Molecular cloning, sequence analysis and homology modeling of *galE* encoding UDP-galactose 4-epimerase of *Aeromonas hydrophila*

Shivani Agarwal [1], Keshav Gopal [1], Gagan Chhabra [2], Aparna Dixit [1,*]

[1] Gene Regulation Laboratory, School of Biotechnology, Jawaharlal Nehru University, New Delhi – 110067, India; [2] Gene Regulation Laboratory, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi -110067, India; Aparna Dixit - Email: adix2100@mail.jnu.ac.in; Phone: +91-11-26742164; Fax: +91-11-26704085; *Corresponding author

Abstract:

A. hydrophila, a ubiquitous gram-negative bacterium present in aquatic environments, has been implicated in illness in humans, fish and amphibians. Lipopolysaccharides (LPS), a surface component of the outer membrane, are one of the main virulent factors of gram-negative bacteria. UDP-galactose 4-epimerase (GalE) catalyses the last step in the Leloir pathway of galactose metabolism and provides precursor for the biosynthesis of extracellular LPS and capsule. Due to its key role in LPS biosynthesis, it is a potential drug target. The present study describes cloning, sequence analysis and prediction of three dimensional structure of the deduced amino acid sequence of the *galE* of *A. hydrophila* AH17. The cloned *galE* consists of the putative promoter-operator region, and an open reading frame of 338 amino acid residues. Sequence alignment and predicted 3D-structure revealed that the GalE of *A. hydrophila* consists of the signature sequences of the epimerase super family. The present study reports the molecular modeling / 3D-structure prediction of GalE of *A. hydrophila*. Further, the potential regions of the enzyme that can be targeted for drug design are identified.

Keywords: *Aeromonas hydrophila*, lipopolysaccharide, virulence, UDP-galactose 4- epimerase, molecular phylogenetic.

Abbreviations: UDP-GlcNAc, UDP-N-acetylglucosamine; GalE, UDP-galactose 4-epimerase; LPS, Lipopolysaccharides.

Background:

Aeromonas hydrophila is a member of the family Aeromonadaceae, associated with disease conditions mainly in fish, amphibians and humans [1]. Identification of strains of *A. hydrophila* capable of causing illness in apparently healthy individuals, by infecting open wounds and possibly by ingestion of the microorganism in food or water, has generated immense interest in this organism [2]. Some of the known virulence factors responsible for pathogenesis of *A. hydrophila* are O-antigen lipopolysaccharide, capsules, exotoxins, enterotoxins, and certain exoenzymes [3 & 4]. LPS has been reported to be involved in adherence and may play a role in antigenic variation [5-7]. Importance of the enzymes involved in galactose metabolism in bacterial virulence has been demonstrated [8-10]. GalE is one of the enzymes involved in galactose metabolism that mediates the incorporation of galactose in extra cellular polysaccharide material such as the O-side chain of lipopolysaccharide. The essential role of UDP-galactose 4-epimerase in virulence of many other gram-negative bacteria is well documented [11-15]. The fact that the epimerase mutants have altered LPS core biosynthesis with significant reduction in their ability to adhere and invade the host cells makes epimerase a potential drug target. GalE from different species exhibits a significant degree of interspecies variation at their gene and quaternary structure. In the present study, we report cloning, characterization of the *galE,* including its putative promoter, and structure modeling of the deduced amino acid sequence of the GalE of *A. hydrophila*

Methodology:

Bacterial strains and vector:

A. hydrophila (AH17) isolated from pond water was obtained from Dr. I. Karunasagar, College of Fisheries, Mangalore, India. *Escherichia coli* DH5α and BL21 (DE3) strains were from GIBCO BRL, USA and Novagen, USA, respectively. Plasmid pBCKS+ was procured from Stratagene (USA).

Cloning and sequencing of *galE* of *A. hydrophila*:

Genomic DNA from *A. hydrophila* (AH17) was isolated essentially as described earlier [16]. The *galE* of *A. hydrophila* was PCR amplified using the genomic DNA as a template and the forward and reverse primers (5′-AGTCTGAGAAAAAGCGCGTGTG -3′, 5′-TTAATCGGGATATCCCTGTGGATGG-3′, respectively), designed on the basis of available sequence information of *galE* of *E. coli* (Acc. No. NC_000913) obtained from Microsynth, Switzerland. The PCR amplified product was purified and the ends of purified PCR product were phosphorylated using T4 polynucleotide kinase (NEB, USA), followed by ligation to *Sma* I digested dephosphorylated pBCKS (+) vector. Competent *E. coli* DH5α cells (Novagen, USA) were then transformed with the ligation mix and the transformants were analyzed by colony PCR and were further confirmed by restriction enzyme digestion for the release of the insert. The construct thus made was designated as pAHG*alE.* The integrity of the *galE* insert was verified by automated DNA sequencing (Applied Biosystem Model 393A).

