BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 6(7)

Characterization of Lovastatin biosynthetic cluster proteins in *Aspergillus terreus* strain ATCC 20542

Thankaswamy Kosalai Subazini* & Gopal Ramesh Kumar

Bioinformatics Lab, AU-KBC Research Centre, M.I.T campus of Anna University Chennai, Chromepet, Chennai 600 044, India; Thankaswamy Kosalai Subazini - Email: subazinitk@au-kbc.org; Phone: 91 44 22234885, 91 9487114579; Fax: 91 44 22231034; *Corresponding author

Received December 21, 2010; Accepted February 22, 2011; Published June 23, 2011

Abstract:
Aspergillus terreus is a filamentous ascomycota, which is prominent for its production of lovastatin, an antihypercholesterolemic drug. The commercial importance of lovastatin with annual sales of billions of dollars made us to focus on lovastatin biosynthetic cluster proteins. The analysis of these lovastatin biosynthetic cluster proteins with different perspectives such as physicochemical property, structure based analysis and functional studies were done to find out the role and function of every protein involved in the lovastatin biosynthesis pathway. Several computational tools are used to predict the physicochemical properties, secondary structural features, topology, patterns, domains and cellular location. There are 8 unidentified proteins in lovastatin biosynthetic cluster, in which 6 proteins have homologous partners, and annotation transfer is done based on the closely related homologous genes, and their structures are also modeled. The two other proteins that do not have homologous partners are predicted as PQ loop repeat protein that may be involved in glycosylation machinery and as thiolase-acyl activity by the integrated functional analysis approach.

Keywords: *Aspergillus terreus*, Functional annotation, HMG CoA reductase inhibitor, Lovastatin, Sequence analysis, Structure analysis.

Background:
Aspergillus terreus, an important fungus is a major source for lovastatin, which treats cardiovascular disease by effectively inhibiting HMG CoA reductase activity, the rate limiting step in cholesterol biosynthesis. In addition, lovastatin is also involved in anti-inflammatory antioxidant activities and has the ability to inhibit proliferation and induce apoptosis in a variety of tumor cell lines **[1]**. Experimental evidences from literature reveal the presence of 18 proteins involved in lovastatin biosynthesis of *A. terreus*, ATCC 20542. Out of the 18 lovastatin biosynthetic cluster proteins it is found that 2 proteins are involved in regulatory mechanisms, 3 in transportation, 9 enzymes, 2 unknown proteins of thiolases acyl-enzyme intermediate signature and PQ loop repeat, and 2 megasynthases, **[2]** Lovastatin Nonaketide Synthase (LNKS) and Lovastatin Diketide Synthase (LDKS). The *insilico* analysis of physicochemical properties along with their topology and the functional analyses of these proteins are done by integrating standard functional annotation tools available online based on different methodologies. The final verification of functions of unknown proteins is done by structure based analysis.

Materials and Methodology:
The protein sequences involved in lovastatin biosynthesis in ATCC 20542 strains are retrieved from NCBI (www.ncbi.nlm.nih.gov) database.

Sequence analysis:
Primary sequence analysis is done using sequence manipulation suite **[3]**. By using the suite the physicochemical properties like grand average of hydropathicity (GRAVY), pH, acidicity, basicity, Instability index, aliphatic index, extinction coefficient and molecular weight are calculated.

Secondary structure prediction:
The secondary structure of protein features such as helical content, beta sheet formations and turns, loops, and coil regions are predicted by GOR **[4]**, available within Antherprot **[5]**. A number of trans-membrane helices are also identified from the sequences.

Functional annotation:
The functional annotation is done using different standard functional annotation tools available online such as KOGnitor **[6]** which incorporates orthologous information for eukaryotic sequence, Pfam **[7]**, a tool that is based on protein families, ScanProsite **[8]** for identifying biologically significant sites, ProDom **[9]** for domain based analysis, and BLAST **[10]** for annotating function based on homology. SignalP **[11]** is used for determining whether the protein is secretory or non-secretory and ProtFun **[12]** for characterizing function by integrating post-translational and localization aspects of the protein.

Tertiary structure prediction:
The 3D structures of the unknown proteins are predicted by using MODELLER9V8 **[13]**. The templates for unknown proteins, which do not have structural homologs in Protein Data Bank (PDB) **[14]** (through blastp analysis) is determined using mgenThreader **[15]**. The model of the proteins was subjected to Swiss-PDB Viewer **[16]** for refinement (of side chains, problematic loops, removal of amino acid clashes and energy minimization). Finally, the refined models were subjected for validation. Backbone conformations are evaluated by Psi/Phi Ramachandran plot obtained from SAVS **[17]** and the Z-Score is obtained from Prosa web server **[18]**.

