

# 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 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 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: finding important sequences in the genome and annotating them.
-  Phylogenetics: “tree of life”.
-  Systems biology: piecing together various control networks.
-  DNA microarrays: what genes are turned on under what conditions.
-  Proteomics: what proteins are present in a mixture.
-  Protein structure prediction.



# 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 EKVSEKSGYS 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.
- 🦖 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 is neural networks.



# Contact prediction

- 🦖 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  $\log(\text{e-value})$  within protein.
- 🦖 Feed  $\log(\text{e-values})$ , 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

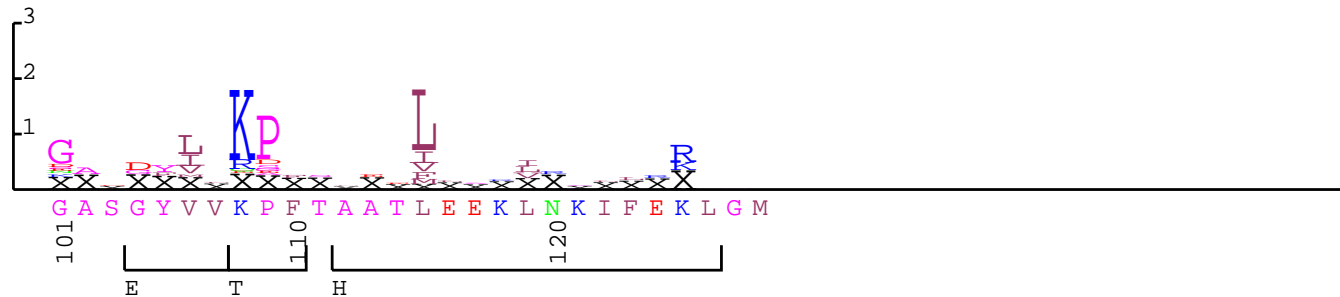
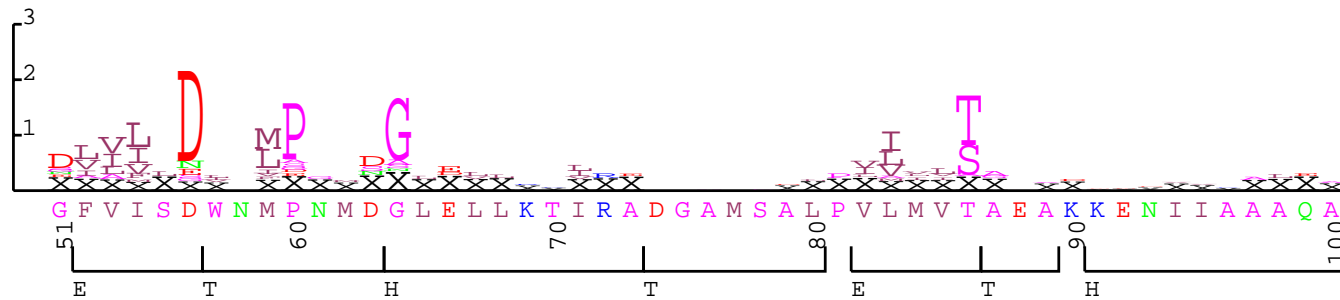
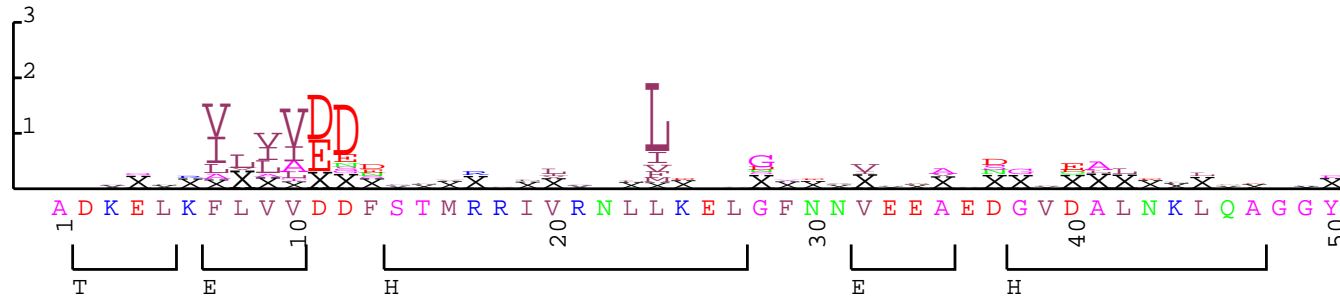
- 🦖 New direction for lab.
- 🦖 Use local-structure neural nets in reverse (find sequences highly predicted to have right local structure).
- 🦖 Use undertaker to build models.
- 🦖 Use RosettaDesign to modify sequences.
- 🦖 Target application: specific binding of carbon nanotubes.



