

Identifying Model-Target Similarity of Proteins Using Local Descriptors of Protein Structure



Pawel Daniluk, Lukasz Szajkowski, Andriy Kryshchovych, and Krzysztof Fidelis

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

andriy@llnl.gov, fidelis@llnl.gov

Abstract

Most of the structure comparison programs are based on a "rigid body" type superposition and are therefore not very well suited to track structural deformations or domain re-orientation. They also tend to perform poorly when asked to identify model-target similarity for the most difficult prediction targets.

We have developed an approach based on local descriptors of protein structure (Kryshchovych & Fidelis, in prep.) that is capable of performing structure comparison tasks operating with residue-attached sets of short segments of protein backbone (mainly 5-7 residues long). Our method searches for local similarities in a systematic way using geometric criteria only, and then identifies all such regions of similarity in a 2D (sequence, sequence) map and in 3D superpositions of model and target. It also reports the total number of residues that are part of such regions of similarity.

Since the local structure of the target protein is used in all comparisons, the approach preserves the structural context and is therefore impervious to model compaction or internal clashes that some other approaches are capable of misinterpreting. Results are reported for all FR and NF CASP6 targets.

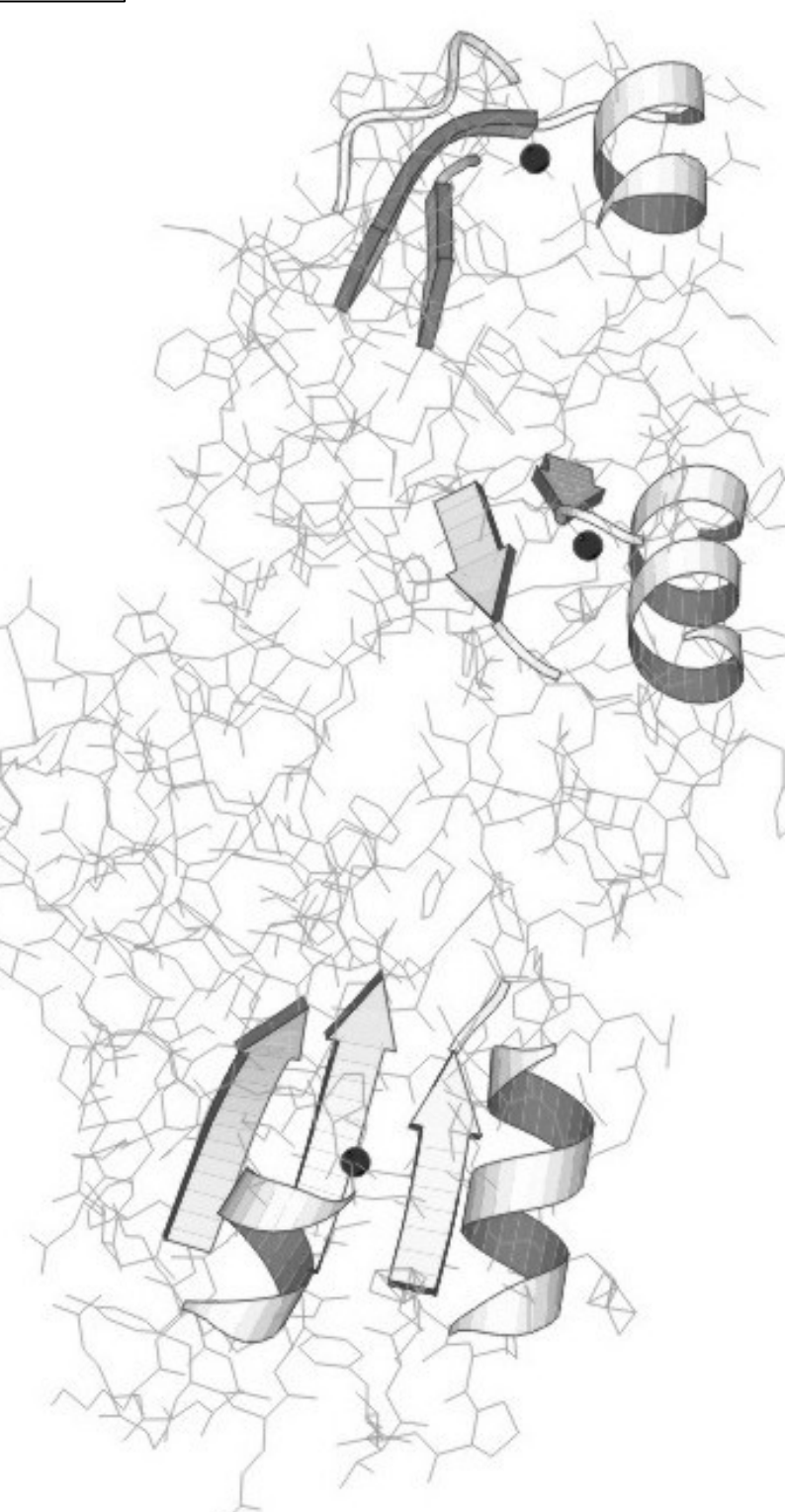
Descriptors

Descriptors are local regions of a tertiary protein structure. Each descriptor 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 the vicinity of a particular residue.

Descriptors are constructed from 5-residue-long fragments of a polypeptide chain (*elements*) that

are close in space relative to the central residue. Elements that overlap in sequence are joined into continuous *segments*.

Cartoons illustrate 3 selected descriptors (centered on residues: 79 - bottom, 173 - middle and 192 - top) of the anaerobic cobalt chelatase (PDB code 1qgo, chain A). C β atoms of the central residues of the descriptors are shown as small dark spheres.



Comparing descriptors

When comparing descriptors we consider all partial bijections between their elements that induce bijections between amino acids (i.e. they have to be valid alignments). We compute all such maximal alignments that satisfy the following conditions:

- optimal superposition of elements containing central residues of the descriptors yields $\text{RMSD} < 1.5 \text{ \AA}$
- all other corresponding elements have $\text{RMSD} < 2.0 \text{ \AA}$
- total RMSD of descriptors on all corresponding fragments is $< 2.5 \text{ \AA}$