Phylogenetic analysis:

Sequence analysis tools of the ExPASy Molecular Biology Server of Swiss Institute of Bioinformatics were used to process nucleic acid sequence for the deduced amino acid sequence. The deduced amino acid sequence of the GalE of *A. hydrophila* was aligned with the GalE of other species and *A. hydrophila* AH3 strain was carried out using ClustalW (Version 1.83) [17]. The phylogenetic tree was inferred using Phylip's inference package, Version 3.5c.

Structure Modeling and Visualization of Model:
The most appropriate template for Homology modeling of *A. hydrophila* GalE (Accession No. AJ785765) was identified using BlastP analysis. The available structure of GalE from *Escherichia coli* in the Protein Database (PDB) (PDB entry 1udcA, resolution =1.65, R value =0.177) was referred **[18]**. The target and the template sequences were aligned using ClustalW. Homology modeling program Swiss-Model was employed to generate a comparative 3D- structure model of *A. hydrophila* GalE **[19]**. Swiss-Model **[20]** is a server for automated comparative modeling of three-dimensional (3D) protein structures. No other refinements were applied. Swiss PDB viewer software **[21]** was employed as a tool to envisage the generated structural model.

Validation of the generated model:
The generated 3D-model was assessed/reviewed evaluated at various structure verification servers viz. PROCHECK **[22]** that relies on Ramachandran plot **[23]**, WHAT_CHECK, a subset of WHATIF programme **[24 & 25]**, and VERIFY3D **[26 & 27]**.
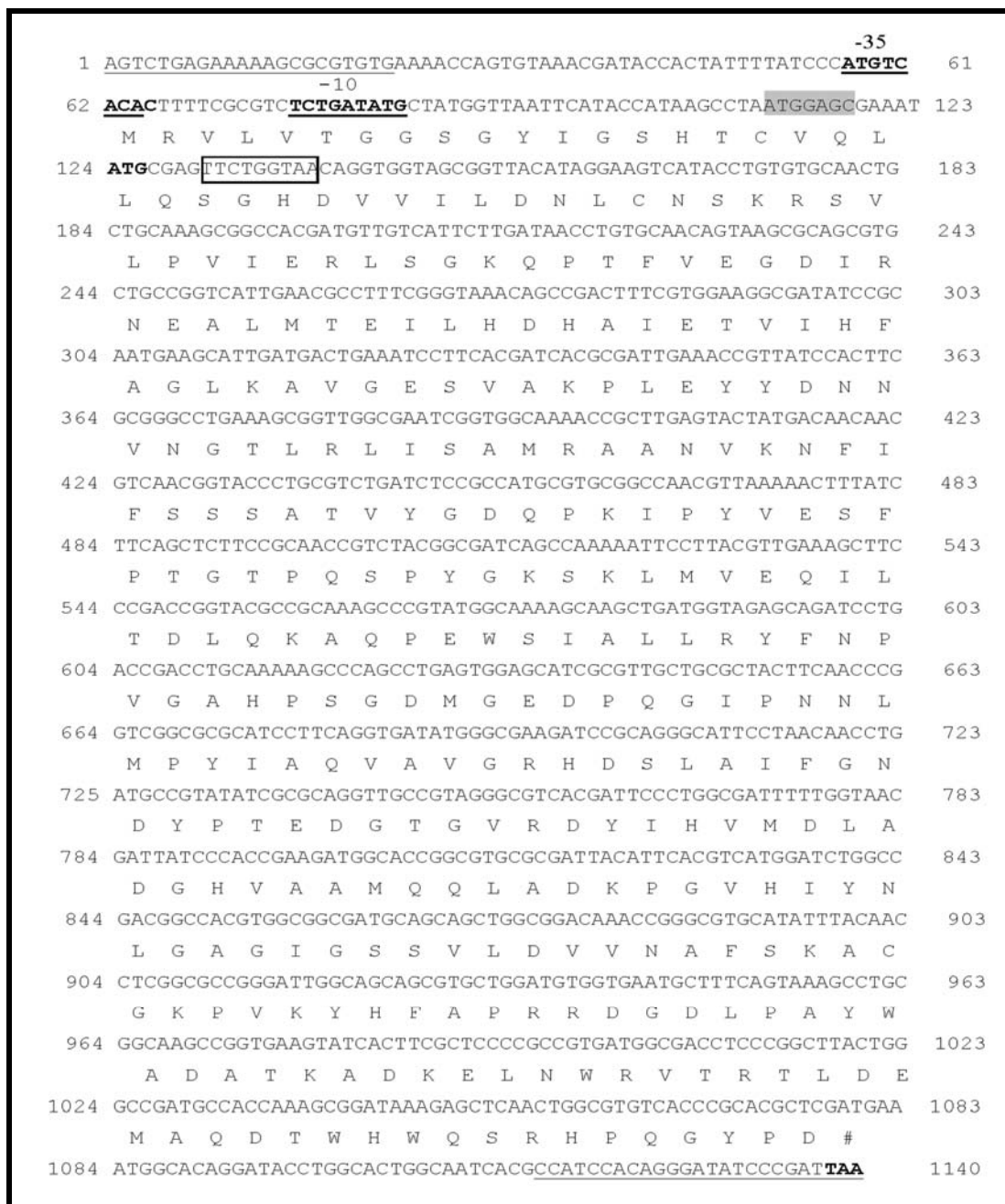

