BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(8)

Homology modeling of an antifungal metabolite plipastatin synthase from the *Bacillus subtilis* 168

Maria Batool[1, 2], Mohammad Hassan Khalid[1], Muhammad Nadeem Hassan[1], Fauzia Yusuf Hafeez[1]*

[1]Department of Biosciences, COMSATS Institute of Information Technology (CIIT), Islamabad, Pakistan; [2]Department of Bioinformatics & Biotechnology, Government College University, Faisalabad, Pakistan; Fauzia Yusuf Hafeez – Email: fauzia@comsats.edu.pk; phone: +92-03006602614; *Corresponding author

Received December 2, 2011; Accepted December 6, 2011; Published December 21, 2011

Abstract:
Lipopeptides have a widespread role in different pathways of Bacillus subtilis; they can act as antagonists, spreader and immuno-stimulators. Plipastatin, an antifungal antibiotic, is one of the most important lipopeptide nonribosomly produced by Bacillus subtilis. Plipastatin has strong fungitoxic activity and involve in inhibition of phospholipase A2 and biofilm formation. For better understanding of the molecule and pathway by which lipopeptide plipastatin is synthesized, we present a computationally predicted structure of plipastatin using homology modeling. Primary and secondary structure analysis suggested that ppsD is a hydrophilic protein containing a significant proportion of alpha helices, and subcellular localization predictions suggested it is a cytoplasmic protein. The tertiary structure of protein (plipastatin synthase subunit D) was predicted by homology modeling. The results suggest a flexible structure which is also an important characteristic of active enzymes enabling them to bind various cofactors and substrates for proper functioning. Validation of 3D structure was done using Ramachandran plot ProsA-web and QMEAN score.This predicted information will help in better understanding of mechanisms underlying plipastatin synthase subunit D synthesis. Plipastatin can be used as an inhibitor of various fungal diseases in plants.

Keywords: Bacillus subtilis; plipastatin synthase; homology modeling

Background:
Bacillus subtilis is a Gram-positive endospore forming bacteria. It produces broad spectrum intoxicating lipopeptides known as biosurfactants and has antifungal, antibacterial and antiviral activities. Cyclic lipodecapeptide plipastatin is an example of such compound. Cyclic lipopeptides (CLPs) are structurally and functionally diverse molecules. CLPs are produced by large, complex, and multifunctional non-ribosomal peptide synthetases (NRPSs) using thioesterase mechanism, showing highly conserved structural organization into specific peptide-binding domain **[1].** NRPSs acquire a modular structure and each module act as building block resulting in a stepwise incorporation and modification of one amino acid unit **[2].** Fengycin is a lipodecapeptide termed as plipastatin when it's Tyr3 and Tyr9 residues are present in the L-and D-form, respectively. It is produced by different strains of Bacillus species and shows moderate surfactant properties. It is an anti-fungal metabolite and inhibits the filamentous fungi but it has no effect on yeast and bacteria. Plipastatin synthase subunit D (ppsD) contains five NRPS subunits which assemble to form a co-linear chain in the order ppsC-ppsD-ppsE-ppsA-ppsB **[3, 4].** The structure is composed of a beta-hydroxy fatty acids connected to a peptide part comprising ten amino acids, where eight of them are structured in cyclic form **[5].** Five multifunctional peptide synthases act in synergy to activate all fengycin amino acid components to form aminoacyl adenylates or aminoacyl thioesters. Disruption of ppsD results in loss of plipastatin production **[6].** Selected open reading frame (ORF) transcribes the ppsD gene out of the pps gene cluster encoding plipastatin synthase enzyme in Bacillus subtilis 168. Plipastatin

Synthase is a large, complex, and multifunctional enzyme, which activates and polymerizes the amino acids Pro, Gln, and Tyr as part of synthesis of plipastatin which is an antibiotic lipopeptide. The gain D-isomer form its tyrosine residue is epimerized **[7].** 3D structure of ppsD was not known. Structure is more evolutionary conserved than sequence; therefore, analysis of three-dimensional (3D) structures holds great potential.

