BABEŞ-BOLYAI UNIVERSITY
FACULTY OF MATHEMATICS AND COMPUTER
SCIENCE

Machine Learning Models for additional insights in Proteomics

Abstract Of PhD Thesis

PhD student: Silvana Albert
Scientific supervisor: Prof. Dr. Gabriela Czibula

2019

Keywords: machine learning, proteomics, protein folding, clustering, self organizing maps, unsupervised learning, data mining, open source software

Contents of the abstract

Contents of the thesis

List of publications

All rankings are listed according to the 2014 classification of journals[1] and conferences[2] in Computer Science

1. [AC19] **Silvana Albert**, Gabriela Czibula. $ProteinA$: An approach for analyzing and visualizing protein conformational transitions using fuzzy and hard clustering techniques, 12th International Conference on Knowledge Science, Engineering and Management (KSEM 2019), to be published. (**ISI Proceedings**)

   Rank B, 4 points.

2. [ATC18] **Silvana Albert**, Gabriela Czibula, Mihai Teletin. Analysing Protein Data Using Unsupervised Learning Techniques. *International Journal of Innovative Computing, Information and Control*, Volume 14, Number 3, June 2018 14(3):861–880, 2018. (**indexed Scopus**)

   Rank B, 4 points.

3. [TCAB18] Mihai Teletin, Gabriela Czibula, **Silvana Albert**, Maria-Iuliana Bocicor Using unsupervised learning methods for enhancing protein structure insight. *International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES 2018)*, Volume 126, 19-28, 2018 (**ISI Proceedings**)

   Rank B, 2 points.

4. [TCB$^+$18] Mihai Teletin, Gabriela Czibula, Maria-Iuliana Bocicor, **Silvana Albert**, and Alessandro Pandini. *Deep autoencoders for additional insight into protein dynamics*. International Conference on Artificial Neural Networks (ICANN), Rhodes, Greece, LNCS, volume 11140, pp. 78-89, 2018.

   Rank B, 1.33 points.

5. [ACT18] **Silvana Albert**, Gabriela Czibula, Mihai Teletin. Analyzing the impact of protein representation on mining structural patterns from protein data. *IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI 2018)* , Volume 12, 533-538, 2018 (**ISI Proceedings**)

[1] http://informatica-universitaria.ro/getpfile/16/CSafisat2.pdf;http://hfpop.ro/standarde/doctorat/2014-jurnale.pdf

[2] http://informatica-universitaria.ro/getpfile/16/CORE2013_Exported.xlsx;\http://hfpop.ro/standarde/doctorat/2014-conferinte.xlsx

Rank C, 2 points.

6. [Alb17b] **Silvana Albert**. Visualizing mutation occurrence using Big Data. *Proceeding Collection on the workshop NETTAB 2017 "Methods, tools and platforms for Personalized Medicine in the Big Data Era"*, 2017. (**indexed DBLP**)

   Rank D, 1 point.

7. [Alb17a] **Silvana Albert**. A Big Data Approach in Mutation Analysis and Prediction. *Studia Universitatis Babes-Bolyai Series Informatica*, 62(1):75–89, 2017. (**indexed Mathematical Reviews**)

   Rank D, 1 point.

8. [BPC^{+}17] Maria Iuliana Bocicor, Alessandro Pandini, Gabriela Czibula, **Silvana Albert**, Mihai Teletin. Using Computational Intelligence Models for Additional Insight into Protein Structure. *Studia Universitatis Babes-Bolyai Series Informatica*, 62(1):107–119, 2017. (**indexed Mathematical Reviews**) **Rank D, 0 point.**

Publications score: **15.33 points**.

Introduction

This Ph.D. thesis is the result of my research conducted since 2016 under the supervision of Prof. Dr. Gabriela Czibula. It consists mainly of developing Machine Learning (ML) models and applying them on biological data along with the analysis of the results with the bigger goal of discovering underlining patterns and motifs in Genomic and Proteomic data.

