Computational Structural Biology

in

Post-genomic Era

Ivet Bahar

Department of Computational Biology
and Department of Molecular Genetics & Biochemistry
School of Medicine, University of Pittsburgh

References

- C. Branden and J. Tooze. ***Introduction to Protein Structure.*** 2nd edition, Garland Publishing Inc., New York, 1999.

- C. L. Brooks, III, M. Karplus, and B.M. Pettitt. ***A Theoretical Perspective of Dynamics, Structure, and Thermodynamics.*** Wiley Interscience, New York, 1988.

- C.R. Cantor and P.R. Schimmel. ***Biophysical Chemistry.*** Vol.1,2,3. W.H. Freeman and Company, San Fransisco, 1980.
- T.E. Creighton, Editor. ***Protein Folding.*** W.H. Freeman & Company, New York, 1992.

- A. Fersht. ***Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding.*** W. H. Freeman and Company, New York, 1999.

- L.M. Gierasch and J. King, Editors. ***Protein Folding, Deciphering the Second Half of the Genetic Code.*** AAAS, Washington D.C., 1990.

- A.Y. Grosberg and A.R. Khokhlov. ***Giant Molecules. Here, There, and Everywhere...*** Academic Press, San Diego, California, 1997.

- A. R. Leach. ***Molecular Modelling. Principles and Applications.*** Addison Wesley Longman, Essex, England, 1996.

- J.A. McCammon and S.C. Harvey. ***Dynamics of Proteins and Nucleic Acids.*** Cambridge University Press, Cambridge, 1987.

- G.E. Schulz and R.H. Schirmer. ***Principles of Protein Structure.*** Springer Advanced Texts in Chemistry, Springer-Verlag, New York, 1990.

- L. Stryer. ***Biochemistry.*** W.H. Freeman, New York, latest edition.

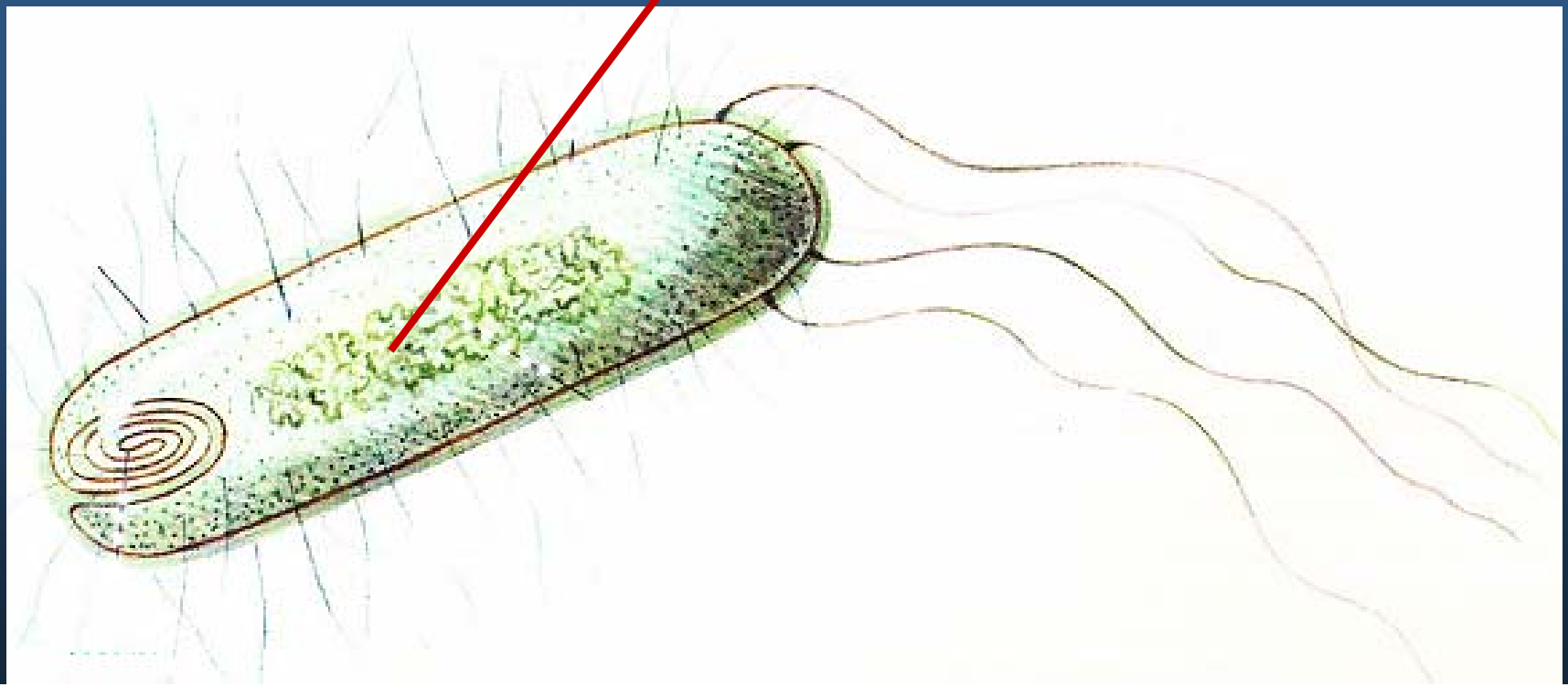
2001 – Draft version of human genome published

- 1990 - Human Genome Project (HGP) launched
- 1993 – 1st five-year plan published
- 1995 - first bacterial genome published (*haemophilus influenza*)
- 1996 – yeast genome sequenced
- 1997 – *E coli* genome sequenced
- 1998 – *C elegans* genome
- 1998 – 2nd five-year plan for HGP
- 2000 – Fruit fly genome
- 2002 – rice genome – 1st draft
- 2002 –mouse genome – 1st draft
- 2003 – HGP completed