Results and Discussion:

Primary sequence analysis:

The primary sequence analysis revealed that the GRAVY indices of 13 lovastatin biosynthetic proteins are ranging from – 0.03 to -0.5, which indicates they are hydrophilic and mostly predicted to be localized in extracellular and mitochondrial regions, the other 5 proteins (PQ loop repeat containing protein, HMG-CoA reductase, the 2 Permeases of the major facilitator superfamily, tricarboxylate transport protein) are hydrophobic (GRAVY values range from 0.05 to 0.6) and are localized in plasma membrane and basic in nature with pH value greater than 7. There are 2 more basic proteins, cytochrome P450 monooxygenase and protein containing thiolase activity region. The rest of proteins are acidic in nature. The aliphatic index value of lovastatin cluster proteins ranges between 60°C and 110°C, indicating that the proteins are highly thermostable for wide temperature range. 11 proteins have instability index greater than 40 (**Figure 1**) and are unstable.

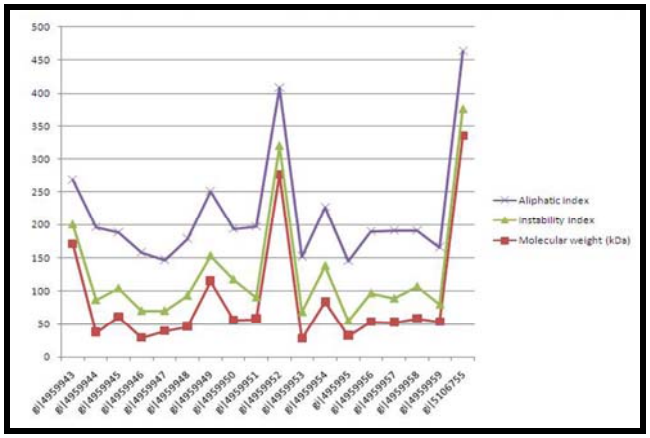

Figure 1: Physicochemical analysis on lovastatin biosynthetic cluster proteins. The graph shows the correlation observed in aliphatic, instability index and molecular weight of lovastatin biosynthetic cluster proteins.

Secondary structure:

The secondary structures of lovastatin biosynthetic cluster proteins revealed that all proteins in cluster have random coils dominated among secondary structure elements alpha helix, extended strand and beta turns.

Topology predictions:

In lovastatin biosynthetic cluster, there are 4 protein sequences with transmembrane helices, in which 2 protein sequences (efflux pump (gi|4959951) and Major facilitator superfamily (gi|4959957)) are involved in membrane transport and has 12 transmembrane helices. The other two sequences with 6 transmembrane helices are enzymes predicted to be PQ loop repeat membrane protein (gi|4959944) and HMG-CoA reductase (gi|4959949).

Functional analysis:

The ScanProsite results showed that presence of unknown protein sequence with accession number gi|4959953 has the thiolases acyl-enzyme intermediate signature with the pattern [LIVM]-[NST]-{T}-x-C-[SAGLI]-[ST]-[SAG]-[LIVMFYNS]-x-[STAG]-[LIVM]-x(6)-[LIVM]. The BLAST results revealed that even though gi|4959959 is highly homologous with immunoglobulin I-set domain protein it is also homologous with Glycosyl hydrolase family 67 N-terminus with 98% query coverage and E-value 9e-86. The presence of sequons NXS/T also confirms that gi|4959959 codes for Glycosyl family protein. The sequons are verified by using Net-N-Glyc server **[19] (Figure 2).** Gi|4959944 is distantly homologous to uroporphyrinogen decarboxylase with query coverage of 33% and E-value of 0.11 is predicted as a PQ-loop repeat family protein. The other proteins are having close homology and their functional prediction is done by annotation transfer from homologous sequences. The patterns and signature of lovastatin biosynthetic proteins along with their predicted functional roles of ATCC strain are shown in **Table 1 (see Supplementary material).**

3-D structure prediction and analysis:

Homology based tertiary structure prediction for unknown proteins are done using Modeller so as to verify the function predicted. Totally 5 models are generated for each of the individual protein and the best model of each protein is refined using Swiss PDB Viewer. The refined structures are validated and the scores obtained are shown in **Table 2 (see Supplementary material)**. More than 80% of residues in modeled structures obey Ramachandran plot, and Prosa Z- scores are negative, which indicates the predicted structure correlated with the structural features of PDB. Homology modeling for the unknown protein gi|4959944 failed, because of the lack of homologous structures. The 3-D structure of modeled protein gi|4959959 (**Figure 3**) with the template Beta-N-acetylhexosaminidase enzyme confirms that gi|4959959 belongs to Glycosyl hydrolase family.

