Deep Neural Network for Classification and Prediction of Oxygen Binding Proteins

Soumiya Hamena
Computer Science and Applications Department
FNTIC
Constantine, Algeria
Soumiya.hamena@univ-constantine2.dz

Souham Meshoul
Computer Science and Applications Department
FNTIC
Constantine, Algeria
Souham.meshoul@univ-constantine2.dz

ABSTRACT

The accurate annotation of a protein function is important for understanding life at molecular level. Nowadays, powerful high throughput proteomics technologies provide an unprecedented understanding of the human biology and disease. These technologies are generating a deluge of protein sequences available in public databases. However, a critical challenge in making sense of these sequences is the assignment of functional roles to newly discovered proteins. The approaches proposed to address this problem use a variety of biological information, such as amino acid sequence, gene expression and protein-protein interaction. By another way, deep learning has emerged as the innovation of this last decade as it uses deep architectures to learn representations of high level entities and creates an improved functional space. In this paper, we propose an approach that proposes a deep neural network to achieve classification of oxygen binding proteins using amino acid composition for protein function prediction. Two alternatives are investigated. The first one casts the tackled problem as a multiclass classification problem and the second one as a binary classification problem. The validation of the approach is achieved using Keras platform and very promising and encouraging results that outperform other state of the art results have been obtained.

CCS Concepts

•Computing methodologies → Neural networks; **•Computing methodologies → Supervised learning by classification**; **•Applied computing → Proteomics**

Keywords

Oxygen binding proteins; Protein function prediction; Proteomics; Classification; Deep neural network

1. INTRODUCTION

The unique use of genomic and transcriptomic information may be insufficient to fully understand a complex organism [1]. Proteomics research is the logical next step after genomics in understanding the working of the cell and therefore is highly demanded by biomedical research and pharmaceutical applications [2]. Proteomics includes knowledge of the structure, function, and expression of all proteins in organism that underlie growth, development, and interactions with the environment. Thus, it is increasingly important to discover the function for new proteins to advance our understanding for life at molecular level. In other words, automated annotation of protein function has become a critical task in the post-genomic era [3]. Meanwhile, protein function prediction is a great challenge in bioinformatics and computational biology [4].With the difficulty and the complexity of the accurate annotation of protein function, biochemical properties of protein function cannot accommodate the huge amount of sequence data already available. The fact that more information becomes available in the databases, more powerful tools and more efficient methods for finding relationships within these data will need to be implemented [5].

According to UniProt, the largest public database of protein sequences, less than one percent of all known proteins correspond directly to experimental annotation, with the rapid explosion of sequences, the difference between known sequences and annotated sequences decreases the pace of scientific discoveries [6]. Classification of the protein presented in sequenced genomes is very important for drawing useful knowledge from the genomic sequences. Therefore, bioinformatics fills this need by developing computational methods for protein function and structure prediction, using statistics, machine learning, data mining and particularly supervised learning algorithms. Classification is one of the most common applications for supervised learning. Indeed, several articles have been published on this subject [7]. A new implementation of InterProScan has been described which aims to improve previous versions in terms of scalability and usability. This tool is applied to classify sequences into protein families and to predict the existence of important domains and sites [8], [9]. Phyre2 is the update of tools available on the web for the prediction and analysis of the structure, function and mutations of proteins [10]. Also, an extended evaluation of methods for predicting protein function has been performed by conducting the second critical assessment of functional annotation (CAFA) whose 126 methods from 56 research groups were evaluated for their ability to predict biological functions using Gene Ontology and gene disease association using Human Phenotype Ontology on a set of 3681 proteins of 18 species [11]. PROVEAN is a prediction tool for protein sequence from any organisms that has been presented as a web server that can predict the functional effect of single or multiple amino acid substitutions, insertions, and deletions [12].