# Sequence logos (MSA)

Summarize multiple alignment:

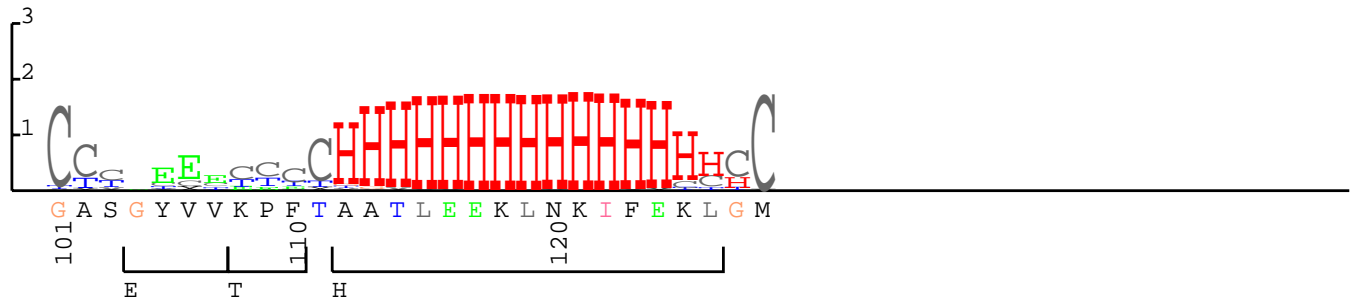
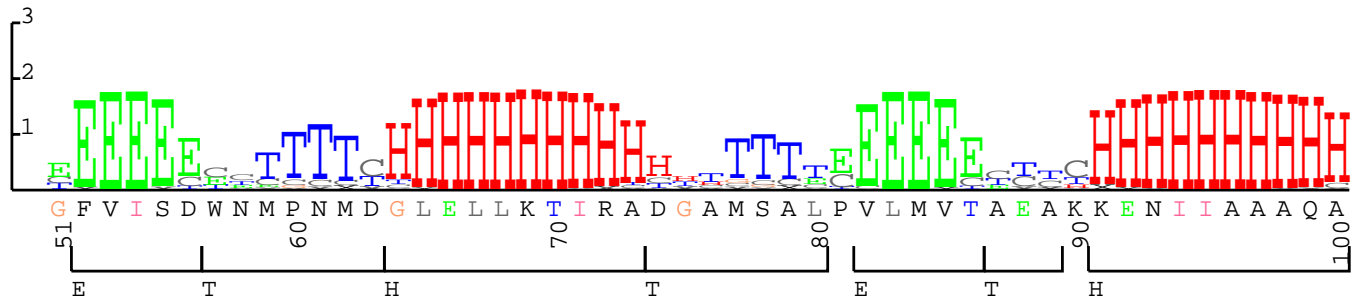
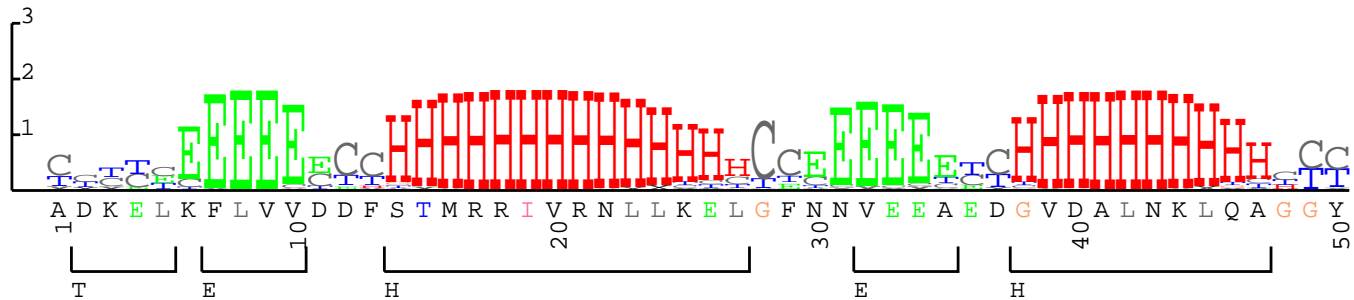
nostruct-align/3chy.t2k w0.5



