Using PERL in Projects Related to Protein Structure Prediction

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins University

Email: ingo@jhu.edu

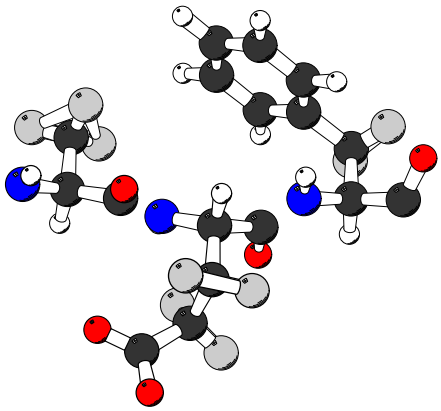
http://biostat.jhsph.edu/∼iruczins

Collaborators

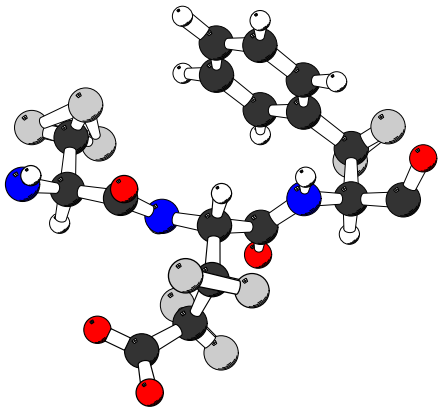
David Baker	University of Washington
Richard Bonneau	Institute for Systems Biology
Chris Bystroff	Rensselaer Polytechnic Institute
Charles Kooperberg	Fred Hutchinson Cancer Research Ctr
Carol Rohl	UC Santa Cruz
Kim Simons	Harvard University
Charlie Strauss	Los Alamos National Laboratory
Jerry Tsai	Texas A&M

What are Proteins?

Without peptide bonds

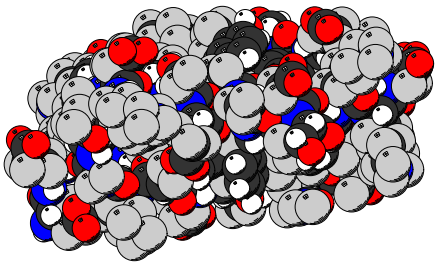


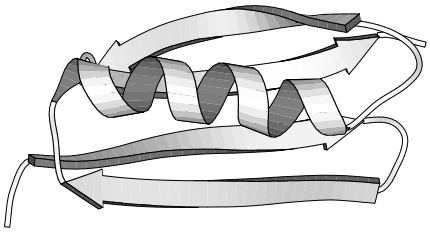
With peptide bonds



The building blocks of proteins are amino acids.