By another way, with deep learning, artificial intelligence has never been better named because it is inspired by the neural network of the human brain to reproduce truly intelligent learning methods [13]. Deep learning is a machine learning model that has demonstrated high performance for a wide range of applications; it has progressed impressively since the early 2000s and has facilitated great advances in many areas such as image recognition, speech recognition and natural language processing [14]. Also,

ICCBB 2018, October 11–13, 2018, Bari, Italy
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6552-9/18/10…$15.00
DOI: https://doi.org/10.1145/3290818.3290821

deep learning has been applied successfully to solve several prediction problems in bioinformatics [15], and particularly for the prediction of protein function and structure. Effectively, DeepBind is a fully automatic tool that handles a million of sequences per experiments; it is applied for the prediction of the sequence specificities of DNA and RNA binding proteins, it can address microarray and sequencing data [16]. Also, deep learning framework has been developed to model structural features of RNA binding proteins targets by integrating the primary sequence, predicted secondary and tertiary structural representation of the targets sites [17]. DeepCNF-D is applied for prediction protein Order/Disorder regions using the long range sequential information and the interdependency between adjacent order/disorder labels [18].

In this context, we propose an approach that proposes a deep learning based model to achieve classification of oxygen binding proteins using amino acid composition to predict function of proteins. Oxygen binding proteins are ancient molecules that have probably evolved from enzymes that preserved the organism against the toxic oxygen [19]. To the best of our knowledge; this problem has not been tackled yet using deep learning. Most of proposed models make use of support vector machines (SVM) [20] and Random Forest [21]. An accuracy of 89.22% has been obtained. Using a deep learning solution has the potential to achieve better classification accuracy, the reason for which we investigate the use of deep learning. The remainder of the paper is organized as follows. Section 2 presents some background material and related work. Section 3 describes the proposed deep learning model for oxygen binding proteins prediction. In section 4, we present and discuss the experimental study and results. Finally conclusions and plans for future work are drawn in the last section.

2. BACKGROUND AND RELATED WORK

2.1 Oxygen Binding Proteins Classification

Oxygen binding proteins bind oxygen reversibly and store it or deliver it from the lung and deposit it through the body's cells to allow cellular respiration, which through metabolism provides the energy of biological processes essential to life. The classification of oxygen binding proteins includes 6 different types namely Erythrocruorin, Myoglobin, Hemerythrin, Hemocyanin, Hemoglobin and Leghemoglobin [19], [22], [23]. At present, oxygen binding proteins have been discovered in all kingdoms of life. These proteins are present mainly in the red blood cells of vertebrates and in the tissues of certain invertebrates, in mammals, in many prokaryote and protozoan. These proteins also carry other ligands, some of which are competitive inhibitors such as carbon monoxide, and this can lead to hypoxia that causes vertigo, nausea, headache, and tachycardia and can even lead to death [24]. Moreover, some mutations of these proteins can cause genetic diseases called hemoglobinopathies such as sickle cell disease, which was the first human disease whose mechanism has been elucidated at the molecular level [25]. Thalassemias form another type of hemoglobinopathies involving an alteration of globins gene regulation, all these diseases result in anemia. For this, oxygen binding proteins are essential for the survival of any living organism. To the best of our knowledge, Oxypred is the first proposed tool for the classification and the prediction of oxygen binding proteins based on support vector machine (SVM) using amino acid and dipeptide composition, achieved 85.5% and 87.8%, respectively [20]. Recently, Random Forest machine learning tool has been proposed and achieved 89.22% of accuracy using amino acid composition [21].

2.2 Deep Neural Network

Deep learning is a sub-domain of machine learning that focuses on algorithms inspired by the structure and function of the brain named Artificial Neural Network. The concept of deep learning algorithms is to extract representations at high level of abstraction from massive volume of input data [26]. These algorithms are widely motivated by the field of artificial intelligence, and aim to emulate the human brain's ability to observe, analyze, learn and make decisions for complex problems [27]. What differentiates deep learning from other machine learning techniques is that everything is on the scale, that when we build larger neural networks and train them with more data, their performance continue to increase [28]. In fact, the main concept of deep learning consists in the form of feedforwad neural network where the levels of abstraction are modeled by several non-linear hidden layers [29]. Earlier model of neural networks contained one input and one output layer and at most one hidden layer. More than three layers including input and output are referred to as deep learning. Thus, the basic structure of deep neural network consists of an input layer, more than one hidden layer and an output layer. The input layer consists on a set of nodes representing the input features $\{x_i|x_1, x_2, \dots, x_m\}$. Each node in the hidden layer transforms the values from the lower layer with a sum of weighted: $w_1x_1 + w_2x_2 +..+ w_mx_m$. Then, the sum is passed by the activation function of a node to calculate the output values of the layer. Finally, the output layer receives the values from the last hidden layer and transforms them into output values.

