Tutorial on protein structure prediction

Kevin Karplus

`karplus@soe.ucsc.edu`

Biomolecular Engineering Department

Undergraduate and Graduate Director, Bioinformatics

University of California, Santa Cruz

Outline of Talk

- What is Bioengineering? Biomolecular Engineering? Bioinformatics?
- What is a protein?
- The folding problem and variants on it:
  - Local structure prediction
  - Fold recognition
  - Comparative modeling
  - “Ab initio” methods
  - Contact prediction
- Protein Design

What is Bioengineering?

Three concentrations:

- Biomolecular
  - Drug design
  - Biomolecular sensors
  - Nanotechnology
  - Bioinformatics
- Rehabilitation
- Bioelectronics

What is Bioengineering?

Three concentrations:

- Biomolecular
- Rehabilitation
    - Systems to held individuals with special needs
    - Cell-phone-based systems to reach large numbers of people.
    - Novel hardware to assist the blind
    - Robotics for rehabilitation and surgery applications.
- Bioelectronics

What is Bioengineering?

Three concentrations:

- Biomolecular
- Rehabilitation
- Bioelectronics
  - Implantable devices
  - Interfacing between organisms and electronics
  - Artificial retina project

What to take early

- Mathematics
- Chemistry and then biology
- Introductory bioengineering courses:
  - BME80G, Bioethics (F)
  - BME5, Intro to Biotechnology (W, S)
  - CMPE80A: Universal Access: Disability, Technology, and Society (W, S)
- Declare your major immediately!!
  You can always change to another one later.
  Bioengineering is one of the most course-intensive majors on campus. Many courses have prerequisites.
  It's important to get advising office and faculty advice early.

What is Biomolecular Engineering?

Engineering **with**, **of**, or **for** biomolecules. For example,

with: using proteins (or DNA, RNA, . . .) as sensors or for self-assembly.

of: protein engineering—designing or artificially evolving proteins to have particular functions

for: designing high-throughput experimental methods to find out what molecules are present, how they are structured, and how they interact.

What is Bioinformatics?

Bioinformatics: using computers and statistics to make sense out of the mountains of **data** produced by high-throughput experiments.

- Genomics: annotating important sequences in genomes.
- Phylogenetics: tree of life, ancestral genome reconstruction.
- Systems biology: discovering and modeling biological networks.
- Expression profiling: what genes are turned on under what conditions (DNA microarrays, RNAseq).
- Protein structure and function prediction.
- Proteomics: what proteins are present in a mixture.

What is a protein?

- There are many abstractions of a protein: a band on a gel, a string of letters, a mass spectrum, a set of 3D coordinates of atoms, a point in an interaction graph, ...
- For us, a protein is a long skinny molecule (like a string of letter beads) that folds up consistently into a particular intricate shape.
- The individual “beads” are amino acids, which have 6 atoms the same in each “bead” (the *backbone* atoms: N, H, CA, HA, C, O).
- The final shape is different for different proteins and is essential to the function. The protein shapes are important, but are expensive to determine experimentally.

