Article

In Silico Protocol for structure and function prediction of proteins in biological systems

Mohamed Ragab Abdel Gawwad
Genetics and Bioengineering, Faculty of Engineering and Natural Sciences, International University of Sarajevo, Sarajevo, Bosnia and Herzegovina
E-mail: mragab@ius.edu.ba

Received 28 March 2023; *Accepted 8 May 2023*; *Published online 30 June 2023*; *Published 1 December 2023*

Abstract
Proteins carry out the majority of the biological processes of living cells. These macromolecules are almost involved in every cellular process: replication and transcription of DNA, and produce, process, and biosynthesize other proteins. They are controlling cell division, metabolism, and materials and information transportation into and out of the cell. The understanding how living cells work requires revealing the structure and function of proteins. Our protocol aim is to analyze protein(s) structure and its function in biological systems. There are 9 steps of the protocol by which the protein(s) is extensively analyzed structurally and functionally. These steps are bioinformatics tools and software: Retrieving sequence of protein(s), Sequence alignment, Phylogenetic tree construction, Protein(s) 3D Structure Prediction and Refinement, Protein(s) 3D structure validation and visualization, Protein(s) Domain Identification, Subcellular Localization of Protein(s) and Interactome Analysis. The protocol links between the structure and function of the protein(s) and a heuristic to predict protein(s) function.

Keywords proteomics; 3D Structure prediction; bioinformatics.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Proteins are the most abundant macromolecules in the living cells. Proteins perform a variety of functions, including catalyzing biochemical reactions, transporting molecules, and providing structural support. In spite of our better knowledge about genomic but we miss a lot of information about proteins. In this study, the sequential procedure steps are very crucial to understand the structure of the protein which is the initial step to understand the functionality of the protein.

2 Material and Methods

2.1 Retrieving sequence of protein(s)

The sequences of proteins were retrieved from the National Center for Biotechnology Information (NCBI) or The Arabidopsis Information Resource (TAIR) databases.

2.2 Sequence alignment

Pairwise or Multiple sequence alignments (MSA) are tools that are used for prediction and analysis of protein structure and function, and also some other essential tasks like homology and phylogeny inference (Edgar and Batzoglu, 2006). When several sequences are aligned, it shows areas where they are similar and that might be related to specific properties and features that are more conserved than other regions (Sievers et al., 2011). This method is commonly used for aligning several sequences at once (Bacon and Anderson, 1986). In this study we mainly want to check the similarity and homology between the sequences of proteins and also the patterns of conserved regions between the sequences. Clustal Omega was the tool used to obtain the wanted results, which is available at the European Bioinformatics Institute (EMBL-EBI) website. Clustal Omega is the most popular multiple sequence alignment tool that is based on using the various techniques at once in order to do alignment of three or more sequences (Sievers et al., 2011). For obtaining the results default options were used.

2.3 Phylogenetic tree construction

Phylogenetic tree was constructed in order to predict the evolutionary relationships among proteins of the study and to deduce similarity among them. The tool used was the Phylogeny.fr software (Dereeper et al., 2008).

2.4 Protein(s) 3D Structure Prediction and Refinement

In order to determine the functions of a protein and possible interaction with other molecules, protein three-dimensional structure has to be determined. Therefore, structure infers function (Berg et al., 2002). Bioinformatics tools were used in order to predict 3D structure of protein(s) in silico.

Three dimensional (3D) models for protein(s) were built using the SWISS-MODEL server. The SWISS-MODEL represents a completely automated modeling server which is based on protein structure and homology, and it is accessible through ExPASy web server. SWISS-MODEL can automatically build model with using its own modelling algorithms already built in the program (Arnold et al. 2006). It was the first automated modelling server which started in 1993 and today it is the most used web server used for this purpose (Schwede et al., 2003).

2.5 Protein(s) 3D structure validation and visualization

This was shown to be valid by usage of Ramachandran plot on the RAMPAGE webpage (Lovell et al., 2002). The Ramachandran plot is used to visualize the possible conformation angles (phi and psi) of the polypeptides in the protein molecule (Gopalakrishnan et al., 2007). Best model was picked by determining which model has the higher number of residues in the favored region. The expected and best score should be around 98 % (Lovell et al., 2002).