3. THE PROPOSED NEURAL NETWORK ARCHITECTURES FOR OXYGEN BINDING PROTEINS PREDICTION

Solving oxygen binding proteins prediction and classification problem aims at developing a model M that helps identifying the type of a protein and consequently its function. We can formally describe this task as follows: Given S a set of labeled data such as: $S = \{P_1, P_2, \dots., P_N\}$ where each Pi refers to a protein which is defined using amino acid composition as a D-dimensional vector i.e.$P_i = (a_1, a_2, \dots, a_D, c_i)^t$ where each aj for j=1...D, refers to an attribute or a feature and c_i refers to the class of proteinP_i. In the case of our study, six classes are considered namely Erythrocruorin, Hemerythrin, Hemocyanin, Hemoglobin, Leghemoglobin and Myoglobin. Therefore, the model to be developed should be able to predict the type of a new protein:

$$M(P_{new}) = C$$

Where, $C \in \{\text{Erythrocruorine}, \text{Hemerythrin}, \text{Hemocyanin}, \text{Hemoglobin}, \text{Leghemoglobin and Myoglobin}\}$

In our work, we cast the problem as a multiclass classification problem in one side and as a binary classification problem in another side.

3.1 Oxygen Binding Proteins Prediction as a Multiclass Classification Problem

In this case, we propose a deep fully connected neural network which encompasses three hidden layers as shown on figure below:

As shown on the illustrated architecture (Figure 1), the input layer is fed with data from a Fasta file [21] and prepared to fit the purpose. The input layer includes 20 nodes as the number of amino acids. The three fully connected hidden layers contain as well 20 nodes each. The RELU (Rectified Linear Units) function is used in the three hidden layers. RELU is frequently used in neural networks. In our work, we used a one-hot encoding pattern for our datasets which justifies the use of RELU function; the output attribute of a vector that contains values for each class is transformed to be a matrix with a Boolean for each class value. The function RELU is described as follows:$\mathrm{RELU}(\mathrm{x}) = \max(\mathrm{x}, 0)$; It works like a real neuron of an organism; if the input is greater than 0, the output is equal to the input. Furthermore, we added a dropout regularization technique to prevent neural network from overfitting and improve the model's ability to generalize [30]. Dropout is applied between the first and the second hidden layer, and between the last hidden layer and the output layer. The dropout rate is set to 20%. Then, we created 6 nodes in the output layer and the Softmax function is used in this layer. The Softmax function is used in the output layer to ensure that the output values are between 0 and 1 and can be used as predicted probabilities. The Softmax function is defined as follows: $\mathrm{softmax}(\mathrm{x})_\mathrm{i} = \frac{\exp(\mathrm{x_i})}{\sum_{\mathrm{l}=1}^{\mathrm{k}} \exp(\mathrm{x_l})}$; Where$\mathrm{z_i}$ represents the ith element of the input to softmax, which corresponds to the class i and k is the number of classes.

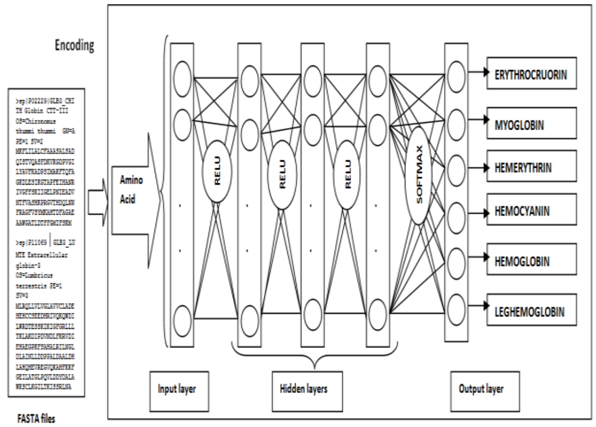
Encoding
>sp|P02229|GLB3_CHITH Globin CTT-III OS=Chironomus thummi thummi GN=A PE=1 SV=2
>sp|P11069|GLB3_LUMTE Extracellular globin-3 OS=Lumbricus terrestris PE=1 SV=3
FASTA files
Amino Acid
RELU
RELU
RELU
SOFTMAX
ERYTHROCRUORIN
MYOGLOBIN
HEMERYTHRIN
HEMOCYANIN
HEMOGLOBIN
LEGHEMOGLOBIN
Input layer
Hidden layers
Output layer

