



## DOLOP—database of bacterial lipoproteins

M. Madan Babu and K. Sankaran\*

Centre for Biotechnology, Anna University, Chennai-600025, India

Received on September 17, 2001; revised on November 19, 2001; accepted on November 30, 2001

### ABSTRACT

**Summary:** Bacterial lipoproteins and lipid modification are gaining importance owing to their essential nature, roles in pathogenesis and interesting commercial applications. We have created an exclusive knowledge base for bacterial lipoproteins by processing information from 510 entries to provide a list of 199 distinct lipoproteins with relevant links to molecular details. Features include functional classification, predictive algorithm for query sequences, primary sequence analysis and lists of predicted lipoproteins from 43 completed bacterial genomes along with interactive information exchange facility.

**Availability:** The website called Database Of bacterial LipOProteins (DOLOP) is available at <http://www.mrc-lmb.cam.ac.uk/genomes/dolop> along with additional information on the biosynthetic pathway, supplementary material and other related figures.

**Contact:** [ksankaran@annauniv.edu](mailto:ksankaran@annauniv.edu); [madanm@mrc-lmb.cam.ac.uk](mailto:madanm@mrc-lmb.cam.ac.uk)

### INTRODUCTION

Covalent modification of a protein with the lipid, *N*-acyl-*S*-diacylglyceryl cysteine, at the *N*-terminus was identified in 1973 in the major outer membrane protein of *Escherichia coli* (Hantke and Braun, 1973). Subsequently, several related and unrelated bacterial proteins with the same lipid modification were discovered and these are now generally referred to as lipoproteins. The initial research in to biosynthesis of these lipoproteins established that bacterial lipoproteins were synthesized as preproteins with a signal sequence (Inouye *et al.*, 1977). Analysis of signal sequences of 26 distinct and well characterized lipoproteins revealed the presence of a distinct sequence at the *C*-terminal end of their signal peptides identifiable with a consensus sequence of LAGC and therefore referred to as lipobox (Sankaran and Wu, 1993); cys (+1 position) that is lipid-modified is invariant and the -3 position is mainly Leu. Mutation studies showed that the *N*-terminal 5–7 residues mostly with two positively charged (Lys/Arg) residues (*n*-region) and the intervening stretch of hydrophobic and uncharged

residues of 7–15 amino acid length (*h*-region) were also features recognized in lipid modification (Braun and Wu, 1993).

Subsequently, the consensus sequence of lipobox in a signal sequence has been effectively used as a signature to predict lipoproteins. To date more than 450 lipoproteins have been identified from a variety of bacterial sources. Research in to its biosynthetic enzymes has revealed that the pathway is essential and ubiquitous in bacteria. Recently, several pathogen-associated virulence lipoproteins have been identified and shown to play a critical role in host–pathogen surface interactions and transport (Zgurskaya and Nikaido, 2000; Chambaud *et al.*, 1999).

Owing to the importance of bacterial lipoproteins, we have compiled the information available, in to an ever expanding database available at <http://www.mrc-lmb.cam.ac.uk/genomes/dolop>. More accurate predictive rules based on the information from the larger set of lipoproteins (199 versus 26) have been used to create an interactive CGI-program, which allows the user to submit their complete protein sequence to find out the presence of a possible lipoprotein signal sequence, using the criteria mentioned subsequently in the text.

### ORGANIZATION OF THE KNOWLEDGE BASE

The bacterial lipoprotein sequences were retrieved from the June 2001 release of Swiss-Prot database (Bairoch and Apweiler, 2000) using combinations of keyword searches and limiting the search to only bacterial species. Additionally, a primary literature search was also carried out to get information about lipoproteins that were missed by the keyword searches. This resulted in 510 sequences. Proteins, which were annotated as putative or hypothetical were not considered for analysis (except when experimentally verified, but not updated in Swiss-Prot). The resulting proteins were then manually checked to remove orthologues from other species. This resulted in 199 distinct lipoproteins. Based on the literature available for each of them, the 199 proteins were classified according to common functions into eight different categories namely structural proteins, binding proteins, transporters, adhesins, toxins, antigens, enzymes and interesting factors. Links to Swiss-Prot, Blast output, PDB, primary

\*To whom correspondence should be addressed.

