Karplus lab: protein structure prediction and design

Kevin Karplus

`karplus@soe.ucsc.edu`

Biomolecular Engineering Department

Undergraduate and Graduate Director, Bioinformatics

University of California, Santa Cruz

Outline of Talk

- 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 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.

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?

```
MTMSRRNTDA ITIHSILDWI EDNLESPLSL EKVSERSGYS KWHLQRMFKK
ETGHSLGQYI RSRKMTEIAQ KLKESNEPIL YLAERYGFES QQTLTRTFKN
YFDVPPHKYR MTNMQGESRF LHPLNHYNS
```

↓
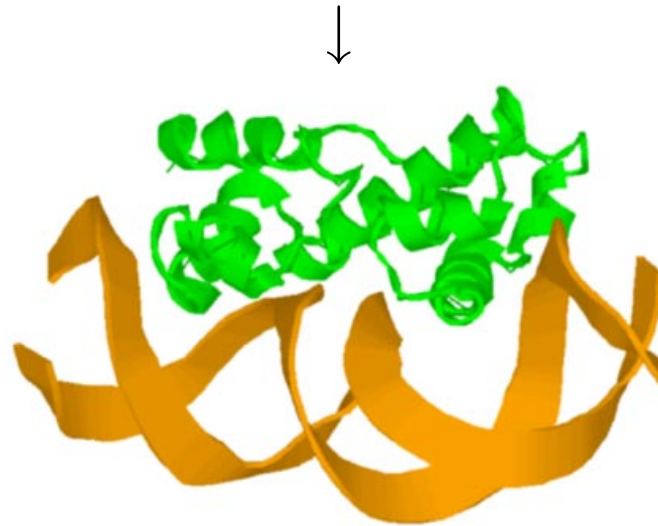
Too hard!

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.

- One popular choice is a 3-valued helix/strand/other alphabet. Typically, predictors get about 80% accuracy on 3-state prediction.

- Many machine-learning methods have been applied to this problem, but the most successful is neural networks.

- Using Conditional Random Fields can improve sampling of sequences, without improving accuracy on individual residues.

Local Structure Alphabets

- What local properties do we choose?

- We want properties that are well-conserved through evolution, easily predicted, and useful for finding and aligning templates.

- We have investigated many alphabets.

- Current favorites are str2, a 13-state secondary-structure alphabet that distingusishes between different β strands, and near-backbone-11, and 11-state burial alphabet.

Contact prediction

- Predict that residues separated along the chain are clos in 3-space.

- 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.

(Rational) Protein Design

- New direction for lab.

- Use local-structure neural nets in reverse (find sequences highly predicted to have right local structure).

- Train new neural nets to take local-structure inputs and provide amino-acid outputs.

- Use undertaker to build models.

- Use RosettaDesign to modify sequences.

