International Journal of Multidisciplinary Approach and Studies

ISSN NO:: 2348 – 537X

Phylogenetic Analysis and Homology Modelling of Betaine Aldehyde Dehydrogenase - An Aroma Producing Protein in Rice (*Oryza Sativa* L.)

Puspita Hore*, Shamsun Nahar, Mina Hasampura***, & Bhaben Tanti******

Department of Botany, Gauhati University,Assam,India.

ABSTRACT:

Increasing demand of aromatic rice varieties are appreciated by Scientist from the MSU rice Genome Annotation Project (MSU) and the International Rice Genome Sequencing Project (IRGSP)/ Rice Annotation Project Database (RAP-DB) generated a unified assembly of the 12 rice pseudo-molecules of Oryza sativa Japonica Group cv. Nipponbare. Betaine aldehyde dehydrogenase (BADH) is an important enzyme having dual roles in rice fragrance and cereals influencing abiotic stress tolerance. Any mutation event in functional BADH2 leads to a premature termination in the gene producing a truncated protein that results in nullification of the function of the enzyme BADH2 and led to the synthesis of 2AP in fragrant rice varieties. The purpose of this study is to analyze the phylogenetic relationships of the gene responsible for the fragrance in aromatic rice with other closely or distantly related species i.e phylogenetic linkage of 'betaine aldehyde dehydrogenase' and to perform homology protein structure modelling through bioinformatic approaches. Understanding the aroma containing protein in details at molecular level could be used for further crop improvement program including rice to impart aroma through bioinformatics approaches. Hence, further research would be required to characterize the betaine aldehyde dehydrogenase (BADH) protein to be used in rDNA technology.

KeyWords: *Betaine aldehyde dehydrogenase (BADH), phylogenetic linkage, protein structure modeling, Oryza etc.*

INTRODUCTION:

Rice *(Oryza sativa* L.),a model plant of Poaceae family, contributes one third of world population. It is grown in tropical and subtropical region of the world. Domesticated rice is of two distinct types, namely - *Oryza sativa* (Asian rice) and *O. glaberrima* (African rice), both of which have unique domestication histories.

The genus *Oryza* contains 21 wild relatives of the domesticated rice (Vaughan *et al.*,2003). The common rice, *O. sativa* and the African rice, *O. glaberrima* are thought to be an example of parallel evolution in crop plants. The wild progenitor of *O. sativa* is the Asian common wild rice, *O. rufipogon,* which shows a range of variation from perennial to annual types. These two subspecies are commonly associated with differences in growth habitat (Khush, 1997**)** and are the products of independent domestication events from ancestral *Oryza rufipogon* populations in different locations and at different times (Vitte *et al.,* 2004; Ma, Bennetzen,2004; Sang, Ge, 2007a,). In a parallel evolutionary path, *O. glaberrima* was

domesticated from annual *O. breviligulata*, which in turn evolved from perennial *O. longistaminata. O. rufipogon* is distributed from Pakistan to China and Indonesia and its populations vary between perennial and annual types, which differ markedly in life history traits (Okra, 1988).

India is rich in rice production. Almost in all the states like Andhra Pradesh, Bihar, Uttar Pradesh, Madhya Pradesh and West Bengal rice is grown. West Bengal and Uttar Pradesh have the highest rice production. The fragrance or aromatic rice is considered to have huge economic importance that determine the premium price in local and export market. Among them, the Basmati rice of India (Bhattacharjee *et al.*, 2002) and Pakistan and the Jasmine type of rice of Thailand (Berner, Hoff, 1986; Sriboonchitta, Wiboonpongse, 2005) are the aromatic cultivars commonly sold in world trade. Some consumers prefer the flavours and qualities of older stored rice, while other consumers have a preference for fresh rice flavours (Zhou *et al.*, 2002).

