Advances in
Biotechnology & Microbiology
ISSN 2474-7637

Juniper Publishers
key to the Researchers

Research Article
Volume 7 Issue 3 – November 2017
DOI: 10.19080/AIBM.2017.07.555715

Adv Biotech & Micro

In Silico Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach

Ruchika Bhat[1,2,3], Rukmankesh Mehra[1,4] and Amit Nargotra[1,5]*

[1]*Discovery Informatics Division, CSIR-Indian Institute of Integrative Medicine, India*

[2]*Department of Chemistry, Indian Institute of Technology, India*

[3]*Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology, India*

[4]*School of Biotechnology, University of Jammu, India*

[5]*Academy of Scientific and Innovative Research, CSIR-Indian Institute of Integrative Medicine, India*

Submission: October 06, 2017**; Published:** November 27, 2017

***Corresponding author:** Amit Nargotra, Discovery Informatics, CSIR-Indian Institute of Integrative Medicine, Canal Road, Jammu 180001, India, Tel: 0191-2569021; Email: ruchikabhat31@gmail.com

Abstract

Streptomyces hygroscopicus subsp. *jinggangensis* 5008, the major producer of anti-fungal aminoglycoside antibiotic jinggangmycin/validamycin (VAL-A), is used to control sheath blight disease of rice plant for more than 40 years. Because of its high level of effectiveness as a fungicide as well as its lack of threat to human and animal health, validamycin has been one of the most important and widely used plant protection reagents against fungus Pellicularia sasakii (*Rhizoctonia solani*) in East Asia. In this work, *S. hygroscopicus* genome has been analyzed particularly with respect to the gene cluster responsible for the production of validamycin to enhance the information about its biochemical pathway, using *in silico* approach. Gene prediction was carried out for the 45kb long genome part reported to be involved in biosynthesis using GeneMark S tool. The research work was aimed at associating the genes/enzymes involved at each step of validamycin synthesis pathway, by the application of protein fold based selection of substrates approach. All the nine reaction steps involved in the biochemical pathway were annotated and the results were confirmed using reported annotated four steps. This work highlights the significance of an *in silico* prediction approach towards annotating novel biochemical pathways

Keywords: Validamycin; S. *hygroscopicus*; Gene prediction; Protein fold based selection; Biochemical pathway and protein folds

Highlights

A. Homology based annotation of genes for characterization of validamycin pathway steps.

B. Identification of probable enzymes working at each step by classifying genes according to known enzyme classes.

C. 3D structure prediction of probable enzymes of the pathway, their validation and the identification of protein folds.

D. Docking studies for the generation and selection of enzyme-substrate complexes.

E. Stability analysis of the proposed substrates-enzyme complexes using MD simulation.

Introduction

Validamycin A (synonyms validamycin, jinggangmycin, VAL-A), which is an antifungal agent, is biosynthesized by the species *Streptomyces hygroscopicus* var. *jinggangensis* 5008, *S. hygroscopicus* var. limoneus and *S. hygroscopicus* 10-22 [1-4]. It is used as crop protectant against fungal action of *Rhizoctonia solani* [2,5]. The mechanism of action of this antifungal drug lies in the inhibition of trehalase enzyme. In fungi, trehalose is a common storage carbohydrate and the enzyme trehalase hydrolyzes it to produce glucose, thus providing energy for the necessary cellular mechanism of fungus. Due to the inhibition of trehalase enzyme, fungus cells are deprived of glucose and eventually die off starving [6-8].

Validamycin belongs to the large family of sugar-derived microbial secondary products. There are various other C7N aminocyclitols natural products secondary metabolites like pyralomicin (used as antibiotics), cetoniacytone (an antitumor agent) and validamycin [9]. Based on the fact that the initial step of cyclization of sedoheptulose-7-phosphate to give 2-epi-5-epi-valiolone remains common, with further incorporation of 2-epi-5-epi-valiolone into various secondary metabolites like acarbose (which is used as an anti-diabetic drug) and validamycin (an antifungal drug), some similarity in their biochemical pathways have been suggested [9-11]. Compounds like 2-epi-5-epi-valiolone, 5-epi-valiolone, valienone, and validone on the basis of experimentation carried out using isotopically labeled precursors, were reported by Floss and co-workers as the intermediates of biosynthetic machinery of validamycin [4,10].

Bioinformatics through its tools and data sets has helped the researchers in various ways to reveal biological mysteries at the gene and protein level. Genome annotation, structural organization, functional proteomics etc. are now a part of exploratory science and are used to answer major biological problems. The data provided by various bioinformatic databases regarding protein folds and domains help in efficiently understanding the biological role of an enzyme, based on which we can predict their mode as well as site of action. Although this won't be wrong to say that bioinformatics just gives a probabilistic solution and in vivo experimentations prove it, but the time saved to screen out the neighborhood path for the solution makes the *in silico* approach appropriate for providing a rational approach to biological research [12-14].

A Study of full genome of species *Streptomyces hygroscopicus* var. jinganggensis revealed that a 45kb region of the *S. hygroscopicus* 5008 chromosome was responsible for the production of secondary metabolite validamycin [1]. Here in, in an attempt to answer the probabilistic enzymatic machinery working behind validamycin pathway, we analyzed the whole cluster reported responsible for the production of validamycin using *in silico* approach. The interpretation of information about domains and the biochemical reaction steps, for which they could be working, helped to acquire an in-depth understanding of the pathway mechanism. Utilizing this knowledge of protein domain regions within the enzymes improved our filtering criterion of the suitable enzyme among the pool of other enzymes present in the gene cluster. Validation of the methodology was done by comparing the results obtained by *in silico* approach to those reported in literature for the annotated steps. The analysis revealed some insights about the probable enzymes working at each biochemical step. This work would help in understanding the application of bioinformatics tools and techniques along with understanding the role of fold and domain analysis for revealing the hidden pathways and also to annotate the existing ones. However, this also opens new options for many wet lab experimentations, to prove and explore the probable pathway suggested as a result of this study.

