In silico sequence analysis of Gibberellin-20 Oxidase-2, the semidwarfing protein of rice through the application of BLAST, FASTA and WU-Blast 2 programs

ASTHA DUHAN, MANOJ KUMAR SHARMA AND TAVINDERJEET KAUR

Accepted : September, 2009

SUMMARY

Computational prediction of evolutionary relationship for genes and proteins remains a fundamental problem in biology. Multiple sequence alignment plays a very important role in the evolutionary relationship prediction. The quality measurement is the most crucial issue in the multiple sequence alignment. An appropriate scoring function, high residue identity and degree of divergence lead to more accurate and reliable multiple sequence alignments. Using two scoring matrices, Blosum and PAM series of substitution and similarity matrices, sequence analysis tools for identifying evolutionary relationships to GA20ox2, the semi-dwarf protein of rice have been selected. The quality of alignment produced by various scoring matrices features the parameters that are important for the selection of right matrix for any alignment tool. We also show that the quality of any alignment and the homologous sequences produced by BLAST has more predictive power in identifying evolutionary conserved residues for the semi-dwarf protein. This was made evident by the conserved residue pattern observed for homologous sequences scanned by BLAST (Blosum80) resulted in 18 conserved residues by multiple sequence analysis in comparison with FASTA and WU-Blast2 that showed 11 and 12 conserved residues, respectively using the scoring matrix PAM 120.

Key words : Rice, GA_{20}, Semidwarf-1, sd-1, Blosum, Pam, FASTA, BLAST, Clustal, Semidwarfing gene

The semidwarfing gene in rice (*sd-1*) is one of the most important genes deployed in modern rice breeding. Its recessive character results in a shortened culm with improved lodging resistance and a greater harvest index, allowing for the increased use of nitrogen fertilizers. The *sd-1* gene was first identified in the Chinese variety Dee-geo-woo-gen (DGWG), and was crossed in the early 1960s with Peta (tall) to develop the semidwarf cultivar IR8, which produced record yields throughout Asia and formed the basis for the development of new high-yielding, semidwarf plant types (IRRI, 1967). Plant architecture refers to the collection of all important agronomic characters which determines the grain production. In rice it is mainly affected by plant height, tillering and panicle morphology (Wang *et al.*, 2005). In rice, short stature plants were developed by altering the plant architecture. Short stature in rice was developed by a recessive semidwarfgene sd-1, responsible for high yielding varieties (Nagano *et al.,* 2005). The sd-1 gene codes for the oxidase enzyme involved in the biosynthesis of gibberellins, Gibberellin-20 oxidase-2(GA20ox2). GA20ox2 is tightly linked to the sd-1 locus of rice chromosome 1. Semidwarfing gene inhibits the elongation of lower internodes which makes rice resistant oto lodging (Ogi *et al.,*1993). The short stature of IR8 is due to a mutation in the plant's sd1 gene, encoding an oxidase enzyme involved in the biosynthesis of gibberellin, a plant growth hormone (Sasaki *et al.,* 2002).Ga20ox2 catalyses the conversion of GA_{53} to GA_{20}. DGWG sd-1 mutants have 383 bp deletion in genomic sequence which encode a non-functional protein in an *indica* semidwarf IR8 (Monna *et al.,* 2002). However, a substitution of Leu-266 that is highly conserved residue resulted in the loss of function in *japonica* semidwarf. Rice represents enormous gene pool for improvement of rice cultivars as well as other crops which show significant similarity to rice genome. By making the use of sequence similarity/ homology genetic relationships among cereal crops can be established (Ishii *et al.,* 1996). Comparative genomic analysis revealed that the sdw1/denso gene that controls plant height, yield and quality has located in the syntenic region of the rice semidwarf gene sd1 on chromosome 1 (Jia *et al.*, 2009).

Protein sequences contain valuable information to predict structure and function of gene product. The comparison of two sets is a fundamental task that infers about how two sequences may be related functionally and genetically. The simultaneous alignment of amino acid sequences is now a major tool in molecular biology. The advent of large genome projects led to an explosion of

Correspondence to:
ASTHA DUHAN, Department of Bioinformatics, Mata Gujri College, FATEHGARH SAHIB (PUNJAB) INDIA
Authors' affiliations:
MANOJ KUMAR SHARMA, Department of Bioinformatics, Janta Vedi College, Baraut, BAGHPAT (U.P.) INDIA
TAVINDERJEET KAUR, Department of Bioinformatics, Mata Gujri College, FATEHGARH SAHIB (PUNJAB) INDIA

sequence data in public databases. Modern genome annotation and analysis tools rely heavily on accurate multiple alignments. The role of multiple sequence alignments in such systems has changed simply transferring annotation from one sequence to another to a genome wide perspective (Timo *et al.*, 2002).

Multiple alignments are used to find diagnostic pattern to characterize protein families to detect homology between new sequences and existing families of sequences to help and predict the function of new sequences and an essential preclude to molecular evolutionary analysis (Altschul, 1993). Therefore, all methods capable of handling larger problems in practical timescales make use of heuristics. Currently, the most widely used approach is to exploit the fact that homologous sequences are evolutionary related. One can build up a multiple alignment progressively by a series of pairwise alignments following the branching order in a phylogenetic tree (Feng *et al.*, 1987). Pairwise alignment of very closely related sequences can be obtained accurately by using a wide range of parameter values like gap penalities and weight matrix and the sequences of different degree of divergenece (Barton *et al.*,1987).

Numerous sequence alignment tools has been developed such as FASTA (Pearson *et al.*, 1988), WU-Blast2 (Altschul *et al.*, 1990) and BLAST (Altschul *et al.*, 1997) to identify the sequential pattern of similar and identical residues over a family of proteins to identify homologues or distant evolutionary relationships. Considering this fact we represent the selection of sequence analysis tools for GA20ox2 protein sequence taken from SWISS-PROT (http://expasy.org) sequence database (ID P0C5H5.1) to find homologous and ancestral sequences that have similar or related functions/ sequences.

