A Bioinformatics Pipeline for Evaluating Protein Misfolding Impact on the Tertiary Structure in Alzheimer's Disease

Antigoni Avramouli, Eleftheria Polychronidou, Panayiotis Vlamos
BiHELab – Bioinformatics and Human Electrophysiology Lab
Department of Informatics of Ionian University
Corfu, Greece
e-mail: c15avra@ionio.gr, c13poly@ionio.gr, vlamos@ionio.gr

***Abstract*— Alzheimer disease (AD) is the most common cause of neurodegenerative disorder in the elderly individuals. To support the biomarker research on Alzheimer's Disease progression, this study describes a bioinformatics pipeline for the evaluation of the mutations impact on the tertiary structure of AD causative genes.**

Keywords: protein structure; protein misfolding; machine learning; Alzheimer's Disease.

I. INTRODUCTION

Proteins are large, complex biomolecules made up of amino acids. Proteins play a significant role in almost all biological processes. The functional properties of proteins rely upon their three-dimensional structures. The three-dimensional structure arises because the polypeptide chains fold to produce (starting from linear sequences) compact and independent structural regions with specific structures. Predicting the three-dimensional structure of proteins by their amino acid sequence contributes to understanding their biological function. Prediction is not always possible: despite the remarkable efforts of recent years, the problem of folding remains one of the major problems in molecular biology. In addition, proteins that do not get the right configuration can bind abnormally to other biomolecules, as well as form aggregates that are highly toxic to the body [1]. Aggregates are organized into fibrillar structures, a common feature of many neurodegenerative diseases [2].

Alzheimer's Disease (AD), characterised as a protein misfolding disease, is the most common progressive form of dementia [3]. Typical pathological findings are misfolded and aggregated amyloid-β (Aβ) peptides and intracellular neurofibrillary tangles of tau protein. The most well-known predisposing genetic factor for the disease is the presence of the e4 allele of apolipoprotein E (ApoE) [4]. In the e4 allele (frequency 13.7%), the codon 112 has been replaced by arginine. However, the frequency of the e4 allele increases dramatically to ~ 40% in patients with AD. This mutation is associated with a change in the tertiary structure of the protein and the accumulation of β-amyloid in neurons, as well as with the induction of inflammatory responses, while it is the most prone isoform to proteolysis. In this context, changes in the tertiary structure of proteins, which are components of major signaling pathways of AD, could justify the genetic background of this heterogeneous disorder.

In recent years, the correlation of the different tertiary structures of the isoforms of the ApoE gene with the pathogenesis of AD has been studied worldwide [5, 6]. In particular, a study published by the Paralvrez-Marin group in Sweden proposed a computational model of the abnormal interaction of the β-amyloid peptide with the e4 isoform of ApoE, due to the incorrect tertiary structure of the second [7]. However, apart from ApoE-related studies, to date, changes in the tertiary form of proteins due to gene mutations have not yet been investigated in AD. Prior to the discovery of mutations in genes associated with disease onset, no molecular signaling pathways were implicated. Recent genetic studies have identified many candidate genes that are associated with an inherited form of AD. Even if mutations in these genes account for a small proportion of Familial AD (FAD), knowledge of these genes and correlated biochemical cascades will provide several potential targets for treatment of AD and aging-related disorders. Also, the different pathogenetic mechanisms of the disease involve a combination of genetic factors (with different severity for the disease from person to person), indicating that it is essentially a set of disorders with common characteristics rather than a distinct disease.

The present research paper aims to contribute to the reduction of the research gap created by the study of the tertiary structure, to understand the pathogenesis of the disease. In recent years, research interest has focused on identifying all the genetic sites associated with the disease and the different alleles of these genes using high-resolution technologies. In contrast, there is the tertiary form of these mutant proteins, which has not yet been studied in depth. In addition, some of the AD-related proteins have not yet had their crystal structure determined.

