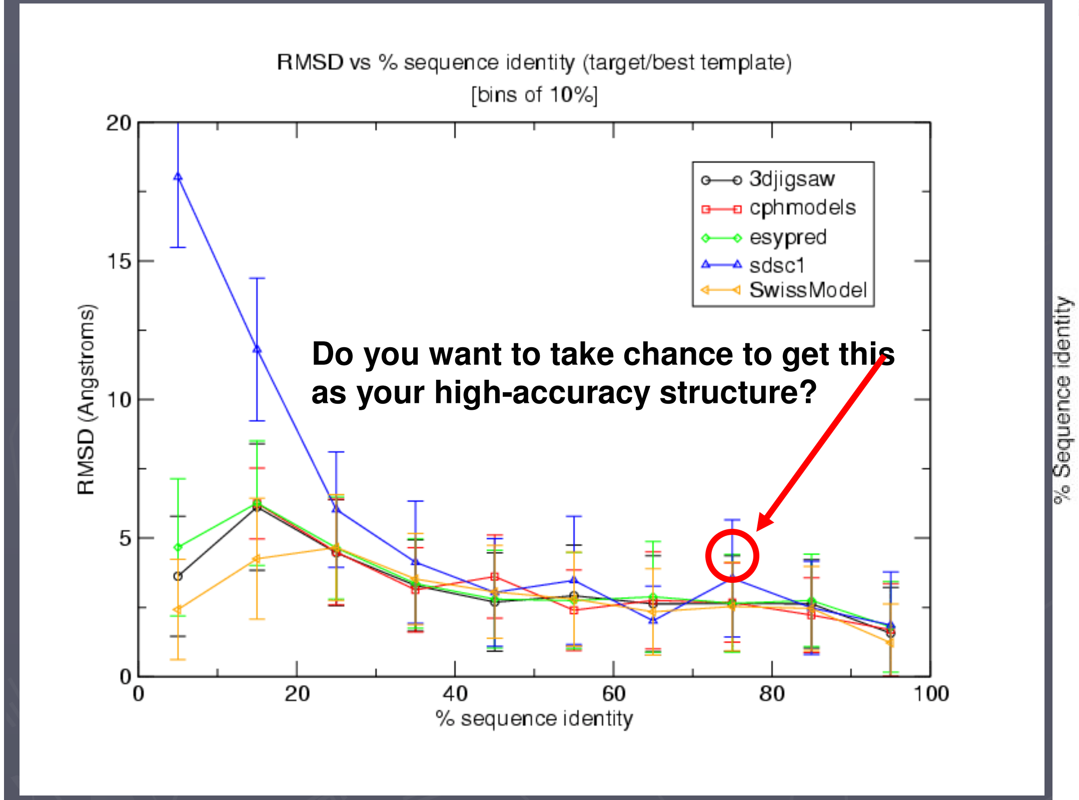
But can simple folds provide all necessary information to define domain orientation in and overall structure of complex proteins?

SV40 T-antigen



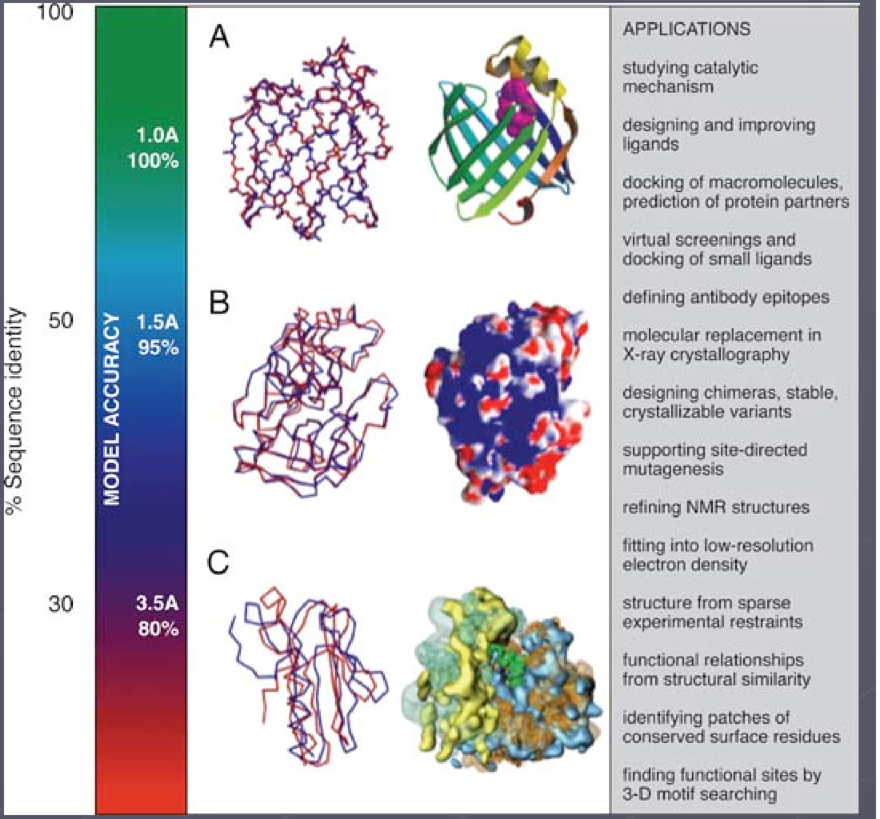
NO!

Accuracy of template-based modeling



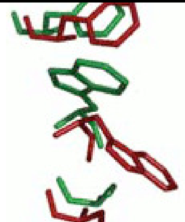
Do you want to take chance to get this as your high-accuracy structure?

http://swissmodel.expasy.org/workspace/tutorial/eva.html

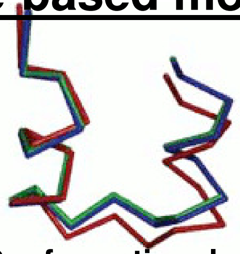

100
50
30
% Sequence identity
MODEL ACCURACY
1.0A
100%
1.5A
95%
3.5A
80%
A
B
C
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screenings and docking of small ligands
defining antibody epitopes
molecular replacement in X-ray crystallography
designing chimeras, stable, crystallizable variants
supporting site-directed mutagenesis
refining NMR structures
fitting into low-resolution electron density
structure from sparse experimental restraints
functional relationships from structural similarity
identifying patches of conserved surface residues
finding functional sites by 3-D motif searching

Curr Protoc Bioinformatics. 2006 Oct;Chapter 5:Unit 5.6.
Comparative protein structure modeling using Modeller.
Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A.

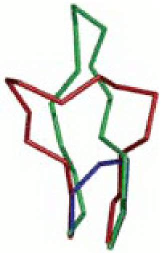
Errors in template-based models:



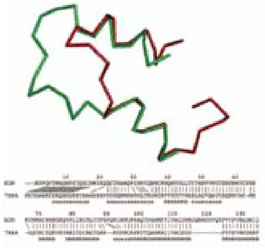
Wrong side-chain packing



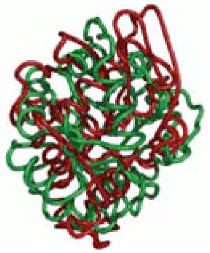
Conformational changes



No-template segments



Mis-alignment

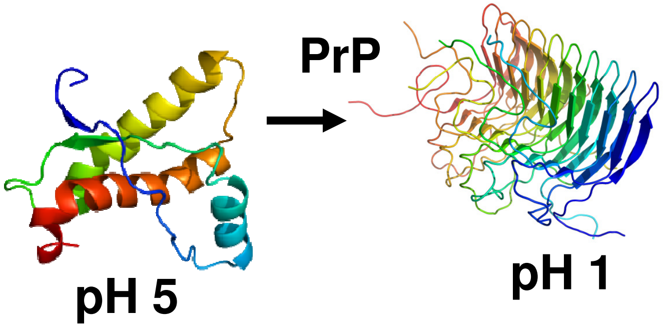


Wrong template

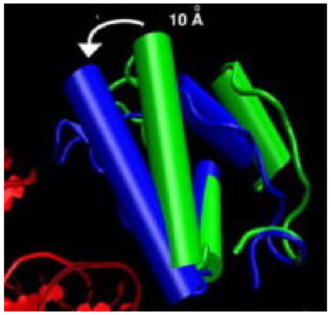
Template-based models have strong 3D bias to the template

Even 100% identical proteins may have very different structures due to:

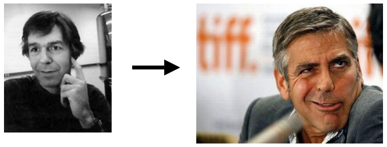
1) Different pHs



2) Different ionic strength

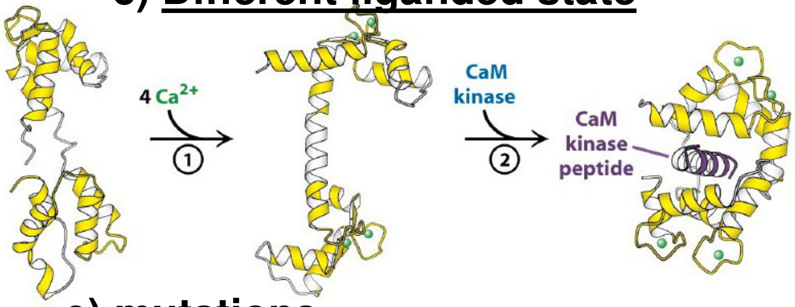


African swine fever virus DNA polymerase X
50mM salt vs 500 mM salt

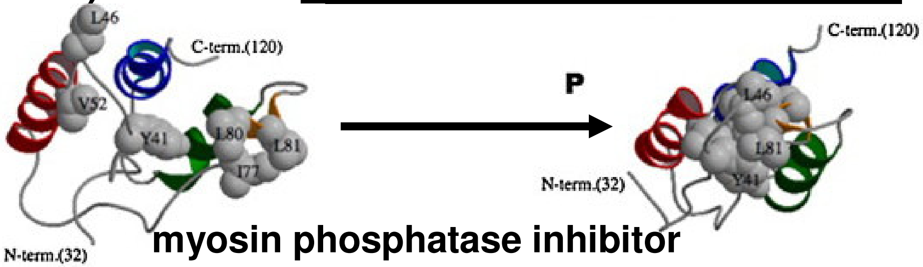


Ad Bax homology-modeled from a George Clooney template

c) Different liganded state

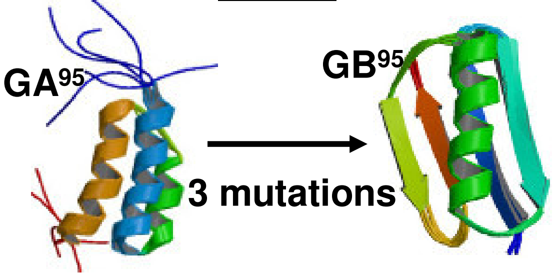


d) Different post-tranlational modification

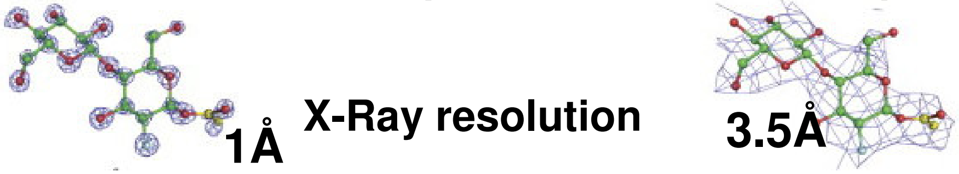


myosin phosphatase inhibitor

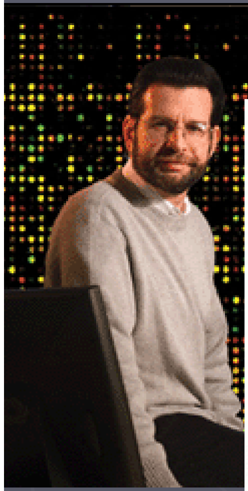
e) mutations



f) Insufficient experimental data of template

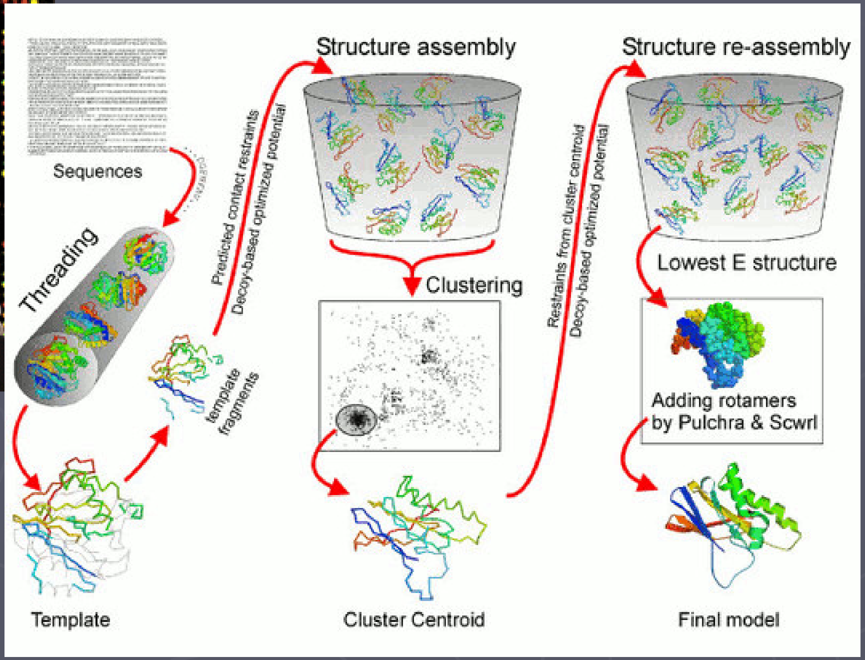


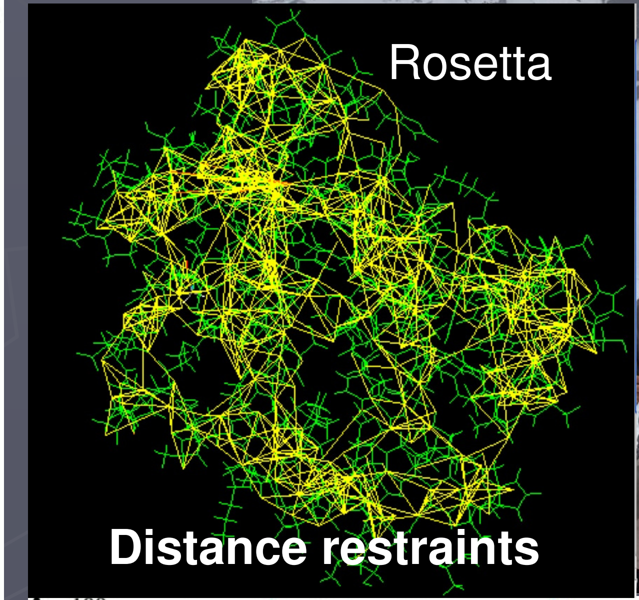
Dealing with 3D bias



Jeffrey Skolnick

TASSER


Sequences
Threading
Template
template fragments
Predicted contact restraints
Decoy-based optimized potential
Structure assembly
Clustering
Cluster Centroid
Restraints from cluster centroid
Decoy-based optimized potential
Structure re-assembly
Lowest E structure
Adding rotamers by Pulchra & Scwrl
Final model