Figure 1. Architecture of the proposed deep neural network for classification of oxygen binding proteins

3.2 Oxygen Binding Proteins Prediction as a Binary Classification Problem

In this case, we propose a binary classification model for each class of oxygen binding proteins. For this, we created six deep fully connected neural network whose only a few changes in the output layer were made compared to the first proposed model; the number of output nodes becomes 1 and the Sigmoid activation function has been applied because the predictions are binary [31]: $Sigmoid(x) = \frac{1}{1+e^{-x}}$.

4. EXPERIMENTAL STUDY

4.1 Description of Used Datasets

The proposed models have been assessed using UniProt database available in [21]. The datasets consist of protein sequences in the standard formats FASTA. There are 16232 sequences of oxygen-binding proteins that belong to 6 different classes distributed as follows: 185 erythrocruorin, 842 myoglobin, 10 hemerythrin, 1515 hemocyanin, 13588 hemoglobin, and 92 leghemoglobin. The numerical representation of the protein sequences was generated using Protr which is a comprehensive R package [32]. Then, the models have been trained and validated on Theano using Keras deep learning library in Python with Anaconda.

4.2 Experimental Results

We used Keras which is a python library for deep learning that covers the efficient numerical computation libraries Theano, Tensorflow and CNTK to develop and validate our proposed models for classification of oxygen binding proteins. The evaluation of the performance of the proposed models is carried out using accuracy and loss measures. To compare our results to others, first we developed a classical neural network with one hidden layer with the same parameters as the proposed models in both cases. Then, we performed the proposed model for the multiple class classification by calculating accuracy and loss and we compared our results with those of the Random Forest model described in [21]. In a second step, we assessed the proposed model for the binary classification using the same performance measures and we compared the obtained results with those of Oxypred [20].We used the efficient Adam gradient descent optimization algorithm to modify the weights of the neural network in order to minimize the errors of network outputs. The reason is that it has been shown to be robust, converge rapidly and able to give good performance [33]. We also used a logarithmic loss function which is called "categorical-crossentropy" for multiple classification and "binary-crossentropy" for binary classification in Keras.

In fact, we split our datasets in a way to get75% for training and 25% for testing purpose. We initialized the random number generator a constant random seed to ensure that the results obtained are achieved again accurately. The models are fit for 100 epochs and the batch size of 10 is used which are the number of instances evaluated before weights are updated within an epoch. The obtained results are presented in Table1 and Table 2, respectively. As can be shown on table 1, our model achieves an accuracy of 95.04% and outperforms the Random Forest based method as well as a classical (non deep) neural network.

Table 1. Experimental results of multiclass classification of oxygen binding proteins

DATA	Neural Network		Proposed model		Random Forest
	ACC	LOSS	ACC	LOSS	ACC
ALL_CLASSES	93.76%	0.2655	95.04%	0.1714	89.22%

As the used datasets are imbalanced (13588 Hemoglobin-10 Hemerythrin), Accuracy is not enough to evaluate our proposed model. Hence,we plotted the area under the ROC (Receiver Operating Characteristic) curve, or AUC and the precision-recall curve for both cases.(Figure 2-Figure 3) show the multiclass classification case and (Figure 4-Figure 8) show the binary classification case, respectively.

