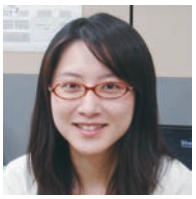
Development of High-Precision Disorder Region Prediction Method

(POODLE: Prediction Of Order and Disorder by machine LEarning)



Tamotsu Noguchi / Kana Shimizu
Leader / Technical Staff
Protein Function Team

Recently it has been experimentally proven that some sequences that do not form certain three-dimensional structures (hence called disorder regions or natively unfolded regions) play important roles especially in higher organisms. Many of them have been discovered in proteins associated with transcriptional regulation and in DNA-binding proteins. In addition, since the disorder regions are known to obstruct crystallization and spectral assignment in the analysis of three-dimensional protein structure by x-ray crystallography and NMR, removing them beforehand will make analysis much easier. Also, time and labor spent in the prediction of protein three-dimensional structure can be greatly reduced by understanding the sequence to be predicted. Thus, in recent years, the disorder regions are watched with keen interest, and the methods for predicting disorder regions is actively developed.

Noticing that the characteristics of disorder regions differ according to their length, we have developed 3 types of prediction method (POODLE-W, L, and S) designed to predict the disorder region(s) of a whole protein, and both long and short disorder regions. Here we would like to introduce POODLE-L (the long disorder region predictor) that was developed first and has indicated the highest prediction performance when compared with five publicly available and well-established disordered regions predictors.

Methods for predicting disorder regions perform processing (learning) via statistical procedures and machine learning of the relationship between disorder regions and the amino acid sequence and/or the PSSM (position-specific scoring matrix). POODLE-L predicts disorder region by using the score matrix based on six physicochemical descriptors from amino acid residue-specific characteristics (10 parameters) and then by learning differentiation of a disorder or order region through a Support Vector Machine (SVM). The special features of this method lie in the fact that the window size is a wide 40 residues and the window moves along the sequence during prediction of the whole window in the first step SVMs, whereas in a general prediction method the entire sequence is divided onto windows and only the center residue(s) are predicted. In the next, the results predicted for each window are tabulated for each residue position on the sequence, and the second step SVM computes the probability of a single residue to be disordered based on their distribution. Prediction tests were performed for all combinations to determine which combinations are optimal for the six descriptors. As a result, instead of the model using all six descriptors, nine models using selected combinations yielded better prediction results. By integrating those results, we successfully increased the accuracy of prediction even more. For long disorder regions, we have surpassed the accuracy of previously available and well-established disordered prediction methods.

POODLE-L is available to the public as a service that can be used by general researchers at http://mbs.cbrc.jp/poodle (Figure 2). Public availability of POODLE-W and S is currently under review.

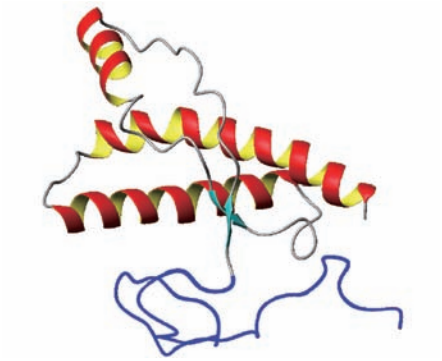


Fig.1 The N-terminal structure (blue) of the normal prion protein PrPC is a disorder region, displayed through modeling

Fig.2 POODLE (http://mbs.cbrc.jp/poodle)
Results of disorder prediction of normal prion protein PrPC. A large peak can be seen at the N-terminus, indicating that a disorder region is predicted there.

Editorial Comment

POODLE-S and -W are available at the POODLE server (http://mbs.cbrc.jp/poodle/). The team from CBRC led by Tamotsu Noguchi has been awarded the 2nd Prize at the "Order-Disorder Regions prediction category" in CASP7 by using the three types of POODLE.