The validation was also shown by global model quality estimation (GMQE) score. It is a score that estimates quality and it makes the combination of properties from the alignment. It is represented in a range from 0 to 1 and it shows the accuracy of the model that was built. Higher accuracy and reliability is represented with higher number (score) (Biasini et al., 2014).

The PDB files downloaded from the SWISS-MODEL were visualized in the PyMOL (The PyMOL Molecular Graphics System) in order to visualize the models and get the high-quality picture. It is a visualization tool on molecular level which enables us to view, customize and download the picture of molecule which is visualized (Secic et al., 2015).

2.6 Protein(s) Domain Identification

The identification of protein domains was done using the Simple Modular Architecture Research Tool

(SMART), which can be found on the website of European Molecular Biology Laboratory (EMBL). SMART can provide the identification and annotation of many mobile domains and also the further analysis of the architecture of domains. This tool can detect more than 400 families of domains that can be found in signaling, chromatin-associated and extracellular proteins. These types of domains are described in detail considering the phyletic distribution, functional class, tertiary structures and residues that have an important function. User interfaces that are part of this database provide possibility to search for proteins that have a specific combination of domains (Schultz et al., 2000).

2.7 Subcellular Localization of Protein(s)

Localization of proteins was determined in subcellular level where two different localization tools were used.

Determining the subcellular localization of a protein is an important step towards discovering its function(s) and also in determining and describing its interaction with other proteins in the cell. For subcellular localization of proteins, Plant Subcellular localization integrative predictor (PSI) tool was used. This tool is currently the most widely used subcellular location tool predictor for plants. The PSI predicts the location using the scores of multiple different predictors and makes the integrated score. It uses 11 predictor softwares to get integrated prediction results and these are: cello, Predotar, MultiLoc, Wolf PSORT, Ipsort, PTS1, mPloc, mitoProt, TargetP, SubcellPredict and YLoc. PSI gives precise results on most of the cellular compartments like vacuole, mitochondria, nucleus, golgi apparatus, membrane, plastid, cytosol, peroxisome and endoplasmic reticulum (Liu et al., 2013).

For subnuclear localization the online tool NucPred was used (MacCallum et al. 2017). NucPred analyses the sequences from eukaryotes and constructs a prediction that tells us if protein is at least sometimes located in the nucleus or never at all. It is a one of the newest web based tools which work on the basis of matching due to regular expression and also based on the work of several classifiers that are the result of genetic programming. The final score is calculated by the programs based on the each sequence that is submitted and position of each residue (Brameier et al., 2007).

2.8 Interactome Analysis

The interaction between proteins and also the proteins they interact with were demonstrated by Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). The STRING database’s main purpose is to show us and provide us with certain protein-protein interactions, which includes physical (direct) interactions as well as functional (indirect) interactions. This database contains information from more than 2000 organisms and operates on new and up-to-date algorithms that are responsible for transfer of information about interaction between different organisms (Szklarczyk et al., 2015).

2.9 Docking site prediction

Computational docking represents a series of processes in which the positioning and binding affinity of ligand (protein) in the binding site of another protein is computationally predicted. Docking methods are mainly based on specific algorithms which are actually responsible for computing the position of ligand in the active site and also calculating the scoring function which shows us the binding affinity (how strong is the interaction between the ligand and receptor) (Dhanik and Kavraki, 2012).

Before performing the docking site prediction, FASTA sequences of proteins that are chosen for docking were retrieved firstly, and using the SWISS-MODEL, models of these proteins were built. These models were validated by Ramachandran plot and GMQE scores, and eventually Protein Data Bank (PDB) files of these proteins are retrieved. Finally, these proteins were visualized using PyMOL.

