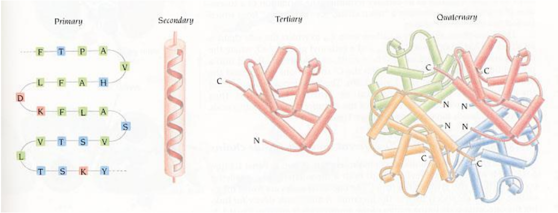
BINF 731

Protein Structure Analysis

Iosif Vaisman

2015

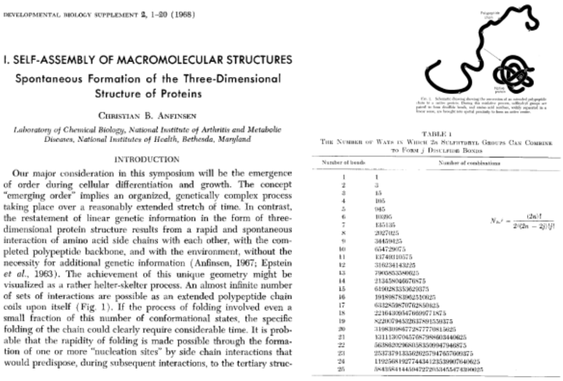
Protein Structure Hierarchy



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.

Protein folding



DEVELOPMENTAL BIOLOGY SUPPLEMENT 2, 1–20 (1968)

I. SELF-ASSEMBLY OF MACROMOLECULAR STRUCTURES

Spontaneous Formation of the Three-Dimensional Structure of Proteins

CHRISTIAN B. ANFINSEN

Laboratory of Chemical Biology, National Institute of Arthritis and Metabolic Diseases, National Institutes of Health, Bethesda, Maryland

INTRODUCTION

Our major consideration in this symposium will be the emergence of order during cellular differentiation and growth. The concept "emerging order" implies an organized, genetically complex process taking place over a reasonably extended stretch of time. In contrast, the restatement of linear genetic information in the form of three-dimensional protein structure results from a rapid and spontaneous interaction of amino acid side chains with each other, with the completed polypeptide backbone, and with the environment, without the necessity for additional genetic information (Anfinsen, 1967; Epstein *et al.*, 1963). The achievement of this unique geometry might be visualized as a rather helter-skelter process. An almost infinite number of sets of interactions are possible as an extended polypeptide chain coils upon itself (Fig. 1). If the process of folding involved even a small fraction of this number of conformational states, the specific folding of the chain could clearly require considerable time. It is probable that the rapidity of folding is made possible through the formation of one or more "nucleation sites" by side chain interactions that would predispose, during subsequent interactions, to the tertiary struc-

FIG. 1. Schematic drawing showing the conversion of an extended polypeptide chain to a native protein. During this oxidative process, sulfhydryl groups are paired to form disulfide bonds, and amino acid residues, widely separated in a linear sense, are brought into spatial proximity to form an active center.

TABLE 1
THE NUMBER OF WAYS IN WHICH 2n SULFHYDRYL GROUPS CAN COMBINE TO FORM j DISULFIDE BONDS

Number of bonds	Number of combinations
1	1
2	3
3	15
4	105
5	945
6	10395
7	135135
8	2027025
9	34459425
10	654729075
11	13749310575
12	316234143225
13	7905853580625
14	213458046676875
15	6190283353629375
16	191898783962510625
17	6332659870762850625
18	221643095476699771875
19	8200794532637891559375
20	319830986772877770815625
21	13113070457687988603440625
22	563862029680583509947946875
23	25373791335626257947657609375
24	1192568192774434123539907640625
25	58435841445947272053455474390625

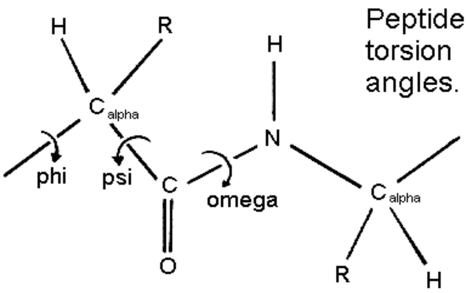
$$N_{n,j} = \frac{(2n)!}{2^j(2n-2j)!j!}$$

