

7.91 Amy Keating

Solving structures using

X-ray crystallography

&

NMR spectroscopy

How are X-ray crystal structures determined?

1. **Grow crystals** - structure determination by X-ray crystallography relies on the repeating structure of a crystalline lattice.
2. **Collect a diffraction pattern** - periodically spaced atoms in the crystal give specific "spots" where X-rays interfere constructively.
3. Carry out a **Fourier transform** to get from "reciprocal space" to a real space description of the electron density.
4. THIS STEP REQUIRES KNOWLEDGE OF THE PHASES OF THE INTERFERING WAVES, WHICH CAN'T BE DIRECTLY MEASURED
"THE PHASE PROBLEM"
4. **Build a preliminary model** of the protein into the envelope of electron density that results from the experiment.
5. **Refine the structure** through an iterative process of changing the model and comparing how it fits the data.

The Phase Problem: we don't know what phases to use to add up all of the contributing waves. BIG PROBLEM.

$$\rho_{(x,y,z)} = \frac{1}{V} \sum_h \sum_k \sum_l F_{(h,k,l)} \exp[-2\pi \cdot i(hx + ky + lz)]$$

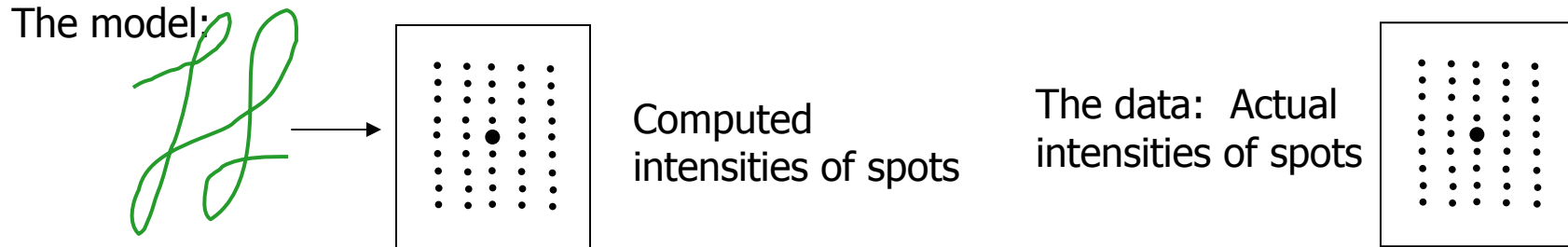
$$|F_{hkl}| \exp(i\alpha_{hkl}) = F_{(h,k,l)} = \sum_{j=1}^{atoms} f_{(j)} \exp[2\pi \cdot i(hx_{(j)} + ky_{(j)} + lz_{(j)})]$$

↑
observable
amplitude
↑
atomic scattering factor - related
to electron density around atom j
↑
the *phase* of F is determined by the
x, y and z coordinates of the atoms

What we *observe* is $I_{hkl} \propto |F_{hkl}|^2$
 we can't measure the phases directly

Get phases from molecular replacement, or heavy atom methods

X-Ray Crystal Structure Refinement



$$U_{\text{X-ray expt}} = \sum_{h,k,l} [|F_{\text{obs}}(h,k,l)| - |F_{\text{calc}}(h,k,l)|]^2$$

Summation runs over spots h, k, l

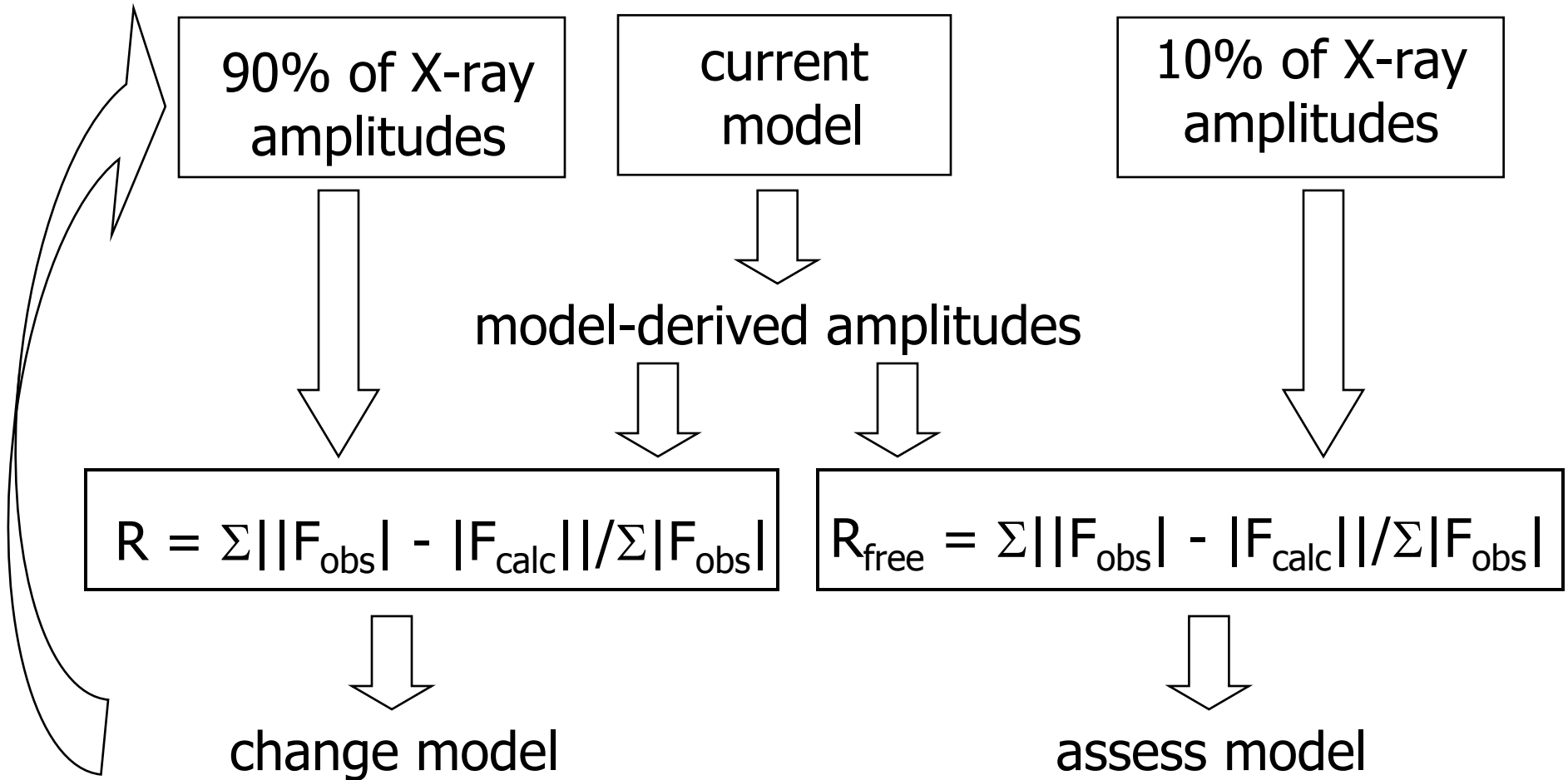
Actual intensity of spot observed in expt

