SAM-T04: what’s new for CASP6

Kevin Karplus

Richard Hughey

Jenny Draper, Sol Katzman, Martina Koeva, George Shackelford

Bret Barnes, Marcia Soriano

`karplus@soe.ucsc.edu`

Biomolecular Engineering

University of California, Santa Cruz

Steps of SAM-Txx Methods

- Iterative search and alignment [rewritten, minor improvements]
- Local structure prediction [new alphabets, minor tweaks]
- Multi-track HMMs [minor tweaks]
- Finding medium-length fragments (fragfinder) [multi-track HMMs, filter implausible]
- Contact prediction [all new]
- Conformation generation (undertaker) [major changes]

Contact prediction: new in 2004!

- 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 z-score of log(e-value) within protein.
- Feed e-values, z-scores, conservation, amino-acid profile, separation along chain into neural net.

Evaluating contact prediction

Two measures of contact prediction:

- Accuracy:

$$\frac{\sum \chi(i,j)}{\sum 1}$$

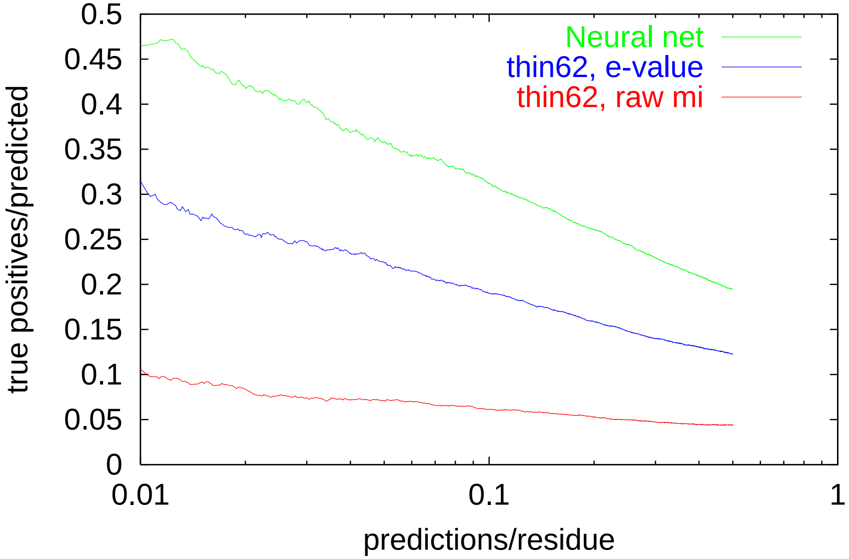
(favors short-range predictions, where contact probability is higher)

- Weighted accuracy:

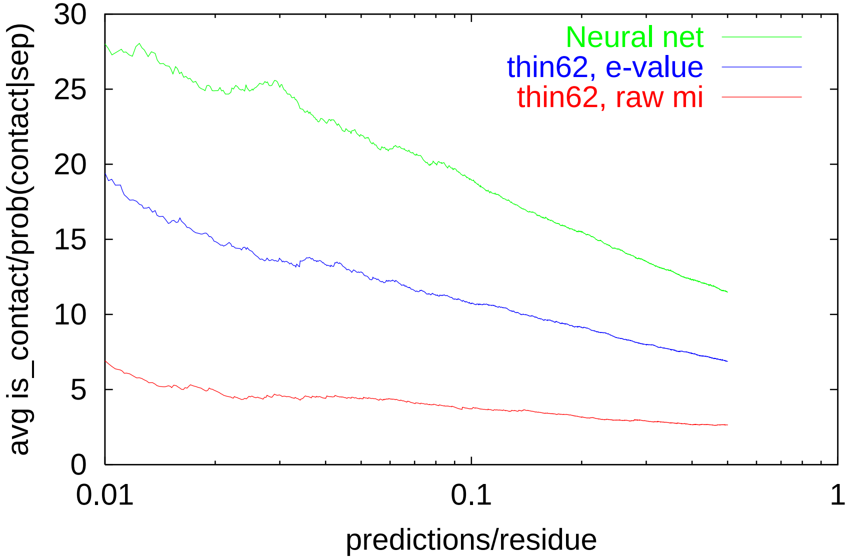
$$\frac{\sum \frac{\chi(i,j)}{\text{Prob}\left(\text{contact}\middle|\text{separation}_{=|i-j|}\right)}}{\sum 1}$$

(1 if predictions no better than chance based on separation).

Contact prediction results



Accuracy of contact prediction, by protein



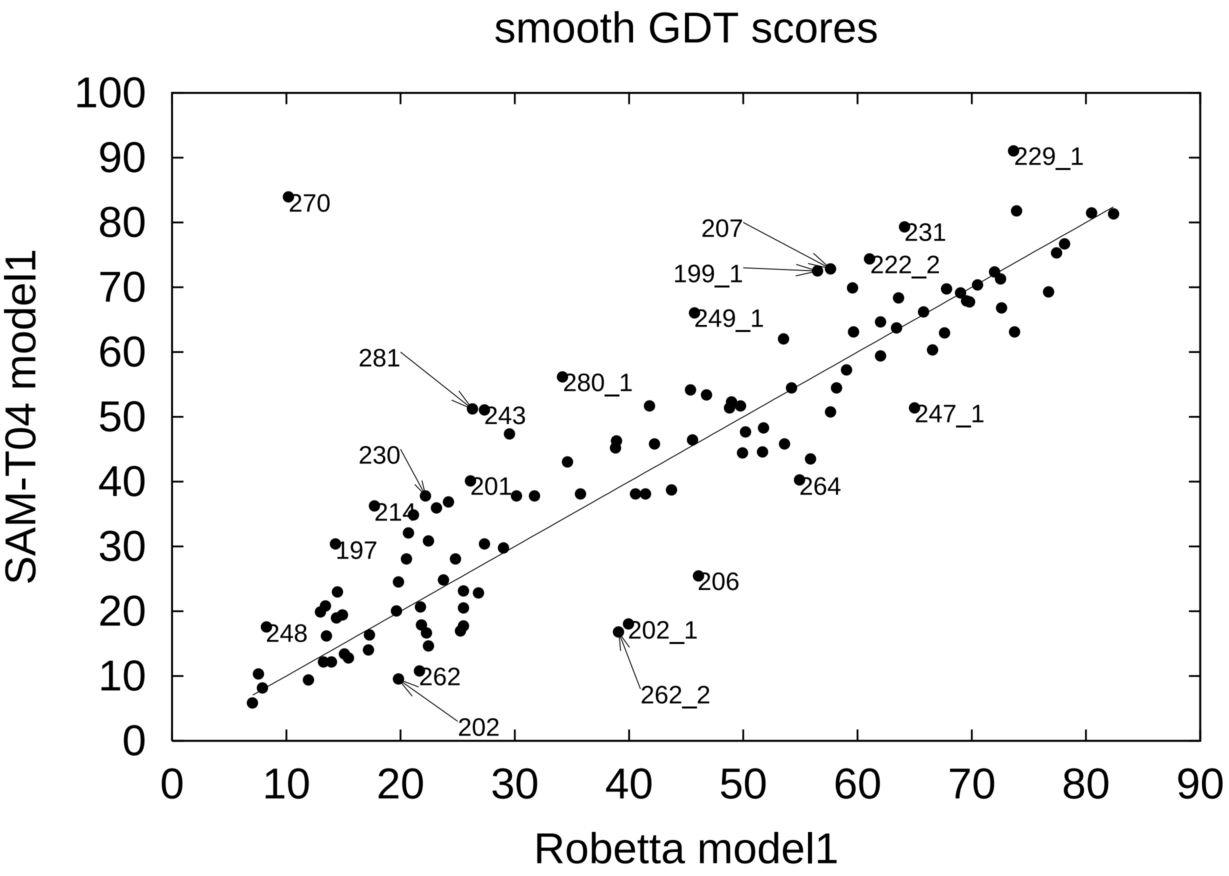
Weighted-accuracy of contact prediction, by protein

Undertaker

Undertaker is UCSC’s attempt at a fragment-packing program (named because it optimizes burial).

- New cost functions (especially H-bonds)
- Improved clash detection.
- New conformation change operators (tweaking torsion angles, rigid body movements of chunks).
- New ways to specify constraints (Hbond, SSbond, HelixConstraint, StrandConstraint, SheetConstraint).
- Improved adaptation of genetic algorithm.