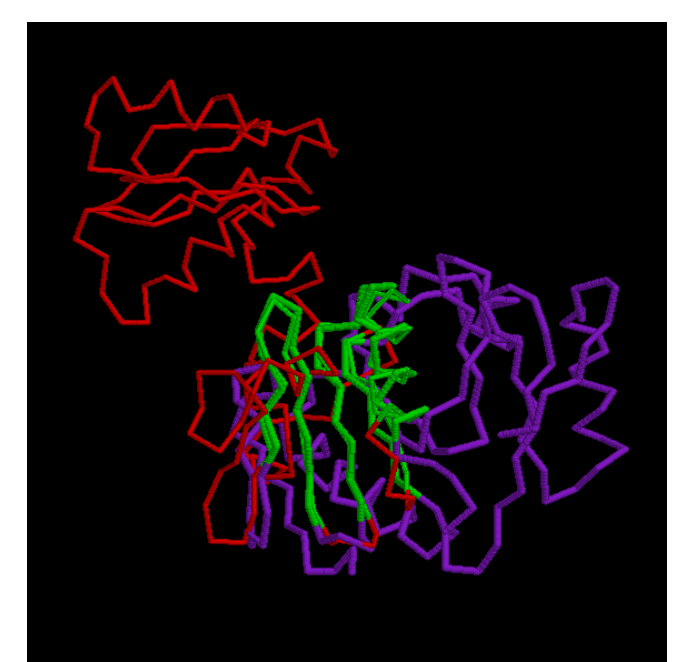
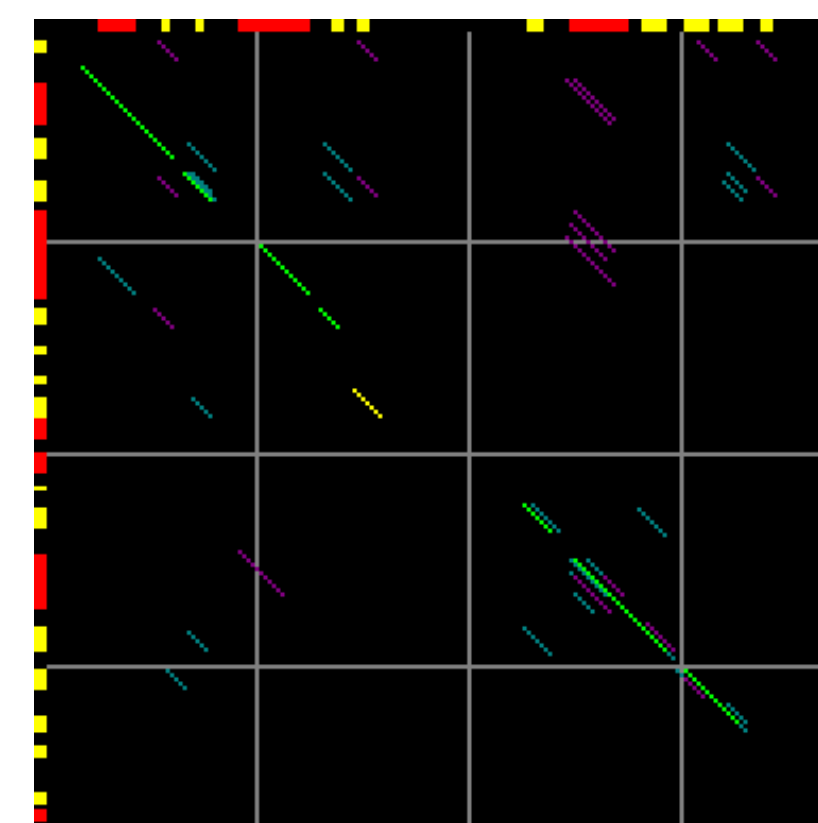