Assam being one of the centres of origin has got wide range of variation of rice cultivars. About 70 % of total agricultural land of Assam is used for rice cultivation. The aromatic rice of Assam is a unique class under '*Sali'* rice traditionally known as '*Joha'*. The *Joha* rice cultivars are known for its unique aroma, superfine kernel, good cooking qualities and excellent palatability. Assam maintains a diverse gene pool of aromatic rice that differs in aroma intensity, durability, grain shape and size, production potentialities etc. The main motto of this study is to analyze the phylogenetic relationships of the gene responsible for the fragrance in aromatic rice with other closely or distantly related species and to perform homology protein i.e., 'betaine aldehyde dehydrogenase' modelling through bioinformatic approaches. Major gene for fragrance (*badh2*) of rice was identified on chromosome 8 (Bradbury *et al.*, 2005.). An 8 bp deletion in the exon 7 of this gene was reported to result in truncation of betain aldehyde dehydrogenase enzyme whose loss of function lead to the accumulation of a major aromatic compounds, 2-acetyl 1-pyrroline *(2AP)* in fragrant rice.

MATERIAL AND METHOD:

Software and databases used in this study were: Uniprot/TrEMBL, NCBI Ref Seq database, RCSB PDB, PMDB (Protein Modelling Database), PROCHECK, WHATCHECK, ERRAT2, ClustalX, Modeller and PHYLIP.

Amino acid sequence for target protein of BADH from *Oryza sativa* has been retrieved from RAP-DB (http://rapdb.dna.affrc.go.jp) having an Accession No Q84LK3 which is located in chromosome no 8 of *Oryza sativa var. japonica* in aromatic rice. A suitable template is selected in BLASTP from NCBI (National Centre for Biotechnology information) in search of homologous structure with respect to the query sequence.

In this study, name of organisms, max score, total score, e-value, identity**,** accession no. etc. were recorded.A database is scanned for homologous sequences and then the query sequence is aligned with them to identify residue-residue correspondence between them, here CLUSTALX2 (Stamatakis *et al.*,2005).and MEGA5

(Felsenstein,1981,1985).softwares are used to aligned. Phylogenetic tree was then constructed(Nei , Kumar 2000). Calculation were done on the basis of pairwise distance between the taxa and overall mean distance

Homology of the target protein sequence to the known 3D structures present in the databases like RCSB PDB is searched through three main classes of protein comparison methods that are useful in fold identification. BLAST & FASTA, PSI-BLAST & PDB-BLAST and Threading (target protein sequence is matched against a library of 3D profiles or threaded through a library of 3D folds).The target protein is input using default parameters but the protein can only be specified as a UniProtKB/SWISS-PROT or UniProtKB/TrEMBL accession no or as sequence identifier (ID) or as sequence of amino acid. After inputing the sequence,click on to "compute parameter" giving the result like number of amino acid, Molecular weight, Theoretical pI, Amino acid composition, Atomic composition, Formula, Total no of atom, Extinction co-efficient, Estimated half-life, Instability index, Aliphatic index and Grand average of hydropathicity (GRAVY) were recorded.

RESULTS AND DISCUSSION:

Amino acid sequence for target protein of betaine aldehyde dehydrogenase (BADH) from *Oryza sativa* was retrieved from RAP-DB (http://rapdb.dna.affrc.go.jp) having an Accession No Q84LK3 which is located in chromosome no. 8 of *Oryza sativa var. japonica* in aromatic rice. BLAST was conducted by BLASTp analysis of *BADH* gene sequence of Nipponbare rice (*O. sativa var. japonica)* genome. A total of 11sequences were selected out of which 6 sequences showed more than 80% similarity with the query sequences (Table 1). It is interesting to note that all the selected sequences belong to the Poaceae family. Their E-values, identity percentage and accession number were noted down. Then, a notepad file was created where at first the query sequence copied from RAP-DB in FASTA format was pasted. Then by clicking on the accession number of one of the selected sequences, the flat file of that organism was obtained from where the protein sequence was copied in FASTA format and pasted in the notepad file just below the protein sequence of the query sequence. Similarly, this process was repeated for all the selected sequences and the sequences were pasted one after the other serially in the notepad file.