Distance restraints

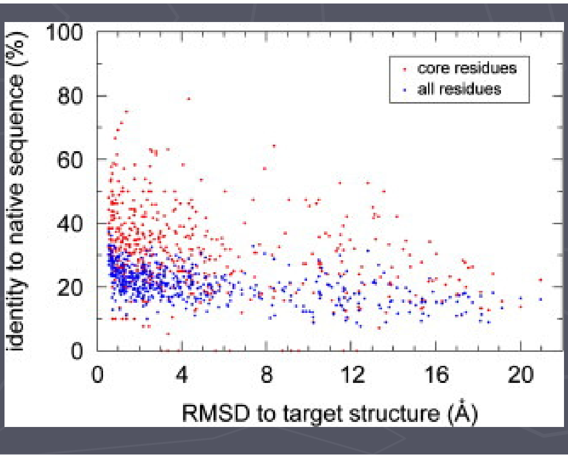
David Baker

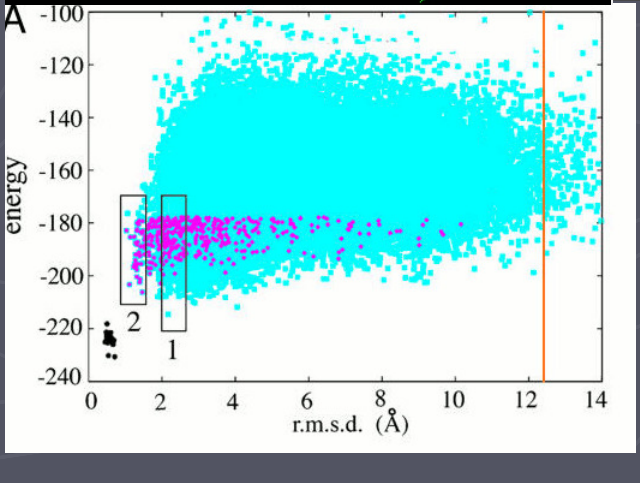
Pros:

-Models are less biased by template

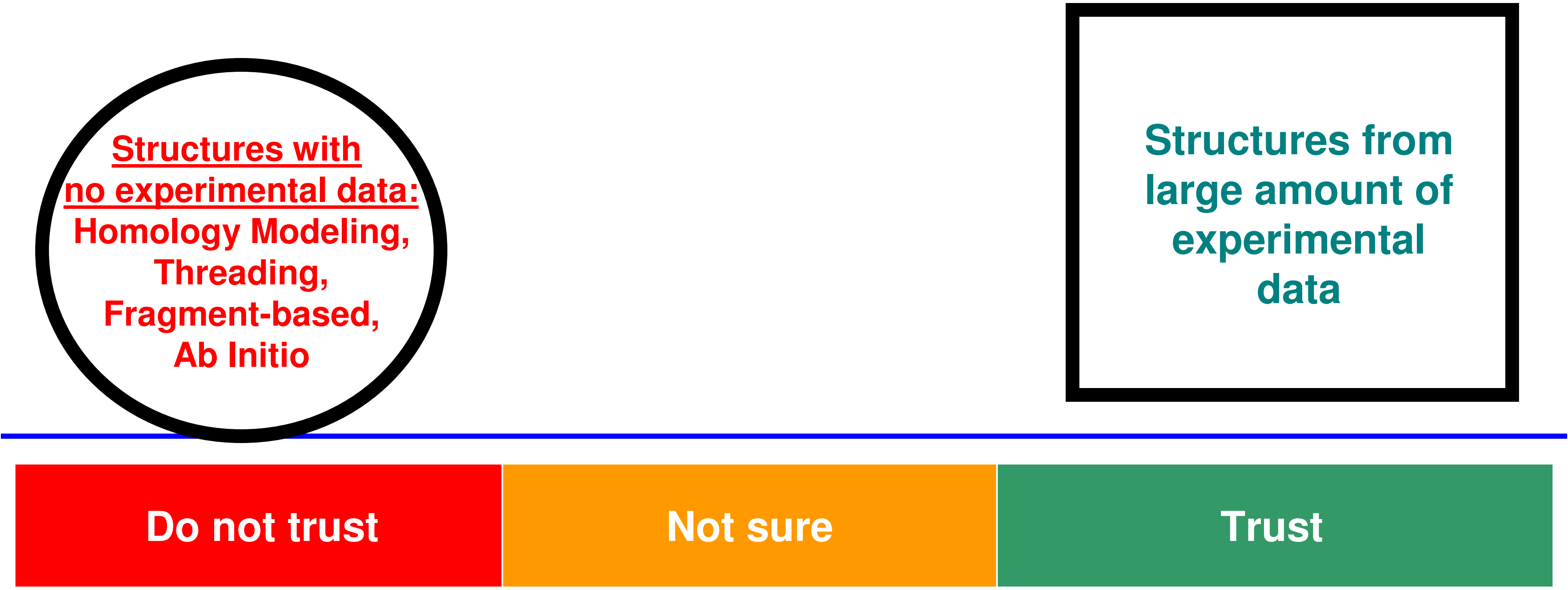
Cons:

- Models are more dependent on imperfect folding scores





Protein structures from the point of view of an experimentalist



Structures with no experimental data:
Homology Modeling,
Threading,
Fragment-based,
Ab Initio

Structures from large amount of experimental data

Do not trust

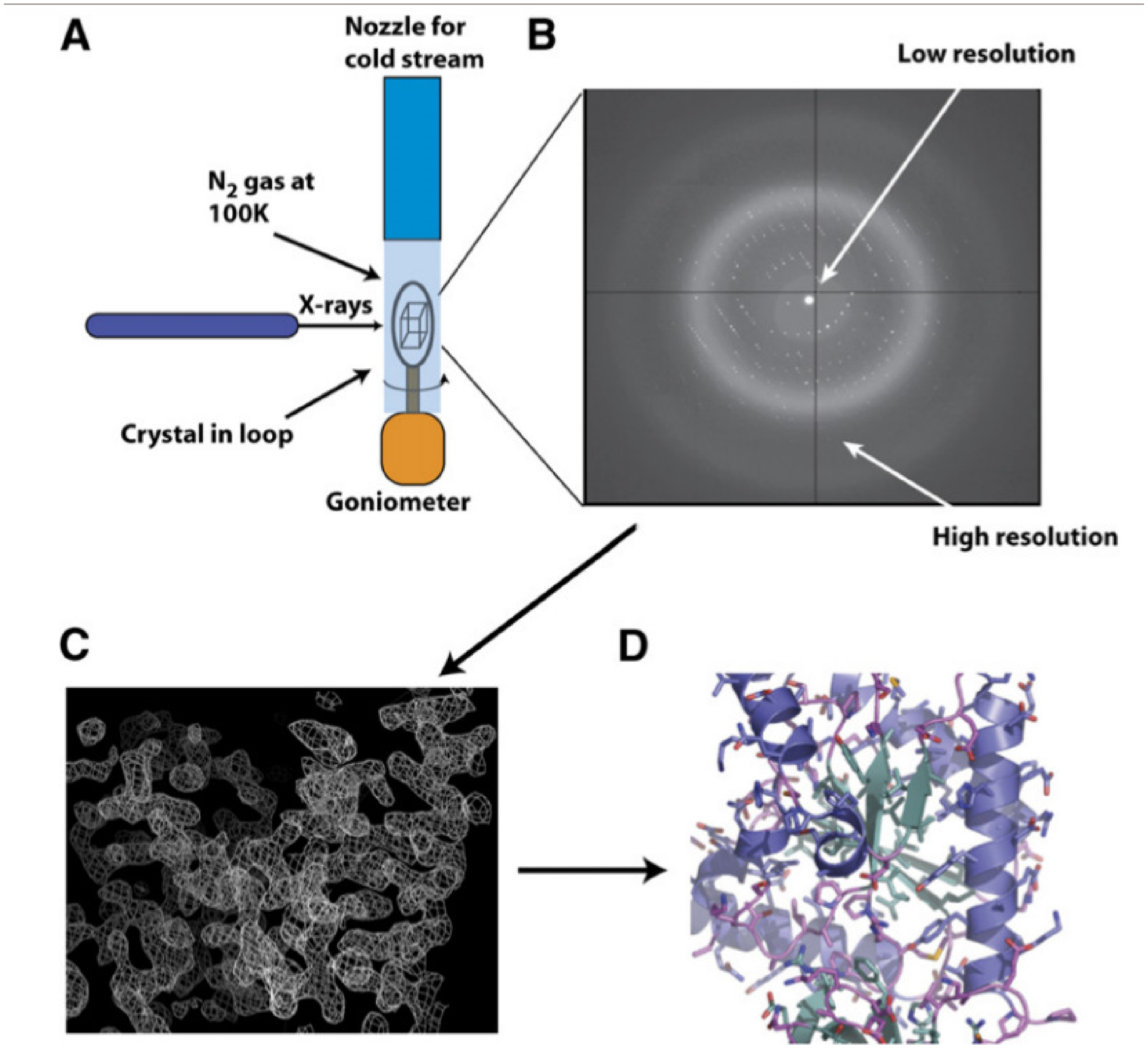
Not sure

Trust

Experimental methods of high-resolution structure determination

X-ray crystallography

X-Ray crystallography



Quality metrics:

1) Experimental data:
-Number of reflections
-Signal to noise ratio

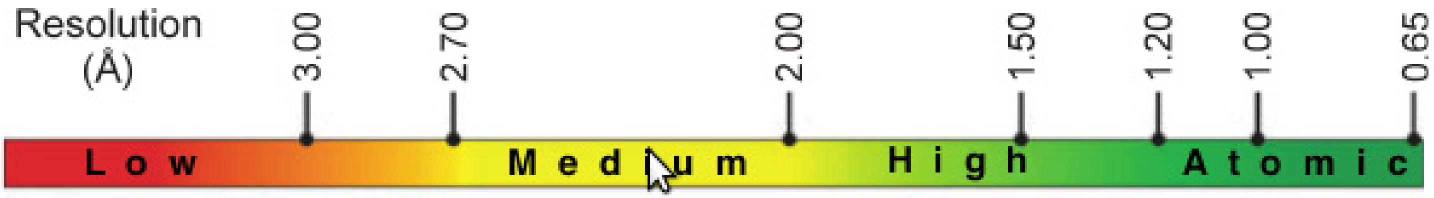
2) Model-to-experiment agreement:
- R factor
-R free factor

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum |F_{obs}|}$$

3) Coordinate uncertainty:
- B-factor

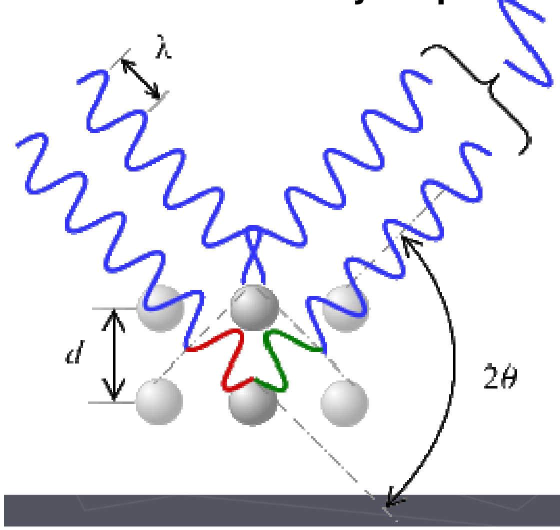
4) Stereo-chemical normality:
- backbone torsion angles (Ramachandran plot)
- bond length, angles
- side-chain torsion angles

X-RayResolution

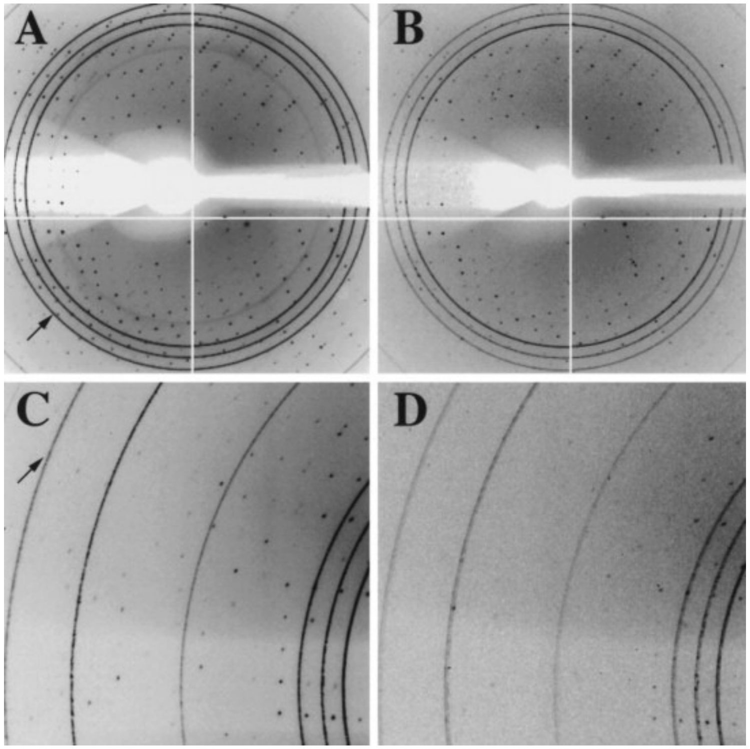


Minimum spacing (d) of crystal lattice planes that still provide measurable diffraction of X-rays.

