Oxypred: Prediction and Classification of Oxygen-Binding Proteins

S. Muthukrishnan, Aarti Garg, and G.P.S. Raghava*

Institute of Microbial Technology, Sector 39-A, Chandigarh 160036, India.

This study describes a method for predicting and classifying oxygen-binding proteins. Firstly, support vector machine (SVM) modules were developed using amino acid composition and dipeptide composition for predicting oxygen-binding proteins, and achieved maximum accuracy of 85.5% and 87.8%, respectively. Secondly, an SVM module was developed based on amino acid composition, classifying the predicted oxygen-binding proteins into six classes with accuracy of 95.8%, 97.5%, 97.5%, 96.9%, 99.4%, and 96.0% for erythrocruorin, hemerythrin, hemocyanin, hemoglobin, leghemoglobin, and myoglobin proteins, respectively. Finally, an SVM module was developed using dipeptide composition for classifying the oxygen-binding proteins, and achieved maximum accuracy of 96.1%, 98.7%, 98.7%, 85.6%, 99.6%, and 93.3% for the above six classes, respectively. All modules were trained and tested by five-fold cross validation. Based on the above approach, a web server Oxypred was developed for predicting and classifying oxygen-binding proteins (available from http://www.imtech.res.in/raghava/oxypred/).

Key words: oxygen-binding proteins, SVM modules, hemoglobin, web server, prediction

Introduction

Oxygen-binding proteins are widely present in eukaryotes ranging from non-vertebrates to humans (*1*). Moreover, these proteins have also been reported to be present in many prokaryotes and protozoans (*2*). The occurrence of oxygen-binding proteins in all kingdoms of organisms, though not in all organisms, shows their biological importance. Extensive studies on oxygen-binding proteins have categorized them into six different broad types, including erythrocruorin, hemerythrin, hemocyanin, hemoglobin, leghemoglobin, and myoglobin, each has its own functional characteristics and structure with unique oxygen-binding capacity. These oxygen-binding proteins are crucial for the survival of any living organism. With the advancement in sequencing technology, the size of protein sequence databases is growing at an exponential rate. Thus it is much needed to develop bioinformatic methods for functional annotation of proteins, particularly for identifying oxygen-binding proteins (*3*, *4*).

Recently, Lin *et al* (*5*) have developed a support vector machine (SVM)-based method for predicting functional classes of metal-binding proteins. However, to the best of our knowledge, no method has been developed specifically for predicting and classifying oxygen-binding proteins. In the present study, we have developed a reliable SVM-based method for predicting and classifying oxygen-binding proteins using different residue compositions.

***Corresponding author.**
E-mail: raghava@imtech.res.in

Results and Discussion

Prediction of oxygen-binding proteins

SVM modules were trained and tested on our dataset of oxygen-binding and non-oxygen-binding proteins. First we developed an SVM module using amino acid composition and achieved a Matthew's correlation coefficient (MCC) value of 0.71 with 85.5% accuracy when evaluated by five-fold cross validation. It has been shown that dipeptide composition provides more information than simple amino acid composition because dipeptide composition encapsulates local order information (*6*). Thus we developed an SVM module using dipeptide composition and achieved an MCC value of 0.76 with 87.8% accuracy, 88.5% sensitivity, and 87.1% specificity. This result demonstrates that the dipeptide composition-based module performs better than the amino acid composition-based module for the prediction of oxygen-binding proteins.

Classification of oxygen-binding proteins

We classified the predicted oxygen-binding proteins into six classes, including erythrocruorin, hemerythrin, hemocyanin, hemoglobin, leghemoglobin, and myoglobin. It was found that the compositions vary significantly from one class to another (Figure 1), indicating that one class of proteins can be discriminated from other classes based on amino acid composition. Therefore, we developed six SVM modules corresponding to the six classes, respectively. First, we developed amino acid composition-based SVM modules and achieved accuracy from 95.8% to 99.4% with an overall accuracy of 97.2% (Table 1). Then we developed dipeptide composition-based SVM modules and achieved accuracy from 85.6% to 99.6% with an overall accuracy of 95.3% (Table 1). It is interesting to note that here the performance of the amino acid composition-based module is better than that of the dipeptide composition-based module (*7*). This study demonstrates that it is possible to predict and classify oxygen-binding proteins using compositional information (amino acid and dipeptide).

Oxypred server

The SVM modules constructed in the present study have been implemented as a web server Oxypred using CGI/Perl script, which is available for academic use at http://www.imtech.res.in/raghava/oxypred/. Users can submit protein sequences in one of the standard formats such as FASTA, GenBank, EMBL, or GCG. The server first predicts oxygen-binding proteins and then classifies them into the six classes.