Approaches that allow the prediction of three-dimensional structures of proteins through computers are relatively new in the medical sciences [8], but their contribution is increasingly recognized as a tool for characterizing changes in the structure of proteins and detecting rare molecular events. These principles make it easier for us to understand how the protein structure is created, to identify common structural issues, to relate structure and function, but also to see the fundamental relationships between different proteins. Deciphering the mechanisms of the loss of the tertiary structure of a protein is essential for understanding the pathogenesis of diseases, such as AD and essential for explaining neuronal damage during aging.

 ISBN: 978-1-61208-859-4

This pipeline is described by four steps: (a) the evaluation of the online prediction tools and the selection of the most suitable for AD protein structures, (b) the prediction of the mutated structures, (c) the AI/ML classification of the tertiary structures into discrete groups and (d) the evaluation of the pathogenicity of each group to gain evidence for the impact of the mutations and to suggest a characterization for the mutations with unclear etiology. This is an on-going research and thus preliminary results on Presenilin one will be presented here.

II. METHODS

The first step towards the implementation of the pipeline is to collect data from biological databases, to evaluate the existing data and finally to apply machine learning approaches and classify proteins into groups with similar characteristics.

A. Data Consolidation

Here some of the most AD pathogenic mutated alleles will be studied. As many of these mutations affect protein stability, modeled protein structures for the mutant proteins will be compared with the native protein to evaluate stability changes. The genetic loci that will be analysed further through protein 3D structure include APP (Amyloid precursor protein), PSEN1 (Presenilin one), PSEN2 (Presenilin two), CLU (Clusterin), CR1 (Complement receptor 1), PICALM (Phosphatidylinositol binding clathrin assembly protein), BIN1 (Myc box- dependent- interacting protein 1), ABCA7 (ATP binding cassette transporter 7), MS4A (Membrane- spanning 4- domains, subfamily A), EPHA1 (Ephrin type-A receptor 1), CD33 (CD33 antigen), CD2AP (CD2 associated protein), SORL1 (Sortilin-related receptor 1), TPEM2 (Triggering receptor expressed on myeloid cells 2) [9]. These genes are linked to inflammation, oxidative stress, vascular regulation, immune system function, and the function of specific proteases.

Successful mapping of these genes and their association with the onset of the disease has led to the formulation of the amyloid hypothesis [10]. This hypothesis sets as the main pathogenetic mechanism the increased production of β amyloid peptide fragments. Nevertheless, there are cases where the onset of symptoms occurs at a much younger age. In a unique clinical case so far, the onset of the disease occurred in the mid-forties and in some people from the age of thirty. Members of this family had a mutation in the PSEN1 gene (Presenilin 1 E280A) [11]. The mutations related to the proteins were identified through literature and used for the next steps of this pipeline. More particular, so far 69 mutations were identified for APP, 112 for MART, 326 for PSEN1, 68 for PSEN2, and 68 for TREM2.

B. Evaluation of Protein Structures

Since the three-dimensional shape of most of the related proteins is not determined through experimental methodologies, the most established servers were evaluated for predicting the mutated structures and estimate the impact of the mutations to the 3-dimensional structure. A list of the selected methodologies is presented on the Table I below:

TABLE I. LIST OF SELECTED METHODOLOGIES

Methodology	Description	How was used
Uniprot [12]	A comprehensive resource for protein sequence and annotation data	To understand the protein function, and the most related protein structures
PolyPhen-2 [13]	A tool which predicts possible impact of an amino acid substitution on the structure and function	To understand how mutations affect the structure and function of the protein
iTASSER [14]	A hierarchical approach to protein structure prediction and structure-based function annotation	To predict the mutated and unmutated 3D protein structures
PDBeFold [15]	An interactive service that allows you to identify structures that are similar to that of your reference protein	To compare the mutated and unmutated structures on residues level
CATH / Gene3D [16]	A protein family classification methodology	To identify if there is any relationship between mutations impact and protein families

The methodologies are currently used based on the order of the table, to determine the protein structures and understand in detail the impact of the mutations to the proteins. Furthermore, STRING [17] server is used to analyse protein-protein association networks and assess any change that might occur on the mutated protein networks (Figures 1&2).

C. Clustering of protein structures

To analyze further the mutated structures, an established methodology from the field of 3D object recognition was applied [18]. The combination of the above local descriptors was applied to the 3D structures to extract the appropriate features for the comparison.