- Target applications: specific binding of carbon nanotubes, mimics for AGRP (agouti-related protein) binding to different melanocortin receptors.
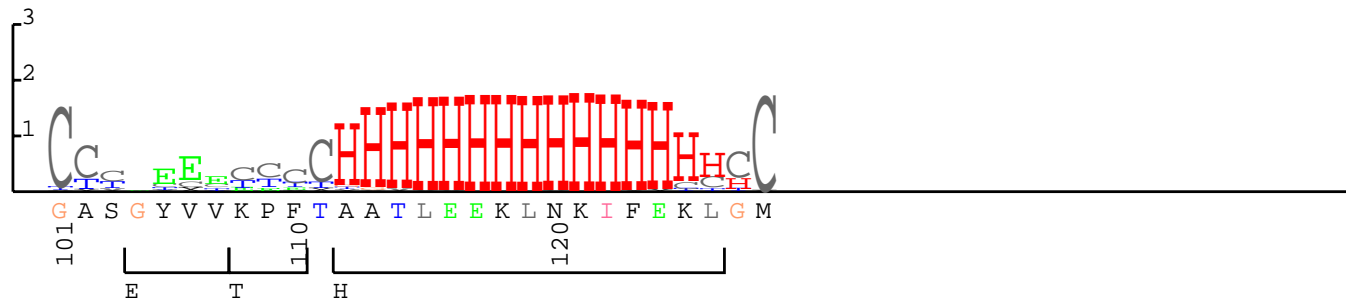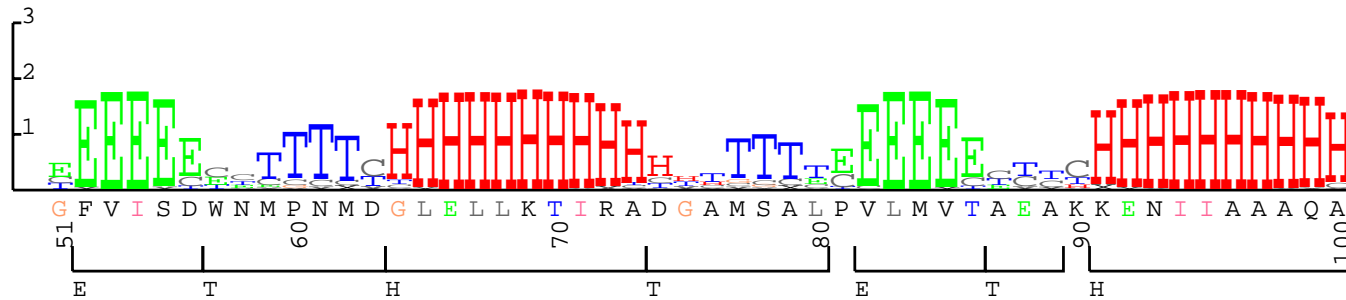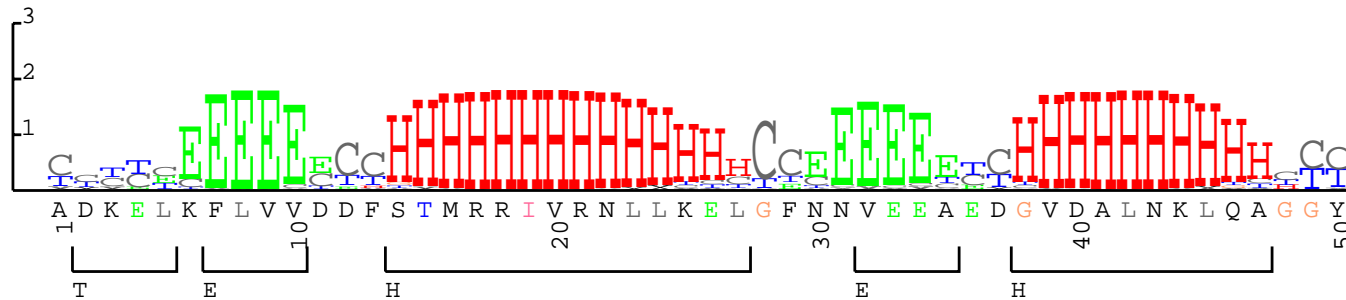
Sequence logos (MSA)