Methodology

Collection of sequence data and gene annotation

The whole genome of S.hygroscopicus and the sequence of the reported gene cluster responsible for validamycin production [2] were taken from the Genbank [15]. Genome annotation of the 45kb gene cluster was done using Genemark S program [16]. Genemark S is a self-training method for prediction of gene-starts in microbial genomes and for finding sequence motifs in regulatory regions. This program was also used to predict the number of genes and related data of the reported gene cluster responsible for validamycin production. The predicted genes were observed for the strands (plus or minus), their lengths, open reading frames they belong to and the protein sequences they code for. To check if the genes cluster contained any rRNA or tRNA forming genes, the RNAmmer 1.2 server [17] and the tRNAscan SE 1.21 [18] were used. RNAmmer predicts 5s/8s, 16s/18s, and 23s/28s ribosomal RNA in the genome sequences and tRNAscan-SE identifies transfer RNA genes in a DNA sequence, both of these are freely available online.

Homology search and structural annotation

All the identified genes from the cluster were used for homology search using BLASTN program [19]. The function of the top hits reported showing maximum similarity with the query sequence was considered as the probable function of the sequence. Besides this, HMMER [20], an online tool to find the probable function of a gene was also used wherein the amino acid sequence of all the genes of the cluster were used as input to find the probable function of the gene. In HMMER, the protein sequences were searched with respect to UniProtKB database using phmmer as the search mode.

All the Genemark S predicted genes of the cluster were then analyzed for their respective location within the whole genome of *Streptomyces hygroscopicus* on the basis of alignment obtained using BLAST 2 sequence tool [19].

Biochemical pathway characterization

Each step of the proposed biochemical pathway [21] of validamycin production was analyzed for possible metabolic reaction so that the probable enzyme function at each step could be hypothetically deduced. The enzyme found to have the similar function that was required for a particular reaction step to occur was chosen as the probable candidate enzyme for that step and classified under one of the six major enzyme class as per the standard enzyme nomenclature [22]. A hypothesis of enzymatic machinery working behind the pathway was generated and analyzed thoroughly using computational studies.

Computational studies for identification of suitable enzyme from the candidate enzymes

3D structure prediction: The 3D structures prediction of all the genes from the reported cluster was carried out using Phyre2 [23]. Phyre2 is an updated version of Phyre that uses

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

fold recognition technique for predicting the structure and/or function of protein sequence. It is used for detecting remotely homologous structures. It makes profiles (or PSSMs) generated by PSI-Blast for query sequence and the sequences of the known structure. It runs profile-profile matching algorithm together with predicted secondary structure matching. Phyre makes changes to the backbone of the known structure (template) only when modeling of insertions or deletions is needed, by searching its loop library for compatible loops. The similarity as well as query coverage values of the target with respect to its template were analyzed. Procheck module of the SAVS server [24] was used for validation of the generated 3D models using its Ramachandran plot [25] values to check the stereochemical properties of each model. The protein models that were shorter in length were removed.

Binding site identification of the predicted structures: The models were analyzed for their folds and domains using Pfam [26] and SCOP databases [27]. The Pfam database has information related to protein domains and families. It consists of two components: Pfam-A and Pfam-B. Pfam-A consists data of manually curated families. Pfam-B families are automatically generated families which are useful for identifying functionally conserved regions when no Pfam-A entries are found. We can look up the domain organization of a protein model using View a Structure module of the Pfam database. Also we used SCOP database to collect information regarding domains as SCOP (The Structural Classification of Proteins) database contains manual classification of protein structural domains based on their structures and sequences similarities. Collection of available domain regions in the templates of the protein models was done.

Binding site predictions on the basis of volume of void space available along the modeled structure of proteins were done using Sitemap module of Schrodinger [28,29]. Five binding sites were predicted using the Sitemap tool. Structure superposition and sequence alignment of each protein model with its template was manually done to select only those binding sites that lie in the domain region. The sites lying in the domain aligned region, showing similar function identified from the BLASTN [19] homology searches, were used for docking.

Molecular docking studies: All the molecular docking studies were carried out using Schrodinger [28] molecular modeling software. Structures of all the ligands (substrates and products) were built using maestro [30]. The ligands were prepared using LigPrep [31] in which their steroisomeric, tautomeric and ionization states were determined. Various conformations of these prepared ligands were generated using ConfGen [32] module.

The docking studies of all the predicted enzymes were carried out with the substrate of the respective position of the proposed biochemical pathway. The predicted protein structures were prepared using Protein Preparation Wizard in which bond orders were given, hydrogens were added, hydrogen bonds were optimized and energy minimization of the protein structure was done. 3D grid was generated on each identified binding site and flexible docking of the generated conformers was performed onto this grid using Extra Precision scoring function of GLIDE [33,34]. Based on the dock score and hence the binding affinity value of each complex, the most probable enzyme was proposed for carrying out the reaction.

Stability evaluation of the docked complexes: In order to evaluate the stability of the docked complexes and proposing the substrate for the enzymes of the pathway steps, molecular dynamics simulation studies were performed on seven protein-substrate complexes. Desmond 3.1 implemented in Schrodinger 2012 was used for this study [35]. The preparation of the system for each complex was done by solvating protein-substrate complex in simple point charge (SPC) water molecules. The system was neutralized by replacing solvent molecules with the counter ions. Each of the prepared systems was relaxed and equilibrated using minimization and then simulated for 5 nanoseconds (ns) using the multistep protocols devised in Desmond. The minimization was performed by applying a hybrid of steepest descent and the limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) methods with a maximum of 2000 iterations. Two steps of minimization were carried out, first with restraints on solute and second without any restraints. Then a short simulation of 12 picoseconds (ps) was done at 10K temperature with restraints on solute heavy atoms using NVT ensemble (constant number of atoms N, volume V and temperature T). The system was further simulated for 12ps in NPT ensemble at 300K with restraints on heavy solute atoms. The final step of equilibration was performed without any restraints for 24ps using NPT ensemble at 300K. Berendsen thermostats and barostats were used for maintaining the temperature and pressure conditions respectively. Each of the equilibrated systems was simulated for 5ns in NPT ensemble with Nose-Hoover thermostat at 300K and Martyna-Tobias-Klein barostat at 1.01325 bar pressure. The time step of 2 femtoseconds (fs) was used with a recording interval of 1.2ps for energy and 4.8ps for trajectory. The short range electrostatic and vanderwall interactions were truncated at 9.0Å and the long-range electrostatic interactions were determined using Smooth particle-mesh Ewald method (PME). Each complex was analyzed for energy plots and RMSD of the protein backbone with respect to the simulation time. The root mean square fluctuations (RMSF) of the backbone were analyzed for each residue.