Model 1 vs. Robetta 1

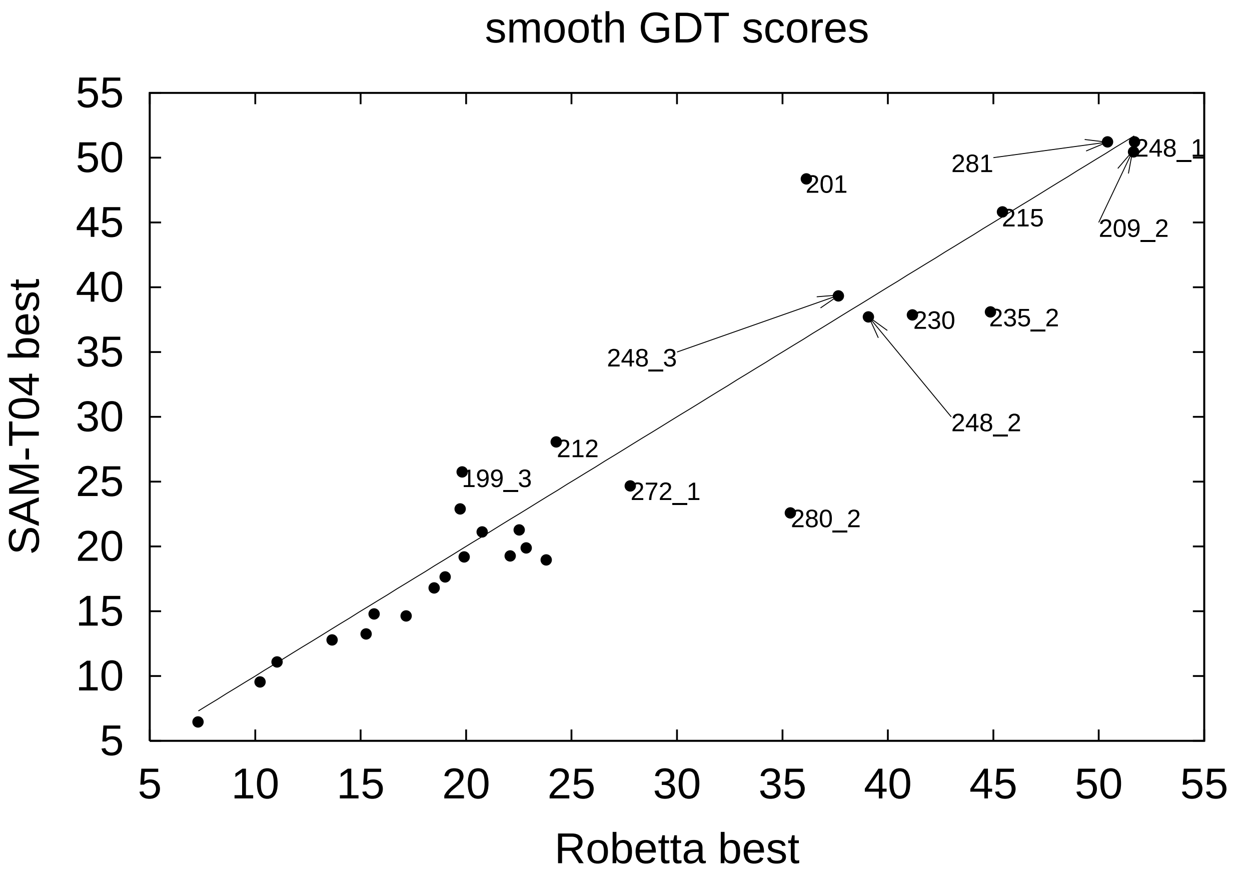


Good stuff from Murzin

We won’t discuss the following:

- T0270: 1t0tA became available after servers ran.
- T0213: Murzin suggested using 1t62A for T0213, T0214, and T0227. T04 scored 1t62A best—we messed up the good alignment.
- T0214: We used 1t62A, but we never got a good alignment.
- T0227: T04 scored 1t62A best, but 2° prediction was poor, so we had bad alignments.
- T0240: We submitted both dimer and monomer, but mistakenly put the dimer first.
- T0245: 1tljA became available, but we don’t have the true structure yet.

Best vs. Robetta best (NF and FR/A)

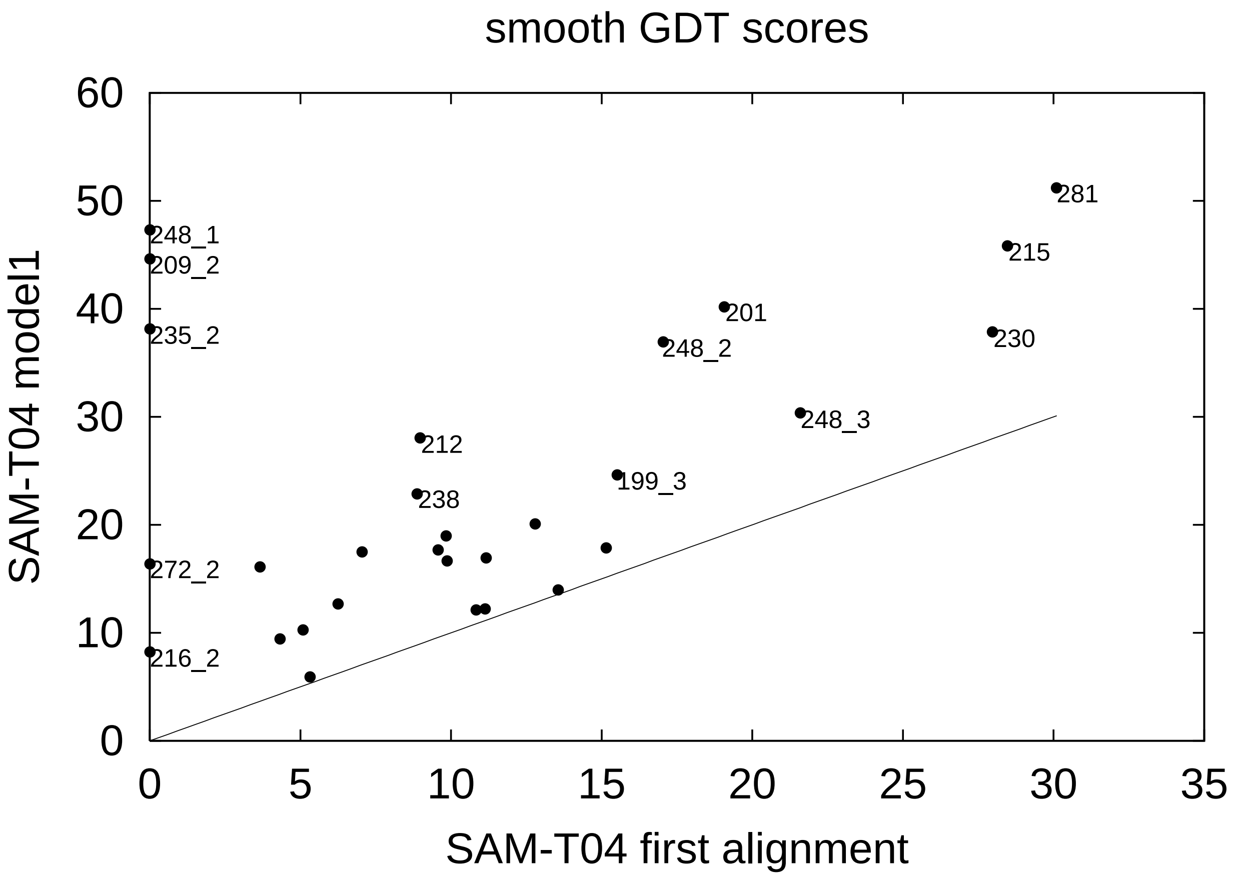


Good stuff from Robetta

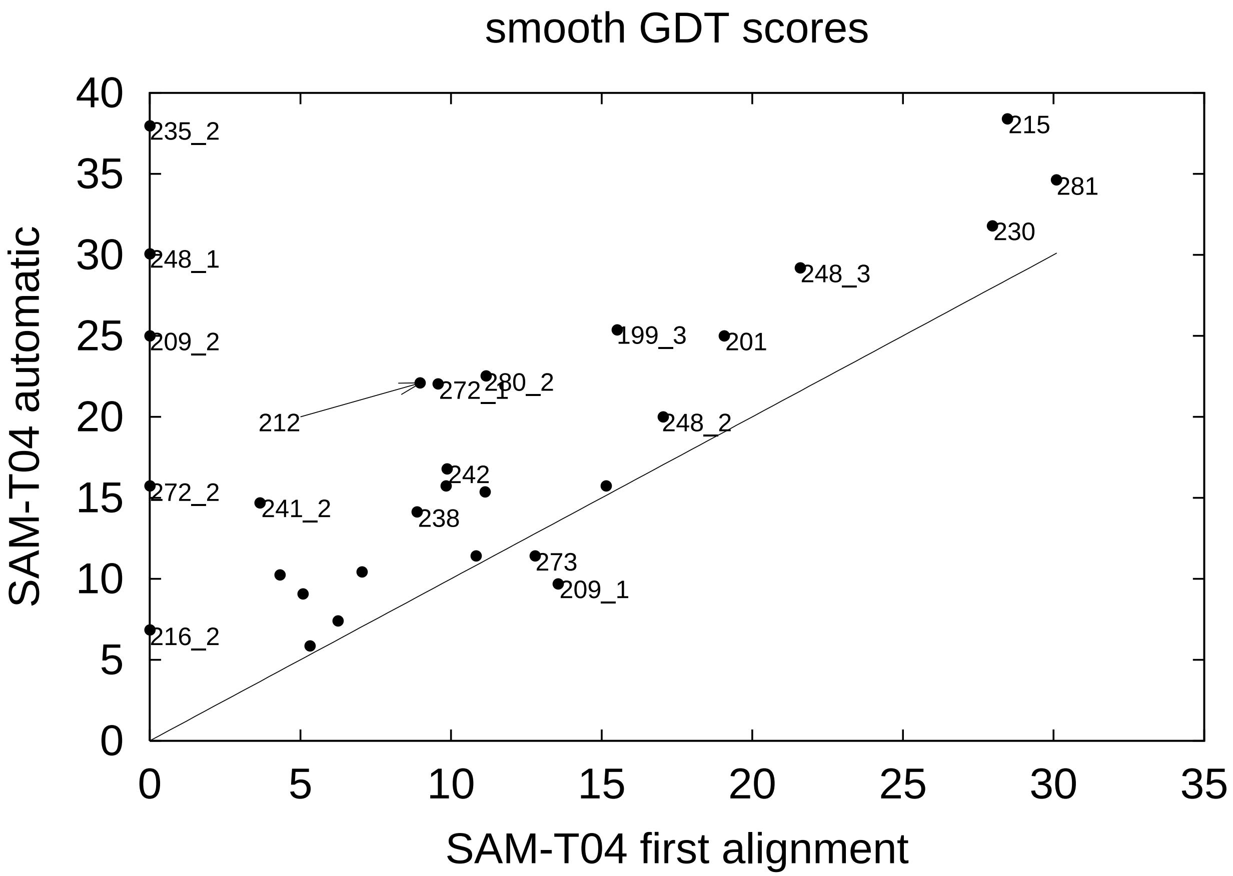
We won’t discuss the following, because the good stuff in them seems to have come from better Robetta models:

- T0209_2: sheet constraints from Robetta-model1
- T0248 (all 3 domains): borrows heavily from Robetta-model2

Model 1 vs. alignment (NF and FR/A)



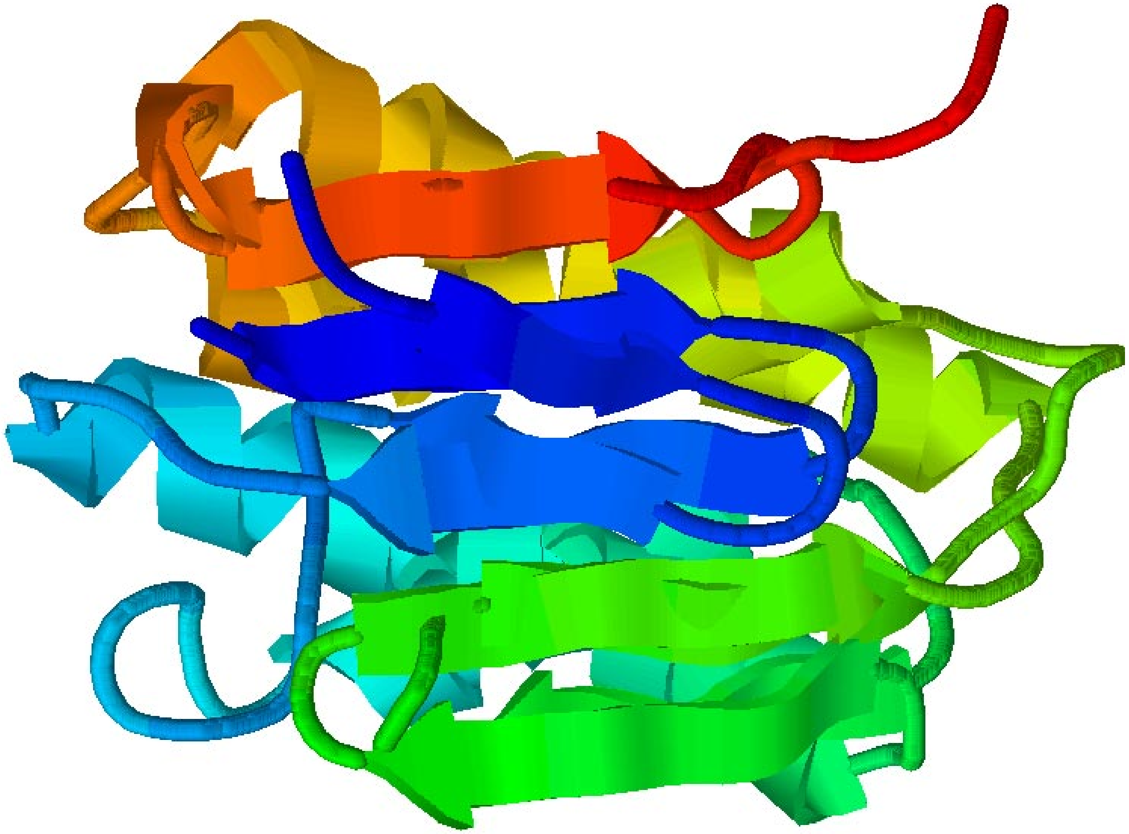
Auto vs. align (NF and FR/A)



Target T0201 (NF)

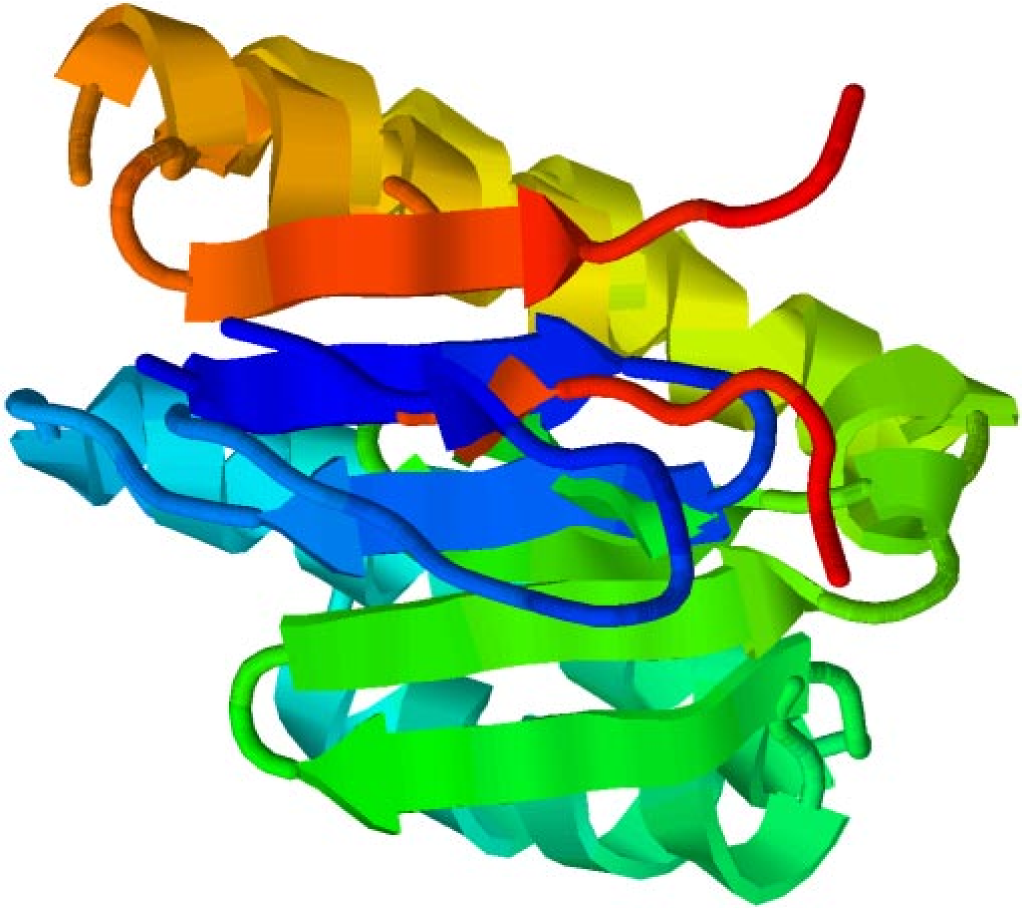
- We tried forcing various sheet topologies and selected 4 by hand.
- Model 1 has right topology (5.9117 all-atom RMSD).
- Unconstrained cost function not good at choosing topology.
- Contact prediction didn’t help, though first prediction right.
- Helices were too short.
- Highest GDT and lowest RMSD model (try41-opt2.repack-nonPC 5.4912 all-atom) has wrong topology.

Target T0201 (NF)



Target T0201 (NF)

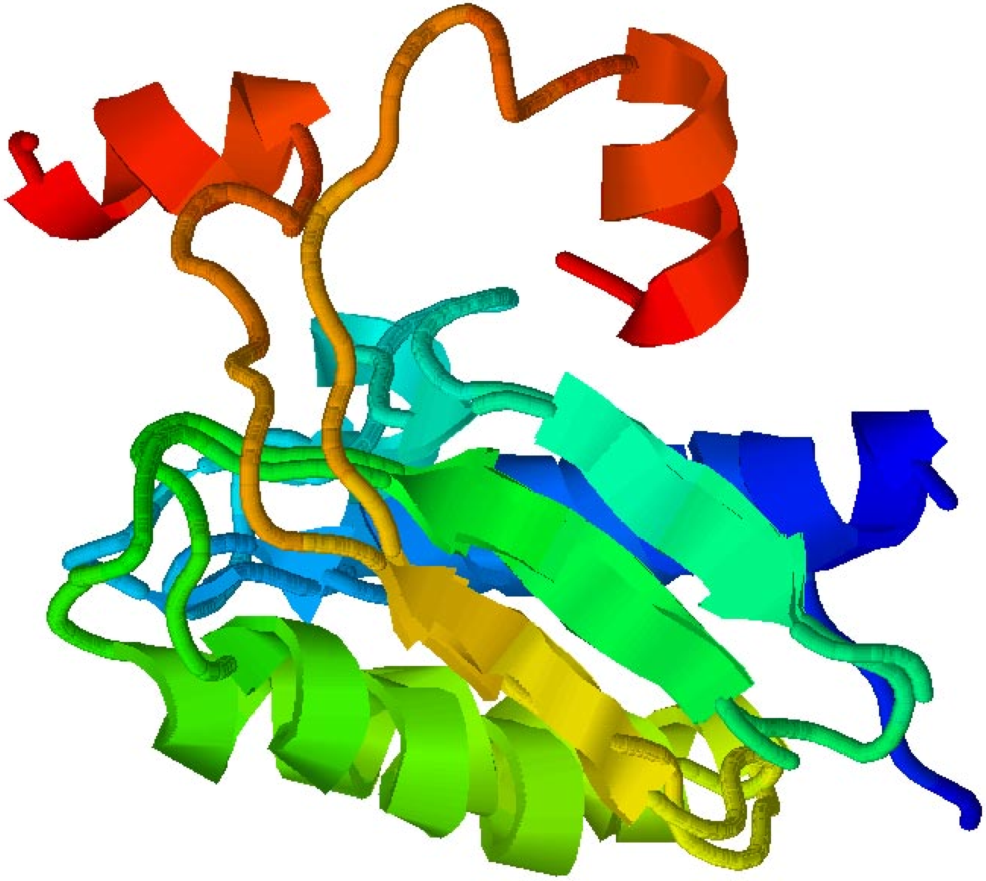
Wrong topology, but best scoring decoy.



