



Identifying Structural Similarity of Proteins Using Local Descriptors of Protein Structure

Pawel Daniluk¹, Andriy Kryshchak¹, Torgeir Hvidsten^{1,2}, Jan Komorowski², and Krzysztof Fidelis¹

(1) Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA, USA

andriy@llnl.gov, fidelis@llnl.gov

(2) Linnaeus Centre for Bioinformatics, Uppsala University, Sweden



Abstract

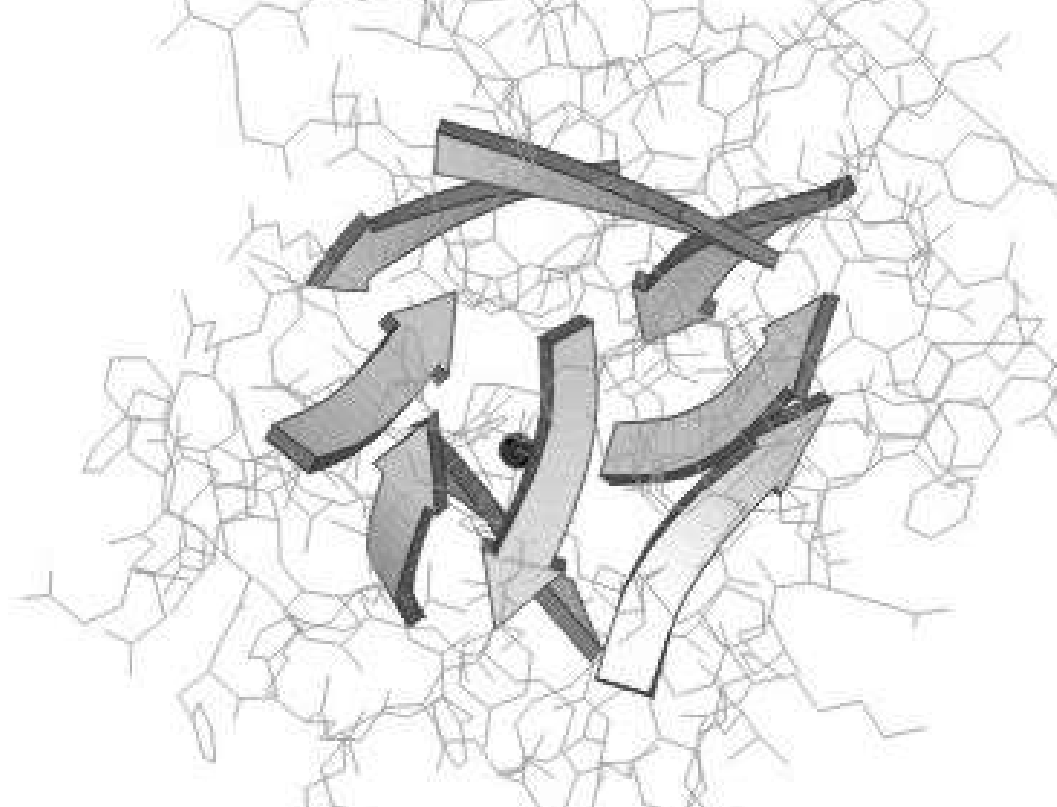
Most of the structure comparison programs are based on a "rigid body" type superposition and are therefore relatively insensitive to protein flexibility. Several approaches have been proposed to evaluate the similarity of proteins taking into account possible conformational change. Such methods reduce the risk of misaligning amino acid residues and allow a more detailed analysis of the compared structures.

We have developed an approach based on local descriptors of protein structure that is capable of performing structure comparison tasks operating with sets of short segments of protein backbone (mainly 5-7 residues long). Our method searches proteins for local similarities in a systematic way using geometric criteria only, and then estimates structural relatedness of proteins based on quality of the local structure descriptor superposition, the number of similar descriptors in both structures, their size, and location along the sequence. Optimal structure-based sequence alignment of the two proteins is also generated.

Because we are interested in recognizing remote similarities between protein structures, we have tested the method by comparing predicted protein models with the corresponding experimentally obtained structures. We have used models submitted for the CASP5 experiment in the fold recognition and new fold categories (<http://predictioncenter.llnl.gov>).