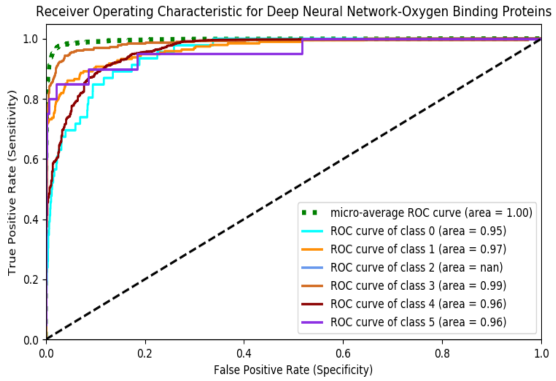

Figure 2. Plot of Receiver Operating Characteristic (ROC) of oxygen binding proteins multiclass classification

Figure 2 shows model performance measured by a ROC plot of Sensitivity (the true positive rate) on the Y axis and Specificity (the false positive rate) on X axis; from which we obtained a micro-average of 1.00. Class 2 (Hemerythrin) run into NaN value because it is totally misclassified to other classes (contains 10 sequences only). Prediction is most accurate for class 3 (Hemocyanin), which obtained an AUC of 0.99,while the other classes had AUCs of 0.95, 0.97, 0.96, and 0.96, respectively.

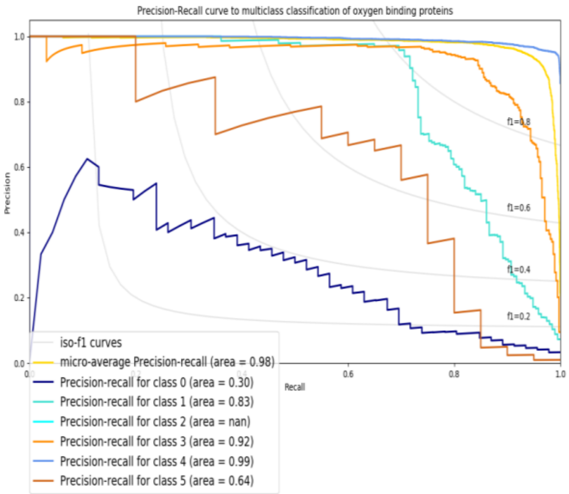

Figure 3. Plot of Recall-Precision curve and iso-f1 curves of oxygen binding proteins to multiclass classification of oxygen binding proteins

Figure 3 illustrates plot of recall-precision curve and iso-f1 curves, which give a micro average of 0.98. We have obtained a precision- recall of 0.30, 0.83, 0.92, 0.99, and 0.64 for Erythrocruorin, Myoglobin, Hemocyanin, Hemoglobin, and Leghemoglobin, respectively.

Table 2. Experimental results of binary classification of each class of oxygen binding proteins

DATA	Neural Network		Proposed model		Oxypred
	ACC	LOSS	ACC	LOSS	ACC
ERYTHROCRUORIN	98.87%	0.0341	98.86%	0.0295	95.8%
MYOGLOBIN	98.48%	0.0488	98.83%	0.0339	96.0%
HEMERYTHRIN	99.96%	7.9259e-04	99.99%	1.902e-04	97.5%
HEMOCYANIN	98.20%	0.0586	98.72%	0.0415	97.5%
HEMOGLOBIN	94.19%	0.1628	95.74%	0.1170	96.9%
LEGHEMOGLOBIN	99.84%	0.0076	99.91%	0.0083	99.4%