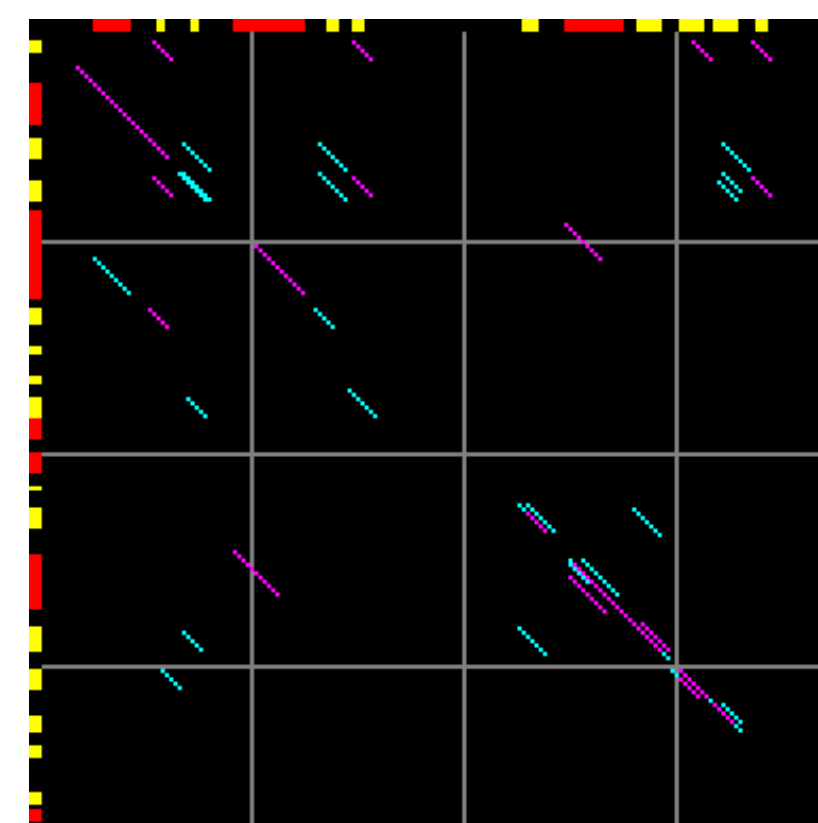
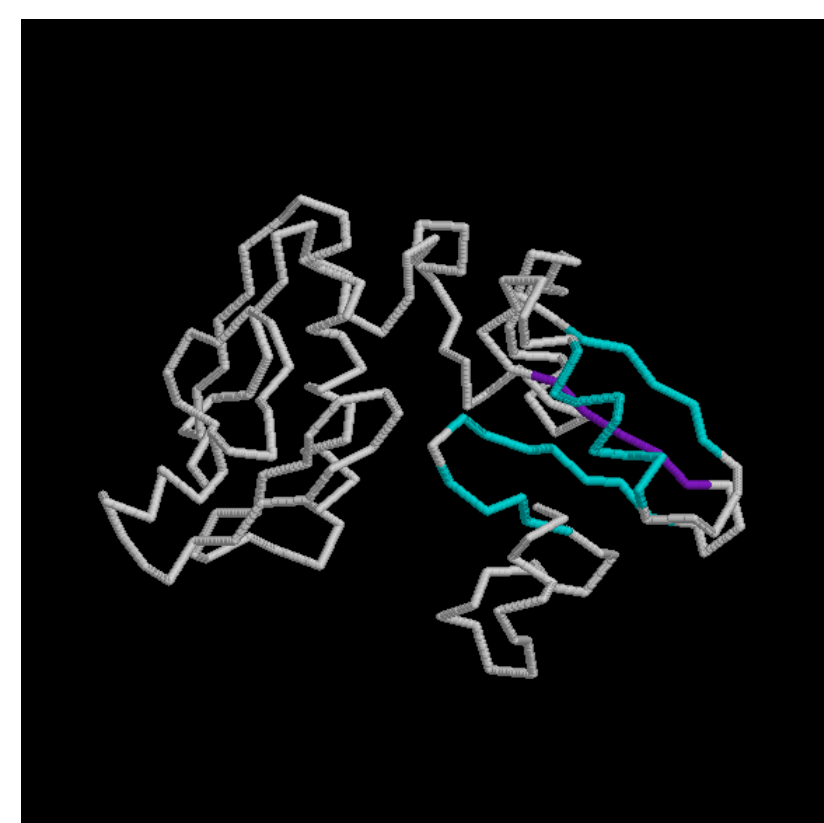