Minimum distance between structural features that can be distinguished in the electron-density maps.

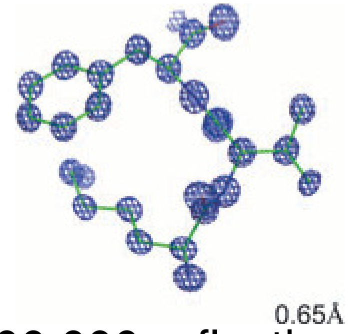


High resolution **Low resolution**



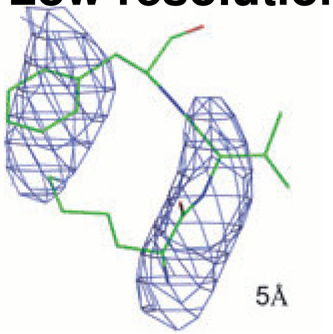
Many reflections Few reflections

High resolution



200,000 reflections

Low resolution



500 reflections

Resolution and protein quality

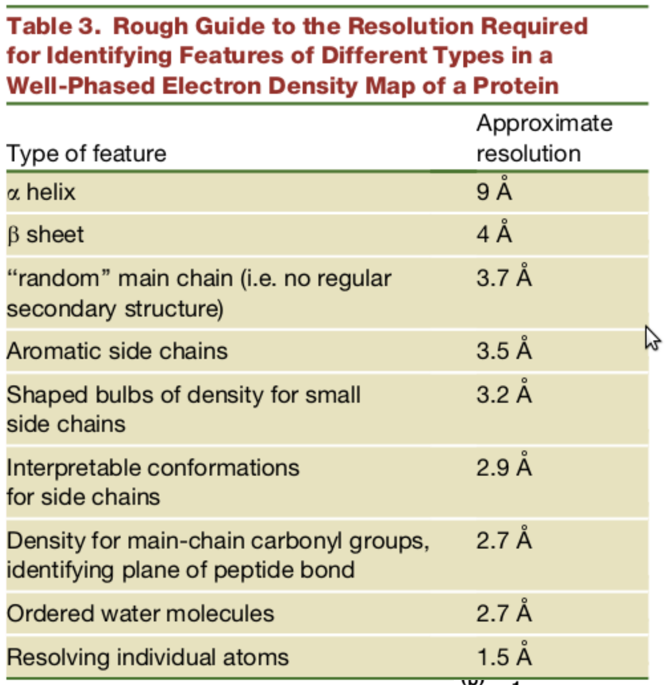
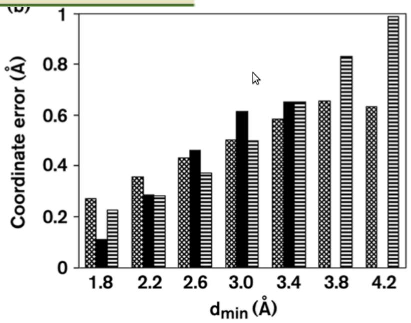


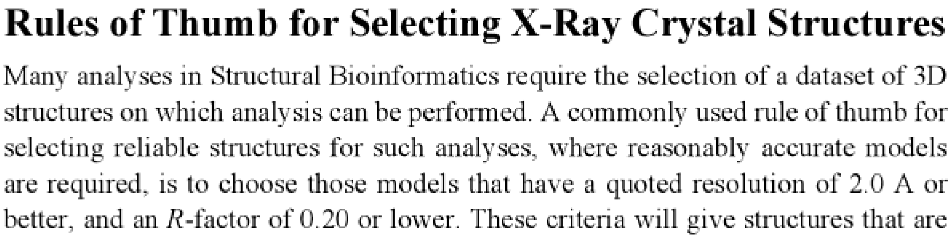
Table 3. Rough Guide to the Resolution Required for Identifying Features of Different Types in a Well-Phased Electron Density Map of a Protein

Type of feature	Approximate resolution
α helix	9 Å
β sheet	4 Å
“random” main chain (i.e. no regular secondary structure)	3.7 Å
Aromatic side chains	3.5 Å
Shaped bulbs of density for small side chains	3.2 Å
Interpretable conformations for side chains	2.9 Å
Density for main-chain carbonyl groups, identifying plane of peptide bond	2.7 Å
Ordered water molecules	2.7 Å
Resolving individual atoms	1.5 Å

Table is taken from Blow (2002).

Blow, D. (2002). Outline of Crystallographyfor Biologists (New York: Oxford University Press).

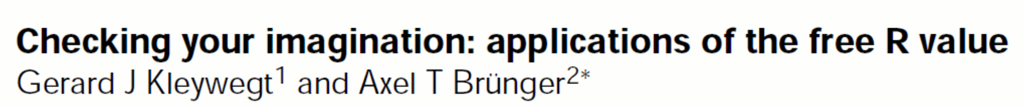




Rules of Thumb for Selecting X-Ray Crystal Structures

Many analyses in Structural Bioinformatics require the selection of a dataset of 3D structures on which analysis can be performed. A commonly used rule of thumb for selecting reliable structures for such analyses, where reasonably accurate models are required, is to choose those models that have a quoted resolution of 2.0 A or better, and an R-factor of 0.20 or lower. These criteria will give structures that are

When X-ray data is incomplete, you have to rely on other sources of structural information: knowledge-based parameters. Imagination, etc.

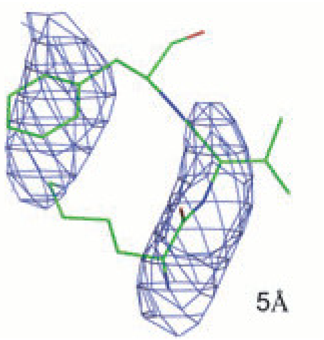


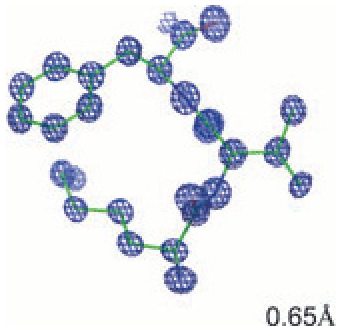
Checking your imagination: applications of the free R value

Gerard J Kleywegt[1] and Axel T Brünger[2*]

Low resolution

High resolution



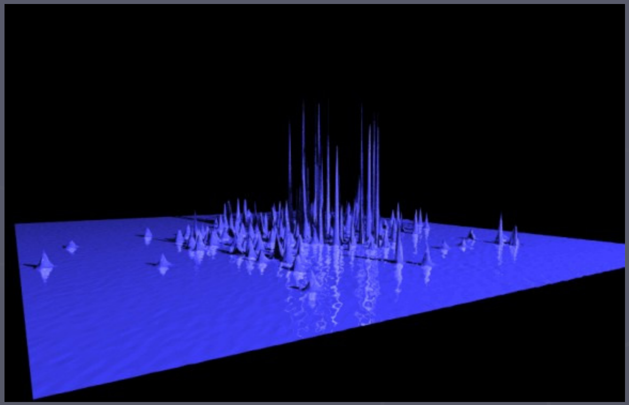


NMR spectroscopy

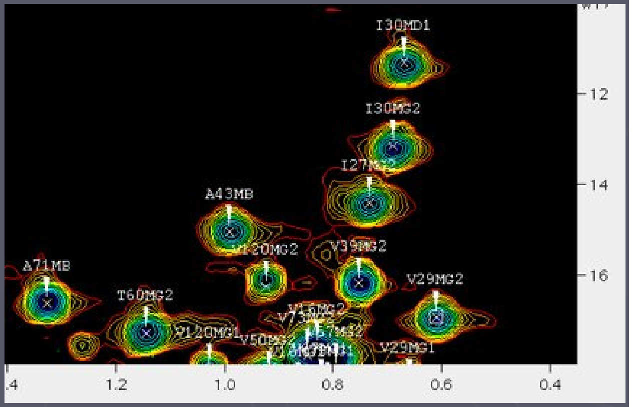
Protein NMR spectroscopy

Experiment

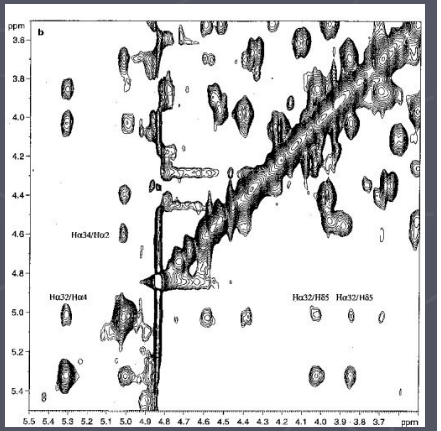
Spectra processing



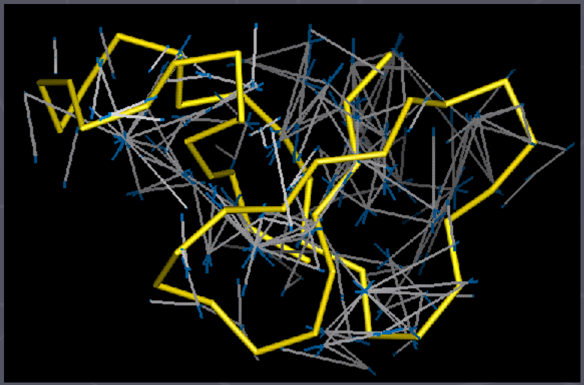
Spectra assignment



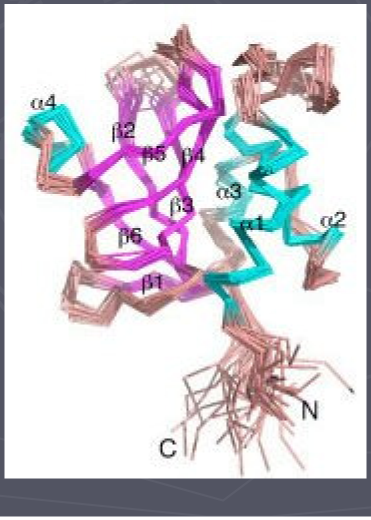
NOE assignment



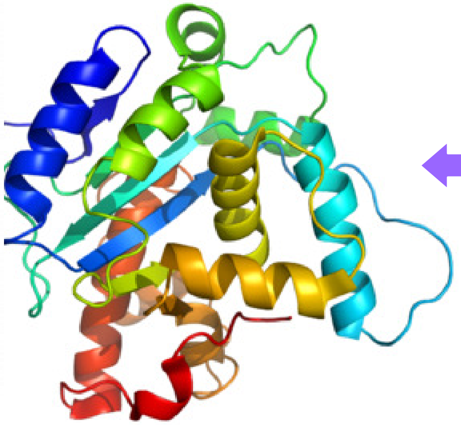
Distance restraints



Model generation



What is typical NMR experimental data?



NMR restraints in the structure calculation	
Intraresidue	333
Sequential (|i - j| = 1)	447
Medium-range (|i - j| <5)	252
Long-range (|i - j| >/ = 5)	369
Hydrogen bonds	66
Total distance restraints	1580
Dihedral angle restraints	113
Residual violations	
CYANA target functions, Å	1.43±0.24
NOE upper distance constrain violation	
Maximum, Å	0.20±0.04
Number >0.2 Å	0±1
Dihedral angle constrain violations	
Maximum, °	3.23±0.72
Number >5°	0±0
Vander Waals violations	
Maximum, Å	0.30±0.00
Number >0.2 Å	3±1
Average structural rmsd to the mean coordinates, Å	
Secondary structure backbone[a]	0.31
Secondary structure heavy atoms[a]	0.80
All backbone atoms[b]	1.30
All heavy atoms[b]	1.79
Ramachandran statistics, %of all residues	
Most favored regions	81.5
Additional allowed regions	18.5
Generously allowed regions	0
Disallowed regions	0

Groups of protein quality parameters:

1) Quality of experimental observables that were used in structure determination*

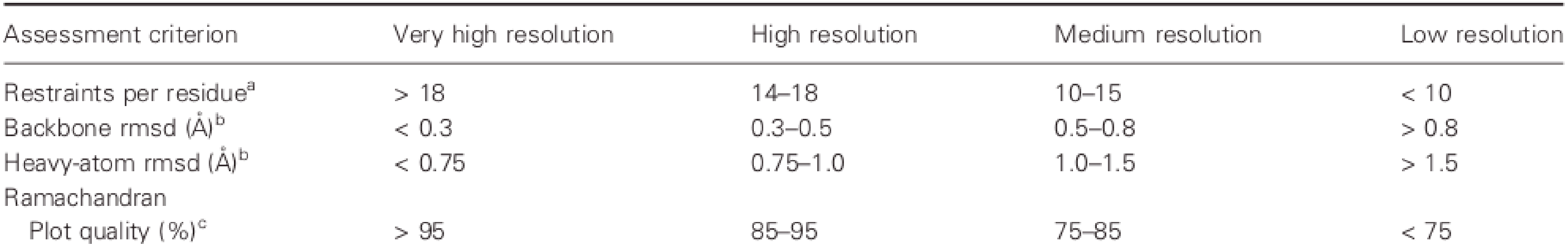
2) Agreement between the structure and experimental observables*

3) Agreement between local geometry of the new structure and parameters of existing high-quality structures

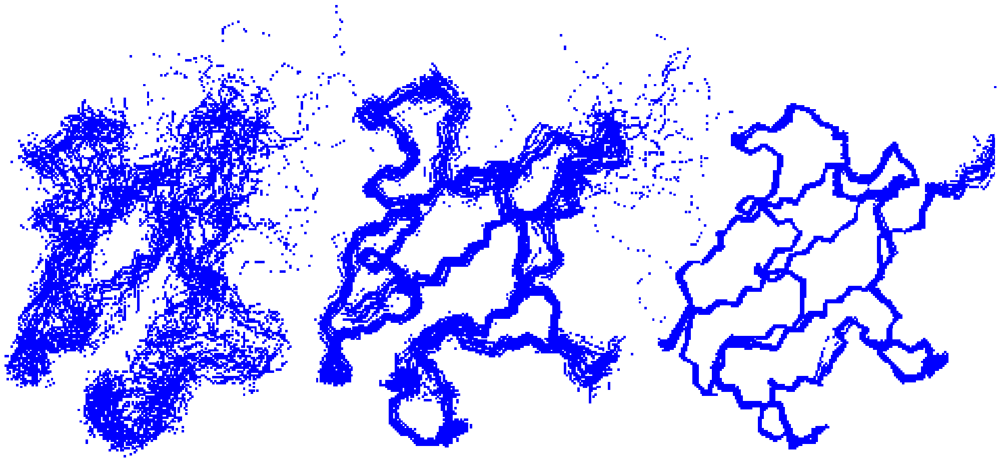
4) Structural uncertainty*

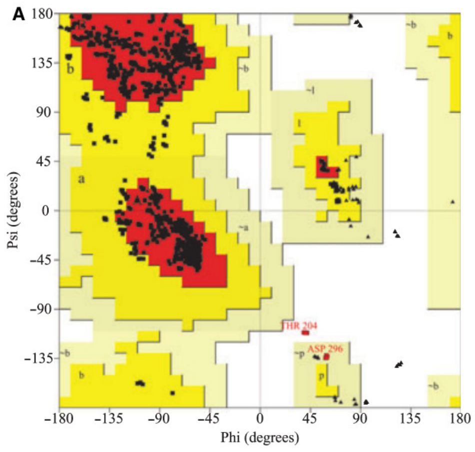
*** Method-dependent**

What is non-sparse NMR data?