Visualizing Proteins

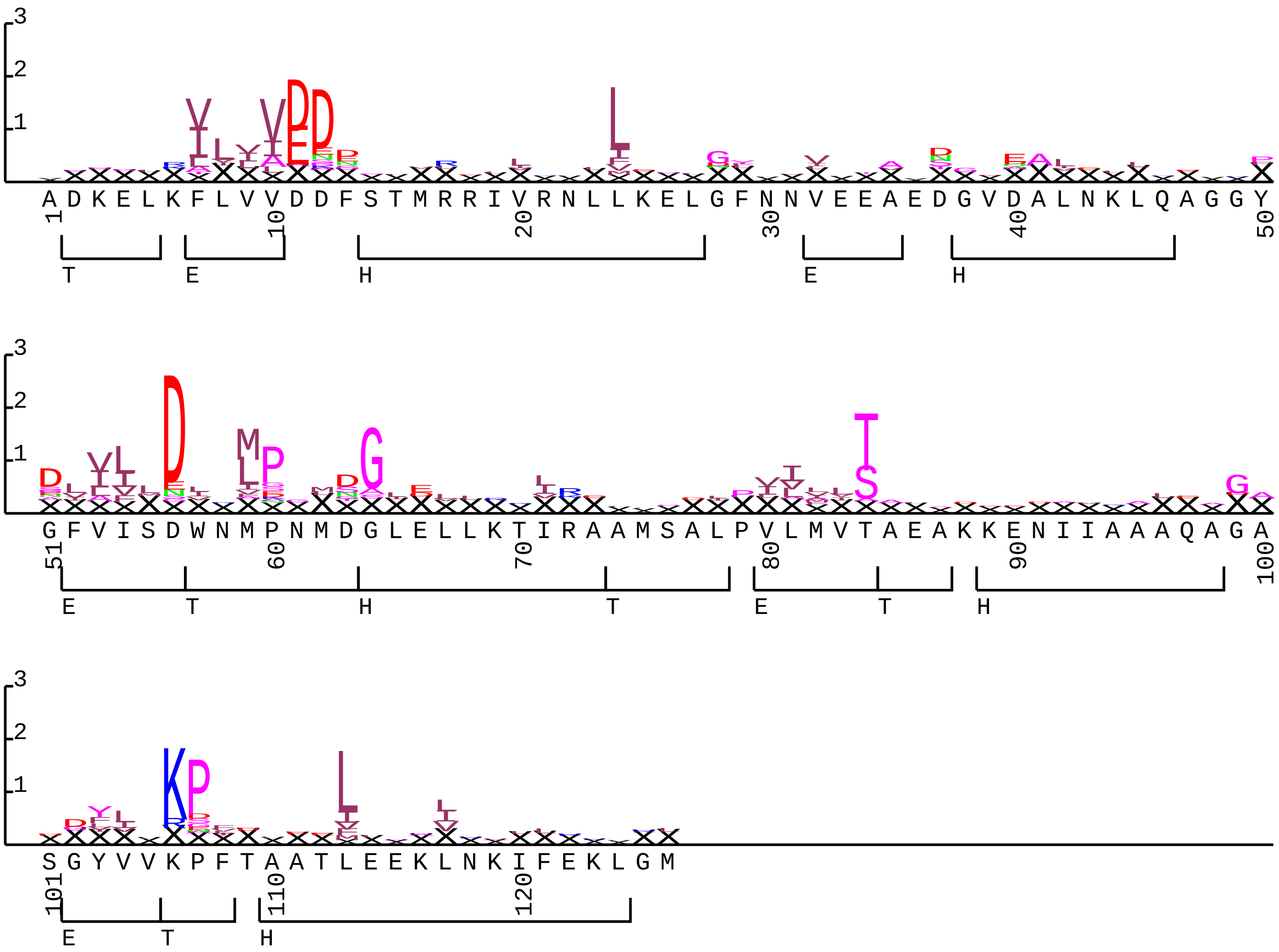
There are many ways to look at proteins:

- Strings of letters.
- Sequence logos: letters plus conservation information.
- Plastic models of structure.
- Computer visualization of structure (rasmol, pymol, vmd, jmol, molmol, . . .)

Sequence logos (MSA)

Summarize multiple alignment for 1jbeA:

nostruct-align/1jbeA.t06 w0.5



DEMO visualization

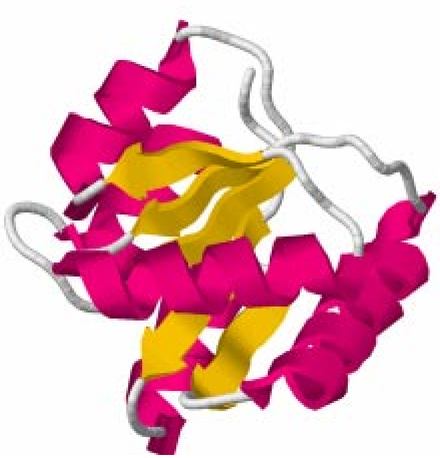
- Demonstrate protein backbone using Darling Models
- Demonstrate different views using Rasmol (or other viewer)

Folding Problem

The *Folding Problem*:
If we are given a sequence of amino acids (the letters on a string of beads), can we predict how it folds up in 3-space?

```
>1jbeA Chemotaxis protein CHEY from E. coli
ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGY
GFVISDWNMPNMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQA
GASGYVVKPFTAATLEEKLNKIFEKLGM
```

↓



Too hard!

Fold-recognition problem

The *Fold-recognition Problem*:
Given a sequence of amino acids A (the *target* sequence) and a library of proteins with known 3-D structures (the *template* library),
figure out which templates A match best, and align the target to the templates.

- The backbone for the target sequence is predicted to be very similar to the backbone of the chosen template.

New-fold prediction

- What if there is *no* template we can use?
- We can try to generate many conformations of the protein backbone and try to recognize the most protein-like of them.
- Search space is huge, so we need a good conformation generator and a cheap cost function to evaluate conformations.