Descriptors

Descriptors are local regions of a tertiary protein structure. Each of them encompasses 3D fragments of protein's backbone and is associated with the specific residue in the protein's amino acid sequence. As opposed to supersecondary structure defined as a common combination of secondary structure elements, descriptors deal only with vicinity of one particular residue. They are organized in 5 residue long elements, of which middle residue is in contact with central residue of descriptor.

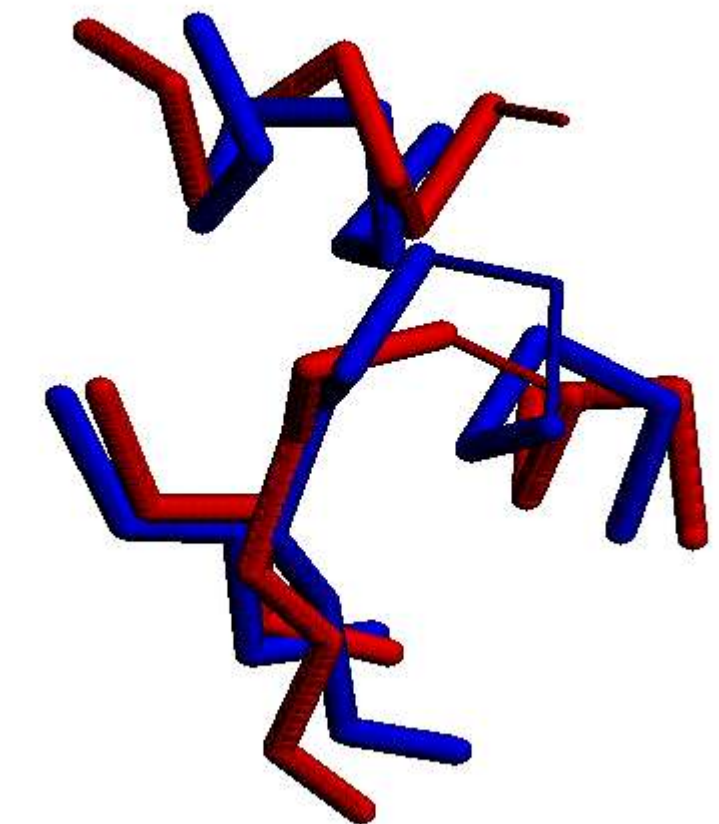


Cartoons illustrate the descriptor for residue #164 from the fibroblast growth factor 9 (PDB code 1ihk, chain A, beta-Trefoil fold b.42); the descriptor's center is shown as a small dark sphere.

Comparing descriptors

Descriptors are considered similar if there exists partial match between elements which satisfies following conditions:

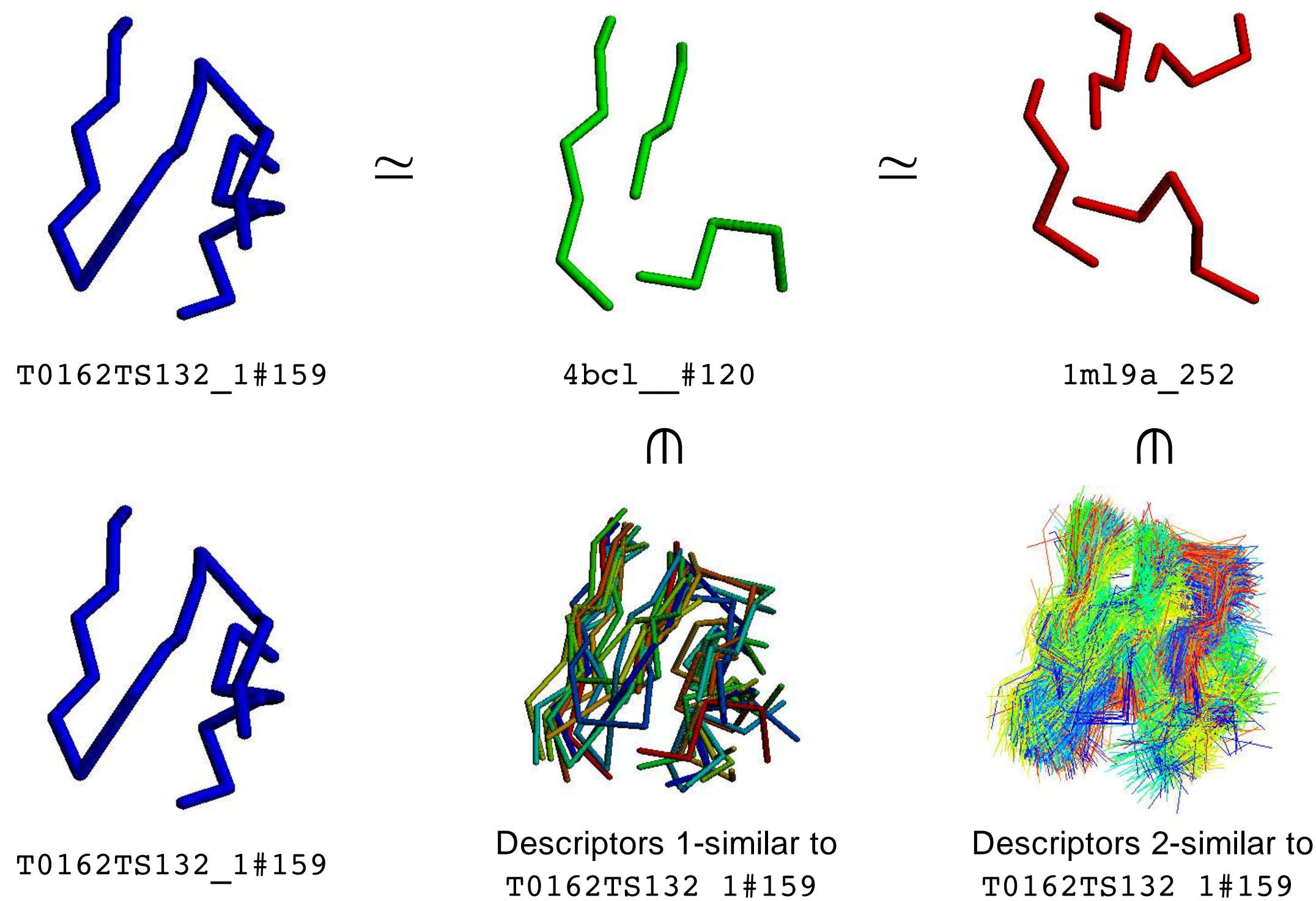
- match is consistent (it yields alignment between residues)
- central residues are aligned
- elements around central residues have RMSD no greater than 1.5 Å
- other matched elements have RMSD no greater than 2.0 Å
- RMSD of aligned residues is no greater than 2.5 Å



1qrra_#202 126-133 -NVLFAIKK 138-142 CHLVK 194-206 APTCKANGIRATD
1dlja3#343 296-304 KQIINVLKE 310-314 KVVGV 336-340 341-347 DILKS-KDIKIII

k-similarity

The similarity criteria we use has to be strict in order to prevent false matches. To detect more distant similarities we explore transitivity of similarity function. We say that descriptors d_1 and d_2 are k -similar if there exists a chain of $k+1$ descriptors, which starts with d_1 , ends with d_2 and subsequent descriptors are similar, and it defines sufficiently big (50%) alignment between first and last descriptors. Intermediate descriptors are taken from the set of 890 thousand descriptors built from ASTRAL 40 domains.

