(Probably) All Possible Protein Folds at Low Resolution

Gordon M. Crippen and Vladimir N. Maiorov

College of Pharmacy, University of
Michigan, Ann Arbor, MI 48109, U.S.A.

Abstract

For decades, a large number of investigators have been sifting the database of experimentally determined three-dimensional protein structures to discover recurring patterns of all types. Now that there are over a thousand such structures available, the natural question is whether we have seen all substantially different protein folds, and if not, how many have yet to be discovered? Answering the question can be broken down into three steps: (1) choose the range and domain for a similarity function, then (2) choose a particular similarity function, and (3) construct a corresponding protein model space that can be searched for dissimilar structures. In our analysis of the problem, we first chose to examine different conformations of the same protein, taking into account only C^{α} atomic coordinates. In particular, we do not compare proteins of different chain lengths on the basis of some kind of gapped alignment. Secondly, we use a measure of conformational similarity based on rigid body superposition that emphasizes overall geometric resemblance, rather than agreement in secondary structure, for example. Third, we employed the discrete cosine transform to construct exhaustive sets of globular self-avoiding C^{α} traces that were all different from each other by a given level. These sets of artificial structures were not too large to explicitly enumerate as long as the level of dissimilarity was high, and the chain flexibility was low. For chains flexible enough to match all experimental structures of 170 residue or less that are not β-barrels, we find 128 artificial structures, of which 28 resemble nothing in the Protein Data Bank.