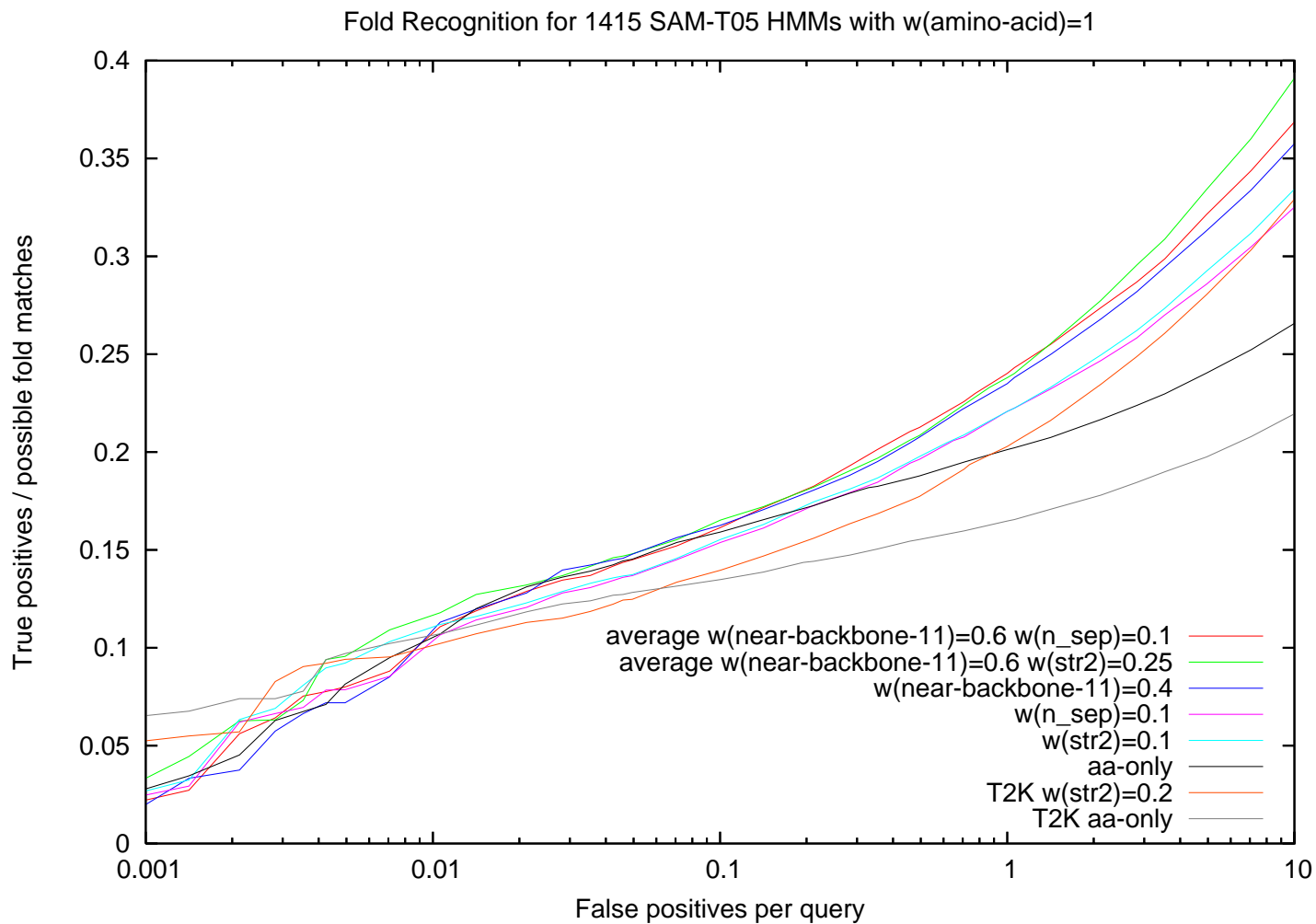Summarize multiple alignment:

nostruct-align/3chy.t2k w0.5
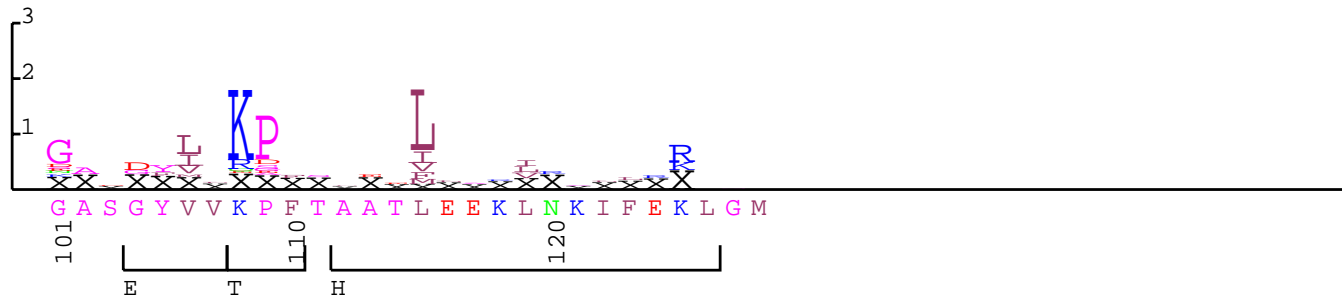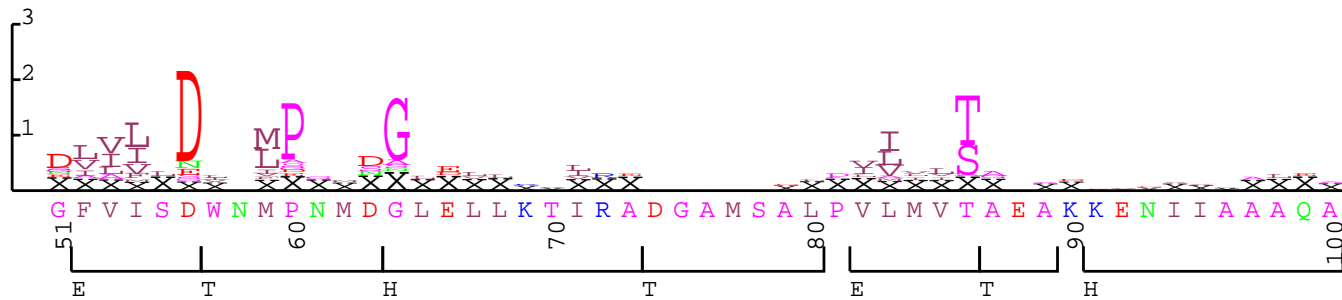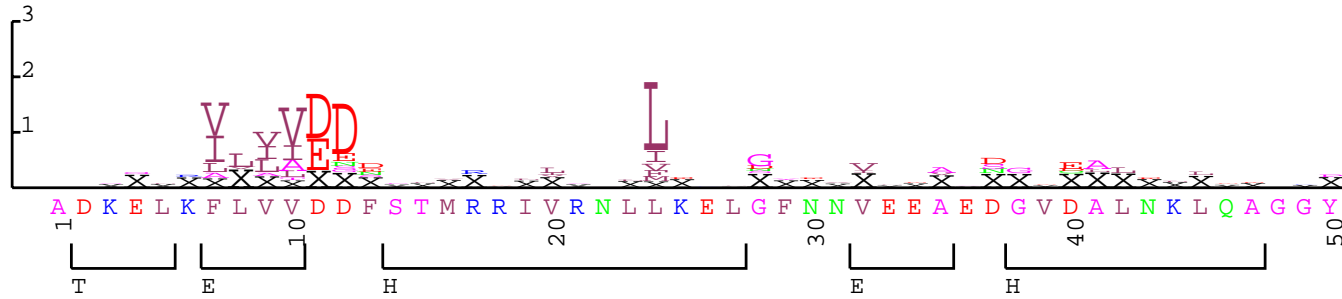
Sequence logos (NN)

Summarize local structure prediction:

nostruct-align/3chy.t2k EBGHTL

Fold recognition results
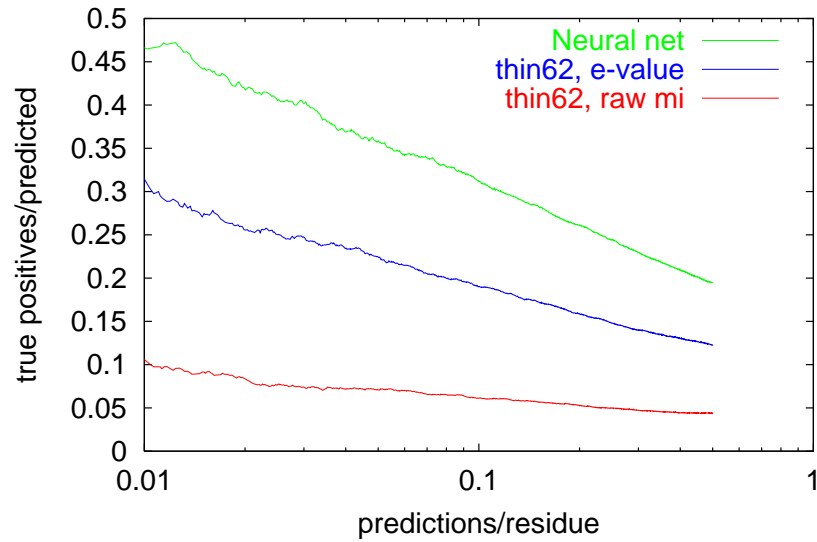

Fold Recognition for 1415 SAM-T05 HMMs with w(amino-acid)=1

True positives / possible fold matches (y-axis)
False positives per query (x-axis)

Legend:
- average w(near-backbone-11)=0.6 w(n_sep)=0.1
- average w(near-backbone-11)=0.6 w(str2)=0.25
- w(near-backbone-11)=0.4
- w(n_sep)=0.1
- w(str2)=0.1
- aa-only
- T2K w(str2)=0.2
- T2K aa-only

Contact prediction results

Accuracy of contact prediction, by protein

Weighted-accuracy of contact prediction, by protein
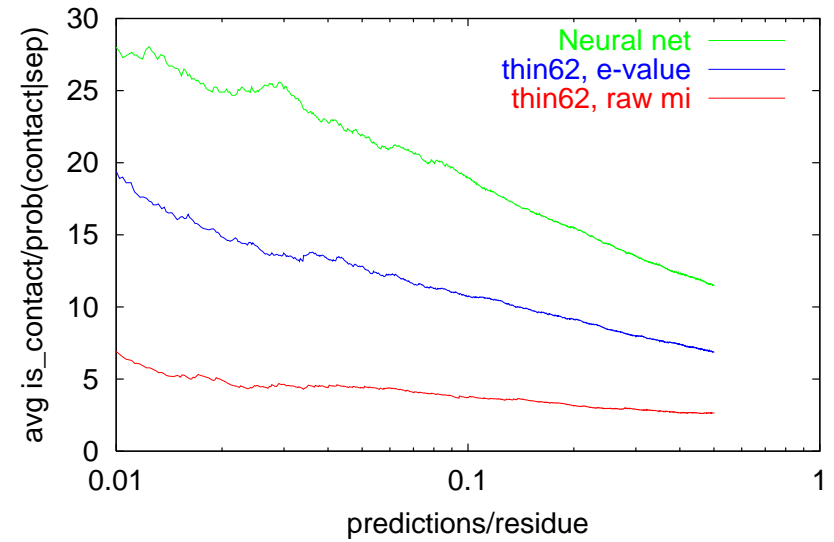
CASP ~~Competition~~ Experiment

- 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.