Results and Discussion

Collection of sequence data

The genome sequence taken from Genbank entry with name *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008, complete Genome and having Accession No.: CP003275.1 was found to be 10,145,835 bp long. The sequence was taken from Bioproject No.: PRJNA81087. The source organism was *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008. It consisted of

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

full length genome sequence of the S.hygroscopicus subsp. *jinggangensis* strain 5008.

The gene cluster found to be responsible for producing validamycin was taken from Genbank entry with name *Streptomyces hygroscopicus* subsp. *jinggangensis* strain 5008 validamycin cluster, genomic sequence and having Accession No.: DQ164098.1 was found to be of 45164 bp linear DNA length. The source organism was *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008. The 45kb sequence contained the 16 structural genes that are reported to run the biosynthetic pathway of validamycin [1].

Gene prediction using Genemark S tool

Initial genome annotation was done using Genemark S program [16], which predicted the genes that were present in the gene cluster. Genemark S analysis was done for the gene cluster sequence which predicted a total of 42 genes as shown in Figure 1. The range of gene lengths varied from 45bps to 2536bps. The class 1 was observed for all predicted 42 genes, showing that all the genes belong to bacterial family. The GC percent of the sequence was reported to be 71.8% in NCBI and as per the Genemark software it was approximately same with 69%.

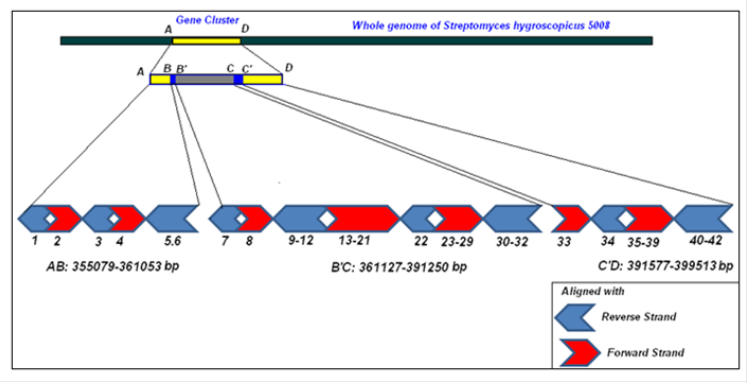

Figure 1: Diagrammatic representation of gene cluster within the full genome showing the location of all Genemark S predicted 42 genes from 1 to 42 as per their alignment found using the BlastN tool. The direction of the arrows shows their location on sense or antisense strand. Reverse arrow means the alignment is done with the reverse strand and the forward arrow means alignment with the forward strand. The region AD denoted the chunk of gene cluster (45kb) taken for gene prediction. The region AB contains gene number 1 to 6 whereas the region B'C contains all the structural genes reported in [1] and region C'D contained the rest genes, namely gene number 33 to 42. The gene numbers obtained as the result of gene prediction by GeneMark S are hereafter used to refer to a particular gene or its product i.e. enzyme depending on their context of reference, nucleotide or protein respectively.

Each gene of the predicted 42 genes of the cluster was in the coding region and none of these were rRNA forming genes or tRNA forming genes as was verified using RNAmmer 1.2 [17] server and tRNAScanSE server [18] as shown in Figure 1.

Homology searches

Table 1: Blast and HMMER results for the 42 genes predicted by Genemark S tool. The organisms with the % id and query coverage that was showing homology are also tabulated. The homological searches by both tools suggested similar functions for most of these genes. Also the last column shows the classification of each gene into main 6 enzyme categories based on their functions predicted by both tools.

Gene Name	Organism	% Identity	QC%	Function Predicted by Blast	Function Predicted by HMMER	Enzyme Classification
Gene1	*Salinispora tropica*	97	100	Transposase	Transposase	Transferases
Gene2	*Burkholderia*	86	29	Hydrolase	Predicted protein	Hydrolase
Gene3	*Streptomyces cattleya*	87	100	Integrase	Integrase	Transferase
Gene4	*Streptomyces cattleya*	91	98	Transposase	Winged helix-turn helix/ endonuclease	Transferase
Gene5	*Streptomyces avermitilis*	95	100	Transposase	Transposase	Transferase
Gene6	*Kitasatospora setae*	85	99	Transposase	Endonuclease	Transferase
Gene7	*Streptomyces davarvensis*	93	79	Hypothetical protein	Hypothetical protein	----
Gene8	*Variovorax paradoxis*	91	29	Peptidoglycan binding lysine domain	No hits	Hydrolase