Secondary structure Prediction

- Instead of predicting the entire structure, we can predict local properties of the structure.
- What local properties do we choose?
- We want properties that are well-conserved through evolution, easily predicted, and useful for finding and aligning templates.
- One popular choice is a 3-valued helix/strand/other alphabet—we have investigated many others. Typically, predictors get about 80% accuracy on 3-state prediction.
- Many machine-learning methods have been applied to this problem, but the most successful are neural networks.

Contact prediction

- Try to predict which residues come close to each other.
- Ones close along the chain are easy (secondary structure prediction).
- Ones far apart along chain, but close in space, are hard to predict, but most useful.
- Correlated mutation is powerful indication of close residues.

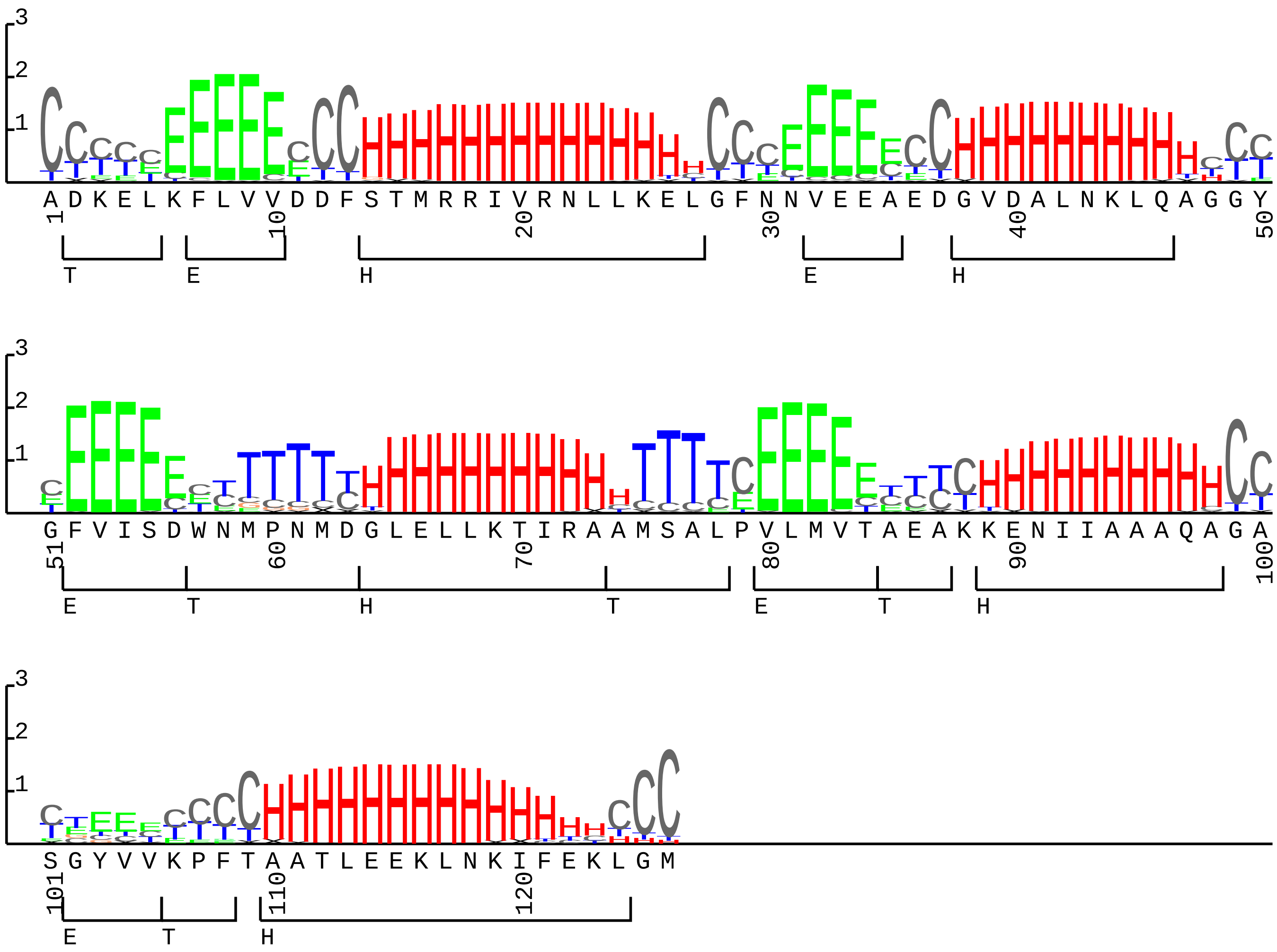
(Rational) Protein Design

- New direction for Karplus lab.
- Use neural nets to predict amino acids from local structure properties.
- Use Undertaker to build models.
- Use RosettaDesign (from Baker lab) to modify sequences.
- Use Undertaker, Rosetta, and Gromacs to validate that designed structure is good.
- Target applications: short proteins that mimic agouti-related protein (and other proteins that bind melanocortin receptors) but which do not have disulfide bridges.