Figure 1: Nucleotide and deduced amino acid sequence of cloned *galE* of *A. hydrophila*. The open reading frame encodes for a protein of 338 amino acid residues. Initiation and termination codons are shown in bold. Putative -35 and -10 regions are shown as bold and underlined nucleotides. The shaded and the boxed bases represent putative ribosome binding site and the GalR binding site, respectively. Nucleotides pertaining to the primers used for amplification of the gene are underlined.
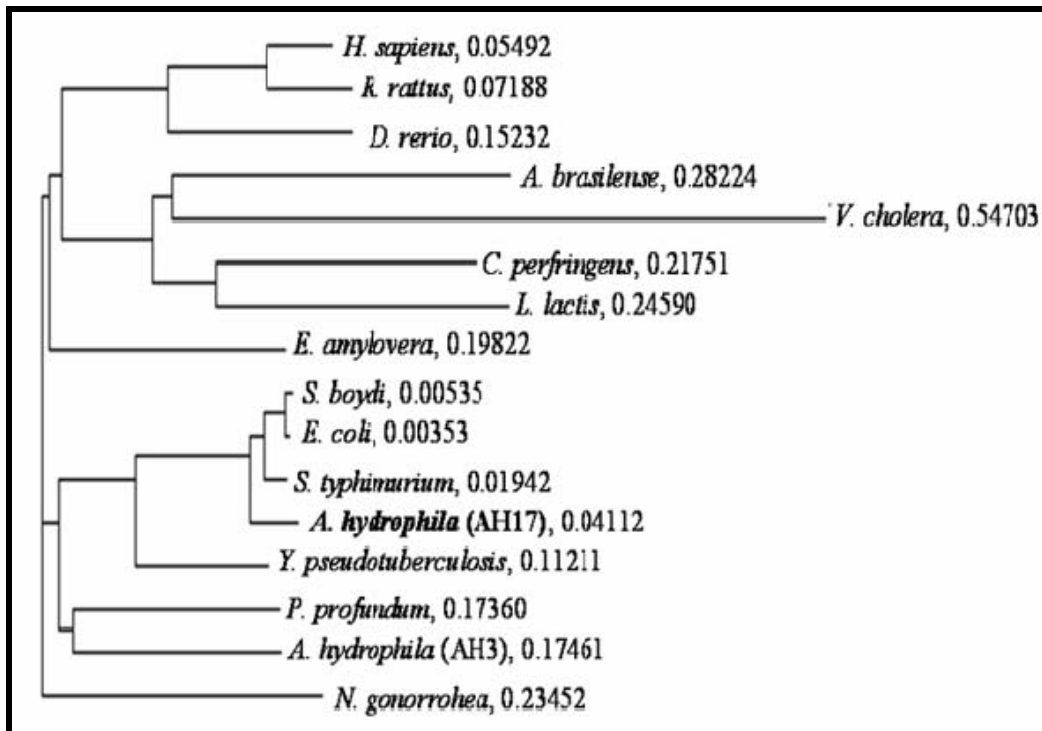
Figure 2: Rooted phylogenetic tree of the deduced amino acid sequence of the GalE of *A. hydrophila* and other organisms. Amino acid sequences for different organisms were obtained from NCBI database and aligned using Clustal W program. The distances from the nodes, i.e. the branch lengths denoted in the tree, correspond to sequence divergence.
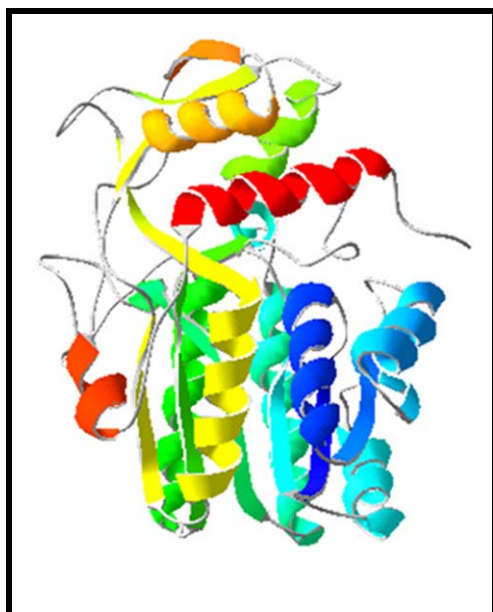

Figure 3: Homology model of the GalE of *A. hydrophila*. This model is produced by Swiss-Model program. Visualization of the structure was done by SWISS PDB VIEWER and is represented in the form of ribbons.

