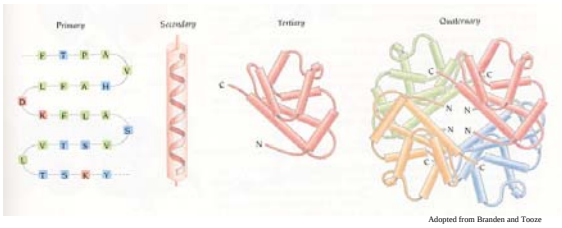
BINF 731

Protein Structure Analysis

Iosif Vaisman

2009

Protein Structure Hierarchy



Adopted from Branden and Tooze

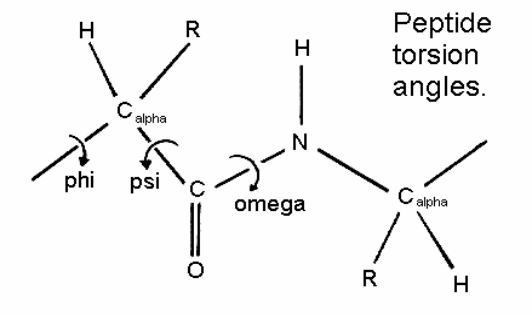
- Primary - the sequence of amino acid residues
- Secondary - ordered regions of primary sequence (helices, beta-sheets, turns)
- Tertiary - the three-dimensional fold of a protein subunit
- Quaternary - the arrangement of subunits in oligomers.

Anfinsen's Dogma

Three-dimensional structure of a protein is determined solely by its amino-acid sequence.

Native conformation of the protein is the global-minimum free energy conformation.

Levinthal paradox



3 conformations per residue is a very conservative estimate

Complexity of protein structure (Levinthal paradox)

100 residue protein
3 conformations per residue

number of distinct conformations:
$3^{100} \cong 10^{48}$

sampling time $\cong 10^{30}$ years

Complexity

P (Polynomial)
complexity class of decision problems for which execution time of a computation is no more than a polynomial function of the problem size

NP (Nondeterministic Polynomial)
complexity class of decision problems for which answers can be checked by an algorithm whose run time is polynomial in the size of the input

Protein Folding Problem

Given: sequence
Find: structure

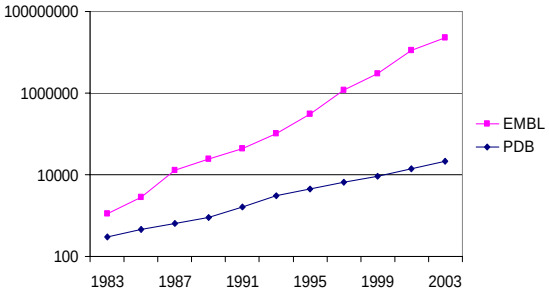
The problem is NP-complete

Protein Folding Problem

Problem for us, not for proteins.
They just fold...

(Ken Dill)

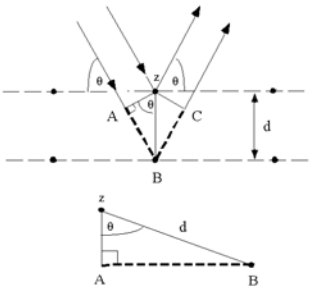
Dynamics of Database Growth



Protein Structure Determination

X-ray crystallography
NMR spectroscopy
Neutron diffraction
Electron microscopy
Atomic force microscopy

X-ray crystallography

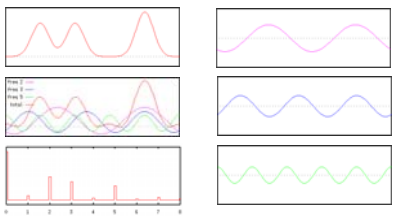


Bragg's Law

$n\lambda = 2d \sin\theta$

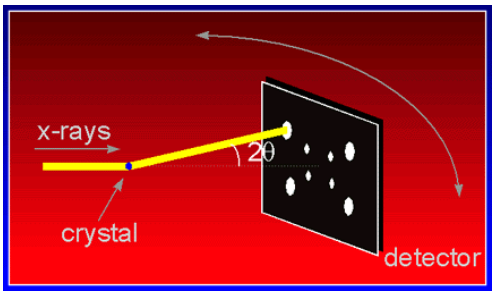
X-ray crystallography

Phase determination: MIR and MAD
(Multiple Isomorphous Replacement and Multiwavelength Anomalous Diffraction)