GENOME SEQUENCING PROJECTS

E. coli
Genome: 4.6 million nucleotides
4289 proteins

Human
Genome: 3 billion nucleotides
~30-40,000 proteins



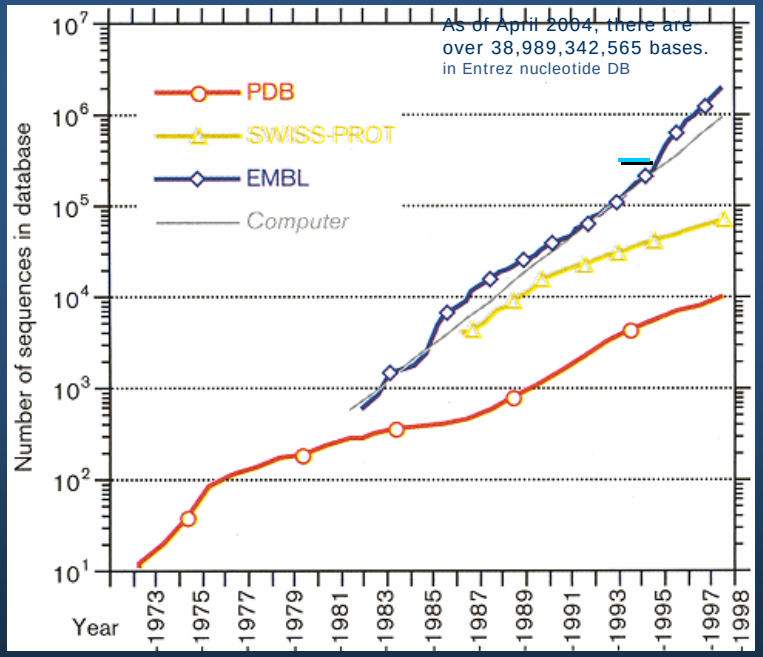
The genomes of many species have been sequenced to date…

… but limited information is conveyed from sequence about how genomes and proteomes give rise to biological function.

Image: Digital Vision, PhotoDisc, Matt Ray/EHP c

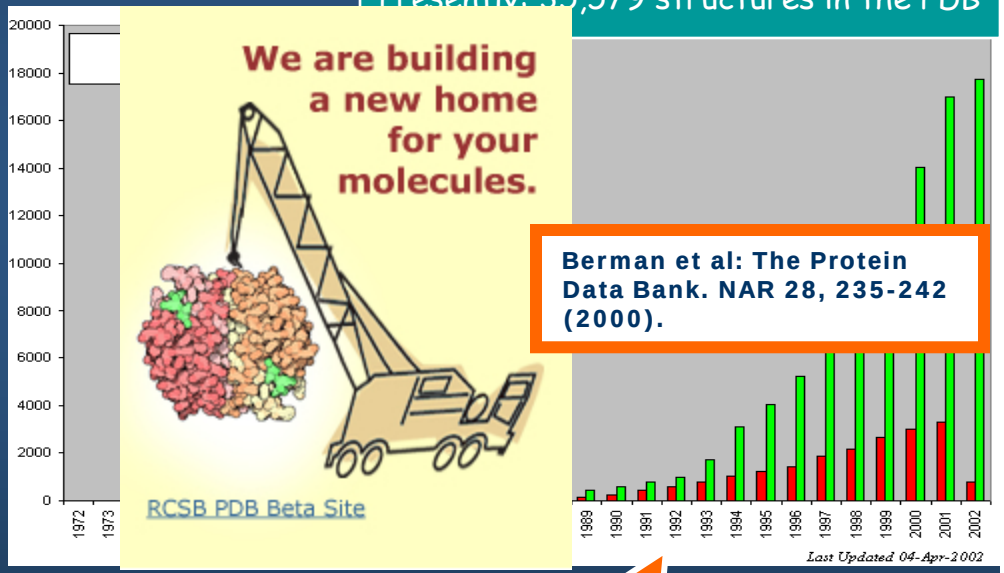
Exponential growth in

- Sequential, structural , genetic and biomedical data
- Computational technology



Rost, B. (1998). Marrying structure and genomics. *Structure* 6, 259-263

Presently: 35,579 structures in the PDB



Berman et al: The Protein Data Bank. NAR 28, 235-242 (2000).

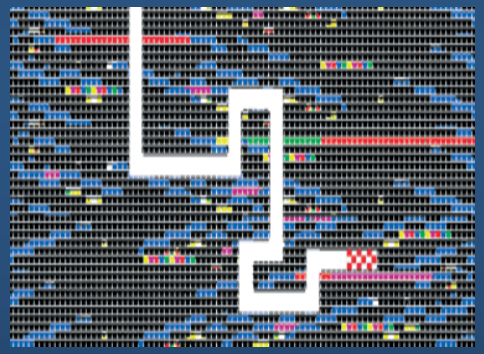
Structural genomics
Functional genomics
Proteomics

Promising Future for Computational Biology

Economist.com

The race to computerize biology

Dec 12th 2002
From The Economist print edition



"In life-sciences establishments around the world, the laboratory rat is giving way to the computer mouse—as computing joins forces with biology to create a bioinformatics market that is expected to be worth nearly $40 billion within three years"

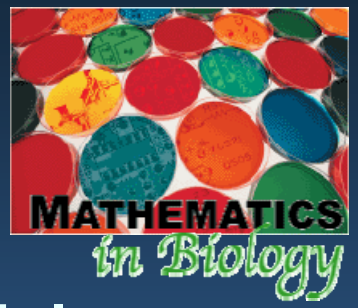
Biotech and pharmaceutical industry became one of the biggest consumers of computing power,

- **supercomputing powers of petaflops (~ 10^{12} floating-point operations per sec)**
- **Storage capacity of terabytes (~ 10^{9} of bytes)**

"A big risk of computer modeling and other tools is to rely too much on them."

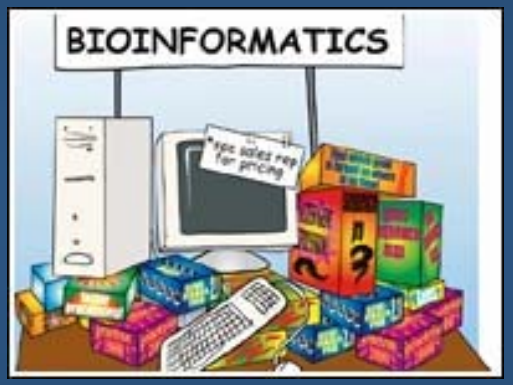
"Wet lab processes are giving way to digital research done *in silico*"

In a special collection of articles published beginning 6 February 2004, *Science* Magazine and its online companion sites team up to explore the interface between mathematics and biology



