Improving the quality of the predictions of protein stability changes upon mutation using a multi-class predictor

Capriotti E (1), Fariselli P (1), Rossi I (1,2), Casadio R (1)

(1) Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna,
via Irnerio 42, 40126 Bologna, Italy
(2) BioDec Srl, Via Calzavecchio 20/2, 40033 Casalecchio di Reno (BO), Italy

Motivation

The accurate prediction of protein stability free energy change (DDG) upon single point mutation is a key problem of Structural Bioinformatics. In the last years several methods were described to predict DDG upon single point mutations in proteins. One common approach is based on the development of different energy functions, starting routinely from the protein known structure and then applied to the mutated protein to compute DDG. Recently the increasing number of experimental thermodynamic data and their availability in the ProTherm database prompted us to develop machine learning-based approaches for predicting both the sign and the value of DDG upon protein mutation starting both from the sequence and/or structure (I-Mutant2.0, http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi).
These automatic methods however suffer from the fact that experimental data are affected by standard deviations associated with the DDG values, when evaluated, and from the fact the most of the experimental data (about 32% of the data set) are close to 0 (-0.5 =< DDG =< 0.5 Kcal/mol). In these cases, considering the associate error, both the value and the sign of DDG may be either positive or negative for the same mutation, leading to ambiguity when evaluating the extent of protein folding stability. In order to overcame this problem we implemented a new predictor able to discriminate between 3 possible classes, dividing the set of experimental data in: destabilizing mutations, stabilizing mutations and neutral mutations. Furthermore we also enriched the training set of experimental data by assuming that for each mutation in the data base also the opposite restoring mutation is present.

Methods

The databases used in this work are derived from the release (September 2005) of the Thermodynamic Database for Proteins and Mutants ProTherm. We select our initial set imposing the following constrains: a) the DDG value was extrapolated from experimental data and reported in the data base; b) the data are relative to single mutations; c) the data are obtained from reversible experiments After this procedure we obtain a larger data set comprising 1681 different single point mutations and related experimental data for 58 different proteins. From the latter by selecting only 55 protein known with atomic resolution we have a subset of 1634 mutations. Adopting a criterion of thermodynamic reversibility for each mutation, we double all the thermodynamic data. Finally, we end up with 3362 mutations for the set containing protein sequences (DBSEQ) and 3268 mutations for the subset of proteins known with atomic resolution (DB3D). According to experimental DDG value each mutation is grouped into one of the following three classes: i) destabilizing mutation, when DDG<-0.5 Kcal/mol; ii) stabilizing mutation when DDG>0.5 Kcal/mol; iii) neutral mutations when -0.5 =< DDG =< 0.5 Kcal/mol. The choice of |0.5| Kcal/mol as a threshold value for DDG classification provides a balanced datasets and is also a limiting value of standard errors reported in experimental works. We developed support vector machines (SVM) and trained them to predict if a given single point protein mutation is classified in one of the three classes defined above. This task is addressed starting from the protein tertiary structure or sequence, adopting a Radial Basis Functions kernel. The input vector consists of 42 values. The first 2 input values account respectively for the temperature and the pH at which the stability of the mutated protein was determined. The next 20 values (for 20 residue types) explicitly define the mutation (we set to -1 the element corresponding to the deleted residue and to 1 the new residue (all the remaining elements are kept equal to 0). Finally, the last 20 input values encode the residue environment: namely a spatial environment, when the protein structure is available, or the nearest sequence neighbours, when only the protein sequence is provided. When prediction is structurebased, the Relative Solvent Accessible Area (RSA) is also coded as input value.

Results

Our methods are trained/tested using a 20-fold cross-validation procedure on the two available sets. After an optimization procedure on different spatial and sequence environments, best predictors score as follows: when based on structural information, an overall accuracy of 58% is achieved with a mean value of correlation to the thermodynamic data of 0.37. In turn, when the prediction is performed considering only sequence information the accuracy is 52% and the mean value of correlation becomes 0.28.

Availability: http://www.biocomp.unibo.it/

Contact email: emidio@biocomp.unibo.it