TABLE 1 Selected organisms along with their e-value, identity percentage and accession number in blastp.

Organism	**Query Sequence**	**E-Value**	**Identity percentage**	**Accession Number**
Oryza brachyantha	99%	0.0	90%	XP015695605.1
Brachypodium distachyon	98%	0.0	90%	XP003574495.1
Sorghum bicolor	99%	0.0	89%	AGZ15751.1
Zoysia tenuifolia	99%	0.0	87%	BAD34948.1
Triticum aestivum	99%	0.0	88%	AAL05264.1
Hordium vulgare	98%	0.0	88%	BAB62846.1
Arabidopsis thaliana	99%	0.0	76%	NP001185399.1

Populus euphratica	99%	0.0	76%	NP001291240.1
Jatropha curcas	99%	0.0	75%	AFY98894.1
Gossypium hirsutum	99%	0.0	78%	AAR23816.2
Zea mays	99%	0.0	77%	ACG29220.1

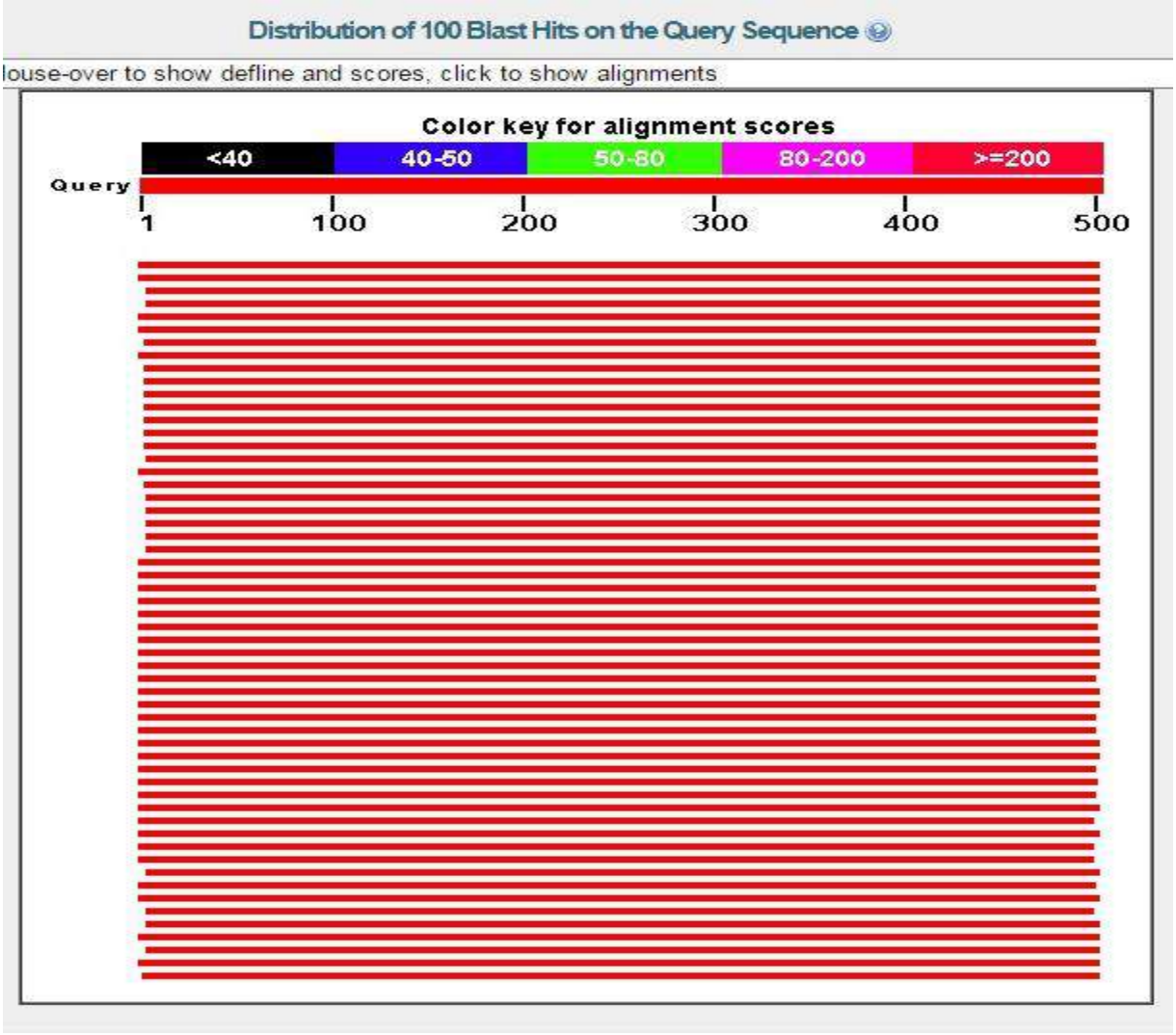

Fig. 1 Distribution of 100 blast hits on the query sequence

Multiple sequence alignment was performed amongst 11 different organisms containing *BADH* genes obtained from the study using MEGA5. Fig. 2 is a part of the multiple sequence alignment of the selected sequences from 1505-1545 base pairs. All the sequences are almost similar in their base composition. The gaps in the alignment are used to optimize the alignment. The numbers of base substitutions per site from between sequences were shown.