**Table 1.** Percentage occurrence of residues at different positions in the lipobox

-3 position		-2 position		-1 position		+1 position	
Residue	Occurrence (%)	Residue	Occurrence (%)	Residue	Occurrence (%)	Residue	Occurrence (%)
L	75.9	A	29.8	G	44.7	C	100
V	7.7	S	25.4	A	40.5		
I	3.7	T	12.2	S	14.8		
A	3.7	V	12.1				
F	3.2	I	9.8				
M	2.8	G	2.6				
T	1.4	M	2.6				
C	1.1	F	1.5				
G	0.5	L	1.5				
		C	0.5				
		N	0.5				
		P	0.5				
		Q	0.5				
		Y	0.5				

sequence analysis page and reference to primary literature for each lipoprotein has been provided for seeking further information on any lipoprotein. Figure 1 schematically represents the organization of the database.

### SIGNAL SEQUENCE ANALYSIS

Analysis of 199 distinct lipoproteins revealed that in 73% of the proteins, the consensus lipobox sequence would be [LV][ASTVI][GAS][C] (Table 1). The *n*-region would contain mostly two positively charged residues (Arg/Lys) within the first 5–7 residues and the *h*-region would contain 7–20 hydrophobic/uncharged residues (with a mean length of 12 residues), followed immediately by the lipobox. More details about the signal sequence analysis are available at the website as supplementary information (<http://www.mrc-lmb.cam.ac.uk/genomes/dolop/lipobox.htm>).

A predictive algorithm has been developed based on the consensus referred above and has been incorporated in to the website to analyze a user given query sequence and to pull out probable lipoproteins from the completed bacterial genomes.

### PREDICTION OF POSSIBLE LIPOPROTEINS FROM COMPLETELY SEQUENCED BACTERIAL GENOMES

Availability of complete genome sequences for the 43 bacterial species served as a good data to re-search for lipoproteins with the improved predictive rules. The translated protein sequence (chromosomal and plasmid encoded proteins) available at the National Centre for Biotechnology Information (NCBI—<http://www.ncbi.nlm.nih.gov/>) for each of the 43 completed bacterial genomes was used as the starting point for the prediction. The list of

predicted lipoproteins and their sequences for each of the completed bacterial genomes is available at <http://www.mrc-lmb.cam.ac.uk/genomes/dolop/compgen.htm>.

It was interesting to note that although as many as 50 lipoproteins have been biochemically characterized in a well studied system like *E. coli*, our prediction results show that there are almost 99 possible lipoproteins in *E. coli*, strain K12, including hypothetical and unannotated ones. When the same set of predictive rules was applied to the pathogenic *E. coli* strain O157:H7, as many as 120 lipoproteins are identified. Most of them still remain annotated as hypothetical proteins or possible Open Reading Frames (ORFs). Our prediction results indicate that a set of probable lipoproteins not investigated so far is present in virulent and non-virulent strains of bacteria including *E. coli* that could aid in the search of essential lipoproteins and additional pathogenic factors.

### CONCLUSION AND FUTURE DIRECTIONS

This newly created knowledge base for the first time attempts to classify the large body of information available on nearly 500 lipoproteins into packages that users could make use of, depending on their requirement. By making it interactive, we look forward to continuous input from users and interactions among them. As the Swiss-Prot gets updated, we will update this site once every six months and include additional features on biosynthetic enzymes and interesting findings on lipoproteins. This knowledge base is dedicated to Professor Henry C. Wu, a pioneer in lipoprotein biosynthesis who has contributed immensely to the advancement of the field.