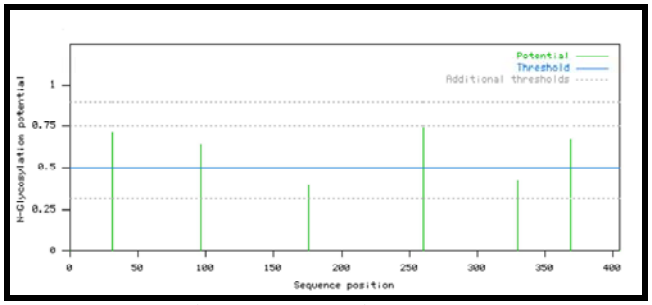

Figure 2: Prediction of glycosylation sites for unknown protein sequence with accession number gi|4959959 using NetNGlyc Server. The vertical lines indicate the presence of 4 sequons crossing the threshold (horizontal line at 0.5) and 2 sequons crossing additional thresholds are predicted to be glycosylated.

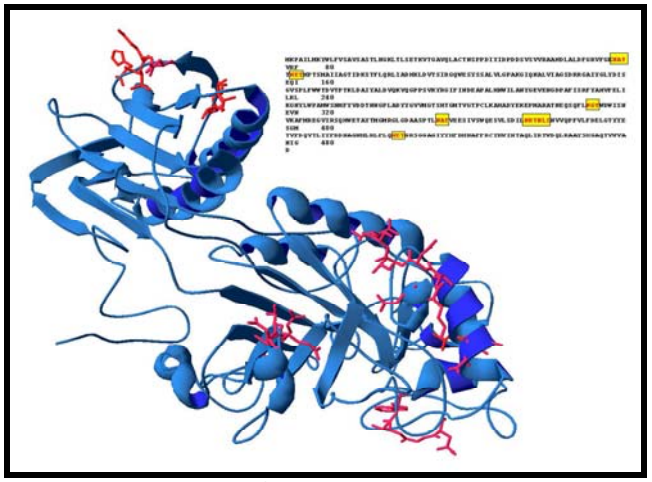

Figure 3: Modelled structure for unknown protein sequence with accession number gi|4959959. The unknown protein sequence of accession number gi|4959959 is modeled using Modeller and the function is predicted as beta-N-acetylhexosaminidases based on structural similarity. The N-glycosylation sites are represented by shaded regions in sequence and as stick in the 3-D structure.

Conclusion:

The *insilico* analysis of lovastatin biosynthetic cluster protein to understand their physicochemical, structural and functional properties are performed. Functional analysis of eight unknown proteins performed by different function analysis tools based on domains, motifs and profiles improved the function annotation capability. The annotation revealed the presence of cis-aconitic acid decarboxylase of itaconic acid biosynthesis in lovastatin biosynthetic cluster, which shows that the two metabolites itaconic acid and lovastatin are related. Gi|4959953 has thiolase activity domain and regulation of it may affect the energy level of the cells in metabolic state. From this analysis it is also found that the sequences with accession number gi|4959944 and gi|4959959 may be involved in glycosylation machinery as tailoring enzyme and can be used in preparing analogues of Lovastatin. The incorporation of these enzymes with cytochrome P450s into heterologous hosts may help in large scale production of lovastatin in microbial fermentations.

Acknowledgment:

Subazini TK is thankful to CSIR, Govt of India, New Delhi for the award of Senior Research fellowship (SRF).