Analyses were conducted using the Poisson Correction Model. The analysis involved 11 amino acid sequences. All position containing gaps and missing data were eliminated. There were total of 463 positions in the final data set. Evolutionary analyses were conducted in MEGA5

Fig. 2 Multiple sequence alignment

```
                ....|....| ....|....| ....|....| ....|....| ....|....|
          1505      1515      1525      1535      1545
O.sativa CTGAGCGTGA AACAGGTGAC CGAATATGCG AGCGATGAAC CGTGGGGCTG
Oryza brac CTGACCGTGA AACAGGTGAC CGAATATGCG AGCGATGAAC CGTGGGGCTG
Triticum CTGAGCATTA AACAGGTGAC CGAATATACC AGCGATGCGC CGTGGGGCTG
Hordeum CTGAGCATTA AACAGGTGAC CGAATATACC AGCGATGCGC CGTGGGGCTG
Zoysia CTGAACGTGA AACAGATTAC CGAATATACC AGCGATGAAC CGTGGGGCTG
Sorghum CTGAGCGTGA AACAGGTGAC CGAATATATT AGCGATGAAC CGTGGGGCTG
Zea CTGACCGTGA AACAGGTGAC CAAATATTGC AGCGATGAAC CGTGGGGCTG
Populus CTGAGCGTGA AACAGGTGAC CCAGTATATT AGCGAAGAAC CGTGGGGCTG
Jatropha CTGAGCGTGA AACAGGTGAC CCAGTATATT AGCAACGAAC CGTGGGGCTG
Gossypium CTGAGCGTGA AACAGGTGAC CCAGTATGTG AGCGATGAAC CGTGGGGCTG
Arabidop CTGAGCGTGA AACAGGTGAC CCTGTATACC AGCAACGATC CGTGGGGCTG
```