MATERIALS AND METHODS

Application of sequence analysis tools to evaluate the scoring matrices that were employed in BLAST (Blastp), FASTA (fasta3), WU-Blast2 (WU-Blast2 protein) and clustal (Higgins *et al.*, 1994) are highlighted. FASTA is a two-step algorithm (Pearson *et al.*, 1988). The first step is a word search with a specific word size which finds regions in a two-dimensional table that are likely to correspond to highly similar segments of two sequences. The second step is a Smith-Waterman alignment (Smith *et al.*, 1981) centered on the diagonals that correspond to the alignment of highly similar sequence segments. FASTA is a heuristic approach to Smith-Waterman algorithm. The heuristics used by FASTA allow it to run much faster than the Smith-Waterman algorithm (Pearson, 1991). WU-Blast2 stands for Washington University Basic Local Alignment Search Tool Version 2.0 (http://www.ebi.ac.uk/blast2). The emphasis of this tool is to find regions of sequence similarity or homology at a faster speed, with minimum loss of sensitivity. This will yield functional and evolutionary clues about structure and function of query sequence. WU-Blast2 is a gapped version of BLASt allowing for gapped alignments and statistics. The BLAST (Altschul *et al.*, 1997) uses a word-based heuristic similar to FASTAthrough Maximal Segment Pairs algorithm. Blast do not allow gaps and have a valuable property that their statistics is more significant (Karlin *et al.*, 1990). P0C5H5.1, a Gibberellin-20 oxidase-2 sequence from *Oryza sativa indica* from SWISS-PROT database was selected. Programs FASTA and WU-Blast2 from European Bioinformatics Institute (http://www.ebi.ac.uk), BLAST from National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov) and Clustal from EBI were used as sequence analysis tools.

The scoring matrices for all the three sequence analysis tools were compared and the common matrices among them were identified and selected. Five matrices that are common among the three tools like Blosum 50, 62, 80 and PAM 120, 250 matrices were selected to study FASTA and WU-Blast2 whereas Blosum 45, 62, 80 and PAM 30, 70 were selected for NCBI BLAST. P0C5H5.1 (389 amino acid length) sequence was scanned against Uni-Prot Knowledge-Base database for similar and identical residues. The methodology is divided into three parts. First, the raw sequence scanned against FASTA with 5 scoring matrices and the per cent identity residues were tabulated. The same procedure was followed for WU-Blast2 and BLAST programs. Second, the parameters like %identity, scores (E-score, bit score, % positives) and amino acid overlaps were compared with respect to the highest and lowest hits in FASTA, WU-Blast2 and BLAST. Third, of all the 5 scoring matrices, the best parameters is evaluated from each tool and highest score from highest hits and lowest hits are identified and subjected to multiple sequence alignment through ClustalX$_2$. For all scoring matrices and tools default options are retained.

RESULTS AND DISCUSSION

A search against a protein database yielded several alignments using five scoring matrices and the scores along with amino acid overlaps accompanying these alignments are used to distinguish sequences related to degree of divergence. Five similarity matrices are used to identify the degree of evolutionary divergence. The output of each scoring matrix for FASTA, WU-Blast2 and BLAST were given in Table 1, 2 and 3, respectively.

Maximum per cent identity using five scoring matrices for FASTA was 100% with GAOX2_ORYSJ(Gibberellin-20 oxidase2 protein from (*Oryza sativa japonica*) for all matrices and a minimum of 22.8 identity observed with ASCL1_RAT (Achaete-scute homolog 1 from rat)for Blosum50. Comparatively 40.5% is having highest identity from lowest hits observed for IRAK1_BOVIN (Interleukin-1 receptor associated protein) when Blosum80 and PAM120 is employed as a scoring matrix. On an average, about 104 protein sequences have shown identity varying from a maximum of 100% to a minimum of 22.8% with five scoring matrices. A maximum of 389 residue overlap and a minimum of 42 residue overlap observed for five scoring matrices using FASTA and the % similarity and the E-valueranges for highest and lowest hits for FASTA are presented in Table 1.

Per cent identity for WU-Blast2 showed GAOX2_ORYSJ for major % identity with P0C5H5.1 with all scoring matrices. SEPS_METM7 (O-phosphoseryl-tRNA(Cys) synthetase) showed the lowest % identity for PAM250 matrix with 118% identity. Interestingly the number of similar proteins varied with different matrices. A highest of 249-250 similar aligned protein sequences was resulted when Blosum 50 80 and PAM 250 were observed in contrast with 103 and 99 similar sequences with Blosum62 and PAM 120, respectively. The %positives, scores and E-value ranges for highest and lowest hits for WU-Blast2 using five scoring matrices have been shown in Table 2.

NCBI blastp results obtained with different scoring matrices are given in Table 3. A maximum of 100% identity observed for all matrices where PAM 30 and 70 replaced against PAM 120 and 250 in above two cases. PAM 30 showed a 100% identity for GAOX2_ORYSJ with lowest 47% limit for VP13B_MOUSE (Vacuolar protein sorting-associated protein 13B). An average of about 95 hits observed when NCBI Blast is used for sd-1 sequences analysis. A maximum of 389 residue overlap and a minimum of 18 residue overlap observed for five scoring matrices. E-value and bit score ranges for highest and lowest hits are presented in Table 3.

The result from PAM 120 matrix with both FASTA and WU-Blast2 are comparable as the maximum per cent identity for both the tools resulted in 100% and 91%, respectively with GAOX2_ORYSJ as a maximum hit and the minimum hit resulted in the 40.5% identity in FASTA and 30% identity in WU-Blast2. Whereas using BLAST, 18 residues is the lowest residue overlap resulted from Blosum 80 matrix.Therefore based on the statistics given in Table 1, 2 and 3, PAM 120 matrix for both FASTA and WU-Blast2 and Blosum 80 matrix from BLAST results were selected to study multiple sequence alignments by ClustalX2.