Structure elucidation is an expensive and time consuming process, and also requires extensive expertise. Currently used techniques to reveal 3D structures are X-ray crystallography and Nuclear Magnetic Resonance (NMR) Imaging. Due to the techniques being expensive and time consuming, there has been an increasing gap of information between DNA/protein sequence information and structure information. Computational methods and molecular dynamic simulations are good alternatives to overcome these problems in protein structure prediction **[8].**Comparative modeling is a computational technique for 3D structure prediction of proteins using known structures as templates **[9].** Hence, the adopted approach enables to formulate a structure from an amino acid sequence in less time. Our current study describes the 3D model of the B.subtilis ppsD protein obtained through homology modeling. In addition, primary and secondary structure analysis, subcellular localization prediction was also performed and is explained.

Methodology:

Sequence analysis and subcellular localization prediction

The amino acid sequence of a Bacillus subtilis gene ppsD was retrieved from the UNIPROT database using the primary accession number P94459_BACSU. The sequence is 3603 amino acids in length. ProtParam **[10]** was used to predict the physiochemical properties. ProtParam computed 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). Secondary structure predictions (helix, sheets, and coils) were using made using PSIPRED **[11].** Prediction of subcellular localization of ppsD was done by CELLO v.2.5 **[12],** a multiclass support vector machine classification system. PSORTb v3.02 **[13]** was used to crosscheck the results.

Domain Analysis of ppsD:

Domains are the functional units of proteins. Plipastatin synthase is a modular protein comprising of three domains. These domain sequences were retrieved from UNIPROT under accession id P94459_BACSU. Each domain sequence was then analyzed using Interproscan **[14].**

Homology modeling and loop optimization of ppsD:

Homology modeling was used to determine the 3D structure of ppsD. As mentioned above ppsD consist of three domains we therefore modeled each domain separately. A BLASTP **[15]** search with default parameters were performed against the Brookhaven Protein Data Bank (PDB) to find suitable templates for homology modeling. PDB ID: 2VSQ_A was identified as the best template based on sequence identity between query and template protein sequence for all three domains. The alignment between template and query sequence was done using MODELLERv9.10 **[16],** which uses global dynamic programming with linear gap penalty function. This aligns the query and template sequences and the output is obtained in PIR format. The PIR format is used by MODELLER in the subsequent model-building stage. A three-dimensional structure was developed from sequence alignment between template and query sequence using MODELLERv9.10. It constructs model by satisfaction of spatial restraints, using its 'automodel' class. Twenty models were generated using the command using python script. Loops of the homology model were then optimized using MODELLERv9.10 to improve the quality of model. Energy minimization, quality assessment and visualization Once the 3D model was generated, structural evaluation and stereochemical analyses were performed using ProSA-web Z-scores **[17],** Qmean plots **[18]** and PROCHECK Ramachandran plots **[19].** Furthermore, Root Mean Squared Deviation (RMSD), superimposition of query and template structure, and visualization of generated models was performed using UCSF Chimera 1.5.3 workbench **[20].**

Discussion:

The present study focused on sequence and structural analysis of Bacillus subtilis protein ppsD. ProtParam was used to analyze different physiochemical properties from the amino acid sequence. The ppsD protein contains 3603 amino acids, with a molecular weight of 406812.3 Daltons and an isoelectric point of 5.46. An isoelectric point below 7 indicates a negatively charged protein corresponds to having more negatively charged residues, and an instability index of 45.99 suggests an unstable protein. The negative GRAVY index of -0.327 is indicative of a hydrophilic and soluble protein. The protein sequence was found to be rich in the amino acid leucine, suggesting a preference for alpha-helices in 3D structure. Secondary structure analysis was performed using PSIPRED and the protein was predicted to contain several helices, consistent with the ProtParam results. The high percentage of helices in the structure makes the protein more flexible for folding, which might increase protein interactions. Subcellular localization is a key functional feature of a protein. Cellular functions are often localized in specific compartments; therefore, subcellular localization prediction of unknown proteins could be used to attain valuable information about their functions, and to select proteins for further study. Moreover, studying the subcellular localization of proteins is also helpful in understanding disease mechanisms which can have application in developing novel drugs. The consensus protein subcellular localization predictions suggest that ppsD is a cytoplasmic protein and had no transmembrane helices.