Bioinformatics Moves to Center Stage in the Genetic Revolution

Computational Biology



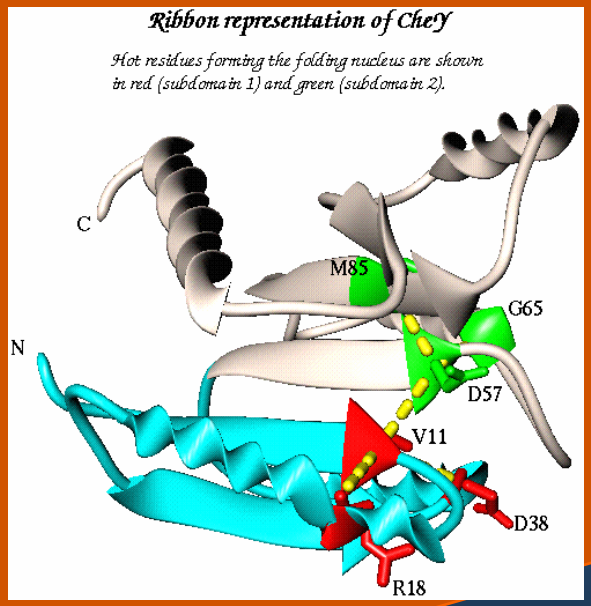
A multidisciplinary field encompassing

- molecular-to-cellular **modeling** of structure and function
- physically inspired **simulation** and visualization of complex processes at multiple scales
- elucidation of the mechanism of operation of biological **systems** (networks of interactions)

- **Biological Pathways, and Networks**
- **Molecular Libraries and Imaging**
- **Structural Biology**
- **Bioinformatics and Computational Biology**
- **Nanomedicine**

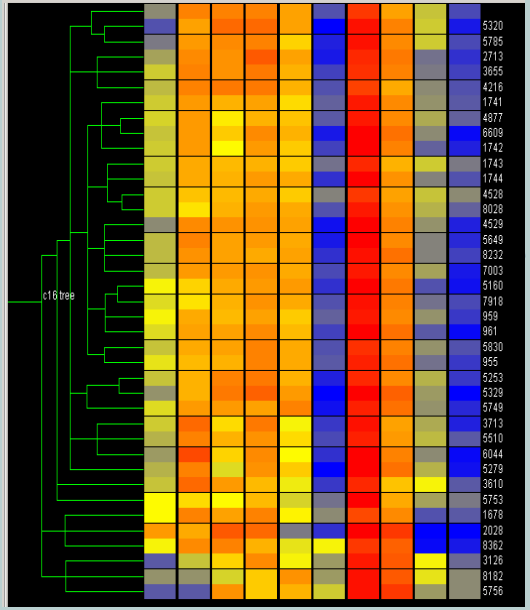
y initiatives, the **NIH Roadmap**
advancing medical research"

Computational Biology



Bioinformatics

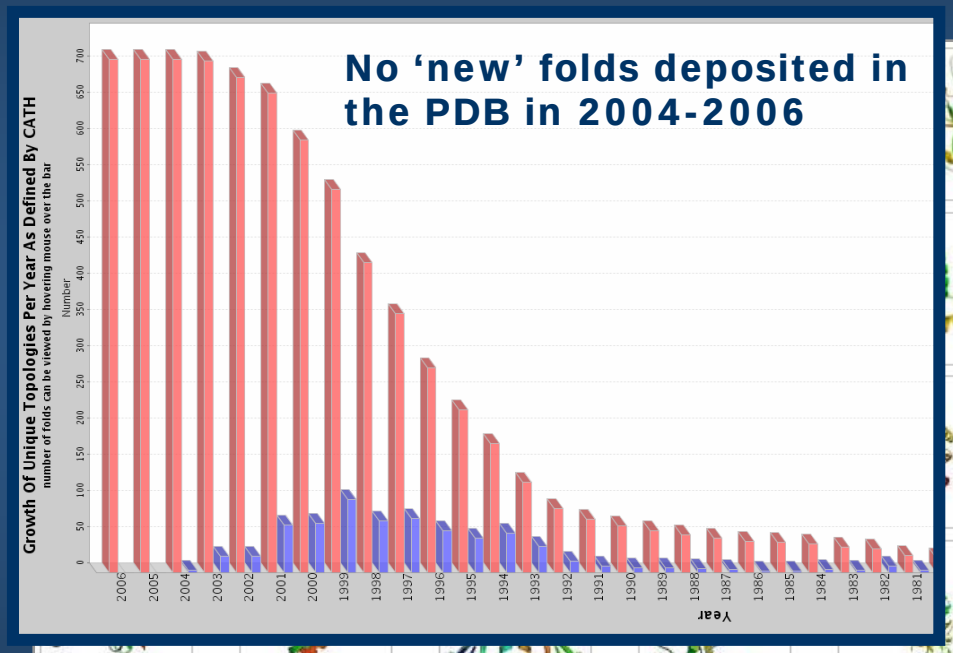
- Database analysis



Five areas of specialization

- Computational Structural Biology
- Computational Genomics
- Systems / Mathematical Biology
- Computational Neurobiology
- Bioimage Informatics

Proteomics – Examining all proteins encoded by a given genome



A more integrated view of

- structural dynamics
- protein-protein interactions

or

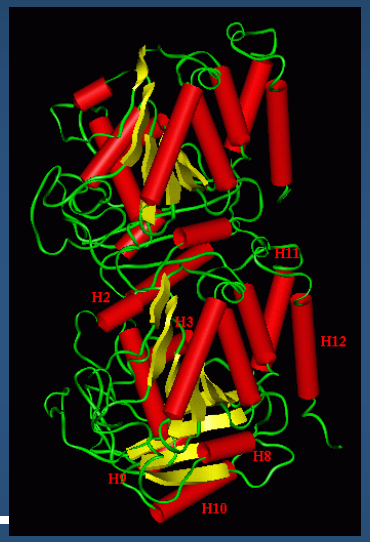
- systems level assessment of pathways, networks, and their dynamics

is now possible.

We need to understand the physical principles that underlie the passage

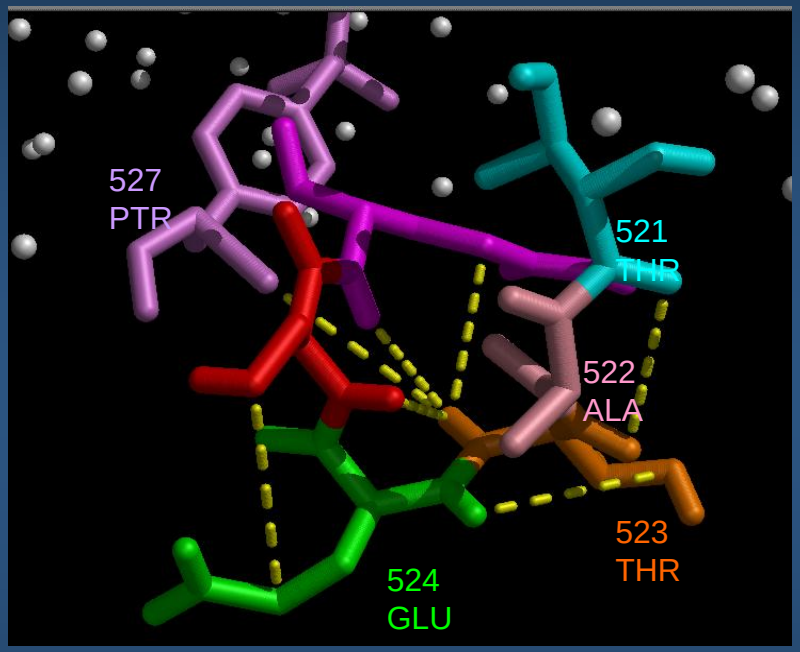
from sequence,

to structure...