Table 1 : Comparison of scoring matrices scanned against FASTA showing percent identity with corresponding parameters *i.e.* % identity, % similarity, overlap residues, E-value and homolog protein id in observed highest and lowest hits

Sr. No.	Matrix	No. of hits	FASTA (Highest hits)					FASTA (Lowest hits)				
			% identity	% similarity	overlap	E-value	Swiss-Prot id	% identity	% similarity	overlap	E-value	Swiss-Prot id
1.	Blosum50	112	100.0	100.0	389	3.5e-161	GAOX2_ORYSJ	22.8	54.0	189	7.6	ASCL1_RAT
2.	Blosum62	106	100.0	100.0	389	4.8e-201	GAOX2_ORYSJ	36.8	61.4	57	9.5	TATB_PSEF5
3.	Blosum80	109	100.0	100.0	389	0	GAOX2_ORYSJ	40.5	54.8	42	9.6	IRAK1_BOVIN
4.	PAM120	91	100.0	100.0	389	0	GAOX2_ORYSJ	40.5	66.7	42	8	IRAK1_BOVIN
5.	PAM250	106	100.0	100.0	389	1.4e-117	GAOX2_ORYSJ	24.1	63.0	108	10	COBS_CLOAB

Table 2 : Comparison of scoring matrices scanned against WU-Blast2 showing percent identity with corresponding parameters *i.e.* %identity, %positives, Score, E-value and homolog protein id in observed Highest and Lowest hits

Sr. No.	Matrix	No. of hits	WUBLAST-2 (Highest hits)					WUBLAST-2 (Lowest hits)				
			% identity	% positives	Score	E-value	Swiss-Prot id	% identity	% positives	Score	E-value	Swiss-Prot id
1.	Blosum50	249	88	88	2247	3.0e-221	GAOX2_ORYSJ	20	37	116	0.026	FAT2_SCHPO
2.	Blosum62	103	91	91	1743	2.0e-179	GAOX2_ORYSJ	29	50	56	8.9	RPON_METKA
3.	Blosum80	250	88	88	2843	5.4e-272	GAOX2_ORYSJ	27	42	138	0.0026	KIF1A_AEDAE
4.	PAM120	99	91	91	1745	5.9e-210	GAOX2_ORYSJ	30	63	70	6.2	COBS_CLOAB
5.	PAM250	250	88	88	1701	6.6e-161	GAOX2_ORYSJ	18	48	131	0.0012	SEPS_METM7

Table 3 : Comparison of scoring matrices scanned against BLAST (Blastp) showing percent identity with corresponding parameters *i.e.* %identity, Bit Score, overlap residues, E-value and homolog protein id in observed Highest and Lowest hits in each matrix

Sr. No.	Matrix	No. of hits	BLAST (Highest hits)					BLAST (Lowest hits)				
			% identity	Bit score	overlap	E-value	Swiss-Prot id	% identity	Bit Score	overlap	E-value	Swiss-Prot id
1.	Blosum4 5	102	100.0	829	389	0.0	GAOX2_ORYSJ	22	31.8	19	6.4	TRA2_ECOLX
2.	Blosum62	101	100.0	793	389	0.0	GAOX2_ORYSJ	52	33.5	18	2.2	DEOC_RHORT
3.	Blosum80	102	100.0	828	389	0.0	GAOX2_ORYSJ	54	34.0	18	2.5	DEOC_RHORT
4.	PAM30	80	100.0	819	389	0.0	GAOX2_ORYSJ	47	32.5	18	9.4	VP13B_MOUSE
5.	PAM70	94	100.0	799	389	0.0	GAOX2_ORYSJ	51	32.0	18	9.6	DEOC_RHORT

ClustalX2 default matrix is Gonnet250scoring matrix that uses an iterative classical distance measures to estimate an alignment of proteins (Gonnet *et al.*, 1992). Out of 91 hits obtained from FASTA with PAM 120, 99 hits obtained from WU-Blast2 with Pam 120 scoring matrix and 102 hits from BLAST with Blosum 80 scoring matrix, we screened the hits by selecting those sequences with % identity greater than or equal to 30. Then in the second round of screening only the crop species were kept. After the screening process only 44, 41 and 21 sequences remained in the hits for FASTA (PAM120), WU-Blast2 (PAM120) and BLAST (Blosum80) programs.

A multiple sequence alignment was done with ClustalX2 for the 21 hits resulted from BLAST program (Blosum80). All sequences were downloaded from SWISS PROT in fasta format and subjected to multiple sequence alignment along with the query sequence. The program was run with the default parameters like a value of 10 for gap open penalty and a value of 0.2 as gap extension penalty. Gaps are represented by "-" symbol. Consensus symbol were given underneath the alignments. Any alignment denoting the symbol "*" means that the residue in that column is identical in all sequences and the symbol ":" means that the conserved substitutions have been observed while "." means that the semi-conserved substitutions have been observed (http://www.ebi.ac.uk/clustalw). From the output result given in Fig. 1, there are a total of 18 residues with "*" which represented identical residues from all the aligned sequences and a total of 22 residues with ":" which represented conserved substitutions and 12 residues with "." which represented semi-conserved substitutions. This feature is taken as the criteria to evaluate the different methods applied to align GA20ox2 protein using five scoring matrices against FASTA, WU-Blast2 and BLAST tools.

Similarly multiple sequence alignment was done for 44 hits resulted from FASTA with PAM120 matrix. The parameters were set as stated earlier. From the output, there are only 11 residues that represented identical feature of aligned sequences, which show a poor evolutionary relationship. About 12 residues represented conserved substitutions and 9 residues represented semi-conserved substitutions. The same procedure applied for aligning the sequences of 41 hits from WU-Blast2 with Pam 120 matrix. The result showed 12 conserved residues for aligned sequences and 16 conserved substitution residues with 6 semi-conserved substitution residues. The relevant data was presented in Table 4.