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

Gene9	*Actinosynnema mirum*	89	81	Alcohol dehydrogenase GroES domain protein	Dehydrogenase	Oxidoreductase
Gene10	*Actinosynnema mirum*	85	95	Aminotransferase classIII	Aminotransferase	Transferases
Gene11	*Actinosynnema mirum*	86	89	Glycosyl transferase family 20	Glycotransferase family	Transferases
Gene12	*Actinosynnema mirum*	91	77	dTDP-4-dehydrorhamnose	RmlD substrate binding domain	Oxidoreductase
Gene13	*Pseudomonas sp*	90	66	Unknown protein	Hypothetical protein	---
Gene14	*Actinoplanes*	89	65	2-epi-5-epi-valiolone synthase	3-dehydroquinate synthase	Lyase
Gene15	*Actinoplanes*	77	96	Nucleotidylyltransferase,cyclitol kinase	Nucleotidyl transferase	Transferase
Gene16	*Streptomyces glaucescens*	69	72	Nucleotidylyltransferase,cyclitol kinase	ROK family	Transferase
Gene17	*Actinoplanes*	75	21	Putative Glyoxylase/ dioxygenase	Glyoxylase like domain/ dioxygenase	Lyase / oxidoreductase
Gene18	*Amycolatopsis mediterrane*	91	28	Oxidoreductase	Non-haem dioxygenase	Oxidoreductase
Gene19	*Streptomyces davawensis*	88	99	Oxidoreductase	Oxidoreductase	Oxidoreductase
Gene20	*Cellulomonas flavigena*	65	50	Glycosyl transferase family 2	Glycosyl transferase	Transferase
Gene21	*Kitasatospora setae*	67	93	Putative major facilitator superfamily transporter	MSF	Hydrolase
Gene22	*Nocardia farcinica*	80	17	Putative ATP/GTP-binding protein	Polyketide biosynthesis	Transporter Protein
Gene23	*Streptomyces griseus*	76	95	Glycosyl hydrolase	Glycosyl hydrolases	Hydrolase
Gene24	*Amycolatopsis mediterranei*	91	37	Oxidoreductase	Non-haem dioxygenase	Oxidoreductase
Gene25	*No significant similarity found*	N.A	N.A	---	Putative oxidoreductase	Oxidoreductase
Gene26	*Conexibacter woesei*	87	70	Hypothetical protein	Unknown protein	---
Gene27	*Streptosporangium roseum*	83	21	Hypothetical protein	Predicted metal binding intergral membrane protein	---
Gene28	*N.A*	N.A	N.A	Unknown	No hit	---
Gene29	*Paenibacillus mucilaginosus*	76	7	Hypothetical protein	SMI1/KNR4 family (SUKH-1)	---
Gene30	*Streptomyces flavogriseus*	94	98	Succinate dehydrogenase, flavoprotein subunit	Predicted protein	Oxidoreductase
Gene31	*Streptomyces avermitilis*	96	93	Phosphatase	StageII sporulation protein(SpoIIE)	Hydrolase
Gene32	*Streptomyces davawensis*	83	78	Histidine kinase	GAF domain	Transferase
Gene33	*Acidothermus cellulolyticus*	80	47	Peptidase/ kinase	No hit	Transferase/ Hydrolase
Gene34	*Streptomyces albus J1074*	82	91	Phosphatidate cytidylyltransferase	ICE-like protease p20 domain containing protein	Transferase
Gene35	*Streptomyces griseus*	94	99	Hypothetical protein	Hypothetical protein	--
Gene36	*Streptomyces avermitilis*	90	100	Putative tellurium resistance protein	Tellurium resistance	--
Gene37	*Streptomyces venezuelae*	80	54	Twin-arginine translocation protein TatA	MttA/Hcf106 family	Functions as transporter protein

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

Gene38	*Streptomyces bingchenggensis*	90	100	Putative integral membrane protein	PAP2 superfamily	Hydrolase
Gene39	*Streptomyces violaceusniger*	90	99	Hypothetical protein	Conserved hypothetical protein	---
Gene40	*N.A*	N.A	N.A	Hypothetical protein	Unknown	---
Gene41	*Streptomyces scabiei*	93	83	Esterase	Alpha/beta hydrolase fold	Hydrolase
Gene42	*Streptomyces flavogriseus*	82	89	Cytochrome P450	Cytochrome	Oxidoreductase

Homology searches for each of these 42 genes from the cluster was done using BLASTN program [19] to predict their probable function. The function associated with the top hits considering their percent similarity and coverage length along with significant E value (max. 0.001) were reported as the probable function of the sequence, as shown in Table 1.

Amino acid sequences obtained from the GeneMark S gene prediction tool were used as input to the HMMER online tool to find the probable function and the results obtained were in agreement with the result of BLASTN searches. In HMMER, the protein sequences were searched with respect to UniProtKB database using phmmer as the search mode as shown in Table 1.

Structural annotation

Annotation using GeneMark S tool resulted a list of 42 predicted genes when the input sequence was 45kb long gene cluster sequence. The position of these 42 predicted genes inside the whole genome of *Streptomyces hygroscopicus* 5008 was found using Blast 2 sequence alignment tool for each gene [19]. The overall region involved was noted in terms of base pair length as shown in Table 1. This gave us an idea about the structural location of the genetic cluster of genes responsible for the validamycin pathway.

Based on the criteria of the alignment, a diagramatic view of the location of the gene cluster was created. This simplified the visualization of the location of the start and end regions of the gene cluster inside the whole genome. The cluster belonged to the region of 355079bps to 399513bps of the whole genome. The gene 1 marked the begining of the sequence cluster at 355079bp followed by genes 2,3,4,5 and 6 in the patch ranging from 355079bp till 361053bp as shown in Figure 1 in region AB. A larger portion B'C of the gene cluster was found in a continous stretch of 361127bp to 391570bp was containing all the 16 reported structural genes. The C'D patch of the Figure 1 streching from 391577bp to 399513bp contained rest of the genes from gene number 33 to gene number 42.

Sixteen structural genes, namely val A to val Q, reported to be responsible for the production of validamycin [1], were aligned with each of the 42 genes of the cluster to identify their particular location, i.e. their corresponding gene number within the 42 predicted genes. Blast 2 sequence tools was used to find the match between the GenBank entry of each val gene with each gene among the 42 predicted genes and the results obtained are discussed in Table 2.

Table 2: Domains present in the template (with Pfam Ids) having similar functions to probable functions of our gene products.

