Statistical and Computational Issues in Ab Initio Protein Structure Prediction

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins University

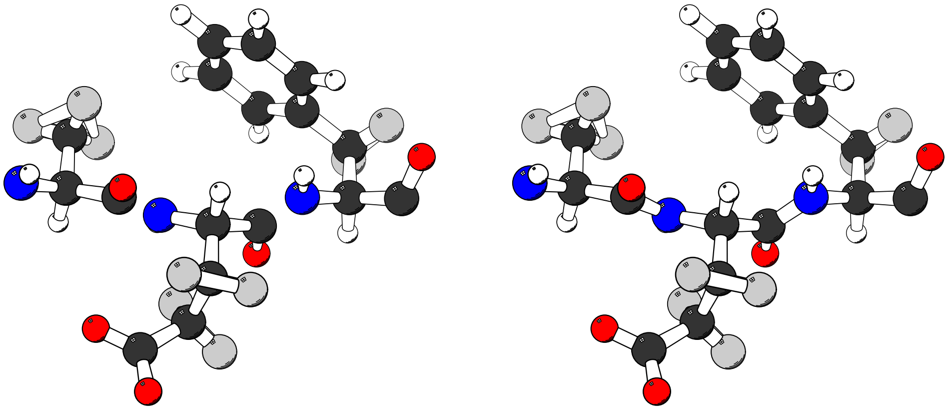
Email: ingo@jhu.edu

http://biosun01.biostat.jhsph.edu/∼iruczins

Collaborators

David Baker	University of Washington
Richard Bonneau	Structural GenomiX
Chris Bystroff	Rensselaer Polytechnic Institute
Charles Kooperberg	Fred Hutchinson Cancer Research Ctr
Kim Simons	Harvard University
Charlie Strauss	Los Alamos National Laboratory
Jerry Tsai	Texas A&M

What are Proteins?

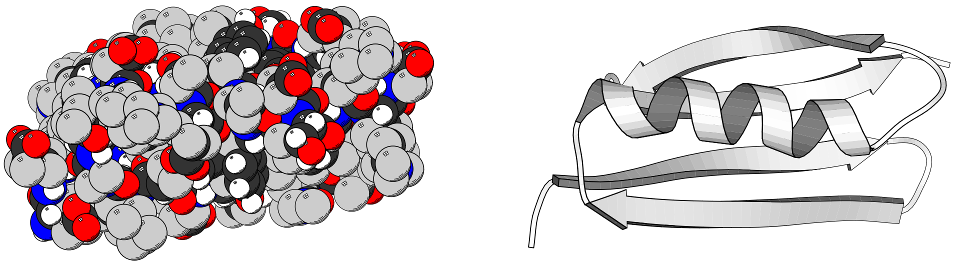


Without peptide bonds

With peptide bonds

The building blocks of proteins are amino acids.

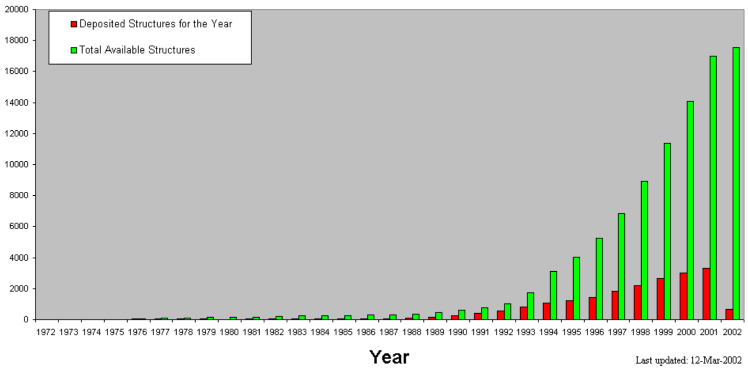
2D and 3D Protein Structure



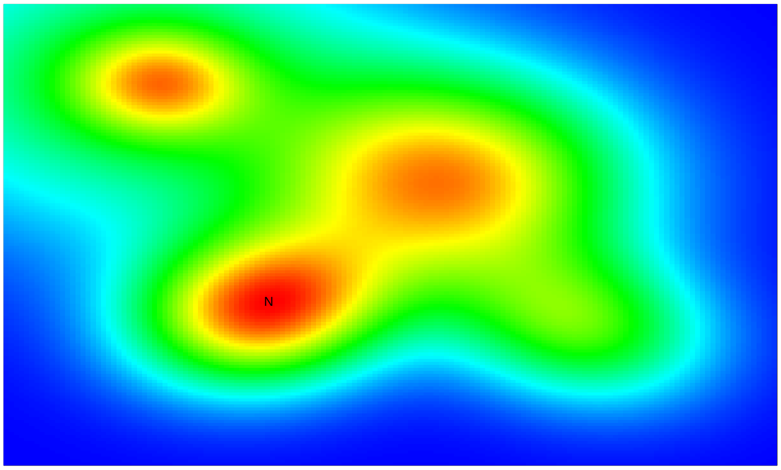
Both figures show the same protein, highlighting the tertiary and secondary structure.

