Journal of Structural Biology 183 (2013) 47–56

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Automatic comparison and classification of protein structures

Janne Ravanti [a,*], Dennis Bamford [a], David I. Stuart [b,c]

[a] Institute of Biotechnology and Department of Biosciences, University of Helsinki, Finland
[b] Division of Structural Biology and The Oxford Protein Production Facility, The Wellcome Trust Centre for Human Genetics, University of Oxford, UK
[c] Diamond Light Source Ltd., Harwell Science and Innovation Campus, Didcot, Oxfordshire, UK

ARTICLE INFO

Article history:
Received 25 January 2013
Received in revised form 2 May 2013
Accepted 8 May 2013
Available online 21 May 2013

Keywords:
Structure comparison
Virus coat proteins
Virus evolution

ABSTRACT

The classification and alignment of multiple three-dimensional protein structures is a powerful way to detect similarities that cannot be discovered from the sequences alone and can help to infer phylogeny. However, the alignment process remains problematic for divergent structures. We have devised a fully automatic pipeline, HSF, drawing its inspiration from well-known structural alignment methods, which given a list of structures not only aligns all pairs but also classifies them fully. We demonstrate proof of principle for the new method by aligning the currently available set of highly diverged virus coat protein structures containing double β-barrels, as well as validating the method with established test sets for multiple structural alignments. The results for the virus proteins are inline with previous observations based on biochemical, genetic and structural studies but go further, since by providing coherent alignments between sets of molecules with marked structural distortion, they facilitate the marshaling of arguments for or against homology. The classification results can therefore be readily interpreted in terms of phylogeny.

1. Introduction

Methods for determining similarities between three-dimensional (3D) protein structures are of considerable importance. Of particular interest to us is the observation that when the differences between proteins are so extensive that their amino acid sequences cannot be reliably compared with currently available methods, we can often still perform useful comparison between the structures, which diverge more slowly than the sequence. Such comparisons have a wide range of uses, giving unique insights from theoretical studies on evolution and the origins of life to structure–function predictions of utility for the discovery of new drugs.

However, comparing 3D objects, such as flexible protein structures is exceptionally challenging (Rossmann and Argos, 1976). A number of methods have been developed to address this fundamental issue, ranging from early methods like HOMOLOGY (Rao and Rossmann, 1973), COMPARER (Sali and Blundell, 1990) and Dali (Holm and Sander, 1993) to more recent algorithms including CE (Shindyalov and Bourne, 1998), FATCAT (Ye and Godzik, 2003), SSM (Krissinel and Henrick, 2004), MUSTANG (Konagurthu et al., 2006) and CLICK (Nguyen and Madhusudhan, 2011). These methods use a variety of approaches, some working at the level of cartesian coordinates, often of Cα atoms, whilst others work with derived properties such as secondary structure, local chain shape or accessible surface area, the usual final action of the comparison is a least squares superposition of Cα atoms deemed to correspond, in order to define the best fit between a pair of structures. For a recent review, see Hasegawa and Holm (2009). The superimposition problem is further complicated when considering not just a comparison of a pair of structures but the analysis of common features among a set of structures, since pair-wise comparison alone will not generate a coherent assignment of equivalent residues across the whole set, which is formally required if we are to derive phylogenetic information from such alignments.

During our research, directed towards the classification of virus families on the basis of virion structure, we found that the currently available methods failed to detect reliably and automatically structural similarities that could be seen by eye and also by the careful application of the program SHP ("Structure Homology Program", Stuart et al., 1979; for a typical example of such a mis-alignment see Supplementary S1 panel A). Detection of these more distant similarities allowed us to show close relationships between various viral coat proteins previously thought to be unrelated (see below). Consequently, we have developed HSF ("Homologous Structure Finder"), a prototype computational platform capable of assimilating various comparison methods and parameterizations. This platform has been validated by applying it to one of the more difficult classification tasks we have come across, for which, as we show below, it successfully performed an automatic

[*] Corrresponding author. Address: Institute of Biotechnology, PL 56, Viikinkaari 5, 6007, University of Helsinki, 00014, Finland. Fax: +358 9 19159098.
E-mail address: Janne.Ravanti@helsinki.fi (J. Ravanti).

1047-8477/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jsb.2013.05.007

comparison and classification of the set of 3D structures (the failure of existing software shown in S1 panel A is for a comparison within this set). The starting point for the method's development was the program SHP (Stuart et al., 1979) which has proven successful for classifying various diverse structures (Abrescia et al., 2010) and has been used as the basis for constructing structure-based phylogenetic trees (Riffel et al., 2002). SHP uses dynamic programming to infer the most probable equivalent residues from a pair of structures, using the criteria for similarity defined by Rossmann and Argos (1976).

Our method not only performs pair-wise comparisons but has an additional level of sophistication, since given a set of structures it starts from a full set of pair-wise comparisons and performs a recursive analysis, successively merging the two closest structures to gradually reduce the pool size and eventually yield a full hierarchical classification. To understand how this is achieved we need to define the concept of structural "cores". At each level of the hierarchy of the comparison (the levels are defined automatically by the program but tend to map onto well understood taxonomic concepts such as order or family) there are groups of structures which have been classified together into that branch. Each group of structures is classified together on the basis of a combined structural core defining the set of residues found, by HSF, to be equivalent across all structures of that group within the hierarchy. A full classification therefore requires establishing a series of such combined cores, which will diminish in size as the group expands to encompass more diversity. The determination of equivalence, which underpins the identification of these cores, depends on both which properties of residues are used to measure similarities (e.g. Cα-distances of superimposed structures, size and charge similarities of the compared residues, etc., see Table 2 for the full list) and also on how they are combined together (Fig. 2). Combining different kinds of information to determine equivalent residues produces more robust results when comparing flexible structures where the overall similarity cannot be properly characterized with strict geometrical metrics alone (e.g. RMSD of equivalent residues; see results).