Sr. No.	Gene No.	Domains in Template	Pfam ID
1	Gene3	Putative snoRNA binding Domain /NoSIC	PF08060/PF01798
2	Gene4	Transposase	PF01359
3	Gene9	Alcohol dehydrogenase/GroES like domain/zinc binding domain	PF08240/ PF00107/ PF13602
4	Gene10	Aminotransferase	PF00202
5	Gene11	Glycol-transf_2	PF00982
6	Gene12	Epimerase/dehydrogenase Aldose1 epimerase/ Short chain dehydrogenase/ dehydrogenase/ isomerase	PF01370/ PF13950/ PF01073/ PF00106/ PF01263/ PF07993
7	Gene14	Alcohol dehydrogenase/ Iron containing Alcohol dehydrogenase/ 3-dehydroquinate synthase	PF00465/ PF13685/ PF01761
8	Gene15	NTP transferase	PF00483
9	Gene16	ROK family domain	PF00480
10	Gene17	Glyoxylase / bleomycin resistance protein/ dehydrogenase family	PF00903
11	Gene18	Non-haem dioxygenase/ 2OG-Fe(II) oxygenase superfamily	PF14226/ PF03171

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

12	Gene19	Oxidoreductase family Rossmann fold/ alpha,beta domain	PF01408/ PF02894
13	Gene20	Glycol-transf_2	PF00535
14	Gene21	MSF/sugar transport protein	PF13347/ PF07690
15	Gene23	Glucodextranase domain N/ glycoside hydrolase family 15	PF09137/ PF00723
16	Gene24	Non-haem dioxygenase/ 2OG-Fe(II) oxygenase superfamily	PF14226/ PF03171
17	Gene31	GHKL domain	PF13581
18	Gene38	PAP2 superfamily	PF01569
19	Gene41	Carboxylesterase family/ alpha-beta hydrolase fold	PF00135/ PF07859
20	Gene42	Cytochrome P450	PF00067

Biochemical pathway characterization

Pathway steps were analyzed and checked for possible metabolic reaction occurring at each step for validamycin biosynthesis so that the probable enzyme working behind it could be hypothetically deduced. There were in all 9 major steps in the pathway as shown in Figure 1. Each reaction of the pathway is discussed in detail (Figure 1) regarding the enzymatic requirement necessary for their catalysis (please refer Supplementary Information).

Candidate enzyme detection/discovery

The results of homology searches helped in elucidating the particular feature of each gene product and that feature was then categorized under the 6 main enzyme classifications. Also the reactions were also generalized based on the same criteria.

However, the proposed pathway steps required enzymatic action belonging to each class of enzyme classification. A total of twelve out of the 42 genes lacked any domain significantly similar to any enzyme class. As nine out of these twelve genes were hypothetical proteins and the rest were having secretory and translocational functions. Thus, these gene products that were not falling under any of the enzyme class required in the pathway reactions were removed from further analysis.

Modeling of probable enzymes

The 3D structures of all the candidate enzymes for each step of the hypothetical pathway were generated. The server used to generate the 3D protein models was Phyre2 and each model was checked for its stability through molecular dynamics simulations of 1ns. The protein model of the gene 16 was not stable as revealed by the stability analysis of the protein structure through the simulation studies. Therefore, the protein sequence of gene 16 was remodeled using Protein Structure Prediction Server (PS2 server) [36]. PS2 is an automated homology modeling server which uses an effective consensus strategy by combining PSI-BLAST, IMPALA, and T-Coffee in both template selection and target-template alignment and it uses MODELLER package to build the final three dimensional structures. The new structure of gene 16 was found to be stable.

Table 3: Candidate enzyme classification along with the candidate gene numbers for each reaction step of the biochemical pathway of validamycin.

Reaction Step Number	Enzyme Mechanism Required	Function/Feature of Enzyme needed at particular step	Probable Candidate Enzyme Name (Gene Number)	No. of Candidate Enzymes
Step1*	Lyase	Synthase	Gene 14, 17	2
Step 2	Isomerase	Isomerase	Gene 12	1
Step 3	Lyase	Dehydratase	Gene 14, 17	2
Step 4*	Transferase	Kinase	Gene 3,4,10,11,15,16,20	7
Step 5	Oxidoreductase	hydrogenase	Gene 9,12,17,18,19,24,42	7
Step 6	Oxidoreductase	hydrogenase	Gene 9,12,17,18,19,24,42	7
Step7*	Transferase	Kinase	Gene 3,4,10,11,15,16,20	7
Step 8	Hydrolase & Transferase	Phosphatase & aminotransferase	Gene3,4,9,10,11,15,16,17, 20,21,23,31,38,41	14
Step 9*	Transferase	Kinase	Gene 3,4,10,11,15,16,20	7

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

The generated models were analyzed with respect to the similarity and query coverage with the template protein used for model building. Also, the Ramachandran plot values for each model were calculated to check the stereochemical properties as discussed in Table 3. Further, ten out of these 30 enzymes were neglected as they were having shorter modeled regions less than 40bps, because of which their binding sites were not generated. Apart from the model length the structure stability (disallowed region of approx. less than 2.6%) was also taken up as a criterion for filtering out the vital enzyme models responsible for the biosynthetic pathway. The ones that were removed were gene numbers 1,2,5,6,8,25,30,32,33 and 34. The remaining 20 candidates were analyzed for domain constituency of their templates.

Domain analysis of templates

The models were analyzed for their folds and domains using databases like Pfam and SCOP. The structure alignment of each model was done with the template to find the position of domains in each model as shown in Table 2. All the domains available in templates were searched using SCOP and Pfam databases and the ones having functionally similarity as required by the modeled enzymes were noted down along with their structural location. Based on the enzyme classification by homology searches and domain based selection criteria, a list of candidate enzymes probably working at each step were tabulated as shown in Table 3.

The region of the modeled proteins having the structural alignment with functionally similar domains were highlighted and manually visualized to find the effective binding site. Each model was then evaluated for their probable binding sites using the Sitemap tool of Schrodinger software. Five to six binding sites were predicted for all these 20 models. The sites lying in the domain aligned region, showing similar function identified from the BLASTN homology searches, were used for docking.

In silico approach to support the results

The docking study for protein-substrate affinity of each enzyme under consideration was done using GLIDE-Ligand docking and their binding affinities were analyzed based on the obtained docking scores. The resultant docking score of enzyme, found to have minimal binding energy values after docking, hinted towards maximum possibility of occurrence of that enzyme than others. This helped in proposing the final biochemical pathway of validamycin.