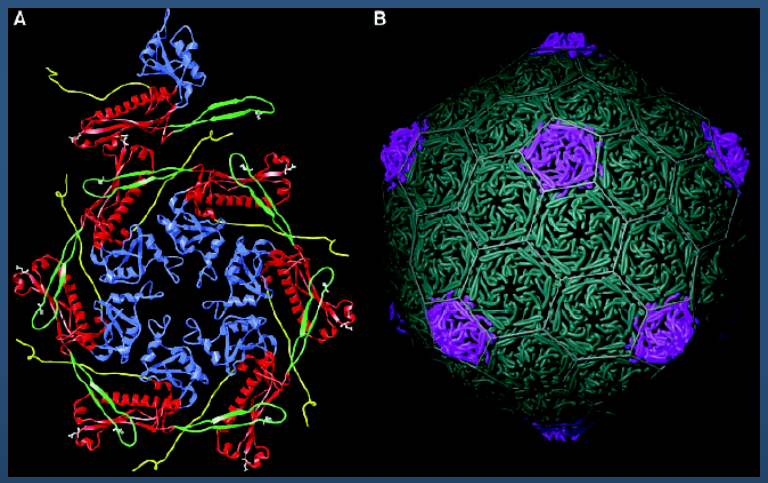


to dynamics...

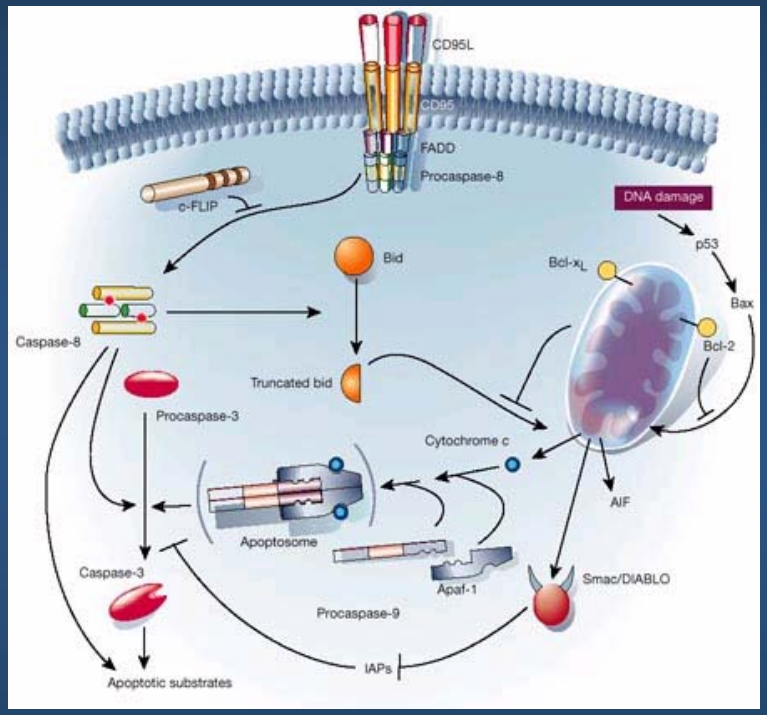
from interacting atoms...

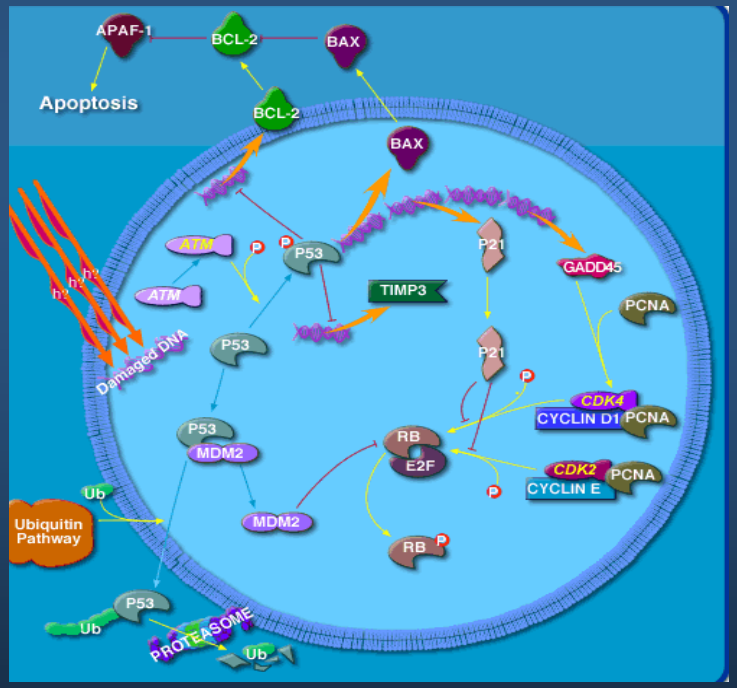


to interacting molecules...