References:
[1] Martirosyan A *et al*. *BMC Cancer*. 2010 **10**: 103 [PMID: 20298590]
[2] Xie X *et al*. *Chem Biol*. 2006 **13**(11): 1161 [PMID: 17113998]
[3] Stothard P. *Biotechniques* 2000 **28**: 1102 [PMID: 10868275]
[4] Garnier J *et al*. *J Mol Biol*. 1978 **120**(1): 97 [PMID: 642007]
[5] Deleage G *et al*. *Comput Biol Med*. 2001 **31**(4): 259 [PMID: 11334635]
[6] Tatusov RL *et al*. *BMC Bioinformatics*. 2003 **4**: 41 [PMID: 12969510]
[7] Bateman A *et al*. *Nucleic Acids Res*. 2004 **32**: D138 [PMID: 14681378]
[8] Gasteiger E *et al*. *Nucleic Acids Res*. 2003 **31**: 3784 [PMID: 12824418]
[9] Servant F *et al*. *Brief Bioinform*. 2002 **3**: 246 [PMID: 12230033]
[10] Altschul SF *et al*. *J Mol Biol*. 1990 **215**: 403 [PMID: 2231712]
[11] Bendtsen JD *et al*. *J Mol Biol*. 2004 **340**: 783 [PMID: 1522332]
[12] Jensen LJ *et al*. *J Mol Biol*. 2002 **319**: 1257 [PMID: 12079362]
[13] Sali A & Blundell TL. *J Mol Biol*. 1993 **234**: 779 [PMID: 8254673]
[14] Berman HM *et al*. *Acta Crystallogr D Biol Crystallogr*. 2002 **58**: 899 [PMID: 12037327]
[15] Jones DT. *J Mol Biol*. 1999 **287**(4): 797 [PMID: 10191147]
[16] Kaplan W & Littlejohn TG. *Brief Bioinform*. 2001 **2**(2): 195 [PMID: 11465736]
[17] Lovell SC *et al*. *Proteins* 2003 **50**(3): 437 [PMID: 12557186]
[18] Colovos C & Yeates TO. *Protein Sci*. 1993 **2**(9): 1511 [PMID: 8401235]
[19] http://www.cbs.dtu.dk/services/NetNGlyc/

Edited by P Kangueane
Citation: Subazini & Kumar. Bioinformation 6(7): 250-254 (2011)

BIOINFORMATION open access

Supplementary material:

Table 1: Functional analysis on lovastatin biosynthetic cluster. The analysis of lovastatin biosynthetic cluster proteins and their predicted domains and functional role are shown in this table.

Accession number	Domains	Predicted from tool esterase	Previous annotation	Current Annotation	Functional role
gi|4959943|gb|AAD34550.1| esterase [Aspergillus terreus].	GDSL-like Lipase/Acylhydrolase	Pfam	esterase	Esterase	Acts on ester bonds, carboxylesterase activity
gi|4959944|gb|AAD34551.1| unknown [Aspergillus terreus].	PQ loop repeat	Pfam	unknown	PQ-loop repeat containing protein	Membrane bound protein involved in Glycosylation machinary
gi|4959945|gb|AAD34552.1| cytochrome P450 monooxygenase [Aspergillus terreus].	Cytochrome P450	Pfam, KOGnitor, Blast	cytochrome P450	Cytochrome P450	Monacolin J are likely to be catalyzed by cytochrome P450 enzymes
gi|4959946|gb|AAD34553| unknown [Aspergillus terreus].	phospholipase/carboxyhydrolase	KOGnitor	unknown	oxidoreductase	Involves in the hydrolysis of hydrophobic phosphonate ester compounds.
gi|4959947|gb|AAD34554| enoyl reductase [Aspergillus terreus].	Alcohol dehydrogenase GroES-like domain ,Zinc-binding dehydrogenase	Scanprosite	enoyl reductase	Lovastatin Polyketide Enoyl Reductase (LovC)	LovC with LNKS synthesizes dihydomonacolin J with successive ketoreduction and dehydration activity.
	Zinc-binding oxidoreductase	Pfam			
gi|4959948|gb|AAD34555| transesterase [Aspergillus terreus].	Beta-lactamase class C and other penicillin binding proteins	KOGnitor	transesterase	Transesterase (LovD)	LovD catalyze last step of biosynthesis by transacylating acl group from LDKS to monacolin J to yield
	Beta-lactamase	Pfam			
gi|4959949|gb|AAD34556| hydroxymethylglutaryl-coenzyme A reductase [Aspergillus terreus].	Sterol-sensing domain (SSD) profile, Hydroxymethylglutaryl-coenzyme A reductases family profile, Hydroxymethylglutaryl-coenzyme A reductases signature 1, 2,3	ScanProsite	3-hydroxy-3-methylglutaryl-coenzyme A reductase	3-hydroxy-3-methylglutaryl-coenzyme A reductase	Involved in fatty acid metabolism
gi|4959950|gb|AAD34557| regulatory protein [Aspergillus terreus].	Fungal Zn(2)-Cys(6) binuclear cluster domain ,Fungal specific transcription factor domain ,	Pfam	Regulatory protein	Fungal-type DNA-binding transcription factor	Regulates a variety of cellular and metabolic processes.
	Zn(2)-C6 fungal-type DNA-binding domain profile, Zn(2)-C6 fungal-type DNA-binding domain signature,	Scanprosite			
gi|4959951|gb|AAD34558| unknown [Aspergillus terreus].	Major Facilitator Superfamily	Pfam, KOGnitor, Blast, Scanprosite	unknown	Permeases of the major facilitator superfamily	Possible roles for efflux pumps include removal of end products of metabolism and buffering the cytoplasm. located in th*e cytoplasmic region*
gi|4959952|gb|AAD34559| polyketide synthase [Aspergillus terreus]	Beta-ketoacyl synthase, N-terminal domain ,Beta-ketoacyl synthase, C-terminal domain ,Acyl transferase domain ,	Pfam	polyketide synthase	Lovastatin diketide synthase	LDKS produce 2-methyl butyrate side chain.
	Acyl carrier protein phosphopantetheine domain profile, Beta-ketoacyl synthases active site, Sugar transport proteins signature 1,	Scanprosite			

