A cautionary note on interpretation of hierarchical classifications of protein folds

Alex CW May

Address: Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK.
E-mail: amay@nimr.mrc.ac.uk

Structure September 1999, 7:R213
http://biomednet.com/electcrel/09692126007R0213
0969-2126/99/$ – see front matter © 1999 Elsevier Science Ltd. All rights reserved.

Recently, Castillo et al. [1] presented an instructive analysis of what they describe as the double-ψβ barrel protein fold. This fold comprises six strands with a characteristic eponymous topology arranged in a closed barrel with shear number ten. Hierarchical classifications are obtained by using the data from pairwise rigid-body superposition of the double-ψβ barrels. The first hierarchical classification is performed at the level of the complete double-ψβ barrel domains, whereas the second is derived from each of the two symmetry-related halves of the fold. A tree diagram describing the latter hierarchical classification is shown in their Figure 4.

Both hierarchical classifications show that the double-ψβ barrels of pyruvoyl-dependent aspartate decarboxylase (Protein Data Bank [PDB] code 1aw8), dimethylsulphoxide reductase (1cxs) and formate dehydrogenase H (1fdo) form a single grouping. In addition, the hierarchical classification of the symmetry-related halves of each domain indicates that each of the first halves of 1aw8, 1cxs and 1fdo and each of their second halves constitute two subgroups: (1aw8-1 1cxs-1 1fdo-1) and (1aw8-2 1cxs-2 1fdo-2). In turn, the cluster of the first halves of barwin and endoglucanase V (1bw4-1 2eng-1) is merged with the partition (1aw8-1 1cxs-1 1fdo-1 1aw8-2 1cxs-2 1fdo-2).

On the basis of these hierarchical classifications, it is suggested that the five proteins 1aw8, 1cxs, 1fdo, 1bw4 and 2eng share a common homodimeric ancestor.

Lately, I have investigated the usefulness of protein structure similarity measures for the hierarchical classification of protein folds [2]. The question was: on hierarchical clustering, which metric provides the highest number of meaningful trees for a set of 24 protein families? A meaningful tree is defined as one where all the clusters are found to be reliable according to a jackknife test [3]. Castillo et al. use a protein structure similarity measure not considered in my work: that of Alexandrov and Go [4]. It seemed helpful, therefore, to assess in this way the hierarchical classifications presented in [1] as these are key to the paper’s conclusions.

I used the unweighted pair-group method using arithmetic averages (UPGMA), the most widely used algorithm for obtaining a hierarchical classification [5]. All of the nested set of six partitions within the hierarchical classification of the seven complete double-ψβ-barrel domains were found to be reliable [3]. However, testing of the tree obtained from the symmetry-related halves demonstrated that three out of the nine internal nodes are unreliable. In fact, the first partition of the ten fragments is not stable: although the cluster (1bw4-2 2eng-2) is reliable, that encompassing the remaining eight halves is not. The grouping of the first and second halves of 1aw8, 1cxs and 1fdo is actually not reliable. Lastly, the merging of 1aw8-2 to the reliable cluster (1cxs-2 1fdo-2) is not stable.

This highlights the need for care in interpretation of hierarchical classifications of protein folds. In particular, if such a tree is to be proposed as support for an evolutionary scenario as in [1] it seems sensible to ensure that the tree is meaningful. After all, published trees derived from sequence information are almost always accompanied by statistics describing the support for a tree representation and individual clusters contained within.

References