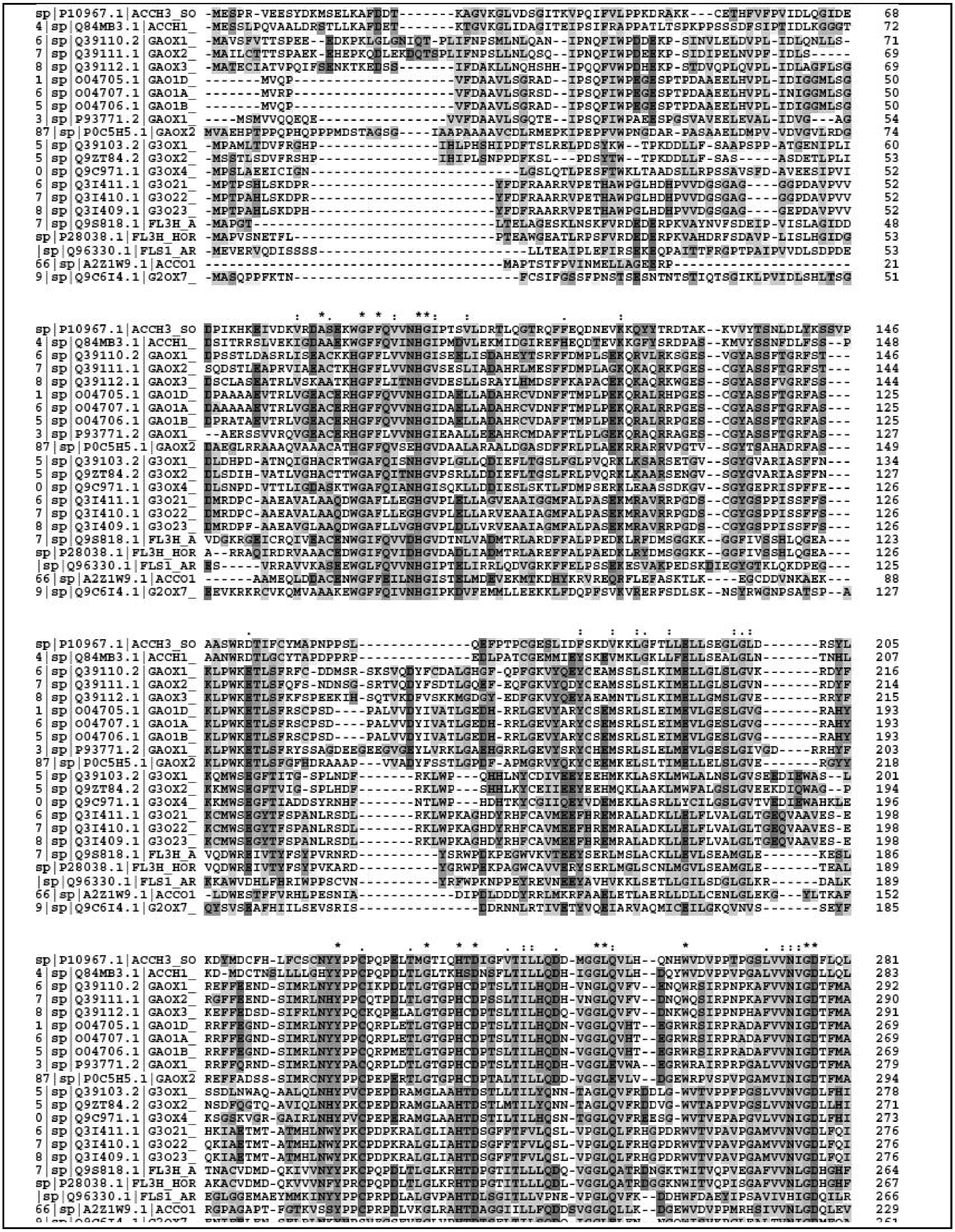
Therefore, from the above given data, it is evident that applying different scoring matrices for 3 sequence alignment tools resulted in evalustion of one scoring matrix that showed a better multiple sequence alignment with ClustalX2.

Conclusion:

In order to use any sequence alignment tool with different scoring matrices, one must have to quantify scoring matrices that may likely to conserve the physical and chemical properties necessary to maintain the structure and function of the protein. Multiple sequence

Table 4 : Comparison of alignment toolsagainst scoring matrix using CluatlX2 multiple sequence analysis using GONNET250matrix. (No. of hits represent the no. of proteins subjected to multiple sequence alignment; '*' represents aligned identical residues: ':' represents conserved substitutions residues and '.'represents semi-conserved substitutions)

