Global Journal of Molecular Sciences 5 (1): 30-36, 2010
ISSN 1990-9241

Secondary Structure Prediction and Phylogenetic Analysis of Salt Tolerant Proteins

[2]*Prashant V. Thakare,* [1]*Uddhav S. Chaudhari,* [1]*Madura S. Makhe,*
[1]*Vishal P. Deshmukh and* [1]*Renuka R. Kurtkoti*

[1]Bioinformatics Laboratory, Department of Botany,
Sant Gadge Baba Amravati University, Amravati. 444 602, India
[2]Genetic Engineering Laboratory, Department of Biotechnology,
Sant Gadge Baba Amravati University, Amravati. 444 602, India

Abstract: Secondary structure prediction and phylogenetic analysis of 14 salt tolerant proteins from different fungi and plants was studied. The secondary structure was predicted by SOPMA method, this method calculate the content of α-helix, β-sheets, turns, random coils and extended strands. The content of α-helix in *A. thaliana* was highest (80%) and lowest (20.16%) found in *C. albicans*. The percentage of ß-turn was the highest in *T. turgidum* (11.76%) and lowest in Arabidopsis thaliana (3.50%). Percentage of extended strands was lowest in *A. thaliana* (0.50%) and highest percentage of extended strands was 29.03% found in *C. albicans*. The percentage of random coil was highest in *C. albicans* (40.75%) and lowest percentage of random coils was 16%, which was found in *A. thaliana*. There were no b-bridges (Bb) found in all the 14 species of salt tolerant plants. Phylogenetic analysis of 14 salt tolerant proteins was performed by using Phylodraw. On cladogram *A. thaliana* was nearest to the origin. *G. hirsutum* and Nax2 forms a cluster, having a largest root distance 0.48675 and *T. aestivum* and Nax1 forming a cluster with lowest root distance and pair distance of 0.46824 and 0.00000 respectively. *A. thaliana* was completely outgrouped, whereas *H. turgidum* was outgrouped but still showing integrity with remaining 12 species under study.

Key words: CLUSTAL-X · Phylogenetic analysis · Salt tolerant protein · Secondary structure · SOPMA

INTRODUCTION

Presence of excessive amount of soluble salts in soil is called as saline soil [1]. The relative growth of plants in the presence of salinity is termed as salt tolerance and the plants are called as salt tolerant plants. Salt tolerant plants either prevent the absorption of sodium or chloride ions by roots and leaves or tolerate the collection of sodium or chloride ions in its tissue [2]. Large numbers of salt tolerant proteins are found in plants and passes specific role to overcome the salinity. Salt tolerant proteins are synthesized in response to salinity and the sequences of these proteins are found to be highly conserved. Many salt tolerant proteins, their roles, activities and location in plant were mentioned in *Arabidopsis thaliana* [3]; *AtNHX1* (*Arabidopsis thaliana* protein) [4]; *Gossypium hirsutum* [5]; *Oryza sativa* [6,7]; *Zygosaccharomyces rouxii* [8]; *Triticum turgidum* [9].

Bioinformatics has revolutionized the field of molecular biology. The raw sequence information of proteins and nucleic acid can convert to analytical and relative information with the help of soft computing tools. Protein prediction is important application of bioinformatics. Studies on genetic relationships and generation of evolutionary tree by comparing amino acid sequences and nucleotide acid sequences have been successively carried out in many species [10]. As single change in sequence of amino acid leads to conformational changes of proteins, hence the information gathered from comparison of members of same protein family can throw some light on path of their evolution.

Among the various softwares available, Multiple Sequence Alignments (MSA) is the most important tool for analysis of amino acid and nucleic acid sequence data. MSA is found to be vital tool in determination of homologies in sequences, analyzing the sequence structure similarity and for phylogenetic analysis [11].

Corresponding Author: Prashant V. Thakare, Genetic Engineering Laboratory, Department of Biotechnology,
Sant Gadge Baba Amravati University, Amravati. 444 602, India.
E-mail: prashantthakare123@rediffmail.com.