Domain analysis of ppsD and homology modeling:

Interproscan predicted the three domains in ppsD gene, which were further divided into subdomains i.e. adenylation domain, condensation domain and epimerization domain. The three main domains predicted by interproscan are ATP-dependent proline adenylase (proline activase), ATP-dependent glutamine adenylase (glutamine activase), and ATP-dependent tyrosine adenylase II (tyrosine activase II). These domains are then modeled by homology modeling. Protein 3D structures can provide us with precise information of how proteins interact and localize in their stable conformation. Homology modeling is one of the most common structure prediction methods in structural genomics and proteomics. Despite minimal

modifications, one initial step that is common in all modeling tools and servers is to find the best matching template by performing a sequence homology search with BLASTP. Templates are experimentally determined 3D structures of proteins that share sequence similarity with the query sequence. The template sequence and the query protein sequence are aligned using pair-wise alignment algorithms **[21, 22].** A well-defined alignment is very crucial for the prediction of a reliable 3D structure using homology modeling. Here we used the protocol of homology modeling to model ppsD along with its three domains. A BLASTP search against the PDB database identified 2VSQ_A as a best template as it shows maximum identity to query sequence of all three domains. The identity between query and template was found to be 99% for all three domains. 2VSQ_A is an X-Ray diffraction model of a srfA-C gene of Bacillus subtilis. 3D structures were constructed using MODELLERv9.10. Ten models were obtained the best model was selected based on discrete optimized protein energy (DOPE). DOPE is based on an improved reference state that corresponds to noninteracting atoms in a homogeneous sphere with the radius dependent on a sample native structure; it thus accounts for the finite and spherical shape of the native structures.

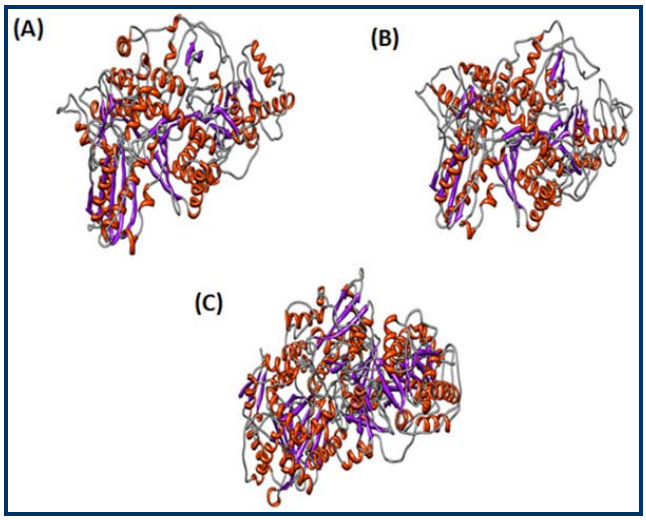

Figure 1: The figure shows modeled structures of respective domains of ppsD **(A)** Proline activase, **(B)** Glutamine activase, **(C)** Tyrosine activase. Secondary structure elements are highlighted in different colors. Image was generated using UCSF Chimera.

Energy minimization, quality assessment and visualization:

Even though there were no steric clashes in the structure generated, it was still subjected to energy minimization and assessed for both geometric and energy aspects. The predicted 3D structures of respective three domains proline activase, glutamine activase and tyrosine activase of ppsD are shown in **Figure 1a, 1b and 1c** respectively. Several structure assessment methods including RMSD, Z-scores, Q-scores and Ramachandran plots were used to check reliability of the 3D models. The RMSD value indicates the degree to which two 3D structures are similar. Lower the value, the more similar the structures. The RMSD value of each domain 3D structure was calculated by superimposing template and query structures. The RMSD values were found to be 0.541, 0.589, 0.745 for the proline activase, glutamine activase and tyrosine activase domains respectively. 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. Z-scores of the template and query models were obtained from PROSAweb. In case of proline activase, z-score was -7.23, for glutamine activase, z-score was -7.27 and for tyrosine activase, z-score was found to be -11.2, were suggesting similarity to their templates. Finally, the Ramachandran plots were obtained for the homology model for quality assessment. In case of proline activase domain, PROCHECK displayed 86.9% of residues in the most favored regions, with 10.2%, 2.0%, and 0.9% residues in additionally allowed, generously allowed and disallowed regions, respectively. In case of glutamine activase domain, PROCHECK displayed 87% of residues in the most favored regions, with 10.1%, 2.3%, and 0.6% residues in additionally allowed, generously allowed and disallowed regions, respectively and in case of tyrosine activase domain, PROCHECK displayed 89.8% of residues in the most favored regions, with 8.2%, 1.3%, and 0.6% residues in additionally allowed, generously allowed and disallowed regions, respectively. The Ramachandran plot for the template structure showed the amino acid residues to be 91.3%, 7.2%, 1.1% and 0.3%, in most favored, additionally allowed, generously allowed and disallowed regions respectively (Data not shown). The comparable Ramachandran plot characteristics, RMSD values, and Z-scores confirm the quality of the homology model of ppsD.