To interaction networks at the cellular scale…





Celllular networks are usually described by simple mass-action kinetics

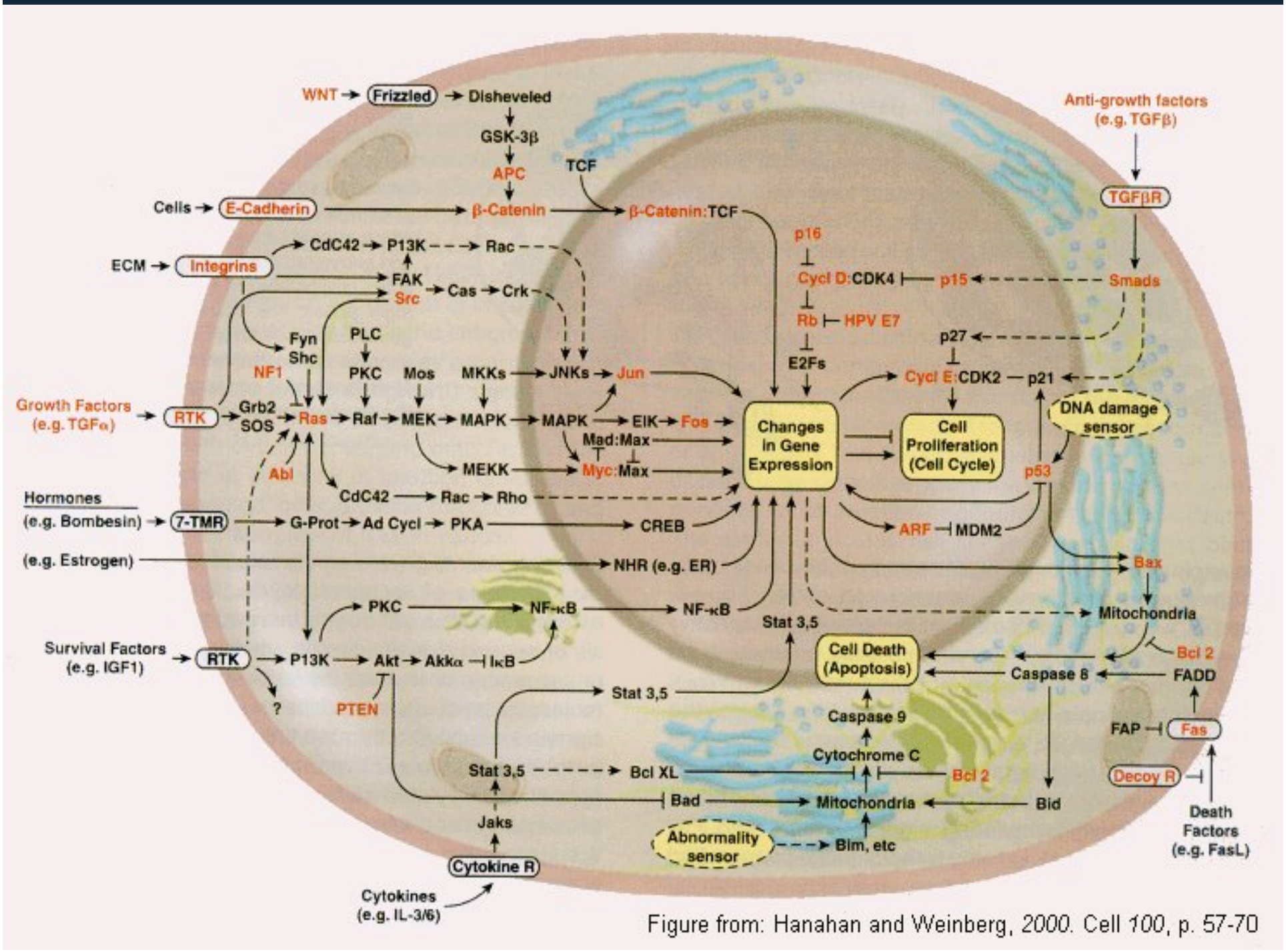
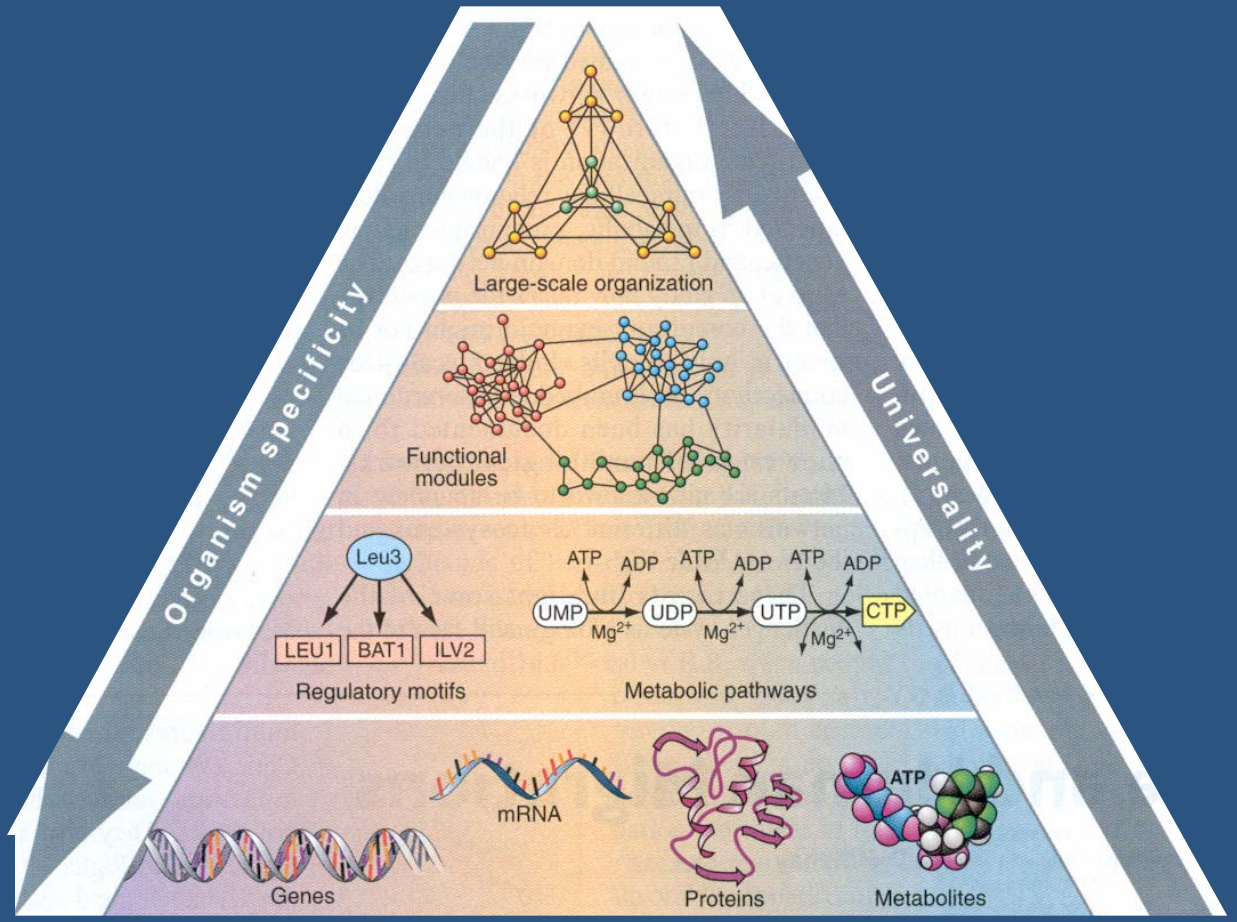