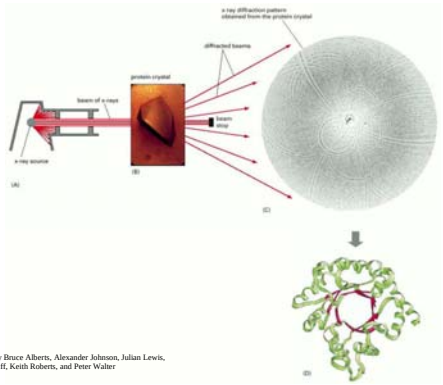


Fourier Transforms

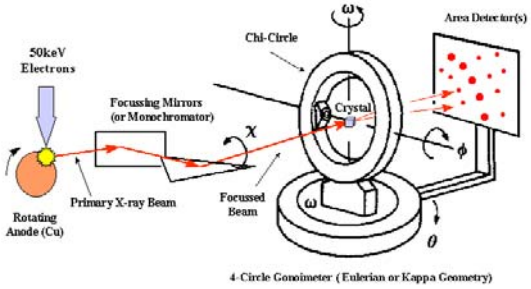
X-ray crystallography



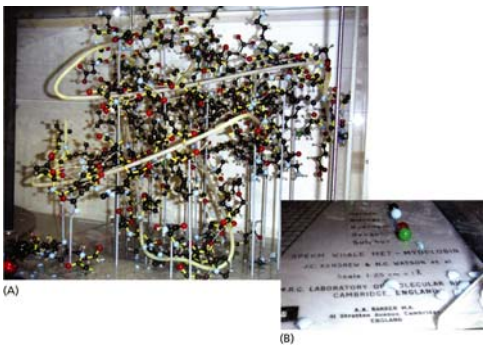
X-ray crystallography



X-ray crystallography

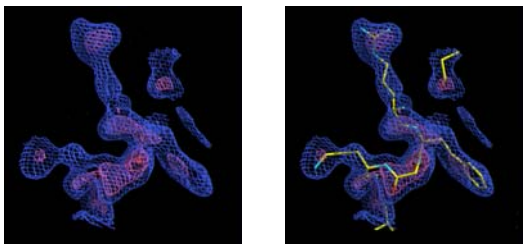


X-ray crystallography



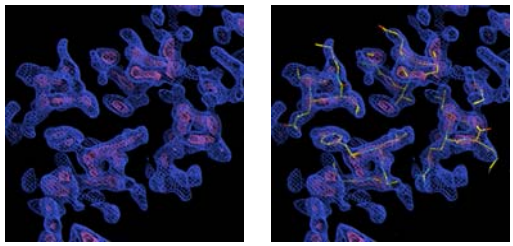
Adopted from *Zvelebil, Baum, 2008*

X-ray crystallography



Electron density map created from multi-wavelength data (Arg)

X-ray crystallography



Experimental electron density map and model fitting
(apoE four helix bundle)

X-ray crystallography

Confidence in structural features of proteins determined by X-ray crystallography

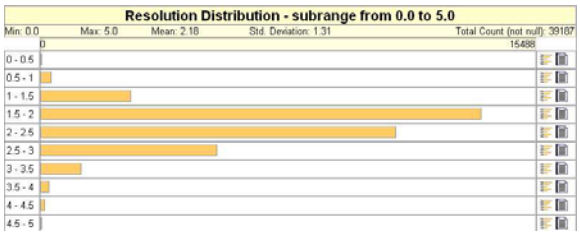
(These are *rough* estimates, and depend strongly on the quality of the data.)

Structural feature	Resolution				
	5 Å	3 Å	2.5 Å	2.0 Å	1.5 Å
Chain tracing	—	Fair	Good	Good	Good
Secondary structure	Helices fair	Fair	Good	Good	Good
Sidechain conformations	—	—	Fair	Good	Good
Orientation of peptide planes	—	—	Fair	Good	Good
Protein hydrogen atoms visible	—	—	—	—	Good

wwPDB statistics

Year	Total Depositions	Deposited To			Processed By		
		RCSB	PDBj	EBI	RCSB	PDBj	EBI
2000	2983	2445	10	528	2294	161	528
2001	3286	2673	118	495	2407	384	495
2002	3563	2769	289	505	2401	657	505
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3083	1613	812
2005	6678	4507	1166	1005	3563	2110	1005
2006	7282	5145	1052	1085	4252	1945	1085
2007	6045	3872	1373	800	3575	1670	800
TOTAL	40175	28695	5581	5899	24714	9562	5899

PDB resolutions



PDB redundancy

Method	Description	# of Clusters
blast	95% identity (one chain)	17575
blast	90% identity (one chain)	16853
blast	70% identity (one chain)	15114
blast	50% identity (one chain)	12886
blast	40% identity (one chain)	11218
blast	30% identity (one chain)	9294

PDB ambiguities

Table 1 The number of PDB structures retrieved by ambiguous chemical component codes

Code	Name	Number of PDB structures[a]
SUL	Sulfate anion	156 (3.6%)
SO4	Sulfate ion	4083 (96.4%)
SUL and SO4	Sulfate anion and sulfate ion	1 (0.03%)
NET	Tetraethylammonium ion	9 (90%)
E4N	Tetraethylammonium ion	1 (10%)
MMC	Methyl mercury ion	8 (66.66%)
HGC	Methyl mercury ion	4 (33.33%)

[a]Percentages of the total number of structures with the chemical component are shown in brackets. Search carried out August 2006.