Discussion:

Sequence analysis of the *galE* of *A. hydrophila*:
Sequencing of the PCR amplified cloned *galE* fragment revealed the insert to be of 1140 bp, representing full length *galE* and its promoter-operator sequences (**Figure 1**). Sequence analysis revealed the presence of putative RNA polymerase binding site (-35 region) at 57-65 bp and a pribnow box (-10 region) at 76-84 bp. Putative binding site for catabolite repressor protein or cyclic AMP receptor protein (CRP) is also present at 49-69 bp, overlapping -35 region. The binding site for the GalR overlaps the translation start site and is present at 131-139 bp. Presence of all the regulatory sequences and components of a promoter upstream of the GalE encoding region of the cloned fragment indicate that the organization of *galETK* operon of *A. hydrophila* is similar to that of other gram-negative bacteria, which is

organized in the order of *galE, galT and galK*. Open reading frame of the cloned *galE* contains a single protein translation start site ATG at 124-126 bp and a termination codon, TAA, is present at the 1138-1140 bp (**Figure 1**). Putative ribosome binding site is located 7 bp upstream of the ATG at 112-117 bp position. The encoded protein is of 338 amino acid residues with a theoretical pI of 5.64 and molecular weight of 36501.36. Genome database search (Blast N) showed varying degrees of similarity to nucleotide sequence of the *galEs* of other species. Blast P of the deduced amino acid sequence showed that GalE of *A. hydrophila* shares the percentage identities ranging from ~60-95% (95% with *Shigella boydii and 63% with Photobacterium profundum*) with different species of bacteria. It is of interest to note that while the GalE of the *A. hydrophila* AH17 shows significant identity with other bacteria, it showed only 59% identity and 85% similarity with the GalE of another strain of *A. hydrophila*, AH3 **[15]**, though the active site and the catalytic sites/residues have remained conserved between the two. Bacteria of *Aeromonas* spp. are highly heterogeneous group of bacteria, and the differences in the GalEs of the two *Aeromonas hydrophila* strains may only be an indication of heterogeneity.

Phylogenetic analysis:
Analysis of amino acid sequence alignment of the GalE (**Figure 1 in supplementary material**) revealed that the *A. hydrophila* GalE consisted of the characteristic Tyr-X-X-X-Lys couple (position 128 to 133) that plays a key role in catalysis and a complete N-terminal NAD binding GXXGXXG motif (position 7 to 13), popularly known as 'Rossman fold'. Both these motifs are conserved in all the family members and among all the species (**Figure 2**, boxed sequences) **[28, 29]**. The signature sequences of the epimerase super family – FSSSATVYG, ALLRYFNPVGAHP, NNLMPXXAQVAXGRR-XXXX-IFGNDYPTEDGTGVRDYIHV, YNLGAGXXXSVLDVVN that have remained conserved across the species are also present in the GalE of *A. hydrophila* (Supplementary figure 1, shaded sequences). Thus, epimerases from all the species appear to have the same evolutionary origin and employ similar catalytic mechanisms though they differ significantly in their subunits, quaternary structure and requirement for NAD. It is also of interest to note that though the signature sequences and catalytic couple have remained conserved in GalE of *A. hydrophila*, it differs significantly from GalE of other species outside these domains. As evident from the Phylogram (**Figure 2**) generated from the sequence alignment, the closeness of *A. hydrophila* GalE with that of *E. coli, S. typhi, S. boydii* and many others is not surprising as these are all enteropathogenic. What is of interest to note that the encoded GalE of *A. hydrophila* exhibited only 51% and 53% identities with that of *H. sapiens* and *D. rerio,* respectively, in which it is an important disease causing bacteria, thus making it a potential drug target. The distance between these species is also evident from the inferred phylogenetic tree.

Structural model and Overall Architecture:
X-ray resolved crystal structure of GalE from *Escherichia coli* (PDB entry 1udc) is available from Protein Data Bank (PDB). Based on the sequence alignment, GalE from *Escherichia coli* was found to be the best template structure for homology modeling of the target sequence. The comparative 3D- structure model of *A. hydrophila* GalE was generated by homology-modeling program Swiss-Model. The predicted model of *A. hydrophila* GalE (**Figure 3**) depicted in the form of ribbons is composed of twelve α-helices and eleven β-strands.

The assessment of the predicted model using the Ramachandran plot showed that the modeled structure has 89.2% residues in the most

