Globularity criteria to evaluate the structural quality of modeled proteins

ID - 104

Costantini Susan[1,2], Facchiano Angelo M[1,2], Colonna Giovanni[1]

[1]Centro di Ricerca Interdipartimentale di Scienze Computazionali e Biotecnologiche, Seconda Università di Napoli, Italy

[2]Laboratorio di Bioinformatica e Biologia Computazionale, Istituto di Scienze dell'Alimentazione, CNR, Avellino, Italy

Motivation

The basic problem of any computational approach to structure prediction is to evaluate the quality of models. CASP results evidenced that most of the submitted models are far from the real protein structure. Our work has been aimed to analyze structural properties that characterize protein globularity, to suggest an operative procedure for the analysis of globular quality of theoretical protein models obtained by computational approaches in the absence of experimental target structures, and, finally, to prevent the diffusion of theoretical models not suitable with the real features of the protein globularity.

Methods

The analyses of experimental structures were performed using the PDBselect set of experimentally determined, non-redundant protein structures with mutual sequence similarity <25%. The secondary structure defined by the DSSP algorithm was used to assign each protein to the right structural class, according to the four main SCOP classes ('all-alpha', 'all-beta', 'alpha/beta', and 'alpha+beta'). Some features related to the globularity of proteins have been evaluated for each structure. Four properties, i.e. the total accessible surface area (ASA), the number of MM-type H-bonds, voids and water molecules in a layer of 5 Angstroms, correlated better then others with the molecular weight. All four properties versus the molecular weights of selected proteins in each structural class fit with linear regressions, and the linear function was defined to estimate the expected value for each feature. Therefore, a globularity score value was calculated for each protein by summing the ratios of the differences between the calculated and predicted values for each of the four properties versus the related errors (i.e. Root Mean Squared Error). On the basis of the most frequent score values, calculated for the proteins belonging to the same class, it was possible to identify a threshold score value specific for each of the four structural classes. These threshold score values were then used as cut-off to evaluate the structural properties of models predicted for the CASP6 protein structure prediction experiment in the New Fold (NF) and difficult Fold Recognition Analogous (FR/A) categories, and used as 'testing dataset'.

Results

We analyzed different structural properties of globular proteins for experimentally solved proteins belonging to the four different structural classes. The properties were found to be linearly correlated to protein molecular weight, but with some differences among the four classes. These results were applied to develop an evaluation test of theoretical models based on the expected globular properties of proteins. In fact, a score value (i.e. globularity score) for all proteins was calculated by using the parameters having the highest correlation coefficients with the protein molecular weights (i.e. MM-type H-bonds, void number, total accessible surface area and water molecules). To verify the success of our test, we applied our globularity score to several protein models submitted to the sixth edition of CASP. Our results surprisingly show that many of the models submitted (54.6%) should be discarded a priori because they do not have the structural properties expected in globular proteins.

Therefore, our study supports the need for careful checks to avoid the diffusion of incorrect structural models and allows the evaluation of models in the absence of experimental reference structures, thus preventing the diffusion of incorrect structural models and the formulation of incorrect functional hypotheses. It can be used to check the globularity of predicted models, and to supplement other methods already used to evaluate their quality.

Availability: http://bioinformatica.isa.cnr.it/GLOBULARITY/

Email: susan.costantini@isa.cnr.it