SOPMA is a self optimizing prediction methods of alignment and is used for prediction of secondary structure of proteins. This method calculates the content of α-helix, β-sheets, turns, random coils and extended strands. SOPMA method predicts 69.5% of amino acids. The prediction of protein secondary structure is improved by 9% to 66%. SOPMA is neural network based methods; global sequence prediction may be done by this sequence method [12].

The present investigation was under taken to study the phylogenetic relationships in salt tolerant proteins of 14 eukaryotic organisms and secondly to predict secondary structure of these salt tolerant proteins.

MATERIALS AND METHODS

Softwares: Windows operating system 98/2000/NT; CLUSTAL-X; GeneDoc; PhyloDraw 8.0; SOPMA (online software); Ms-Word; Paint programme (Bitmap BMP image).

Protein Sequences of Salt Tolerant Proteins: The amino acid sequences of salt tolerance proteins were downloaded from organisms such as *Arabidopsis thaliana, Aspergillus niger, Avicennia marina, Candida albicans, Gossypium hirsutum, Horedum vulgare, Nax1* (*Triticum monoccum* gene), *Nax2* (Wheat gene), *Triticum turgidum, Triticum aestuivum, YDR456W* (Yeast protein), *Saccharomyces cervisae, YLR138W* (Yeast protein), *Zygosaccharomyces rouxii* from NCBI (National Center for Biotechnology Institute) from website http://www.ncbi.nlm.nih.gov/ by giving key words salt tolerant protein.

Methods:

Collection of Sequences Data:

- Amino acid sequences of salt tolerant proteins were downloaded from NCBI (National Center for Biotechnology Institute) by giving salt tolerant as a key word.
- Protein sequences were saved in FASTA format.

Multiple Sequence Alignment:

- CLUSTAL-X was used for sequence alignment.
- FASTA format sequences were loaded into CLUSTAL-X
- Sequences were aligned using command 'Do complete alignment'.
- Alignment results were saved as 'xxx.aln'.

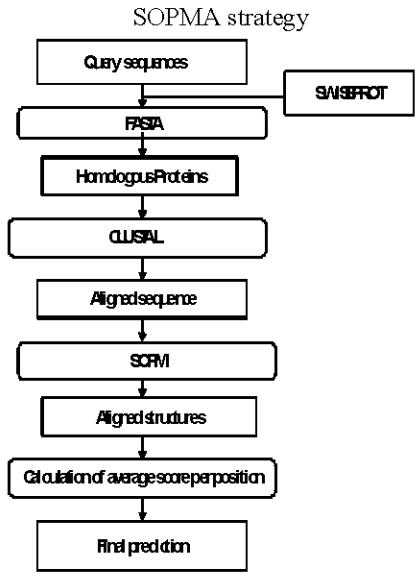


Fig. 1: Logical strategy of SOPMA method (Source: Geourjon and Deleage, 1995)

Determination of Conserved Regions:

- The conserved regions were compared and determined in GeneDoc.
- Clustal sequences were imported to GeneDoc.
- Results were obtained by selecting 'done' option and saved as 'xxx.msf'.

Phylogenetic Relationships/ Analysis:

- Phylogenetic analysis was performed using software's like phylodraw or tree view. These phylogenetic softwares are forwards and files with xx.ph can be imported in phylodraw using rectangle cladogram. The phylogenetic clusters can be visualized along with root distance and pair distance.
- The files were saved as picture image as '.bmpor', '.TIFF' image.

Structural Prediction of Proteins from SOPMA:

- SOPMA (Figure 1) is a secondary structure prediction software which available online on web address, www.expasy.ch.
- After opening the desired web page, the amino acid sequence in FASTA format was imported in a particular SOPMA window and submitted to SOPMA server.
- The results appeared with secondary structure and with percentage of each secondary structure of proteins.