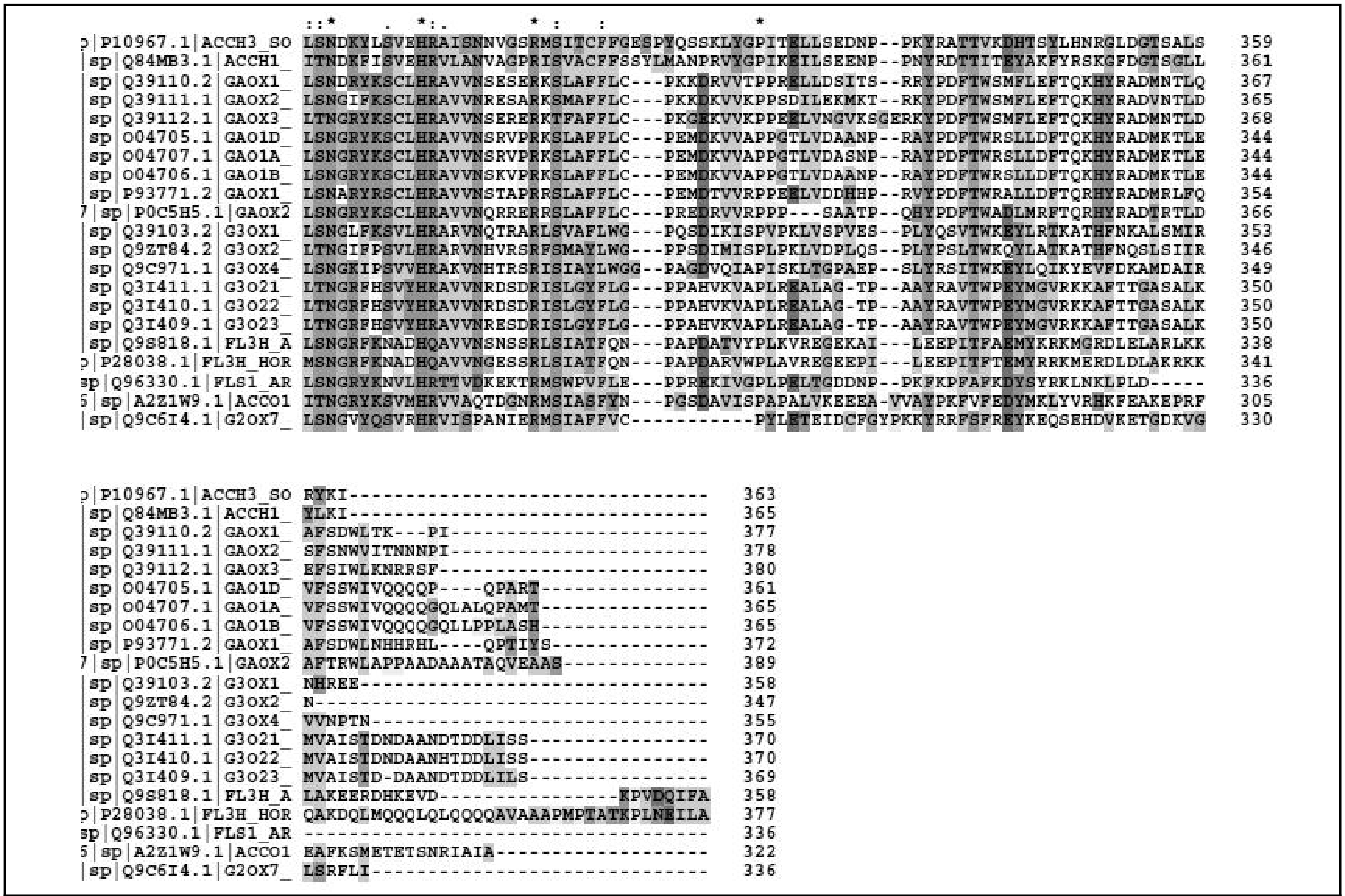
		ClustalX2				
Sr. No.	Alignment tool	Scoring matrix	No. of hits	'*'	':'	'.'
1.	FASTA	PAM 120	44	11	12	9
2.	WU-Blast2	PAM 120	41	12	16	6
3.	BLAST	Blosum 80	21	18	22	12



```
sp|P10967.1|ACCH3_SO -MESPR-VEESYDKMSELKAFDDT-------KAGVKGLVDSGITKVPQIFVLPPKDRAKK---CETHFVFPVIDLQGIDE   68
4|sp|Q84MB3.1|ACCH1_ -MESSLPQVAALDRSTLLKAFDET-------KTGVKGLIDAGITEIPSIFRAPPATLTSPKPPSSSDFSIPTIDLKGGGT   72
6|sp|Q39110.2|GAOX1_ -MAVSFVTTSPEE--EDKPKLGLGNIQT-PLIFNPSMLNLQAN--IPNQFIWPDDEKP-SINVLELDVPL-IDLQNLLS-   71
7|sp|Q39111.1|GAOX2_ -MAILCTTTSPAEK-EHEPKQDLEKDQTSPLIFNPSLLNLQSQ--IPNQFIWPDEEKP-SIDIPELNVPF-IDLS-----   69
8|sp|Q39112.1|GAOX3_ -MATECIATVPQIFSENKTKEDSS-------IFDAKLLNQHSHH-IPQQFVWPDHEKP-STDVQPLQVPL-IDLAGFLSG   69
1|sp|O04705.1|GAO1D_ -------MVQP--------------------VFDAAVLSGRAD--IPSQFIWPEGESPTPDAAEELHVPL-IDIGGMLSG   50
6|sp|O04707.1|GAO1A_ -------MVRP--------------------VFDAAVLSGRSD--IPSQFIWPEGESPTPDAAEELHVPL-INIGGMLSG   50
5|sp|O04706.1|GAO1B_ -------MVQP--------------------VFDAAVLSGRAD--IPSQFIWPEGESPTPDAAEELHVPL-IDIGGMLSG   50
3|sp|P93771.2|GAOX1_ ----MSMVVQQEQE----------------VVFDAAVLSGQTE--IPSQFIWPAEESPGSVAVEELEVAL-IDVG---AG   54
87|sp|P0C5H5.1|GAOX2 MVAEHPTPPQPHQPPPMDSTAGSG----IAAPAAAAVCDLRMEPKIPEPFVWPNGDAR-PASAAELDMPV-VDVGVLRDG   74
5|sp|Q39103.2|G3OX1_ -MPAMLTDVFRGHP---------------IHLPHSHIPDFTSLRELPDSYKW--TPKDDLLF-SAAPSPP-ATGENIPLI   60
5|sp|Q9ZT84.2|G3OX2_ -MSSTLSDVFRSHP---------------IHIPLSNPPDFKSL---PDSYTW--TPKDDLLF-SAS-----ASDETLPLI   53
0|sp|Q9C971.1|G3OX4_ -MPSLAEEICIGN--------------------------LGSLQTLPESFTWKLTAADSLLRPSSAVSFD-AVEESIPVI   52
6|sp|Q3I411.1|G3O21_ -MPTPSHLSKDPR-----------------------YFDFRAARRVPETHAWPGLHDHPVVDGSGAG----GGPDAVPVV   52
7|sp|Q3I410.1|G3O22_ -MPTPAHLSKDPR-----------------------YFDFRAARRVPETHAWPGLHDHPVVDGSGAG----GGPDAVPVV   52
8|sp|Q3I409.1|G3O23_ -MPTPAHLSKDPH-----------------------YFDFRAARRVPETHAWPGLHDHPVVDGSGAG----GEPDAVPVV   52
7|sp|Q9S818.1|FL3H_A -MAPGT------------------------------LTELAGESKLNSKFVRDEDERPKVAYNVFSDEIP-VISLAGIDD   48
sp|P28038.1|FL3H_HOR -MAPVSNETFL-------------------------PTEAWGEATLRPSFVRDEDERPKVAHDRFSDAVP-LISLHGIDG   53
|sp|Q96330.1|FLS1_AR -MEVERVQDISSSS--------------------------LLTEAIPLEFIRSEKEQPAITTFRGPTPAIPVVDLSDPDE   53
66|sp|A2Z1W9.1|ACCO1 -------------------------------------MAPTSTFPVINMELLAGEERP----------------------   21
9|sp|Q9C6I4.1|G2OX7_ -MASQPPFKTN----------------------------FCSIFGSSFPNSTSESNTNTSTIQTSGIKLPVIDLSHLTSG   51

                                :  *.   * * :  **:   :           .      :
sp|P10967.1|ACCH3_SO DPIKHKEIVDKVRDASEKWGFFQVVNHGIPTSVLDRTLQGTRQFFEQDNEVKKQYYTRDTAK--KVVYTSNLDLYKSSVP  146
4|sp|Q84MB3.1|ACCH1_ DSITRRSLVEKIGDAAEKWGFFQVINHGIPMDVLEKMIDGIREFHEQDTEVKKGFYSRDPAS--KMVYSSNFDLFSS--P  148
6|sp|Q39110.2|GAOX1  DPSSTLDASRLISEACKKHGFFLVVNHGISEELISDAHEYTSRFFDMPLSEKQRVLRKSGES--VGYASSFTGRFST---  146
7|sp|Q39111.1|GAOX2_ SQDSTLEAPRVIAEACTKHGFFLVVNHGVSESLIADAHRLMESFFDMPLAGKQKAQRKPGES--CGYASSFTGRFST---  144
8|sp|Q39112.1|GAOX3_ DSCLASEATRLVSKAATKHGFFLITNHGVDESLLSRAYLHMDSFFKAPACEKQKAQRKWGES--SGYASSFVGRFSS---  144