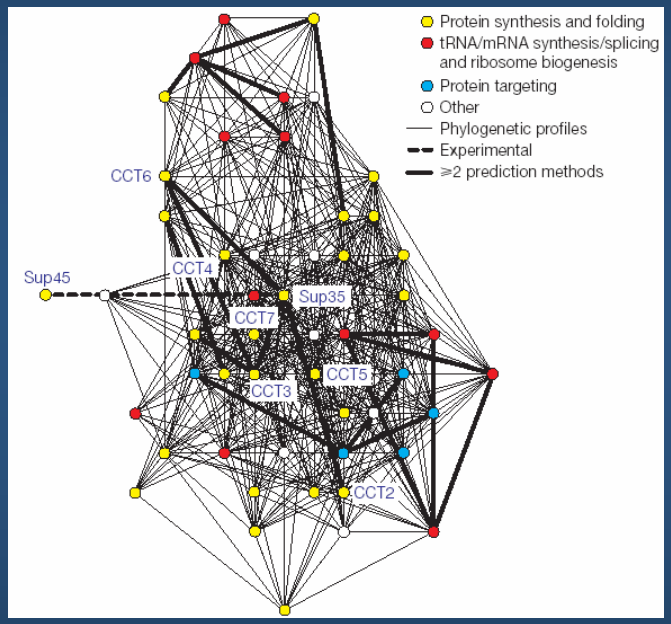
Figure from: Hanahan and Weinberg, *2000*. Cell *100*, p. 57-70

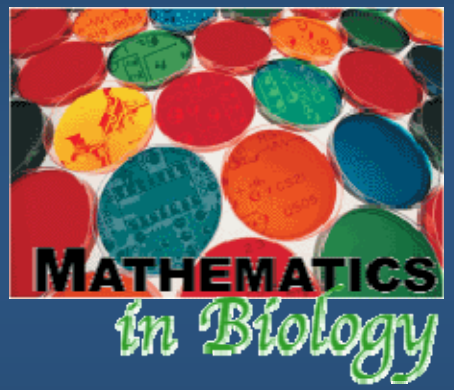
Life's complexity pyramid



Oltvai & Barabasi, *Science* 298, 763-764, 2002.

Interaction networks – at all scales





In a special collection of articles published beginning 6 February 2004, *Science* Magazine and its online companion sites team up to explore the interface between mathematics and biology

Proteomics

Examination of all proteins encoded by a given genome

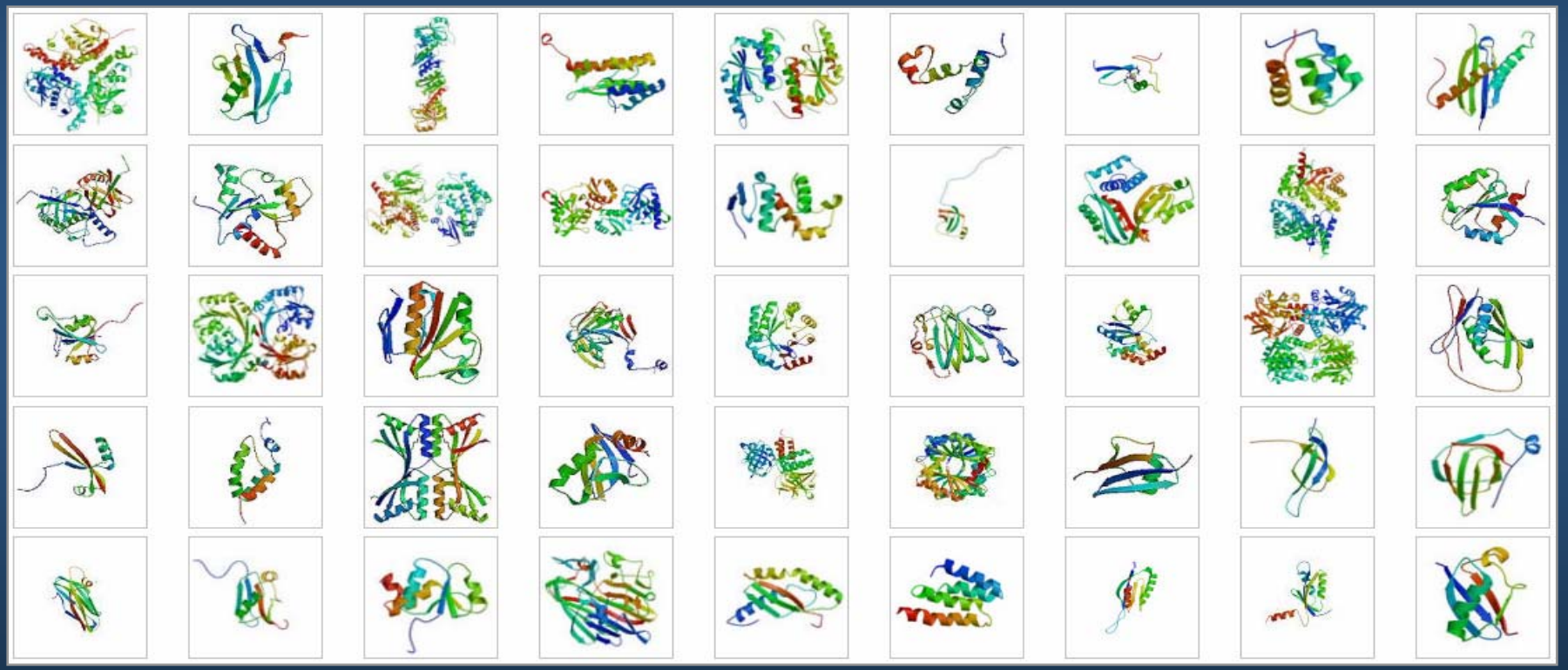
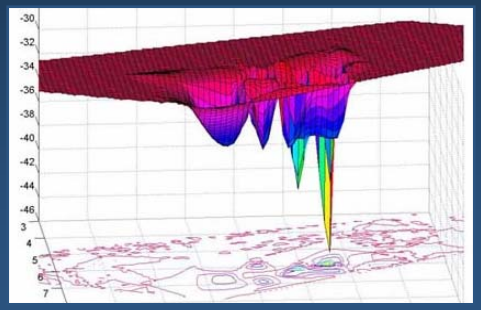


Figure: courtesy of Mark Gerstein 2003, Yale U

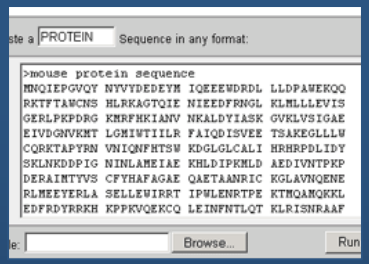
‘Protein folding problem’