RESULTS

To start with present study the amino acid sequences of 14 salt tolerant proteins (Table 1) were retrieved in the FASTA format. Among the 14 salt tolerant proteins largest protein was *YL138W* (Yeast protein) with 985 amino acids while, the smallest protein was *T. turgidum* with 153 amino acids. The secondary structure prediction of salt tolerant proteins among 14 species was obtained by using SOPMA online secondary structure prediction server.

The content of α-helix in *A. thaliana* was highest (80%) and lowest (20.16%) found in *C. albicans* (Table 2). The percentage of β-turn was the highest in *T. turgidum* (11.76%) and lowest in *A. thaliana* (3.50%). Percentage of extended strands was lowest in *A. thaliana* (0.50%) and highest percentage of extended strands was 29.03% found in *C. albicans*. The percentage of random coil was highest in *C. albicans* (40.75%) and lowest percentage of random coils was 16%, which was found in *Arabidopsis thaliana* (Figure 2).

By observing the colours, structural differences could be identified with respective amino acid changes in an individual species. On the basis of α-helical regions, *H. vulgare* was homologus to *Nax1* (*Triticum monoccum* gene), while *T. aestivum* was homologus to Yeast protein i.e. *YDR456W* (Table 2). *C. albicans* was homologus to *T. turgidum*, whereas *A. niger* was homologus to *Z. rouxii*. There were no β-bridges (Bb) found in all the 14 species of salt tolerant plants. The β-turns (Tt) of the *H. vulgare* was homologus to *Nax1*. Extended strands (Ee) in *A. thaliana* were found to be zero.

Table 1: Sequences retrieved from NCBI data repositories

S.No	Organism	Gi Number
1	*Arabidopsis thaliana*	Q65XN9gi|75115288|
2	*Aspergilus niger*	CAK42775gi|134083012|
3	*Avicennia marina*	AAZ04239gi|69880088|
4	*Candida albicans*	CAA09498gi|3702407|
5	*Gossypium hirsutum*	AAM54141gi|22902099|
6	*Horedum vulgare*	BAC54275gi|27531337|
7	*Nax1*	ABK41857gi|117583138|
8	*Nax2*	ABG33946gi|109452932|
9	*Triticum turgidum*	AAY26389gi|63021412|
10	*Triticum aestivum*	ABK41857gi|117583138|
11	*YDr456W*	NP_010744gi|6320663|
12	*Saccharomyces cervisae*	CAN08430gi|147223265|
13	*YLr138W*	CAA97709gi|1360557|
14	*Zygosacchromyces rouxii*	P24545gi|114348|

Phylogenetic analysis of 14 salt tolerant proteins was performed by using phylodraw. The resultant cladogram was divided in to two distinct clusters (Figure 3). On cladogram *A. thaliana* was nearest to the origin and is placed separately forming separate cluster with root distance 0.46023 and pair distance 0.92702. In second cluster, two sub-clusters were formed. Sub-cluster I consisted of only *H. vulgare* and was

Table 2: Percentage of amino acids sequence forming secondary structure in SOPMA prediction

S.No.	Name of species	Number of amino acids	α-helix (Hh)(%)	β-turns (Tt) (%)	Extended strands (Ee) (%)	Random coils (%)
1	*Arabidopsis thaliana*	200	80	3.50	0.50	16
2	*Aspergillus niger*	342	48.83	4.09	8.77	38.30
3	*Avicennia marina*	261	52.49	6.13	14.18	27.20
4	*Candida albicans*	248	20.16	10.08	29.03	40.73
5	*Gossypium hirsutum*	543	42.54	4.60	18.23	34.62
6	*Horedum vulgare*	352	46.02	8.81	15.91	29.26
7	*Triticum turgidum*	153	31.37	11.76	28.76	28.10
8	*Triticum aestivum*	554	44.95	6.32	16.79	31.95
9	*NaX1*	352	46.02	8.81	15.91	29.26
10	*NaX2*	517	37.52	6.00	19.92	36.56
11	*Sacchromyces cervisae*	351	47.01	6.55	11.97	34.47
12	*YDR456W*	633	42.34	4.74	16.59	36.33
13	*Ylr138W*	985	35.53	8.02	19.39	37.06
14	*Zygosacchromyces rouxii*	920	46.20	5.33	17.83	30.65