1|sp|O04705.1|GAO1D_ DPAAAAEVTRLVGEACERHGFFQVVNHGIDAELLADAHRCVDNFFTMPLPEKQRALRHPGES--CGYASSFTGRFAS---  125
6|sp|O04707.1|GAO1A_ DAAAAAEVTRLVGEACERHGFFQVVNHGIDAELLADAHRCVDNFFTMPLPEKQRALRRPGES--CGYASSFTGRFAS---  125
5|sp|O04706.1|GAO1B_ DPRATAEVTRLVGEACERHGFFQVVNHGIDAELLADAHRCVDAFFTMPLPEKQRALRRPGES--CGYASSFTGRFAS---  125
3|sp|P93771.2|GAOX1_ --AERSSVVRQVGEACERHGFFLVVNHGIEAALLEEAHRCMDAFFTLPLGEKQRAQRRAGES--CGYASSFTGRFAS---  127
87|sp|P0C5H5.1|GAOX2 DAEGLRRAAAQVAAACATHGFFQVSEHGVDAALARAALDGASDFFRLPLAEKRRARRVPGTV--SGYTSAHADRFAS---  149
5|sp|Q39103.2|G3OX1_ DLDHPD-ATNQIGHACRTWGAFQISNHGVPLGLLQDIEFLTGSLFGLPVQRKLKSARSETGV--SGYGVARIASFFN---  134
5|sp|Q9ZT84.2|G3OX2_ DLSDIH-VATLVGHACTTWGAFQITNHGVPSRLLDDIEFLTGSLFRLPVQRKLKAARSENGV--SGYGVARIASFFN---  127
0|sp|Q9C971.1|G3OX4_ DLSNPD-VTTLIGDASKTWGAFQIANHGISQKLLDDIESLSKTLFDMPSERKLEAASSDKGV--SGYGEPRISPFFE---  126
6|sp|Q3I411.1|G3O21_ DMRDPC-AAEAVALAAQDWGAFLLEGHGVPLELLAGVEAAIGGMFALPASEKMRAVRRPGDS--CGYGSPPISSFFS---  126
7|sp|Q3I410.1|G3O22_ DMRDPC-AAEAVALAAQDWGAFLLEGHGVPLELLARVEAAIAGMFALPASEKMRAVRRPGDS--CGYGSPPISSFFS---  126
8|sp|Q3I409.1|G3O23_ DMRDPF-AAEAVGLAAQDWGAFLLVGHGVPLDLLVRVEAAIAGMFALPASEKMRAVRRPGDS--CGYGSPPISSFFS---  126
7|sp|Q9S818.1|FL3H_A VDGKRGEICRQIVEACENWGIFQVVDHGVDTNLVADMTRLARDFFALPPEDKLRFDMSGGKK--GGFIVSSHLQGEA---  123
sp|P28038.1|FL3H_HOR A--RRAQIRDRVAAACEDWGIFQVIDHGVDADLIADMTRLAREFFALPAEDKLRYDMSGGKK--GGFIVSSHLQGEA---  126
|sp|Q96330.1|FLS1_AR ES-----VRRAVVKASEEWGLFQVVNHGIPTELIRRLQDVGRKFFELPSSEKESVAKPEDSKDIEGYGTKLQKDPEG---  125
66|sp|A2Z1W9.1|ACCO1 ------AAMEQLDDACENWGFFEILNHGISTELMDEVEKMTKDHYKRVREQRFLEFASKTLK----EGCDDVNKAEK---   88
9|sp|Q9C6I4.1|G2OX7_ EEVKRKRCVKQMVAAAKEWGFFQIVNHGIPKDVFEMMLLEEKKLFDQPFSVKVRERFSDLSK--NSYRWGNPSATSP--A  127

                          .                                        :   :  :.  :       :.:
sp|P10967.1|ACCH3_SO AASWRDTIFCYMAPNPPSL--------------QEFPTPCGESLIDFSKDVKKLGFTLLELLSEGLGLD-------RSYL  205
4|sp|Q84MB3.1|ACCH1_ AANWRDTLGCYTAPDPPRP--------------EDLPATCGEMMIEYSKEVMKLGKLLFELLSEALGLN-------TNHL  207
6|sp|Q39110.2|GAOX1_ KLPWKETLSFRFC-DDMSR-SKSVQDYFCDALGHGF-QPFGKVYQEYCEAMSSLSLKIMELLGLSLGVK-------RDYF  216
7|sp|Q39111.1|GAOX2_ KLPWKETLSFQFS-NDNSG-SRTVQDYFSDTLGQEF-EQFGKVYQDYCEAMSSLSLKIMELLGLSLGVN-------RDYF  214
8|sp|Q39112.1|GAOX3_ KLPWKETLSFKFSPEEKIH-SQTVKDFVSKKMGDGY-EDFGKVYQEYAEAMNTLSLKIMELLGMSLGVE-------RRYF  215
1|sp|O04705.1|GAO1D_ KLPWKETLSFRSCPSD----PALVVDYIVATLGEDH-RRLGEVYARYCSEMSRLSLEIMEVLGESLGVG-------RAHY  193
6|sp|O04707.1|GAO1A_ KLPWKETLSFRSCPSD----PALVVDYIVATLGEDH-RRLGEVYARYCSEMSRLSLEIMEVLGESLGVG-------RAHY  193
5|sp|O04706.1|GAO1B_ KLPWKETLSFRSCPSD----PALVVDYIVATLGEDH-RRLGEVYARYCSEMSRLSLEIMEVLGESLGVG-------RAHY  193
3|sp|P93771.2|GAOX1_ KLPWKETLSFRYSSAGDEEGEEGVGEYLVRKLGAEHGRRLGEVYSRYCHEMSRLSLELMEVLGESLGIVGD----RRHYF  203
87|sp|P0C5H5.1|GAOX2 KLPWKETLSFGFHDRAAAP---VVADYFSSTLGPDF-APMGRVYQKYCEEMKELSLTIMELLELSLGVE-------RGYY  218
5|sp|Q39103.2|G3OX1_ KQMWSEGFTITG-SPLNDF-------RKLWP---QHHLNYCDIVEEYEEHMKKLASKLMWLALNSLGVSEEDIEWAS--L  201
5|sp|Q9ZT84.2|G3OX2_ KKMWSEGFTVIG-SPLHDF-------RKLWP---SHHLKYCEIIEEYEEHMQKLAAKLMWFALGSLGVEEKDIQWAG--P  194
0|sp|Q9C971.1|G3OX4_ KKMWSEGFTIADDSYRNHF-------NTLWP---HDHTKYCGIIQEYVDEMEKLASRLLYCILGSLGVTVEDIEWAHKLE  196