Aligning structures

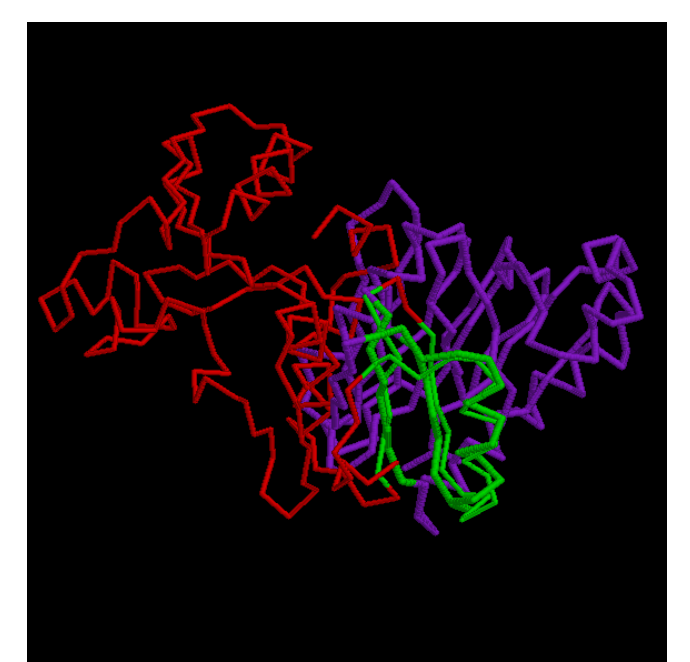
Similar descriptors are identified in both structures. Only descriptors having at least 3 segments are considered.