For docking prediction of a protein with some other chosen proteins, an online protein-protein docking tool was used according to (Kozakov et al., 2017). It represents one of the first web programs that was completely automated and used for computational docking of proteins. User can usually upload PDB files from its computer or type in the PDB codes for certain model which are then automatically downloaded from the PDB server. The algorithms that are responsible for docking evaluation check billions of protein complexes saving only the ones that have similar favorable surfaces. This is followed by a filtering method which is applied to all of the structures, and in the end, the ones with the good electrostatic energies are selected (Comeau et al., 2004).

3 Results

This protocol was successfully used to publish articles in different scientific journals in the field of proteomics and bioinformatics; Network Biology Journal, Current Proteomics Journal, Molecules MDPI Journal and Journal of Infectious and Public Health (Gawwad et al., 2013; Sutkovic et al., 2013; Ćemanović et al., 2014; Sutkovic and Gawwad, 2014; Gawwad et al., 2015; Gawwad and Musrati, 2015; Gawwad et al., 2016; Gawwad et al., 2017; Gawwad et al., 2020).

4 Discussion

Proteins are large, complex molecules that perform a wide range of crucial functions in the biological systems, including catalyzing chemical reactions, transporting molecules, and providing structural support. The structure of a protein is intimately tied to its function, and understanding the relationship between structure and function is essential for understanding many aspects of biology and medicine.

Bioinformatics tools play a critical role in studying proteins, as they enable researchers to analyze and interpret vast amounts of data related to protein structure and function. These tools include sequence analysis software, which can be used to identify and compare protein sequences; structural analysis software, which can be used to visualize and analyze protein structures; and molecular dynamics simulations, which can be used to study the dynamic behavior of proteins.

One particularly powerful tool in the field of bioinformatics is protein structure prediction, which involves using computational methods to predict the three-dimensional structure of a protein based on its amino acid sequence. This technique has revolutionized the study of proteins, as it allows researchers to make predictions about protein function and behavior before conducting expensive and time-consuming experimental studies.

The In Silico Protocol for structure and function prediction of proteins in biological systems is a crucial tool for understanding the behavior and properties of proteins, which are essential components of living organisms. Understanding the structure and function of proteins is critical for many fields, including drug discovery, biotechnology, and molecular biology. The Protocol allows researchers to predict the three-dimensional structure and function of proteins using computational methods, without the need for laborious and time-consuming experimental approaches. This protocol has revolutionized the study of proteins, enabling researchers to make predictions that were previously difficult, leading to a deeper understanding of biological systems.

Acknowledgment

I would like to express my gratitude to my colleagues at the International University of Sarajevo for their valuable contributions to this research. Their insightful feedback and discussions were instrumental in shaping the protocol in this publication. I am grateful for their support and collaboration.

References

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics, 22(2): 195-201

Bacon DJ, Anderson WF. 1986. Multiple sequence alignment. Journal of Molecular Biology, 191(2): 153-161

Berg JM, Tymoczko JL, Stryer L. 2002. Biochemistry (5th edition). WH Freeman, New York, USA

Biasini M, Bienert S, et al. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Research, 42: 252-258

Brameier M, Krings A, MacCallum RM. 2007. NucPred—Predicting nuclear localization of proteins. Bioinformatics, 23(9): 1159-1160

Ćemanović A, Šutković J, Elkhoby W, Gawwad MRA. 2014. 3D structure prediction of histone acetyltransferase (HAC) proteins of the p300/CBP family and their interactome in *Arabidopsis thaliana*. Network Biology, 4(3): 109-122

Cemanovic A, Sutkovic J, Gawwad MRA. 2014. Comparative structural analysis of HAC1 in Arabidopsis thaliana. Network Biology, 4(2): 67-73

Comeau SR, Gatchell DW, Vajda S, Camacho CJ. 2004. ClusPro: a fully automated algorithm for protein–protein docking. Nucleic Acids Research, 32: 96-99

Dereeper A. Guignon V, Blanc G, et al. 2008. Phylogeny.fr robust phylogenetic analysis for the non-specialist. Nucleic Acid Research, 36: 465-469

Dhanik A, Kavraki LE. 2012. Protein–Ligand Interactions: Computational Docking. Retrieved Jun 2017, from http://www.els.net: doi: 10.1002/9780470015902.a0004105.pub2