2D and 3D Protein Structure

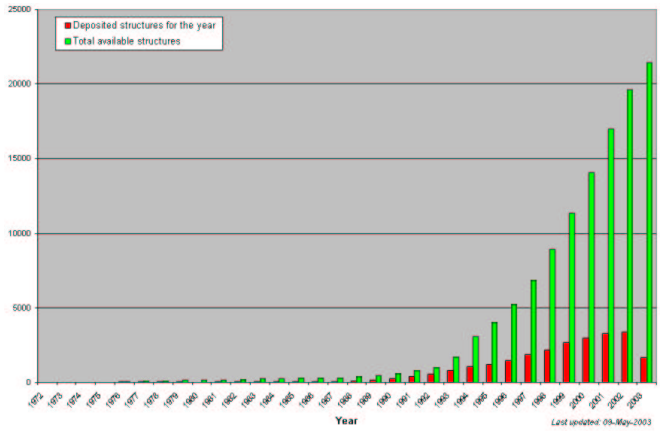




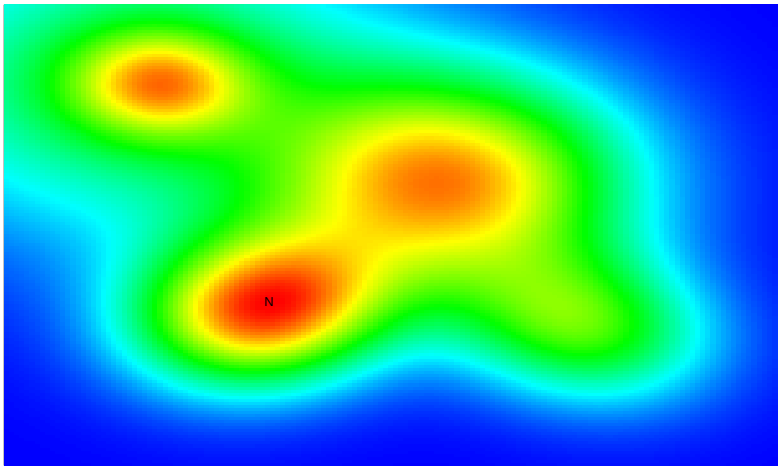
Both figures show the same protein, highlighting the tertiary and secondary structure.

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

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

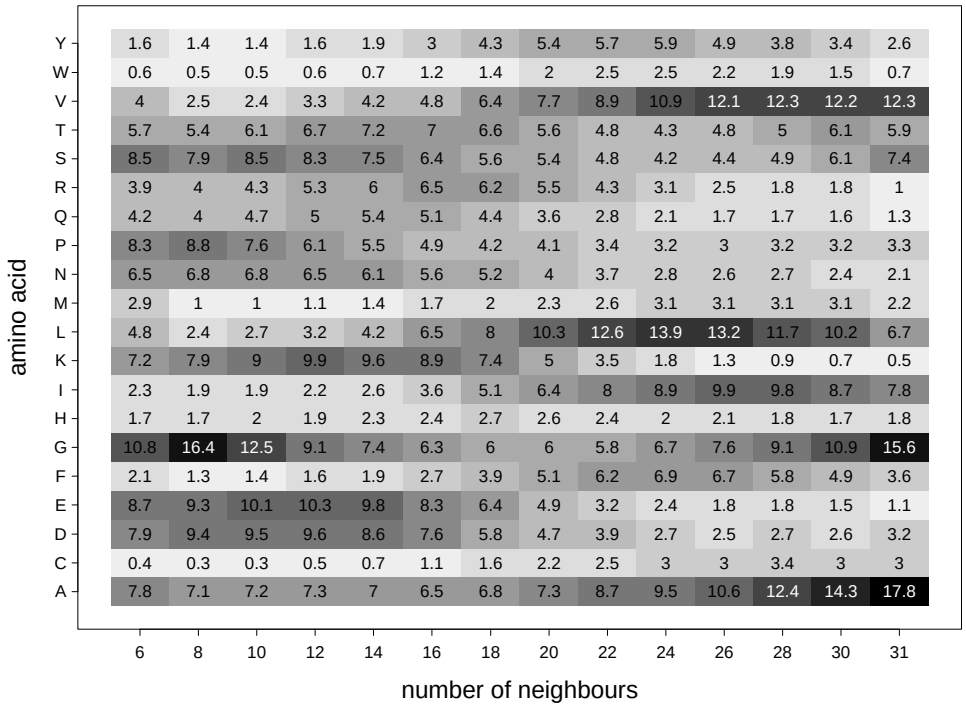
Sequence dependent:

- hydrophobic burial
- residue pair interaction

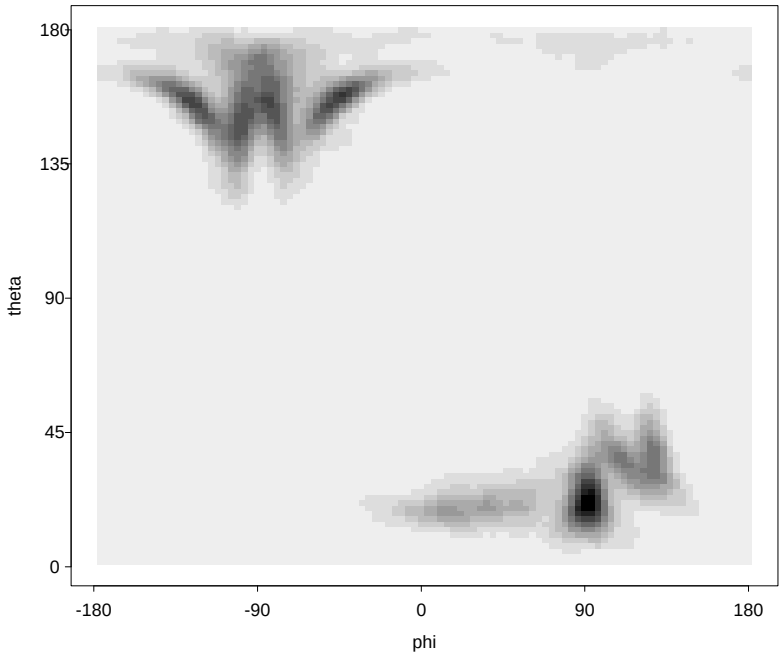
Sequence independent:

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

Hydrophobic Burial

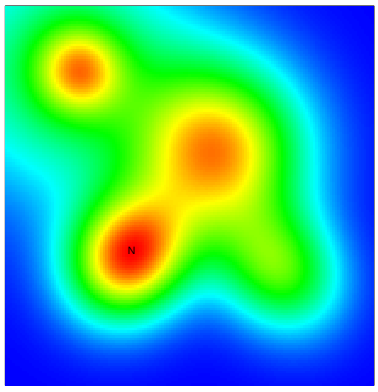


Strand-Strand Interaction

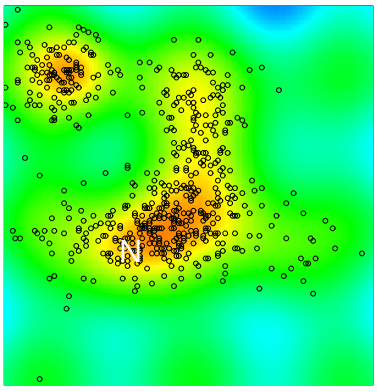


Energy Landscape (2)

True landscape

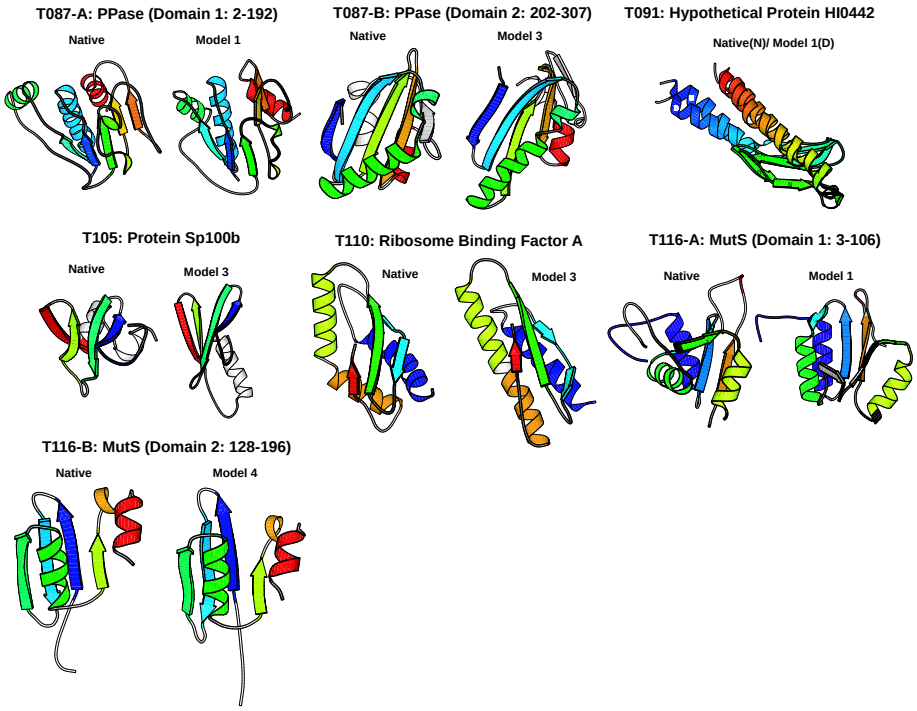


Our scoring function

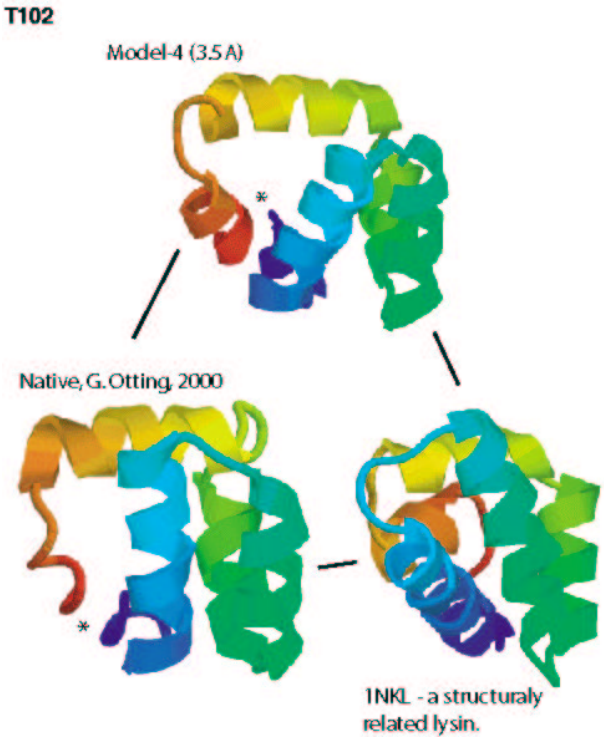


Decoys tend to cluster near low energy states

Structure Predictions



Functional Annotation



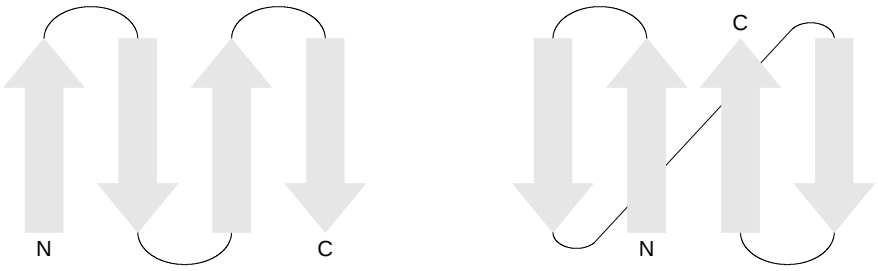