Optimal alignment is found. Alignment is a maximum non-contradictory set of detected similarities. This can be seen as a clique finding problem, and may require exponential time. Fortunately small size makes it bearable. Structure comparison can be performed in sequence dependent, independent and semi-dependent (i.e. with limited shift allowed) modes.

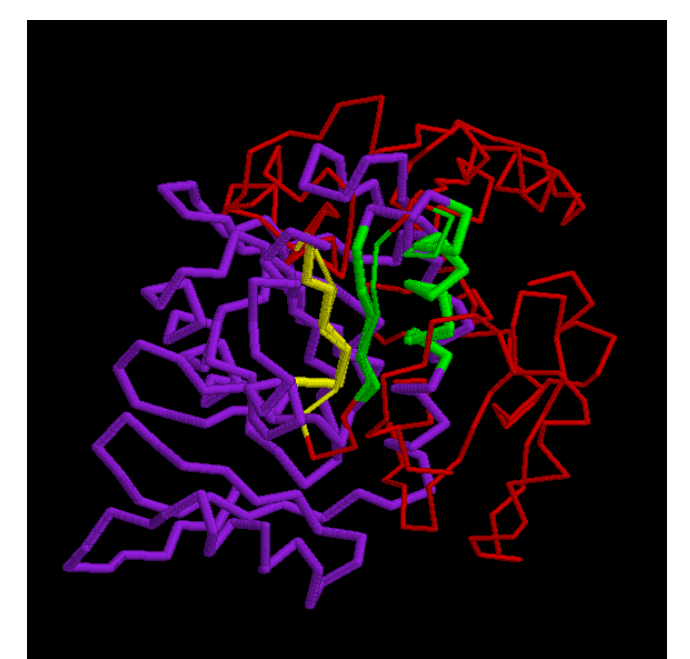
Alignment does not imply that there exists a single rigid-body superposition. In this case there are three locally similar regions



