In silico Sequence Analysis, Homology Modeling and Function Annotation of *Ocimum basilicum* Hypothetical Protein G1CT28_OCIBA

Sobia Idrees*, Shahid Nadeem, Samia Kanwal, Beenish Ehsan, Ayesha Yousaf, Sameera Nadeem, Muhammad Ibrahim Rajoka

Department of Bioinformatics and Biotechnology
Government College University
Faisalabad, Pakistan
E-mail: sobia_binm@live.com

**Corresponding author*

Received: December 05, 2011 **Accepted: July 12, 2012**

Published: July 16, 2012

Abstract: *Ocimum basilicum is commonly known as sweet basil and belongs to the Lamiaceae Family. Ocimum basilicum has great therapeutic benefits and can be used for lowering blood pressure, as an antispasmodic as well as cleansing the blood. In the present study, subcellular localization prediction suggested that it is a cytoplasmic protein. We predicted the 3D structure of protein using homology modeling as 3D structure prediction approach. 3D structure of the protein was determined using Protein Structure Prediction Server (PS)^2 selecting MODELLER as 3D structure prediction method. Quality analysis of the model indicated that it is a reliable model. Furthermore, it was discovered that Ocimum basilicum hypothetical protein G1CT28_OCIBA is involved in two biological processes, oxidation reduction and metabolic process and the biochemical function of the protein is acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor, catalytic activity and oxidoreductase .*

Keywords: *Sweet basil, Sequence analysis, Homology modeling, Function annotation.*

Introduction

The Lamiaceae family, which includes basil, sage, and thyme, has long been recognized as a rich source of diverse and unique anthocyanins. The development of intensely purple pigmented basil in the ornamental and herb trade prompted the examination of eight commercial varieties of purple basils (*Ocimum basilicum* L.) as a potential new source of anthocyanins [1]. The genus Ocimum (Lamiaceae) has a long history of use as culinary and medicinal herbs. Many species are used for their antioxidant and neuroprotective activity in various parts of the world. *Ocimum basilicum* Linn. has been used traditionally for the treatment of anxiety, diabetes, cardiovascular diseases, headaches, nerve pain, as anticonvulsant and anti-inflammatory, and used in a variety of neurodegenerative disorders [2]. Chemical composition, antioxidant and antimicrobial activities of the essential oils from aerial parts of basil (*Ocimum basilicum* L.) are affected by four seasons, namely summer, autumn, winter and spring [3]. Computational analysis of biological sequences has become an extremely rich field of modern science and a highly interdisciplinary area, where statistical and algorithmic methods play a key role [4].

In present study the objectives were set to:

- Perform sequence analysis on *Ocimum basilicum* hypothetical protein G1CT28_OCIBA.
- Perform the primary and secondary structure analysis.

- Perform homology modeling to find the 3D structure of the hypothetical protein.
- Ensure the quality of the predicted model.

Methodology

Protein retrieval and sequence analysis:

The protein sequence of G1CT28_OCIBA protein was retrieved from Uniprot Knowledgebase database using accession No. G1CT28. Physiochemical properties of the protein were computed by ProtParam tool (http://web.expasy.org/protparam/). The parameters computed by ProtParam included the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). Subcellular localization of any protein is important understanding protein function. Prediction of subcellular localization of protein was carried out by CELLO v.2.5 [5, 6].

Secondary structure prediction

PredictProtein [7] was employed for computing and analyzing the secondary structural features of *O. basilicum* protein sequence.

3D structure prediction using homology approach

3D structure of protein was determined by homology modeling. BLASTP search with default parameters against the Protein data bank (PDB) was used to find the best suitable templates for homology modeling. Based on maximum identity and lowest e-value, best suitable template with PDB ID 1J0X_O having 74% identity was selected. This template was used as a reference to determine the 3D structure of G1CT28_OCIBA protein. Comparative amino acid composition of the query and template protein was determined using CLC protein workbench. Protein Structure Prediction Server $(PS)^2$ [8] predicted the homology model based on package MODELLER.

Quality and reliability assessments

Once the 3D model was generated, energy minimization was performed by GROMOS96 force field in a Swiss-PdbViewer. Structural evaluation and stereochemical analyses were performed using ProSA-web [9, 10] Z-scores and Procheck Ramachandran plot. Furthermore, superimposition of query and template structure, and visualization of generated models was performed using UCSF Chimera 1.5.3.