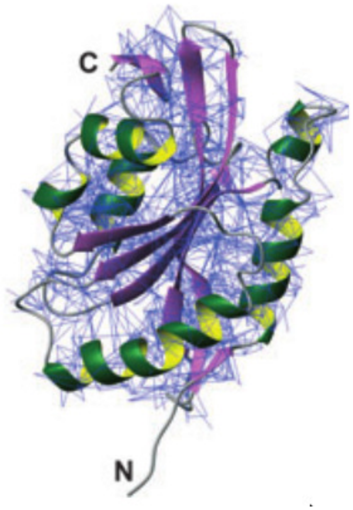
Assessment criterion	Very high resolution	High resolution	Medium resolution	Low resolution
Restraints per residue[a]	> 18	14–18	10–15	< 10
Backbone rmsd (Å)[b]	< 0.3	0.3–0.5	0.5–0.8	> 0.8
Heavy-atom rmsd (Å)[b]	< 0.75	0.75–1.0	1.0–1.5	> 1.5
Ramachandran Plot quality (%)[c]	> 95	85–95	75–85	< 75





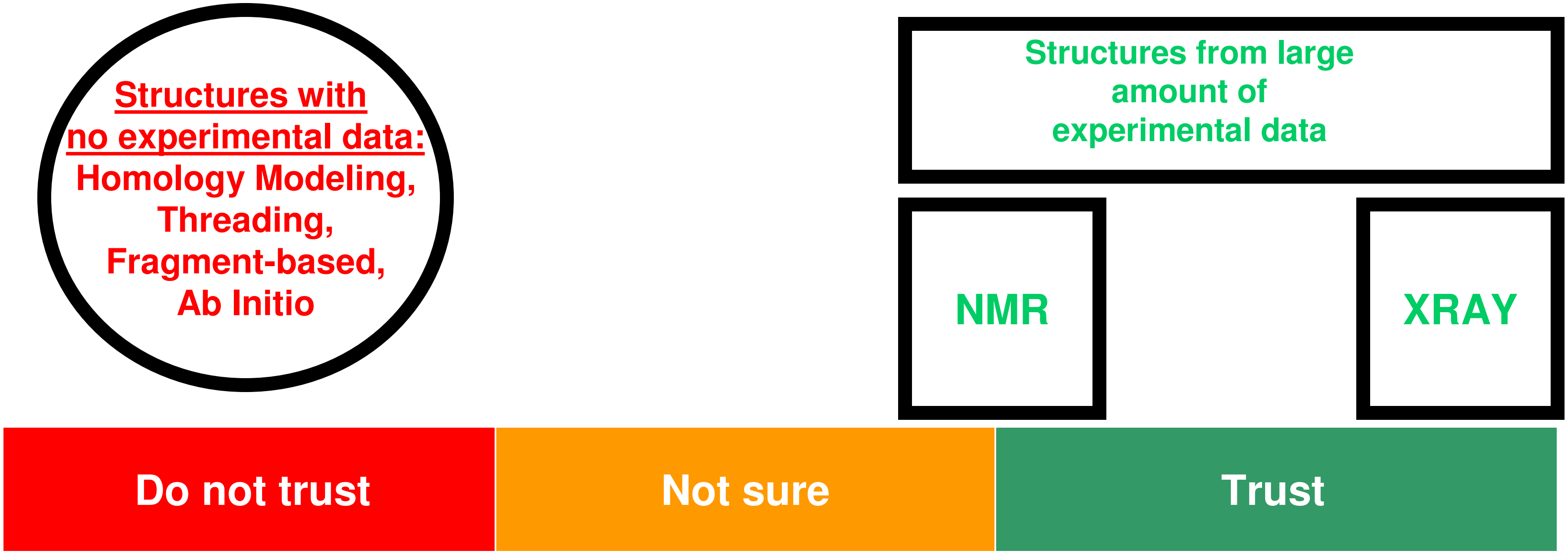
Plot statistics

Residues in most favoured regions [A,B,L]	1013	90.8%
Residues in additional allowed regions [a,b,l,p]	95	8.5%
Residues in generously allowed regions [~a,~b,~l,~p]	4	0.4%
Residues in disallowed regions	4	0.4%



Macromolecular NMR spectroscopy for the non-spectroscopist. Kwan AH, Mobli M, Gooley PR, King GF, Mackay JP. FEBS J. 2011 Mar;278(5):687-703

Protein structures from the point of view of an experimentalist



Structures with no experimental data:
Homology Modeling,
Threading,
Fragment-based,
Ab Initio

Structures from large amount of experimental data

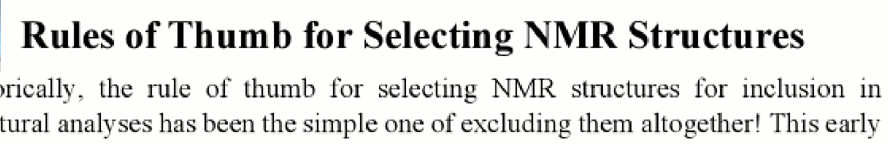
NMR

XRAY

Do not trust

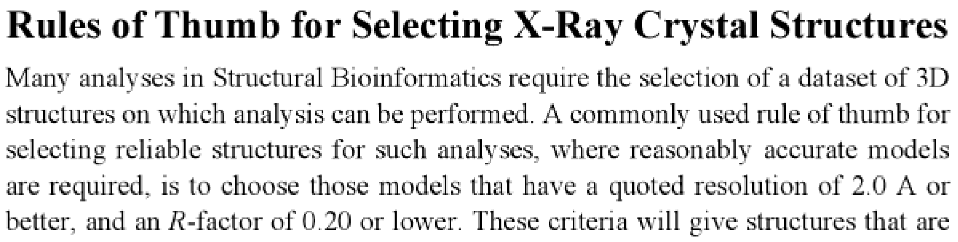
Not sure

Trust



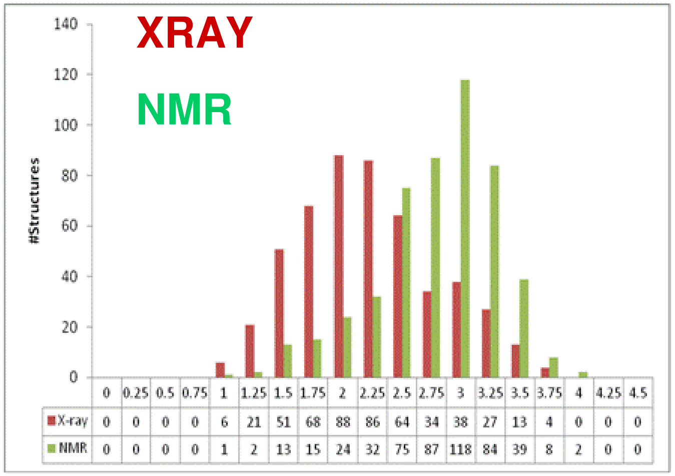
Rules of Thumb for Selecting NMR Structures

Historically, the rule of thumb for selecting NMR structures for inclusion in structural analyses has been the simple one of excluding them altogether! This early



Rules of Thumb for Selecting X-Ray Crystal Structures

Many analyses in Structural Bioinformatics require the selection of a dataset of 3D structures on which analysis can be performed. A commonly used rule of thumb for selecting reliable structures for such analyses, where reasonably accurate models are required, is to choose those models that have a quoted resolution of 2.0 A or better, and an *R*-factor of 0.20 or lower. These criteria will give structures that are

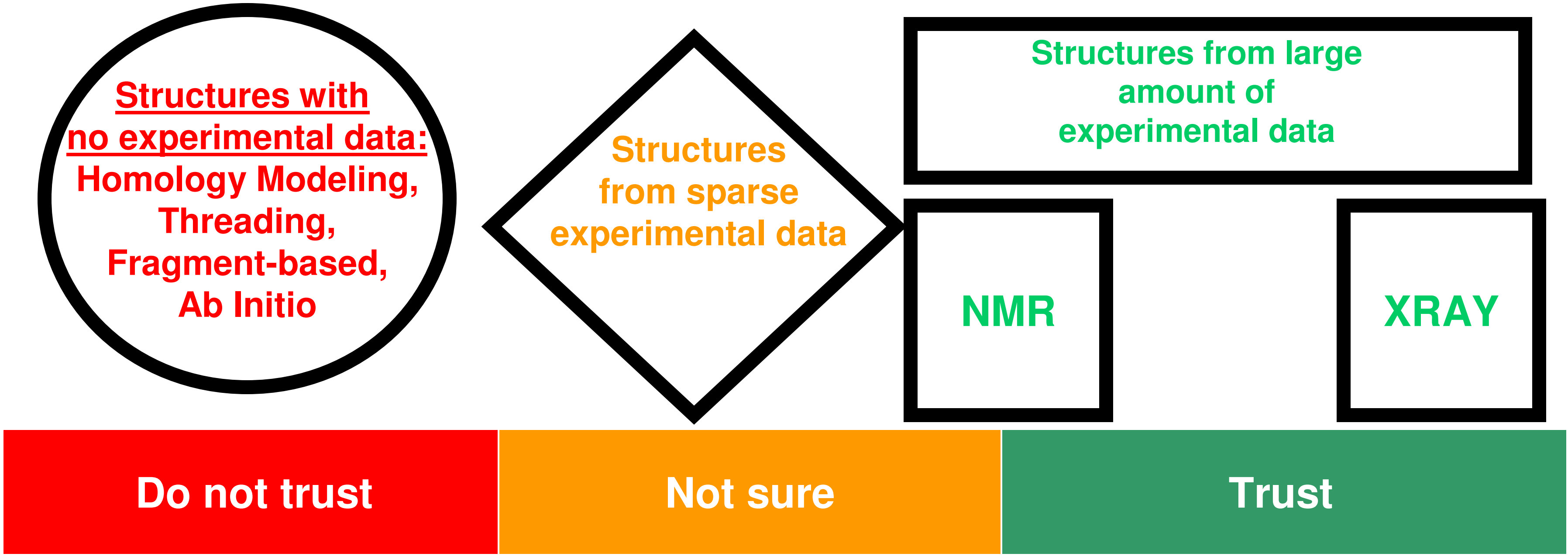


	0	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5	2.75	3	3.25	3.5	3.75	4	4.25	4.5
🟥 X-ray	0	0	0	0	6	21	51	68	88	86	64	34	38	27	13	4	0	0	0
🟩 NMR	0	0	0	0	1	2	13	15	24	32	75	87	118	84	39	8	2	0	0

[Equivalent] Resolution

Sparse experimental data

Protein structures from the point of view of an experimentalist



Structures with no experimental data:
Homology Modeling,
Threading,
Fragment-based,
Ab Initio

Structures from sparse experimental data

Structures from large amount of experimental data

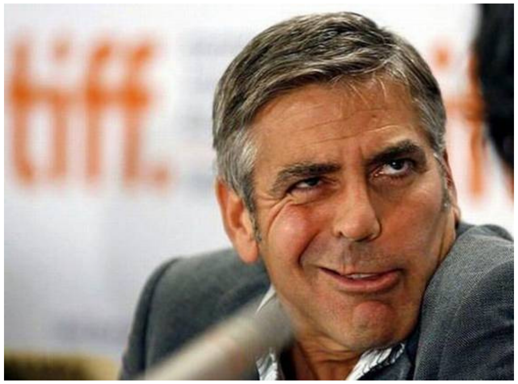
NMR

XRAY

Do not trust

Not sure

Trust

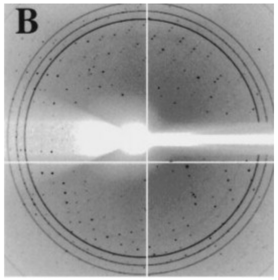


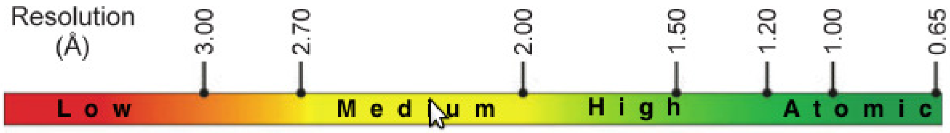
What is sparse experimental data?