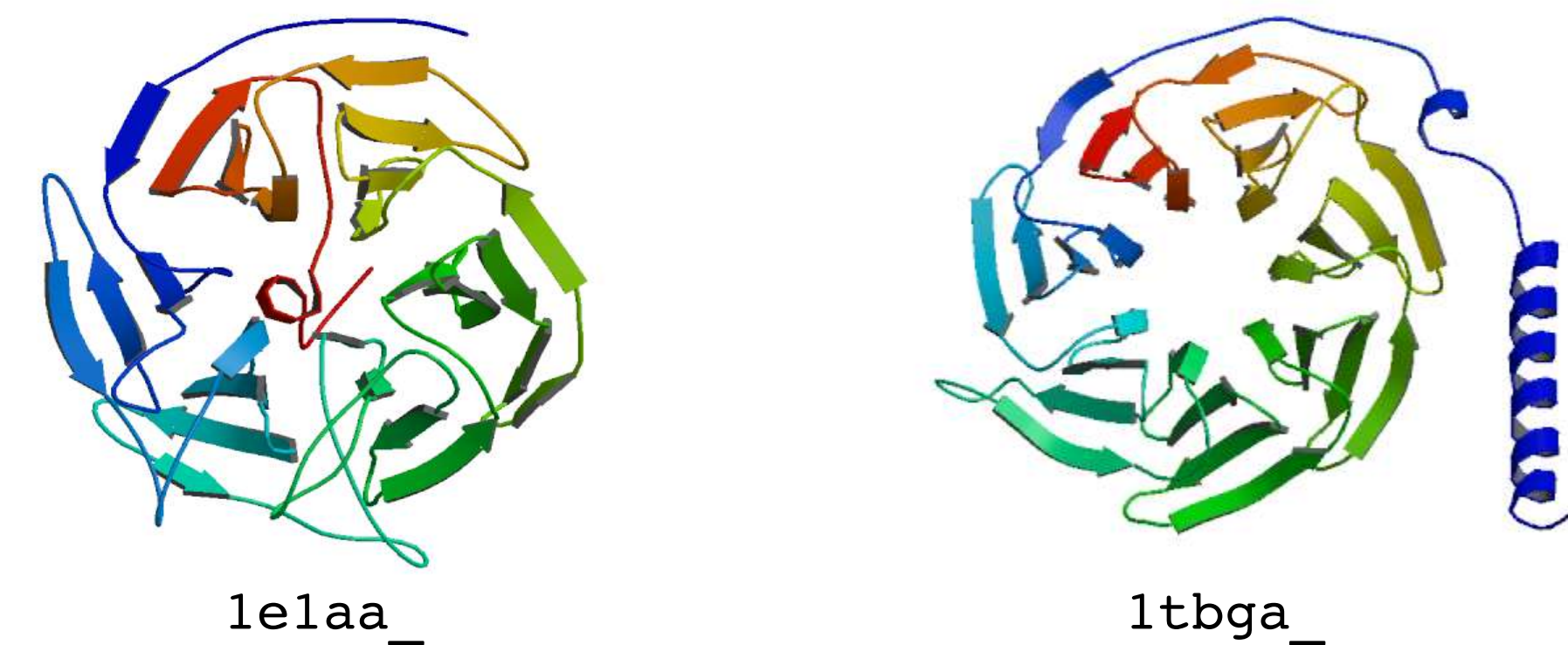
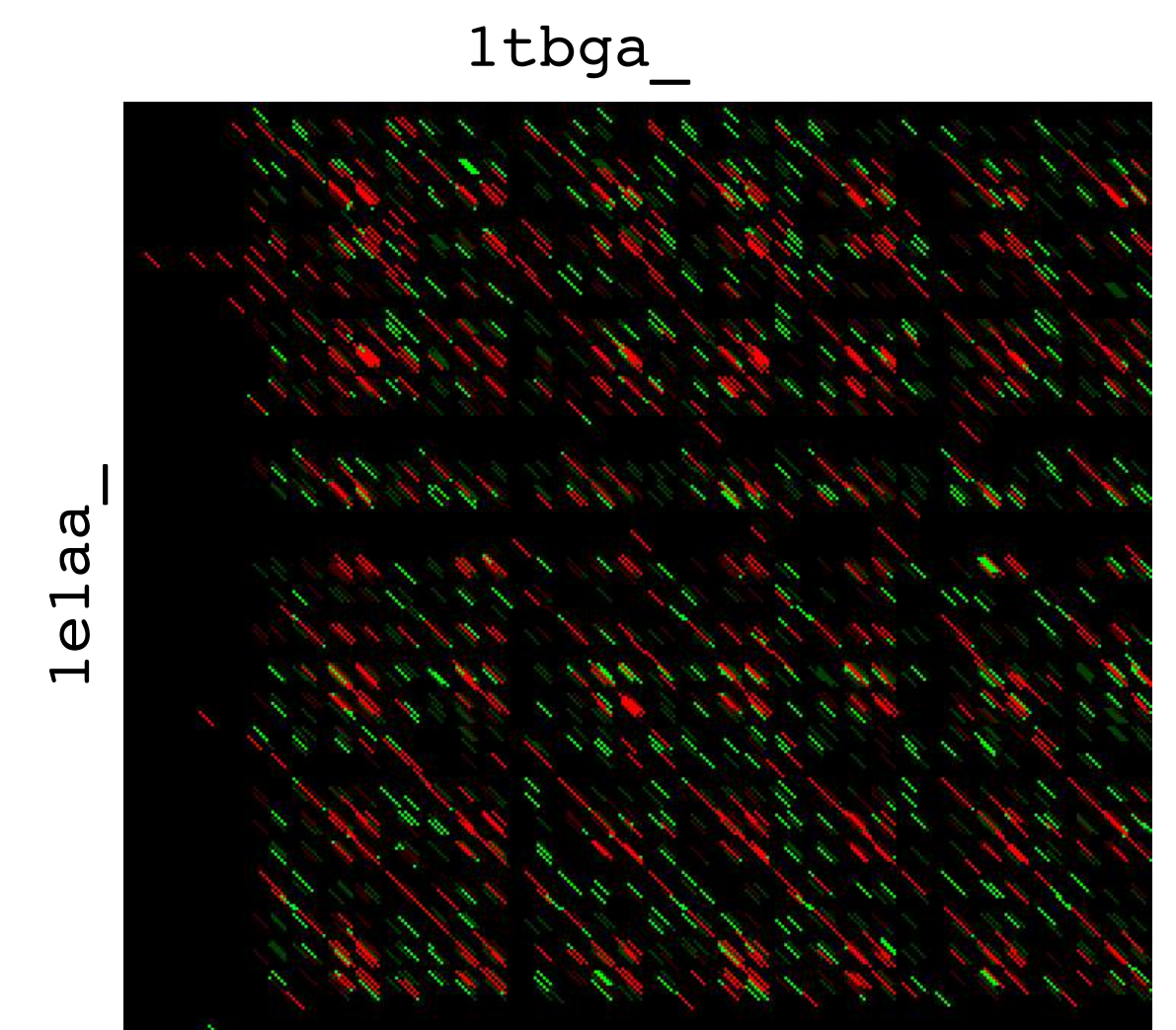


