Computers in Biology and Medicine 43 (2013) 1941–1948

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Disulfide connectivity prediction based on structural information without a prior knowledge of the bonding state of cysteines

Hsuan-Hung Lin [a,1], Jiin-Chyr Hsu [b,1], Yuan-Nian Hsu [c], Ren-Hao Pan [a,d], Yung-Fu Chen [e,f,*], Lin-Yu Tseng [d,g,**]

[a] Department of Management Information System, Central Taiwan University of Science and Technology, Taichung 40601, Taiwan
[b] Department of Internal Medicine, Taoyuan General Hospital, Ministry of Health and Welfare, Taoyuan 33004, Taiwan
[c] Department of Director Center, Taoyuan General Hospital, Ministry of Health and Welfare, Taoyuan 33004, Taiwan
[d] Department of Computer Science and Engineering, National Chung Hsing University, Taichung 40227, Taiwan
[e] Department of Healthcare Administration, Central Taiwan University of Science and Technology, Taichung 40601, Taiwan
[f] Department of Health Services Administration, China Medical University, Taichung 40402, Taiwan
[g] Department of Computer Science and Communication Engineering, Providence University, Taichung 43301, Taiwan

ARTICLE INFO

Article history:
Received 24 April 2013
Accepted 10 September 2013

Keywords:
Disulfide bonding pattern
Support vector machine
Multiple trajectory search
Ensemble classifier

ABSTRACT

Previous studies predicted the disulfide bonding patterns of cysteines using a prior knowledge of their bonding states. In this study, we propose a method that is based on the ensemble support vector machine (SVM), with the structural features of cysteines extracted without any prior knowledge of their bonding states. This method is useful for improving the predictive performance of disulfide bonding patterns. For comparison, the proposed method was tested with the same dataset SPX that was adopted in previous studies. The experimental results demonstrate that bridge classification and disulfide connectivity predictions achieve 96.5% and 89.2% accuracy, respectively, using the ensemble SVM model, which outperforms the traditional method (51.5% and 51.0%, respectively) and the model that is based on a single-kernel SVM classifier (94.6% and 84.4%, respectively). For protein chain and residue classifications, the sensitivity, specificity, and accuracy of ensemble and single-kernel SVM approaches are better than those of the traditional methods. The predictive performances of the ensemble SVM and single-kernel models are identical, indicating that the ensemble model can converge to the single-kernel model for some applications.

1. Introduction

Disulfide bonds constitute an important cross-linkage between cysteine side chains in proteins and are known to play a key role in stabilizing protein conformations and functions. Protein folding simulations demonstrate that correctly predicted disulfide bonding patterns can efficiently reduce the search space [1,2]. The disulfide bonds impose geometrical constraints on the protein backbones; therefore, the correct prediction of the disulfide bonding pattern may greatly help to predict the three-dimensional structure of a protein, which, in turn, can manifest its function.

Disulfide bonding patterns can be divided into inter- and intra-chain disulfide bonds. Niu et al. [3] provided a method for the classification of inter- and intra-chain disulfide bonds. In previous works, the prediction of disulfide bonding patterns only focused on intra-chain disulfide bonds because the cysteines that are contained in inter-chain disulfide bonds are considered free cysteines. The realm of disulfide bond predictions can be divided into four problems [4]: (i) protein chain classification to determine whether a protein contains disulfide bridges; (ii) residue classification to categorize the bonding state of cysteines; (iii) bridge classification to predict whether a pair of cysteines is linked by a disulfide bond; and (iv) disulfide bonding pattern prediction to predict the cysteine pairs that are bonded to each other.

Recently, several approaches have been proposed to predict cysteine bonding states and disulfide bonding patterns. These approaches can be grouped into the following three categories: (i) methods for predicting the cysteine bonding states [5–8]; (ii) approaches for predicting the disulfide bonding patterns with a prior knowledge of the cysteine bonding states [4,9–17]; and (iii) methods for predicting both cysteine bonding states and

* Corresponding author at: 666 Buzih Road, Beitun District, Taichung 40601, Taiwan, ROC. Tel.: +886 4 22391647x7285; fax: +886 4 22394465.
** Corresponding author at: 666 Buzih Road, Beitun District, Taichung 40601, Taiwan, ROC. Tel.: +886 4 26328001x18314
E-mail addresses: shlin@ctust.edu.tw (H.-H. Lin), jinchyr.hsu@msa.hinet.net (J.-C. Hsu), hsu2101@mail.tygh.gov.tw (Y.-N. Hsu), bphouse.tw@gmail.com (R.-H. Pan), yfchen@ctust.edu.tw (Y.-F. Chen), lytseng@pu.edu.tw (L.-Y. Tseng).
[1] These authors contributed equally to this work.

0010-4825/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiomed.2013.09.008

disulfide bonding patterns [4,11]. In the prediction of cysteine bonding states, several methods that are based on statistical analysis [5], neural networks [6,7], and support vector machines [8] have been proposed, with significant progress being made in the prediction of the cysteine bonding state, with achieved accuracies ranging from 81% to 90%. The computational approaches that are used to predict disulfide bonding patterns have also been proposed recently. Fariselli and Casadio [9] proposed a method to convert the prediction problem into a graph matching problem with vertices that indicate the oxidized cysteines and edge weights that are labeled as the contact potentials between the corresponding pairs of cysteines. The optimal values of contact potentials were obtained using the Monte Carlo simulated annealing method, and then the disulfide bonds were located by finding the maximum weight perfect matching. Vullo and Frasconi [10] applied an ad-hoc recursive neural network to improve the prediction accuracy from 34% to 44%. Cheng et al. [4] used two-dimensional recursive neural networks for predicting the connectivity probabilities between cysteine pairs to further improve the accuracy. Ferrè and Clote [11] used the secondary structure information and diresidue frequency to train the predictive model that was designed based on the diresidue neural network to predict connectivity probabilities between cysteine pairs. Tsai et al. [12] used the local sequence profiles and the linear distance of cysteines as the features for training the support vector machine (SVM) model to predict connectivity probabilities between cysteine pairs.