Sequence logos (NN)

Summarize local structure prediction:

nostruct-align/1jbeA.t06 EBGHTL



CASP ~~Competition~~ Experiment

- Everything published in literature “works”
- CASP set up as true blind test of prediction methods.
- Sequences of proteins about to be solved released to prediction community.
- Predictions registered with organizers.
- Experimental structures compared with solution by assessors.
- “Winners” get papers in *Proteins: Structure, Function, and Bioinformatics*.

Overview of Prediction Method

- Look for homologs.
  - Homologs = proteins with common ancestral sequence.
  - Can’t really determine algorithmicly, so we look for “sufficiently similar” sequences.
- Make multiple sequence alignment (MSA).

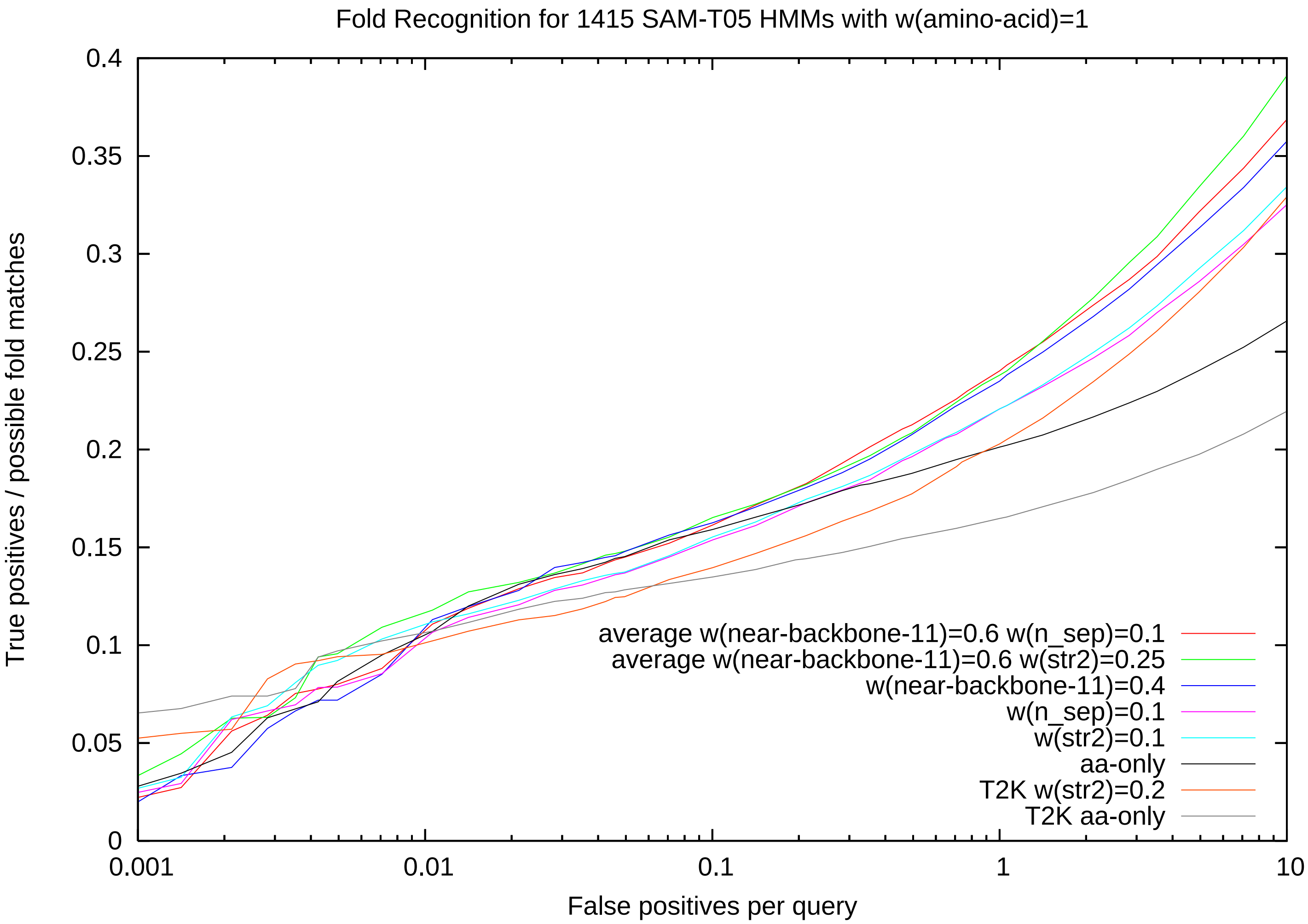
Overview of Prediction Method 2

- Use MSA to make local structure predictions.
- Use MSA (and local structure predictions) to make Hidden Markov Models (HMMs).