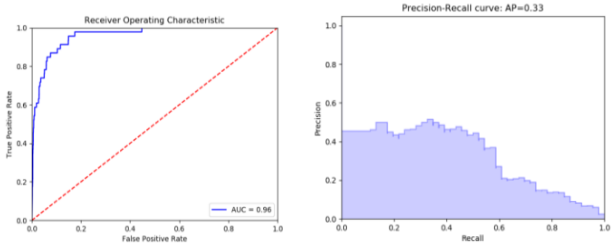

Figure 4. Erythrocruorin

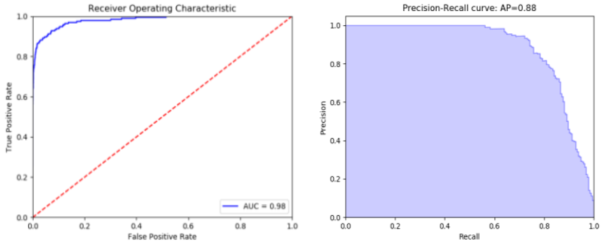

Figure 5. Myoglobin

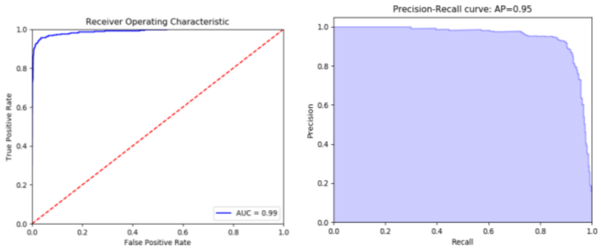

Figure 6. Hemocyanin

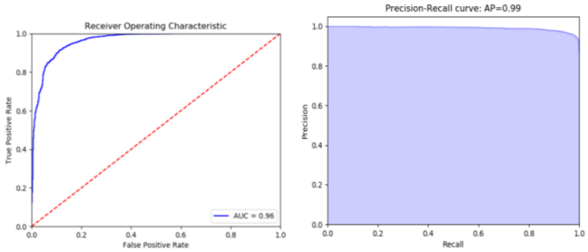

Figure 7. Hemoglobin

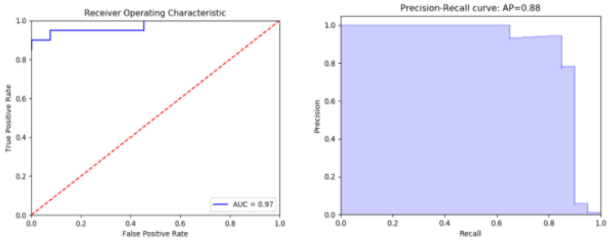

Figure 8. Leghemoglobin

Figure 4-Figure 8 show plot of ROC curves (AUC) and precision-recall curves (PR) for binary classification case of erythrocruorin, myoglobin, hemocyanin, hemoglobin and leghemoglobin, respectively. We have obtained: AUC= 0.96, PR= 0.33, AUC= 0.98, PR= 0.88, AUC= 0.99, PR= 0.95, AUC= 0.96, PR= 0.99, AUC= 0.97, PR= 0.88, respectively.

5. CONCLUSION

In this paper, we proposed an approach for classification of oxygen binding proteins using deep neural network, with the aim to predict function of proteins. We obtained very encouraging and promising results by training and validating the proposed models. In this study, we used the amino acid composition of proteins. As ongoing work, we intend to classify oxygen binding proteins using other residue composition such as: dipeptide composition which is a protein represented by 400 attributes. Indeed, the fact that, these compositions provide more information than amino acid composition would help to improve the accuracy of the classification.

6. REFERENCES

[1] Wright, P. C., Noirel, J., Ow, S. Y., and Fazeli, A. A review of current proteomics technologies with a survey on their widespread use in reproductive biology investigations. *Theriogenology*. 77(4): 738-765.e52 (2012)

[2] Front Matter: Defining the mandate of proteomics in the post-genomics era. Workshop report, the National Academy of Sciences (2002)

[3] Wang, S., Qu, M., and Peng, J: PROSNET: Integrating homology with molecular networks for protein function prediction. *Pac SympBiocomput*. 22, 27-38 (2017)

[4] Cao, R. and Cheng, J. Integrated protein function prediction by mining associations, sequences, and protein-protein and gene-gene interaction networks. *Methods*. 93, 84-91 (2016)

[5] Mousumi Debnath, Godavarthi B. K. S. Prasad, Prakash S. Bisen: Molecular diagnostics: Promises and possibilities: Omics Technology. pp. 11-31. Springer. Dordrech Heidelberg London (2010)

[6] Michele Magrane, UniProt Consortium: UniProt knowledgebase: a hub of integrated protein data. Database (Oxford) (2011)

[7] Gaurav, P., Vipin, K., Michael, S. Computational approaches for protein function prediction: a survey. technical report, University of Minnesota (2006)

[8] *Nucleic Acids Res*. The InterPro protein families database: the classification resource after 15 years. Database issue. 43, D213-D221 (2014)