Table 2 Estimation of evolutionary divergence among the sequences

	1	2	3	4	5	6	7	8	9	10	11
1. Oryza sativa japonica											
2. Oryza brachyantha	0.033										
3. Zoysia tenuifolia	0.129	0.131									
4. Sorghum bicolor	0.114	0.114	0.114								
5. Hordeum vulgare	0.124	0.131	0.124	0.129							
6. Triticum aestivum	0.124	0.131	0.126	0.129	0.011						
7. Arabidopsis thaliana	0.260	0.263	0.274	0.263	0.274	0.280					
8. Populus euphratica	0.274	0.283	0.286	0.288	0.291	0.294	0.184				
9. Jatropha curcas	0.235	0.238	0.260	0.254	0.271	0.274	0.174	0.109			
10. Gossypium hirsutum	0.243	0.249	0.271	0.246	0.283	0.283	0.172	0.166	0.141		
11. Zea mays	0.263	0.266	0.294	0.280	0.306	0.300	0.318	0.291	0.280	0.297	

Table 2 is a matrix showing the pair-wise evolutionary distance between the organisms. For example, the distance between *Oryza brachyantha* and *Zoysia tenufolia* is 0.129 and that of *Hordium vulgare* and *Jatropha curcas* is 0.235 and so on. The overall mean distance of the taxa was calculated as 0.126.

