Expert Systems with Applications 41 (2014) 1672–1679

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Predicting the protein solubility by integrating chaos games representation and entropy in information theory

Niu Xiaohui, Shi Feng, Hu Xuehai, Xia Jingbo, Li Nana *

College of Science, Huazhong Agricultural University, Wuhan, PR China

ARTICLE INFO

Keywords:
Protein solubility
Pseudo amino acid composition
Entropy in information theory
Chaos game representation
Support Vector Machine

ABSTRACT

Protein solubility is a prerequisite for many structural, functional studies. Predicting the propensity of a protein to be soluble or to form inclusion body is a challenging and crucial problem. In order to formulate the protein samples which can reflect the intrinsic correlation with protein solubility, triangle, quadrangle and 12-vertex polygon CGR, the concept of entropy in information theory, together with amino acid and dipeptide compositions are applied based on a different mode of pseudo amino acid composition (PseAAC). The mathematical expressions involving with seven CGR methods and amino acid, dipeptide compositions with their corresponding entropies are evaluated with 10-fold cross validation and re-substitution test. The numerical results confirm that the introduction of the entropy can significantly improve the performance of the classifiers. Triangle CGR method surpass the two other CGR methods in classifier construction. It can provide complementary sequence-order information on the basis of dipeptide composition. The optimal mathematical expression is dipeptide composition, triangle CGR and their entropies. With the 2-level triangle polygon CGR + dipeptide composition together with their corresponding entropies as the mathematical feature, the classifier achieved the best accuracy 88.45% and MCC achieved 0.7588 in 10-fold cross validation test. In the re-substitution test, the 3-level triangle polygon CGR, dipeptide composition and their entropies perform best, its accuracy was 92.38%, MCC achieved 0.8387.

1. Introduction

Protein solubility is an important property for the proteomic researches. With the wide usage of *Escherichia coli* as the host, most of proteins form inclusion bodies on overexpressions (Idicula-Thomas & Balaji, 2005; Ventura, 2005; Zhang, Zhang, & He, 2004). It leads to failing to study the proteins' biophysical, structural and functional properties (Pawel et al., 2007). It becomes a huge barrier in proteomics.

There are many experimental tools to deal with this problem, such as, fusion proteins (Kapust & Waugh, 1999; Davis, Elisee, Newham, & Harrison, 1999), co-expression of chaperones (Tresaugues et al., 2004), modified growth media (Georgiou & Valax, 1996), weak promoters (Makrides, 1996) and so on. However, the experimental effectiveness is far from success. All of these methods cannot produce the desire effect for many insoluble proteins.

With the development of the bioinformatics, it is possible to predict the protein solubility with its primary sequence. It can avoid the blindness in the trial-and-error procedure. It can identify the promising candidate protein to improve the experimental success rate. There have been many successful studies on this problem. For example, Wilkinson and Harrison (1991) extracted the proteins' features from turning-forming residue, cysteine fraction, hydrophilicity, average charge, poline fraction and total number of residues. Subsequently, based on the former study, Davis found turning-forming residue, cysteine fraction were the significant factors for predicting the protein solubility (Davis et al., 1999). With the rapid development of the machine learning algorithm, this kind of methods was widely used in the follow-up researches. Artificial Neural Network (ANN) was used to predict the protein solubility in the lysozyme–NaCl–H_2O system (Zhang et al., 2004). As a powerful prediction engine, the Support Vector Machine (SVM) was widely used in the recent studies. Thomas et al. extracted six physicochemical features together with residue- and dipeptide-compositions to predict the protein solubility (Thomas, Kulkarni, Kulkarni, Jayaraman, & Balaji, 2006). Smialowski et al. developed a two-layered classifier with SVM and Naïve Bayes to assess the chance of a protein to be soluble (Pawel et al., 2007). Magnan et al. proposed a two-stage SVM based on the 23 groups of features derived directly from the primary sequence to improve the performance (Magnan, Randall, & Baldi, 2009). For combination the advantages of ANN and SVM, a hybrid approach was proposed by

* Corresponding author.
E-mail address: nnl1010@126.com (L. Nana).

0957-4174/$ - see front matter © 2013 Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.eswa.2013.08.064

Niu, Li, Shi, Hu, and Xia (2010). Agonstini et al. predicted the solubility of proteins with their physicochemical properties of the proteins' thermodynamic stability (Agostini, Vendruscolo, & Tartaglia, 2012). Recently, Niu et al. extracted new feature with the concept of fractal dimension to reveal the sequence-order information which was computed directly from primary sequence to improve the performance of classifier (Niu, Hu, Shi, & Xia, 2012).

According to the recent review (Chou, 2011), to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us follow these steps to construct our predictor.

Considering the mathematical expression for protein samples, Amino Acid Composition (ACC), dipeptide composition, and so on, are the simplest and important discrete models for the prediction of proteins' attributes. However, the obvious shortcoming of these models is losing the sequence-order information, which limits the accuracy of these models. To overcome this shortcoming, the different discrete model, the so-called 'pseudo amino acid composition' (PseAAC) model is proposed by Chou (Chou, 2001; Chou & Shen, 2009). So far, the concept of PseAAC was widely used in the various studies on the problems in proteins and protein-related systems from the introduction of this concept (Chou, Wu, & Xiao, 2011; Chou, Wu, & Xiao, 2012; Du, Wang, Xu, & Gao, 2012; Mei, 2012; Nanni, Lumini, Gupta, & Garg, 2012). Following the general idea of PseAAC, the Chaos Game Representation (CGR) and Entropy in information theory were introduced to explore the further information hidden in the sequence order.