Materials and Methods

Dataset

We extracted the sequences of oxygen-binding and non-oxygen-binding proteins from Swiss-Prot database (http://www.expasy.org/sprot/) (*8*). In order to obtain a high-quality dataset, we removed all those proteins annotated as "fragments", "isoforms", "potentials", "similarity", or "probables" (*9*, *10*), and created a non-redundant dataset where no two proteins have a similarity more than 90% using PROSET software (*11*). Our final dataset consisted of 672

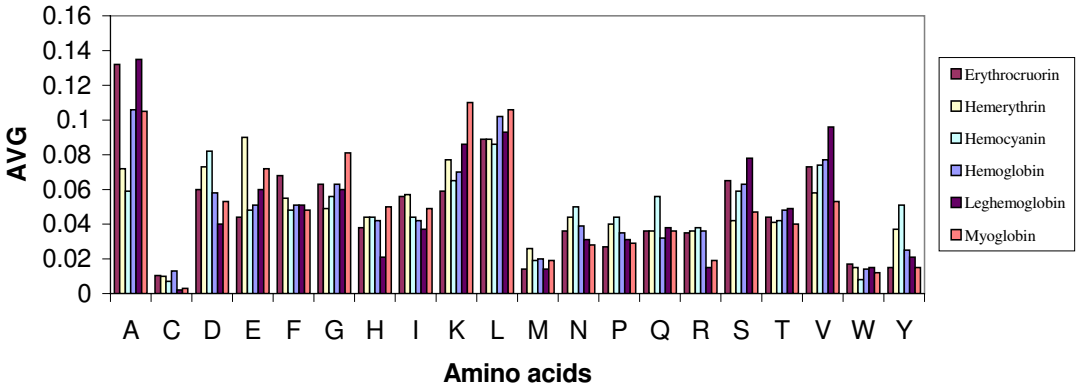

Fig. 1 Average (AVG) amino acid composition of six different classes of oxygen-binding proteins. Amino acids are denoted by their single letter codes.

Table 1 Performance of SVM modules for classifying oxygen-binding proteins

Protein class	Accuracy (%)	
	Amino acid composition	Dipeptide composition
Erythrocruorin	95.8	96.1
Hemerythrin	97.5	98.7
Hemocyanin	97.5	98.7
Hemoglobin	96.9	85.6
Leghemoglobin	99.4	99.6
Myoglobin	96.0	93.3
Average	97.2	95.3

oxygen-binding proteins and 700 non-oxygen-binding proteins. These 672 oxygen-binding proteins were then classified into six different classes, consisting of 20 erythrocruorin, 31 hemerythrin, 77 hemocyanin, 486 hemoglobin, 13 leghemoglobin, and 45 myoglobin proteins.

Support vector machine

SVM modules were implemented by a freely downloadable package of SVM^{light} (http://www.cs.cornell.edu/people/tj/svm_light/). The software enables users to define a number of parameters as well as inbuilt kernel functions such as linear kernel, radial basis function and polynomial kernel (of a given degree). In order to develop the prediction method, we trained SVMs using oxygen-binding proteins as positive labels and non-oxygen-binding proteins as negative labels. For classifying oxygen-binding proteins, we used the one-versus-rest SVM strategy.

Input features and performance evaluation

We used amino acid composition and dipeptide composition as input features. For amino acid composition, a protein is represented by a vector of 20 dimensions, while for dipeptide composition a protein is represented by a vector of 400 dimensions. We used the five-fold cross validation technique to evaluate the performance of SVM modules (*12*, *13*). The performance of these modules were measured with standard parameters like accuracy, sensitivity, specificity, and MCC (*14*).

Acknowledgements

We thank Dr. K.L. Dixit for her valuable inputs. This work was supported by the Council of Scientific and Industrial Research (CSIR) and the Department of Biotechnology, Government of India.

Authors' contributions

SM and AG created datasets, developed various modules, and evaluated all modules. SM and AG also developed the web server. GPSR conceived the idea, coordinated it and refined the manuscript drafted by SM and AG. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Zhang, L., *et al.* 2007. Recent developments and future prospects of *Vitreoscilla* hemoglobin application in metabolic engineering. *Biotechnol. Adv.* 25: 123-136.
2. Wu, G., *et al.* 2003. Microbial globins. *Adv. Microb. Physiol.* 47: 255-310.
3. Garg, A., *et al.* 2005. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.* 280: 14427-14432.
4. Kumar, M., *et al.* 2006. Prediction of mitochondrial proteins using support vector machine and hidden markov model. *J. Biol. Chem.* 281: 5357-5363.
5. Lin, H.H., *et al.* 2006. Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. *BMC Bioinformatics* 7: S13.
6. Bhasin, M. and Raghava, G.P. 2004. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* 32: W414-419.
7. Saha, S. and Raghava, G.P. 2006. VICMpred: SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics* 4: 42-47.
8. Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28: 45-48.
9. Saha, S. and Raghava, G.P. 2007. BTXpred: prediction of bacterial toxins. *In Silico Biol.* 7: 0028.
10. Saha, S., *et al.* 2007. VGIchan: prediction and classification of voltage-gated ion channels. *Genomics Proteomics Bioinformatics* 4: 253-258.
11. Brendel, V. 1992. PROSET—a fast procedure to create non-redundant sets of protein sequences. *Mathl. Comput. Modelling* 16: 37-43.
12. Bhasin, M. and Raghava, G.P. 2004. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.* 32: W383-389.
13. Saha, S. and Raghava, G.P. 2006. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* 34: W202-209.
14. Kaur, H. and Raghava, G.P. 2004. Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. *Proteins* 55: 83-90.