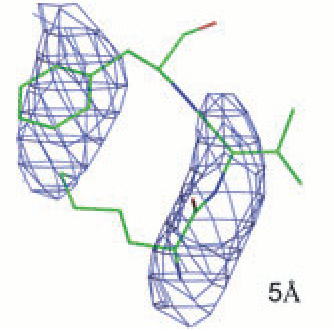
Low resolution

In XRAY:

- When you do not have enough XRAY reflections



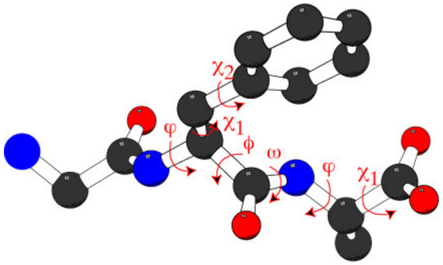


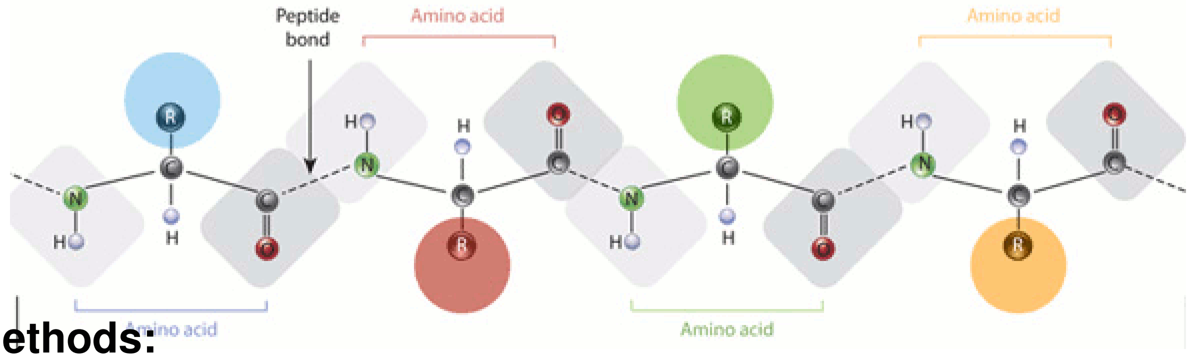


500 reflections

In NMR:

- When you do not have enough NOEs (e.g. no side-chain NOEs) or any NOEs





Sparse data from other methods:

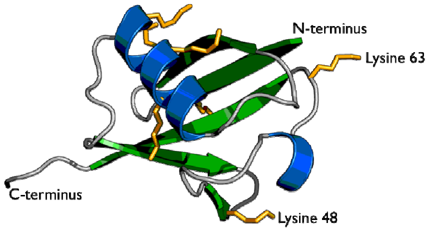
-Distance restraints from cross-linking and mass-spectroscopy

-Distance restraints from spin-labeling and electron paramagnetic resonance (EPR) spectroscopy

- Protein size, shape, radius of gyration from small angle Xray scattering (SAXS)

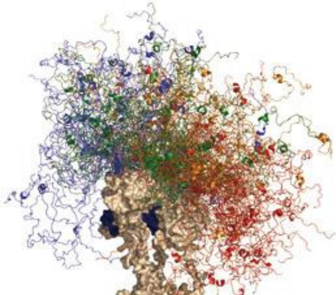
Why use incomplete experimental data for protein structure determination?

1) Some biological questions (e.g. prediction of function from protein fold) may not require high structural accuracy

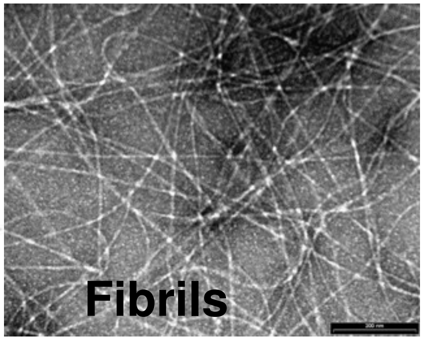


Ubiquitin

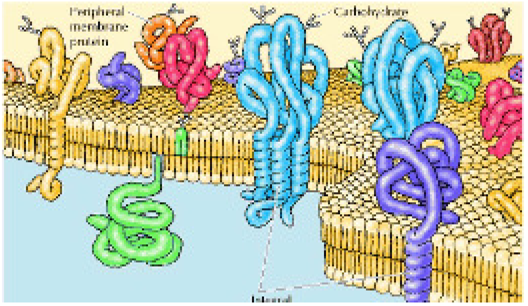
2) Some biologically interesting proteins are too difficult to study by any high-res method:
- proteins with extended flexible regions
-**large proteins**
- fibrillar and membrane proteins