Table 4: Docking score obtained with different candidate enzymes for the annotated reaction step 1.

Substrate	XP GScore for Gene14 Product	XP GScore for Gene17 Product
D-sedoheptulose-7-phosphate	-9.537	-8.024

Analyzing the reaction step 1 in detail: Based on the binding score, as shown in Table 4, obtained by docking the two candidate enzymes namely gene 14 and gene 17, the probable enzyme working as Lyase in step 1 is product of gene 14. The XP GScore of gene14 enzyme product for the substrate is better as compared to the gene 17 enzyme product, which is in agreement with the already annotated result for the step [4]. The ligand interaction of gene14 and D-sedoheptulose-7-phosphate is shown in the Figure 1 and docking image is shown in Figure 1. The OH groups are interacting with various residues like ASP, LYS, ILE and ARG. A total of eight interactions of hydrogen bonds (Table 4) are occurring between the ligand and the receptor.

Analyzing the reaction step 2 in detail: The activity required for the step to occur was that of an isomerase enzyme. Among all the 42 gene products i.e. 42 enzymes only gene 12 was having a functional domain matching to an epimerase domain and showed a possible isomerase character. So, we propose that enzyme synthesised by gene 12 is responsible for the step no. 2 of the biochemical pathway of validamycin synthesis.

Table 5: Docking score obtained with different candidate enzymes for the unannotated reaction step 3.

Substrate	XP GScore for Gene14 Product	XP GScore for Gene17 Product
5-epi-valiolone	-6.319	-5.515

Analyzing the reaction step 3 in detail: The dock scores as shown in Table 5 for the reaction step 3 suggest that the probable enzyme working as Lyase in step 3 is product of gene14 as the XP GScore of gene14 enzyme product for the substrate is better as compared to the gene 17 enzyme product.

The ligand Interaction of gene14 and 5-epi-valilone is shown in the Figure 1 and docking image is shown in Figure 1. The OH groups are interacting with various residues like ASN, LYS, GLU and ASP. A total of six interactions of hydrogen bonds (Table 5) are occurring between the ligand and the receptor.

Analyzing the reaction step 4 in detail: The dock scores obtained for each candidate enzyme for reaction step 4 are shown in Table 6. The binding affinity analysis suggests that the probable enzyme working as transferase in step 4 is product of gene 16. Since the XP GScore of gene16 product for the substrate of the reaction is high as compared to the other candidate

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

enzymes, so enzyme of gene 16 is the enzyme working behind step number 4 which is in agreement with the already annotated data [36,37]. The ligand interaction of gene16 and valienone is shown in the Figure 1 and docking image is shown in Figure 1. The OH groups are interacting with various residues like ASN, ALA, ILE and GLY. A total of six interactions of hydrogen bonds (Table 6) are occurring between the ligand and the receptor.

Table 6: Docking Scores obtained with different candidate enzymes for the annotated reaction step 4.

Substrate	XP GScore for the Enzymes of Gene Number							
	Gene1	Gene3	Gene4	Gene10	Gene11	Gene15	Gene16	Gene20
Valienone	-4.754	-6.578	-6.501	-6.229	-6.232	-5.830	-7.919	-6.547

Analyzing the reaction step 5 in detail: The dock scores as shown in Table 7 for the reaction step 5 suggest that the probable enzyme working as oxidoreductase in step 5 is enzyme product of gene 12. Since the XP GScore of gene 12 for the substrate is high as compared to other candidate enzymes, so gene 12 becomes the proposed enzyme for step number 5. The ligand Interaction of gene12 and valienone-7-phosphate and docking image is shown in Figure 1. The OH groups are interacting with various residues like ALA, LEU, SER and GLY. A total of six interactions of hydrogen bonds (Table 7) are occurring between the ligand and the receptor.

Table 7: Docking Score obtained with different candidate enzymes for the unannotated reaction step 5.

Substrate	XP GScore for the Enzymes of Gene Numbers						
	Gene9	Gene12	Gene17	Gene18	Gene19	Gene24	Gene42
valienone-7-phosphate	-5.454	-8.036	-5.551	-5.294	-4.366	-5.119	-5.346

Analyzing the reaction step 6 in detail: The dock scores generated for each candidate enzyme for step 6 as shown in Table 8 suggest that the probable enzyme working as oxidoreductase in step 6 is enzyme product of gene 19. Since the XP GScore of gene 19 for the substrate is high as compared to other candidate enzymes, so gene 19 becomes the proposed enzyme for step number 6. The ligand Interaction of gene19 and valienone is shown in the Figure 1 and docking image is shown in Figure 1. The OH groups are interacting with various residues like ALA, LYS and GLY. A total of three interactions of hydrogen bonds (Table 8) are occurring between the ligand and the receptor.

Table 8: Docking Score obtained with different candidate enzymes for the unannotated reaction step 6.

Substrate	XP GScore for the Enzymes of Gene Numbers						
	Gene9	Gene12	Gene17	Gene18	Gene19	Gene24	Gene42
Valienone	-4.244	-5.606	-5.702	-4.474	-7.133	-4.222	-5.746

Analyzing the reaction step 7 in detail: Based on the dock score data generated as shown in Table 9 for the XP docking of each candidate enzymes with validone as ligand it could be suggested that the probable enzyme working as transferase in step 7 is product of gene 16. Since the XP GScore of gene16 product for the substrate of the reaction is high as compared to the other candidate enzymes and comparable to gene 11, so enzyme of gene 16 is the enzyme working behind step number 6 which is in agreement with the already annotated data [37]. The ligand Interaction of gene16 and validone is shown in the Figure 1 and docking image is shown in Figure 1. The OH groups of the ligand are interacting with various residues like ASN, ILE and GLU (Table 9).

Table 9: Docking Score obtained with different candidate enzymes for the annotated reaction step 7.