Motivation

- What are proteins? Why do we care about them?
- Why do we care about protein structure?
- Why do we need to predict protein structures?
- How does the computational approach work?



Energy Landscape



The free energy of a structure changes with its geometry.

A Scoring Function for Ab Initio Protein Folding

$$P(\text{structure}|\text{sequence}) \propto P(\text{sequence}|\text{structure}) \times P(\text{structure})$$

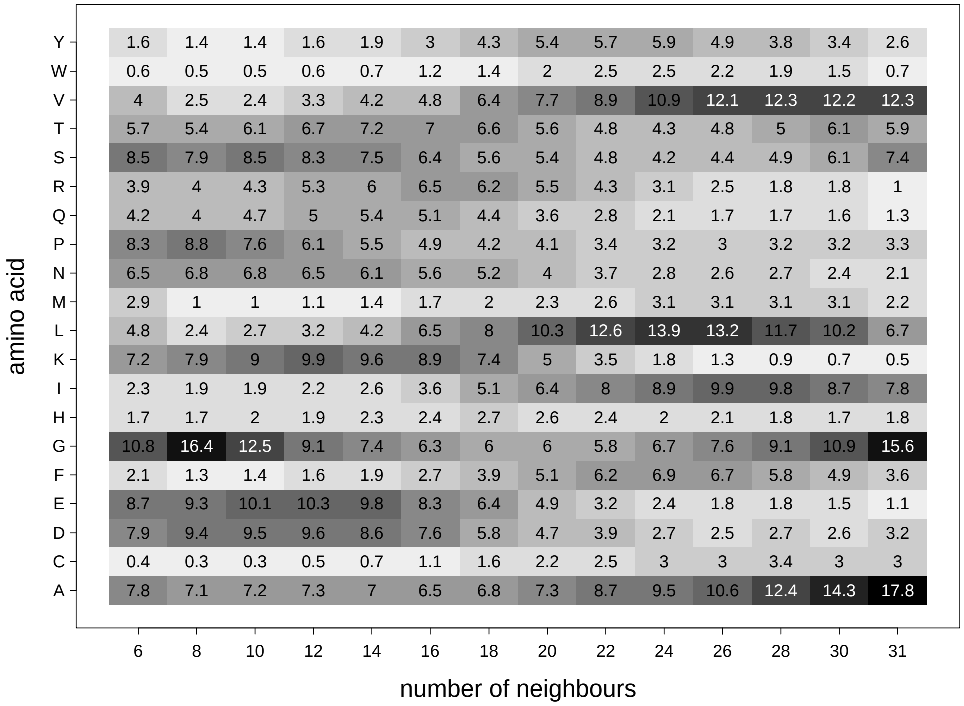
Sequence dependent:

- hydrophobic burial
- residue pair interaction

Sequence independent:

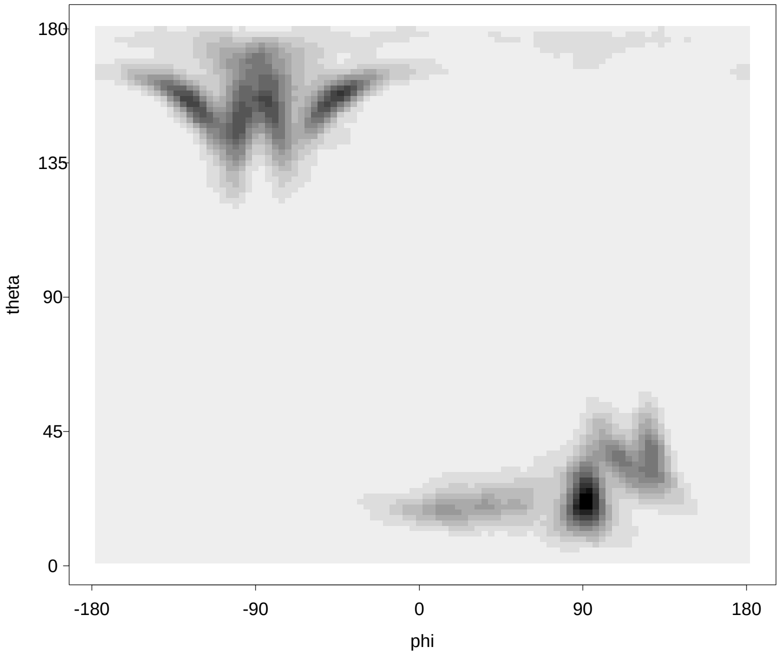
- helix-strand packing
- strand-strand packing
- sheet configurations
- vdW interactions