Function annotations of the protein

To functionally annotate the *O. basilicum* hypothetical protein G1CT28_OCIBA, Profunc was used and to find the conserved domains in protein to identify its family, it was searched against close orthologous family members. NCBI Conserved Domain Database (NCBI CDD) [11] was used to find the conserved domains or ancient domains in the protein sequence.

Submission of the model in protein model database (PMDB)

The models generated for *O. basilicum* hypothetical protein G1CT28_OCIBA was successfully submitted in Protein model database (PMDB) [12] having PMID PM0077760.

Results and discussions

The present study was to perform sequence and structure analysis of *O. basilicum* hypothetical protein G1CT28_OCIBA.The protein sequence was retrieved using accession No. G1CT28 from uniprot database.

Protein sequence analysis

ProtParam was used to find out the physiochemical properties from protein sequence. The hypothetical protein was predicted to have 195 amino acids, with molecular weight of 20778.6 Daltons and theoretical isoelectric point (PI) of 7.02. An isoelectric point above 7 indicates a positively charged protein. The instability index (II) is computed to be 32.78. This classifies the protein as stable. The N-terminal of the sequence considered is F (Phe). Therefore estimated half-life is 1.1 hours (mammalian reticulocytes, in vitro), 3 min (yeast, in vivo) and 2 min (*Escherichia coli*, in vivo). The negative Grand average of hydropathicity (GRAVY) of -0.137 indicates that the protein is hydrophilic and soluble in nature. Valine and alanine were found in rich amounts in the protein. Alignment and comparative composition of query sequence and template sequence was computed using CLC protein workbench (Table 1).

Table 1. Comparative amino acid composition of query and template proteins

Amino acid	1J0X	G1CT28_OCIBA
Alanine (A)	33	19
Cysteine (C)	3	2
Aspartic Acid (D)	21	14
Glutamic Acid (E)	13	11
Phenylalanine (F)	14	8
Glycine (G)	32	15
Histidine (H)	11	3
Isoleucine (I)	20	12
Lysine (K)	26	19
Leucine (L)	18	11
Methionine (M)	9	3
Asparagine (N)	17	7
Proline (P)	11	8
Glutamine (Q)	6	1
Arginine (R)	10	6
Serine (S)	19	17
Threonine (T)	21	15

Cellular functions are often localized in specific compartments; therefore, predicting the subcellular localization of unknown proteins can give information about their functions and can also help in understanding disease mechanisms and developing drugs. The subcellular localization prediction using CELLO predicted that our protein is a cytoplasmic protein and this protein does not contain a nuclear localization signal. PredictProtein was used to predict the secondary structure of the protein. Results showed that protein is a mixed protein having composition of Helix = 29.2%, Strand = 26.7%, Loop = 44.1%.

3D structure prediction using homology modeling approach

Protein 3D structure is very important in understanding the protein interactions, functions and their localization. Homology modeling is the most common structure prediction method. To perform the homology modeling, first and basic step is to find best matching template using similarity searching program like BLASTP against PDB database. Templates are selected on the basis of their sequence similarity with query sequence. PDB ID 1J0X_O was selected for

homology modeling which is an X-Ray diffraction model of Crystal Structure of the rabbit muscle glyceraldehydes-3-phosphate dehydrogenase (Gapdh). The query sequence and template ID was then given as input to the $(PS)^2$ server for homology modeling using MODELLER. 3D structure of protein showed that it has 544 hydrogen bonds (Fig. 1). Quality and reliability of structure was checked by several structure assessment methods including Z-score and Ramachandram plots. The Z-score is indicative of overall model quality and is used to check whether the input structure is within the range of scores typically found for native proteins of similar size. PROSAweb was used to find the Z-score of template and query. Z-score of query protein was -7.39 and Z-score of template was -9.98. Procheck checks the stereochemical quality of a protein structure by analyzing residue-by-residue geometry and overall structure geometry. This tool was used to determine the Ramachandran plot to assure the quality of the model. The result of the Ramachandran plot showed 93.5% of residues in favorable region representing that it is a reliable and good quality model (Fig. 2). A model having more than 90% residues in favorable region is considered as good quality model. Reliability of the model was further checked by ERRAT that analyzes the statistics of non-bonded interactions between different atom types and plots the value of the error function versus position of a 9-residue sliding window, calculated by a comparison with statistics from highly refined structures. Results from ERRAT showed 93.583 overall model quality (Fig. 3). The Z-scores (Fig. 4) confirm the quality of the homology model of *O. basilicum* hypothetical protein G1CT28_OCIBA.