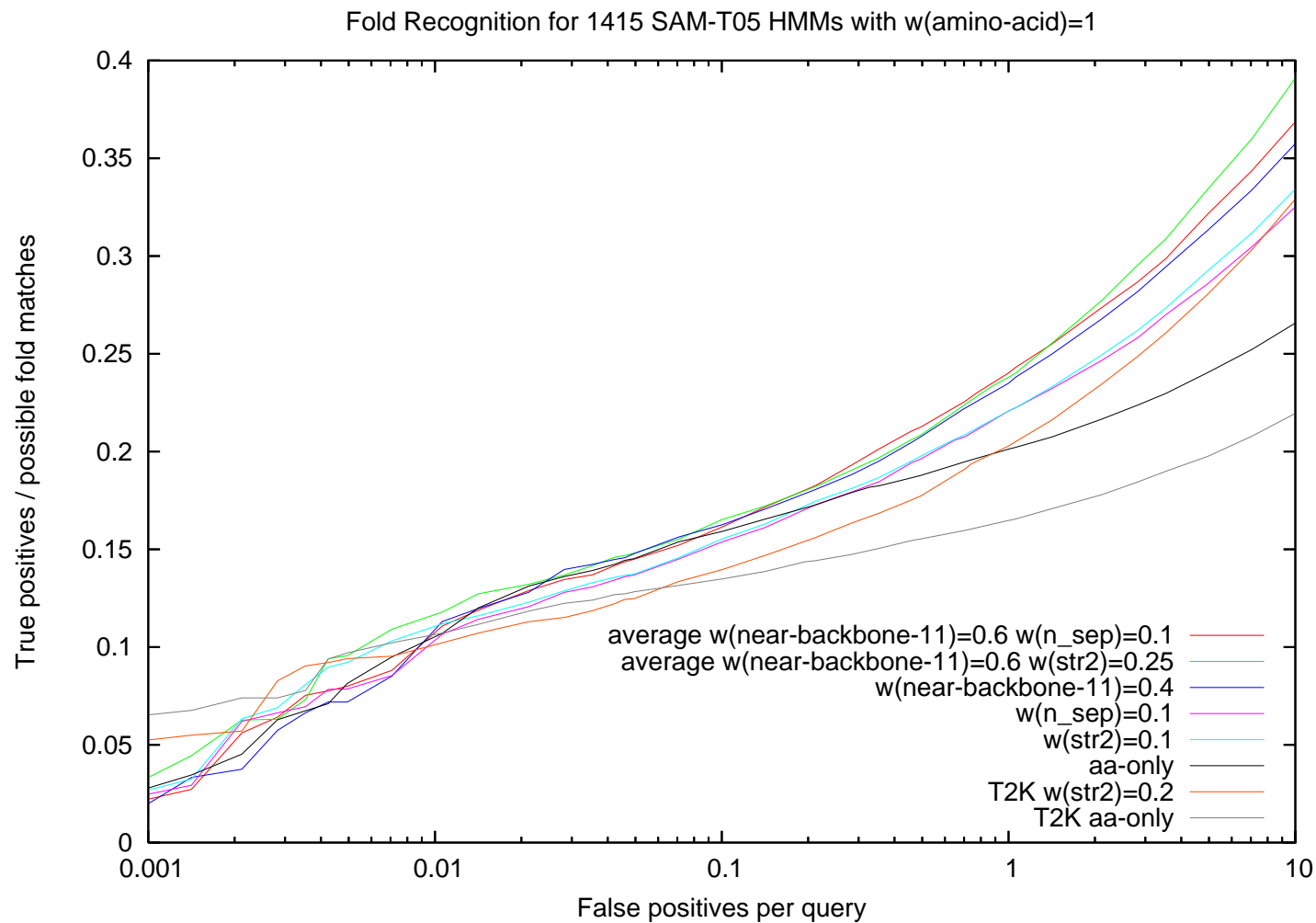
# Sequence logos (NN)