Edgar RC, Batzoglu S. 2006. Multiple sequence alignment. Current Opinion in Structural Biology, 16(3): 368-373

Gopalakrishnan K, Sowmiya G, Sheik SS, Sekar K. 2007. Ramachandran plot on the web (2.0). Protein and Peptide Letters, 14(7): 669-671

Gawwad MRA, et al. 2019. Interactome analysis and docking sites of Muts homologs reveal new physiological roles in *Arabidopsis thaliana*. Molecules, 24(13): 2493

Gawwad MRA, Musrati MA. 2015. In silico structural and functional analysis of *Arabidopsis thaliana*'s XPB homologs. Current Proteomics, 12(4): 236-244

Gawwad MRA, Kadunic A, Adilović M, Kaljanac AM, Maric A. 2016. Analysis of DNA Damage-Binding Proteins (DDBs) in *Arabidopsis thaliana* and their protection of the plant from UV radiation. Current Proteomics, 14(2): 146-156

Gawwad MRA, Ucuncu D, MAdilović M, Marić A. 2017. Interactome analysis and docking site prediction of Cockayne Syndrome A (CSA) proteins in *Arabidopsis thaliana*. Current Proteomics, 14(3): 242-251

Gawwad MRA, Šutković J, Mataković L, Musrati M, Zhang LZ. 2013. Functional interactome of Aquaporin 1 sub-family reveals new physiological functions in *Arabidopsis thaliana*. Network Biology, 3(3): 87-96

Gawwad MRA, Alpdemir S, Eminagic E. 2015. Interactome analysis and docking sites of PCNA subunits reveal new function in *Arabidopsis thaliana*. Current Proteomics, 12(3): 152-167

Gawwad MRA, Šutković J, Zahirović E, Akcesme FB, Akcesme B, Zhang L. 2013. 3D structure prediction of replication factor C subunits (RFC) and their interactome in *Arabidopsis thaliana*. Network Biology, 3(2): 74-86

Gawwad MRA, Elshikh MS, Lokvancic H. 2020. Interactome analysis and docking site prediction of DNA X-ray repair cross-complementing protein (XRCC) in *Arabidopsis thaliana*. Network Biology, 10(2): 32-44

Gawwad MRA, Mahmutovic´ E, Dunia A. Al Farraj, Elshikh MS. 2020. In silico prediction of silver nitrate nanoparticles and Nitrate Reductase A (NAR A) interaction in the treatment of infectious disease causing clinical strains of *E. coli*. Journal of Infection and Public Health, 13(10): 1580-1585

Kozakov D, et al. 2017. The ClusPro web server for protein-protein docking. Nature Protocols, 12(2): 255-278

Liu L, Zhang Z, Chen M. 2013. PSI: A comprehensive and integrative approach for accurate plant subcellular localization prediction. Plos One, 8(10): 0075826

Lovell SC, de Bakker P, et al. 2002. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. Proteins: Structure, Function & Genetics, 50(3): 437-450

MacCallum B, Brameier M, Krings A, Heddad A. 2017. NucPred - Predicting Nuclear Localization of Proteins. Retrieved June 2017, from Stockholm Bioinformatics Center: http://www.sbc.su.se/~maccallr/nucpred/cgi-bin/single.cgi

Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. 2000. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Research, 28(1): 231-234

Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Research, 31(13): 3381-3385

Secic E, Sutkovic J, Gawwad MRA. 2015. Interactome analysis and docking sites prediction of Radiation Sensitive 23 (RAD 23) proteins in *Arabidopsis thaliana*. Current Proteomics, 12(1): 28-44

Sievers F, Wilm A. et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology, 7(1): 539

Sutkovic J, Gawwad MRA. 2014. In silico prediction of three-dimensional structure and interactome analysis of Tubulin α subfamily of *Arabidopsis thaliana*. Network Biology, 4(2): 47-57

Szklarczyk D, Franceschini A, et al. 2015. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research, 43(Database issue): 447-452