Intensity of spot calculated from trial structure

$$U_{\text{hybrid}} = U_{\text{Molec Model}} + sU_{\text{X-ray expt}}$$

- Simulated annealing on hybrid potential rapidly improves correspondence between structure and X-ray observations while maintaining reasonable chemistry (large radius of convergence)
- Previous method effectively used local minimization which became trapped in local minima (small radius of convergence)

The Free R factor



What parameters do you refine?

- Atomic coordinates X, Y, Z
- The temperature factor of each atom, B
- Can also refine the occupancy

$$B = 8\pi^2 \times u^2$$

u^2 = mean square atomic displacement

B results from atomic vibrations and disorder
units = \AA^2

Example:

$B = 20 \rightarrow 0.5\text{\AA}$ displacement

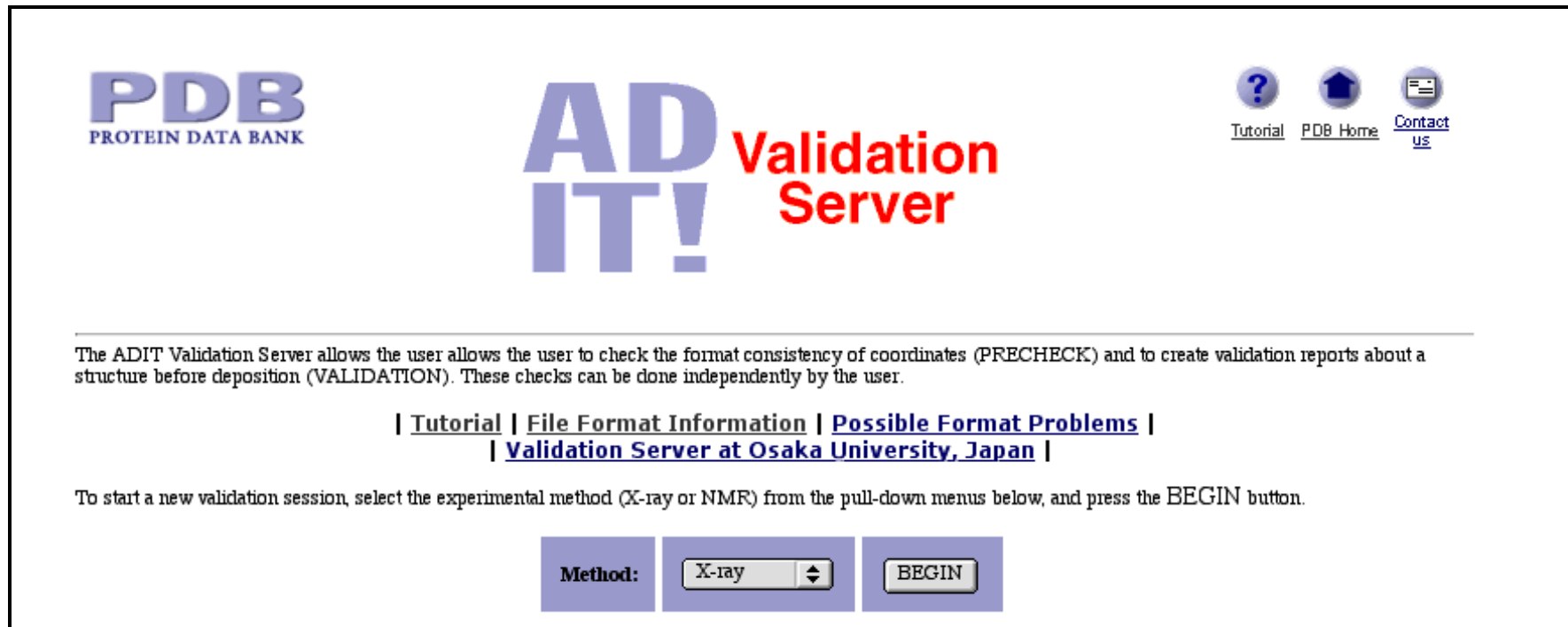
$B = 80 \rightarrow 1\text{\AA}$ displacement

Atomic coordinates in the PDB file

					X	Y	Z	occ	B
ATOM	1	N	GLU	4	28.492	3.212	23.465	1.00	70.88
ATOM	2	CA	GLU	4	27.552	4.354	23.629	1.00	69.99
ATOM	3	C	GLU	4	26.545	4.432	22.489	0.00	67.56
ATOM	4	O	GLU	4	26.915	4.250	21.328	0.00	68.09
ATOM	5	CB	GLU	4	28.326	5.683	23.680	0.00	72.34
ATOM	6	CG	GLU	4	27.447	6.910	23.973	0.00	75.98
ATOM	7	CD	GLU	4	28.123	8.247	23.659	0.00	78.43
ATOM	8	OE1	GLU	4	29.375	8.299	23.604	0.00	79.32
ATOM	9	OE2	GLU	4	27.393	9.251	23.468	0.00	79.58
ATOM	10	N	ARG	5	25.274	4.610	22.852	1.00	63.77
ATOM	11	CA	ARG	5	24.179	4.807	21.907	1.00	59.83
ATOM	12	C	ARG	5	23.411	3.698	21.219	1.00	56.20
ATOM	13	O	ARG	5	23.987	2.808	20.596	1.00	57.33
ATOM	14	CB	ARG	5	24.604	5.784	20.812	1.00	60.86
ATOM	15	CG	ARG	5	23.926	7.127	20.866	1.00	61.89
ATOM	16	CD	ARG	5	24.295	7.944	19.647	1.00	62.21

Is your structure correct?