Hydrophobic Burial

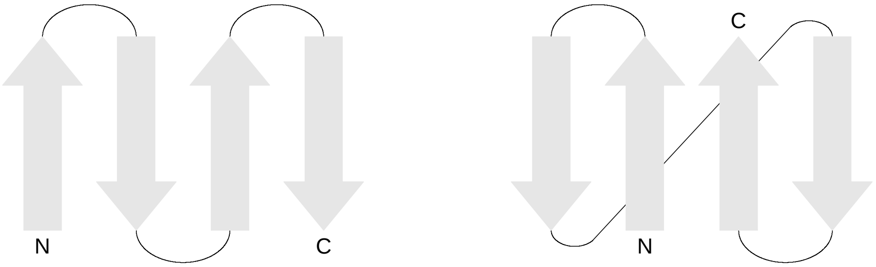


amino acid	6	8	10	12	14	16	18	20	22	24	26	28	30	31
Y	1.6	1.4	1.4	1.6	1.9	3	4.3	5.4	5.7	5.9	4.9	3.8	3.4	2.6
W	0.6	0.5	0.5	0.6	0.7	1.2	1.4	2	2.5	2.5	2.2	1.9	1.5	0.7
V	4	2.5	2.4	3.3	4.2	4.8	6.4	7.7	8.9	10.9	12.1	12.3	12.2	12.3
T	5.7	5.4	6.1	6.7	7.2	7	6.6	5.6	4.8	4.3	4.8	5	6.1	5.9
S	8.5	7.9	8.5	8.3	7.5	6.4	5.6	5.4	4.8	4.2	4.4	4.9	6.1	7.4
R	3.9	4	4.3	5.3	6	6.5	6.2	5.5	4.3	3.1	2.5	1.8	1.8	1
Q	4.2	4	4.7	5	5.4	5.1	4.4	3.6	2.8	2.1	1.7	1.7	1.6	1.3
P	8.3	8.8	7.6	6.1	5.5	4.9	4.2	4.1	3.4	3.2	3	3.2	3.2	3.3
N	6.5	6.8	6.8	6.5	6.1	5.6	5.2	4	3.7	2.8	2.6	2.7	2.4	2.1
M	2.9	1	1	1.1	1.4	1.7	2	2.3	2.6	3.1	3.1	3.1	3.1	2.2
L	4.8	2.4	2.7	3.2	4.2	6.5	8	10.3	12.6	13.9	13.2	11.7	10.2	6.7
K	7.2	7.9	9	9.9	9.6	8.9	7.4	5	3.5	1.8	1.3	0.9	0.7	0.5
I	2.3	1.9	1.9	2.2	2.6	3.6	5.1	6.4	8	8.9	9.9	9.8	8.7	7.8
H	1.7	1.7	2	1.9	2.3	2.4	2.7	2.6	2.4	2	2.1	1.8	1.7	1.8
G	10.8	16.4	12.5	9.1	7.4	6.3	6	6	5.8	6.7	7.6	9.1	10.9	15.6
F	2.1	1.3	1.4	1.6	1.9	2.7	3.9	5.1	6.2	6.9	6.7	5.8	4.9	3.6
E	8.7	9.3	10.1	10.3	9.8	8.3	6.4	4.9	3.2	2.4	1.8	1.8	1.5	1.1
D	7.9	9.4	9.5	9.6	8.6	7.6	5.8	4.7	3.9	2.7	2.5	2.7	2.6	3.2
C	0.4	0.3	0.3	0.5	0.7	1.1	1.6	2.2	2.5	3	3	3.4	3	3
A	7.8	7.1	7.2	7.3	7	6.5	6.8	7.3	8.7	9.5	10.6	12.4	14.3	17.8

Strand-strand Interaction

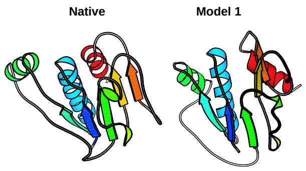


Beta-Sheet Motifs

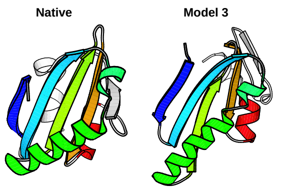


Two possible motifs for 4-stranded sheets.