Protein structure accession number:

The predicted 3D structures of Bacillus subtilis protein ppsD domains were submitted to the Protein Model Database (PMDB) **[23]** and assigned the PMDB ID PM0077747 for proline activase, PM0077748 for glutamine activase and PM0077746 for tyrosine activase domain.

Conclusion:

We used homology modeling to solve the 3D structure of ppsD, an important antifungal protein, from Bacillus subtilis. Further investigation in this line of research may include interaction studies with other metabolites involved in the pathway. The predicted information is hoped to help in better understanding of the mechanisms underlying plipastatin synthesis as well as the production of novel therapeutic drugs for the treatment of long term diseases. Docking experiments may suggest additional lead compounds and medicinally significant conformations. Ultimately, the identification of pharmacologically active conformations via simulation will be great leap towards use of NRPSs as the new generation drugs.

Acknowledgments:

We would like to express special gratitude to Mr. Azeem Mehmood Butt, PhD researcher in Molecular Biology at National Centre of Excellence in Molecular Biology (CEMB), University of the Punjab, Lahore for his assistance in proof reading of manuscript.

References:

[1] Tosato V *et al. Microbiology* 1997 **143**: 3443 [PMID: 9387222]
[2] Raaijmakers JM *et al. Mol Plant-Microbe interact* 2006 **19**: 699 [PMID: 16838783]

[3] Lin TP *et al, Biochim Biophys Acta.* 2005 **1730**: 159 [PMID: 16102594]
[4] Wu CY *et al. J Biol Chem*. 2007 **282**: 5608 [PMID: 17182617]
[5] Ongena M *et al. Appl Microbiol Biotechnol* 2005 **69**: 29 [PMID: 15742166]
[6] Steller S *et al. Chem Biol.* 1999 **6:** 31 [PMID: 9889147]
[7] Tsuge K *et al. Antimicrob Agents Chemother.* 1999 **43**: 2183 [PMID: 10471562]
[8] Liu HL & Hus JP, *Proteomics* 2005 **5:** 2056 [PMID: 15846841]
[9] Sali A & Blundell TL, *J Mol Biol.* 1993 **5**: 779 [PMID: 8254673]
[10] Wilkins MR *et al. Methods Mol Biol*. 1999 **112**: 531 [PMID: 10027275]
[11] McGuffin LJ *et al. Bioinformatics* 2000 **16**: 404 [PMID: 10869041]
[12] Yu CS *et al. Proteins.* 2006 **64**: 643 [PMID:16752418]
[13] Yu NY *et al. Bioinformatics.* 2010 **26:** 1608 [PMID:20472543]
[14] Quevillon E *et al. Nucleic Acids Res*. 2005 **1**: W116 [PMID: 15980438]
[15] Altschul SF *et al. Nucleic Acids Res.* 1997 **25**: 3389 [PMID: 9254694]
[16] Sali A & Blundell TL. *J Mol Biol.* 1993 **5**: 779 [PMID: 8254673]
[17] Wiederstein M & Sippl MJ, *Nucleic Acids Res*. 2007 **35**: W407 [PMID: 17517781]
[18] Benkert P *et al. Nucleic Acid Res*. 2009 **37**: W510 [PMID: 19429685]
[19] Laskowski RA *et al. J Biomol NMR*. 1996 **8**: 477 [PMID: 9008363]
[20] Pettersen EF *et al. J Comput chem*. 2004 **25**: 1605 [PMID: 15264254]
[21] Butt *et al. Bioinformation* 2011 **7**: 303
[22] Butt AM *et al. African Journal of Biotechnology*. 2011 **10**: 38
[23] Castrignano T *et al. Nucleic Acids Res*. 2006 **34**: D306 [PMID: 16381873]

Edited by P Kangueane
Citation: **Batool** et al. Bioinformation 7(8): 384-387 (2011)