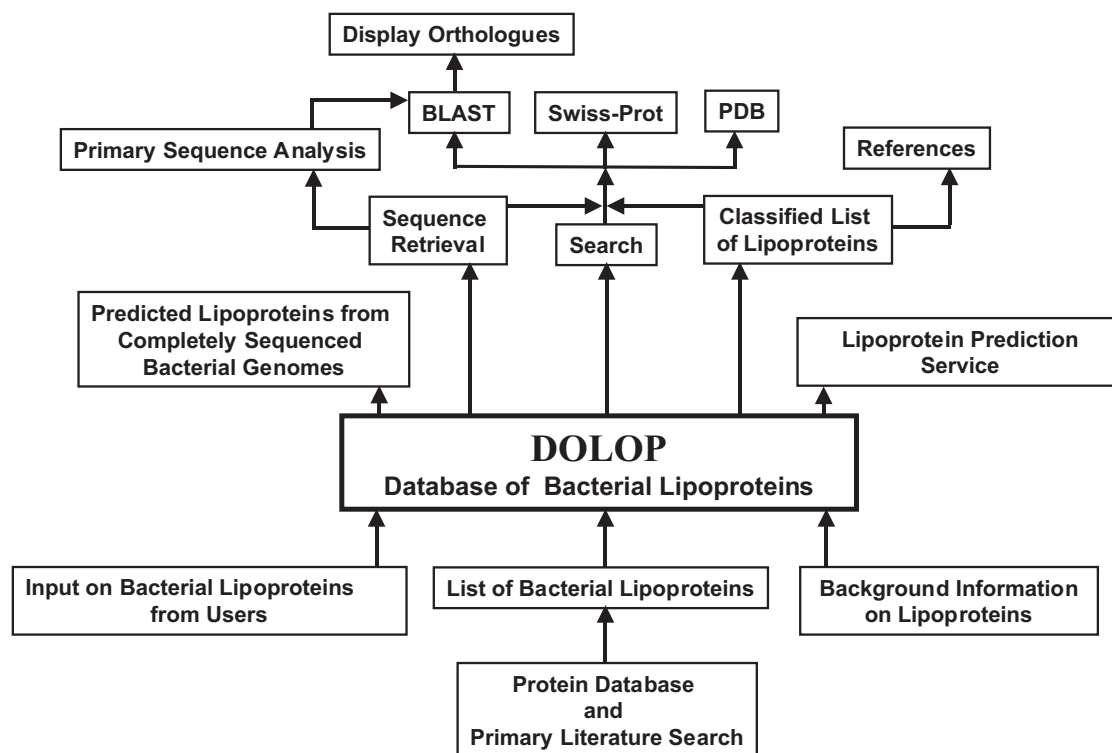


Fig. 1. Organization of the database.

## ACKNOWLEDGEMENTS

We thank Professor P.Kaliraj, Director, Centre for Biotechnology, Anna University for providing the computational facilities and encouragement to create the website and prepare this paper. We thank Dr Sarah Teichmann and Dr Cyrus Chothia for useful discussions and MRC-Laboratory of Molecular Biology for kindly providing web space to host DOLOP. We also thank the anonymous reviewers for their valuable suggestions.

## REFERENCES

- Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Braun,V. and Wu,H.C. (1993) Lipoproteins, structure, function, biosynthesis and model for protein export. In Ghuysen,J.-M. and

- Hakenback,R. (eds), *Comprehensive Biochemistry*, vol 27. *Bacterial Cell Wall*, Elsevier, Amsterdam, pp. 319–342.
- Chambaud,I., Wroblewski,H. and Blanchard,A. (1999) Interactions between *Mycoplasma* lipoproteins and the host immune system. *Trends Microbiol.*, **7**, 493–499.
- Hantke,K. and Braun,V. (1973) Covalent binding of lipid to protein. Diglyceride and amide-linked fatty acid at the *N*-terminal end of the murien lipoprotein of the *Escherichia coli* outer membrane. *Eur. J. Biochem.*, **34**, 284–296.
- Inouye,S. *et al.* (1977) Amino acid sequence for the peptide extension on the prolipoprotein of the *Escherichia coli* outer membrane. *Proc. Natl Acad. Sci. USA*, **74**, 1004–1008.
- Sankaran,K. and Wu,H.C. (1993) Bacterial lipoproteins. In Schlesinger,M.J. (ed.), *Lipid Modifications of Proteins*. CRC Press, Boca Raton, pp. 163–181.
- Zgurskaya,H.I. and Nikaïdo,H. (2000) Multidrug resistance mechanisms: drug efflux across two membranes. *Mol. Microbiol.*, **37**, 219–225.