6|sp|Q3I411.1|G3O21_ KCMWSEGYTFSPANLRSDL-------RKLWPKAGHDYRHFCAVMEEFHREMRALADKLLELFLVALGLTGEQVAAVES-E  198
7|sp|Q3I410.1|G3O22_ KCMWSEGYTFSPANLRSDL-------RKLWPKAGHDYRHFCAVMEEFHREMRALADKLLELFLVALGLTGEQVAAVES-E  198
8|sp|Q3I409.1|G3O23_ KCMWSEGYTFSPANLRSDL-------RKLWPKAGHDYRHFCAVMEEFHREMRALADKLLELFLVALGLTGEQVAAVES-E  198
7|sp|Q9S818.1|FL3H_A VQDWREIVTYFSYPVRNRD----------YSRWPDKPEGWVKVTEEYSERLMSLACKLLEVLSEAMGLE-------KESL  186
sp|P28038.1|FL3H_HOR VQDWREIVTYFSYPVKARD----------YGRWPEKPAGWCAVVERYSERLMGLSCNLMGVLSEAMGLE-------TEAL  189
|sp|Q96330.1|FLS1_AR KKAWVDHLFHRIWPPSCVN----------YRFWPKNPPEYREVNEEYAVHVKKLSETLLGILSDGLGLKR------DALK  189
66|sp|A2Z1W9.1|ACCO1 -LDWESTFFVRHLPESNIA------------DIPDLDDDYRRLMKRFAAELETLAERLLDLLCENLGLEKG---YLTKAF  152
9|sp|Q9C6I4.1|G2OX7_ QYSVSEAFHIILSEVSRIS---------------DDRNNLRTIVETYVQEIARVAQMICEILGKQVNVS-------SEYF  185

                                     *  .     . *   * *   . ::  .    **:        *          . :::**
sp|P10967.1|ACCH3_SO KDYMDCFH-LFCSCNYYPPCPQPELTMGTIQHTDIGFVTILLQDD-MGGLQVLH--QNHWVDVPPTPGSLVVNIGDFLQL  281
4|sp|Q84MB3.1|ACCH1_ KD-MDCTNSLLLLGHYYPPCPQPDLTLGLTKHSDNSFLTILLQDH-IGGLQVLH--DQYWVDVPPVPGALVVNVGDLLQL  283
6|sp|Q39110.2|GAOX1_ REFFEEND-SIMRLNYYPPCIKPDLTLGTGPHCDPTSLTILHQDH-VNGLQVFV--ENQWRSIRPNPKAFVVNIGDTFMA  292
7|sp|Q39111.1|GAOX2_ RGFFEEND-SIMRLNHYPPCQTPDLTLGTGPHCDPSSLTILHQDH-VNGLQVFV--DNQWQSIRPNPKAFVVNIGDTFMA  290
8|sp|Q39112.1|GAOX3_ KEFFEDSD-SIFRLNYYPQCKQPELALGTGPHCDPTSLTILHQDQ-VGGLQVFV--DNKWQSIPPNPHAFVVNIGDTFMA  291
1|sp|O04705.1|GAO1D_ RRFFEGND-SIMRLNYYPPCQRPLETLGTGPHCDPTSLTILHQDN-VGGLQVHT--EGRWRSIRPRADAFVVNIGDTFMA  269
6|sp|O04707.1|GAO1A_ RRFFEGND-SIMRLNYYPPCQRPLETLGTGPHCDPTSLTILHQDN-VGGLQVHT--EGRWRSIRPRADAFVVNIGDTFMA  269
5|sp|O04706.1|GAO1B_ RRFFEGND-SIMRLNYYPPCQRPMETLGTGPHCDPTSLTILHQDN-VGGLQVHT--EGRWRSIRPRADAFVVNIGDTFMA  269
3|sp|P93771.2|GAOX1_ RRFFQRND-SIMRLNYYPACQRPLDTLGTGPHCDPTSLTILHQDH-VGGLEVWA--EGRWRAIRPRPGALVVNVGDTFMA  279
87|sp|P0C5H5.1|GAOX2 REFFADSS-SIMRCNYYPPCPEPERTLGTGPHCDPTALTILLQDD-VGGLEVLV--DGEWRPVSPVPGAMVINIGDTFMA  294
5|sp|Q39103.2|G3OX1_ SSDLNWAQ-AALQLNHYPVCPEPDRAMGLAAHTDSTLLTILYQNN-TAGLQVFRDDLG-WVTVPPFPGSLVVNVGDLFHI  278
5|sp|Q9ZT84.2|G3OX2_ NSDFQGTQ-AVIQLNHYPKCPEPDRAMGLAAHTDSTLMTILYQNN-TAGLQVFRDDVG-WVTAPPVPGSLVVNVGDLLHI  271
0|sp|Q9C971.1|G3OX4_ KSGSKVGR-GAIRLNHYPVCPEPERAMGLAAHTDSTILTILHQSN-TGGLQVFREESG-WVTVEPAPGVLVVNIGDLFHI  273
6|sp|Q3I411.1|G3O21_ HKIAETMT-ATMHLNWYPKCPDPKRALGLIAHTDSGFFTFVLQSL-VPGLQLFRHGPDRWVTVPAVPGAMVVNVGDLFQI  276
7|sp|Q3I410.1|G3O22_ QKIAETMT-ATMHLNWYPKCPDPKRALGLIAHTDSGFFTFVLQSL-VPGLQLFRHGPDRWVTVPAVPGAMVVNVGDLFQI  276
8|sp|Q3I409.1|G3O23_ QKIAETMT-ATMHLNWYPKCPDPKRALGLIAHTDSGFFTFVLQSL-VPGLQLFRHGPDRWVTVPAVPGAMVVNVGDLFQI  276
7|sp|Q9S818.1|FL3H_A TNACVDMD-QKIVVNYYPKCPQPDLTLGLKRHTDPGTITLLLQDQ-VGGLQATRDNGKTWITVQPVEGAFVVNLGDHGHF  264
sp|P28038.1|FL3H_HOR AKACVDMD-QKVVVNFYPRCPQPDLTLGLKRHTDPGTITLLLQDL-VGGLQATRDGGKNWITVQPISGAFVVNLGDHGHF  267
|sp|Q96330.1|FLS1_AR EGLGGEMAEYMMKINYYPPCPRPDLALGVPAHTDLSGITLLVPNE-VPGLQVFK--DDHWFDAEYIPSAVIVHIGDQILR  266
66|sp|A2Z1W9.1|ACCO1 RGPAGAPT-FGTKVSSYPPCPRPDLVKGLRAHTDAGGIILLFQDDSVGGLQLLK--DGEWVDVPPMRHSIVVNLGDQLEV  229
```