favorable regions, 10.8% residues occurring in the allowed regions and none of the residues in the disallowed regions. Such figures assigned by Ramachandran plot represent a good quality of the predicted model (**Figure 4**). All Ramachandrans show 6 labelled residues out of 336, whereas chi1-chi2 plots show 0 labelled residues out of 192. The main chain and side chain parameters for all of them were found to be concentrated/convoluted in the 'better' region. No bad contact was detected in the modeled structure. To define a model reliable, the score for G-factor (a log odds score based on the observed distribution of stereochemical parameters such as main chain bond angles, bond length and phi-psi torsion angles) should be above -0.50. The observed G-factor score for the present model was 0.04 for dihedrals bonds, 0.38 for covalent bonds and 0.18 overall. The distribution of the main chain bond lengths and bond angles were 99.9% and 99.1% within the limits, respectively. The modeled *A. hydrophila* GalE structure was also validated by other structure verification servers as such WHAT_CHECK and Verify-3D. For the modeled structure of GalE of *A. hydrophila,* 96.76% of the residues had an averaged 3D-1D score > 0.2 indicating a good quality of modeled structure. **T**he modeled structure of *A. hydrophila* GalE is comparable to the structurally resolved GalE from *Escherichia coli,* wherein structural motifs have been identified to remain conserved. Since *A. hydrophila* has also been reported to infect humans, it is important to compare the depicted model with that of human UDP-Galactose 4-Epimerase, with which it shares only 51% identity. A superimposition of the *A. hydrophila* GalE onto the human epimerase monomer along with UDP-GlcNAc and NADH is shown in **Figure 5**.

Superimposition of modeled structure of *A. hydrophila* GalE onto a *Homo sapiens* GalE monomer (subunit A of PDB entry 1HZJ) matches 335 Cα atoms with rms distance of 1.35Å and there is high conservation of sequence and structure between the two (**Figure 5A**). The core of the GalE subunit is highly conserved in both the structures, with differences confined to active site and some areas distant from the active site. In *Homo sapiens* GalE, three residues Ala305, Ala306 and Cys307 form a beta strand (ß12) in C-terminal domain, whereas corresponding residues Pro297, Ala298 and Tyr299 of *A. hydrophila* GalE form coiled