- How unusual is the structure geometry?
- Does it contain rare conformations?
- Does it make chemical sense?



PDB
PROTEIN DATA BANK

ADIT! Validation Server

[Tutorial](#) [PDB Home](#) [Contact us](#)

The ADIT Validation Server allows the user to check the format consistency of coordinates (PRECHECK) and to create validation reports about a structure before deposition (VALIDATION). These checks can be done independently by the user.

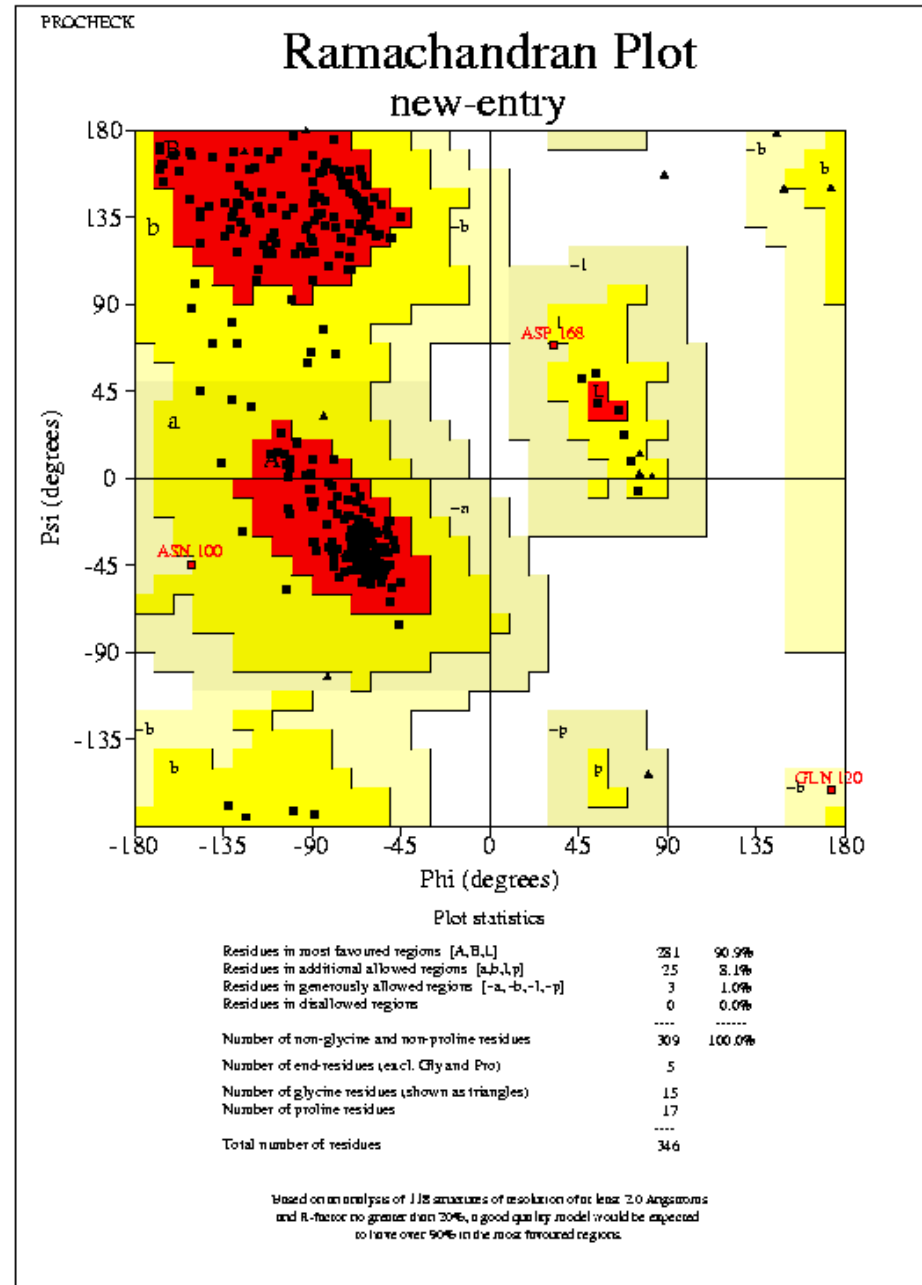
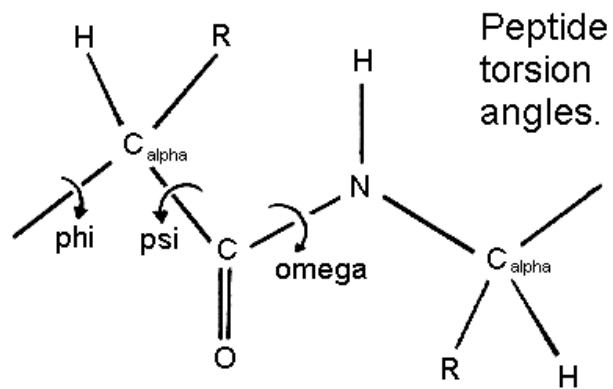
| [Tutorial](#) | [File Format Information](#) | [Possible Format Problems](#) |
| [Validation Server at Osaka University, Japan](#) |

To start a new validation session, select the experimental method (X-ray or NMR) from the pull-down menus below, and press the BEGIN button.