Target T0230 (FR/A)

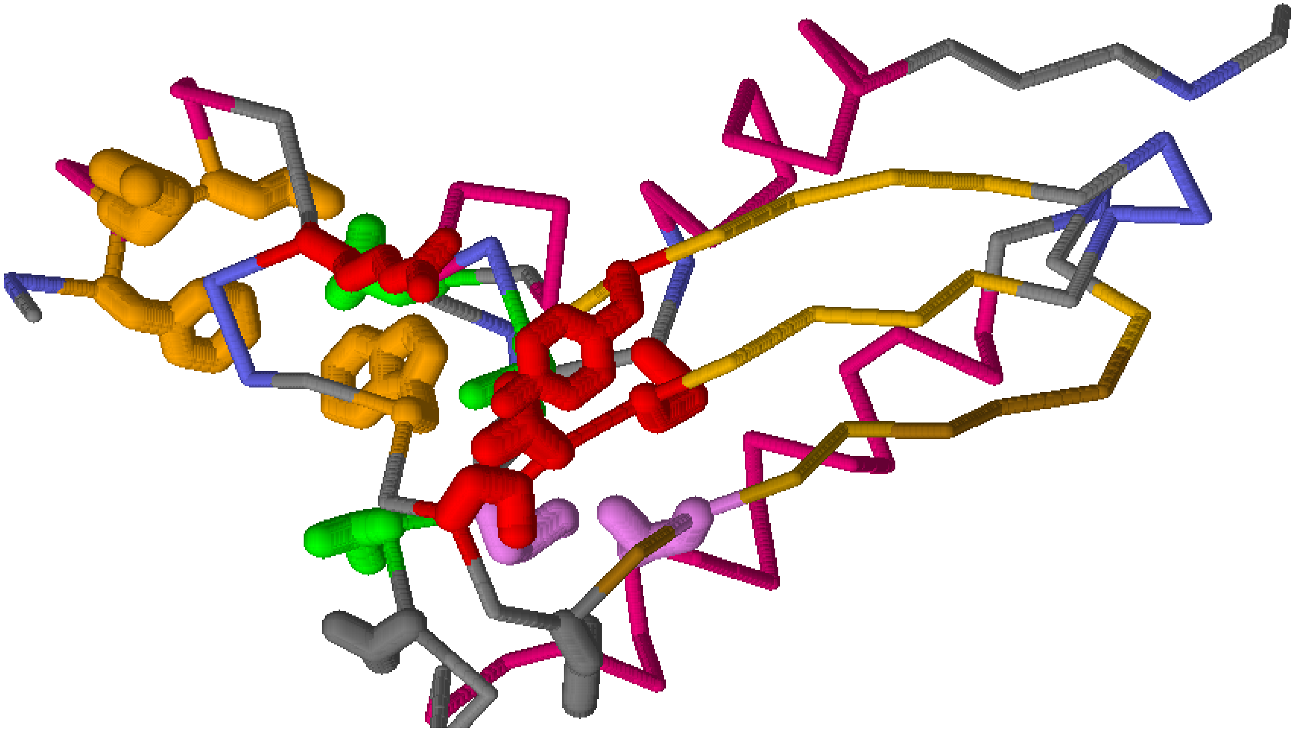
- Good except for C-terminal loop and helix flopped wrong way.
- We have secondary structure right, including phase of beta strands.
- Contact prediction helped, but we put too much weight on it—decoys fit predictions better than real structure does.

Target T0230 (FR/A)



Target T0230 (FR/A)

Real structure with contact predictions:

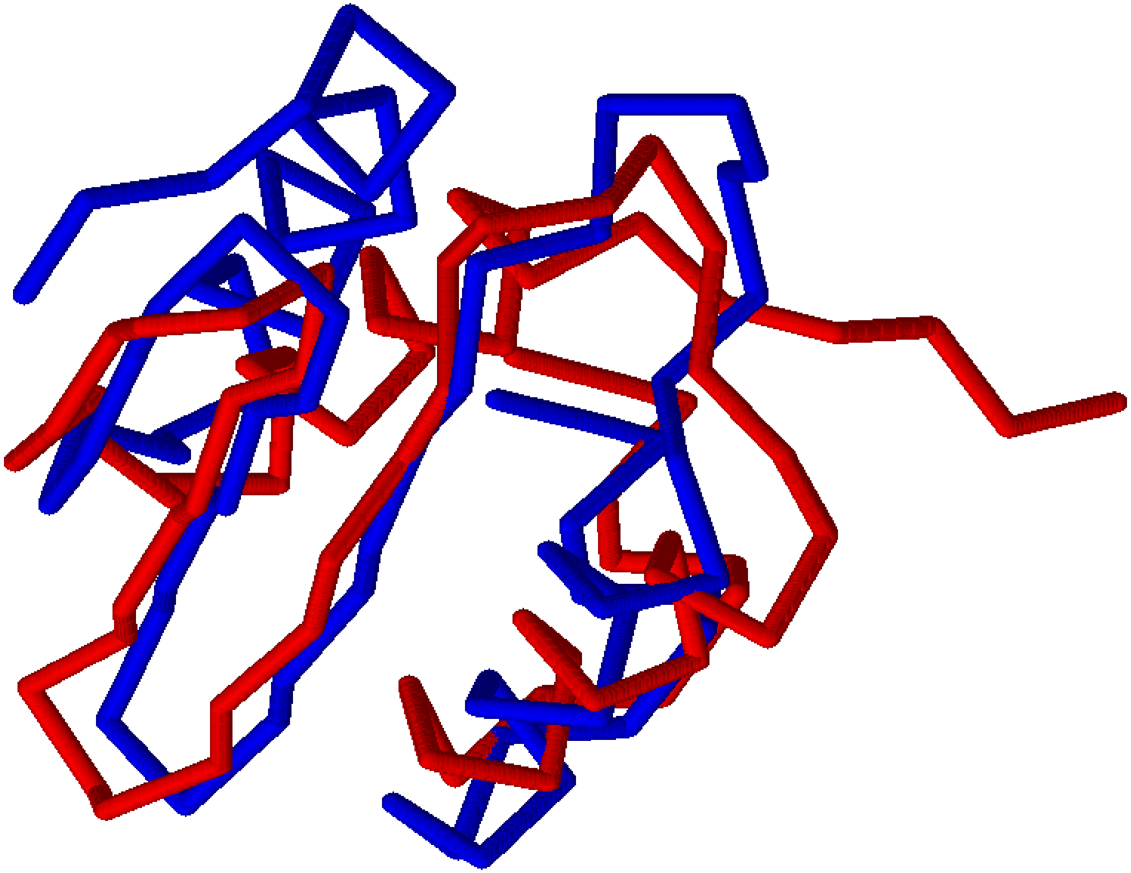


Target T0281 (FR/A)

- Third strand has off-by-one error.
- Top T04 hit (1gefA) is good, T2K put it 3rd.
- We submitted the best model we had (in GDT score, try7-opt1 had better rmsd).
- Sol’s hand work helped, but my attempts to force M1-P4 as a first strand and to remove the bulge at R22 were misguided.

Target T0281 (FR/A)

Red is real structure.

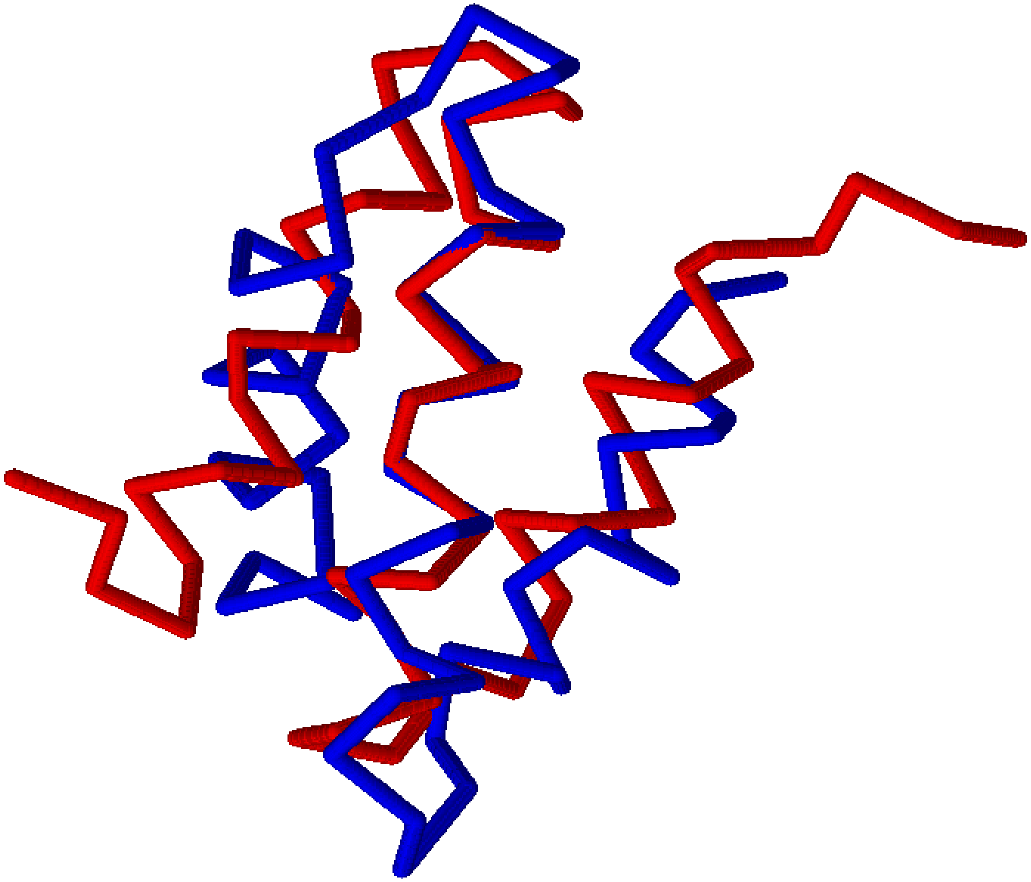


Target T0215 (FR/A)

- Secondary structure good, but helix packing angles wrong.
- Need helix packing info in undertaker—hand-added constraints were wrong.
- Too few homologs for contact prediction.