Perl Scripts

Perl scripts are used all over the place! Some applications include:

- Accessing and transforming sequence data from local and remote databases
- Comparing amino acid sequences
- Extracting information from sequence alignments
- Comparing protein structures
- Parsing structures (sequences) into domains
- Creating decoy files and managing them
- . . .

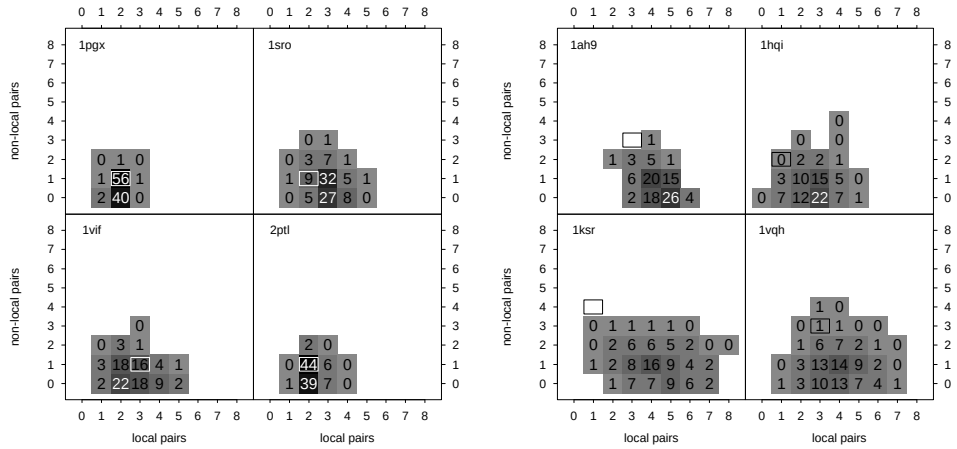
Check out the bioperl tutorial (http://bio.perl.org/).