- Use HMMs to find and align proteins of known structure (PDB).
- Use model-building program to change alignments into 3D models.
- Clean up models (close gaps, rebuild loops, adjust sidechains, ...)
- Choose best model(s) (Model Quality Assessment).
- Maybe use contact predictions to select among models.

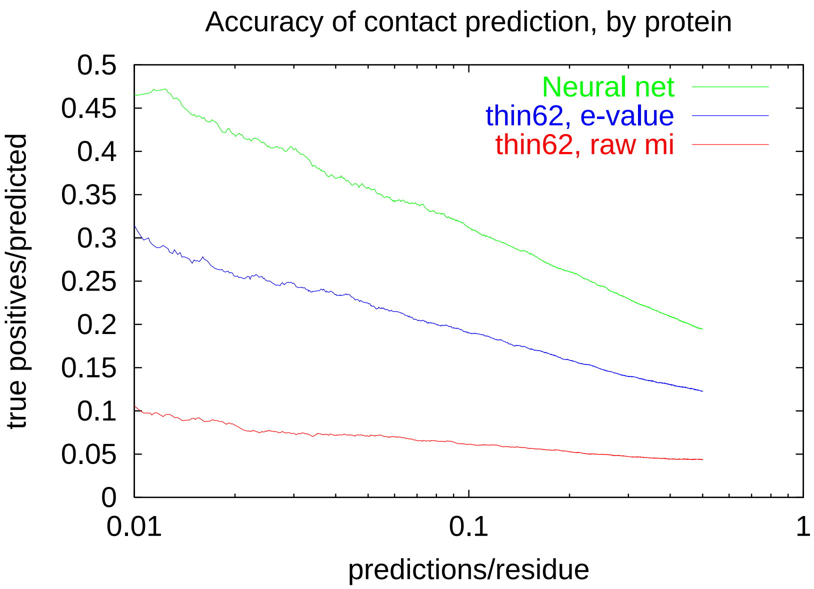
Contact Prediction Method

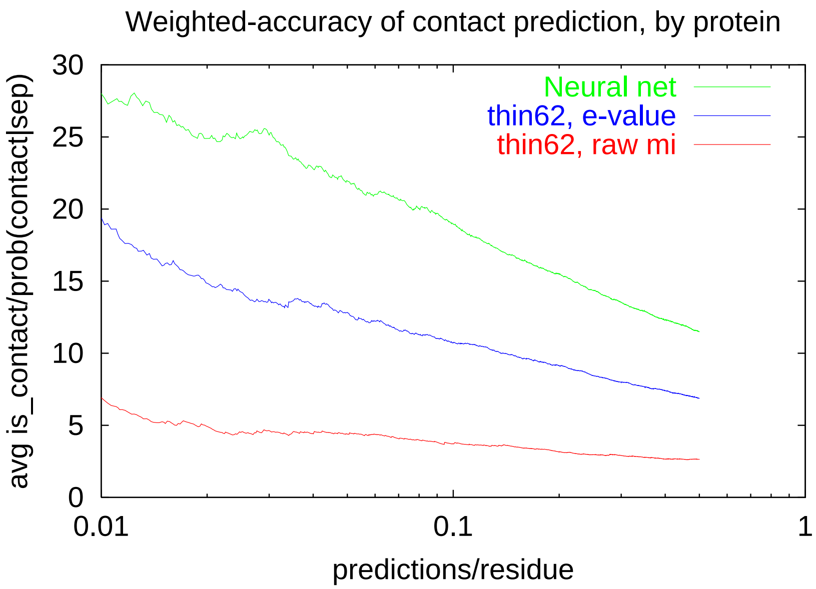
- Use mutual information between columns.
- Thin alignments aggressively (30%, 35%, 40%, 50%, 62%).
- Compute e-value for mutual info (correcting for small-sample effects).
- Compute rank of log(e-value) within protein.
- Feed log(e-values), log rank, contact potential, joint entropy, and separation along chain for pair, and amino-acid profile, predicted burial, and predicted secondary structure for each residue of pair into a neural net.

