Analytical Biochemistry 491 (2015) 18–22

Contents lists available at ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Prediction of protein disorder on amino acid substitutions

P. Anoosha, R. Sakthivel, M. Michael Gromiha*

Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, Tamilnadu, India

ARTICLE INFO

Article history:
Received 11 May 2015
Received in revised form
27 July 2015
Accepted 27 August 2015
Available online 6 September 2015

Keywords:
Disorder
Mutation
Stability
Machine learning
Neighboring residue

ABSTRACT

Intrinsically disordered regions of proteins are known to have many functional roles in cell signaling and regulatory pathways. The altered expression of these proteins due to mutations is associated with various diseases. Currently, most of the available methods focus on predicting the disordered proteins or the disordered regions in a protein. On the other hand, methods developed for predicting protein disorder on mutation showed a poor performance with a maximum accuracy of 70%. Hence, in this work, we have developed a novel method to classify the disorder-related amino acid substitutions using amino acid properties, substitution matrices, and the effect of neighboring residues that showed an accuracy of 90.0% with a sensitivity and specificity of 94.9 and 80.6%, respectively, in 10-fold cross-validation. The method was evaluated with a test set of 20% data using 10 iterations, which showed an average accuracy of 88.9%. Furthermore, we systematically analyzed the features responsible for the better performance of our method and observed that neighboring residues play an important role in defining the disorder of a given residue in a protein sequence. We have developed a prediction server to identify disorder-related mutations, and it is available at http://www.iitm.ac.in/bioinfo/DIM_Pred/.

Disordered proteins lack intrinsic three-dimensional structures under physiological conditions, and most of them are known to play important roles in regulatory functions of a cell [1]. Locally disordered regions are common in proteins and are observed in many three-dimensional structures. Few of them are involved in proteins' biological function [2]. Thousands of structures are deposited in the PDB, which are known to have disordered chains that become ordered in the presence of their respective interacting partners [3]. The details of these disordered proteins are being deposited in the DisProt database [4], which contains more than 1500 intrinsically disordered regions from nearly 700 proteins.

Increased availability of naturally disordered protein data triggered the development of computational methods for prediction analysis. Several bioinformatic algorithms were developed for predicting disordered regions from the sequence based on different principles such as physicochemical properties of amino acids and their conservation during evolution [5–11]. Intrinsically disordered proteins are implicated in numerous human diseases, including cardiovascular and neurodegenerative disorders, and diabetes [12]. In addition, missense mutations in disordered regions are involved in cancer development [13,14]. This suggests that the study of disorder-related mutations in proteins is important for understanding their functional outcome in the human body. Although significant progress has been made over decades to predict the functional effects of point mutations in proteins [15–18], there is only one prediction program available for disorder-related variants [19], and the prediction performance is rather poor (accuracy of 70.5% and 50% for 5-fold cross-validation and test set, respectively). Hence, it is necessary to understand the factors influencing the disorderness of a protein on mutation and to improve the prediction performance with high sensitivity and specificity.

In this study, we mainly focused on missense mutations that influence the disorder of proteins. We used a dataset of 90 mutants [19] and analyzed the effect of mutation on protein disorder. We observed that most of the mutants causing protein disorder destabilize the protein and are located in the buried interior of the protein. Furthermore, we derived a set of 208 features from amino acid sequences of the considered proteins and obtained a set of the 9 best discriminative features using feature selection methods. Interestingly, the features are the combinations of amino acid properties, mutation matrices, and the effect of neighboring residues. Using these 9 features, a support vector machine (SVM)-based model was developed for predicting the class of mutants (order to order or order to disorder). The model achieved an accuracy of 90.0% in 10-fold cross-validation on a dataset of 90

Abbreviations used: SVM, support vector machine; ASA, accessible surface area.

* Corresponding author.
E-mail address: gromiha@iitm.ac.in (M.M. Gromiha).

http://dx.doi.org/10.1016/j.ab.2015.08.028
0003-2697/

mutants. In addition, we used a test set of 18 mutants to assess the performance of our method and observed an average accuracy of 89.9% using 10 iterations, which outperformed other available methods. Hence, we suggest that the proposed method could be used as an efficient tool to predict disorder-related mutations.

Materials and methods

Dataset

We obtained a dataset of 101 mutants from the literature [19]. These mutations belong to four classes: O → D, O → O, D → O, and D → D (where O is order and D is disorder). The numbers of mutations in two classes, D → D and D → O, are not sufficient to carry out the analysis and are not considered in the current work. Hence, our dataset contained 90 mutants with 59 and 31 mutants in the O → D and O → O classes, respectively. In the current method, we used the dataset in two aspects: (i) used the entire dataset of 90 mutants for developing the model and validating it using 10-fold cross-validation and jackknife tests and (ii) randomly selected 80% of the data for training and the rest of the data (20%, 18 mutants) for testing the method. The data were shuffled 10 times and repeated the prediction to evaluate the performance of the method.