Method:

<http://pdb.rutgers.edu/validate/>

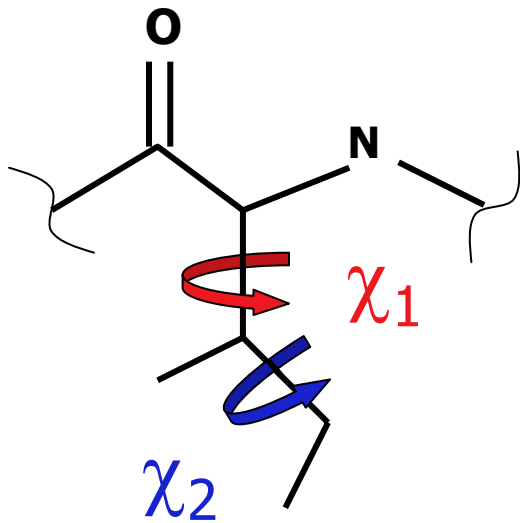
Backbone geometry



new-entry_01.ps

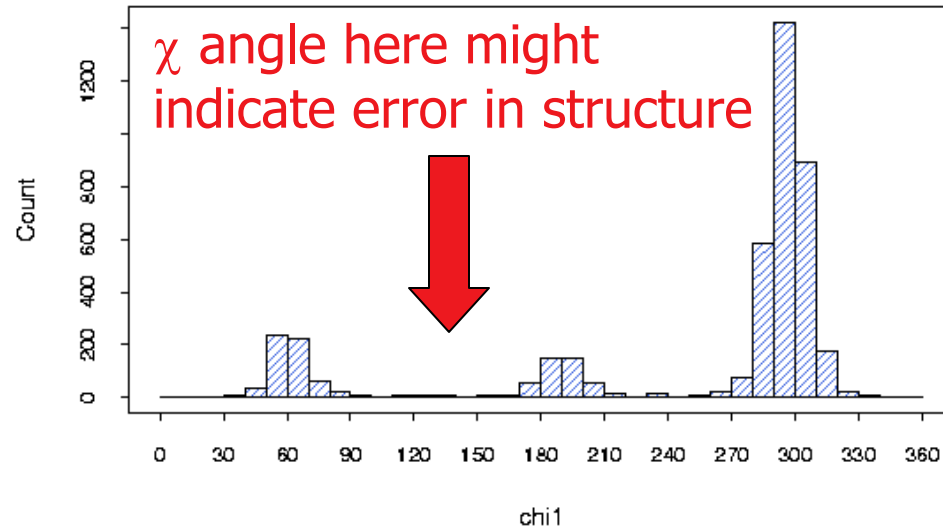
<http://pdb.rutgers.edu/>

Side chain geometry

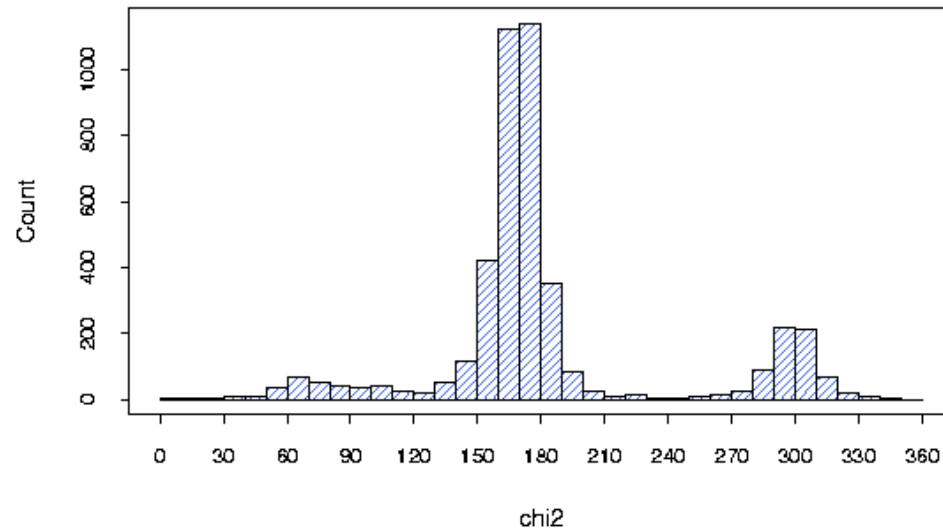


isoleucine

Ile chi1 distribution



Ile chi2 distribution

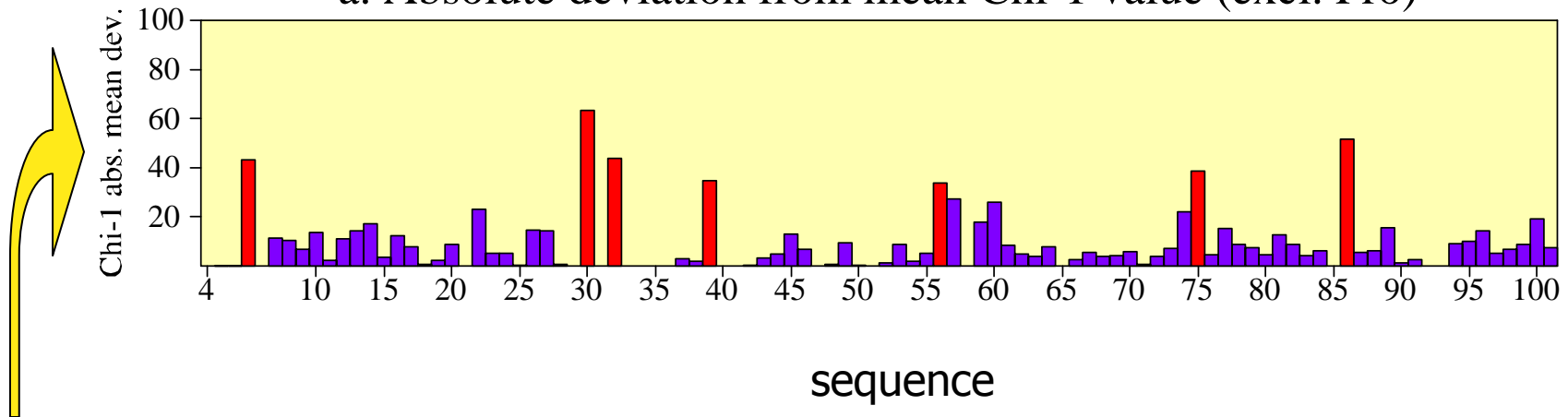


<http://pdb.rutgers.edu/validate/>

PROCHECK

Residue properties new-entry

a. Absolute deviation from mean Chi-1 value (excl. Pro)



χ_1 absolute deviation from values determined for high-resolution X-ray structures

Laskowski, R A, M W MacArthur, D S Moss, and J M Thornton. "PROCHECK: A Program to Check The Stereochemical Quality of Protein Structures." *J. Appl. Cryst.* 26 (1993): 283-291.