Fig. 1 Predicted 3D structure of *O. basilicum* hypothetical protein G1CT28_OCIBA

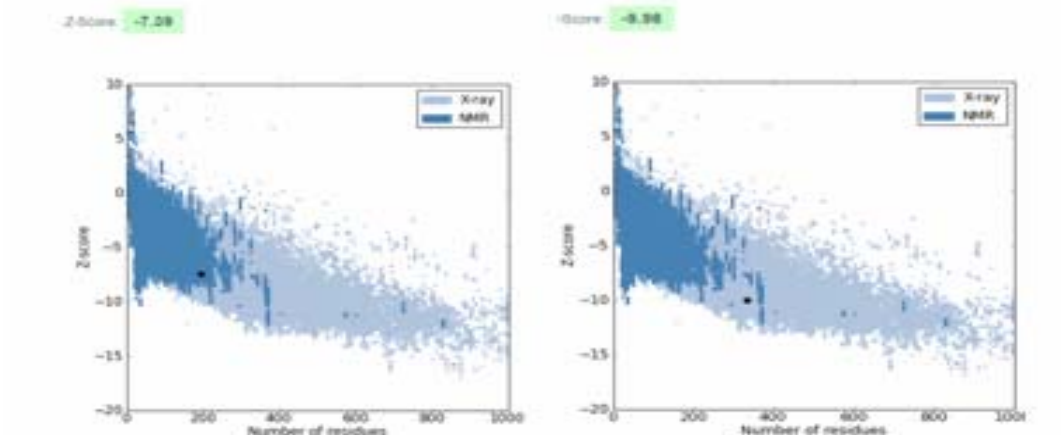

Fig. 2 Z-score of query and template protein using PROSA web

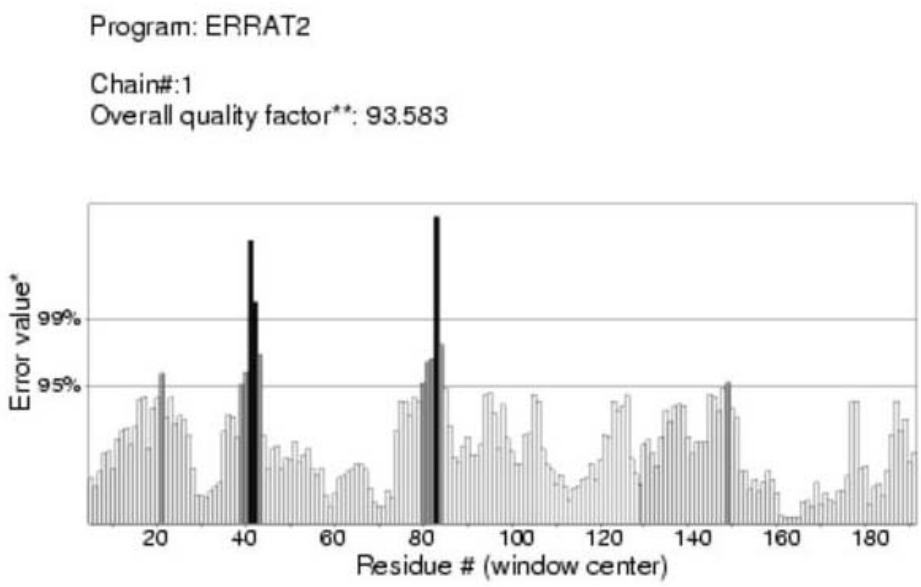

Fig. 3 Overall quality factor checked by ERRAT

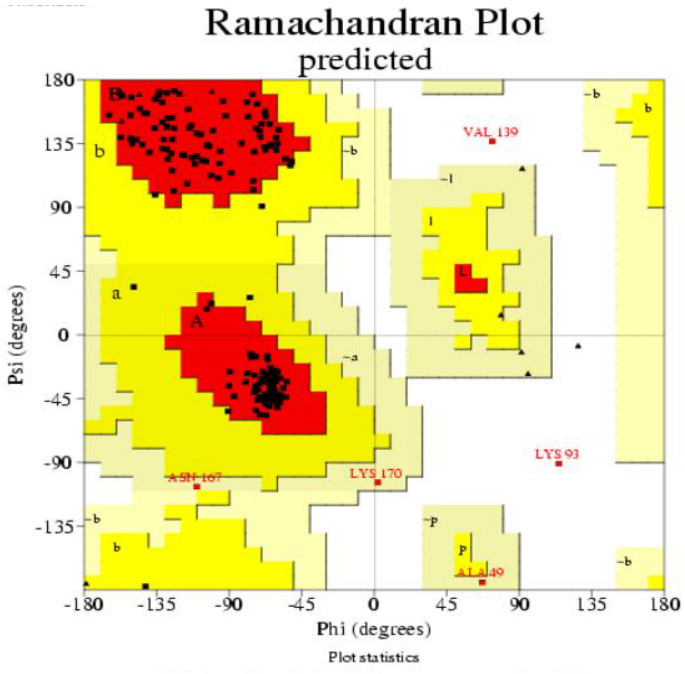

Fig. 4 Ramachandran plot using Procheck

Function annotation of the protein

To hypothetically annotate the function of the *O. basilicum* hypothetical protein G1CT28_OCIBA ProFunc was used. It was discovered that protein is involved in two biological processes, oxidative reduction and metabolic processes and the biochemical function of the protein is oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor and catalytic activity. To futher investigate about the function of protein by finding its family; it was searched in the NCBI Conserved Domain Database (NCBI CDD) to find conserved domains so that its family can be identified. The results showed that *O. basilicum* hypothetical protein G1CT28_OCIBA has NAD (P)

binding domain and belongs to Gp_dh_C super family i.e Glyceraldehyde 3-phosphate dehydrogenase, C-terminal domain; GAPDH is a tetrameric NAD-binding enzyme involved in glycolysis and glyconeogenesis.

Submission of the model in protein model database (PMDB)

The predicted structure of the protein was submitted to protein model database (PMDB) and can be find using PMID PM0077760.

Conclusion

Our main objective of this study was to perform sequence analysis, structure analysis and homology modeling on *O. basilicum* hypothetical protein G1CT28_OCIBA. We have used various sequence and structure analysis tools that helped in understanding of the sequence and its structure. Furthermore, protein was functionally annotated by using ProFunc and by searching conserved domain of the protein. As a part of present study, we used homology modeling approach to propose the first 3D structure of the *O. basilicum* hypothetical protein G1CT28_OCIBA. The predicted 3D structure will provide more insight in understanding the structure and function of the protein. Moreover this structure can be used for drug designing or understanding the interactions between proteins.