Calculation of features

Physicochemical properties

We collected a set of 49 physical, chemical, energetic, and conformational properties of amino acid residues from the literature [20]. The change in property was calculated as

$$\Delta P = P_{\text{Snp}} - P_{\text{wild-type}}, \quad (1)$$

where $P_{\text{wild-type}}$ and P_{Snp} are the property values of a residue at the mutation site in wild type and mutant, respectively, and ΔP is the change in property due to mutation.

AAindex database matrices

We collected 75 amino acid mutation matrices from the AAindex2 database [21], and the numerical value from a given matrix corresponding to each mutation is considered as a feature. Furthermore, 43 pairwise contact potential matrices were collected from the AAindex3 database [21], and the difference of amino acid contact potential for a mutation is obtained by subtracting the contact potential value of N/C neighbor of mutation position to wild-type residue from N/C neighbor to mutant residue. Along with the neighboring residue information, we also considered the wild-type and mutant residue information for computing the contact potential. Hence, for the variation at the same position with same neighboring residues, we obtained different contact potentials because the mutant residue is different. For example, variants M → L and M → K at position 1663 in BRCA1 protein have the same neighboring residues in the N and C terminals (F and L, respectively). However, the contact potentials with N − 1 neighbor, F for these mutants (FM-FL and FM-FK), are different (−0.08 and 0.38 obtained with AAindex3 code MOOG990101) because the mutant residue is different in both cases.

Neighboring residue information

We obtained the information about the neighboring residues of a mutant position on both N and C directions along the sequence at different window lengths from 3 to 13 residues. The information about one residue on both sides of the mutant gives the data for the window length of three residues. For each window length, the numbers of residues on different amino acid types—polar, nonpolar, positively charged, negatively charged, aromatic, aliphatic, and sulfur-containing residues—were considered as separate features. Our analysis shows that the information about only one residue along the N and C terminals of the mutant in terms of contact potentials is sufficient for discrimination.

Feature selection and classification method

We used WEKA data mining software [22] for feature selection and classification. It includes several algorithms based on Bayes function, neural network, logistic function, support vector machine, regression analysis, nearest neighbor, meta learning, decision tree, and rules. We evaluated and analyzed various machine learning algorithms available in WEKA for classifying mutations based on the given features. Based on the performance of all the methods in our dataset, we chose the SMO (sequential minimal optimization) classifier, which is an SVM-based algorithm for the classification of mutations. Support vector machines can efficiently handle imbalanced data. There are several parameters in the algorithm that can influence the data imbalance. We optimized the parameters such as the complexity parameter (C), which is very important in handling such data [23].

Various feature selection methods are also available in WEKA. We used various combinations of attribute evaluator and search methods in WEKA to select the best feature set, and for our dataset feature selection was done using the SVM attribute evaluator and Ranker search methods. This combination of evaluator and search method has already proved to be efficient in the feature selection procedure [24]. SVM attribute evaluator assesses the worth of an attribute by using SVM classifier, and the Ranker search method ranks the attributes by their individual evaluations. We derived a set of 9 features that could best discriminate the mutants.

Performance evaluation

The model is evaluated using n-fold cross-validation in which $n-1$ data will be used to train the classifier and the rest of the data is used for testing. The classification performance of the model was assessed by the following measures:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (2)$$

$$\text{Sensitivity or True Positive Rate} = \text{TP}/(\text{TP} + \text{FN}) \quad (3)$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}) \quad (4)$$

$$\text{False Positive Rate} = \text{FP}/\text{FP} + \text{TN}, \quad (5)$$

where TP, TN, FP, and FN refer to the numbers of true positives, true negatives, false positives, and false negatives, respectively.

Results and discussion

Distribution of amino acid substitutions and influence of secondary structure and solvent accessibility

The distribution of amino acid substitutions of 90 mutants is shown in Fig. 1. We observed that the amino acid substitutions of the mutants belonging to two different classes are very specific, that is, A → L, A → M, and so forth in O → O and M → K, N → K, and so forth in O → D. There are 63 amino acid substitution types present in the entire dataset of 90 mutants in which only 7 substitutions are present in both classes of mutants.

We further studied the distribution of these mutants with respect to accessible surface area (ASA) using the prediction program SABLE [25], and the results are presented in Table 1. Interestingly, mutations causing disorder to the proteins are mostly