gi|4959953|gb|AAD34560| unknown [Aspergillus terreus].	Thiolases acyl-enzyme intermediate signature,	ScanProsite	Unknown	Enzyme involved in Thiolase activity	Enzyme involved in Thiolase activity
	Thiolase	ProDom			
gi|4959954|gb|AAD34561| regulatory protein [Aspergillus terreus].	Fungal Zn(2)-Cys(6) binuclear cluster domain ,Fungal specific transcription factor domain	Pfam,	regulatory protein	Fungal-type DNA-binding transcription factor	Regulates a variety of cellular and metabolic processes.
	Zn(2)-C6 fungal-type DNA-binding domain profile, Zn(2)-C6 fungal-type DNA-binding domain	Scanprosite			
gi|4959955|gb|AAD34562| unknown [Aspergillus terreus].	Mitochondrial carrier protein	Pfam	unknown	tricarboxylate transport protein, mitochondrial precursor	Involved in citrate-H+ /malate exchange. Provides a carbon source for fatty acid and sterol biosyntheses, and NAD+ for the glycolytic pathway.
	Solute carrier (Solcar) repeat profile,	Scanprosite			
	tricarboxylate transport protein, mitochondrial precursor	Blast			
gi|4959956|gb|AAD34563| unknown [Aspergillus terreus].	MmgE/PrpD family	Pfam	unknown	cis-aconitic acid decarboxylase	Involved in itaconic acid biosynthesis
gi|4959957|gb|AAD34564| unknown [Aspergillus terreus].	Major Facilitator Superfamily	Pfam, KOGnitor, Blast, Scanprosite	unknown	Permeases of the major facilitator superfamily	Transportation
gi|4959958|gb|AAD34565| cytochrome P450 monooxygenase [Aspergillus terreus]	cytochrome P450 monooxygenase	Pfam, KOGnitor, Blast, Scanprosite	cytochrome P450 monooxygenase	PHENYLACETATE P450 HYDROXYLASE	cytochrome P450 monooxygenase
gi|4959959|gb|AAD34566| unknown [Aspergillus terreus].			unknown	Glycosidases family 67 N terminus	Alpha-Glucuronidases are family 67 glycosidases that cleave the Alpha-1,2-glycosidic bond between 4-*O*-methyl-D-glucuronic acid and xylose units as part of an array of hemicellulose-hydrolyzing enzymes
gi|5106755|gb|AAD39830.1|AF151722_1 lovastatin nonaketide synthase [Aspergillus terreus].	Beta-ketoacyl synthase, N-terminal domain ,Thiolase, N-terminal domain ,Beta-ketoacyl synthase, C-terminal domain ,Acyl transferase domain	Pfam	hypothetical protein similar to polyketide synthase	lovastatin nonaketide synthase	LNKS is in iterative form and along with lovC; it synthesizes dihydomonacolin J
	Acyl carrier protein phosphopantetheine domain profile, Beta-ketoacyl synthases	Scanprosite			

Table 2: Structure validation report. Structure validation of different models for 6 unknown proteins obtained from MODELLER

Accession number	RAMACHANDRAN PLOT				PROSA (Z-Score)	TEMPLATE	Template Function
	Core (%)	Allowed (%)	General (%)	Disallowed (%)			
gi|4959946	90.50	7.20	1.80	0.50	-4.26	1YCD	Yhr049w/FSH1
gi|4959951	86.10	11.3	0.90	1.80	-2.73	1PW4	Glycerol-3-Phosphate Transporter
gi|4959955	90.60	8.60	0.40	0.40	-2.61	1OKC	Mitochondrial ADP/ATP carrier
gi|4959956	84.70	12.0	1.40	1.90	-5.84	2HP0	Aminopeptidase N
gi|4959957	84.70	11.0	2.90	1.40	-1.1	2GFP	multidrug transporter EmrD
gi|4959959	81.20	14.5	3.1	1.20	-3.74	3GH5	Beta-N-acetylhexosaminidase