Substrate	XP GScore for the Enzymes of Gene Number							
	Gene1	Gene3	Gene4	Gene10	Gene11	Gene15	Gene16	Gene20
Validone	-5.067	-4.644	-4.760	-4.500	-5.275	-4.052	-8.466	-4.247

Analyzing the reaction step 8 in detail: It was seen that the step required a phosphatase as well as aminotransferase activities to catalyze the reaction. The aminotransferase was only enzyme 10 as per the homology searches and domain analysis. Enzyme 31 had BLAST result matching the feature of phosphatase also its domain analysis showed a transferase domain which is an additional advantage as the step requires both the functions of transferase and hydrolase enzyme. Thus, enzyme 10 and enzyme 31 are the proposed enzyme candidates for the step 8 of the validamycin biosynthetic pathway.

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

Table 10: Docking score obtained with different candidate enzymes for the annotated reaction step 9.

Substrate	XP GScore for the Enzymes of Gene Number							
	Gene1	Gene3	Gene4	Gene10	Gene11	Gene15	Gene16	Gene20
ValidoxylamineA	-6.555	-5.426	-7.663	-8.868	-10.090	-8.756	-10.914	-11.345

Analyzing the Reaction Step 9 in detail: The dock scores as shown in Table 10 for the reaction step 9 suggest that the probable enzyme working as transferase in step 4 is product of gene 20. Since the XP GScore of gene20 product for the substrate of the reaction is high as compared to the other candidate enzymes, so enzyme of gene 20 is the enzyme working behind step number 6 which is in agreement with the already annotated data [1]. The ligand Interaction of gene20 and validoxylamine A and docking image is shown in Figure 1. The OH groups are interacting with various residues like GLU, THR and ARG. A total of five interactions of hydrogen bonds (Table 10) are occurring between the ligand and the receptor.

The verification of the methodology was done by comparing the results obtained through our *in silico* approach with the results reported for the annotated steps proven through wet lab experimentations. The results obtained by both methods were in agreement with each other which imply that using this fold based *in silico* approach for analysis of pathway has driven us to a more concrete platform of obtaining dependable results. The verification suggests that the results obtained for the unannotated steps, using this fold based approach in our *in silico* analysis of enzymes responsible for the biochemical pathway, would also be reliable.

Molecular dynamics simulation of the docked complexes: Molecular Dynamics (MD) simulations of the seven protein-substrate complexes were carried out. For performing the MD simulation of the gene20 complex in step 9, the domain region of the protein model bound with the substrate was used. The part of the protein between Arg246 and Leu270 was not modeled by the Phyre2 server and analysis of the domain region (using Pfam) of the modeled structure suggested that the domain part has been modeled properly which was lying between Leu270 and Arg420. Sitemap binding site prediction and docking results also suggested that the proposed substrate of the gene20 was binding on the functional domain part of the model and the binding site has no involvement of the other part of the structure (from residues Val8 to Arg246). Therefore the domain region of the gene20 (from Leu270 to Arg420) along with the bound substrate at the binding site was used for carrying MD simulation.

The analysis of the RMSD plots of the protein backbone showed that the protein-substrate complexes of the gene12-step 5, gene14-step 1, gene14-step 3, gene19-step 6 and gene20-step 9 were stable as shown in Figure 1 respectively. Stable trajectories of the RMSD were obtained below 3.5Å for these complexes. MD simulations were performed for the new gene 16 model complexed with the substrates of the reaction step 4 and step 7. These two complexes showed stable trajectories of RMSD below 4Å (Figure 1) that demonstrated the stability of these protein substrate complexes.

Conclusion

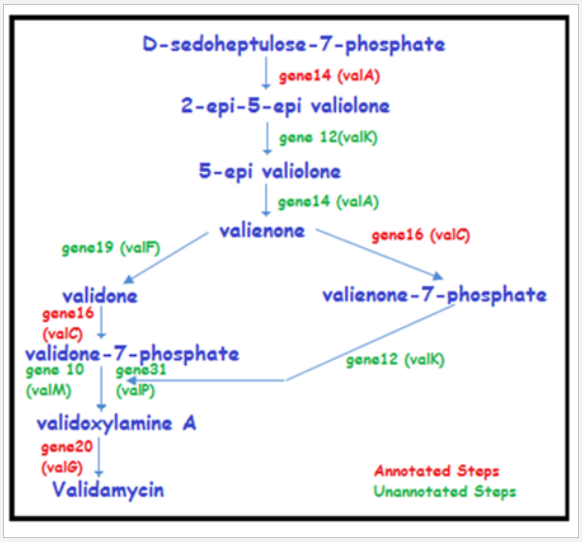

Figure 2: Proposed final functionally annotated biochemical pathway for validamycin. Green coloured enzymes are the ones proposed through this work whereas red ones are the already annotated ones.

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

The genes predicted using GeneMark S tool for the reported cluster of *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008 were structurally and functionally annotated. Both homology based search and protein fold based selection criteria was adopted to filter out a pool of probable candidate enzymes from the predicted genes for each one of the nine biochemical reaction steps of validamycin pathway, five of which are not annotated till date. The study was aimed to find enzymes involved in the five unannotated pathway reactions of the nine step long validamycin pathway. Homology based searches provided a direct correlation between the enzymatic function and the biochemical changes required for pathway step 8 to occur, thus gene31 (phosphatase) and gene10 (aminotransferase) were hypothesized to be the enzymatic cause of that reaction step. The protein fold based selection helped in finding the possible enzyme i.e. gene12 for the reaction step 2 of the pathway which required an isomerase activity. The remaining three unannotated steps had more than one candidate enzyme, so an *in silico* based docking approach was used to find the best one among them based on the affinity of enzyme for the substrate. The four previously annotated steps were also analyzed using the same *in silico* protocol and the results corroborated with the reported in vitro studies for these steps, except for gene16 (valC) in reaction step 7. Further MD simulation studies revealed that all the docked protein-substrate complexes were stable. This study proposes five new enzymes namely gene12, gene14, gene19, gene10 and gene31 for the unannotated steps of validamycin pathway (Figure 2).

Acknowledgement

The authors acknowledge the support provided by the projects BSC0117 and GAP0141.