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

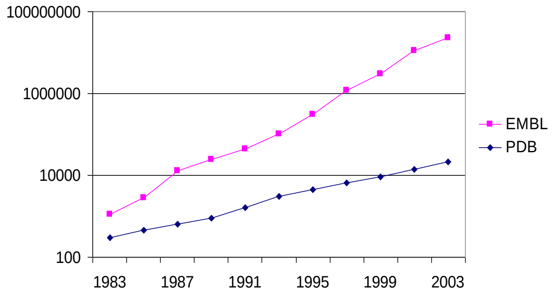
Given: sequence
Find: structure

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

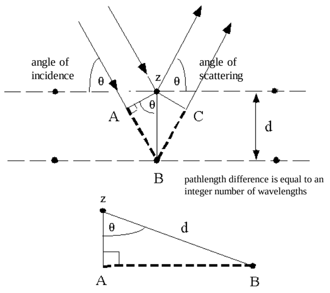
NMR spectroscopy

Neutron diffraction

Electron microscopy

Atomic force microscopy

X-ray crystallography

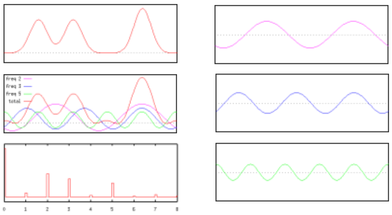

angle of incidence
θ
z
θ
angle of scattering
θ
A
C
d
B
pathlength difference is equal to an integer number of wavelengths
z
θ
d
A
B

Bragg's Law

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

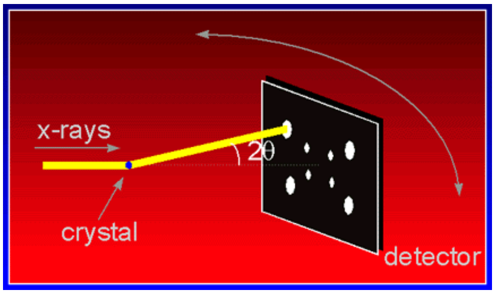
X-ray crystallography

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

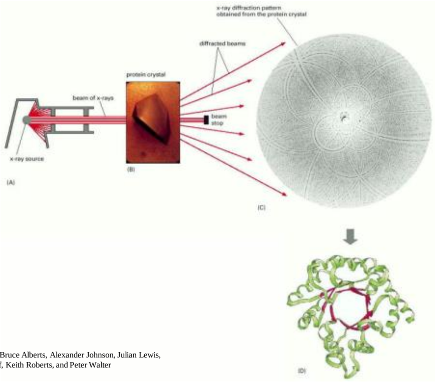


Fourier Transforms

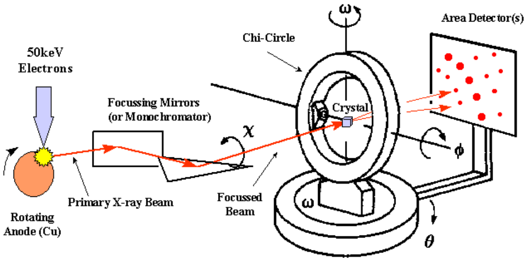
X-ray crystallography



X-ray crystallography

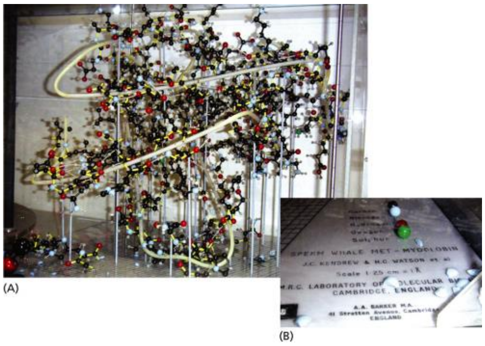


X-ray crystallography



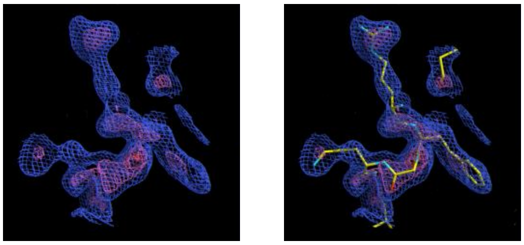
4-Circle Goniometer (Eulerian or Kappa Geometry)

