Assessment of structure-based functional annotation methods on protein 3D models

I. Mangone, M. Helmer-Citterich✉, G. Ausiello
Department of Biology, University of Tor Vergata, Roma, Italy

Motivations

The three-dimensional structure is more informative of the sole aminoacidic sequence to assign a molecular function to a new protein. For this reason many automated methods have been developed to infer the function of a protein structure using comparison approaches or analyzing its physicochemical characteristics. Unluckily while the genome sequencing projects of organisms have considerably increased the number of available protein sequences, protein structure determination with X-ray crystallography and NMR is still a complex and rather costly procedure. There are indeed more than 20 thousand entries in the database of protein sequences (UniProt) and only 79.600 entries in the database of protein structures (PDB). The big gap separating the number of known sequences from the number of solved structures is increasing every year and is strengthening the need for structure-based functional annotation methods capable to work on homology models instead of crystal structures. The applicability of the existing functional prediction methods to protein models has never been explored so far, even if most of the structural information now available is stored in 3D models. The aim of this work is to study the reliability of different structure-based functional annotation methods when used on protein models and to analyze how the prediction methods performance is correlated with the overall quality of the available homology model.

Methods

We used an automated procedure to compare the performances of many structure-based functional prediction methods when they work on a set of homology models of different quality or on a crystallographic solved structure. Each different method is tested on the same dataset proposed by the authors in the original publication and on a set of homology models built for each structure in the dataset. All models were generated using MODELLER (v9.9) and evaluated using the GDT _ TS score. To obtain models of different quality only templates are used having a sequence similarity with the solved structures under a set of fixed thresholds.

Results

We have evaluated five methods: PDBinder, Concavity and Fpocket for the prediction of protein binding pockets and Pfinder and FINDSITE _ metal for the prediction of phosphate and metals binding sites. The performances have been measured using the F-score or the MCC where applicable. Preliminary results show that when using models with a GDT higher than 99% on average the performances drop by about the 22%. When models quality decreases we have a significant decrease of prediction method performances up to 50% (with a GDT of 50%), with some methods that have shown a greater resilience to the decrease of the model quality. These are only to be considered as preliminary results since a number of other methods are being added to the analysis.