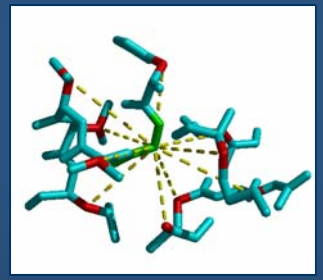
Sequence ----------> Structure

Bioinformatics. Sequence alignments



Modelling and simulations





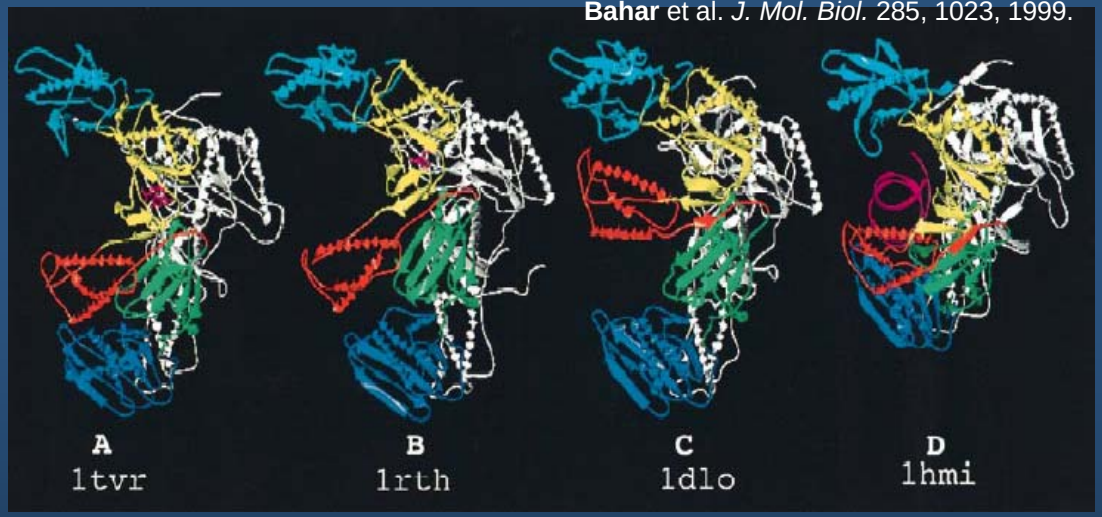
Function

Fundamental paradigm: Sequence encodes structure; structure encodes function

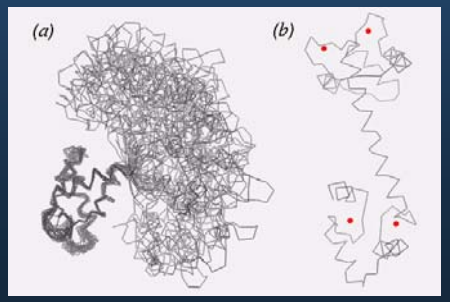
Stuctures suggest mechanisms of function

A. Comparison of static structures available in the PDB for the same protein in different form has been widely used as an ***indirect*** method of inferring dynamics.

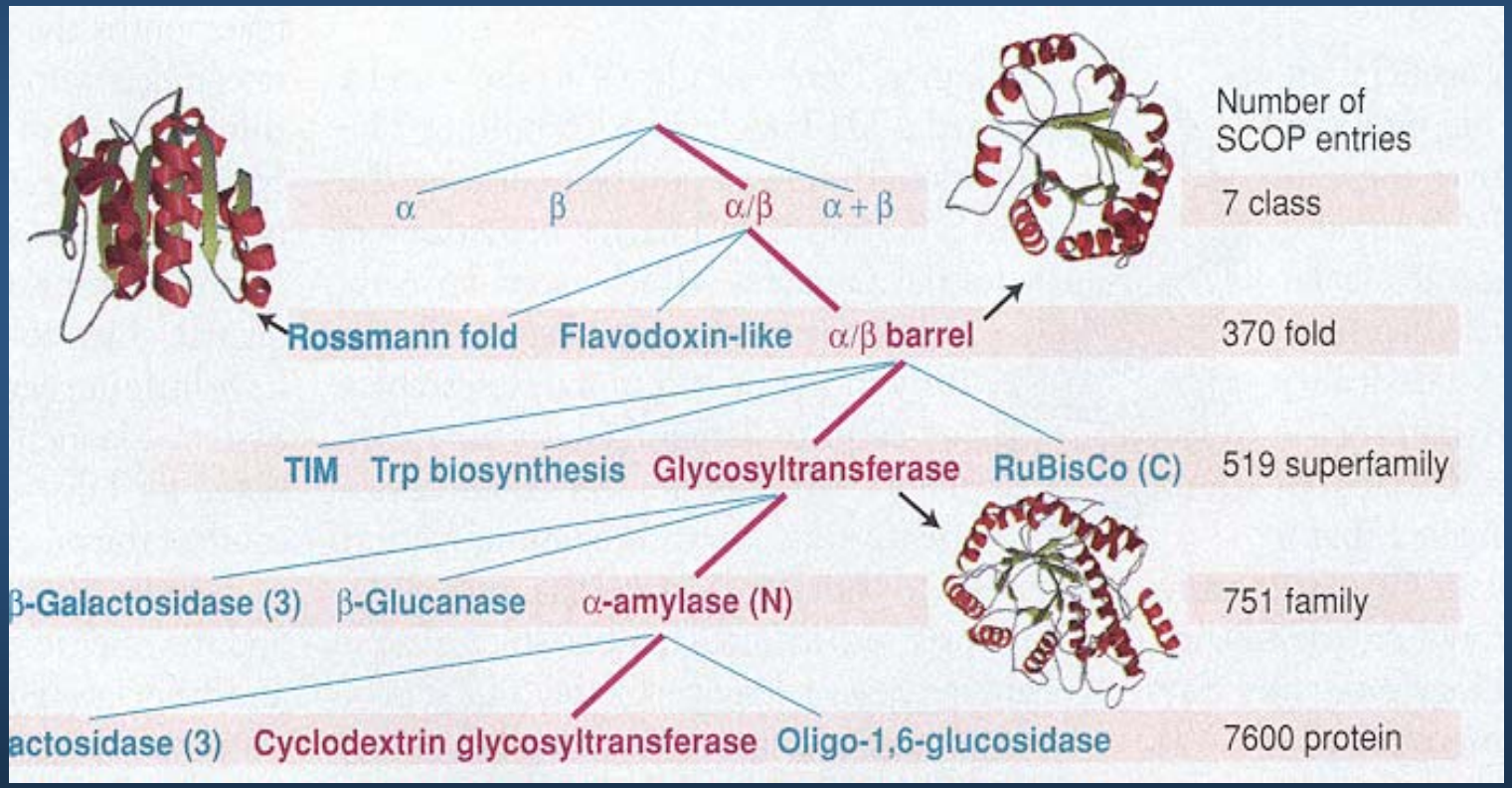
Bahar et al. *J. Mol. Biol.* 285, 1023, 1999.