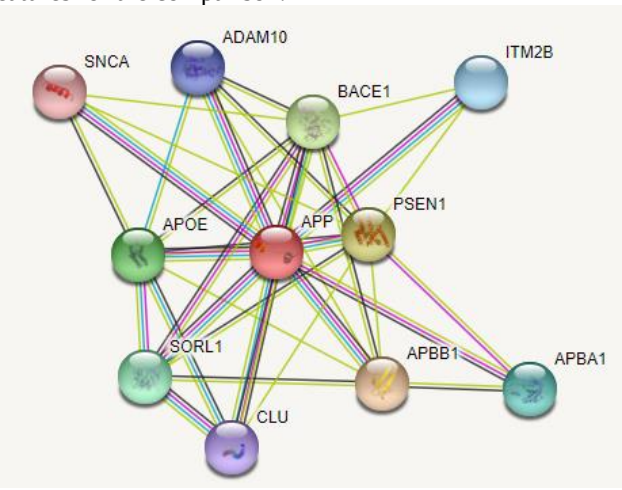

Figure 1. Example of APP network in STRING network analysis.

In order to evaluate the accuracy of clustering using the 3D descriptors, a first round of experiments was conducted, using an annotated dataset. This dataset included every mutated structure while the label of each structure was aligned with the pathogenetic impact of each mutation according to literature data.

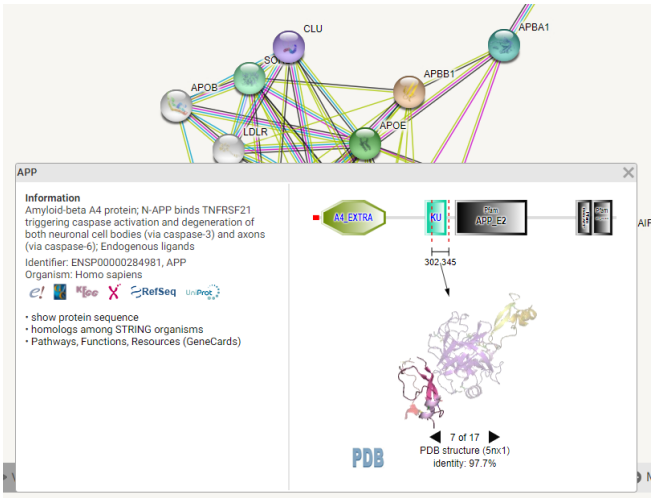

Figure 2. Structural Analysis of the protein. Each protein is mapped to the experimental determined structures (one or more) included in PDB(e).

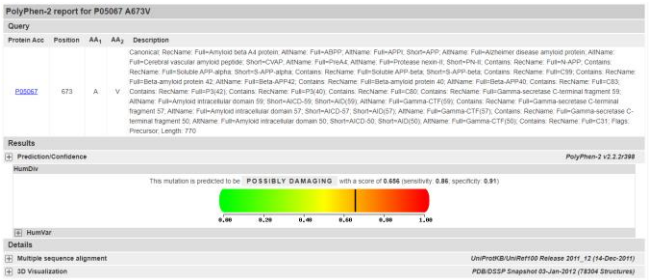

Figure 3. Example of PolyPhen-2 output for the A673V mutation of the APP protein. Percentage of prediction is taken into consideration for the annotation of the clustering output.

The k-medoids, Agglomerative Hierarchical clustering and Density-based spatial clustering of applications with noise (DBSCAN) methods were used to cluster the data using the extracted descriptors [19]. After extracting descriptors from each pair of aligned proteins, the root-mean-square distances (RMSD) between each pair of descriptors is computed, forming a square distance matrix.

In this computational analysis work, preliminary results of our study on PSEN1 mutations are presented and are compared to available clinical data for PSEN1 variants known to cause AD (Figures 3&4). To the best of our knowledge, this is the first study of its kind investigating performing comparative and ab initio prediction of protein structure for mutated forms of PSEN1. The experimental results verify that the use of 3D descriptors can be effectively applied to distinguish structural differences of proteins based on the pathogenic categories of the mutations.