Target T0215 (FR/A)

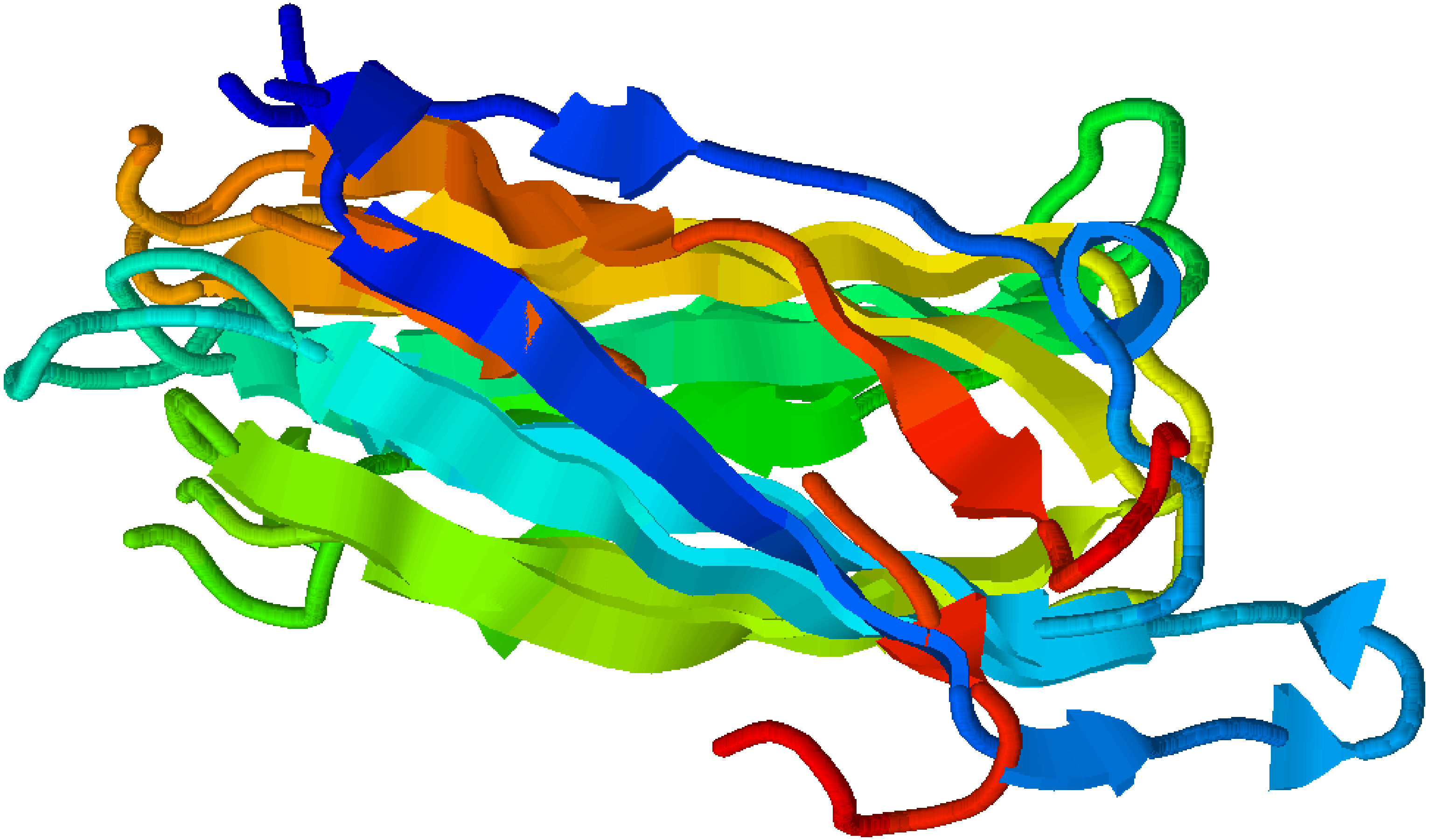
Red is real structure.



Target T0212 (FR/A)

- We tried to force a jelly-roll structure with the N-terminal strand omitted.
- Swapping the N- and C-terminal strands of our model would make it almost right.
- Strand T60-A66 is off by one.

Target T0212 (FR/A)



Web sites

UCSC bioinformatics degrees:

`http://www.soe.ucsc.edu/programs/bionformatics/`

SAM tool suite info:

`http://www.soe.ucsc.edu/research/compbio/sam.html`

HMM **servers:** `http://www.soe.ucsc.edu/research/compbio/HMM-apps/`

These slides:

`http://www.soe.ucsc.edu/~karplus/papers/casp6-slides.pdf`

CASP6 all working files: `http://www.soe.ucsc.edu/~karplus/casp6`

Iterative search using HMMS

SAM-T98, T99, T2K, and T04 methods all use similar method for building a target HMM, given a single sequence (or a seed alignment). The `target04` script

- uses perl modules to encapsulate programs, for greater flexibility.
- uses fastacmd instead of grep for counting and retrieving sequences.
- uses blastpgp on each iteration to prefetch sequences for hmmscore.
- uses cheap_gaps transition regularizer throughout.

Local Structure Alphabets

- Use more backbone alphabets:
  - DSSP & DSSP-ehl2
  - Str2
  - Stride
  - Bystroff
  - alpha
- Use burial alphabets:
  - CB-14-7
  - near-backbone-11

Neural Net

- We use neural nets to predict local properties.
- Input is profile with probabilities of amino acids at each position of target chain, plus insertion and deletion probabilities. New in 2004 is additional 20 inputs with one-hot encoding of amino acid in the target sequence.
- Neural nets were retrained using T04 alignments and better training set.

Multi-track HMMs

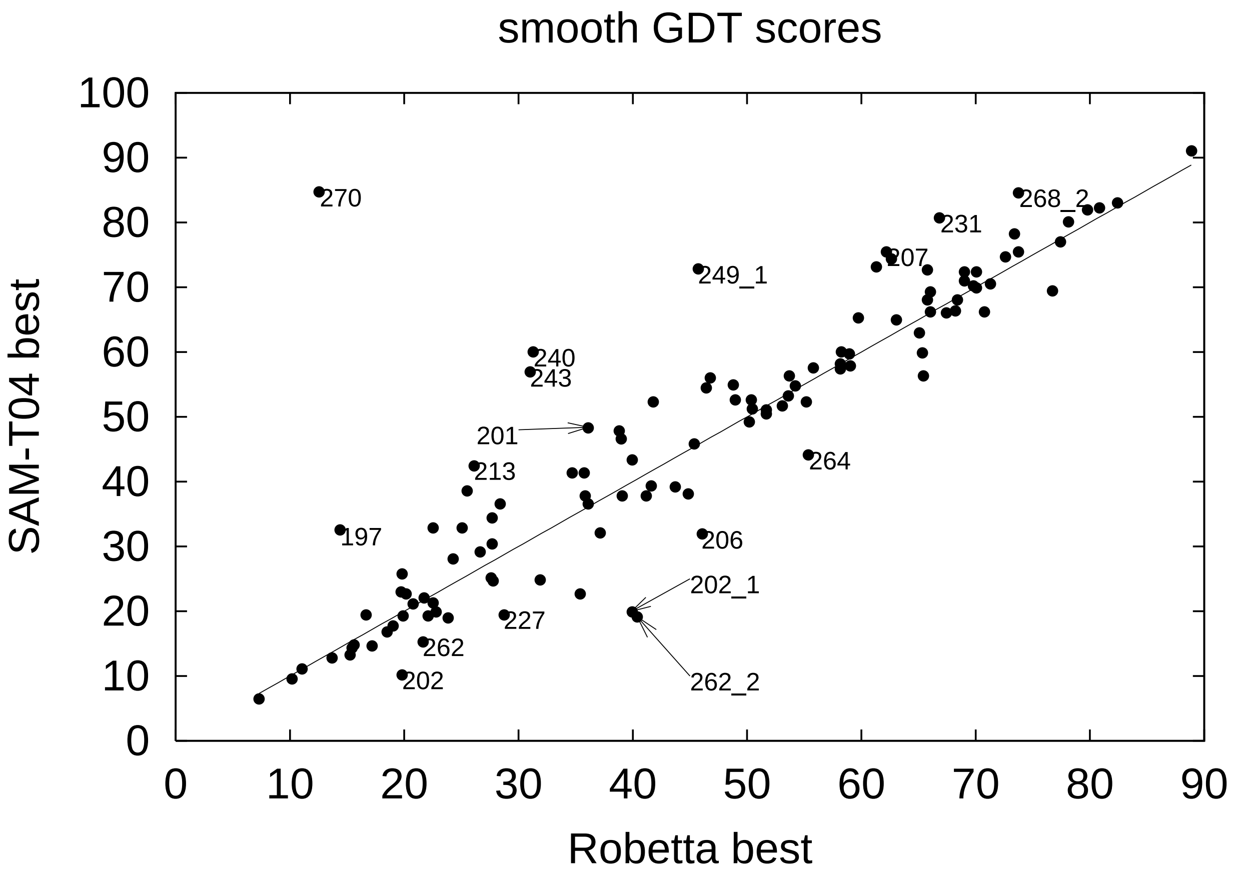
- Using more 2-track HMMs: amino acid plus each local structure alphabet.
- Using 3-track HMMs: amino acid, backbone (str2), burial (CB-14-7)
- Generate many alignments for each potential template.
  - use different HMMs.
  - use both local and global.
  - use both Viterbi and posterior decoding.