Comparing domains

To compare two structures we compute k -similarity of descriptors. Each pair of residues can belong to one or more alignments between descriptors (in the example to the right each pixel corresponds to one residue, main segments of similar descriptors are marked red, other segments are marked green). These data can be analyzed in two modes.

In sequence dependent mode we can summarize score along a given alignment (in the map to the right we can see similarity between all propeller blades – 6x7 pattern of short diagonals). We have used this mode with a trivial alignment to compare CASP5 models with respective targets.

In sequence independent approach (which is our future work) we would search for alignment with the highest score.

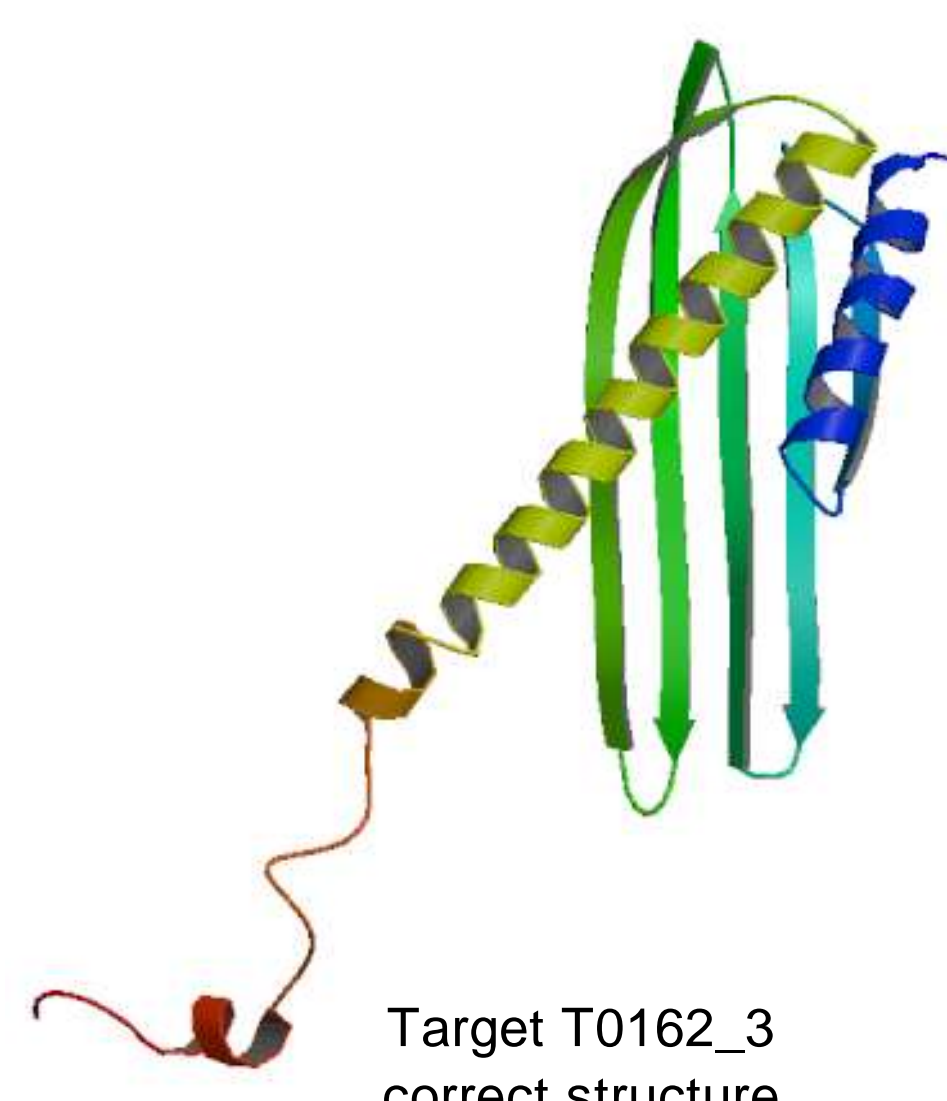


Application to CASP5 predictions

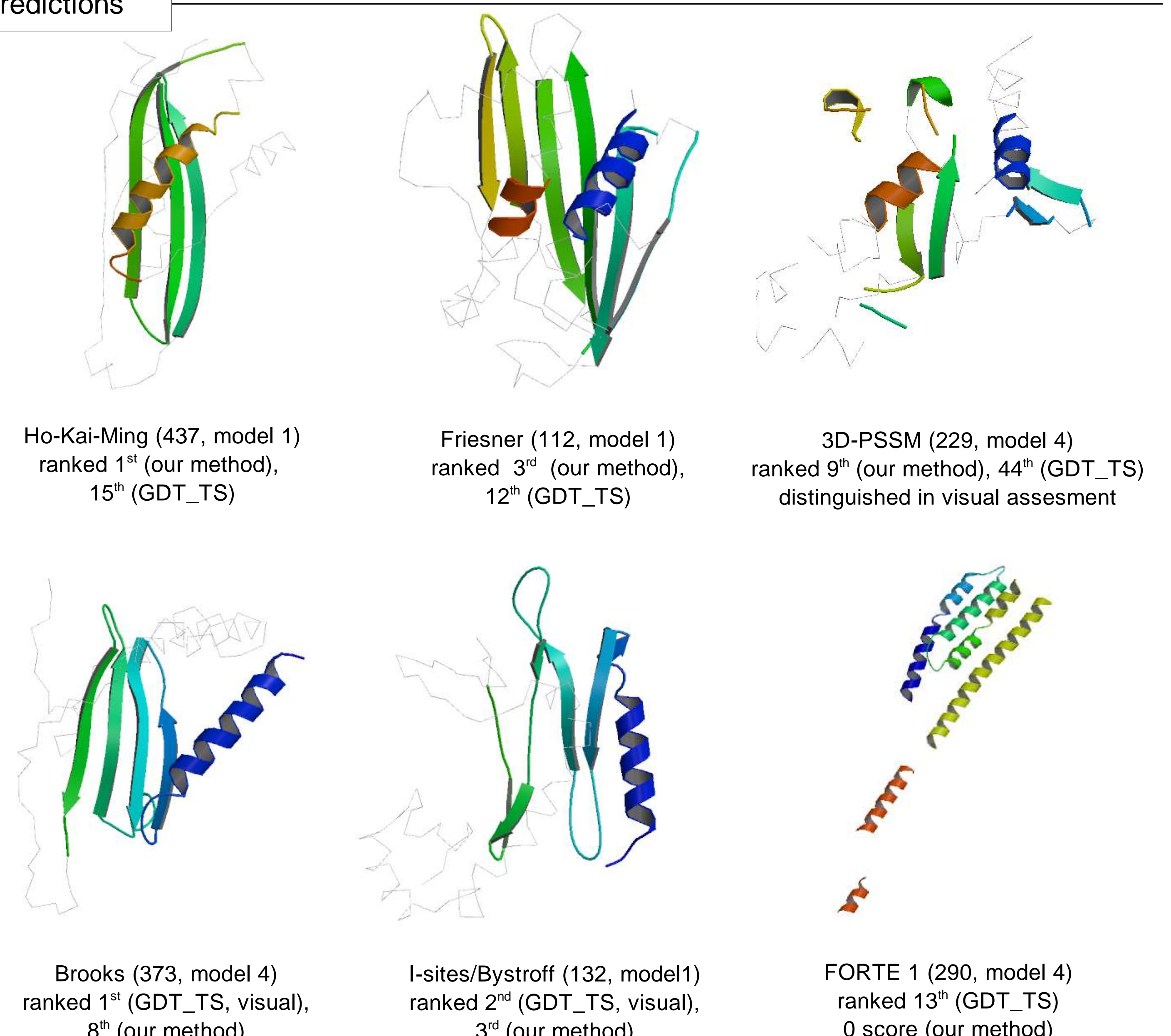
T0129 was a relatively well solved *ab initio* target. Rankings computed using our method, GDT_TS and visual assessment are almost identical. Prediction by Doniach (401, model 2) has both domains predicted correctly but severely misarranged.

