UNIVERSITY OF ILLINOIS AT CHICAGO
CS559 - NEURAL NETWORKS - FALL 2015

Project Report

M.S. student *Umberto Di Fabrizio*

***Abstract*— The protein superfamily classification problem, which consists of determining the superfamily membership of a given unknown protein sequence, is very important for a biologist. In this work, the classification task is tackled using several ANN (Backprop, CNN, LAMSTAR) which are compared with respect to their performances. The goal is to predict the family of novel protein sequences based on the sequence only. The results obtained are comparable to the state of the art and the methods can be further investigated to be adapted to a larger dataset of families.**

I. INTRODUCTION

Bioinformatics has been growing in the last three decades[2] given the huge amount of biological data that is continuously gathered, mainly about DNA, RNA and proteins. The volume of data generated from project such as The Human Genome Project[3](1990-2003) has strengthen the collaboration between the computer scientist community and the biologist one.
One of the most challenging problem is to classify protein accordingly to their function or by their family or superfamily. Protein sequences are composed by an unique sequence of 20 amino acids which determines the protein function. They carry out fundamental roles for the cells functions (basically represent the blueprint of the cell), infact they determine the shape and the structure of the cell.
Each protein encode a certain function which depends on its structure and amino acids sequence but can only be completely understood with experiments. Those experiments are costly and slow thus they cannot keep pace with the amount of information available and which needs to be annotated.
The challenge that has been tackled is to classify proteins into functional or structural existing superfamilies so that the annotation process can be automated.
A protein superfamily is a set of proteins for which common ancestry can be inferred so they possed sequence or structural homology. In a superfamily classification, an unlabeled protein sequence may belong to any of the superfamily from a set of known superfamilies. This classification is enormously useful because similar protein sequences exhibit almost the same biological structure and function, more importantly one of the main reason is treating and preventing genetic disease as well as drug discovery, prediction of molecular function and medical diagnosis.

II. BACKGROUND

Several methods have been investigated in order to solve the superfamily classification problem: determining the superfamily membership of a given unknown protein sequence. BLAST (Basic Local Alignment Search Tool) [4] is a tool that uses direct modelling, performing a search of homologie between sequences. This software ex-

plores the local alignment in pairs to measure the similarity between sequences. The classification is done based on the alignment which had the greatest punctuation.

Another method that uses direct modelling is the HMM Hidden Markov Models that is widely used for probabilistic modelling of family of proteins. It uses probabilistic values to score how much an unkown protein belongs to a given family.

A Fuzzy ARTMAP model, a machine learning method has been proposed used to classify the protein sequence[8].

The use of Neural Networks to tackle this problem has been successfully presented in the work by C.H.Wu at al.[5][6][7] and an introductory survey of neural networks applied to genome problems can be found in the book by C.H.Wu and Jerry W. McLarty[9]. The book explains the basic idea of neural networks presenting the different kinds of network that can be used and gives meaningful example on how to use them.

The process of protein classification using ANN can be divided in two parts: (i) pre-processing: protein sequence encoding; (ii) processing the protein classification using the ANN.

The first step is very challenging, the reason is that protein sequences have different lengths and can even be very long (˜1000) whilst the neural network has a fixed amount of inputs and can hardly handle missing inputs. The methods of encoding can be basically dived in two types[9]: direct or indirect. The direct encoding basically translate each amino acid into a binary vector[1] accordingly to the one-hot encoding, this means that given that we have 20 possible amino acids then each of them will be represented by a vector of 20 bits with only one position at '1' and all the rest with '0'. Of course this kind of encoding is not feasible in the common case because a protein sequence with 300 amino acid will be translated with 20*300=6000 binary inputs.