Fold recognition results



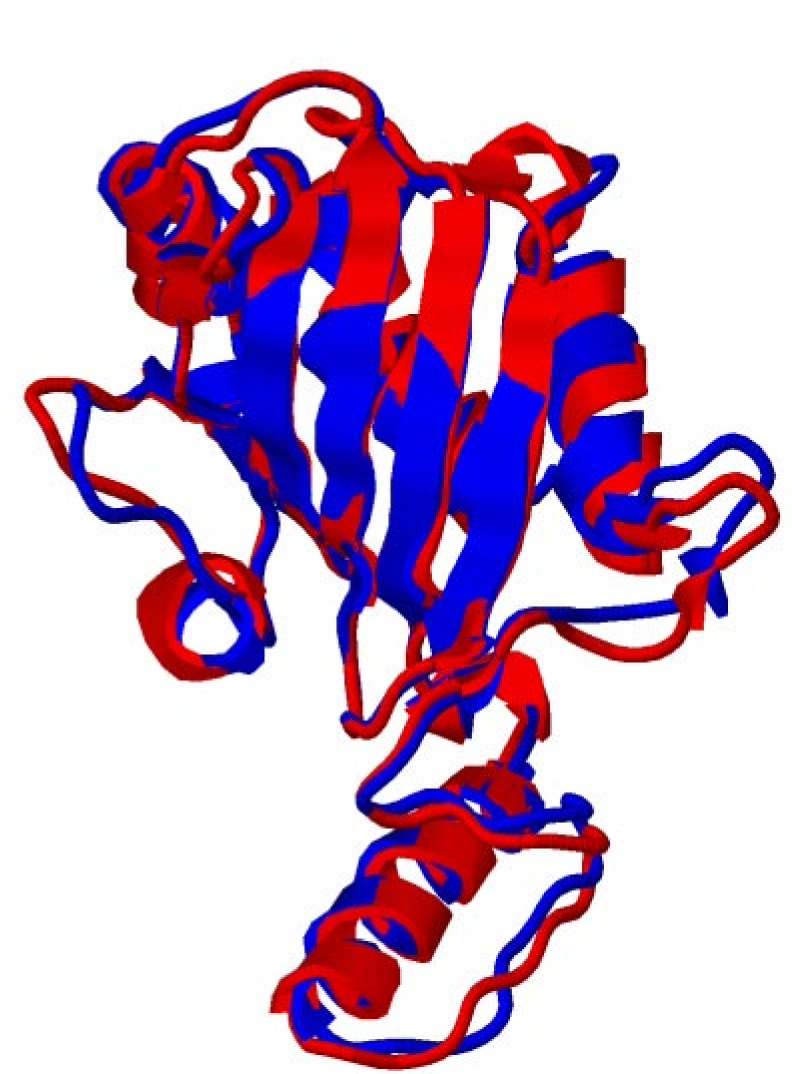
Contact prediction results





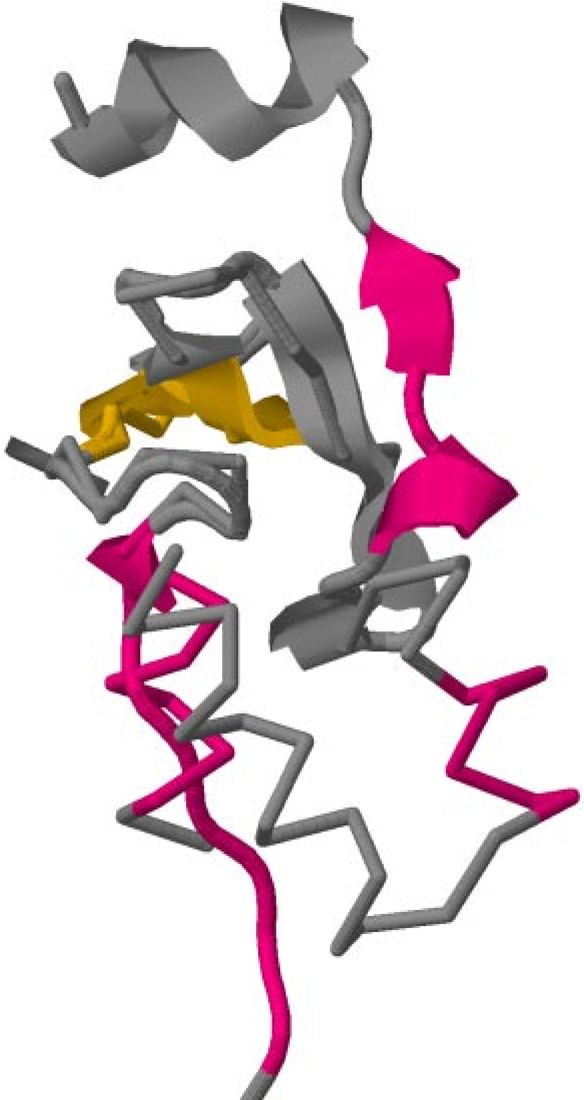
T0298 domain 2 (130–315)