Group	Model	GDT_TS	Visual	Our method
BAKER	T0129TS002_4	37.94	2	54.63
Shortle	T0129TS349_1	23.68	1	46.56
Jones-NewFold	T0129TS068_4	29.26	1	33.37
SAM-T02-human	T0129TS001_3	26.77	1	30.07
Scheraga-Harold	T0129TS314_1	24.70	0	29.30
Brooks	T0129TS373_3	19.85	0	21.29
Doniach	T0129TS401_2	26.17	1	20.00
BAKER-ROBETTA	T0129TS029_3	22.80	0	17.65
I-sites/Bystroff	T0129TS132_5	16.18	0	14.62
Skolnick-Kolinski	T0129TS010_1	22.65	0	12.57
ORNL-PROSPECT	T0129TS012_1	19.41	0	8.24
SAMUDRALA-NF	T0129TS051_5	20.00	0	7.06
PROTINFO-AB	T0129TS140_5	20.00	0	7.06
Pmodel3	T0129TS045_1	22.35	0	5.00
Levitt	T0129TS016_1	22.06	0	0.00
ATOME	T0129TS464_3	21.47	0	0.00
Dunbrack	T0129TS327_1	20.88	0	0.00
GeneSilico	T0129TS517_1	17.64	0	0.00
3D-PSSM	T0129TS229_4	17.50	0	0.00
BioInfo.PL	T0129TS006_1	17.50	0	0.00
Bilab	T0129TS080_1	16.18	0	0.00

There were no accurate predictions of target T0162_3. Accurate predictions of β -sheet without C-terminal α -helix had relatively high GDT_TS score. Our method promotes predictions with at least partial prediction of the long helix, correctly positioned across β -sheet. Also prediction 290_4 which scores high in GDT_TS, scores 0 in our contact-based method.



Target T0162_3 correct structure



Ho-Kai-Ming (437, model 1) ranked 1st (our method), 15th (GDT_TS)

Friesner (112, model 1) ranked 3rd (our method), 12th (GDT_TS)

3D-PSSM (229, model 4) ranked 9th (our method), 44th (GDT_TS) distinguished in visual assessment

Brooks (373, model 4) ranked 1st (GDT_TS, visual), 8th (our method)

I-sites/Bystroff (132, model 1) ranked 2nd (GDT_TS, visual), 3rd (our method)

FORTE 1 (290, model 4) ranked 13th (GDT_TS), 0 score (our method)