B. NMR structures provide information on fluctuation dynamics



Classification of structural data (SCOP)



Pennisi, E. (1998) Science 279, 978; Hubbard et al. (1999) Nucleic Acids Res 254.

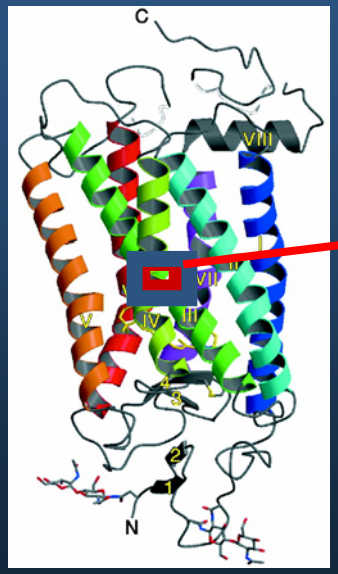
PROTEINS

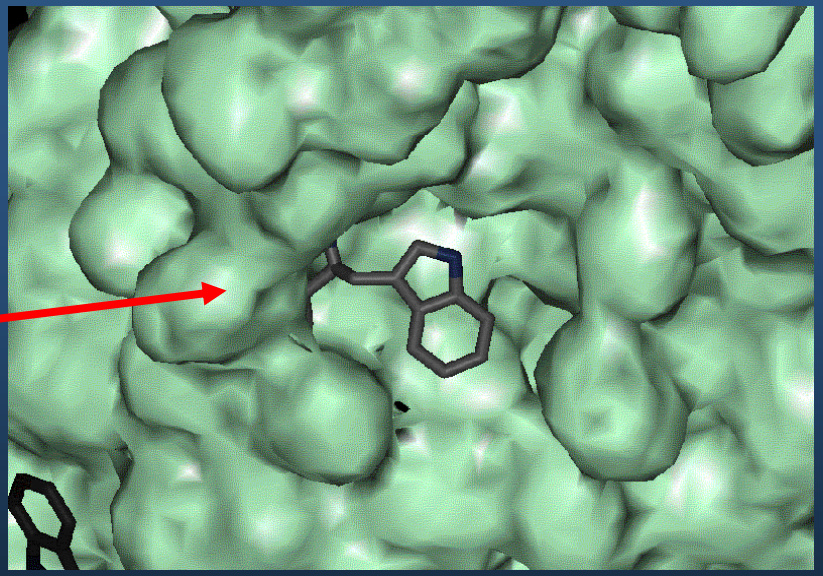
Sequence → Structure

Primary Sequence

```
MNGTEGPNFY VPFSNKTGVV RSPFEAPQYY LAEPWQFSML AAYMFLLIML GFPINFLTLY VTVQHKKLRT
PLNYILLNLA VADLFMVFGG FTTTLYTSLH GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC
KPMSNFRFGE NHAIMGVAFT WVMALACAAP PLVGWSRYIP EGMQCSCGID YYTPHEETNN ESFVIYMFVV
HFIIPLIVIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV TRMVIIMVIA FLICWLPYAG VAFYIFTHQG
SDFGPIFMTI PAFFAKTSAV YNPVIYIMMN KQFRNCMVTT LCCGKNPLGD DEASTTVSKT ETSQVAPA
```

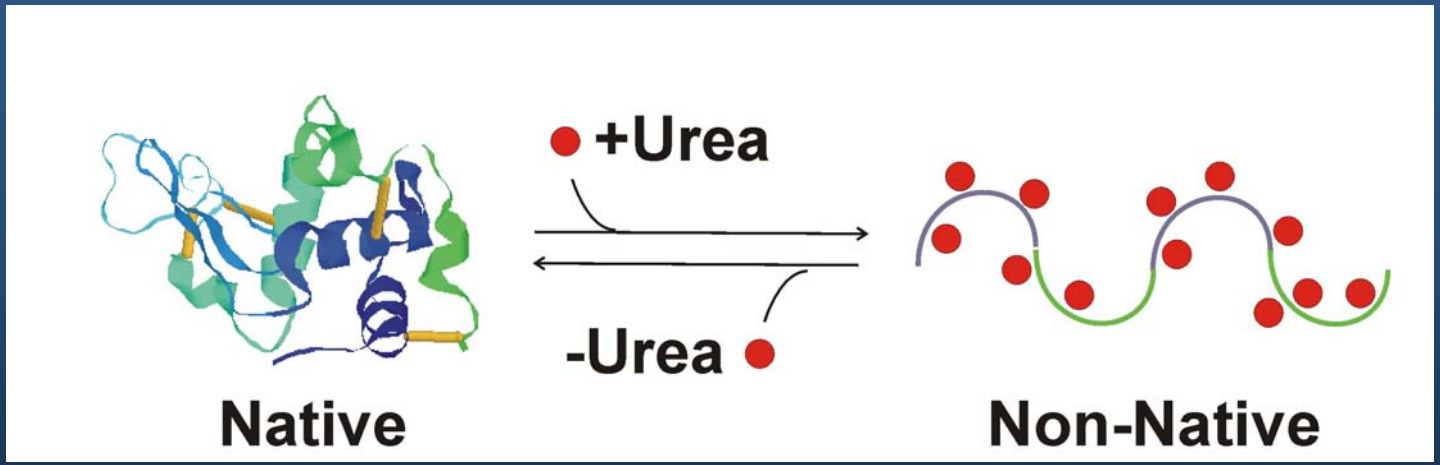
3D Structure





FACTS:

- Each sequence folds into a **unique** structure – native structure
- Proteins are functional only in their native state
- Sequence → structure mapping is not yet understood
- Folding is reversible – unfolding and re-folding is possible



Protein folding problem:

"Predicting 3-dimensional structure from sequence"

- A unique folded structure (native conformation, native fold) is assumed by a given sequence, although infinitely many conformations can be accessed.
- Which? (Protein folding problem)
- How, why? (Folding kinetics)

Basic postulate:
Thermodynamic equilibrium → Global energy minimum