In this thesis we will employ various *ML* techniques for better insight into biological processes. Protein data space in particular will be explored with the help of different algorithms such as clustering, self organizing maps, long short-term memory networks, principal component analysis and autoencoders. Uncovering patterns in protein conformations helps to better understand how proteins change and evolve. It is already common knowledge that protein shape dictates the biological function, so any anomaly during the folding process can lead to a malformed, defective protein that, in time, can cause illness or even death for the organism. That is why we need to investigate computational methods and develop solutions that can help the researchers to gain a better understanding into the mysterious world of protein folding. The author's increased interest, passion and fascination for the biological processes that create and ensure life has been one of the driving points for choosing this theme. Both the importance and the complexity of the problem motivates us to explore the utility of *machine learning* models and methods for the analyzing and detecting the conformational changes in proteins. Our work contains various applications of machine learning algorithms on biological data, along with the analysis of the results within the bigger goal of discovering underlining patterns and motifs in Proteomic data. We are also aiming our studies on enzymes that have interesting properties when it comes to the biodegradation of waste materials. On the *Machine Learning* side, there has also been unprecedented growth in interest and popularity. During the *protein folding* process, the protein undergoes changes from one conformation to another and it is influenced both internally by its initial structure and composition, and externally by other factors such as temperatures and nearby elements that interact with it. Understanding the dynamics of the protein leads to creating better targeted medicine, slow down the aging process and improve life quality all together.

With the main target of understanding the importance of the protein folding problem and uncovering hidden patterns in protein data, we have analyzed protein conformational transitions with unsupervised learning tools, by applying different types of hard and fuzzy clustering algorithms and comparing the results. The RSA values have been studied and their relevancy has been put to test when it comes to proteins' internal transitions prediction. On these values, we have constructed multiple case studies in the realm of unsupervised learning. For better visualization, we have employed *Principal component analysis* and we were able to see how proteins evolve through their RSA values from one conformation to another.

Deep learning was used, specifically Long Short Term Memory networks for exploring protein conformational transitions while employing some other protein variable: the width of the binding cleft that helps gaining more knowledge. We have proposed two open source software solutions which allow domain experts to easily replicate our experiments leading to a better collaboration. Our approaches have been published in journals and conferences: [Alb17a, BPC+17, ATC18, TCAB18, ACT18, TCB+18, AC19, Alb17b].

Chapter 1: Analysing protein data using unsupervised learning

In this chapter we investigate the usefulness of *machine learning* models and methods for analyzing and detecting the conformational changes in proteins. The presentation from this chapter is based on our original papers [BPC^{+}17], [ATC18] and [ACT18].

1.1 A theoretical model for analyzing protein conformational transitions

We tackle the problem of determining conformational transitions in proteins from a different angle and we derive a different formalization for it, starting from a data set of more than 300 proteins and their associated conformations. Our focus is to test if the conformational transitions of the proteins provide useful information regarding their three-dimensional structure and if an *unsupervised learning* model is able to capture this type of biological relationships between the proteins. Our chosen *unsupervised learning* model is a *self-organizing map* (SOM) because it is a considered a strong tool from the *data mining* domain which aids in visualizing high-dimensional data. Using a *data mining* experiment, we show that the information obtained through analyzing proteins conformational transitions is able to capture the relationships between related proteins, relations which are confirmed from a biological perspective.

1.2 Protein Data Analysis using Self Organizing Maps and Principal Component Analysis

Proteins have essential roles in the biological processes of living organisms by contributing to maintaining cellular environments. Understanding the conformational transitions of proteins may help identifying situations when incorrect folding or mutations can occur and thus, it may contribute to inhibit possible uncontrolled and undesired behaviour. The structural similarity between proteins is unsupervisedly uncovered using crisp and fuzzy *self-organizing maps*, based on proteins conformational transitions. We propose a method for modelling a protein based on its conformational transitions and we also examine how feature selection impacts the performance of the proposed models.

1.3 Analyzing the impact of protein representation on mining structural patterns from protein data

We are performing a study on how different protein representations impact the process of mining relevant patterns from protein related data. Two representations are used for the proteins, one using the structural alphabet and the second using the relative solvent accessibility values of the amino acids from the proteins' primary structure. Using these representations, two case studies are performed to emphasize the effectiveness of using the proposed protein representations to unsupervisedly learn structural patterns from on a protein data set. The RSA values seem to be very relevant in representing the conformations of proteins.

Chapter 2 - Clustering approaches for protein data analysis and visualization

In this chapter we employ *clustering* as an unsupervised classification method in order to study the relevance of the residues' *relative solvent accessibility* RSA values to analyze protein internal transitions. We provide two approaches involving hard and soft clustering and we are comparing the results [TCAB18], [AC19].