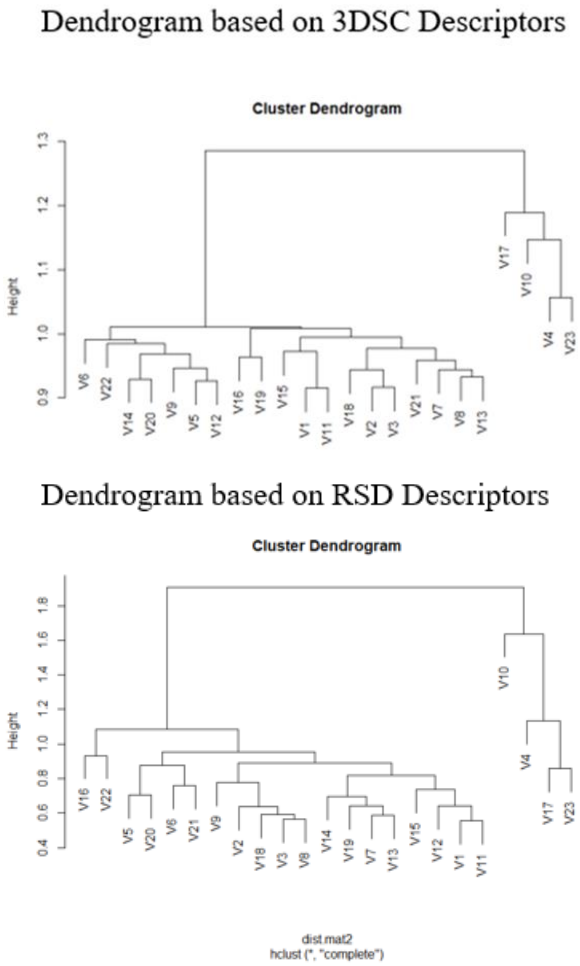

Figure 4. Results of hierarchical clustering are presented for the PSEN1 protein based on the 3DCS and the RSD descriptors type that was applied to each case.

The same process will be repeated for all the other proteins related to the AD progression described in the Data Consolidation section. However, due to the limitations in the prediction time of the online servers, the proof of concept of the PSEN1 is presented here.

III. CONCLUSION

It is known that there is no cure for AD to date. The collective failure of recent clinical trials in the treatment of AD suggests the need for a fuller understanding of the complex biological processes underlying this disease to develop effective, targeted therapeutic approaches. To date, several genetic sites have been identified that are involved in the onset or evolution of AD. Also, AD, like other neurodegenerative diseases, seems to be a biological phenomenon distinct from the phenomenon of normal aging and not an accelerated and pathological version of it. These

 ISBN: 978-1-61208-859-4

data indicate the multiplicity of etiological factors that contribute to the occurrence of AD.

The therapeutic targeting of protein folding has created unique challenges for the discovery and development of new drugs. To achieve this, we must first understand the dynamic nature of the protein species involved and discover the structure and folding of each protein (formation of monomers, oligomers or insoluble aggregates) as well as whether this leads to cell toxicity. To date, our lack of understanding of how proteins interact with other cell proteins and the lack of well-characterized biomarkers that can be used in clinical trials is another bet for the research community.

In the present study, a comprehensive methodology for the analysis of the impact of the AD related proteins is presented. Based on the approach, a combination of well-established online tools can support the prediction of 3D protein structures that have not been determined experimentally yet. Furthermore, the use of Poly-phen2 and CATH can support the identification of evidence of the impact of mutations to the protein structure. Finally, a combination of bioinformatic and object recognition clustering methodology is applied to group the tertiary structures. The annotation of the groups based on the pathogenic characterization of the mutations along with the networks produced by STRING server can reveal evidence on how each mutation affects the protein network.

As mentioned in Section II, the prediction process through online servers consumes significant time and thus a proof of concept is presented here. Since this is an on-going work, the complete analysis will be available as soon as the models are obtained.

ACKNOWLEDGMENT

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning 2014-2020» in the context of the project “Analysis of the tertiary protein structure and correlation of mutations with the clinical characteristics of Alzheimer's disease”, Project no. 5067210.