Schierz et al., Nature Biotechnology 25:437

Structure verification and validation

Biotech Validation Suite
ERRAT
Verify3D
Procheck

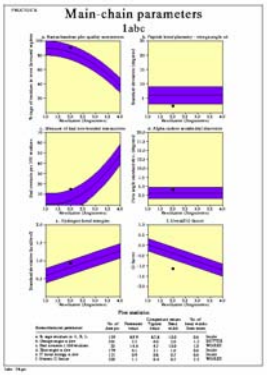
Bond lengths (Procheck)

Bond	labeling		Value	sigma
C-N	C-NH1	(except Pro)	1.329	0.014
	C-N	(Pro)	1.341	0.016
C-O	C-O		1.231	0.020
Calpha-C	CH1E-C	(except Gly)	1.525	0.021
	CH2G*-C	(Gly)	1.516	0.018
Calpha-Cbeta	CH1E-CH3E	(Ala)	1.521	0.033
	CH1E-CH1E	(Ile,Thr,Val)	1.540	0.027
	CH1E-CH2E	(the rest)	1.530	0.020
N-Calpha	NH1-CH1E	(except Gly,Pro)	1.458	0.019
	NH1-CH2G*	(Gly)	1.451	0.016
	N-CH1E	(Pro)	1.466	0.015

Bond angles (Procheck)

Angle	labeling		Value	sigma
C-N-Calpha	C-NH1-CH1E	(except Gly,Pro)	121.7	1.8
	C-NH1-CH2G*	(Gly)	120.6	1.7
	C-N-CH1E	(Pro)	122.6	5.0
Calpha-C-N	CH1E-C-NH1	(except Gly,Pro)	116.2	2.0
	CH2G*-C-NH1	(Gly)	116.4	2.1
	CH1E-C-N	(Pro)	116.9	1.5
Calpha-C-O	CH1E-C-O	(except Gly)	120.8	1.7
	CH2G*-C-O	(Gly)	120.8	2.1

Procheck output



a. Ramachandran plot quality - percentage of the protein's residues that are in the core regions of the Ramachandran plot.

b. Peptide bond planarity - standard deviation of the protein structure's omega torsion angles.

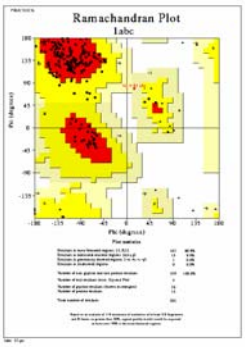
c. Bad non-bonded interactions - number of bad contacts per 100 residues.

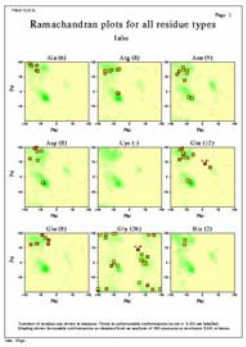
d. Cα tetrahedral distortion - standard deviation of the ζ torsion angle (Cα, N, C, and Cβ).

e. Main-chain hydrogen bond energy - standard deviation of the hydrogen bond energies for main-chain hydrogen bonds.

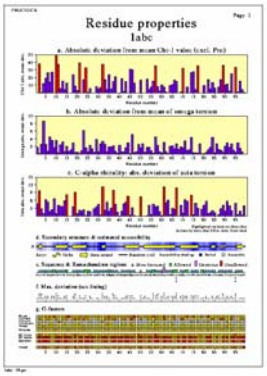
f. Overall G-factor - average of different G-factors for each residue in the structure.

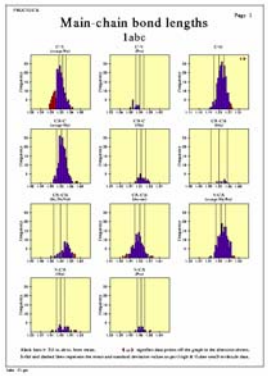
Procheck output



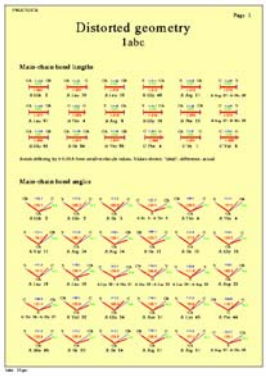


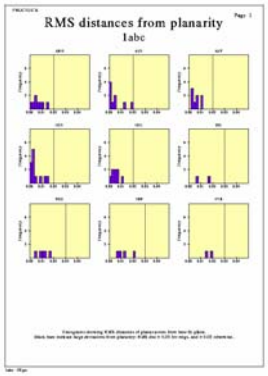
Procheck output



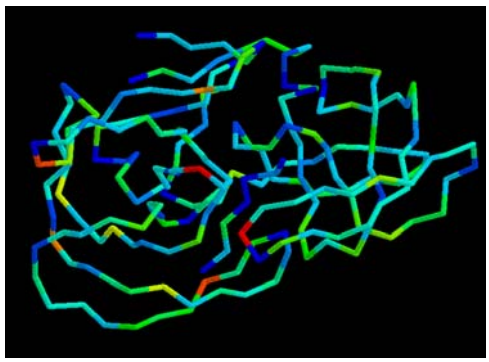


Procheck output





Procheck output - backbone G factors



Procheck output - all atom G factors