structure. When comparing *A. hydrophila* GalE to *Homo sapiens* GalE, and ignoring single amino acid differences, a stretch of six amino acids GGSLPE make a loop between β2 and α2 in *Homo sapiens* GalE, which is absent in *A. hydrophila* GalE, . Moreover, α2 of *A. hydrophila* GalE adopts a slightly different orientation compared with *Homo sapiens* GalE. *A. hydrophila* GalE also differs from that of fish species (*D. rerio*) by about 47%, therefore, the regions of differences between the two, can be targeted for drug design against the pathogen enzyme.

Superimposition of the catalytic site of the *Homo sapiens* and *A. hydrophila* GalE along with the UDP-GalNAc and NADH is shown in Figure 5B. It is well known, that in addition to catalyzing the interconversion of UDP-galactose and UDP-glucose, the human epimerase is also capable of interconverting UDP-GalNAc and UDP-GlcNAc **[30]**. Markedly, *E. coli* epimerase has not been reported for this activity. It is clear from the superimposition of *A. hydrophila* GalE onto *Homo sapiens* GalE (**Figure 5B**) Tyr299, a conserved residue in *A. hydrophila* GalE as well as in *E. coli* GalE has been is replaced with a Cys307 in *Homo sapiens* GalE. It can be suggested that the substitution of more bulky Tyr299 in the *A.* epimerase with a Cys307 in the human epimerase most likely prohibit UDPGalNAc from binding in the *A. hydrophila* GalE active site as has been reported for the *E. coli* GalE. These points can be taken into consideration for designing suitable inhibitors against *A. hydrophila* GalE.
hydrophila exhibited greater degree of differences between the hosts, fish and human. Structure modeling of the *A. hydrophila* GalE resulted in identification of the structural differences between the GalE of the host and the pathogen. These differences can be targeted for drug design against the pathogen.
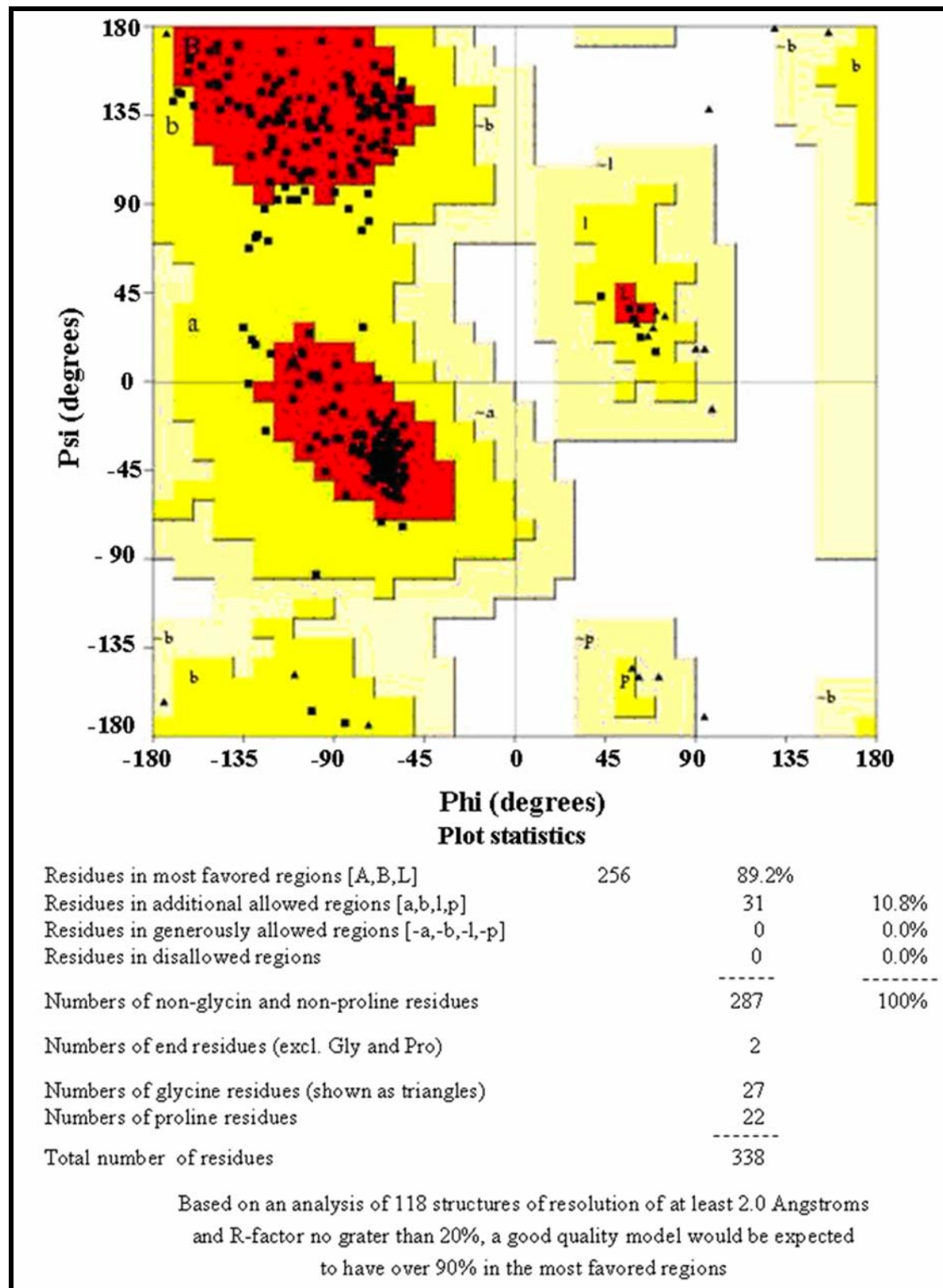
Conclusion
The GalE activity is crucial for lipopolysaccharide biosynthesis, one of the virulent factors of *A. hydrophila*, and GalE mutants exhibit altered core LPS biosynthesis and reduced ability to infect the host cell. Therefore, inhibition of this enzyme can result in controlling *Aeromonas* infection. In the present study, cloning and sequence analysis of GalE of one of the Indian isolate of *A. hydrophila* revealed it to be different from other strains of the bacterium. The GalE of *A.*

