The effect of mutations on protein stability changes: a three class pair residue-discrimination study

ID - 115

Capriotti Emidio[1], Fariselli Piero[2], Rossi Ivan[2,3], Casadio Rita[2]

[1]Structural Genomic Unit, Department of Bioinformatics, Centro de Investigacion Principe Felipe (CIPF) Valencia, Spain

[2]Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, via Irnerio 42, 40126 Bologna, Italy

[3]BioDec Srl, via Fanin 48, 40127 Bologna, Italy

Motivation

A basic question in protein structural studies is to which extent mutations affect protein folding stability. This can be asked starting from protein sequence and/or from structure. In proteomics and genomics studies prediction of protein stability free energy change (DDG) upon single point mutation may also help the annotation process. So far methods to address this problem are based on to two different approaches: development of optimised and different energy functions (Gilis and Rooman, 1997; Guerois et al., 2002; Zhou and Zhou, 2002) for proteins whose structure is known and implementations of machine learning approaches when sequences and/or structures are available (Capriotti et al., 2004; Capriotti et al 2005a; Capriotti et al 2005b; Cheng et al., 2006). On the other hand experimental DDG values are affected by uncertainty as measured by standard deviations. Most of the DDG values are nearly zero (about 32% of the DDG data set ranges from -0.5 to 0.5 Kcal/mol); furthermore both the value and sign of DDG may be either positive or negative for the same mutation blurring the relationship among mutations and expected DDG value. In order to overcome this problem we describe a new predictor that discriminates between 3 mutation classes: destabilizing mutations, stabilizing mutations and neutral mutations, being neutral mutations all those substitutions whose effect is to promote a protein stability free energy change (DDG) ranging from -0.5 to 0.5 Kcal/mol.

Methods

In this paper a support vector machine (SVM) starting from the protein sequence or structure discriminates between stabilizing, destabilizing and neutral mutations. The machine learning method here presented was trained and tested considering experimental data selected from the new release of the ProTherm Database (Kumar et al. 2006). We collect more than 1600 mutations and according to the criterion of thermodynamic reversibility for each mutation, we double all the thermodynamic data.

Finally, we end up with more than 3200 mutations to train our method with a cross-validation procedure. Following other previous works of ours (Capriotti et al 2005a; Capriotti et al 2005b) two different SVM methods were developed depending on the provided information. If the predictions are sequence-based an input vector of 42 elements is used. The input accounts for the residue mutation (encoded in the first 20 elements), the sequence environment (encoded in second 20 elements) and experimental conditions (temperature and pH reported in the last two element). When the 3D structure of the mutated proteins is known is it also possible to perform the prediction with the structure-based method that considers an input vector of 43 elements. The input vector is similar to that of the sequence-based predictor: however the 20 element vector encoding for the sequence environment is replaced considering the structural environment and adding also one element accounting for the relative solvent accessible area is added.

Results

We rank all the possible substitutions according to a three class-classification system that aside prediction indicates also the rate of occurrence of the mutations in the data base and the list of proteins where the mutation effect has been experimentally detected. We show that the overall accuracy of our predictor is as high as 52% when performed starting from sequence information and 58% when the protein structure is available with a mean value correlation coefficient of 0.30 and 0.39 respectively. These values are about 20 points per cent higher than those of a random predictor; when selecting only the mutations with high effect on the protein stability (|DDG|>0.5 Kcal) the prediction of the destabilizing and stabilizing mutations are well balanced and reaches the accuracy values of 71% and 76% with correlation coefficient of 0.43 and 0.52, respectively when sequence-based and structure-based predictions are provided.

Email: ecapriotti@cipf.es