Fragfinder

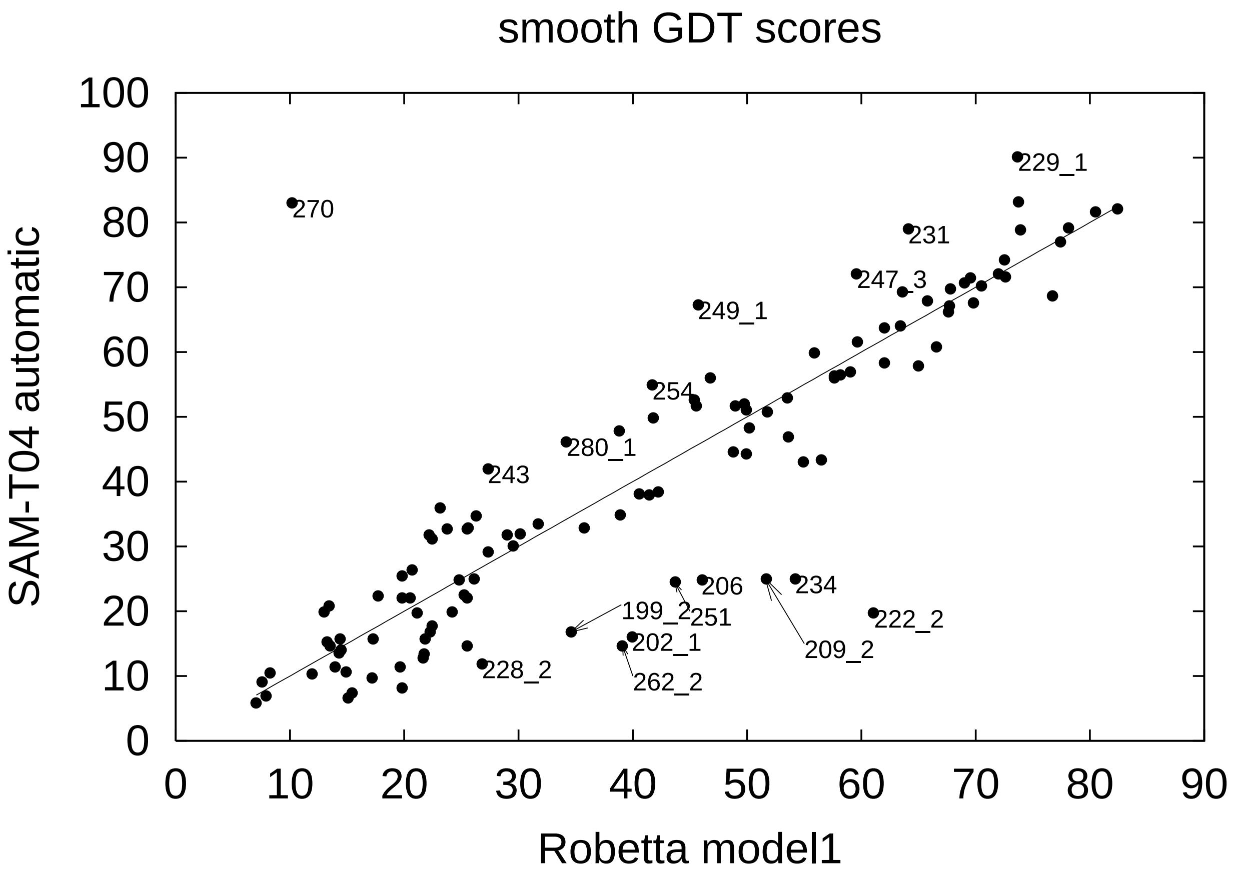
Medium-length fragments (9 long) for every position

- Generated from 3-track HMMs.
- Residues filtered to remove improbable ϕ-ψ pairs (creating smaller fragments).