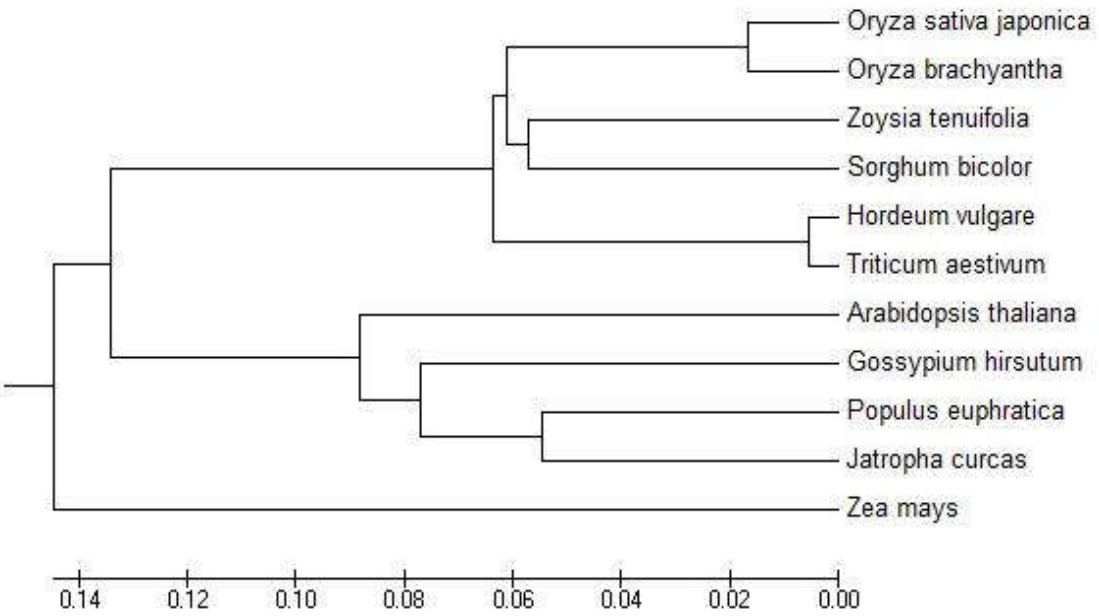

Fig. 3 Phylogenetic tree of the organisms having the gene BADH2

The phylogenetic tree was constructed showed the evolutionary lineage among the selected organisms. Below the tree is a scale showing the mutation rate in the evolutionary line. By the analysis of the table we have found that the line diverged through three distinct clusters from an unknown ancestor. In terms of aroma producing protein, *Zea mays* showed a separate lineage forming a single cluster alone, whereas the others were grouped in separate clusters by the remaining organism. By analysing the *BADH* sequences of the organisms, the phylogenetic tree was constructed as shown in Fig. 3, which clearly revealed the 11 taxa diverged into three clades, *O. sativa, O. brachyantha, Triticum, Hordeum, Zoisia Sorghum,* were formed in 1st clade. *Populus, Jatropha, Gossypium, Arabidopsis* were formed in 2nd clade and the 3rd clade was formed by *Zea mays*. Our query *BADH* gene is closely related to *Zoysia, Triticum, Hordeum,Sorghum, Populus and Jatropha.* The tree could be broken down into nodes (represented in the tree as circles) and branches (lines connecting them). Nodes represented common ancestors for the descendents. There were two types of nodes; external nodes, also called 'tips' or 'leaves' and internal nodes. Two descendents that splited from the same node were called sister groups. From Fig. 3, we could infer that *Populus euphratica* and *Jatropha curcas* were sister groups. They had a common ancestor that is unique to them (in terms of *BADH* gene). *Zea mays* constituteed an outgroup (taxa outside the group of interest). All the members of the group of interest were closely related to each other.

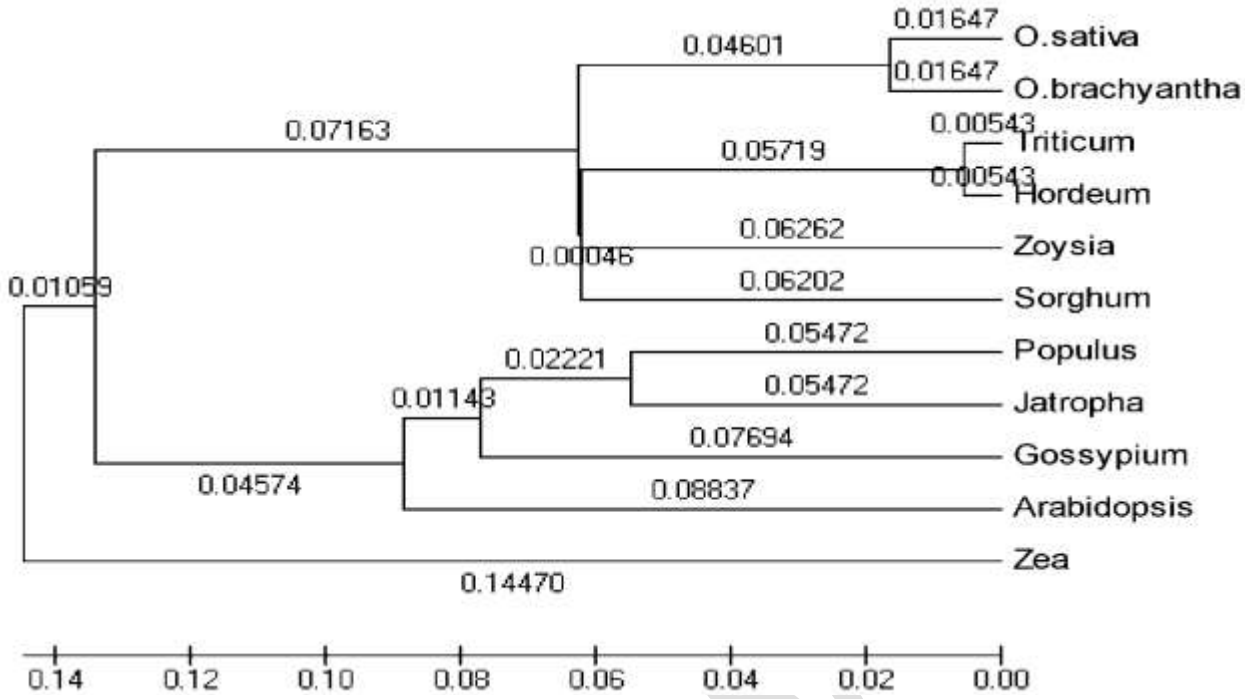

Fig. 4 Phylogenetic tree showing the branch length

In Fig. 4, the horizontal dimension gave the amount of genetic changes. The horizontal lines are branches and represent evolutionary lineages changing over time.