On the one hand, the Chaos Game Representation (CGR) is a graphical tool to provide intuitive picture for helping the analyzing the complicated information hidden in the primary sequence. It was firstly proposed for the representation of DNA sequences by Jeffery in 1990 (Jeffrey, 1990). From the mathematical view, CGR is an iterative mapping technique by regarding biological sequence as chaotic system. Jeffrey's success demonstrated that CGR method can explore the information hidden in primary sequence and find discriminate the genes with different functions. Afterwards, CGR of DNA sequences has been extended to represent protein sequences. Basu et al. used 12-sided regular polygon to generate the pictorial representation of protein sequence, each vertex of which represents a group of amino acid residues leading to conservative substitutions (Basu, Pan, Dutta, & Das, 1997). The grid is deemed as the discriminative and diagnostic signature to predict the proteins' attributes. Following this idea, CGR was widely used to study a series of important biological problems (Liu, Lu, & Hu, 2011; Xia et al., 2011; Yang et al., 2009). Among them, Xia et al. (2011) have proposed triangle, quadrangle and 12-vertex polygon CGR based on the Basu's CGR with different classification of the 20 native amino acids. This upgrade CGR method could reflect much more meaningful information hidden in CGR figures, which was demonstrated by performance of the numerical experiments.

On the other hand, the Entropy in the information theory is a measure of unpredictability or information content. When the outcome of random event is relatively unpredictable, in this case the entropy is large. From another point of view, it has more new information. On the contrary, if it is fairly predictable, it has no new information. So it is the expected value of unpredictability for a random variable. It can reveal the regularity hidden in the primary sequence.

In this paper, following the frame of PseACC, a novel method based on the upgrade CGR feature together with the concept of entropy in information theory is developed to improve the accuracy of protein solubility prediction. According to the upgrade CGR method, triangle, quadrangle and 12-vertex polygon CGR in plus ACC and dipeptide composition were taken into the feature pool. SVM was applied as prediction engine. Numerical experiments were carried out to evaluate the efficiency of using these features in classifiers. The performance of the different predictors is objectively evaluated with 10-fold cross validation test and re-substitution test. The numerical results show that among the three polygon methods, triangle method performed the best in the classifier construction. With the 2-level triangle polygon CGR + dipeptide composition together with their corresponding entropies as the mathematical feature, the classifier achieved the best accuracy, 88.45% in 10-fold cross validation test and MCC achieved 0.7588. In the re-substitution test, the 3-level triangle polygon CGR, dipeptide composition and their entropies won the championship, its accuracy was 92.38%, MCC achieved 0.8387.

2. Materials and methods

2.1. Benchmark dataset

The dataset in this paper follows our previous work (Niu et al., 2012). Let us recall the procedures to obtain the benchmark data. Firstly, we screened the related protein sequences with 'soluble' and 'insoluble' as in National Center for Biotechnology Information database (NCBI, http://www.ncbi.nlm.nih.gov/). There were 69686 soluble proteins and 18034 insoluble proteins to hit. Subsequently, in order to reduce the size of dataset, 5000 soluble bacteria protein sequences and 4500 insoluble bacteria nucleotide sequences were randomly picked out. Finally, we removed the homologous sequences with the threshold (90% homologous similarity) by CD-HIT (Huang, Niu, Gao, Fu, & Li, 2010), thus there were 2448 soluble sequences and 3244 insoluble sequences in the final dataset.

In the recent studies, a cutoff threshold of 25% was imposed in (Chou & Shen, 2010a; Chou & Shen, 2010b; Chou et al., 2012) to reduce the redundancy. The proteins in the benchmark datasets that had equal to or greater than 25% sequence identity to any other in a same subset were excluded. Taking the final size of the final dataset into consideration, in this study we did not use such a stringent criterion. Otherwise, there were so limited proteins left for some subsets. It would be too few to have statistical significance.

2.2. Chaos Game Representation (CGR)

At first, let us recall the general Chaos Game Representation (CGR) algorithm for protein sequence. CGR is a method developed by Basu and updated by Xia (Basu et al., 1997; Xia et al., 2011) to generate visually identifiable distinct patterns of protein sequences. With classifying 20 native amino acids into 12 different groups by conservative substitutions, a given protein sequence is mapped onto a 12-sided regular polygon graph. In a further step, the percentages of points plotted in 24 different segments of the CGR graph are extracted to quantify the CGR patterns.

According to the Xia's work, two more different category modes were introduced. It led to three different kinds of CGR methods based on the same procedure of Basu's CGR algorithm. All the three category modes were show as follow:

(1) Category mode for 12-vertex ploygen CGR (Dayhoff, 1978): I, L, V, M; R, K; D, E; N; Q; H; S, T; P; A, G; C; F, Y; W.