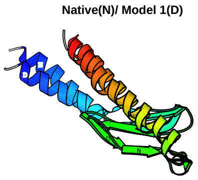
Structure Predictions



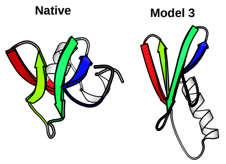
T087-A: PPase (Domain 1: 2-192)



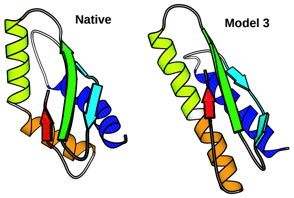
T087-B: PPase (Domain 2: 202-307)



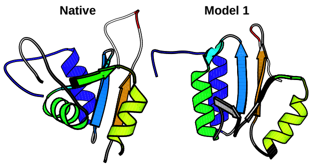
T091: Hypothetical Protein HI0442



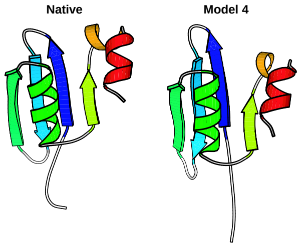
T105: Protein Sp100b



T110: Ribosome Binding Factor A

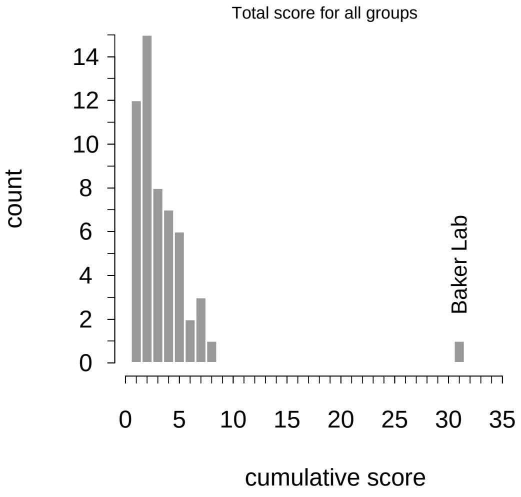


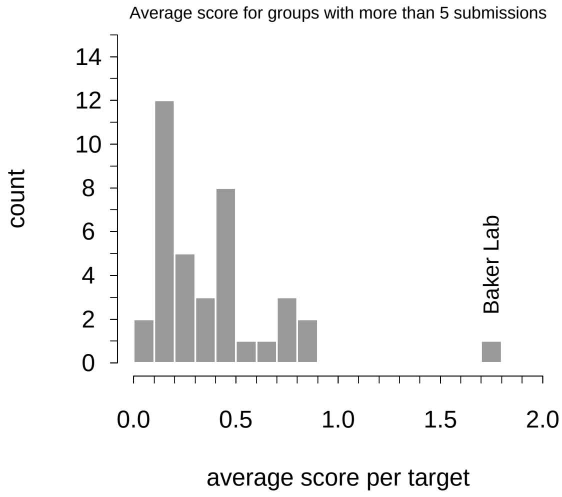
T116-A: MutS (Domain 1: 3-106)



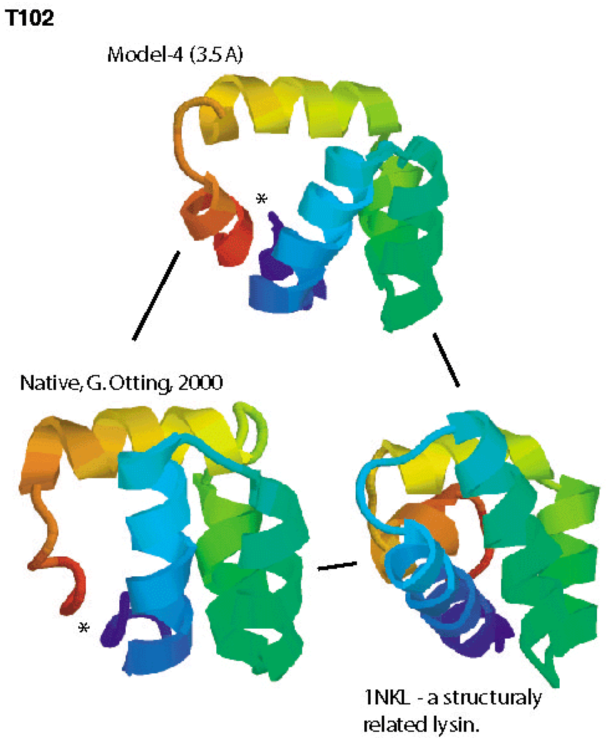
T116-B: MutS (Domain 2: 128-196)

CASP4 Results


Total score for all groups


Average score for groups with more than 5 submissions

Functional Annotation



References

http://biosun01.biostat.jhsph.edu/∼iruczins

- Bonneau R, Tsai J, Ruczinski I, Baker D:
  Functional inferences from blind ab initio protein structure predictions,
  Journal of Structural Biology 134(2-3):186-90, 2001.
- Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss C, Baker D:
  Rosetta in CASP4: progress in ab protein initio structure prediction,
  Proteins: Structure, Function and Genetics 45:S5, 2001.
- Ruczinski I, Kooperberg C, Bonneau R, Baker D:
  Distributions of beta sheets in proteins with application to structure prediction,
  Proteins: Structure, Function and Genetics (to appear).
- Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D:
  Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins,
  Proteins 34(1):82-95, 1999.