REFERENCES

[1] C. M. Dobson. “Protein folding and misfolding”. *Nature* 426, pp. 884–890, 2003.

[2] C. Soto and S. Pritzkow. “Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases”. *Nature neuroscience*, 21(10), pp. 1332–1340, 2018.

[3] V. Vingtdeux, N. Sergeant and L. Buee, “Potential contribution of exosomes to the prion-like propagation of lesions in Alzheimer's disease”. *Front Physiol.* 3 pp. 229, 2012.

[4] B. V. Zlokovic, “ Cerebrovascular effects of apolipoprotein E: implications for Alzheimer disease”. *JAMA Neurol.* 70 pp. 440–444, 2013.

[5] V. V. Giau, E. Bagyinszky, S. S. An and S. Y. Kim, “Role of apolipoprotein E in neurodegenerative diseases”. *Neuropsychiatric disease and treatmen*t, 11, pp. 1723–1737, 2015.

[6] P. Huebbe and G. Rimbach, “Evolution of human apolipoprotein E (APOE) isoforms: Gene structure, protein function and interaction with dietary factors”. *Ageing Research Reviews*, 37, pp. 146–161, 2017.

[7] J. Luo, J. D. Maréchal, S. Wärmländer, A. Gräslund and A. Perálvarez-Marín, “In silico analysis of the apolipoprotein E and the amyloid beta peptide interaction: misfolding induced by frustration of the salt bridge network”. PLoS Comput Biol. 5;6(2) pp. e1000663, 2010.

[8] E. Polychronidou, I. Kalamaras, A. Agathangelidis, L. A., Sutton, X. J. Yan, V. Bikos, et. al, “Automated shape-based clustering of 3D immunoglobulin protein structures in chronic lymphocytic leukemia”. *BMC bioinformatic*s 19.14: 414, 2018.

[9] M. Calabrò, C. Rinaldi, G. Santoro, and C. Crisafulli, “ The biological pathways of Alzheimer disease: a review”. *AIMS neuroscience*, 8(1), pp. 86–132, 2020.

[10] D. J. Selkoe and J. Hardy, “The amyloid hypothesis of Alzheimer’s disease at 25 years”. *EMBO Mol Med.* 8(6), pp. 595–608, 2016.

[11] D. Sepulveda-Falla, L. Chavez-Gutierrez, E. Portelius, J. I. Vélez, S. Dujardin, A. Barrera-Ocampo, F. Dinkel, et al, “A multifactorial model of pathology for age of onset heterogeneity in familial Alzheimer's disease”. *Acta neuropathologica*, 141(2), pp. 217–233, 2021.

[12] UniProt Consortium “UniProt: a worldwide hub of protein knowledge”. Nucleic acids research, 47 (D1), pp. D506-D515, 2019.

[13] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork and S. R. Sunyaev, “ A method and server for predicting damaging missense mutations”. *Nature* methods, 7(4), pp. 248-249, 2010.

[14] J. Yang and Y. Zhang, “I-TASSER server: new development for protein structure and function predictions”. *Nucleic Acids* Research, 43: W174-W181, 2015.

[15] E. Krissinel and K. Henrick, “Protein structure comparison in 3D based on secondary structure matching (PDBeFold) followed by Ca alignment, scored by a new structural similarity function. In: Andreas J. Kungl & Penelope J. Kungl (Eds.)”, Proceedings of the 5th International Conference on Molecular Structural Biology, Vienna, September 3-7, p.88, 2003.

[16] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, et.al., “CATH: increased structural coverage of functional space”. *Nucleic Acids Res.* 49(D1) pp. D266-D273,2021.

[17] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, et. al, “STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. *Nucleic Acids Research*, 47(D1), pp. D607-D613, 2019.

[18] E. Polychronidou, A. Avramouli and P. Vlamos, “Alzheimer’s Disease: The Role of Mutations in Protein Folding”. Adv Exp Med Biol., 1195, pp. 227-236, 2020.

[19] M. Ester, H.P. Kriegel, J. Sander, X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. *KDD Proceedings* pp. 226-231, 1996.

 ISBN: 978-1-61208-859-4