Table 3: Clusters form on cladogram with their root and pair distance

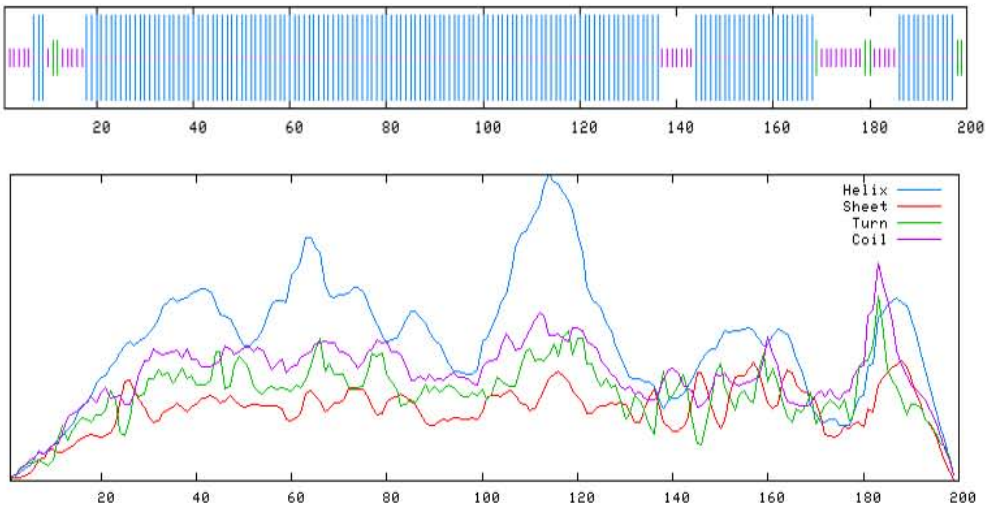
Sn.No	Cluster among Organisms	Root distance	Pair Distance
1	*A. thaliana* (nearest to origin)	0.46023	0.92702
2	*H. vulgare*	0.468240	----
3	*A.niger/ C. albicans*	0.478290	0.905000
4	*S. cervisae / YLR138W*	---	0.910000
5	*T. aestivum/ Nax1*	0.46824	0.000000
6	*A.marina (differed)*	0.473240	0.949270
7	*G. hirsutum/ Nax2*	0.48675	0.900000

```
          10        20        30        40        50        60        70
          |         |         |         |         |         |         |
MASHQDQASYRAGETKAHTEEKAGQVMGASKDKASEAKDRASEAAGHAAGKGQDTKEATKEKAQAAKERA
ehccccchhhcttccccchhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
SETAQAAKDKTSSTSQAARDKAAESKDQTGGFLGEKTEQAKQKAAETAGAAKQKTAETAQYTKDSAIAGK
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhccc
DKTGSVLQQASEQVKSTVVGAKDAVMSTLGMTEDEAGTDDGANKDTSATAAATETTARDH
ccchhhhhhhhhhhhhhhhhhhhhhhhhtccccccccccttccccchhhhhhhhhhhhtt
```

Sequence length : 200

SOPMA:

Alpha helix (Hh)	:	160 is 80.00%
3_{10} helix (Gg)	:	0 is 0.00%
Pi helix (Ii)	:	0 is 0.00%
Beta bridge (Bb)	:	0 is 0.00%
Extended strand (Ee)	:	1 is 0.50%
Beta turn (Tt)	:	7 is 3.50%
Bend region (Ss)	:	0 is 0.00%
Random coil (Cc)	:	32 is 16.00%
Ambigous states (?)	:	0 is 0.00%
Other states	:	0 is 0.00%



Parameters:

Window width	:	17
Similarity threshold	:	8
Number of states	:	4

Fig. 2: Secondary structure prediction of *Arabidopsis thaliana* from SOPMA server

diverged with root distance 0.468240. Sub-cluster II comprised remaining 12 species and was divided in to two small sub-sub clusters. *G. hirsutum* and *Nax2* formed a cluster, having a largest root distance 0.48675 and *T. aestivum* and *Nax1* forming a cluster with lowest root distance and pair distance of 0.46824 and 0.00000 respectively. *A. marina* was differed and has a largest pair distance 0 .9492 (Table 3). Though *H. vulgare* showed integrity with two clusters it was outgroup.

