Supporting protein tertiary structure modelling with contact sites prediction method

1. Introduction

Knowledge about three dimensional structure of proteins is a prerequisite for studies on their behaviour or, for example, their role as a target in drug design. Only 0.15% of known amino acid sequences, though, have also a known structure. This is mainly due to the technical limitations of experimental structure determination methods, which include X-ray crystallography, NMR, or electron microscopy. The application of computer methods can provide a solution and support the process of protein structure determination.

One of the most popular methods used in computational protein structure prediction are based on the prediction of residue-residue contacts in proteins as an initial step for structure reconstruction. A residue-residue contact is usually defined as two amino acids that have the distance between their C_β atoms (C_α for glycine) not higher than 8 Å and that are separated in sequence by more than 4 residues.

There are many contact prediction methods. The best of them are based on the direct coupling analysis (DCA) of correlated mutations that occur in the multiple sequence alignment (MSA) of an analysed protein sequence. The main assumption of methods using correlated mutations is that if two sequence positions are significant for a protein's function or structure, and one of the residues on these positions mutates, the other one should also mutate to maintain proper interactions between them. Since contacts hold important information about the protein structure, their occurrence can be indicated using methods based on correlated mutations.

Unfortunately, even the best DCA based methods achieve an accuracy of only 40% for the 100 strongest predicted contacts. That is not enough for routine, reliable *ab initio* protein structure prediction. Also, the reported accuracy is an average over a large test set. This value, thus, can be much higher but also much lower for one particular protein of interest. This can be a real issue, especially for scientists who use predicted contacts to steer their research on that protein. Therefore, the main goal of the PhD studies was to make predicted residue-residue contacts more reliable and useful.

2. Results

The first step in the studies involved examining the characteristics of residue-residue contacts in proteins in relation to a protein's structural class and the contact definition. Parameters were proposed that describe the contact propensity of residue types and the differences between these propensities relate to the protein's structural groups - class and its topology. One of the main results of these analyses was the discovery of residue pairs that differ the strongest in contact propensity between the Alpha and Beta classes. Also, contact characteristics within topologies and such characteristics of their classes were compared. The topologies that differ the most from their classes were indicated and discussed.

The second part of the studies dealt with the application of the residue-residue contact characteristics to improve the contact prediction accuracy of one of the best methods of those days - mfDCA. An algorithm was proposed that aimed to match contact characteristics among the 200 strongest predicted contacts for one single protein with those calculated for the entire structure class of that protein. The results showed that application of the algorithm on contacts predicted with mfDCA can increase the average prediction accuracy by as much as 5% for the 10 strongest predicted contacts. The largest improvement was obtained for proteins in the

Alpha class. These results were published in (**Wozniak and Kotulska, 2014**) and proved the **first thesis** proposing that "*the use of residue-residue contact characteristics in contacts prediction can increase contact prediction accuracy*".

In the next step, it was shown that DCA can support protein structure determination methods by selecting between properly folded and misfolded forms of a protein. Contacts were predicted with one of the best recent DCA-based methods - gplmDCA. The decisive step of the method is a comparison between the number of properly predicted contacts for the native structure and a wrong model. In general, that structure was called a proper one which had more contacts predicted properly. Results showed that the method greatly favours the native structure over the wrong or less accurate one. It shows a perfect 100% accuracy when the selection has to be done between properly folded and intentionally misfolded protein structures. The method was further evaluated using a group of obsolete and successor PDB entry pairs. The PDB obsolete-successor pairs for which the method failed were investigated in detail showing that it is very likely that the successor structure was still not yet properly folded, so the designed method most likely is not wrong. The method published in **(Wozniak et al., 2017A)** can be especially useful for experimentalists who need to choose between two folds of the same protein. The results proved the **second thesis** proposing that "*predicted residue-residue contacts enable distinguishing between properly folded and misfolded protein structures*".

In the last step of the PhD studies regression models were designed that forecast the accuracy of residue-residue contact prediction for an individual protein. The method was tested for two well-working DCA-based methods: gplmDCA, and PSICOV. The forecasting method uses various regression models and their combinations. The regression models were built on parameters describing the MSA, the predicted secondary structure, the predicted solvent accessibility, and the contact prediction scores for the target protein. The best performance was achieved for the regression model that was the average of Lasso, SVM, and Random Forest Algorithm. For gplmDCA, this model forecasted the percentage of correctly predicted contacts among the 200 strongest predictions with a root mean square error of only about 7 percentage points. It was also shown that the method holds great potential for meta-methods such as RaptorX. Furthermore, the study included the analysis of individual parameters related to the protein with regard to their importance in the performance of the method. The designed regression models can significantly enhance the usefulness of predicted residue-residue contacts in many fields of life sciences. The results published in **(Wozniak et al., 2017B)** prove the **third thesis** proposing that "*prediction accuracy of residue-residue contacts can be forecasted for each protein individually*".

3. Conclusions

The proposed theses were proved. First, a procedure that supports residue-residue contact prediction methods and increase their prediction accuracy by using contact statistics has been designed. Second, the ability of residue-residue contact prediction methods to differentiate between properly folded structures and their decoys has been examined. Third, a method was designed that forecasts an accuracy of residue-residue contacts prediction DCA-based methods for individual protein.

Results presented in the PhD dissertation increase the reliability and usefulness of contacts predicted with the best methods available. In order to apply contacts in automatic way, though, the studies must be continued.