Summarize local structure prediction:

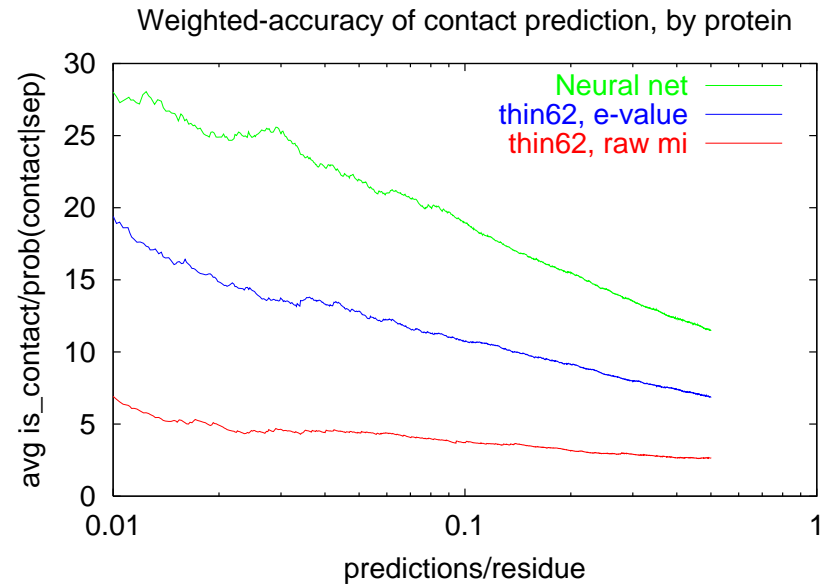
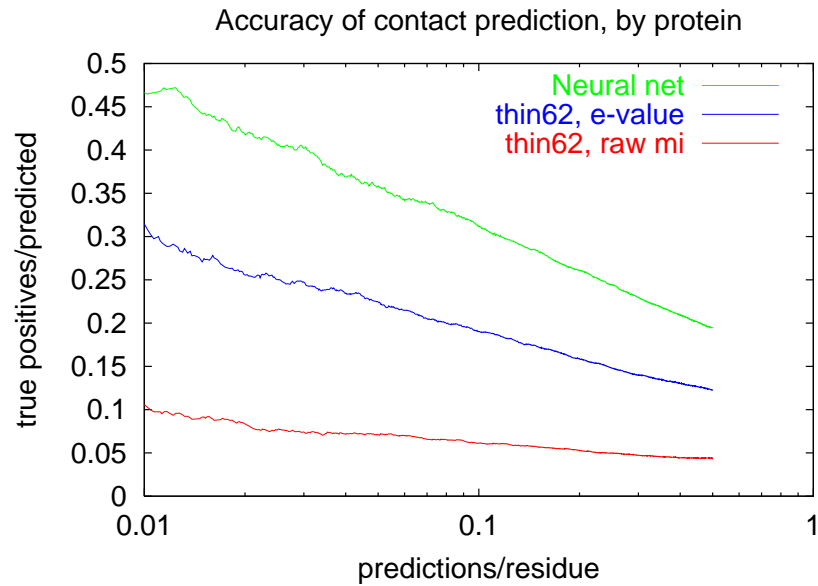
nostruct-align/3chy.t2k EBGHTL