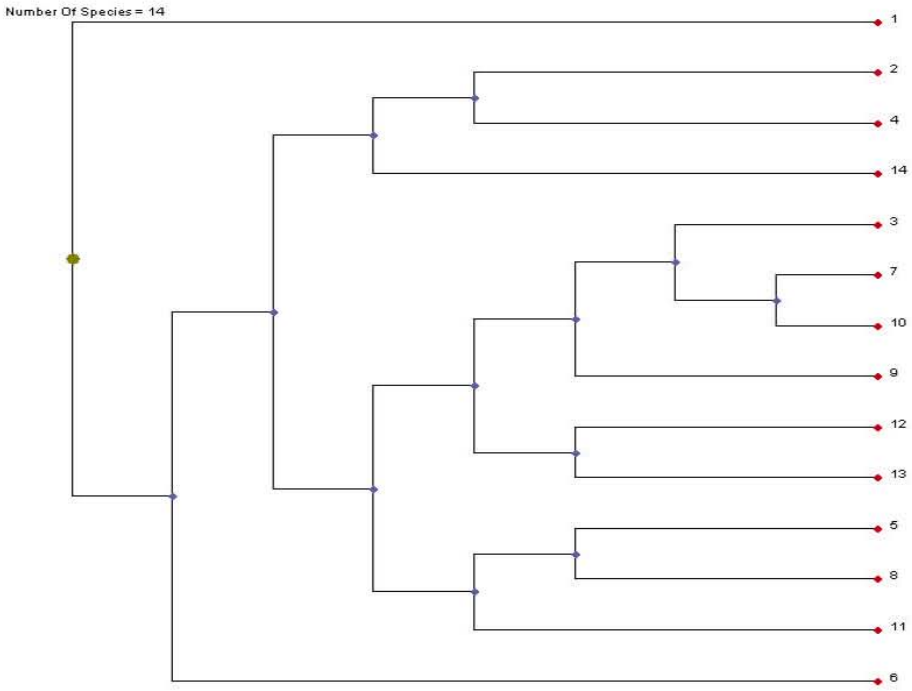


Fig. 3: Cladogram of 14 salt tolerant proteins
1-*Arabidopsis thaliana*; 2- *Aspergillus niger*; 3- *Avicennia marina*; 4- *Candida albicans*; 5- *Gossypium hirsutum*; 6- *Horedum vulgare*; 7- *Nax1* (Wheat gene); *8- Nax2* (Wheat gene); *9- Triticum turgidum; 10- Triticum aestuivum; 11- YDR456W* (Yeast protein); *12- Saccharomyces cervisae; 13- YLR138W* (Yeast protein); *14- Zygosaccharomyces rouxii*

DISCUSSION

Multiple sequence alignment (MSA) is of fundamental importance in all aspects of DNA and protein sequence analysis. It is used as a first and critical step in protein structure prediction, classification, phylogenetic reconstruction analysis of protein domain and identification of functional sites in genomic sequences, to mention just a few important applications [13]. The assembly of MSA has become of the most common tasks while dealing with sequence analysis [14] and it became useful technique for studying molecular relationship [11]. Sarin *et al*. [15] used PSI-BLAST for interation output in homologous protein of the *Presniline* in different organisms for establishment of phylogenetic relationship among them. In present investigation FASTA format of amino acid sequences was preferred rather than BLAST, as Gi number is present in former format where as absent latter. A versatile sequence colouring scheme allows the user highlight conserved feature in the alignment [16]. The present investigator used N-J method for estimation of phylogenetic trees. Generally UPGMA was not used because, it performed well only when branch lengths were close in length [17].