X-ray crystallography


SPERM WHALE MET - MYOGLOBIN
J.C. KENDREW & H.C. WATSON et al
Scale 1·25 cm = 1 Å
M.R.C. LABORATORY OF MOLECULAR BIOLOGY
CAMBRIDGE, ENGLAND
A.A. BARKER M.A.
(A)
(B)

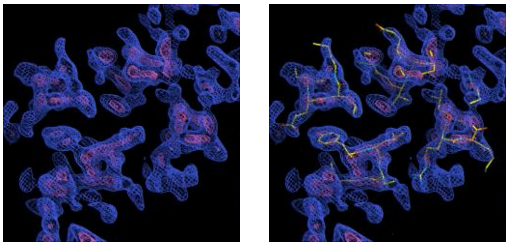
Adopted from *Zvelebil, Baum, 2008*

X-ray crystallography



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

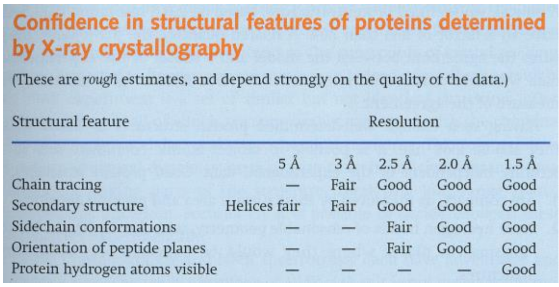
X-ray crystallography



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

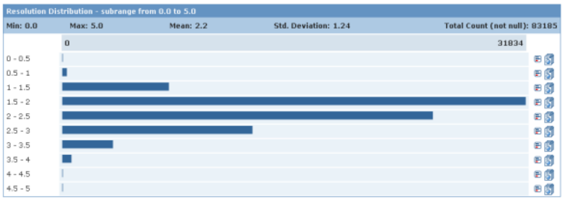
wwPDB statistics

Year	Total Depositions	Deposited To			Processed By		
		RCSB PDB	PDBj	PDBe	RCSB PDB	PDBj	PDBe
2001	3287	2673	118	496	2408	383	496
2002	3565	2769	289	507	2401	657	507
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3082	1614	812
2005	6678	4507	1166	1005	3563	2110	1005
2006	7282	5145	1052	1085	4252	1945	1085
2007	8130	5399	1603	1128	4703	2299	1128
2008	7073	5452	648	973	4106	1994	973
2009	8300	6715	527	1058	5069	2173	1058
2010	8878	6912	593	1373	5464	2041	1373
2011	9250	7172	582	1496	5938	1816	1496
2012	9972	7695	601	1676	6408	1888	1676
2013	10566	8031	749	1786	6652	2128	1786
2014	10364	8178	501	1685	6040	1779	2545
2015	8070	6880	49	1141	3692	1411	2968
TOTAL	114736	87257	10061	17418	69210	25422	20105

PDB statistics

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	93956	1668	4692	4	100320
NMR	9751	1130	227	8	11116
ELECTRON MICROSCOPY	619	29	204	0	852
HYBRID	76	3	2	1	82
other	168	4	6	13	191
Total	104570	2834	5131	26	112561

PDB resolutions



Resolution Distribution - subrange from 0.0 to 5.0

Min: 0.0 Max: 5.0 Mean: 2.2 Std. Deviation: 1.24 Total Count (not null): 83185

PDB redundancy

Description	# of Clusters
100% identity	63082
95% identity	44244
90% identity	42214
70% identity	37420
50% identity	32207
40% identity	28542
30% identity	24304

PDB ambiguities

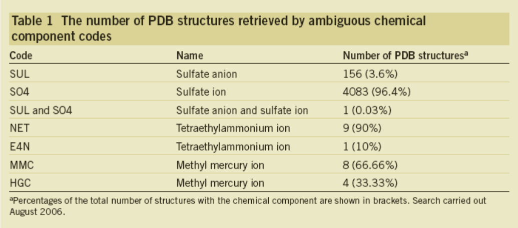


Table 1 The number of PDB structures retrieved by ambiguous chemical component codes

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