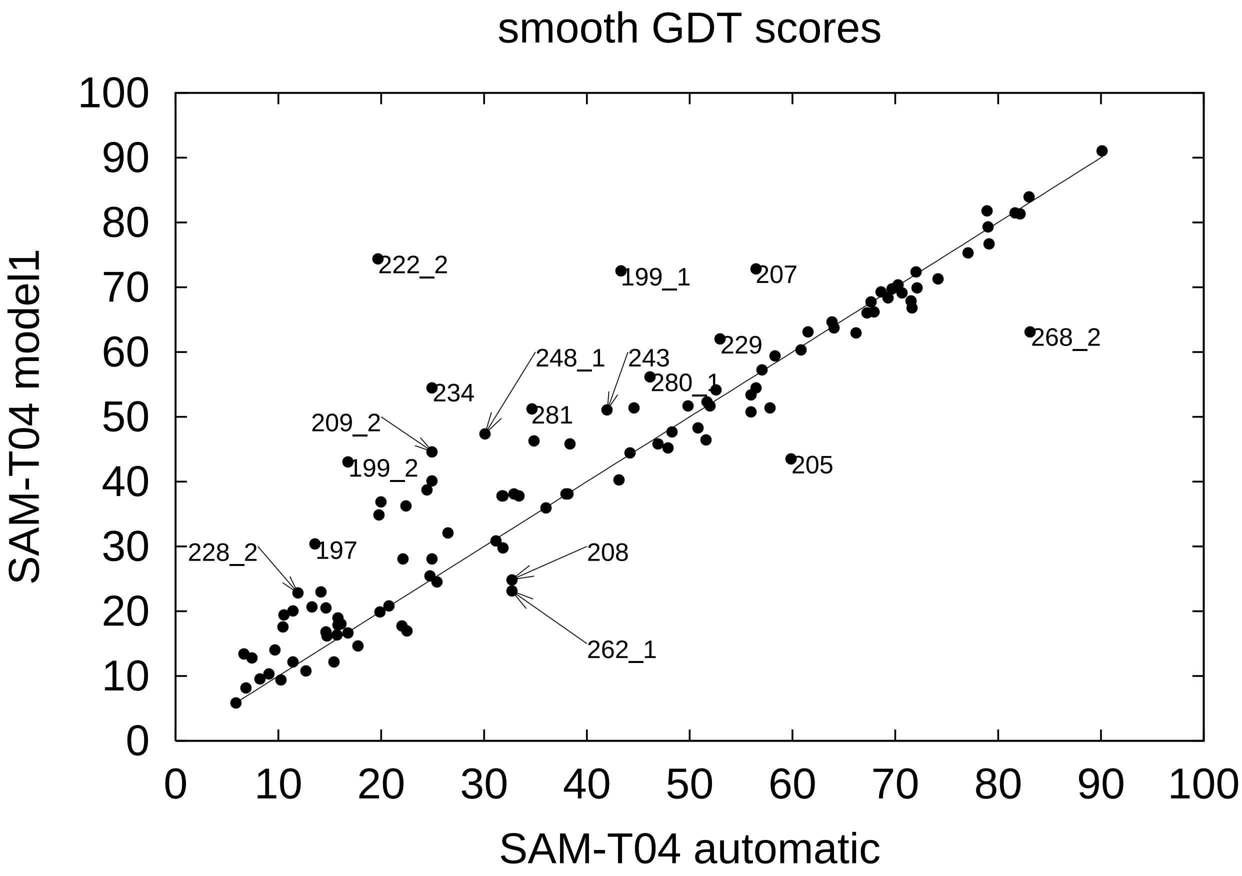
Best vs. Robetta best



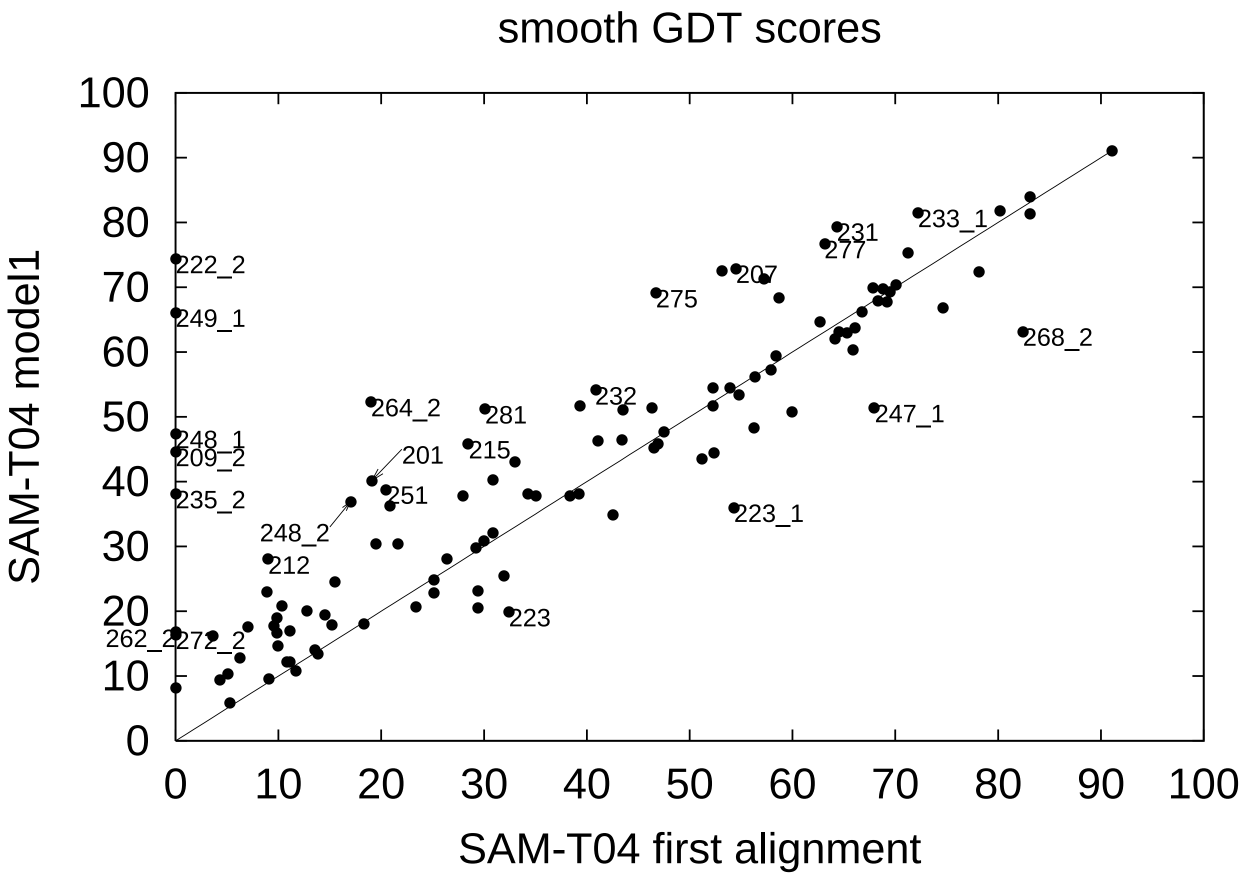
SAM-T04 auto vs. Robetta 1



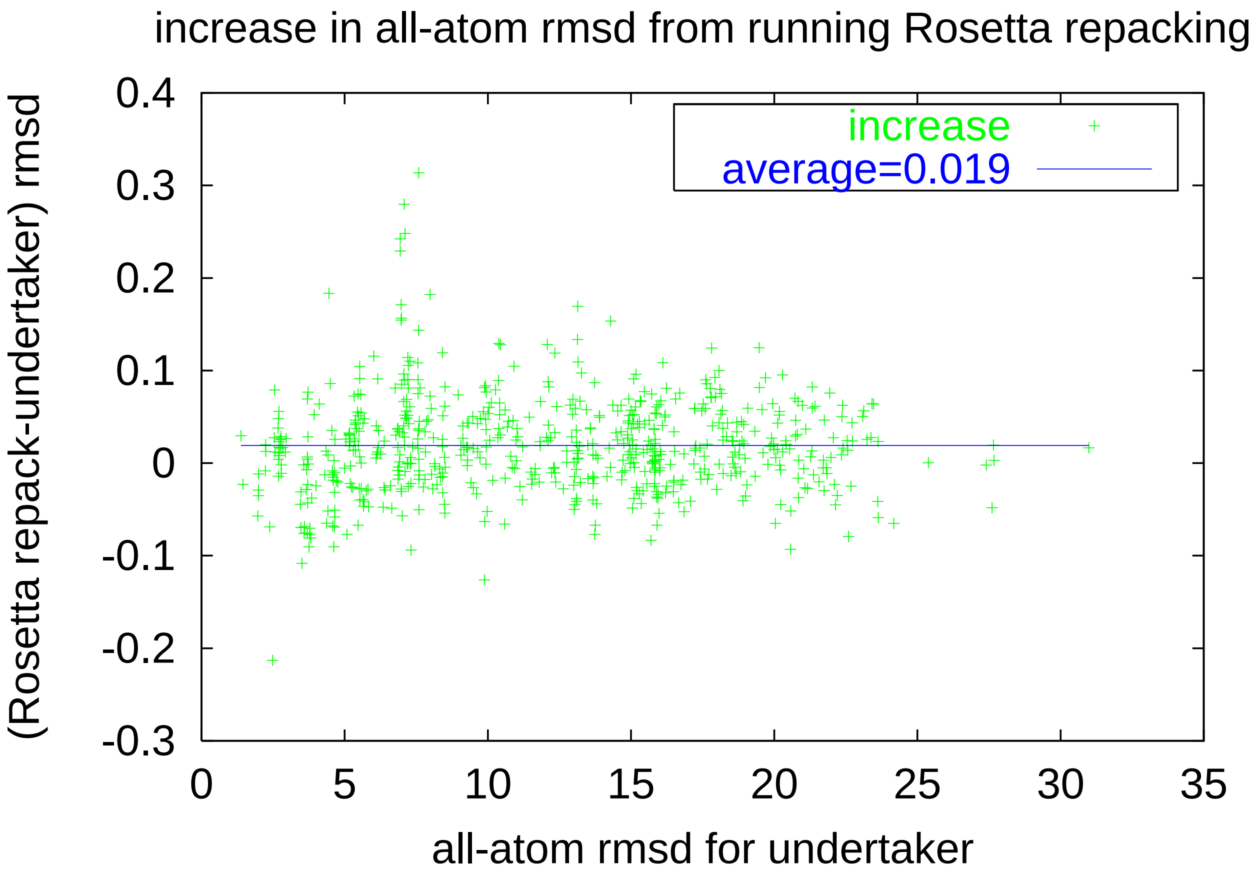
Model 1 vs. SAM-T04 auto



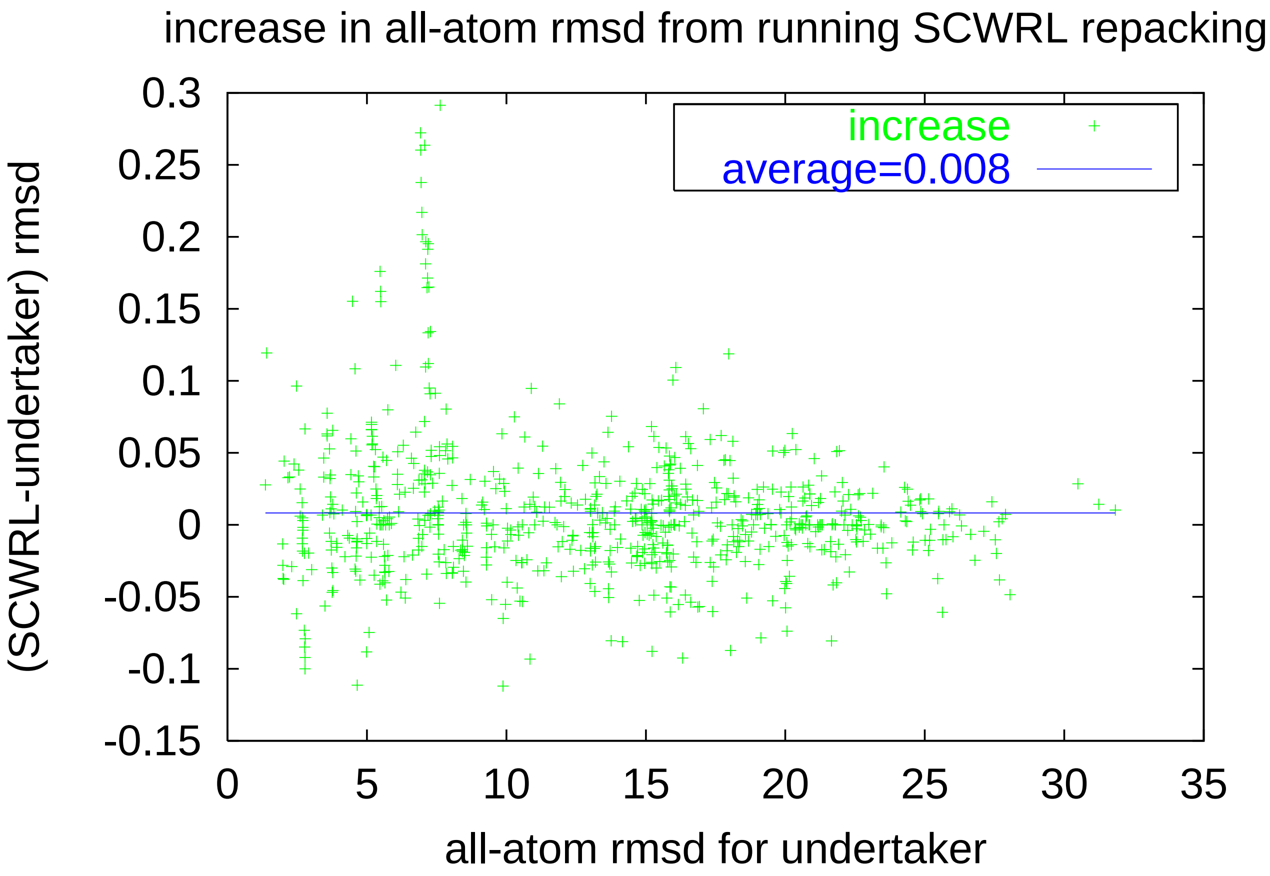
Model 1 vs. alignment



Undertaker sidechains vs. Rosetta



Undertaker sidechains vs. SCWRL

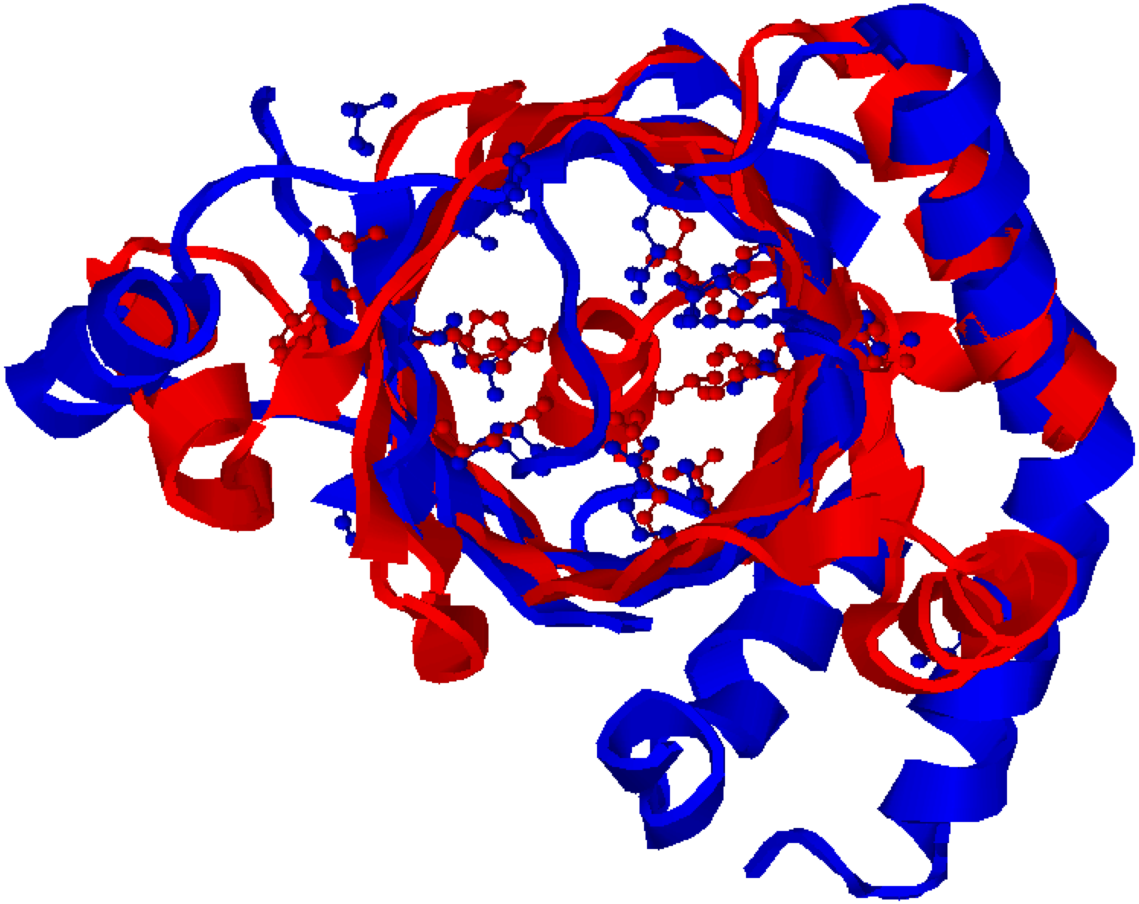


Target T0197 (FR/H)

- Robetta did surprisingly poorly for an FR/H model.
- Our scores indicated more distant relationship, and meta-servers got wrong family.
- SAM-T04’s secondary prediction better than SAM-T02’s.
- We tried assembling sheets into various barrels, based on top few fold-recognition hits.
- We used conserved residues, but not contact predictions.

Target T0197 (FR/H)

Real structure is red.

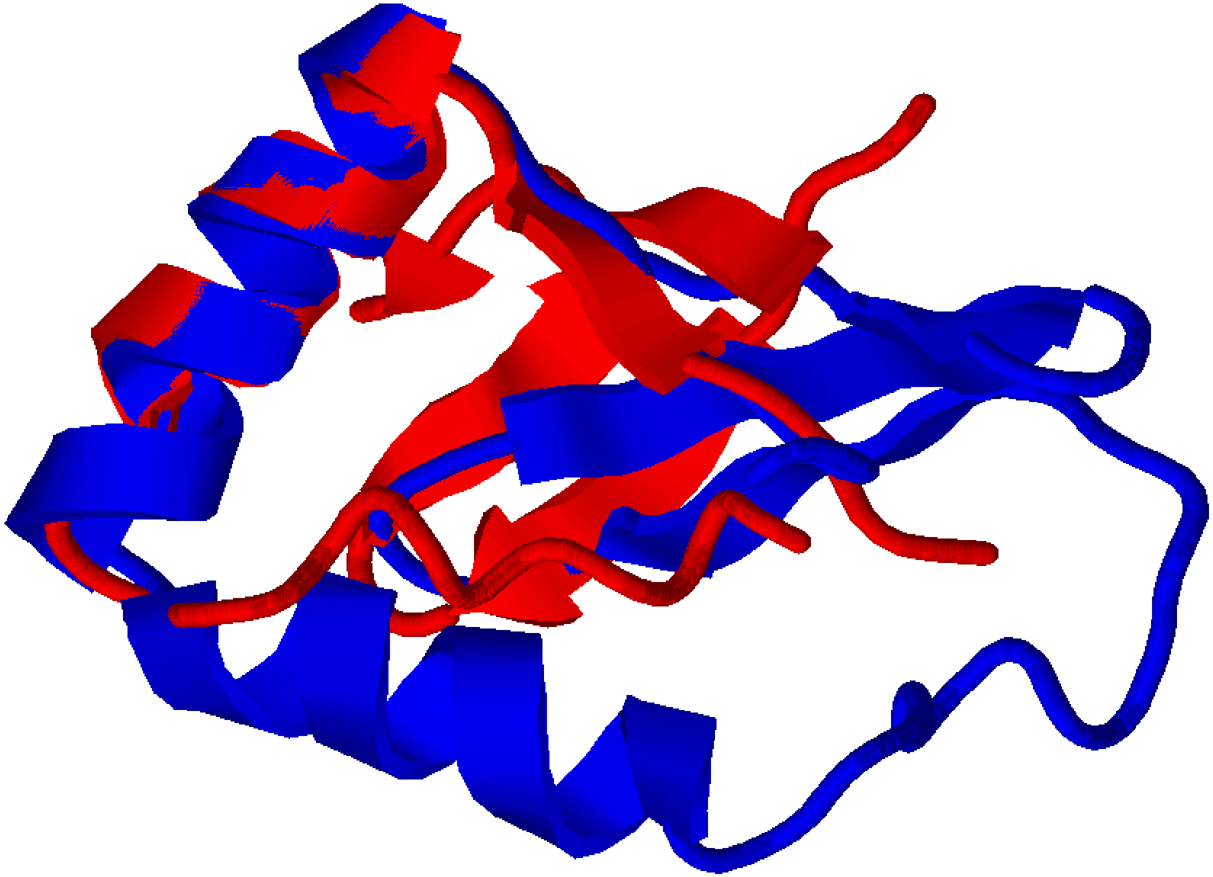


Target T0209_2 (NF)