# Fold recognition results



# Contact prediction results



# CASP Competition Experiment

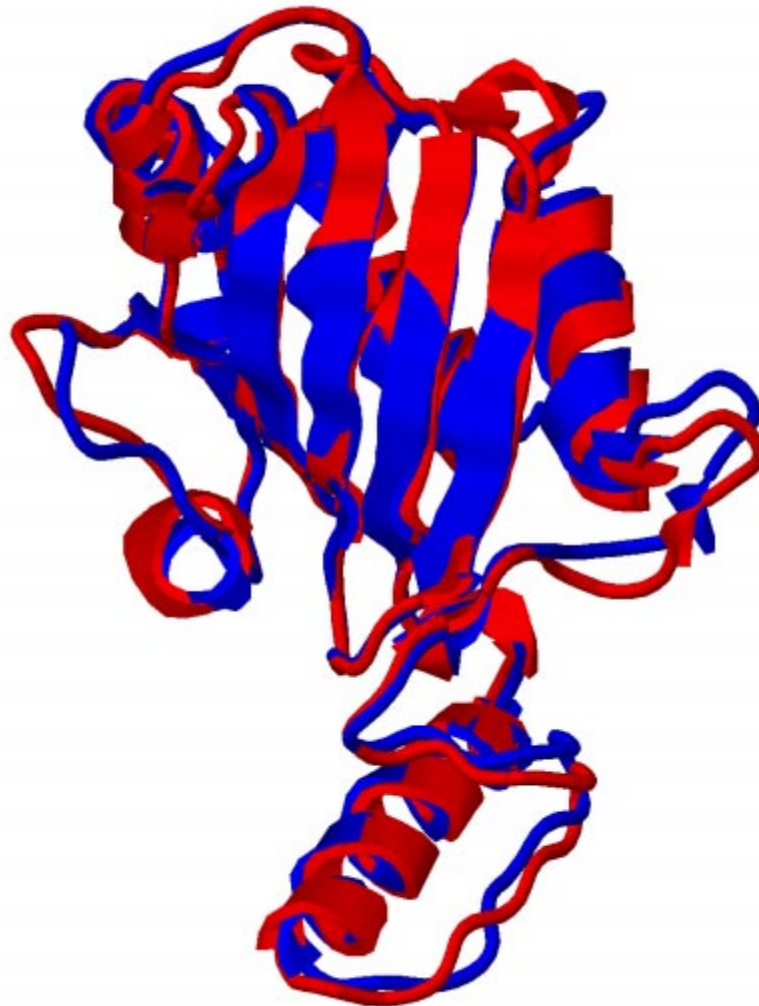
- 🦖 Everything published in literature “works”
- 🦖 CASP set up as true blind test of prediction methods.
- 🦖 Sequences of proteins about to be solved released to prediction community.
- 🦖 Predictions registered with organizers.
- 🦖 Experimental structures compared with solution by assessors.
- 🦖 “Winners”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





# T0298 domain 2 (130–315)

RMSD= 2.468Å all-atom, 1.7567Å  $C_{\alpha}$ , GDT=82.5%  
best model 1 submitted to CASP7 (red=real)



# Comparative modeling: T0348

RMSD= 11.8 Å  $C_{\alpha}$ , GDT=58.2% (cartoon=real)  
best model 1 by CASP7 GDT, Robetta1 slightly better.