2.1 Approach 1: A theoretical model enhancing proteins structure insight

We design a study directed towards investigating how proteins conformational transitions evolve in time, with the goal of broadening the knowledge into internal protein dynamics. It employs clustering as an unsupervised learning technique for inspecting the importance of RSA values in decoding protein internal transitions. For each of the proteins, we have 10000 conformations available, along with their associated RSA values. They were obtained through molecular dynamics simulations, a process that is considered fairly expensive from the required resources point of view. The 10000 conformations are consecutive and they can be viewed as a time lapse in the formation stage of the protein. Our assumption is that the changes a protein undergoes from a conformation to the next one are very small (if there are such changes) as a deduction from the biological perspective, meaning that close consecutive conformations are fairly similar. That is why we are performing *Euclidian distance* for computing the dissimilitude between two successive conformational transitions. The data sets remain unlabelled throughout the whole process, unsupervised learning being our strategy. As the protein undergoes conformational changes, certain parts of its structure are subjected to minor modifications, which are reflected in the positions of the amino acid residues and consequently, in their RSA values. Thus, consecutive conformations are fairly similar from the perspective of their considered representations (RSA values). This is also reflected in our obtained clustering results. One also observes that the proteins are structurally similar and because of that, it is anticipated to have a high degree of similarity on their represented shape and dynamics. We are highlighting the potential of clustering models to accurately model biological transitions, from conformations represented via RSA values. From a biological viewpoint, in a transition between two successive conformations, the protein might remain unchanged, or slight modifications can appear in certain parts of it.

2.2 Approach 2: A model for analyzing protein conformational transitions using fuzzy and hard clustering techniques

We are conducting several experiments on two protein data sets with the goal to empirically show, using fuzzy c-means and Birch clustering methods, that RSA values slowly change when a protein goes through conformational changes. The end goal is proving that consecutive conformations are closer and the protein evolves linearly. The two proteins used in our experiments are: 6EQE - a “High resolution crystal structure of a polyethylene terephthalate degrading hydrolase from Ideonella sakaiensis” and 4CG1 - a “Thermostable polyethylene terephthalate degrading hydrolase from Thermobifida fusca” [WOT+14]. Both proteins are being investigated for their roles in PET degradation and the fact that they have a lot in common leads us to expect similar results when applying unsupervised algorithms. We note that the fuzzy operation does not improve as expected. This is possibly due to the fact that the input data is not necessary suitable for data fuzzification, considering the chosen representation. Future work will be carried out in this direction, for identifying other enhanced representations more for the fuzzy perspective because it does not improve the effectiveness of the clustering process. More experiments will be further carried out in this direction.

Chapter 3 - Deep learning approaches for protein data analysis

This Chapter is oriented around *Neural Networks* and how they can be employed for capturing hidden patterns on protein data. We explore *Deep Autoencoder Neural Networks* and their ability in capturing aspects related to a protein's structure. The presentation represents a sequel of our work presented in [TCB+18].

3.1 Using Autoencoder's for uncovering protein dynamics

Because molecular dynamics approaches are so expensive, data dimensional representation reduction is our focus as well. Denoising sparse autoencoders are trained on each protein data set with the main purpose of reducing the dimensionality of the datasets and aid visualization. For validating our results, we are computing similarities in the original data set and finally we will compute the similarities of the two-dimensional data outputted by the autoencoder.

3.2 Predicting the width of the binding cleft by employing Long short-term memory networks

Our intuition is that taking into account additional information about each protein conformation would help gain more insight into the internal working of a protein during the folding process. Because we are using a 2D representation of a 3D object, we are losing valuable information about proteins. Including other values that describe each conformation is our attempt to regain some of it. That is why we exploit a property called the *width of the binding cleft* (also known as *active-site cleft*), that characterizes both *cutinases* and *PETase*. Studies were made to prove that by narrowing down this property for PETase to look more like the Cutinase's active site cleft would lead to better PET Degradation [AAD+18]. We are employing a Long Short Term Memory network for learning important information about a protein's transition and mainly how each conformation can be evaluated when it comes to open/closed states based on the width of the binding cleft. Using multiple train/test data splits in our learning and testing, we attempt to enhance the model's performance on unseen data, due to the fact that more models are being trained. We are aware nonetheless that the multi train-test split approach has a limitation: each of the training models remains fixed as it is evaluated in the test set. After training our LSTM Network, we are able to predict new values for the above described d property, based on the *Angles representation* of the protein conformations. We conclude that it is safe to assume that LSTMs are suitable for the presented problem and the developed computational model is able to predict future values for the investigated property. From a biological perspective, this is useful because, by making small alterations to the *PETase* proteins, thus adjusting it's binding cleft, it can perform better when it comes to PET Degradation [AAD+18]. Similar models could be used by scientists as they alter the composition of proteins, to emulate the *Open/Closed* state. Future work includes a classification LSTM network that could easily determine if a new instance (protein conformation in our case) has the state Open or Closed.

Chapter 4 - Software development

We are introducing in this Chapter two software solutions which represent our original work published in [Alb17a], [Alb17b] and [AC19]. We attempt through these applications to contribute to the open-source community and to allow domain experts to easily replicate our experiments leading to a better collaboration.