References

1. Phippen B. W., E. J. Simon (1998). Anthocyanins in Basil (*Ocimum basilicum* L.), Journal of Agricultural and Food Chemistry, 46(5), 1734-1738.
2. Bora K. S., S. Arora, R. Shri (2011). Role of *Ocimum basilicum* L. in Prevention of Ischemia and Reperfusion-induced Cerebral Damage, and Motor Dysfunctions in Mice Brain, Journal of Ethnopharmacology, 137(3), 1360-1365.
3. Hussain I. A., F. Anwar, H. T. S. Sherazi, R. Przybylski (2007). Chemical Composition, Antioxidant and Antimicrobial Activities of Basil (*Ocimum basilicum*) Essential Oils Depends on Seasonal Variations, Food Chemistry, 108(3), 986-995.
4. Giancarlo R., A. Siragusa, E. Siragusa, F. Utro (2007). A Basic Analysis Toolkit for Biological Sequences, Algorithms for Molecular Biology, 2(10), 404-406.
5. Yu C. S., C. J. Lin, J. K. Hwang (2006). Predicting Subcellular Localization of Proteins for Gram-negative Bacteria by Support Vector Machines based on n-peptide Compositions, Protein Science, 13, 1402-1406.
6. Yu C. S., Y. C. Chen, C. H. Lu, J. K. Hwang (2004). Prediction of Protein Subcellular Localization, Proteins: Structure, Function and Bioinformatics, 64, 643-651.
7. Rost B., G. Yachdav, J. Liu (2004). The PredictProtein Server, Nucleic Acids Research, 32(Web Server issue), W321-W326.
8. Chen C. C., J. K. Hwang, J. M. Yang (2006). $(PS)^2$: Protein Structure Prediction Server, Nucleic Acids Research, 34, W152-W157.
9. Wiederstein M., M. J. Sippl (2007). ProSA-web: Interactive Web Service for the Recognition of Errors in Three-dimensional Structures of Proteins, Nucleic Acids Research, 35, W407-W410.
10. Sippl M. J. (1993). Recognition of Errors in Three-Dimensional Structures of Proteins, Proteins, 17, 355-362.
11. Marchler-Bauer A., S. H. Bryant (2004). CD-Search: Protein Domain Annotations on the Fly, Nucleic Acids Research, 32(W), 327-331.
12. Arnold K., F. Kiefer, J. Kopp, J. N. Battey, M. Podvinec, J. D. Westbrook, H. M. Berman, L. Bordoli, T. Schwede (2009). The Protein Model Portal, Journal of Structural and Functional Genomics, 10, 1-8.