Flexible



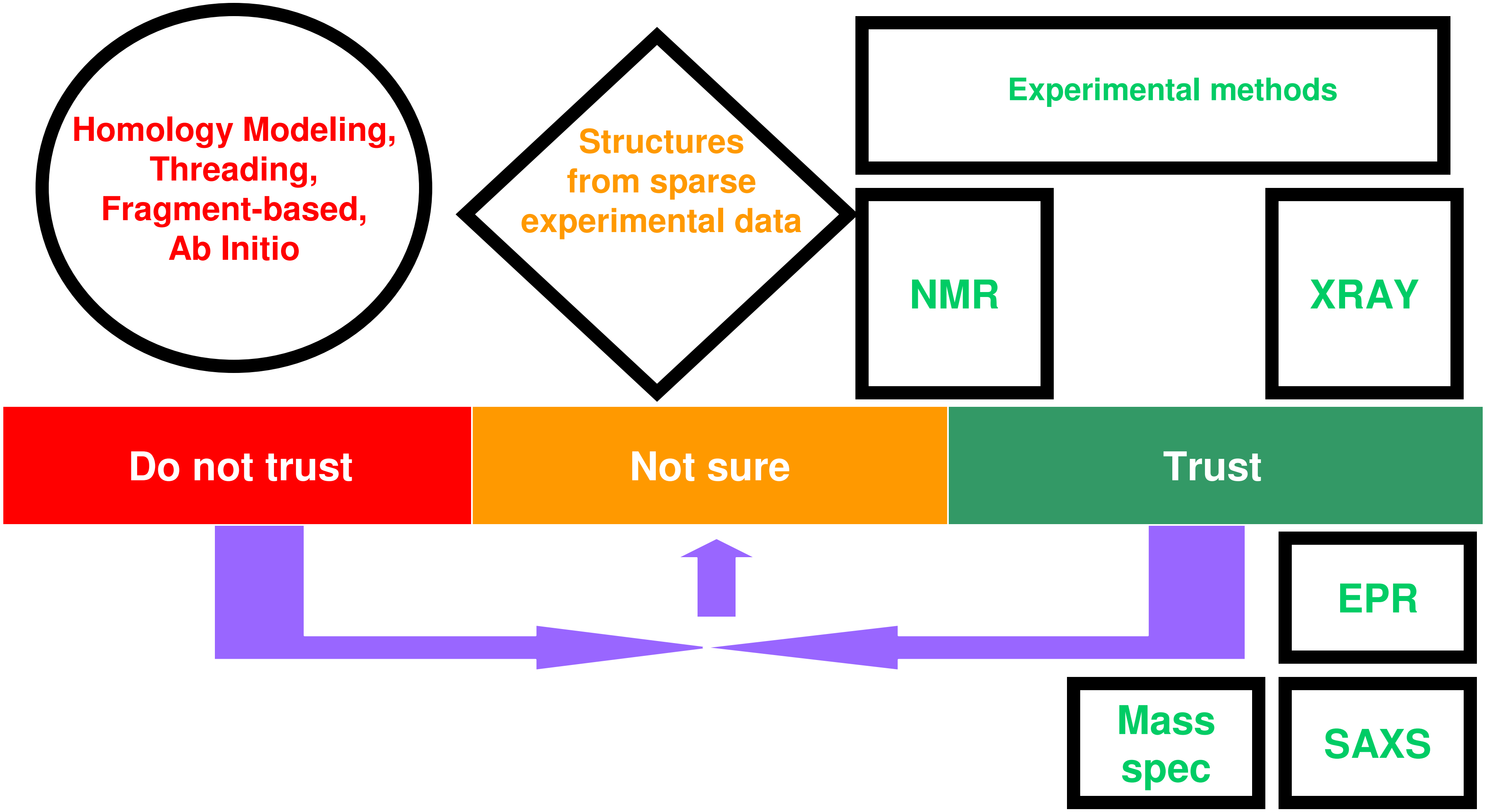
Fibrils



Membrane proteins

Large proteins

Protein structures from the point of view of an experimentalist



Homology Modeling, Threading, Fragment-based, Ab Initio

Structures from sparse experimental data

Experimental methods

NMR

XRAY

Do not trust

Not sure

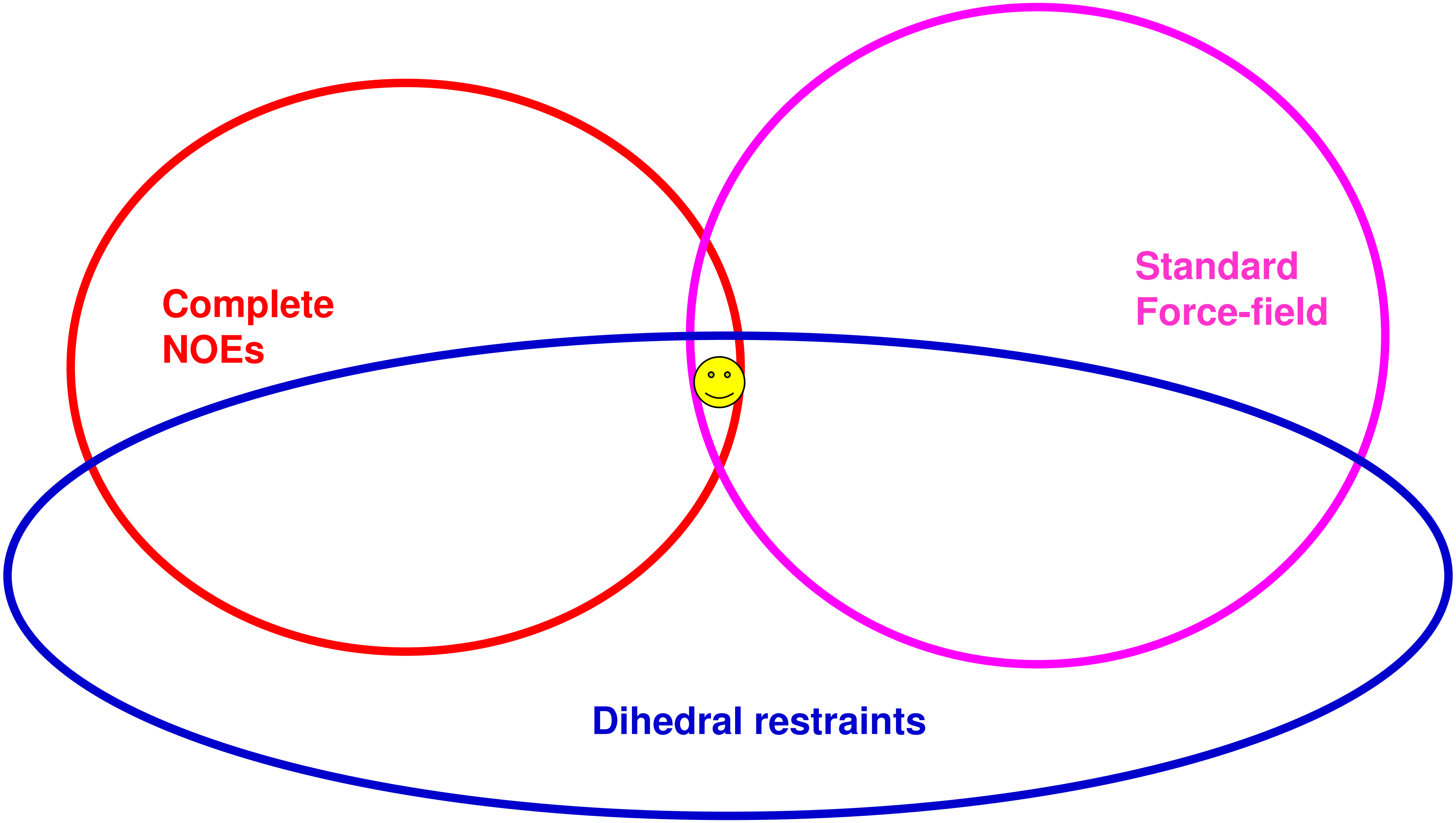
Trust

EPR

Mass spec

SAXS

Regular NMR structure determination

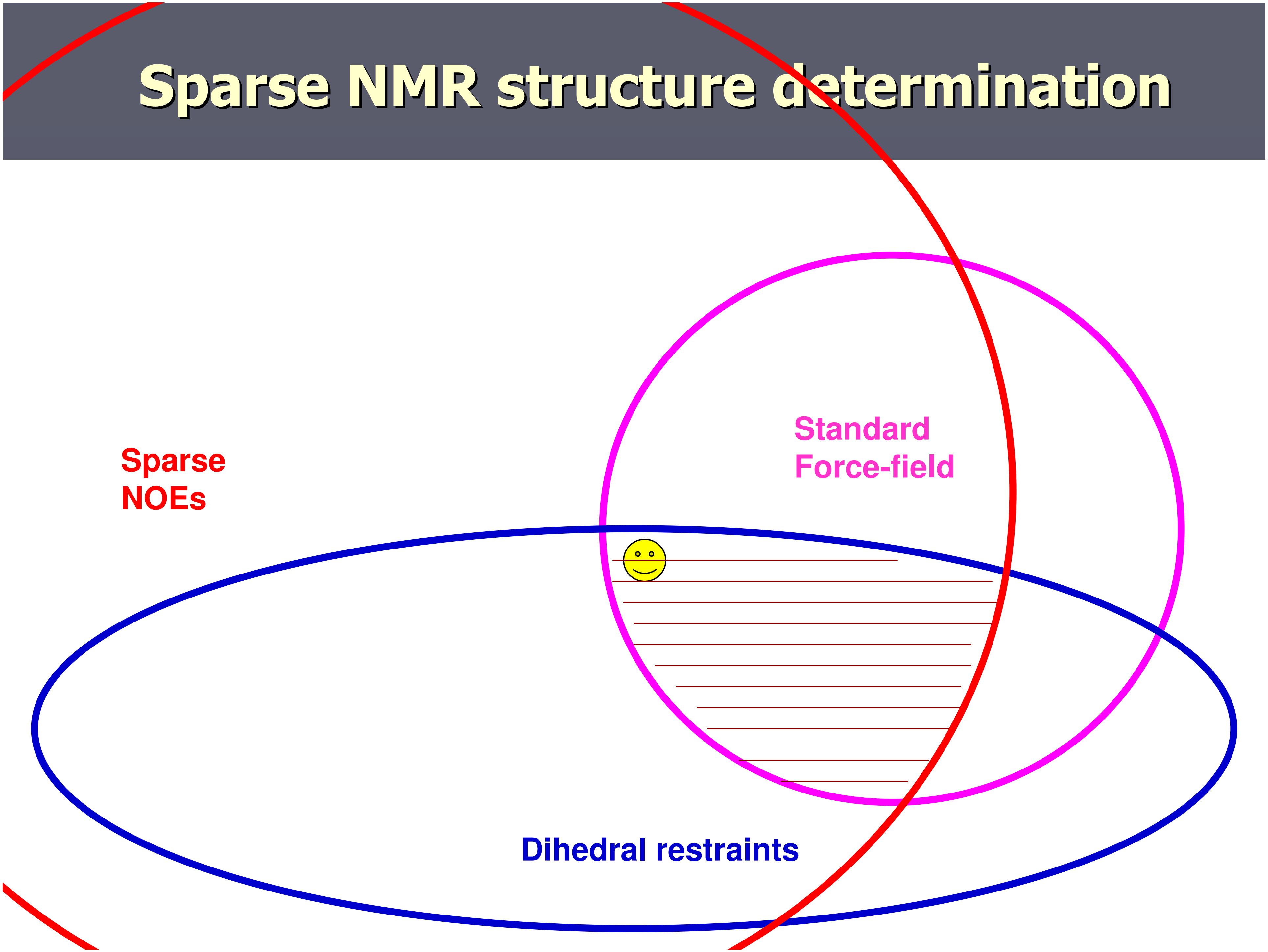


Complete NOEs

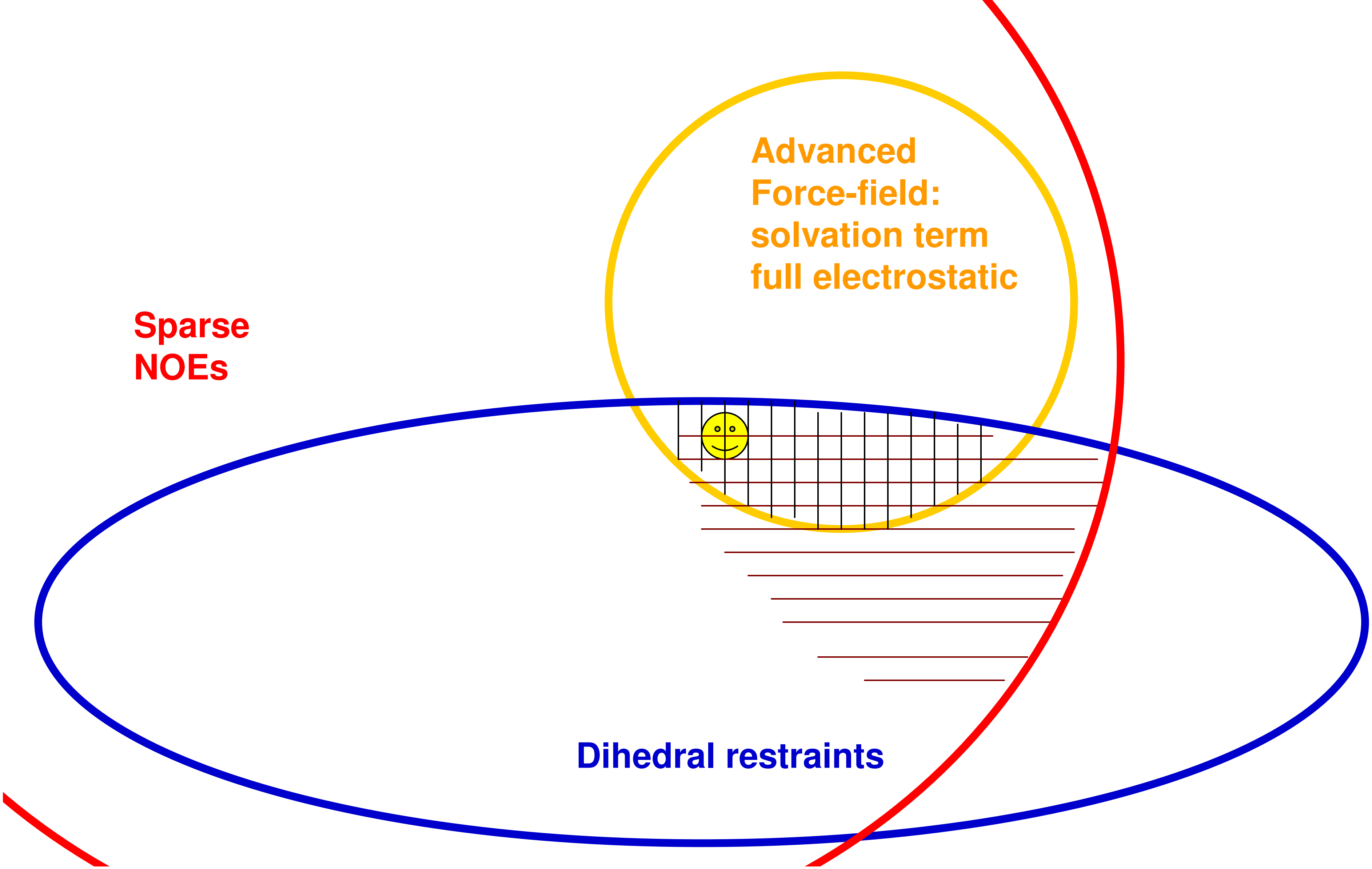
Standard Force-field

Dihedral restraints

Sparse NMR structure determination



Sparse NMR structure determination

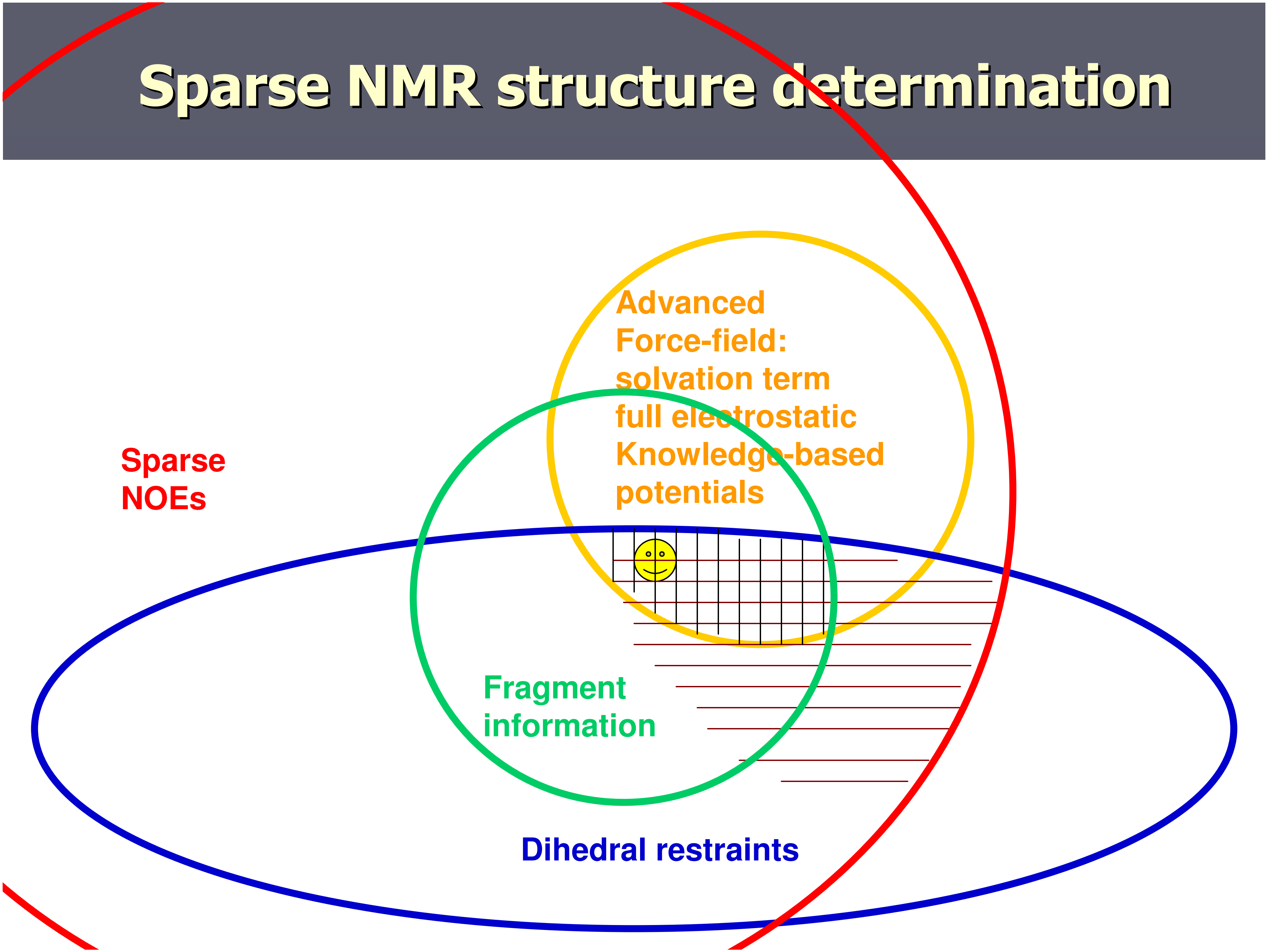


Advanced
Force-field:
solvation term
full electrostatic

Sparse
NOEs

Dihedral restraints

Sparse NMR structure determination



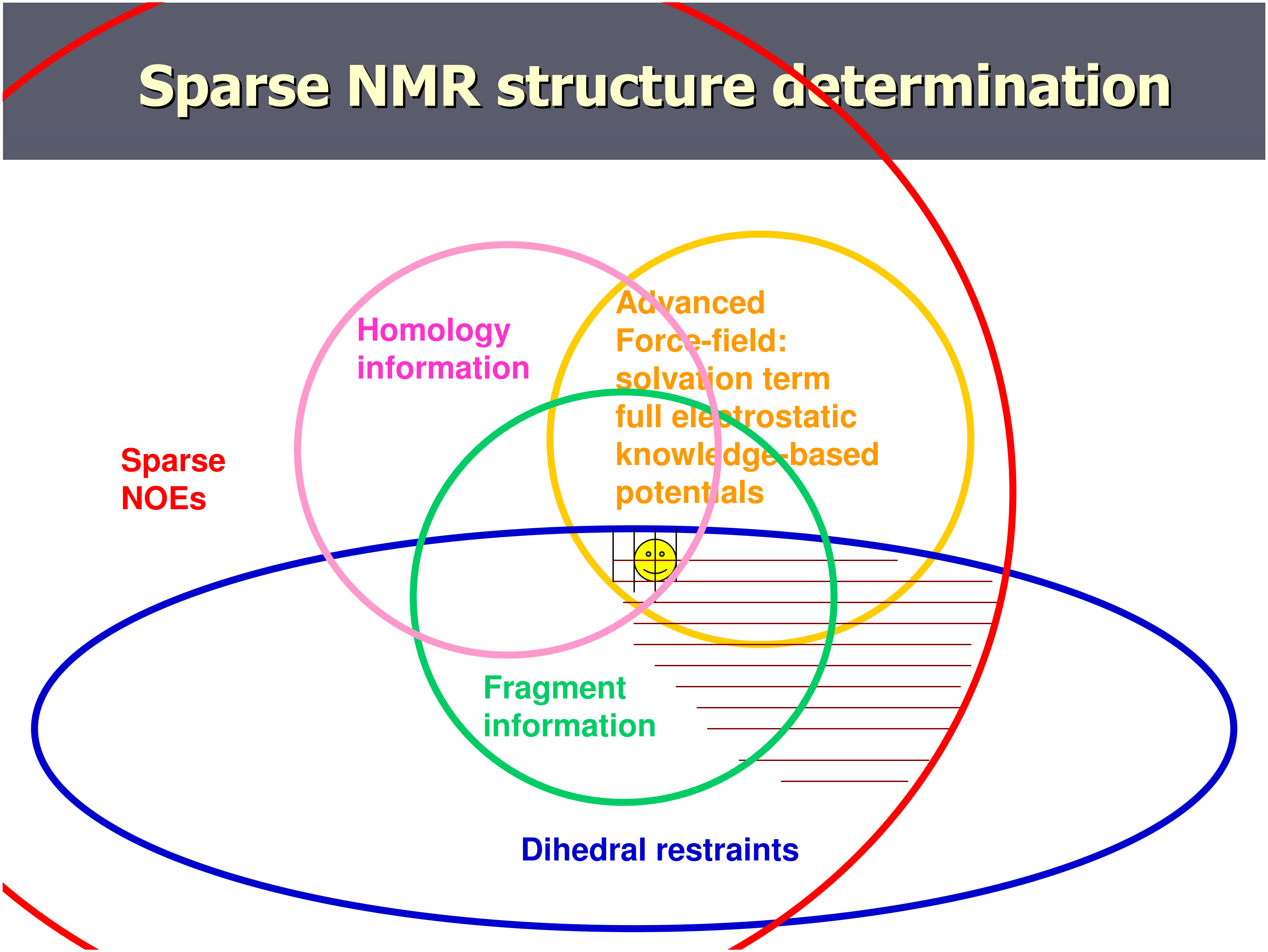
Sparse NOEs

Advanced Force-field:
solvation term
full electrostatic
Knowledge-based potentials

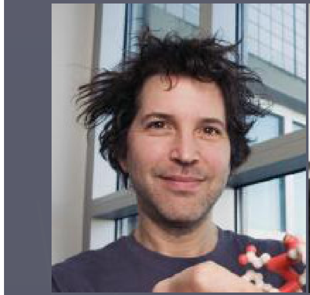
Fragment information

Dihedral restraints

Sparse NMR structure determination



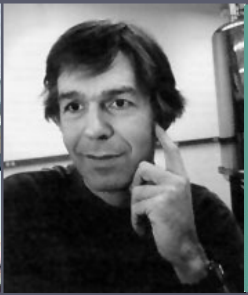
Pushing the boundaries of protein structure determination