Plot statistics

Residues in most favored regions [A,B,L]	256	89.2%
Residues in additional allowed regions [a,b,l,p]	31	10.8%
Residues in generously allowed regions [-a,-b,-l,-p]	0	0.0%
Residues in disallowed regions	0	0.0%
	-------	--------
Numbers of non-glycin and non-proline residues	287	100%
Numbers of end residues (excl. Gly and Pro)	2	
Numbers of glycine residues (shown as triangles)	27	
Numbers of proline residues	22	

Total number of residues	338	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no grater than 20%, a good quality model would be expected
to have over 90% in the most favored regions

Figure 4: Ramachandran plot of the predicted model of *A. hydrophila* GalE: This figure is generated by PROCHECK. The red regions in the graph indicate the most allowed regions whereas the yellow regions represent allowed regions. Glycine is represented by triangles and other residues are represented by squares.
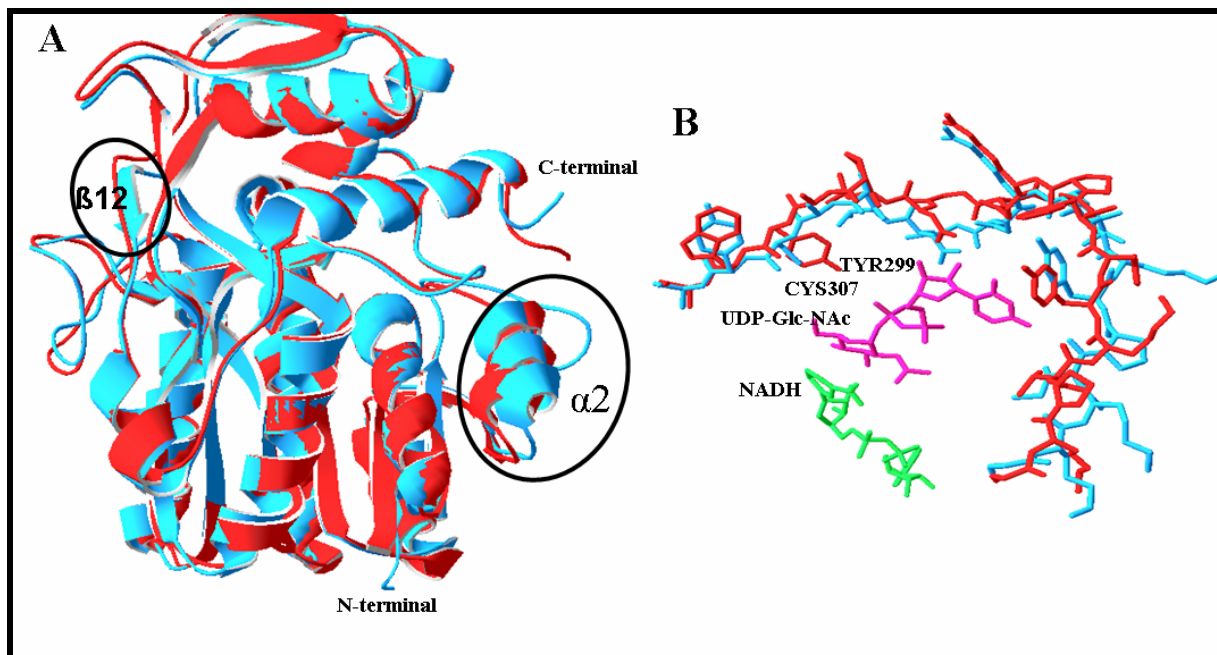
Figure 5: (A) Superimposition of homology modeled structure of *A. hydrophila* GalE onto a *Homo sapiens* GalE monomer. *A. hydrophila* GalE is shown in red and *Homo sapiens* GalE in blue. α2 and β12 correspond to human GalE. Structural differences between the two structures are encircled. (B). Superimposition of the catalytic site of human epimerase/UDPGlcNAc/NADH structure onto the catalytic site of the predicted model of the *A. hydrophila* GalE**.** The human protein is shown in blue with UDPGlcNAc (in CPK color scheme) and NADH (in green) while the *A. hydrophila* protein is depicted in red. Figure shows the substitution of Tyr 299 in the *A. hydrophila* epimerase with a Cys307 in the human enzyme.

References:
[1] JM Janda *Clinical Microbiology Reviews* **4:** 397 (1991) [PMID: 1747858]
[2] JM Janda, SL Abbott, *Clinical Infectious Diseases* **27**: 332 (1998) [PMID:9709884]
[3] S Merino *et al., Microbial Pathogenesis* **20**: 325 (1996) [PMID: 8831828]
[4] YL Zhang *et al., Infection and Immunity* **70**: 2326 (2002) [PMID: 11953367]
[5] DE Pierson, S Carlson, *Journal of Bacteriology* **178**: 5916 (1996) [PMID: 8830687]
[6] DH Kwon *et al., Current Microbiology* **37**: 144 (1998) [PMID: 9662617]
[7] BN Fry *et al., Infection and Immunity* **68**: 2594 (2000) [PMID: 10768949]
[8] HS Houng *et al., Journal of Bacteriology* **172**: 4392 (1990) [PMID: 2198256]
[9] MD Potter, RY Lo, *Infection and Immunity* **64**: 855 (1996) [PMID: 8641792]
[10] HM Holden *et al., Journal of Biological Chemistry* **278**: 43885 (2003) [PMID: 12923184]
[11] R Germanier, E Fuer, *The Journal of Infectious Diseases* **131**: 553 (1975) [PMID: 1092768]
[12] RC Clarke, CL Gyles, *Canadian Journal of Veterinary Research* **50**: 165 (1986) [PMID: 3530414]
[13] J Nesper *et al., Infection and Immunity* **69**: 435 (2001) [PMID: 11119535]
[14] R Canals *et al., Infection and Immunity* **74**: 537 (2006) [PMID: 16369010]
[15] R Canals *et al., Journal of Bacteriology* **189**: 540 (2007) [PMID: 17098903]
[16] T Upadhyaya *et al., DNA sequence: The journal of DNA sequencing and mapping* **18**: 302 (2007) [PMID: 17541836]
[17] JD Thompson *et al., Nucleic Acids Research* **22**: 4673 (1994) [PMID: 7984417]
[18] http://www.rcsb.org/pdb/home/home.do.
[19] T Schwede *et al., Nucleic Acids Research* **31**: 3381 (2003) [PMID:12824332]
[20] http://swissmodel.expasy.org
[21] http://www.expasy.org/spdbv
[22] RA Laskowski *et al., Journal of Applied Crystallography* **26**: 283 (1993)
[23] GN Ramachandran, V Sasisekharan, *Advances in Protein Chemistry* **23**: 283 (1968) [PMID: 4882249]
[24] G Vriend *Journal of Molecular Graphics* **8**: 52 (1990) [PMID: 2268628]
[25] http://swift.cmbi.ru.nl/gv/whatcheck/
[26] D Eisenberg *et al., Methods in Enzymology* **277**: 396 (1997) [PMID: 9379925]
[27] http://nihserver.mbi.ucla.edu/Verify_3D/
[28] Y Liu *et al., Biochemistry* **36**: 10675 (1997) [PMID: 9271498]
[29] JB Thoden *et al., Biochemistry* **39**: 5691 (2000) [PMID: 10801319]
[30] JB Thoden *et al., Journal of Biological Chemistry* **276**: 15131 (2001) [PMID: 11279032]

