Poster Presentation

MS055.P03

Shape descriptors for domain classification and structure solution

Michal Tykac[1], Robert A. Nicholls[1], Fei Long[1], Garib N. Murshudov[1]
[1]*MRC Laboratory Of Molecular Biology, Cambridge, United Kingdom*
E-mail: mtykac@mrc-lmb.cam.ac.uk

The BALBES database [1] contains over 13, 000 protein domain structures derived from the PDB. Since this database was built specifically for use in a molecular replacement pipeline, the focus of the domain classification was on structural similarity. This feature makes the database amenable for use in such shape-based applications as finding molecular replacement candidates using Patterson map matching (instead of sequence matching) or automatic fitting of domains into cryo-EM maps.

Both applications are computationally expensive due to their requirement for multiple searches against the whole database and for this reason are not used routinely. Here we present two translationally and rotationally invariant fast shape descriptors based on spherical harmonic map decomposition. The advantage of these shape descriptors is twofold: 1) they can be used for reclassification of the BALBES domain database, thus making its use more efficient; and 2) these descriptors can be used as fast pre-filters for finding matches between a query map (be it determined by X-ray crystallography or cryo-EM) and the BALBES database.

The shape descriptors presented here require transforming the query data into a
Patterson-like map, which is subsequently interpolated onto a set of concentric spheres (shells) serving to retain the radial information, while the angular information is represented using the spherical harmonic decomposition of each shell. One of the shape descriptors then uses the correlation between spherical harmonic coefficients from different shells to describe the shape, while the other descriptor borrows from the rotation function [2] optimisation and uses the singular value decomposition to obtain shape descriptors. Both methods result in rotation invariant shape descriptors, while the use of Patterson-like maps ensures translation invariance. Combination of these features removes the requirement for a six-dimensional search, replacing it with matrix comparisons, thus making these methods computationally efficient.

It should be noted that the calculation of the presented descriptors is irreversible, meaning that there is a loss of information. This feature is a trade-off between computational speed and accuracy. The optimisation of parameters for descriptor calculation was performed using over 100 protein domain groups from the BALBES database; each group contains different, but very similarly shaped domains. The resulting Area Under the Receiver Operator Characteristic curve (AUROC – which is closely related to Mann-Whitney U test, but has more interpretable value [3]) values were found to be 0.998 and 0.997 in our tests, respectively. While these results show the ability of both descriptors to distinguish between similar and different protein domain shapes, the information loss means that they should be used as pre-filters for one-against-many domain structure comparisons, rather than as a replacement for existing methods.

[1] Long F., Vagin A.A., Young P. and Murshudov G.N. (2008). Acta Cryst, D64, 125-132.
[2] Navaza J. (1994). Acta Cryst, A50, 157-163.
[3] Hanley J.A. and McNeil B.J. (1982). Radiology, 143, 29-36.

Keywords: PROTEIN DOMAINS, SHAPE DESCRIPTORS, MOLECULAR REPLACEMENT