David Baker

University of Washington

Rosetta



Ad Bax

NIH

CS-Rosetta

Gaetano Montelione

Rutgers University

CS-DP-Rosetta

Jens Meiler

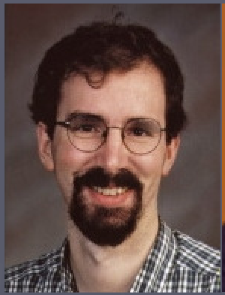
Vanderbilt University

RosettaEPR

David Wishart

University of Alberta

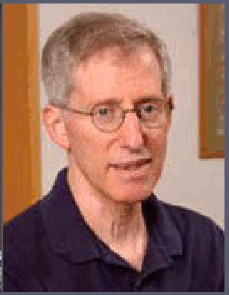
CS23D



Michele Vendruscolo

University of Cambridge

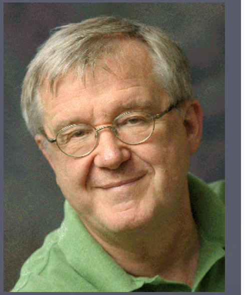
CHESHIRE



George Rose

Johns Hopkins University

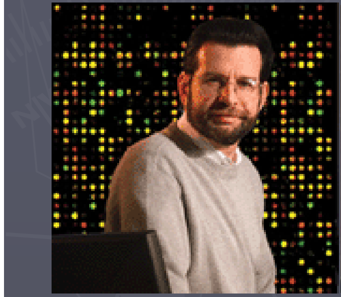
LINUS



Klaus Schulten

University of Illinois

NAMD



Jeffrey Skolnick

Georgia Institute of Technology

TOUCHSTONEX

TASSER

Yang Zhang

University of Michigan

iTASSER

Andrzej Kolinski

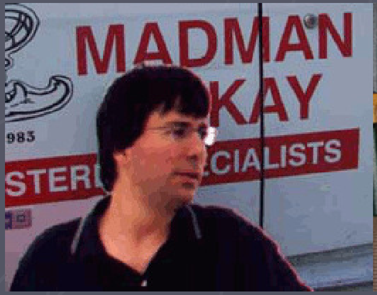
Warsaw University

CABS-NMR

Hans Kalbitzer

Universität Regensburg

PERMOL



Lewis Kay

University of Toronto

CHESHIRE
Rosetta

Julia Forman-Kay

University of Toronto

ENSEMBLE

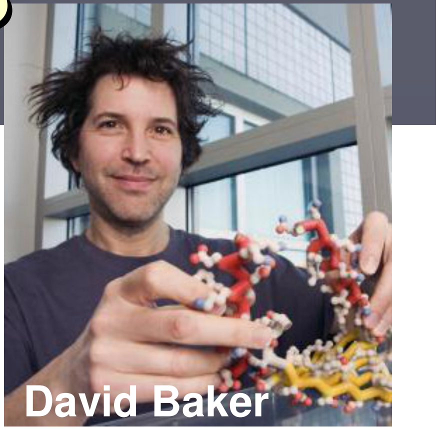
Charles L. Brooks III

University of Michigan

CHARMM

Rosetta-family methods

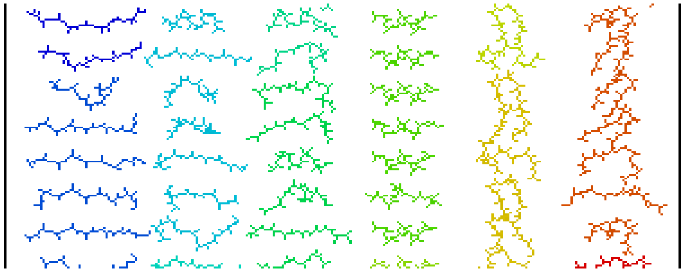
Developed by 12 labs
50 people



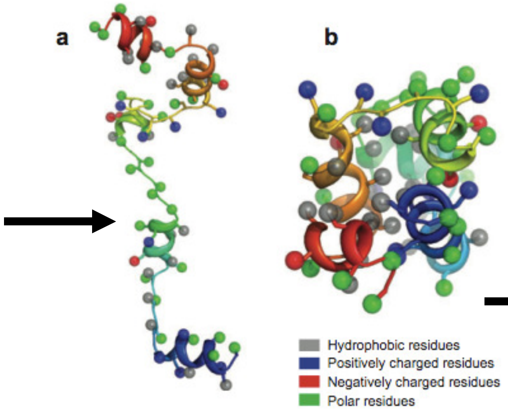
Sequence

Secondary Structure

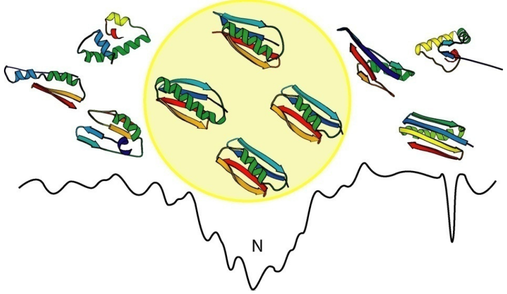
3- and 9-residue fragments



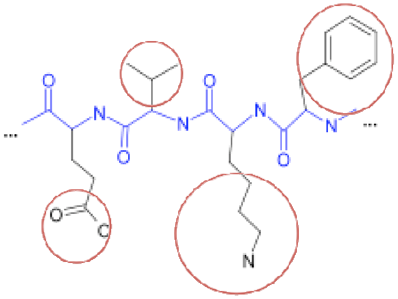
Low-resolution folding



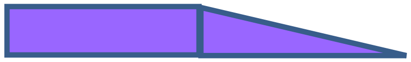
Best low-res decoy selection



Full-atom side-chain restoration

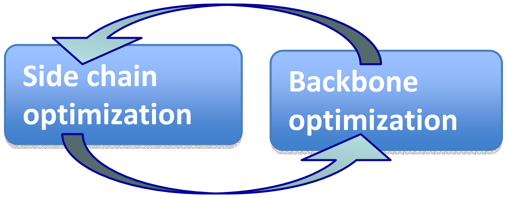


Full-atom refinement



vdW repulsive

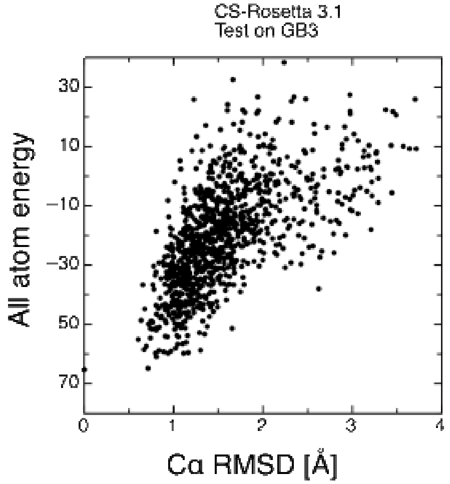
Small backbone moves



Side chain optimization

Backbone optimization

Model quality evaluation



Rosetta-family methods

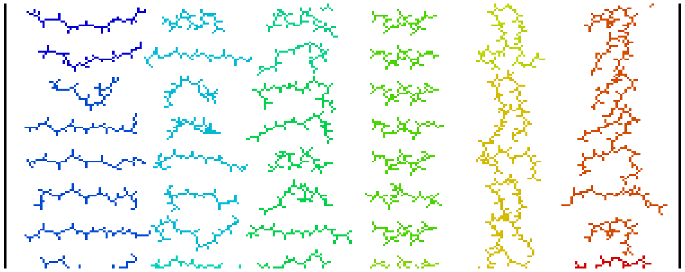
Method	**Year**	**Restraints**
Rosetta	1996-1999	
Rosetta-NMR	2000	NOEs
Rosetta-NMR-RDC	2002	NOEs, RDCs
CS-Rosetta	2008	CS
CS-DP-Rosetta	2010	CS, unassigned NOEs
iterative-CS-RDC-NOE Rosetta	2010	CS, RDC, backbone NOEs
PCS-ROSETTA	2011	Pseudo-contact shifts
Rosetta-EPR	2011	EPR data
CS-HM Rosetta	2012	CS, homology
RASREC Rosetta	2011-2012	CS, methyl NOEs, RDCs

How is sparse data used in Rosetta-family methods?