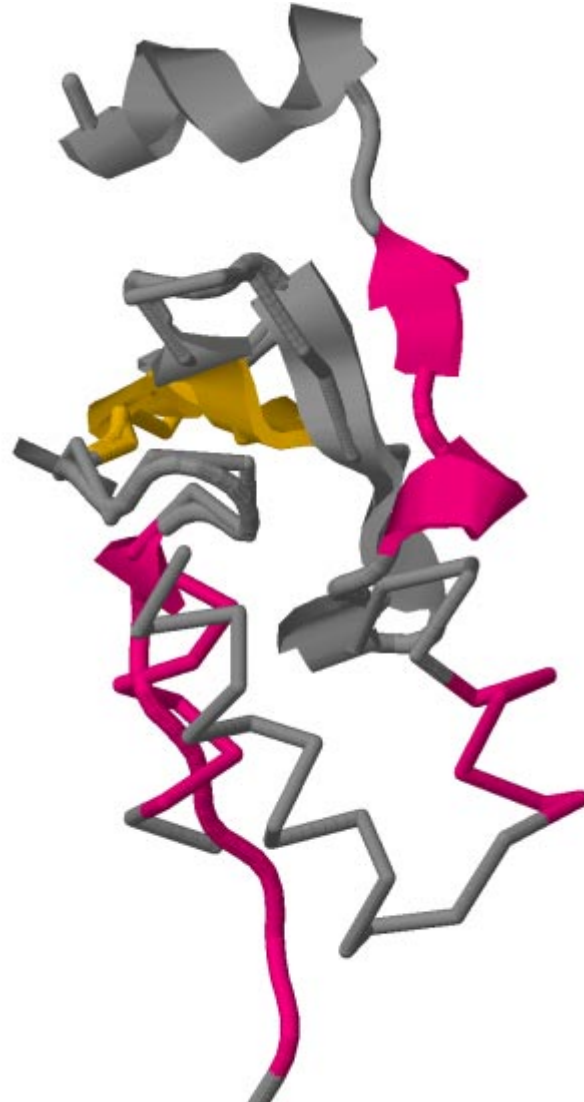- "Winners" get papers in *Proteins: Structure, Function, and Bioinformatics*.

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)
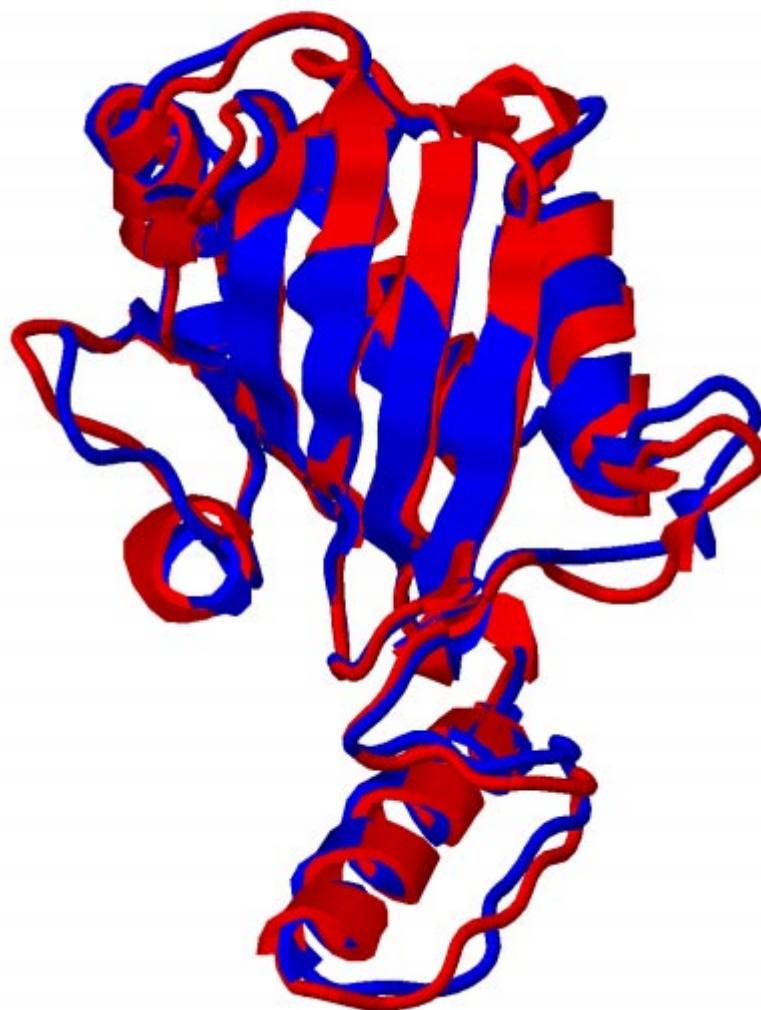
Comparative modeling: T0348