In contrast to the aforementioned approaches that were based on the conversion of a disulfide bonding pattern prediction problem to a maximum weight perfect matching problem, Cheng and Hwang [13] directly predicted the disulfide bonding patterns by a model that was designed with the SVM model, which was based on the following features: coupling between the local sequence environments of cysteine pairs, cysteine separations, and amino acid contents. Conversely, Zhao et al. [14] used a simple feature, i.e. cysteine separation profiles (CSPs), which is based on the assumption that similar protein structures have similar disulfide bonding patterns, to predict the disulfide bonding pattern. Chen et al. [15] proposed a two-level model, consisting of a pair-wise level and a pattern-wise level, by extending local information regarding cysteine pairs (pair-wise) to global information, including protein length, cysteine separation, and disulfide connectivity frequency (pattern-wise) for the prediction of disulfide bonding patterns. Lu et al. [16] adopted the genetic algorithm (GA) to optimize the feature selection for training the SVM model, achieving an accuracy of 73.9%. Song et al. [17] used the multiple sequence vectors and secondary structure information to train a support vector regression model for the prediction of disulfide bonding patterns, with an accuracy of 74.4%. Recently, Lin and Tseng [18] used features, including position-specific scoring matrix (PSSM), normalized bond lengths, predicted secondary structure of proteins, and indices of the physicochemical properties of amino acids for training a SVM model for the prediction of disulfide bonding pattern with an accuracy of 79.8%.

The computational methods for the prediction of disulfide bonding patterns can be divided into sequence-based and structural-based [19]. In general, the features used in the aforementioned methods are primarily extracted from the protein sequence. In this study, the structural information, i.e. *X*, *Y*, and *Z* coordinates of the C_α of each amino acid, which was contained in the protein that was predicted by the MODELLER software [20] was used to calculate the feature *NPD*. In the work that was performed by Lin and Tsang [21], a single-kernel SVM was used for the bridge classification and prediction of disulfide bonding patterns, which mainly focused on the web service to provide a tool for biologists and medical scientists. In contrast, this study applied an ensemble SVM consisting of three kernels to train the model to compute the connectivity probabilities of cysteine pairs, followed by the modified maximum weight perfect matching algorithm to find the disulfide bonding pattern. This SVM has been demonstrated to have better predictive performance than the methods that were proposed by Cheng et al. [4] and the single-kernel SVM model [21]. The method proposed in this work aims to improve the predictive performance of the disulfide bonding pattern of a protein sequence that does not have cysteines involved in the metal-binding sites. An updated version of the web service [21] is available at http://biomedical.ctust.edu.tw/edbcp.

2. Materials and methods

2.1. Datasets

To compare the predictive performance of our proposed method with previously reported methods [4,7,8], the same dataset, which was denoted as MART and provided by Martelli et al. [7], was employed for the experiments. In the dataset MART, a total of 4136 segments containing cysteines in 969 protein sequences were extracted from PDB [22] with sequence identities less than 25% and without chain breaks (non-homologous). The segments with cysteines that were inter-chain disulfide bonded were included as 'free' cysteines (non-disulfide-bonded). Among the 4136 segments with cysteines, 2690 were in the free state, and the other 1446 were in the disulfide bonded state. The dataset was split into 20 subsets of roughly equal size by Martelli et al. for 20-fold cross-validation to verify the performance of their proposed method.

The dataset SPXC was adopted to address the problem of protein chain classification, whereas the dataset SPX was used to pinpoint the problems regarding residue classification, bridge classification, and disulfide bonding pattern prediction. The datasets SPXC and SPX were prepared by Cheng et al. [4], with all the proteins in both SPXC and SPX datasets that were extracted from the PDB on May 17, 2004. To remove overrepresentation of particular protein families, the UniqueProt tool [23], which was designed based on the HSSP distance [24] to reduce protein redundancy, was adopted [4]. In dataset SPXC, there are 897 positive sequences that were selected with an HSSP cutoff distance of 5 and 1650 negative sequences that contained no disulfide bridges with an HSSP cutoff distance of 0. The dataset SPX is a collection of 1018 proteins, which contain at least one intra-chain disulfide bond and at least 12 amino acids that were obtained by setting the HSSP cutoff distance to 10. To compare the methods that were proposed in this study and the method that was proposed by Cheng et al. [4], the protein sequences were randomly divided into 10 subsets of roughly equal size for 10-fold cross-validation.

2.2. Methodology

The method proposed by Cheng et al. [4] can be divided into two stages. This method first predicted the bonding state of cysteines, and then oxidized cysteines were used for the prediction of disulfide bonding patterns. In this study, in contrast, the bonding probability of all the cysteine pairs was directly predicted. The normalized pair distance (*NPD*) was adopted as the feature, and then the SVM model was trained to compute the connectivity probabilities of cysteine pairs. Afterward, the MTS [25] was used to evolve the SVM parameters, i.e., C and γ, the *NPD* window sizes, and the weights of individual models in the ensemble SVM classifier. Finally, the modified maximum weight perfect matching algorithm was then used to find the disulfide connectivity pattern without a prior knowledge of the bonding state of cysteines.