Morris, A L, M W MacArthur, E G Hutchinson, and J M Thornton. "Stereochemical Quality of Protein Structure Coordinates." *Proteins* 12 (1992): 345-364.

Summary of Structure Assessment

problem

diagnostic

structure is incomplete	PDB file header & coordinates, occupancies
residues are disordered	B-factors
model doesn't match data	R value Free R value
model has unusual stereochemistry	Ramachandran plots, side chain analysis

How are NMR structures solved?

1. **Solution phase technique** - protein at mM concentration in a buffer. Currently limited to proteins \leq 30-50 kDa.
2. **Measure resonant frequencies** of ^1H , ^{13}C , ^{15}N atoms in a magnetic field.
3. **Assign peaks** observed in the spectrum to individual amino acids.
4. **Measure distances** between different residues $< 6\text{\AA}$ apart to get restraints. Need many restraints per residue.
5. **Build structures** consistent with the experimental distance restraints and principles of stereochemistry.
6. Yields a **set of structures** consistent with the data.

- Please refer to <http://public-1.cryst.bbk.ac.uk/PPS2/projects/schirra/html/home.htm> for an NMR Tutorial.

Types of restraints available from NMR experiments

1. NOEs give rough distances between assigned atoms - given as upper and lower bounds.
2. COSY spectra and J-couplings give dihedral angle restraints

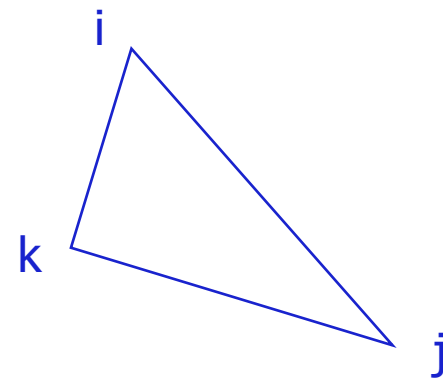
Also have constraints from what you know about the protein:

1. Connectivity due to known aa geometry & sequence
2. Standard bond lengths and angles

Building a structure from NMR data I: Distance Geometry

Given: a set of labeled distance constraints

1. Bounds smoothing using the triangle inequality
given upper bounds u and lower bound l (e.g. from NOEs and bond lengths)
if $u_{ij} > u_{ik} + u_{kj}$ then set u_{ij} to $u_{ik} + u_{kj}$
2. Specific distances d_{ij} that are compatible with the bounds and the triangle inequalities are chosen (metrization).
3. "Embedding" is used to compute a 3D model from the distances - often the distances are not all compatible with a 3D model but instead with a higher-dimensional one. In this case it is necessary to project into three dimensions (-> error).
4. Initial models contain many errors that must be iteratively corrected by refinement.



Building a structure from NMR data II: Simulated Annealing

$$U(\mathbf{R}) = E_{\text{empirical}} + E_{\text{effective}}$$

$$E_{\text{effective}} = E_{\text{NOE}} + E_{\text{torsion}} \text{ *derived from NMR experiment*}$$

$$E_{\text{empirical}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}} + E_{\text{vdW}} + E_{\text{elec}}$$

as previously-discussed

$$E_{\text{NOE}} = c \cdot |r_{ij} - r'_{ij}|^2$$

$$c = kTS/2 \cdot \Delta^2$$

where Δ is an error estimate on the experimental constraint r'_{ij}
S is chosen to balance the effective energy with the empirical energy

Assessing NMR structure quality

1. Number of restraints used

want ~10-20 per residue

2. Number of restraint violations

3. RMS deviation from restraints

4. RMS differences between models

want main chain atom rmsd < 0.4 Å, side chain < 1.0Å

5. Stereochemical quality

*e.g. use the validation server at the PDB to
check for bad backbone and side chain torsions*

Methods for Protein Structure Prediction

Homology Modeling

Threading

Ab Initio Prediction

Studying protein structure

... without a structure

Comparative modeling - inferring the structure of a protein from a homolog

Fold *recognition* - an easier problem than fold prediction!

Ab initio prediction - prediction of structure from sequence

Translating structure between members of the same family - **Homology Modeling**

- Identify a protein with similar sequence for which a structure has been solved (the *template*)
- Align the target sequence with the template
- Use the alignment to build an approximate structure for the target
- Fill in any missing pieces
- Fine-tune the structure
- Evaluate success

An excellent review:

Marti-Renom et al. *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000): 291-325.

Identifying a good template

- By sequence similarity
 - Use FASTA, BLAST, PSI-BLAST or threading
 - Best performance from high sequence identity, but can try distant homologues and assess performance later
- The closer the evolutionary relationship, the better
 - Consider a phylogenetic tree
- Generally better to have many templates to use as models
- Consider the structure quality (R, resolution, average B)
- Consider particulars of the structure
 - Quaternary structure
 - Any ligands bound?
 - pH
- The probability of finding a template is ~20-50%

You have cloned a new Pombe gene that is a putative protein kinase

Blast against **PDB**, hit = 1DM2

Score = 250 bits (638), Expect = 6e-67

Identities = 136/302 (45%), Positives = 185/302 (61%), Gaps = 17/302 (5%)

- Query: 71 IDDYELLEKIEEGSYGIVYRGLDKSTNTLVALKKIKFDPNGIGFPITSLREIESLSSIRH 130
- +++++ +EKI EG+YG+VY+ +K T +VALKKI+ D GP T++REI L + H
- Sbjct: 1 MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTEGVPSTAIRESLLKELNH 60

- Query: 131 DNIVELEKVVVGKDLKDVYLVMEFMEHDLKTL~~LD~~----NMPEDFLQSEVKTLMLQLLAA 185
- NIV+L V+ ++ +YLV EF+ DLK +D +P +K+ + QLL
- Sbjct: 61 PNIVKLLDVIHTEN--KLYLVFEFLHQDLKKFMDASALTGIPLPL---IKSYLFQLLQG 114

