Estimating the probability of a protein to have a new fold based on a map of all protein sequences

Elon Portugaly and Michal Linial Dept. of Biological Chemistry, Institute of Life Sciences, The Hebrew University Jerusalem, 91904, Israel. elonp@cs.huji.ac.il;michall@leonardo.ls.huji.ac.il

In the foreseeable future it will be impossible to experimentally solve the structure of hundreds of thousands of proteins. Therefore, it is necessary to find ways to predict a protein's structure based on data derived from structurally solved proteins. Current attempts to computationally determine a protein structure still have a limited success, partly due to the shortage of solved structures that can be used as models. The goal of this study is to predict which proteins have new, currently unknown structural folds. To this end, a comparison between ProtoMap (release 2.0) and SCOP (release 1.37) had been performed. ProtoMap is an automatically-generated hierarchical and relational classification of all protein sequences in Swissprot database. SCOP is a hierarchical classification of all known protein structural domains. Our goal thus is to estimate the probability that a protein sequence has a new, unknown fold. We compared ProtoMap's most relaxed level of classification, and SCOPs classification of folds. We first showed that ProtoMap is selective in terms of SCOP folds. Thus, we can unequivocally speak of the "representative fold of a cluster" and hence about the chance that a cluster's fold is a new, unknown fold. We say that a cluster is *vacant* when it contains no known structures, and is *occupied* otherwise. Of the 13,354 clusters in ProtoMap, only 756 are occupied. A vacant cluster is said to be new when its (presently undetermined) corresponding fold is absent from SCOP, and old otherwise. Using this terminology, our target is to estimate, for each cluster, its probability to be new.

Each cluster of sequences in ProtoMap has a weighted list of related clusters. The weights reflect relatedness among clusters. A threshold on this weight defines a graph describing the relationship between all the clusters of ProtoMap. We studied the distribution of distances among occupied clusters in this graph. Based on this, we derived a statistical estimate for two distributions. One consists of distances from new clusters to occupied clusters. The other of distances from old clusters to occupied clusters. Finally, we employed Bayes' rule to calculate, on the basis of these two distributions, the probability of a cluster being new, given the distance from this cluster to occupied clusters.

Our calculations yield an estimated probability function that partitions all clusters to six classes according to their probability of being new (one class consisting of the occupied clusters). Of the 13,354 clusters, only 5.7% where assigned to the two highest probability classes.

Our results were tested using clusters of specific biological characteristics. We tested membranous proteins (1,048 clusters) and transcription related clusters (479 clusters). A preferred probability (6.45 fold) for proteins to have a new fold was calculated for membranous proteins and the opposite preference was calculated for the transcription-related clusters. We evaluated our results against a more recent release - SCOP 1.39. We show a linear relation between our predicted probability of being new and the proportion of new folds among the recently released structures. The clusters in the most probable classes form a selective list of proteins (473 clusters without the membranous ones) which are a suggested target list for the purpose of discovering new protein folds.

Preliminary results from this study are presented at the National Instate of General Medical Studies, USA (http://www.structuralgenomics.org). Our complete list of target proteins can be accessed at (http://www.ls.huji.ac.il/~michall/Structural-Genomics/Target).