The bar at the bottom of the figure provided a scale. In this case, each unit segment of the line had a measure of 0.02. This showed the length of branch that represented an amount of genetic changes of 0.02. The units of branch length were usually protein substitutions per site i.e. the number of changes or 'substitutions' divided by the length of the sequence. The vertical dimension in this figure had no meaning and was used simply to lay out the tree visually with the labels evenly spaced vertically. Here, we could see that the branch length for the divergence of our query sequence from the ancestor that was common to Populus *euphratica* and *Jatropha curcus*, *Triticum aestivum* and *Hordeum vulgare* and *O. sativa* and *O. brachyantha* were 0.05472, 0.00543 and 0.01647 respectively.

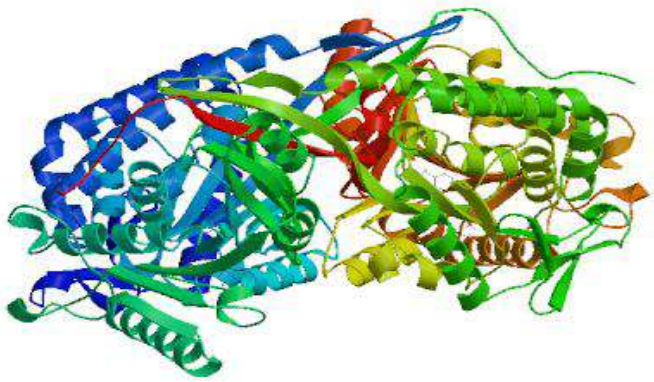

Fig. 5 Predicted 3D structure of the target protein sequence i.e., betaine aldehyde dehydrogenase (BADH) from *Oryza sativa*

The structure of the target protein having a similarity with the template of **4i84p.1 A (Zea mays)**
Sequence identity **- 89.98%**
Resolution **- 1.95Å**
Sequence similarity **- 0.59**
Range **- 5-503**
Coverage **- 0.99**

The general property of the protein BADH of *Oryza sativa* was predicted through the online tool ProtParam. This tool could be accessed from the Expasy site. The obtained result was given in the following.

Number of amino acid: 503
Molecular weight: 54682.7
Theoretical pI: 5.36
The amino acid composition of the target protein was:
Ala (A) 56 11.1% ,Arg (R) 25 5.0% ,Asn (N) 12 2.4% ,Asp (D) 23 4.6% ,Cys (C)14 2.8%, Gln (Q) 13 2.6%, Glu (E) 43 8.5% ,Gly (G) 43 8.5% ,His (H) 4 0.8% ,Ile (I) 26 5.2% , Leu (L) 42 8.3% ,Lys (K) 32 6.4%,Met (M) 8 1.6% ,Phe (F) 15 3.0% ,Pro (P) 30 6.0% , Ser (S) 29 5.8% ,Thr (T) 24 4.8% ,Trp (W) 13 2.6% ,Tyr (Y) 11 2.2% ,Val (V) 40 8.0%, Pyl (O) 0 0.0% ,Sec (U) 0 0.0% ,(B) 0 0.0% ,(Z) 0 0.0% ,(X) 0 0.0%
Total number of negatively charged residues was found to be (Asp + Glu) 66 and positively charged residues (Arg + Lys) was 57. The atomic composition was revealed as below with the formula = C2440H3852N656O725S22 and the total number of atoms were 7695. Extinction coefficients were measured in water in units of M-1 cm-1 at 280 nm.

Ext. coefficient was calculated as 88765. Abs 0.1% (= 1 g/l) 1.623, assuming all pairs of Cys residues form cystines.

Again, Ext. coefficient was calculated as 87890. Abs 0.1% (=1 g/l) 1.607, assuming all Cys residues are reduced.

The half life of the predicted protein was estimated as below.