The indirect encoding tries to extract useful features from the protein sequence in order to give meaningful inputs to the neural networks, one example is usually the n-gram hashing method[5]. This ngram method computes residue frequencies, the basic idead is that if we have a sequence *Seq*='ACACTGAC' then the possible bi-gram are 'AA', 'AC', 'AT', 'AG', 'CC', 'CA, 'CT', 'CG', 'TT', 'TA, 'TC', 'TG', 'GG', 'GA', 'GT', 'GC', and for each of them the occurrences are counted, usually the approach involves also the mapping of the original sequence to a smaller alphabet. A summary table on the encoding pros and cons is shown in Table I

[1]other techniques will be presented later

TABLE I

COMPARISON BETWEEN DIRECT AND INDIRECT PROTEIN SEQUENCE ENCODING

	Pros	Cons
Direct	Keeps the sequence order	Encoding is too large
Indirect	Independent of length of the biosequence	It is hard to find meaningful feature, this requires domain knowledge. The order of the sequence is not mantained

III. OUTLINE

The idea is to exploit the power of the Deep Learning Neural Network in order to classify the protein in its family.
In section DATA COLLECTION it is explained where the data was collected from and which tools have been used, in section DATA ENCODING the encoding method to transform the sequence to numbers for the ANN is discussed. Sections BACKPROPAGATION CONVOLUTIONAL NEURAL NETWORK, LAMSTAR discuss the design and structure of the ANNs employed to solve the problem, together with the choices that were made to optimize the execution time and the accuracy of the nets. Finally section RESULTS presents the results and in CONCLUSION the contribution of the work are summarized and compared to the state of the art techniques.

IV. DATA COLLECTION

The objective is to obtain for each superfamily or family the list of all the proteins that belong to that family and their complete protein sequence.
This step is particularly tricky because there is not a database of protein sequences and families, or better there are several but they all have to be queried manually through a web interface. For this reason the data is scraped from the web sites (http://prosite.expasy.org/, http://www.uniprot.org/uniprot/) with the use of the Kimono platform[10] and Rscripts. Automatizing is particularly useful in case any new superfamily data has to be collected.
With this process 4 families have been chosen: *ADH_SHORT*, *ABC_TRANSPORTER_2*, *G_TR_2*, *globin*, and for each of them 600 protein sequences have been collected. The overall dataset is thus made by 2400 protein sequences and it is divided between training(90%) and testing(10%).

V. DATA ENCODING

During this phase the sequences of proteines will be analyzed and encoded through the indirect method. Different solution will be explored as regard the n-gram value (2,3) and the possibility to include any other valuable information. The objective is to extract as many information as possible from the protein without creating too many inputs for the neural network.

VI. DESIGN OF THE NN

The input of the neural network will be the encoded protein sequence (the estimated size will be around 500 inputs), the output is the superfamily of the protein or 'Other' if the protein does not belong to any of the available families (possibly 3 or 4). The neural network two design are still uncertain, the basic approach will be to first create a backpropagation network in order to have a classifier to use as baseline and to test the underlining assumptions and the encoding method.
Then a Convolutional Neural Network will be designed, during the step the input format will be re-formatted from the vector form to the matrix form that CNN handles and the choice of the convolution filters and pooling method will be made by comparing different solution.
At this point, if I will still have time the LAMSTAR will be designed and used although given the amount of time available this step may be omitted.

VII. BACKPROPAGATION

VIII. CONVOLUTIONAL NEURAL NETWORK

IX. LAMSTAR

X. RESULTS

XI. CONCLUSION

REFERENCES

[1] D.Graupe, Principles of Artificial Neural Networks, 3nd ed., Advanced Series in Circuits and System-Vol.7

[2] Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics,Muhammad Javed Iqbal et al., 2014

[3] https://www.genome.gov/12011238

[4] Basic local alignment tool,S. Altschul et al., 1990

[5] Protein classification artificial neural system, C. H. Wu et al, 1992

[6] Neural Networks for Full-Scale Protein Sequence, C. H. Wu et al, 1995

[7] Motif identification neural design for rapid and sensitive protein family search, C. H. Wu et al,1996

[8] Multi-class Protein Sequence Classification Using Fuzzy ARTMAP,Shakir Mohamed et al.,2006

[9] Neural Networks and Genome Informatics

[10] https://www.kimonolabs.com/