RMSD= 2.468Å all-atom, 1.7567Å C_α, GDT=82.5%
best model 1 submitted to CASP7 (red=real)



Comparative modeling: T0348

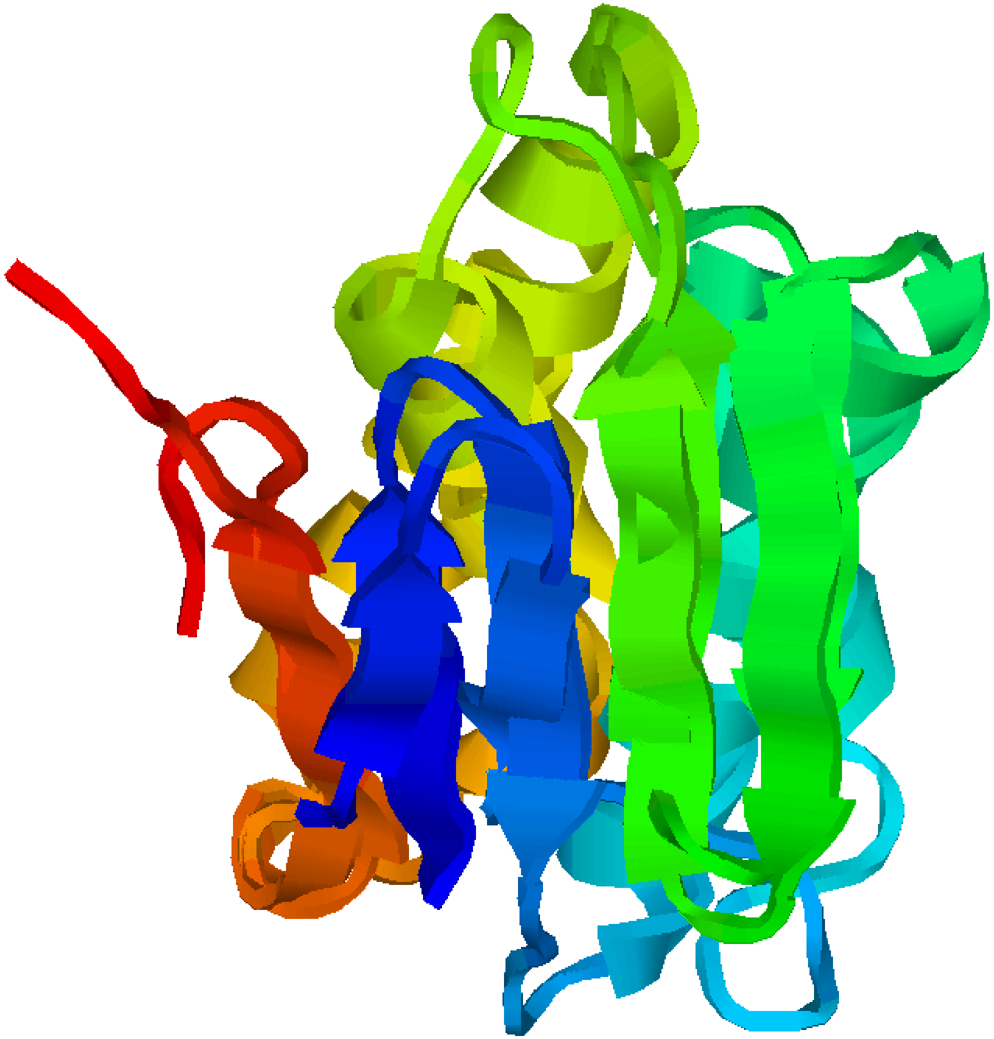
RMSD= 11.8 Å C_α, GDT=58.2% (cartoon=real)
best model 1 by CASP7 GDT, Robetta1 slightly better.



Target T0201 (NF, CASP6)

- We tried forcing various sheet topologies and selected 4 by hand.
- Model 1 has right topology (5.912Å all-atom, 5.219Å C_α).
- Unconstrained cost function not good at choosing topology (two strands curled into helices).
- Helices were too short.