In present study PhyloDraw was applied for viewing the phylogenetic tree because, it supports various kinds of multi alignment formats (Dialign 2, CLUSTAL-W, PHYLIP format, NEXUS, MEGA and pairwise distance matrix) and various kinds of tree diagrams i.e. rectangular cladogram, slanted cladogram, phylogram, unrooted tree and radial tree can be visualized. By using several control parameters, users can easily and interactively manipulate the shape of phylogenetic trees [18]. In dendrogram, two distinct clusters were formed and the homologies between related species was observed. Rice proteins like *OsHKT1* and *OsHKT2* were closer and formed a cluster, similar observations were noted by [19] where *Eucalyptus* proteins *EcHKT1;1* and *EcHKT1;2* forming similar cluster. Based on salt tolerant gene of *Nax1* (wheat gene) and *T. aestiuvum* were closer and formed one cluster, similarly *YLR138W* (Yeast protein) and *S. cervisae* were also placed in one cluster. Most of the proteins from similar origin were placed in clusters, the only exception is *G. hirsutum* and *Nax2* formed a cluster of different species, one of them was cotton and other was wheat gene. Position of

NhX1 protein was placed another cluster as it was diverged from a common ancestor [3] and *A. thaliana* was outlier in dendrogram, showing its complete divergence from common ancestor.

The secondary structure of proteins can be predicted by various methods like, Self optimized prediction methods of alignment (SOPMA), Predict Heidelberg Deutscland method (PHD). SOPMA predicted the percentage amount α-helix, β-sheets, random coils, extended stands at a time. Birve *et al.* [20] was applied SOPMA for the prediction of secondary structure of proteins in *T. aestivum.* However present study directed towards the prediction of secondary structure of *14* different salt tolerant proteins. The SOPMA method was found to be suitable as it correctly predicts 69.5% of amino acids for three state description of secondary structure viz. α-helix, β-sheet and coil helix transitions [12]. Secondary structure prediction recently has surpassed 70% level of average accuracy evaluated on single residue states helix, strands and loop [21]. The predicted secondary structure of p1 of polymerase1 of human influenza virus consisted of 33% predicted α-helices, 26% β-pleated sheets, 23% β-reverse turns and 18% undefined structure [22] and that of PCOR 33% α-helix, 13% β-sheets and 54% turns and random coils were reported by Birve *et al.* [20]. In contrast to this, it was significant to note that, there was absence of β-bridges in 14 salt tolerant proteins while, α-helix, β-turns, coils and extended strands were present on different amino acid residues.

The overall picture of cladogram reveals that *A. thaliana* was completely outgrouped, whereas *H. turgidum* was outgrouped but still showing integrity with remaining 12 species under study.