References

1. Bai L, Li L, Xu H, Minagawa K, Yu Y, et al. (2006) Functional analysis of the validamycin biosynthetic gene cluster and engineered production of validoxylamine A. Chem Biol 13(4): 387-397.
2. Singh D, Seo MJ, Kwon HJ, Rajkarnikar A, Kim KR, et al. (2006) Genetic localization and heterologous expression of validamycin biosynthetic gene cluster isolated from *Streptomyces hygroscopicus* var. limoneus KCCM 11405 (IFO 12704). Gene 376(1): 13-23.
3. Singh D, Kwon HJ, Rajkarnikar A, Suh JW (2007) Glucoamylase gene, vldI, is linked to validamycin biosynthesis in *Streptomyces hygroscopicus* var. limoneus, and vldADEFG confers validamycin production in Streptomyces lividans, revealing the role of VldE in glucose attachment. Gene 395: 151-159.
4. Yu Y, Bai L, Minagawa K, Jian X, Li L, et al. (2005) Gene cluster responsible for validamycin biosynthesis in *Streptomyces hygroscopicus* subsp. *jinggangensis* 5008. Appl Environ Microbiol 71: 5066-5076.
5. Zheng L, Zhou X, Zhang H, Ji X, Li L, et al. (2012) Structural and functional analysis of validoxylamine A 7′-phosphate synthase vall involved in validamycin a biosynthesis. PLOS one 7(2): e32033.
6. Asano N, Yamaguchi T, Kameda Y, Matsui K (1987) Effect of validamycins on glycohydrolases of *Rhizoctonia solani*. J Antibiot (Tokyo) 40(4): 526-532.
7. Xue YP, Zheng YG, Shen YC (2005) Preparation of trehalase inhibitor validoxylamine A by biocatalyzed hydrolysis of validamycin A with honeybee (Apis cerana Fabr.) BETA-glucosidase. Appl Biochem Biotechnol 127(3): 157-171.
8. Li H, Su H, Kim SB, Chang YK, Hong SK, et al. (2011) Enhanced production of trehalose in Escherichia coli by homologous expression of otsBA in the presence of the trehalase inhibitor, validamycin A, at high osmolarity. J Biosci Bioeng 113(2): 224-232.
9. Mahmud T, Flatt PM, Wu X (2007) Biosynthesis of unusual aminocyclitol-containing natural products. J Nat Prod 70(8): 1384-1391.
10. Dong H, Mahmud T, Tornus I, Lee S, Floss HG (2001) Biosynthesis of the validamycins: identification of intermediates in the biosynthesis of validamycin A by *Streptomyces hygroscopicus* var. limoneus. J Am Chem Soc 28: 2733-2742.
11. Naganawa H, Hashizume H, Kubota Y, Sawa R, Takahashi Y, et al. (2002) Biosynthesis of the cyclitol moiety of pyralomicin 1a in Nonomuraea spiralis MI178-34F18. J Antibiot 55: 578-584.
12. Marina C (2002) Bioinformatics: Bringing it all together technology features. Nature 419: 751-757.
13. Edwards JS, Ibarra RU, Palsson BO (2001) *In silico* predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19(2): 125-130.
14. Graaf DC, Vermeulen NP, Feenstra KA (2005) Cytochrome P450 *in silico*: An integrative modeling approach. J Med Chem 48(8): 2725-2755.
15. Benson DA, Cavanaugh M, Clark K, Mizrachi IK, Lipman DJ, et al. (2013) GenBank. Nucleic Acids Res 41: D36-D42.
16. Besemer J, Alexandre L, Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29(12): 2607-2618.
17. Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T, et al. (2007) RNammer: consistent annotation of rRNA genes in genomic sequences. Nucleic Acids Res 35(9): 3100-3108.
18. Lowe TM, Eddy SR (2005) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5): 955-964.
19. Mount DW (2004) Using the basic local alignment search tool (BLAST), Cold spring harbor laboratory press, USA.
20. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-W37.
21. Karp PD, Riley M, Paley SM, Toole PA (2002) The MetaCyc database. Nucleic Acids Res 30(1): 59-61.
22. Moss GP (1992) Enzyme nomenclature, world wide web updated version of 1992, School of biological and chemical sciences, Queen mary university of london, Mile end road, London, UK.
23. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the web: a case study using the Phyre server. Nat Protoc 4(3): 363-371.
24. (2012) SAVS: Version.
25. Sheik SS, Sundararajan P, Hussain AS, Sekar K (2002) Ramachandran plot on the web. Bioinformatics 18(11): 1548-1549.
26. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Res 40: D290-D301.
27. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4): 536-540.
28. (2012) Schrodinger, Version.

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715

29. (2012), Site Map, version 2.6, Schrödinger, New York, USA.

30. (2012), Maestro, Version 9.3, Schrödinger, New York, USA.

31. (2012), LigPrep Version 2.5, Schrödinger, New York, USA.

32. (2012) ConfGen Version 2.3, Schrödinger, New York, USA.

33. Friesner RA, Banks Jl, Murphy RB, Halgren TA, Klicic JJ, et al. (2004) GLIDE: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47: 1739-1749.

34. (2012) GLIDE Version 5.8, Schrödinger, New York, USA.

35. (2012) Desmond molecular dynamics system, version 3.1, DE Shaw Research, New York, USA.

36. Chen CC, Hwang JK, Yang JM (2006) (PS) 2: Protein structure prediction server. Nucleic Acids Res 34: W152-157.

37. Minagawa K, Zhang Y, Ito T, Bai L, Deng Z, et al. (2007) A new type of C7-Cyclitol kinase involved in the biosynthesis of the antifungal agent validamycin A. Chembiochem 8(6): 632-641.

DOI: 10.19080/AIBM.2017.07.555715

How to cite this article: Ruchika B, Rukmankesh M, Amit N. *In Silico* Characterization of Various Steps of Validamycin Pathway Using Gene Annotation and Protein Fold Based Selection Approach. Adv Biotech & Micro. 2017; 7(3): 555715. DOI: 10.19080/AIBM.2017.07.555715