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

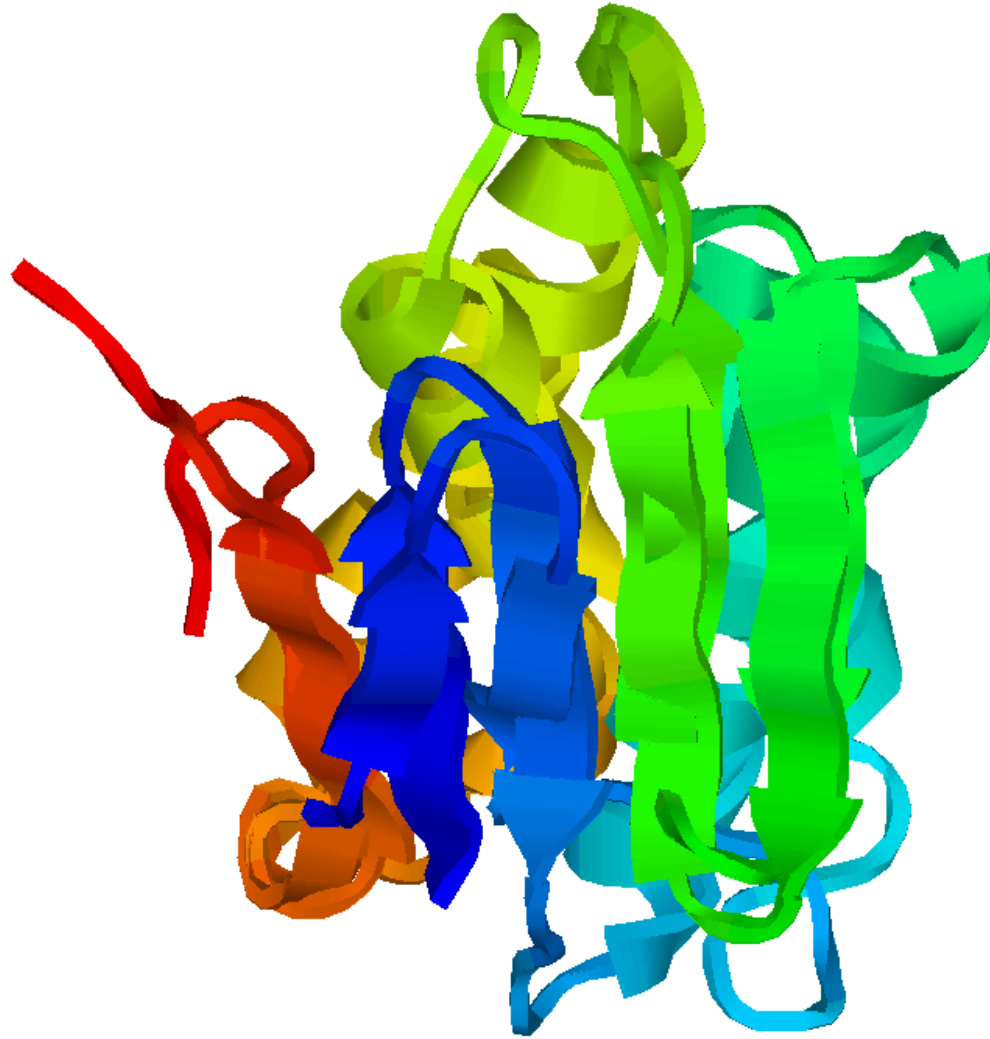# Target T0201 (NF, CASP6)

- 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)