- Query: 186 TAFMHHH WYLHRDLKPSNLLMNNTGEIKLADFGLARPVSEPKSSLTRLVVTLWYRAPELL 245
- AF H H LHRDLKP NLL+N GIKLADFGLAR P +T VVTLWYRAPE+L
- Sbjct: 115 LAFCHSHRVLHRDLKPQNLLINTEGAIKLADFGLARAFGVPVRTYTHEVVTW YRAPEIL 174

- Query: 246 LGAPSYGKEIDMW SIGCIFAEMITRTPLFSGKSELDQLYKIFNLLGYPTREEWPQYFLLP 305
- LG Y +D+WS+GCIFAEM+TR LFG SE+DQL++IF LG P WP +P
- Sbjct: 175 LGCKYYSTAVDIWSLGCIFAEMVTRRALFPGDSEIDQLFRIFRTLGTDPDEVVW PGVTSMP 234

- Query: 306 YANKIKHPTVPTHSKIRTS--IPNLTGNAYDLLNRLSLNPAKRISAKEALEHPYFYESP 363
- P+P ++ S +P L + LL+++L +P KRISAK ALHP+F +
- Sbjct: 235 DYK—PSFPKWARQDFSKVVPPLDEDGRSLLSQMLHYDPNKRISAKAALAHPPFQDVT 290

- Query: 364 RP 365
- +P
- Sbjct: 291 KP 292

Aligning the target to the template sequences

- **A GOOD ALIGNMENT IS ABSOLUTELY ESSENTIAL**
- For $> 40\%$ sequence identity the alignment is usually clear
- For $< 40\%$ sequence identity usually have to deal with gaps

OBSERVATION: at 30% sequence only 20% of residues are correctly aligned!

- How could you try to improve the alignments over those provided by BLAST?

Aligning the target to the template sequences

- **A GOOD ALIGNMENT IS ABSOLUTELY ESSENTIAL**
- For > 40% sequence identity the alignment is usually clear
- For < 40% sequence identity usually have to deal with gaps

OBSERVATION: at 30% sequence only 20% of residues are correctly aligned!

- Try to use structural information

OBSERVATION: most insertions/deletions occur in loops, not in secondary structure elements

- Do a structure-based sequence alignment of all possible templates (e.g. with DALI)
- Add the target sequence to the alignment, using its predicted secondary structure to choose gap placement
- do the alignment over the known extent of a single protein domain in the template

To improve the alignment: check secondary structure of 1DM2 (given in the pdb entry)

```
1 MENFQKVEKI GEGTYGVVYK ARNKLTGEVV ALKKIRLDTE TEGVPSTAIR
   EEE EE  B SSSEEEE EEETTT EE EEEE          HHHH
```

```
51 EISLLKELNH PNIVKLLDVI HTENKLYLVF EFLHQDLKKF MDASALTGIP
   HTTTTTT  TTB B EEE EETTEEEEE E SEEHHHH HTTTTTT
```

```
101 LPLIKSYLFQ LLQGLAFCHS HRVLHRDLKP QNLLINTEGA IKLADFGLAR
   HHHHHHHHHH HHHHHHHHHH TT  S  G GGEEE TTS  EEE
```

```
151 AFGVPVRTYT HEVVTLWYRA PEILGCKYY STAVDIWSLG CIFAEMVTRR
   TT  HHHHTT SS  THHHHHHHH HHHHHHHHSS
```

```
201 ALFPGDSEID QLFRIFRTLQ TPDEVVWPGV TSMPDYKPSF PKWARQDFSK
   SS SSHHH HHHHHHHH  TTTSTTG GGTTTTTTS  GGG
```

```
251 VVPPLDEDGR SLLSQMLHYD PNKRISAKAA LAHPFFQDVT KPVPHLRL
   TTTT HHHH HHHHHHS SS TTTS  HHHH TTTGGGTT
```

Compare to the PREDICTED secondary structure
of the target
(from PHD, PREDATOR, JPRED, etc.)

Build a model from the alignment - I

- Construct a backbone framework
 - If you have only one model, copy the backbone coordinates for the aligned part of the target
 - If you have multiple models, average the C α positions, then fit a backbone trace to those positions by
 - using the template with highest sequence identity at each site
 - OR
 - selecting a hexapeptide from a database that fits

Build the model - II

- Add the side chains
 - For positions with identical sequence, copy the template structure
 - For positions with different sequence select the side chain placement from a list of commonly-observed conformers (known as “rotamers”)
 - Side chain positions may need to be iteratively refined so as to be consistent (more on this later!)

Build the model - III

- Build in the loops
 - Often the target differs from the templates in the loop region
 - Local sequence doesn't uniquely determine loop structure
 - Often loops contain important functional residues!
 - Loops can be built two ways
 - using a database of loop structures found in the pdb
 - Match the "stem" of the loop with a known segment, then transfer the coordinates to the target structure ("knowledge based" approach)
 - Do a conformational search using a molecular mechanics energy function (physics based approach)
 - These methods work reasonably for short loops (4-5 residues) and for specialized classes of loops (e.g. IgG hypervariable regions)

Refine the model

- The model as built in steps I - III may have poor stereochemistry (e.g. clashes)
- Can improve severe *local* errors through molecular mechanics minimization

**OBSERVATION: EXTENSIVE MINIMIZATION GIVES
WORSE MODELS**

- At this point side chain conformations can be adjusted to be consistent with the entire model