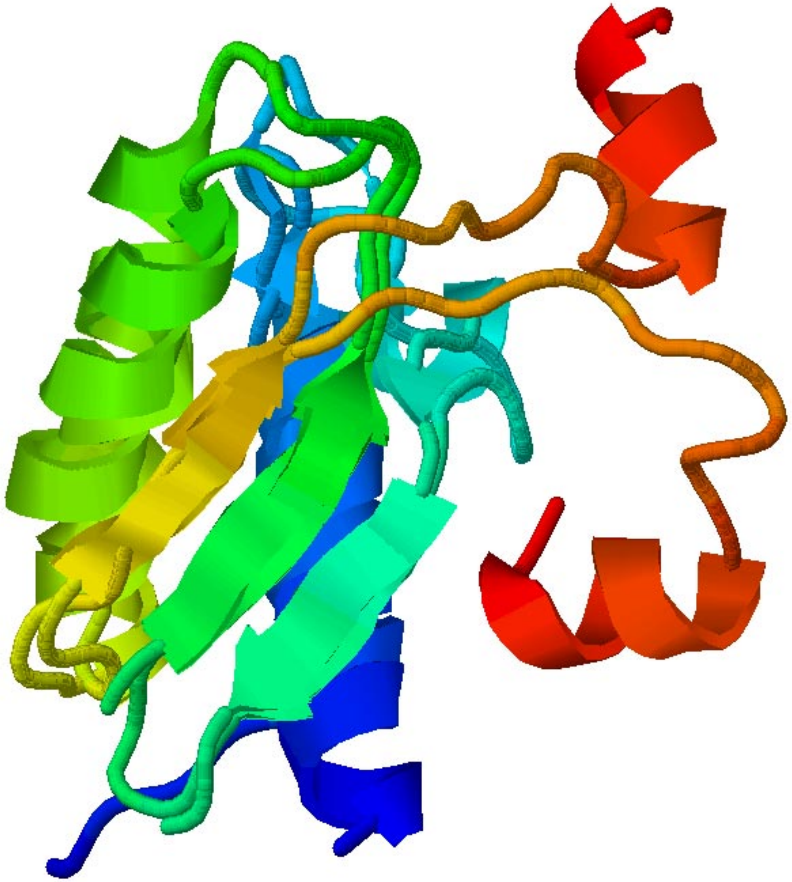
Target T0201 (NF, CASP6)



Target T0230 (FR/A, CASP6)

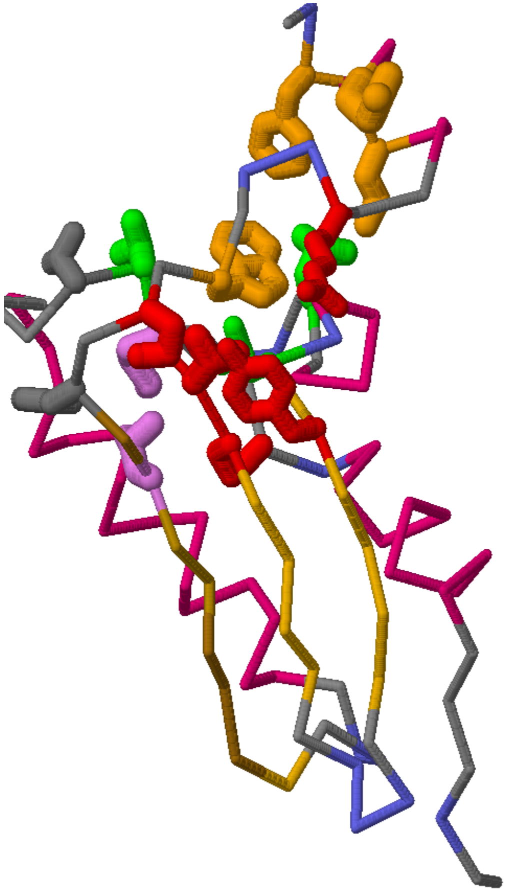
- Good except for C-terminal loop and helix flopped wrong way.
- We have secondary structure right, including phase of beta strands.
- Contact prediction helped, but we put too much weight on it—decoys fit predictions better than real structure does.

Target T0230 (FR/A, CASP6)



Target T0230 (FR/A)

Real structure with contact predictions:



Web sites

These slides: `http://www.soe.ucsc.edu/~karplus/papers/structure-prediction-tutorial-jul-2009.pdf`

Old CASP results—all our results and working notes:

`http://www.soe.ucsc.edu/~karplus/casp6/`

`http://www.soe.ucsc.edu/~karplus/casp7/`

`http://www.soe.ucsc.edu/~karplus/casp8/`

SAM-T08 prediction server:

`http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html`

UCSC bioinformatics and bioengineering degree programs:

`http://www.bme.ucsc.edu/bioinformatics/`

`http://beng.soe.ucsc.edu/`