- Our best model was try15-opt2 (model3) (5.7115 Ang all-atom RMSD).
- Good, but final strand misregistered (off by 2).
- Model is more complete than crystal.
- Sheet constraints came from robetta-model1, which outperformed it.

Target T0209_2 (NF)

Real structure is red.

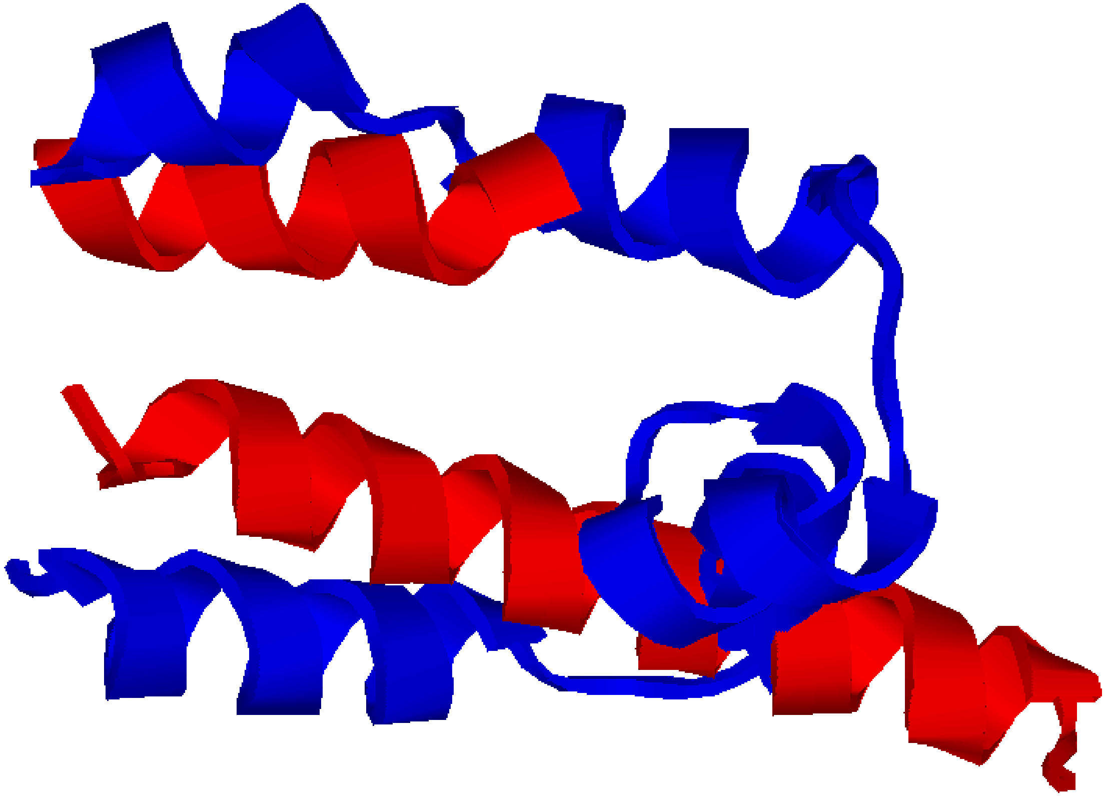


Target T0235_2 (FR/A)

- 43-residue inserted domaim—not fully resolved in crystal.
- We had made separate predictions for P347-P426, and had a good alignment to 1occJ, which we then messed up. We ended up not using the separate domain prediction.
- Good score only because first and last helix constrained by surrounding domain.
- We made last helix of domain too short, despite prediction that it was longer.

Target T0235_2 (FR/A)

Real structure is red.



Target T0248

Borrows heavily from robetta model2, which beats it.