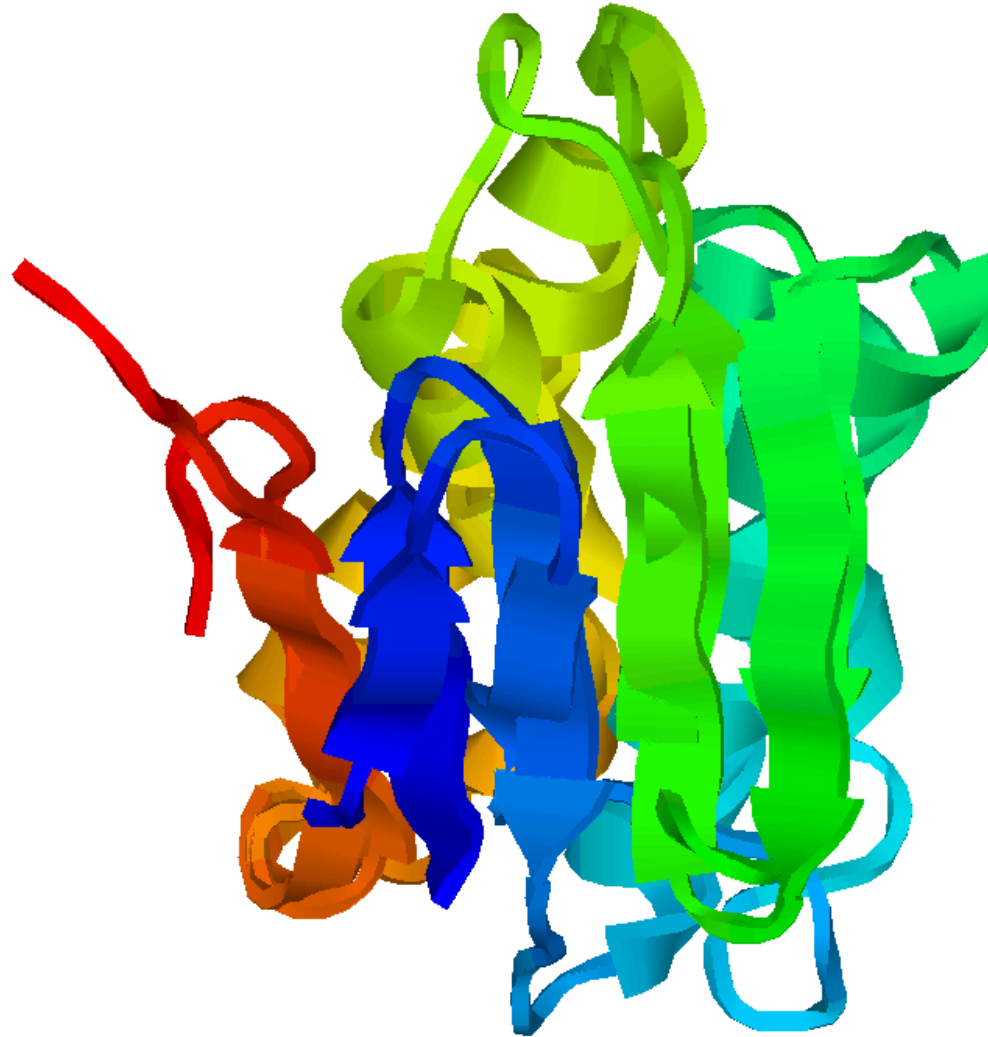


# Target T0201 (NF, CASP6)

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



# Target T0201 (NF, CASP6)

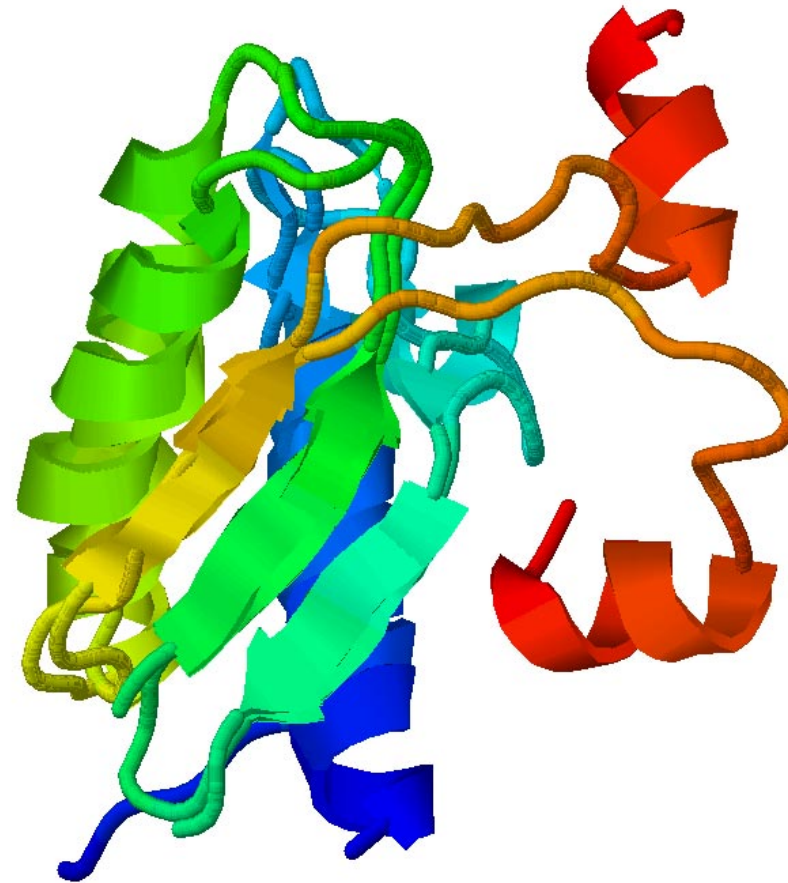


# Target T0230 (FR/A, CASP6)

- 🦖 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.

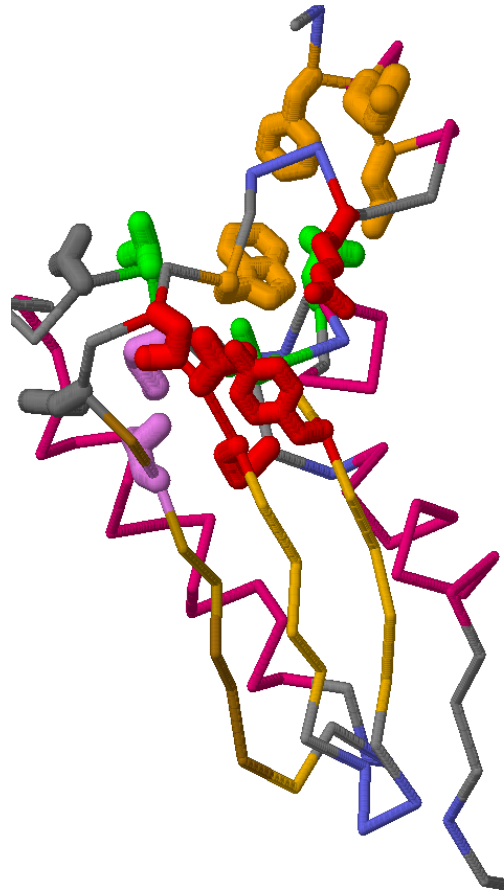


# Target T0230 (FR/A, CASP6)



# Target T0230 (FR/A)

Real structure with contact predictions:



# Web sites

## These slides:

<http://www.soe.ucsc.edu/~karplus/papers/what-lab-does-mar-2007.pdf>

## CASP6 and CASP7—all our results and working notes:

<http://www.soe.ucsc.edu/~karplus/casp6/>

<http://www.soe.ucsc.edu/~karplus/casp7/>

## SAM-T06 prediction server:

[http://www.soe.ucsc.edu/research/compbio/SAM\\_T06/T06-query.html](http://www.soe.ucsc.edu/research/compbio/SAM_T06/T06-query.html)

## Predictions for all yeast proteins:

<http://www.soe.ucsc.edu/~karplus/yeast/>

## UCSC bioinformatics (research and degree programs) info:

<http://www.soe.ucsc.edu/research/compbio/>