Optimization using constraints

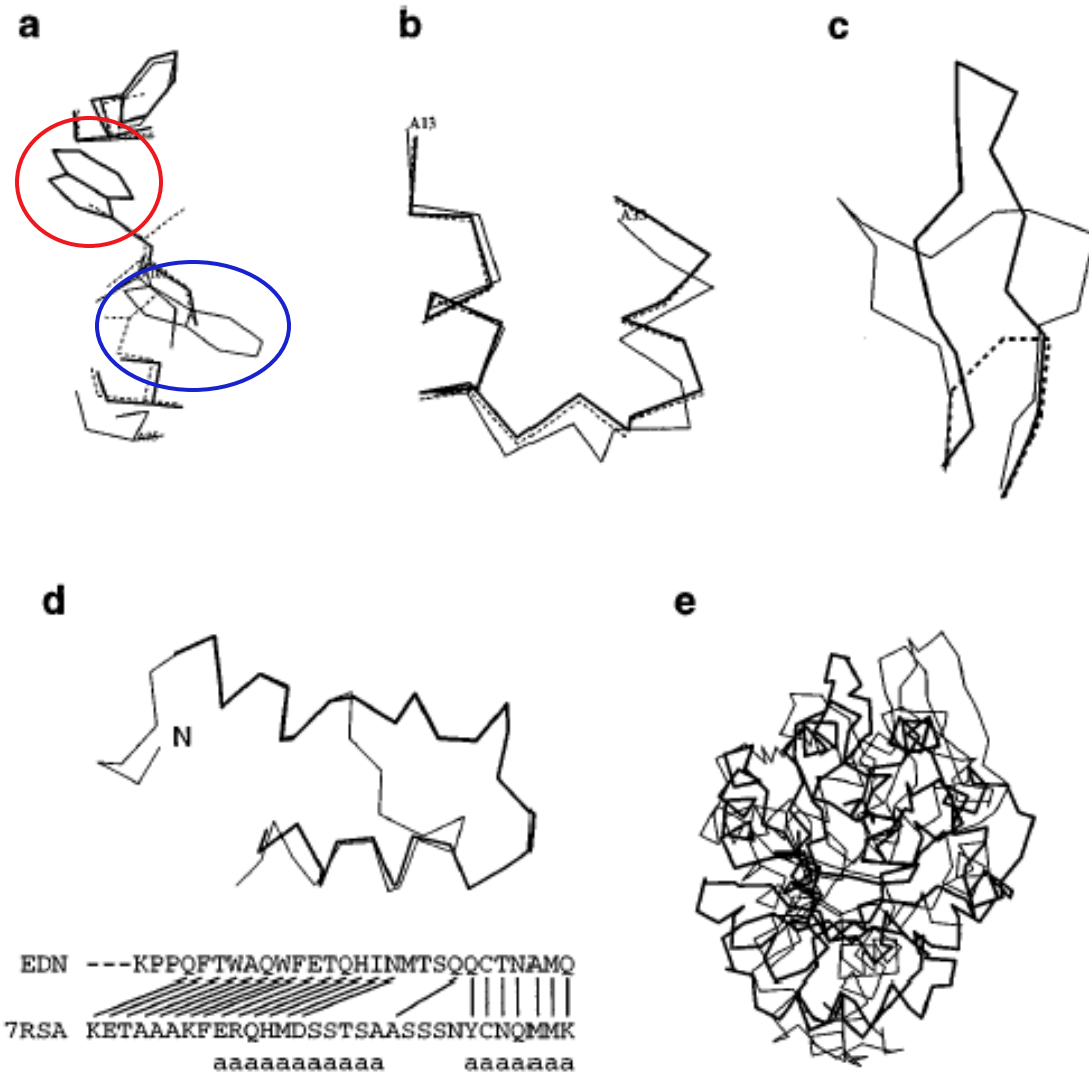
- A. Sali's MODELLER, G. Montelione's HOMA
- Uses the template to generate constraints
 - Atom distances, dihedral angles
- Uses molecular mechanics to introduce other constraints
 - Bond lengths, angles, dihedrals, non-bond terms
- Combine constraints into an objective function
- Minimize in Cartesian space
- Advantages: combines model building & refinement, can incorporate many types of data (e.g. NMR constraints)

Sali, A, and TL Blundell. "Comparative Protein Modelling by Satisfaction of Spatial Restraints." *J Mol Biol.* 234, no. 3 (5 December 1993): 779-815.

There are many places to go wrong...

- Bad template - it doesn't have the same structure as the target after all
- Bad alignment (a very common problem)
- Good alignment to good template still gives wrong local structure
- Bad loop construction
- Bad side chain positioning

Pitfalls in comparative modeling



Marti-Renom et al. *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000): 291-325.

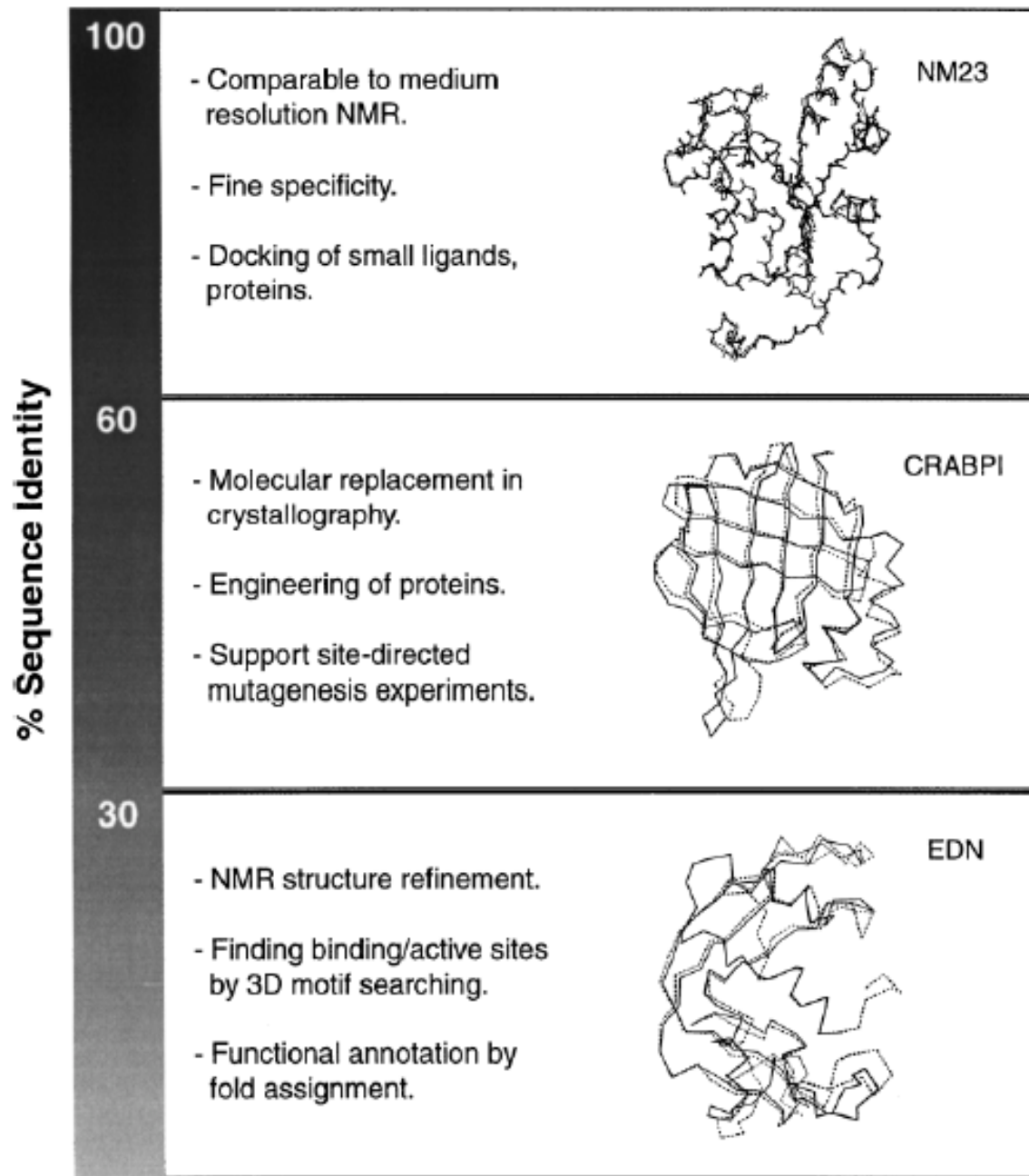
Courtesy of Annual Reviews Nonprofit Publisher of the Annual Review of TM Series. Used with permission.