4.1 ProteinA: A software solution for visualizing clustered protein conformational transitions

The software presented in this section has been created in order to allow any user to try various combinations or parameters and independently analyze the results. It has been first introduced in our original paper [AC19]. We are proposing the tool *ProteinA* for capturing protein conformational transitions by clustering. It is a web application allowing users to start custom analyses and download the results. A clustering analysis takes about 5 minutes, however the idea behind the software is to allow more complex processing and delivering the results when ready. The solution it is publicly running at [Alb19][1]. The code is available on Github at [Alb18b].[2] Another option for easily running it on a local machine is by accessing the public docker image at [Alb18a]. [3]

4.2 Novel software for visualizing genetic mutations

The presented solution helps by aggregating all the precedent mutations correlated with a series of external factors. The doctor is able to narrow it down to a reasonable number of possibilities based on the cases that were already solved. This leads to making an informed decision of which mutations to test for. After successfully determining the current case, the specialist will introduce it to the global database, this way, helping future doctors. As a proof of concept, it demonstrates the huge role that Big Data has in genetic mutations aggregation and it can be considered a starting point for similar solutions that aim to continuously innovate genetics.

[1] Protein clustering online http://proteinclusters.online/proteins.

[2] Protein clustering web application https://github.com/albusilvana/proteinclusteringwebapp.

[3] Protein clustering docker image on Docker hub https://hub.docker.com/r/salbert/proteinclustering.

Bibliography

[AAD$^+$18] Harry P. Austin, Mark D. Allen, Bryon S. Donohoe, Nicholas A. Rorrer, Fiona L. Kearns, Rodrigo L. Silveira, Benjamin C. Pollard, Graham Dominick, Ramona Duman, Kamel El Omari, Vitaliy Mykhaylyk, Armin Wagner, William E. Michener, Antonella Amore, Munir S. Skaf, Michael F. Crowley, Alan W. Thorne, Christopher W. Johnson, H. Lee Woodcock, John E. McGeehan, and Gregg T. Beckham. Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proceedings of the National Academy of Sciences*, 115(19):E4350–E4357, April 2018.

[AC19] S. Albert and G. Czibula. *proteina*: An approach for analyzing and visualizing protein conformational transitions using fuzzy and hard clustering techniques. In *12th International Conference on Knowledge Science, Engineering and Management (KSEM 2019)*. Springer, May 2019.

[ACT18] S. Albert, G. Czibula, and M. Teletin. Analyzing the impact of protein representation on mining structural patterns from protein data. In *IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI 2018)*, page To be published, 2018.

[Alb17a] S. Albert. A big data approach in mutation analysis and prediction. *Studia Universitatis Babes-Bolyai Series Informatica*, 62:75–89, 2017.

[Alb17b] S. Albert. Visualizing mutation occurrence using big data. In *ProceedingCollection on the workshop NETTAB 2017 Methods, tools platforms for PersonalizedMedicine in the Big Data Era, 2017*, 2017.

[Alb18a] Silvana Albert. Protein clustering docker image. https://hub.docker.com/r/salbert/proteinclustering, 2018.

[Alb18b] Silvana Albert. Protein clustering git repository. https://github.com/albusilvana/proteinclusteringwebapp, 2018.

[Alb19] Silvana Albert. Protein clustering analysis, 2019.

[ATC18] S. Albert, M. Teletin, and G. Czibula. Analysing protein data using unsupervised learning techniques. *International Journal of Innovative Computing, Information and Control*, page 861–880, 2018.

[BPC$^+$17] M.I. Bocicor, A. Pandini, G. Czibula, S. Albert, and M. Teletin. Using computational intelligence models for additional insight into protein structure. *Studia Universitatis Babes-Bolyai Series Informatica*, 62:107–119, 2017.

[TCAB18] M. Teletin, G. Czibula, S. Albert, and I. Bocicor. Using un-supervised learning methods for enhancing protein structure insight. In *InternationalConference on Knowledge Based and Intelligent Information and EngineeringSystems*, page To be published, 2018.

[TCB$^+$18] M. Teletin, G. Czibula, M.I. Bocicor, S. Albert, and A. Pandini. Deep autoencoders for additional insight into protein dynamics. In *22nd International Conference on Knowledge-Based and Intelligent Information Engineering Systems*, pages 79–89. Springer, 2018.

[WOT$^+$14] Ren Wei, Thorsten Oeser, Johannes Then, Christina G. Föllner, Wolfgang Zimmermann, and Norbert Sträter. Structural and functional studies on a thermostable polyethylene terephthalate degrading hydrolase from thermobifida fusca. *Applied Microbiology and Biotechnology*, 98:7815–7823, 2014.