Edited by P. Kangueane
Citation: Agarwal *et al.,* Bioinformation 4(5): 216-222 (2009)
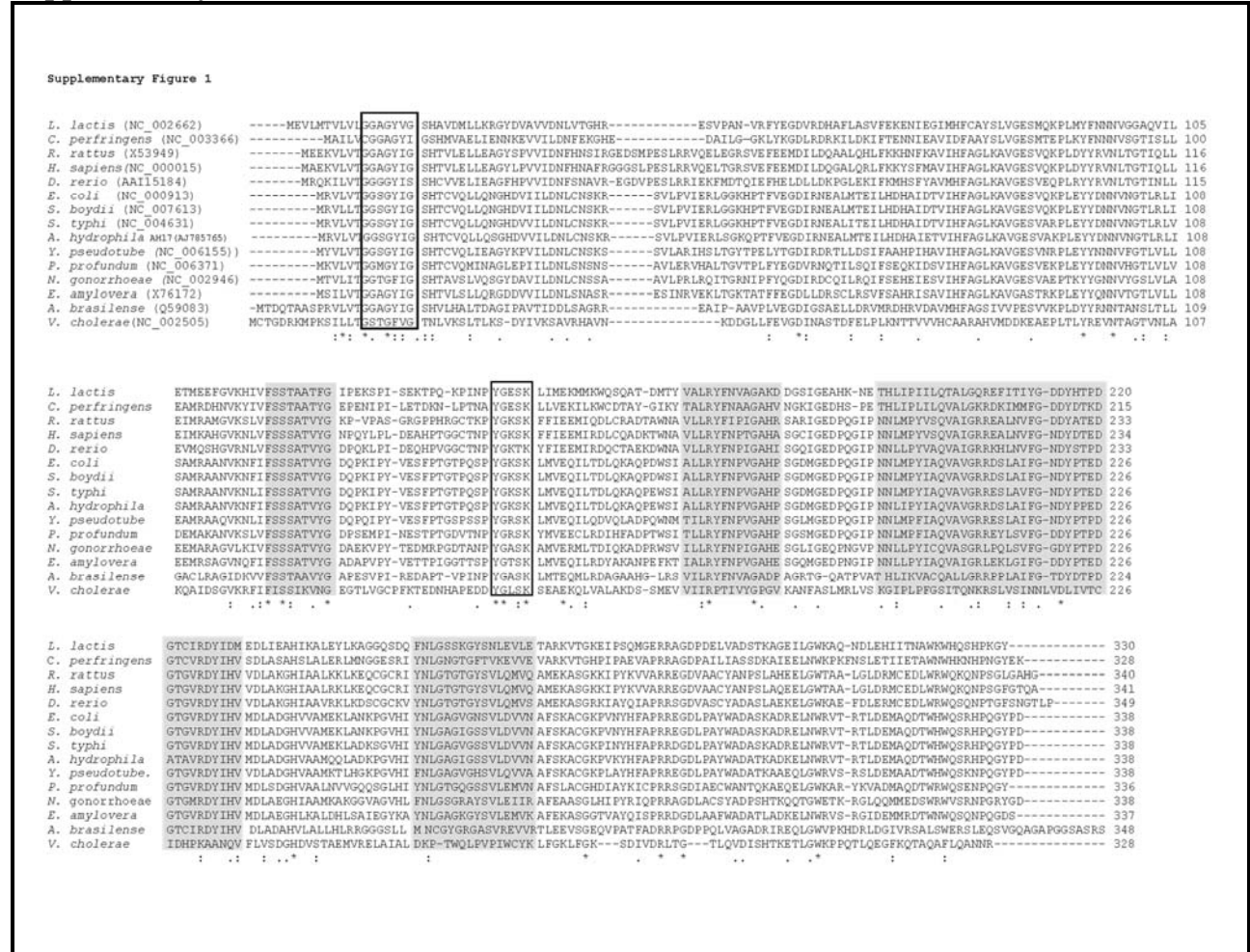
Supplementary material:

Figure 1: Multiple sequence alignment of the deduced amino acid sequence of epimerase.
The deduced amino acid sequence of *A. hydrophila* GalE was aligned with amino acid sequences of GalE from other species. Genbank Accession Numbers are given in bracket in the top section. The signature sequences of the epimerase super family are shown as shaded residues. The conserved N-terminal NAD binding domain (GXXGXXG) and the catalytic couple (TyrXXXLys) are shown as boxed sequences. (*), (:) and (.) denote single fully conserved residues; residues with conserved strong groups and residues with conserved weak groups, respectively.