How do you know if you can trust your model?

Model Assessment

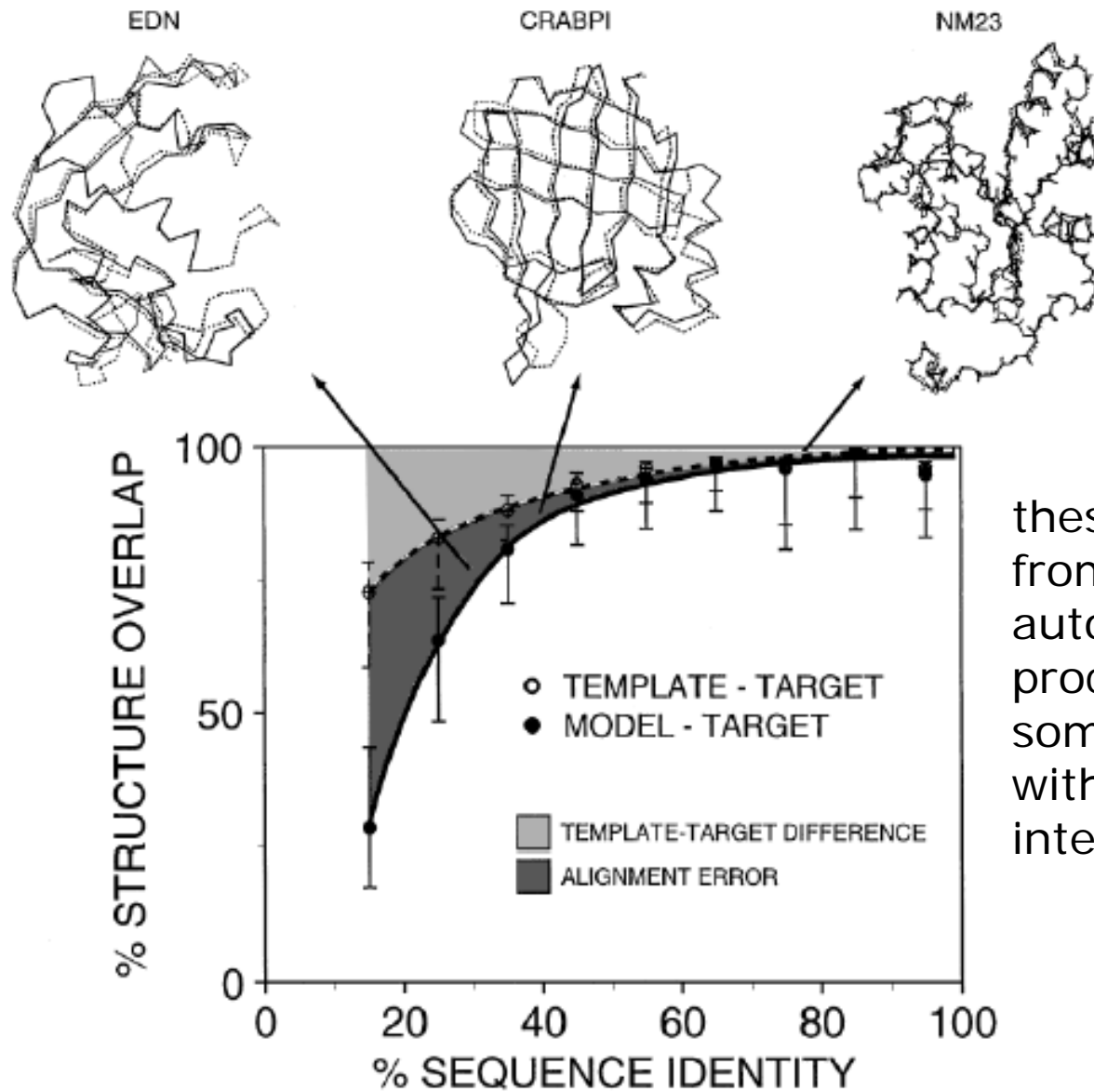
- The sequence identity between target and template
- Structural tests similar to those used for new crystal structures
 - backbone & side chain conformations, H-bonding
- Is the structure “protein-like”?
 - does it have good H/P patterning?
- Does it score better than alternate models according to some energy function?

$$\text{Z score} = \frac{S - \langle S \rangle}{\sigma}$$



Marti-Renom et al. *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000): 291-325.

Courtesy of Annual Reviews Nonprofit Publisher of the Annual Review of TM Series. Used with permission.



these numbers from an entirely automated process - can do somewhat better with manual intervention

Marti-Renom et al. *Annu. Rev. Biophys. Biomol. Struct.* 29 (2000): 291-325.

Courtesy of Annual Reviews Nonprofit Publisher of the Annual Review of TM Series. Used with permission.