Beta-Sheet Motifs



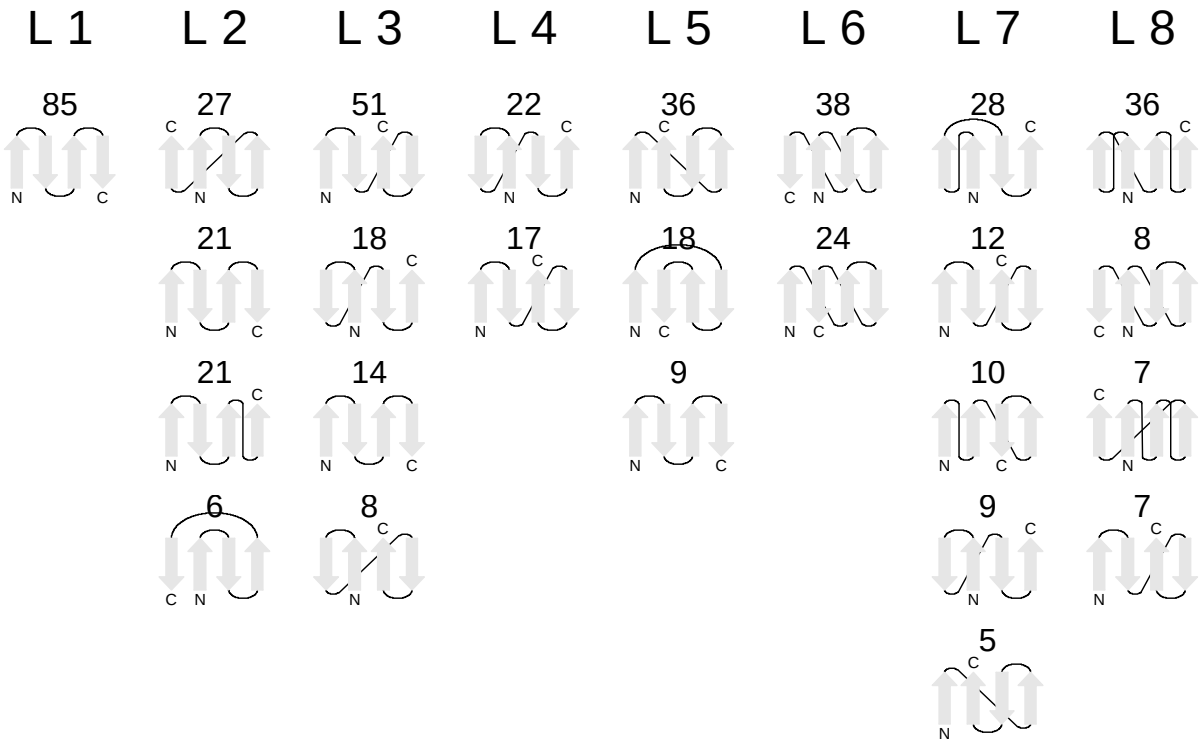
Two possible motifs for 4-stranded sheets.

Bias towards Local Conformations



Local structures are easier to generate than non-local structures.

Likely Sheet Topologies



References

Publications:

- Simons KT, Ruczinski I, Kooperberg C, Fox B, Bystroff C, and Baker D (1999), *Improved Recognition of Native-like Protein Structures using a Combination of Sequence-dependent and Sequence-independent Features of Proteins*, Proteins: Structure, Function and Genetics 34 (1) 82-95.
- Bonneau R, Tsai J, Ruczinski I, and Baker D (2001), *Rosetta in CASP4: Progress in Ab Initio Protein Structure Prediction*, Proteins: Structure, Function and Genetics 45 (S5), 119-126.
- Ruczinski I, Kooperberg C, Bonneau R, and Baker D (2002), *Distributions of Beta Sheets in Proteins with Application to Structure Prediction*, Proteins: Structure, Function and Genetics 48, 85-97.
- Bonneau R, Tsai J, Ruczinski I, and Baker D (2002), *Contact Order and Ab Initio Protein Structure Prediction*, Protein Science 11 (8), 1937-1944.

References

Web Pages:

My Home Page	http://www.biostat.jhsph.edu/∼iruczins/
Class Notes and Stuff	http://www.biostat.jhsph.edu/∼iruczins/teaching/misc/misc.html
The Protein Data Bank	http://www.rcsb.org/pdb/
The Dunbrack Lab	http://www.fccc.edu/research/labs/dunbrack/
The Baker Lab	http://depts.washington.edu/bakerpg/
The CASP Page	http://predictioncenter.llnl.gov/
The Bioperl Page	http://bioperl.org/
The Beta Sheet Page	http://www.biostat.jhsph.edu/∼iruczins/sheets/sheets.html