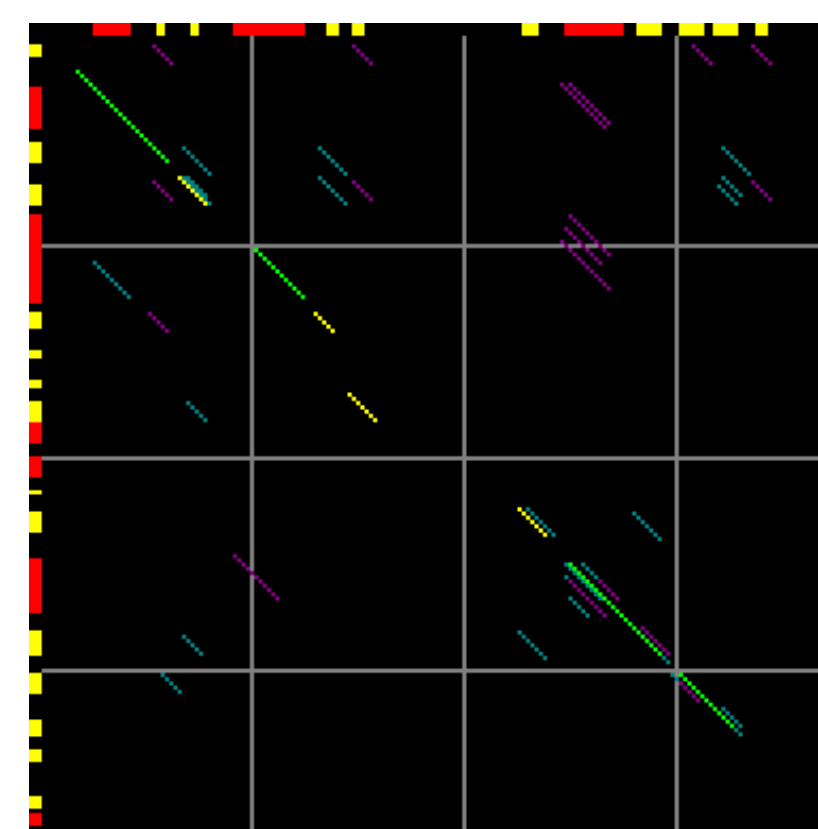
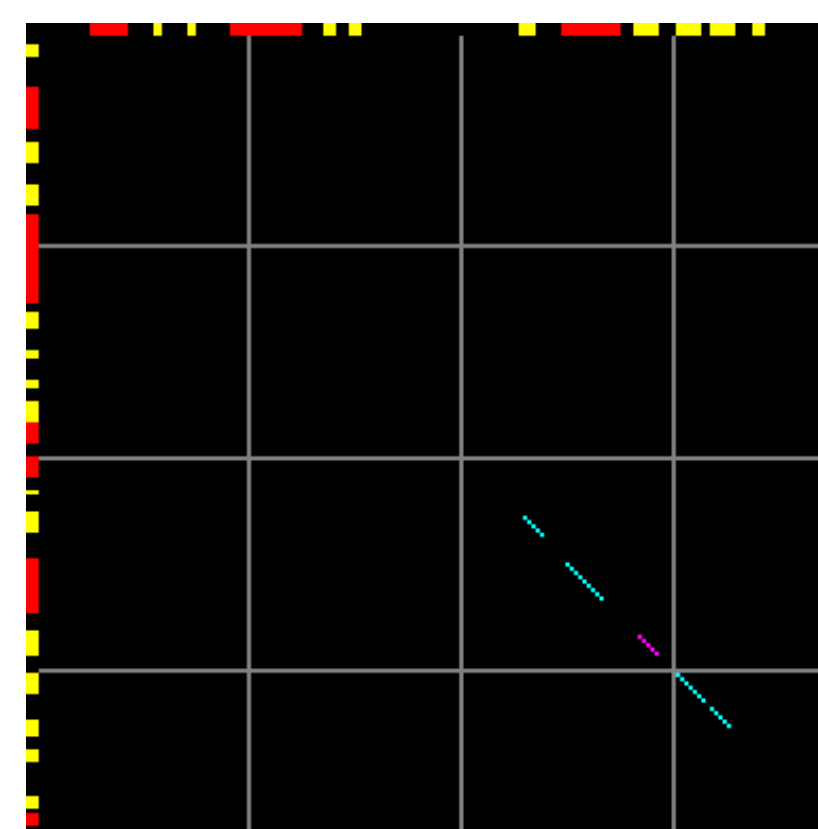
Blue