Sobia Idrees, M.Phil. in Biotechnology
E-mail: sobia_binm@live.com

Sobia Idrees is a Master student at Government College University, Faisalabad, Pakistan. Scientific Interests: Software Engineering, Web and Database Development, Bioinformatics Sequence Analysis, Drug Designing, Health Biotechnology and Molecular Biology.

Shahid Nadeem, Ph.D.
E-mail: snadeem63@yahoo.com

Dr. Shahid Nadeem got his Ph.D. from the Punjab University, Lahore in 2002. He is working as a senior scientist in NIAB, Faisalabad. In 2009 he joined the Department of Bioinformatics and Biotechnology as a Chairman. He is among the founders of Department of Bioinformatics and Biotechnology in GC University Faisalabad. He is also the founder of Database and Software Engineering Research Group in the Department of Bioinformatics and Biotechnology. He is also an active member of teaching Faculty of the department, teaching some of the interesting subjects like Biotechnology and its applications, Social, Ethical aspects of Biotechnology also publishing textbooks. Scientific interests: Biotechnology, Plant breeding and Genetics, Microbiology.

Samia Kanwal, M.Sc. in Bioinformatics
E-mail: samiabioinfo@yahoo.com

Samia Kanwal is a Master student at Government College University, Faisalabad, Pakistan. Scientific Interests: Web and Database Development, Homology Modelling and Protein Docking.

Beenish Ehsan, M.Phil. in Biotechnology
E-mail: merjeena_196@yahoo.com

Beenish Ehsan is pursuing Master in Biotechnology at Government College University, Faisalabad. Scientific Interests: Phylogenetic Analysis, Molecular Biology, Health Biotechnology.

Ayesha Yousaf, M.Sc. in Bioinformatics
E-mail: aishayousaf@live.com

Ayesha Yousaf is pursuing Master in Bioinformatics at Quaid e Azam University, Islamabad. Scientific Interests: Phylogenetic Analysis, Structure Prediction and Software Development.

Sameera Nadeem, M.Sc. in Bioinformatics
E-mail: chocolate_hits@yahoo.com

Sameera Nadeem is pursuing Master in Bioinformatics at Government College University, Faisalabad. Scientific Interests: Structure Prediction, Protein-protein interactions and Drug Designing.

Muhammad Ibrahim Rajoka, Ph.D. in Biochemistry
E-mail: mibrahimrajoka47@gmail.com

M. I. Rajoka graduated from the University of New South Wales – Australia in 1981 with the distinction of securing second highest percentage in the session. He did Ph.D. in Biochemistry from University of the Punjab in 1990. After serving 41 years in Nuclear Institute for Agriculture and Biology – Faisalabad (19 years) and National Institute for Biotechnology and Genetic Engineering – Faisalabad (22 years), he joined as Professor in the Department of Bioinformatics and Biotechnology. He has supervised 25, 21 and 7 M.Sc., M.Phil. and Ph.D. students respectively and has published 140 research papers and review articles in International and national peer review journals and proceedings of conferences/workshops. He is also an active member of teaching Faculty of the department, teaching bioprocess technology, research methods in biological sciences, technical writing, communication skills and guiding faculty in writing research proposals. Scientific interests: Industrial strain development through conventional breeding and rDNA technology, bioprocess engineering.