The N-terminal of the sequence considered is M (Met)
The estimated half-life is: 30 hours (mammalian reticulocytes, *in vitro*).
>20 hours (yeast, in vivo).
>10 hours (Escherichia coli, in vivo).
The instability index (II) was computed to be 35.87, which classified the protein as stable with the aliphatic index as 86.92 and grand average of hydropathicity (GRAVY) as -0.124.

CONCLUSION:

From the overall analysis, we could conclude that, in terms of *BADH* gene, the two rice *viz.*, *O. glaberrima* (wild rice) and *O. sativa* (cultivated rice) are different from each other. Both the species have already been distinguished by different attributes like morphological, physiological and genetically. Here, very interesting observation was observed that with respect to aroma producing *BADH* gene, both the species were showing ~55% differences. Further, the entire aroma producing gene sequences distinctly grouped into three major

clusters. In terms of *BADH* gene, the query sequence has highest distance with *Zea mays* (0.14470) with respect to *BADH* gene. The query gene also shares common ancestry with *Triticum and Hordeum*. Amongst the sequences of study, *Zea mays* are significantly different from the other taxa and the model presented here can serve as a guide for the allocation of amino acid residues involved in each fold, which is important for further investigations on active sites and molecular mechanism of function. The study was performed for sequence analyses and prediction of 3D structure of *BADH* gene *O. sativa* using the comparative (homology) modelling due to high level sequence identity. Understanding the aroma containing protein in details at molecular level could be used for further crop improvement program including rice to impart aroma through bioinformatics approaches. Hence, further research would be required to characterize the betaine aldehyde dehydrogenase (BADH) protein to be used in rDNA technology.

REFERENCES

i. Berner D.K., Hoff B.J. (1986). Inheritance of scent in American long grain rice. *Crop Science*, **26:** 876-878.

ii. Bhattacharjee P, Singhal R.S., Kulkarni P.R. (2002). Basmati rice*:* a review. *International Journal of Food Science and Technology,* **37:** 1-12.

iii. Bradbury L.M.T., Fitzgerald T.L., Henry R.J., Jin Q.S., Waters D.L.E. (2005). The gene for fragrance in rice. *Plant Biotechnology,* **3:** 363-370.

iv. Felsenstein J. **(**1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17:** 368–376.

v. Felsenstein J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, **39:** 783–791.

vi. Khush G.S. (1997). Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.*, **35:** 25–34.

vii. Ma J., Bennetzen J.L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci.* **101:** 12404–12410.

viii. Nei M., Kumar S. (2000). Molecular evolution and phylogenetics. *Oxford: Oxford University Press*.

ix. Okra H.I. (1988). Indica-japonica differentiation of rice cultivars. In: Origin of cultivated rice. Tokyo/Amsterdam. *Japan Science Society Press/Elsevier* pp. 141–179.

x. Sang T., Ge S. (2007a). Genetics and phylogenetics of rice domestication. *Curr. Opin.Genet. Dev.*, **17:** 1–6.

xi. Sriboonchitta S., Wiboonpongse A. (2005). On the Estimation of Stochastic Production Frontiers with Self-Selectivity*:* Jasmine and Non-Jasmine Rice in Thailand. *Chiang Mai University Journal* ,**4:** 105-124.

xii. Stamatakis A., Ludwig T., Meier H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456–463.

xiii. Vaughan D.A., Morishima H., Kadowaki K. (2003). Diversity in the *Oryza* genus. *Current Opinion in Plant Molecular Biology,* **6:** 139–146.

xiv. Vitte C., Ishii T., Lamy F., Brar D., Panaud O. (2004). Genomic paleontology provides evidence for two distinct origins of Asian rice (*Oryza sativa* L.). *Mol. Genet. Genomics,* **272:** 504–511.

xv. Zhou Z., Robards K., Helliwell S., Blanchard C. (2002). Ageing of stored rice: Changes in chemical and physical attributes. *Journal of Cereal Science,* **35:** 65-78.