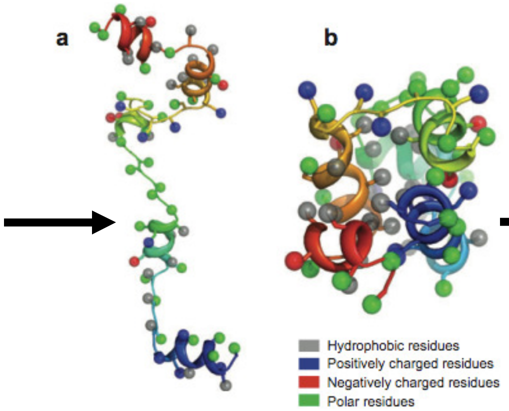
Sequence

Secondary Structure

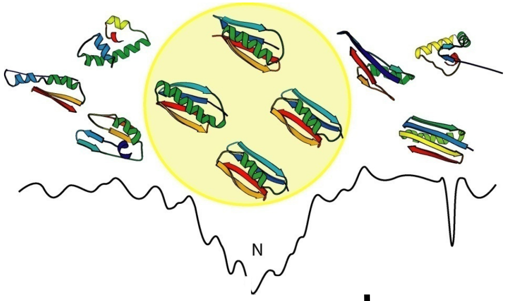
3- and 9-residue fragments



Low-resolution folding



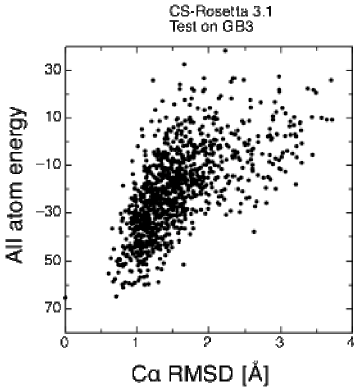
Best low-res decoy selection



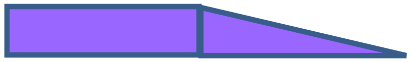
Chemical shifts

NOEs, RDCs, PCSs, Homology, etc

Model quality evaluation

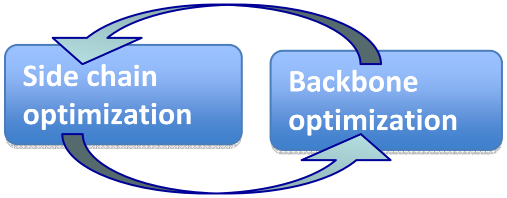


Full-atom refinement



vdW repulsive

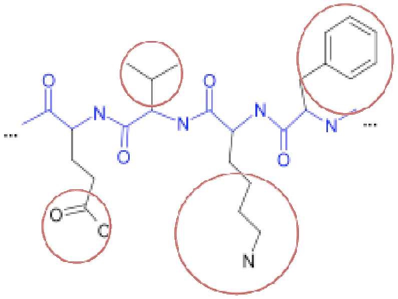
Small backbone moves



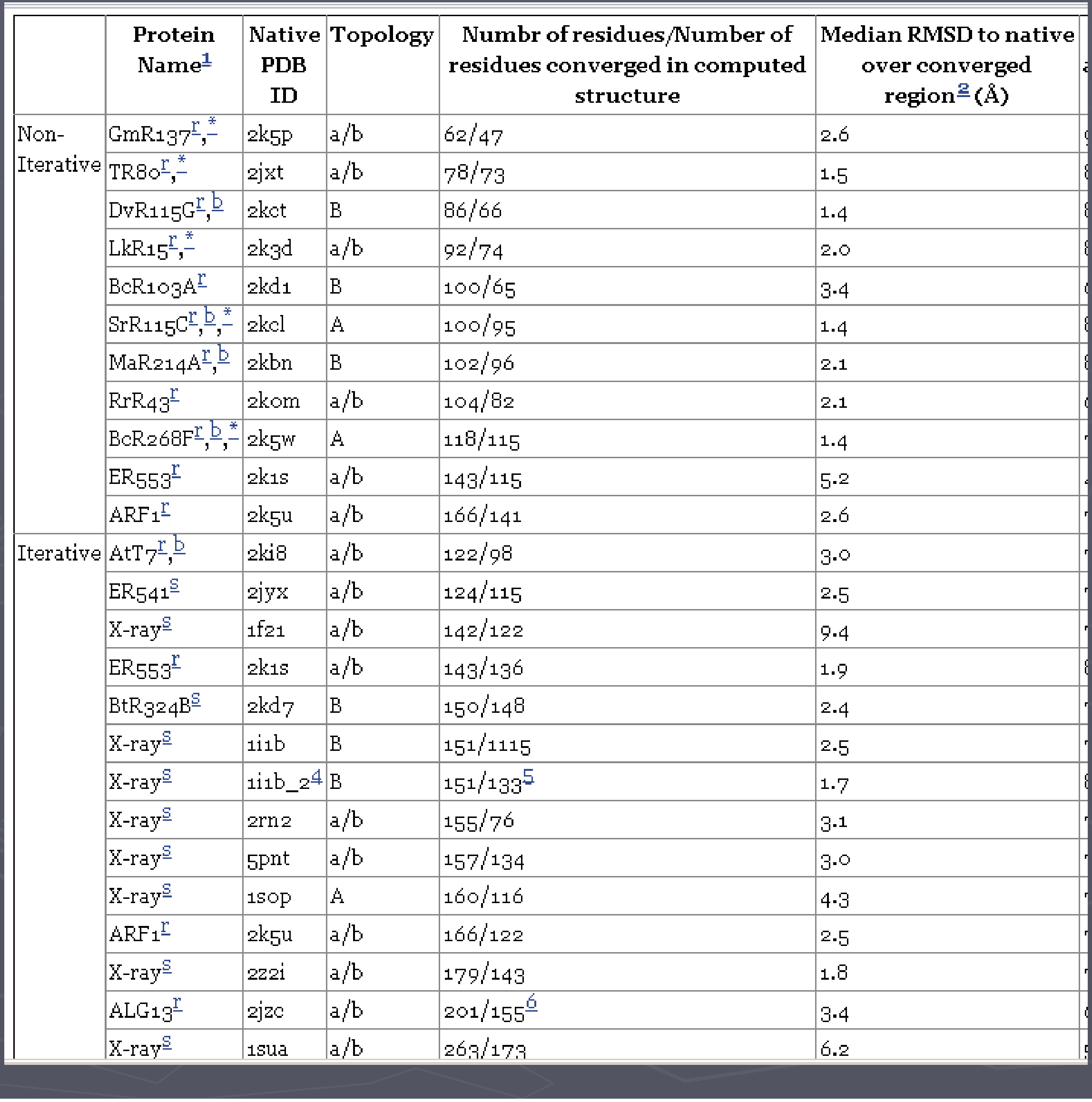
Side chain optimization

Backbone optimization

Full-atom side-chain restoration



Performance of iterative Rosetta for backbone-only NMR data

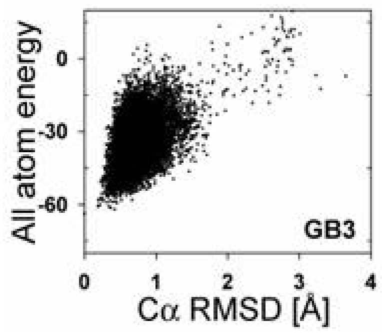


	Protein Name[1]	Native PDB ID	Topology	Numbr of residues/Number of residues converged in computed structure	Median RMSD to native over converged region[2] (Å)
Non-Iterative	GmR137[r,*]	2k5p	a/b	62/47	2.6
	TR80[r,*]	2jxt	a/b	78/73	1.5
	DvR115G[r,b]	2kct	B	86/66	1.4
	LkR15[r,*]	2k3d	a/b	92/74	2.0
	BcR103A[r]	2kd1	B	100/65	3.4
	SrR115C[r,b,*]	2kcl	A	100/95	1.4
	MaR214A[r,b]	2kbn	B	102/96	2.1
	RrR43[r]	2kom	a/b	104/82	2.1
	BcR268F[r,b,*]	2k5w	A	118/115	1.4
	ER553[r]	2k1s	a/b	143/115	5.2
	ARF1[r]	2k5u	a/b	166/141	2.6
Iterative	AtT7[r,b]	2ki8	a/b	122/98	3.0
	ER541[s]	2jyx	a/b	124/115	2.5
	X-ray[s]	1f21	a/b	142/122	9.4
	ER553[r]	2k1s	a/b	143/136	1.9
	BtR324B[s]	2kd7	B	150/148	2.4
	X-ray[s]	1i1b	B	151/1115	2.5
	X-ray[s]	1i1b_2[4]	B	151/133[5]	1.7
	X-ray[s]	2rn2	a/b	155/76	3.1
	X-ray[s]	5pnt	a/b	157/134	3.0
	X-ray[s]	1sop	A	160/116	4.3
	ARF1[r]	2k5u	a/b	166/122	2.5
	X-ray[s]	2z2i	a/b	179/143	1.8
	ALG13[r]	2jzc	a/b	201/155[6]	3.4
	X-ray[s]	1sua	a/b	263/173	6.2

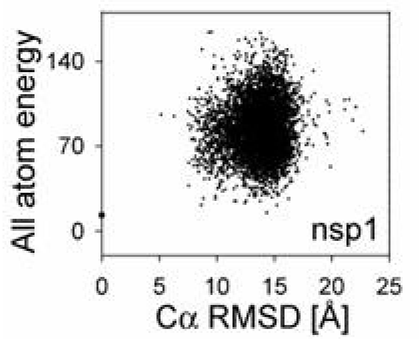
What are the criteria of Rosetta simulation success?

1) RMSD with respect to the lowest/best score model should be within 2Å for more than 60% of models

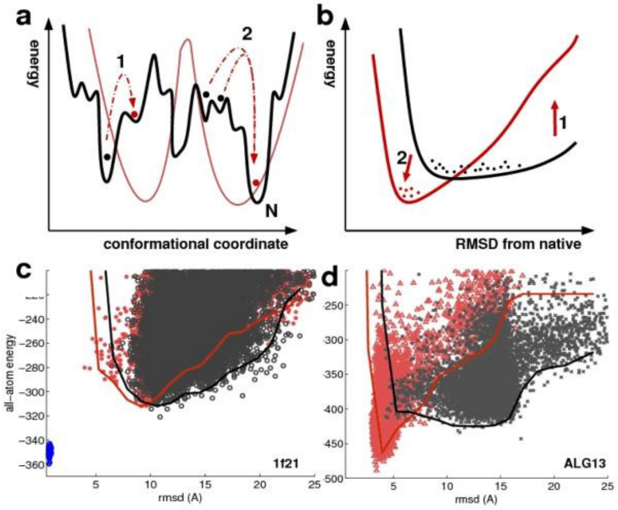
Successful simulation



Failed simulation



Effect of experimental data



2) The converged structures should be clearly lower in energy than all significantly different (RMSD greater than 7 Å

3) The structures generated with experimental data should be at least as low in energy as those generated without experimental data or even lower/better

Problems with validation of Rosetta models.

1) It is not clear if the Rosetta success criteria are universal for all scenarios

2) Agreement with experimental data is not very meaningful because

the data is sparse.

3) Rfree like validation is difficult (and also not meaningful) because experimental data is sparse

4) Independent experimental data for validation will likely be unavailable

5) Normality-based scores for model validation (e.g. ResProx) will likely fail

to detect inaccurate but highly-idealized Rosetta models

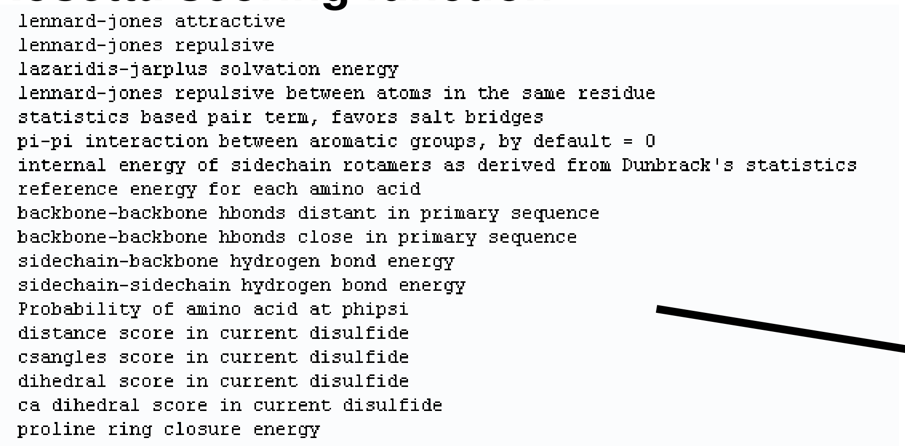
⬇

Need for developing an independent model validation protocol

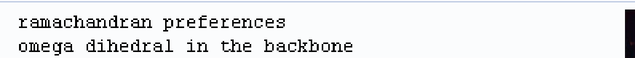
Problem with informational content of protein models from sparse data

The experimental data is over-powered by knowledge based information

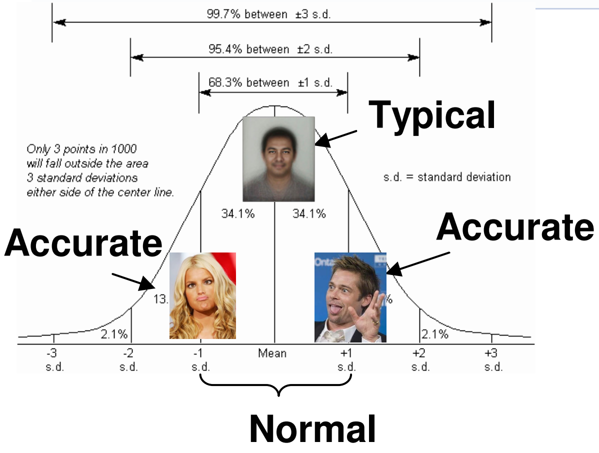
Rosetta scoring function



```
lennard-jones attractive
lennard-jones repulsive
lazaridis-jarplus solvation energy
lennard-jones repulsive between atoms in the same residue
statistics based pair term, favors salt bridges
pi-pi interaction between aromatic groups, by default = 0
internal energy of sidechain rotamers as derived from Dunbrack's statistics
reference energy for each amino acid
backbone-backbone hbonds distant in primary sequence
backbone-backbone hbonds close in primary sequence
sidechain-backbone hydrogen bond energy
sidechain-sidechain hydrogen bond energy
Probability of amino acid at phipsi
distance score in current disulfide
csangles score in current disulfide
dihedral score in current disulfide
ca dihedral score in current disulfide
proline ring closure energy
```



```
ramachandran preferences
omega dihedral in the backbone
```



Fragment idealization

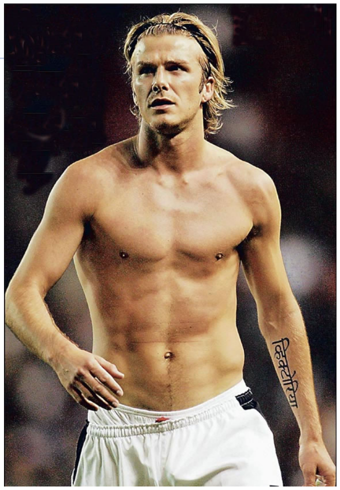
Bias by fragments from other proteins

Biased by restraints from homologous proteins

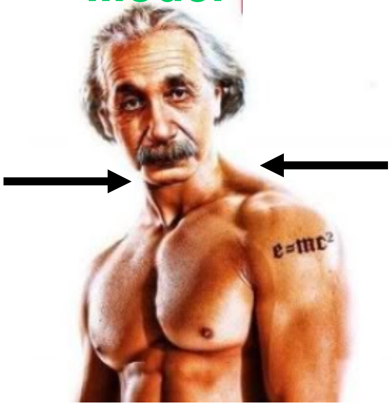
VS

Sparse experimental data

Typical male

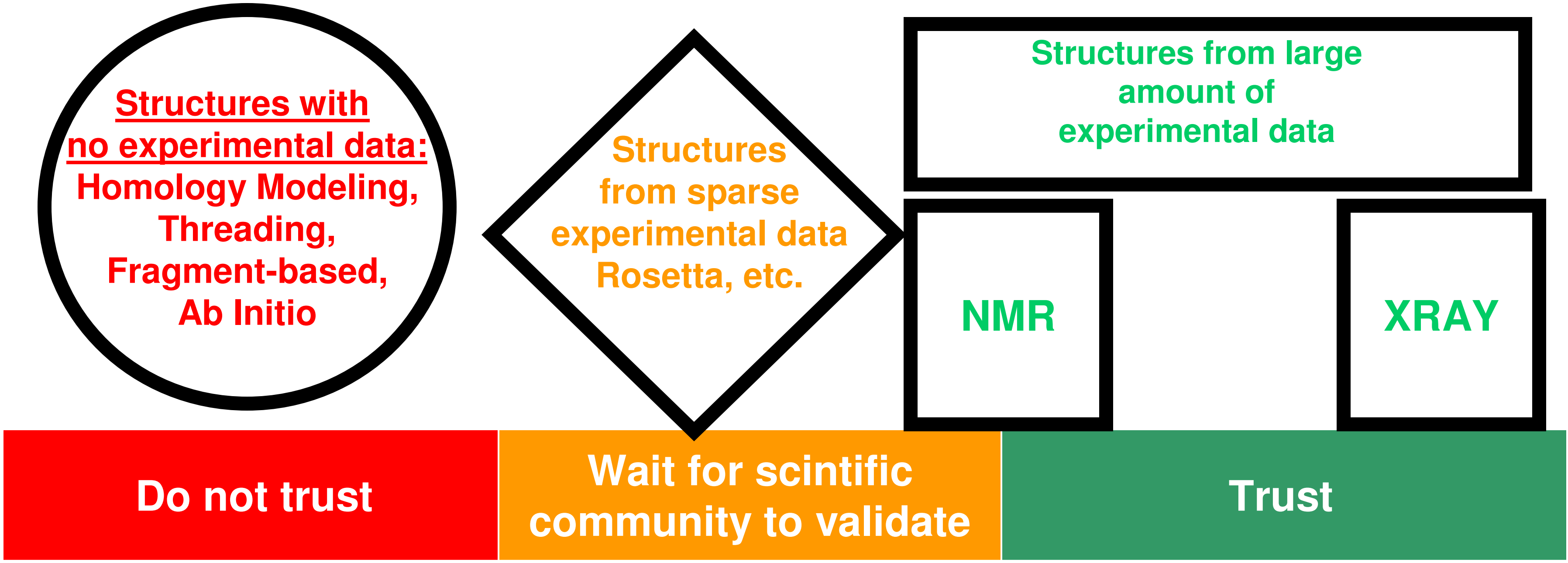


Model



Sparse data

Protein structures from the point of view of an experimentalist



Structures with no experimental data: Homology Modeling, Threading, Fragment-based, Ab Initio

Structures from sparse experimental data Rosetta, etc.

Structures from large amount of experimental data

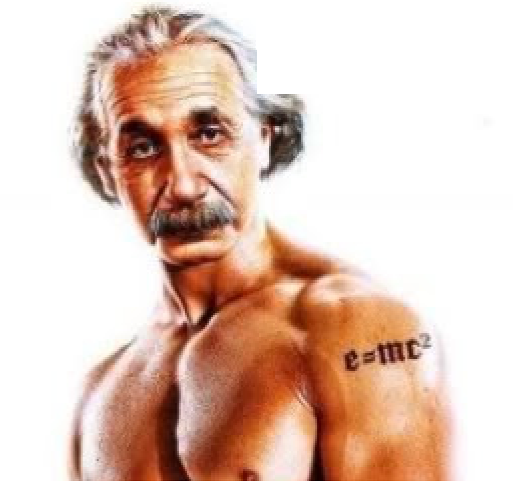
NMR

XRAY

Do not trust

Wait for scintific community to validate

Trust



What’s up with the “no free lunch” thing?

You can not build an accurate high-resolution model of protein structure without getting high-quality experimental data with your sweat and blood

1) There is no substitute for a large amount of experimental data. If you do not do experiment, you do not get the information relevant to your specific experimental conditions (e.g. protein construct, sample conditions, etc).

You can not get the same level of accuracy with sparse data or theoretical models

2) If you have an easy protein, do a full-blown structure determination

3) If you have no choice other than using sparse data, do not over-interpret your structure model.

This is not gonna happen any time soon

Success of theoretical methods is still limited to very small proteins.

Many theoretical models are biased, over-normalized, low-resolution, or simply inaccurate.

Accuracy and high-resolution of models from sparse data is questionable.