The in-built classification process produces, in the cores, exactly the information required for automatically inferring a plausible phylogenetic tree. As we describe below not only is the method as good, in terms of typical accuracy metrics (e.g. geometrically defined measures such as the number of equivalent residues and their RMSD, or the number equivalent residues within a pre-defined cutoff distance), as the best existing methods at performing pairwise comparisons, but the recursive multiple comparison method produces trees which agree with already established classification and structural similarities identified manually by structural biologists, on a case-by-case basis. The recursive method

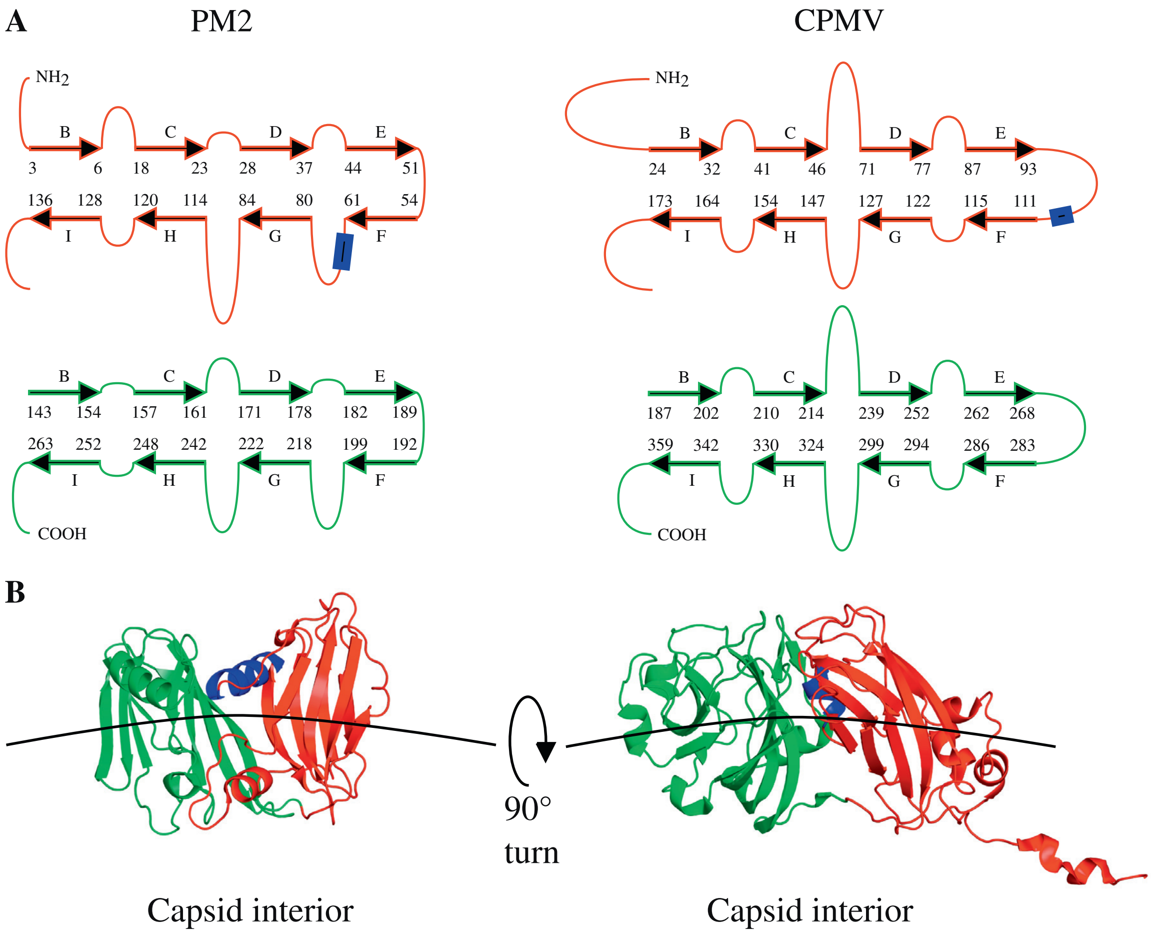

Fig.1. Visualizations of selected capsid proteins from PM2 and CPMV. (A) Arrangement of β-strands as diagrammatic representations. The first and second β-barrels are colored red and green, respectively. The β-strands are shown as arrows, labeled in canonical order, numbered at the beginning and at the end. The topological difference between the functionally analogous α-helices in PRD1-type capsid proteins and Comovirus capsid proteins in the first β-barrels is shown with a blue bar in PM2 and CPMV, respectively. (B) Cartoon diagrams of the capsid protein structures. Coloring as in (A). The proteins are viewed along virus capsid surfaces towards the 5-fold axis. PM2 exhibits an upright orientation whereas CPMV lies nearly flat on the icosahedral facet highlighted with a 90° degree arrow. The relative orientation of the capsid protein to the capsid itself affects how the protein and its double-β barrels form contacts to the neighboring proteins (see text). The secondary structures are visualized with program ALINE (Bond and Schüttelkopf, 2009) and the structures with PYMOL (Schrödinger, 2010).