Analysis of four different sets of predictive features for metalloproteins

Huseyin Seker[1] and Parvez I. Haris[1, 2]
[1]*Bio-Health Informatics Research Group, Centre for Computational Intelligence, School of Computing, De Montfort University, Leicester, LE1 9BH, UK*
[2]*Faculty of Health and Life Sciences, De Montfort University, Leicester, LE1 9BH, UK*
(hseker@dmu.ac.uk) (pharis@dmu.ac.uk)

Abstract

Metals bound to the protein are important for functional or structural roles. Despite their importance there is a distinct lack of research for identification of metalloproteins from sequence data and their predictive features that help distinguish them from non-metal binding proteins. In this study, four sets of features were analysed in order to see their ability to distinguish between metal and non-metal binding proteins. The analysis was carried out using a novel fuzzy logic method. The results show that the amino acid composition is more capable of distinguishing metal from non-metal binding proteins, than any of the other three features, yielding a predictive accuracy of 69.4%. Cofactors were the least useful feature for distinguishing metalloproteins. However, better results were obtained when physico-chemical and secondary structure features are used, yielding accuracies of 67.8% and 67.1%, respectively. Although the amino acid composition yields the highest predictive accuracy, considering the number of features, the latter two sets of features may be more appropriate for such analysis.

1. Introduction

Metalloproteins are proteins that bind one or more metal ions. Metals bound to the protein may be important for functional or structural roles in biological systems. In some cases metal binding to a protein, as often observed *in vitro*, may not have a physiological relevance. It is estimated that as much as 30% of most genomes encode metalloproteins, and the term "metallomics" has recently been used to define a new scientific field encompassing metal binding molecules including metalloproteins [1]. Despite their importance there is a distinct lack of research for identification of metalloproteins from sequence data and their predictive features that help distinguish them from non-metal binding proteins.

For protein function identification, a number of different sets of features have been explored. They are mainly based on amino acid composition and secondary structure information. The later set of the features consisting of four predictive attributes (α-helix, β-strand, 3_{10}-helix and others) has previously been used for identification of metalloproteins from the sequence data [2]. It has also been shown that cofactors (ATP, NAD, AND FAD), and physico-chemical parameters (hydropathicity of proteins) and theoretical isoelectric point (pI) are potentially useful sets of features that can be used to distinguish metal binding proteins from non-metal binding proteins [3].

A reliable and accurate identification for protein functions depends on a well-defined feature set as well as a robust computational method [4]. Therefore, it is important to use a robust computational intelligence tool to identify reliably a set of features that characterize the structure.

In this study, a fuzzy logic-based method, namely enhanced version of the fuzzy k-nearest neighbour (EFK-NN) was used to assess four different sets of the features to identify if any of them can be used as a reliable feature set for predicting metal binding proteins.

2. Methodology

2.1 Data Collection and Data Structure

Information regarding the proteins used in this study was collected from the protein data bank (www.pdb.org). Metal binding proteins were identified as those proteins, deposited in the protein data bank that had one or more metal ions bound to it. Non-metal binding proteins were classified as those not containing a metal in their 3D structure. Data was collected for 301 proteins. Of these, 150 samples contained one or

more metal ions in their 3D structure, and 151 proteins that did not.

The coordinate files of each protein structure were studied to extract a set of predictive features for the two protein functions. In total, 30 different features were extracted, which were grouped into four sub-feature sets, namely (1) amino acid composition of each protein (20 features in total), (2) Secondary structure (α-helix, β-strand, 3_{10}-helix and "others"), (3) Cofactors (ATP, NAD, AND FAD), and (4) physico-chemical parameters (hydropathicity of proteins) and theoretical isoelectric point (pI).

2.2 Enhanced Fuzzy K-Nearest Neighbor Method

Enhanced Fuzzy K-Nearest Neighbor Method (EFK-NN), which has been shown to be a powerful and reliable technique for medical/biological data mining [4], was used to analyse the data. The EFK-NN is defined as a function of the number of neighbourhoods (K), class membership degrees (between 0 and 1), and distances between a pattern to be classified and patterns for which the class membership degrees were previously determined. A class membership degree between 0 and 1 is computed using the first minimum distances and the known class membership degrees of the patterns. The EFK-NN not only gives a class to which the pattern is assigned, but also the class membership degree that provides information about the certainty of the classification decision. The method also provides a membership degree for each subset of the features being analysed. Further details can be found in [4].

3. Results and Discussion

The data was analysed using the EFK-NN with K=1 to 11. The results are based on the leave-one-out cross validation, and listed in Table 1.

The results show that the amino acid composition is more capable of distinguishing metal from non-metal binding proteins, than any of the other three features, yielding a predictive accuracy of 69.4%. The cofactors are the least useful for distinguishing metalloproteins, yielding an accuracy value of only 53.5%. However, better results are obtained when physico-chemical and secondary structure features are used, yielding accuracies of 67.8% and 67.1%, respectively. Although the amino acid composition yields the highest predictive accuracy, considering the number of features, the latter two sets of features may be more appropriate for such analysis.

In our previous study [2], where all combinations of the secondary structure features were analysed using the same method, only two secondary structure features consisting of "β-strand" and "others" reached the accuracy of 73.1%, which is higher than that of amino acid composition. This may indicate that the secondary structure features could be potentially a robust set of predictive features for identifying metalloproteins.

4. Conclusions

Prediction of metal binding proteins based on the four sets of features was carried out by using the EFK-NN. Among these sets, the amino acid composition yields the highest predictive accuracy. However, considering number of features, physico-chemical and secondary structure features may be more appropriate for such analysis. Further research is being carried out to investigate interaction between these features, and then to establish a robust and reliable model for identification of metal binding proteins.

TABLE 1
Predictive accuracy results for metalloproteins

Feature sets (# of features)	K	Predictive accuracy (%)
Amino Acid (20)	11	69.4
Secondary Structure (4)	7	67.1
Co-factors (4)	7	53.5
Physico-chemical (2)	4	67.8

References

[1] H. Haraguchi, "Metallomics as integrated biometal science", Journal of Anal. Atom. Spectrometry, vol. 19(1), pp: 5-14, 2004.

[2] H. Seker, R. Abdulla, and P. Harris, "Prediction of metal binding proteins based on secondary structure information using a fuzzy logic-based method", Proc. of 2004 Congress Proteomics for Health, 2004, Switzerland, pp: 170-172.

[3] H. Seker, R. Abdulla, and P. Harris, "Developing a non-alignment based approach for identifying predictive features in metal binding proteins", Proc. of 2004 Congress Proteomics for Health, 2004, Switzerland, pp: 57-59.

[4] H. Seker, M.O. Odetayo, D. Petrovic, and R.N.G. Naguib, "A fuzzy logic based-method for prognostic decision making in breast and prostate cancers", IEEE Trans. on Information Technology in Biomedicine, vol. 7(2), pp: 114-122, 2003.