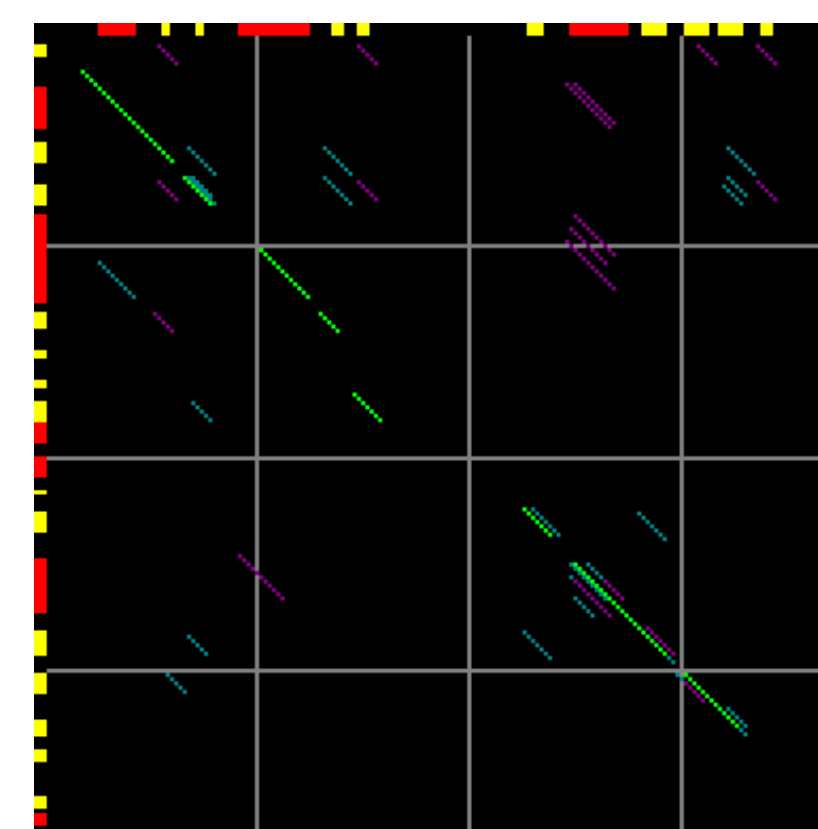
Green



Red



DAL_4



DAL_I



Model

Gallery of examples

Target T0201

Descriptor type alignments are capable of detecting slight relative shifts of secondary structure elements, as well as erroneous connectivity of secondary structure elements (not shown).

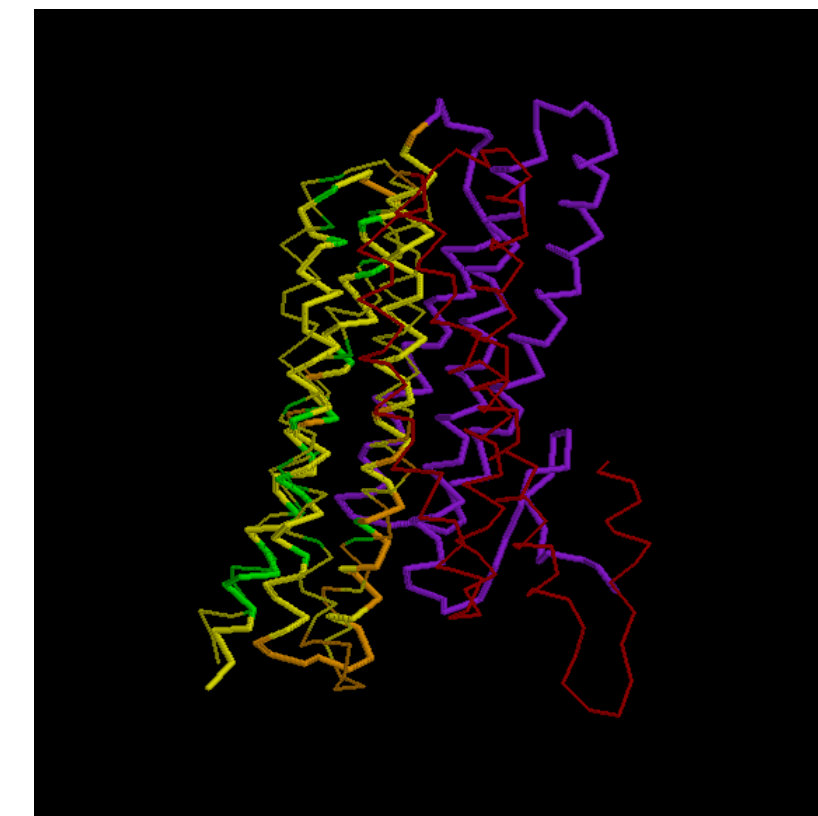
GDT - 51.06 (right)
DAL_0 - 19.15 (lower left)
DAL_4 - 57.45 (lower right)



Target T0201

Descriptor type alignments detect two separate similar local substructures. In general it is possible to detect erroneously oriented subdomains.

GDT - 32.11 (right)
DAL_4 - 78.22 (bottom)
(Target - left, Model - right)



Target T0273

Descriptor type alignments require local structure similarity, while methods that extend alignments without such constraints often fail to capture structural context.

GDT - 22.31 (left)
DAL_4 - 7.53 (right)