[9] Sequence analysis: InterProScan 5: genome-scale protein function classification. Bioinformatics. 30, 1236-1240 (2014)

[10] Lawrence A. Kelley, Stefans Mezulis, Christopher M Yates, Mark N Wass, Michael J E Sternberg: The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols. 10, 845-858 (2015)

[11] Jiang et al: An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biology. 17: 184 (2016)

[12] Choi, Y. W. and Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 31, 2745-2747 (2015)

[13] Nikhil, B. Fundamentals of deep learning: Designing next-generation machine intelligence algorithms. O'Reilly Media, Beijing, Boston, Farnham, Sebastopol, Tokyo (2017)

[14] Cun, Y. L., Bengio, Y. S., and Geoffrey, H. Deep learning. *Nature*. 521, 436-444 (2015)

[15] Min, S., Lee, B., Yoon, S. Deep learning in bioinformatics. *Brief Bioinform*. 18(5), 851-869 (2017)

[16] Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J.Predicting the sequence specificities of DNA and RNA binding proteins by deep learning. *Nat Biotechnol*. 33(8), 831-838 (2015)

[17] Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., and Zeng, J. A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res*. 44(4), e32 (2016)

[18] Wang, S, Weng, S., Ma, J., and Tang, Q. DeepCNF-D: Predicting protein order/disorder regions by weighted deep convolutional neural fields. *Int J Mol Sci*. 16(8), 17315-17330 (2015)

[19] Costa-Paiva, E. M., Schrago, C. G., and Halanych, K. M. Broad phylogenetic occurrence of the oxygen-binding Hemerythrin in Bilaterians. *Genome BiolEvol*. 9(10), 2580-2591 (2017)

[20] Muthukrishnan, S., Garg, A., G.P.S. Raghava: Oxypred: Prediction and classification of oxygen-binding proteins. *Genomics Proteomics Bioinformatics*. 5(3-4), 250-252 (2007)

[21] Matt, C. Classification of oxygen binding proteins using Random Forest Machine Learning. http://rpubs.com/oaxacamatt/Random_Forest_Oxygen_Binders (2017)

[22] Decker, H. andTerwilliger, N. Cops and robbers: putative evolution of copper oxygen-binding proteins. [J]*ExpBiol*. 203(Pt12), 1777-1782 (2000)

[23] Cinzia, V. et al: Structure, function and molecular adaptations of haemoglobins of the polar cartilaginous fish Bathyrajaeatonii and Raja hyperborea. Biochem J. 389(Pt2), 297-306 (2005)

[24] Struttmann, T., Scheerer, A., Prince, T. S., and Goldstein L. A. Unintentional carbon monoxide poisoning from an unlikely source. J Am Board FamPract. 11(6), 481-484 (1998)

[25] Senan, J. Y., Vivian Irene Ravn Berg, Asimahmad, Donald Doll: Hemoglobin Titusville: a rare low oxygen affinity hemoglobinopathy. Clin Case Rep. 5(6), 1011-1012 (2017)

[26] Najafabadi et al: Deep learning applications and challenges in big data analytics. *Journal of Big Data*. 2:1 (2015)

[27] Liu, W. et al: A survey of deep neural network architectures and their applications. *Neurocomputing*. 234, 11-26 (2017)

[28] Juergen, S. Deep learning in neural networks: An overview. *Neural Networks.* 61, 85-117 (2015)

[29] Mariette, A. and Rahul, K. Deep Neural Networks. In: Efficient learning Machines. pp.127-147. SpringerLink, Apress, Berkeley, CA (2015)

[30] Nitish, S., Geoffrey, H. et al: Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*. 15, 1929-1958 (2014)

[31] Cybenko, G. Approximation by superpositions of a Sigmoidal function. Math. Control Signals Systems. 2, 303-314 (1989)

[32] Xiao, N., Cao, D. S., Zhu, M. F., and Xu, Q. S.protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics. 31(11), 1857-1859 (2015)

[33] Diederic, P. K., and Jimmy, B. Adam: A method for stochastic optimization. In*the 3rd International Conference for Learning Representations,* San Diego (2015)