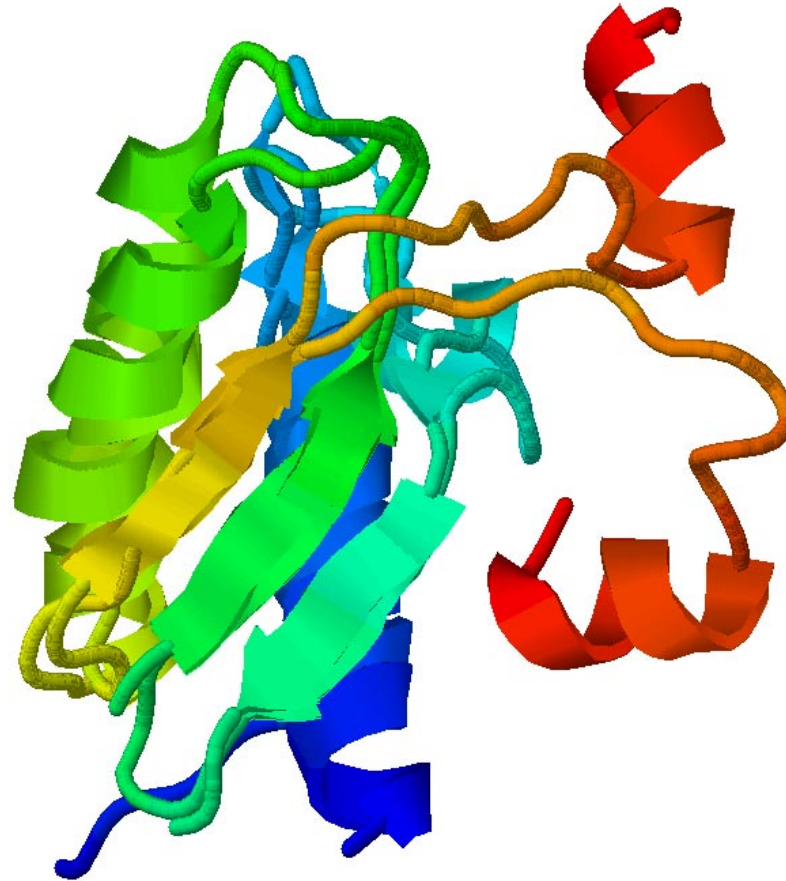# Target T0230 (FR/A, CASP6)

- Good except for C-terminal loop and helix flopped wrong way.

- We have secondary structure right, including phase of beta strands.

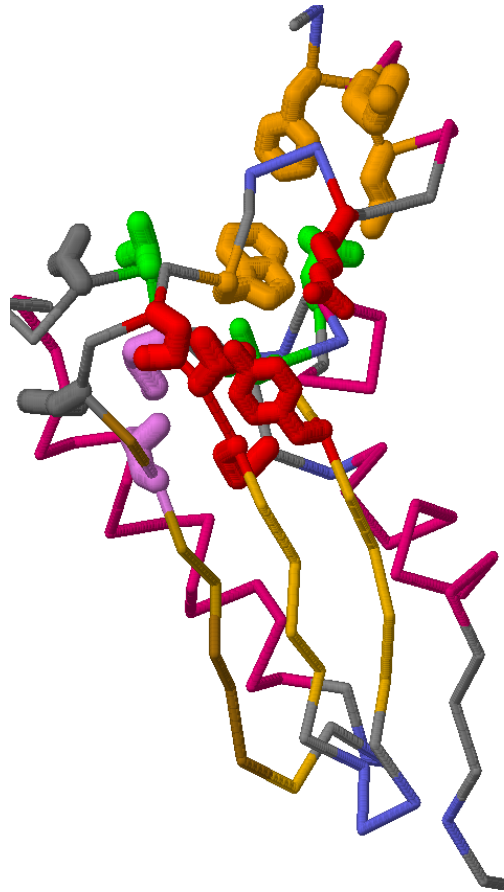- Contact prediction helped, but we put too much weight on it—decoys fit predictions better than real structure does.

Target T0230 (FR/A, CASP6)

Target T0230 (FR/A)

Real structure with contact predictions:

Web sites

These slides:

`http://users.soe.ucsc.edu/~karplus/papers/what-lab-does-oct-2009.pdf`

CASP2 through CASP8—all our results and working notes:

`http://users.soe.ucsc.edu/~karplus/casp2/`

`...`

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

SAM-T08 prediction server:

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

UCSC bioinformatics (research and degree programs) info:

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