REFERENCES

1. Blaylock, A.D., 1994. Soil salinity, salt tolerance and growth potential of horticultural and landscape plants. University of Wyoming Cooperative Extension Service Dept. of Plant, Soil and Insect Science College of Agriculture.
2. Bezona, N., D. Hensley, J. Yogi, J. Tavares, F. Rauch, R. Iwata, M. Kellison and M. Wong, 2001. Salt and wind tolerance of landscape plants for Hawaii. Cooperative extension service CTAHR.
3. Wang, W., Y. Li, Y. Zang, C. Yang, N. Zheng and Q. Xie, 2007. Comparative expression analysis of three genes from the *Arabidopsis* vacuolar Na^+/ H^+ antiporter (*AtNHX*) family in relation to abiotic stresses. Chinese Science Bulletin, 52: 1754-1763.
4. Rus, A., S. Yokoi, A. Sharkhuu, M. Reddy, B.H. Lee, T.K. Matsumoto, H. Koiwa, J.K. Zhu, R.A. Bressan and P.M. Hasegawa, 2001. *AtNHX1* is a salt tolerance determinant that controls Na^+ entry into plant roots. Proceedings of National Academy of Sciences, USA., 98: 14150-14155.
5. Wu, C.A., G.D. Yang, Q.W. Meng and C.C. Zheng, 2004. The cotton *GhNHX1* gene encoding a novel putative tonoplast Na^+/ H^+ antiporters plays an important role in salt stress. Plant Cell Physiol., 45: 600-607.
6. Xiang, Y., Y. Huang and L. Xiong, 2007. Characterization of stress-responsive *CIPK* genes in rice for stress tolerance improvement. Plant Physiol., 144: 1416-1428.
7. Golldack, D., H. Su, F. Quigley, C.B. Michalowski, U.R. Kamasani and H.J. Bohnert ,2003. Salinity stress-tolerant and sensitive rice (*Oryza sativa* L.) regulate *AKT1*-type potassium channel transcripts differently. Plant Molecular Biol., 51: 71-81.
8. Wantabe, Y., M. Hirasaki, N. Tohnai, K. Yagi, S. Abe and Y. Tamai, 2003. Salt shock enhances the expression of *ZrATP2*, the gene for the mitochondrial ATPase β-subunit of *Zygosaccharomyces rouxii.* J. Bioscience and Bioenergy, 96: 193-195.
9. Brini, F., M. Hanin, I. Mezghani, G.A. Berkowitz and K. Masmoudi, 2007. Overexpression of wheat Na^+/H^+ antiporter *TNHX1* and H^+-pyrophosphatase *TVP1* improve salt-and drought -stress tolerance in *Arabidopsis thaliana* plants. J. Experimental Botany, 58: 301-308.
10. Rastogi, R.S., N. Mendirutta and P. Rastogi, 2003. Bioinformatics, Concept, Skill & applications, CBS Publishers and distributors New Delhi (India).
11. Lipman, D.J., S.F. Altschul and J.D. Kececioglu, 1989. A tool for multiple sequence alignment. Proceedings of National Academy of Sciences USA., 86: 4412- 4415.
12. Geourjon, C. and G. Deleage, 1995. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. CABIOS., 11: 681-684.
13. Morgenstern, B., 2004. DIALIGN : multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res., 32: W32-W36.
14. Notredame, C., 2002. Recent progressess in multiple sequence alignment : a survey. Pharmacogenomics 3: 1-14.
15. Sarin, K., C. Bagchi, A. Kumar and M.S. Bisht, 2004. Phylogenetic analysis of *Persenilin.* Bioinformatics India J., 2: 17-23.

16. Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin and D.G. Higgins, 1997. The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res., 25: 4876-4882.
17. Huelsenbeck, J.P. and D.M. Hillls, 1993. Success of phylogenetic methods in the four-taxon case. Systematic Biol., 42: 247-264.
18. Choi, J.H., H.Y. Jung, H.S. Kim and H.G. Cho, 2000. PhyloDraw: a phylogenetic tree drawing system. Bioinformatics, 16: 1056-1058.
19. Platten, J.D., O. Cotsaftis, P. Berthomieu, H. Bohnert, R.J. Davenport, D.J. Fairbairn, T. Horie, R.A. Leigh, H.X. Lin, S. Luan, P. Maser, O. Pantoja, A. Rodriguez-Navarro, D.P. Schachtman, J.I. Schroeder, H. Sentenac, N. Uozumi, A.A. Very, J.K. Zhu, E.S. Dennis and M. Tester, 2006. Nomenclature for *HKT* transporters, key determinants of plant salinity tolerance. Trends in Plant Sci., 11(8): 372-374.
20. Birve, S.J., E. Selstam and L.B.A. Johansson, 1996. Secondary structure of NADPH: protochlorophyllide oxidoreductase examined by circular dichrioism and prediction methods. Biochemistry J., 317: 549-555.
21. Rost, B., C. Sander and R. Schneider, 1994. Redefining the goals of protein secondary structure prediction. J. Molecular Biol., 235: 13-26.
22. Sivasubramanian, N. and D.P. Nayak, 1982. Sequence analysis of polymerase 1 gene and secondary structure prediction of polymerase 1 protein of Human influenza virus A/WSN/33. J. Virol., 44: 321-329.