Fig. 1 contd...

Contd... Fig. 1



```
                      ::*    .  *:.       * :   :                *
p|P10967.1|ACCH3_SO  LSNDKYLSVEHRAISNNVGSRMSITCFFGESPYQSSKLYGPITELLSEDNP--PKYRATTVKDHTSYLHNRGLDGTSALS   359
 |sp|Q84MB3.1|ACCH1_ ITNDKFISVEHRVLANVAGPRISVACFFSSYLMANPRVYGPIKEILSEENP--PNYRDTTITEYAKFYRSKGFDGTSGLL   361
 |sp|Q39110.2|GAOX1_ LSNDRYKSCLHRAVVNSESERKSLAFFLC---PKKDRVVTPPRELLDSITS--RRYPDFTWSMFLEFTQKHYRADMNTLQ   367
 |sp|Q39111.1|GAOX2_ LSNGIFKSCLHRAVVNRESARKSMAFFLC---PKKDKVVKPPSDILEKMKT--RKYPDFTWSMFLEFTQKHYRADVNTLD   365
 |sp|Q39112.1|GAOX3_ LTNGRYKSCLHRAVVNSERERKTFAFFLC---PKGEKVVKPPEELVNGVKSGERKYPDFTWSMFLEFTQKHYRADMNTLD   368
 |sp|O04705.1|GAO1D_ LSNGRYKSCLHRAVVNSRVPRKSLAFFLC---PEMDKVVAPPGTLVDAANP--RAYPDFTWRSLLDFTQKHYRADMKTLE   344
 |sp|O04707.1|GAO1A_ LSNGRYKSCLHRAVVNSRVPRKSLAFFLC---PEMDKVVAPPGTLVDASNP--RAYPDFTWRSLLDFTQKHYRADMKTLE   344
 |sp|O04706.1|GAO1B_ LSNGRYKSCLHRAVVNSKVPRKSLAFFLC---PEMDKVVAPPGTLVDAANP--RAYPDFTWRSLLDFTQKHYRADMKTLE   344
 |sp|P93771.2|GAOX1_ LSNARYRSCLHRAVVNSTAPRRSLAFFLC---PEMDTVVRPPEELVDDHHP--RVYPDFTWRALLDFTQRHYRADMRLFQ   354
7|sp|P0C5H5.1|GAOX2  LSNGRYKSCLHRAVVNQRRERRSLAFFLC---PREDRVVRPPP---SAATP--QHYPDFTWADLMRFTQRHYRADTRTLD   366
 |sp|Q39103.2|G3OX1_ LSNGLFKSVLHRARVNQTRARLSVAFLWG---PQSDIKISPVPKLVSPVES--PLYQSVTWKEYLRTKATHFNKALSMIR   353
 |sp|Q9ZT84.2|G3OX2_ LTNGIFPSVLHRARVNHVRSRFSMAYLWG---PPSDIMISPLPKLVDPLQS--PLYPSLTWKQYLATKATHFNQSLSIIR   346
 |sp|Q9C971.1|G3OX4_ LSNGKIPSVVHRAKVNHTRSRISIAYLWGG--PAGDVQIAPISKLTGPAEP--SLYRSITWKEYLQIKYEVFDKAMDAIR   349
 |sp|Q3I411.1|G3O21_ LTNGRFHSVYHRAVVNRDSDRISLGYFLG---PPAHVKVAPLREALAG-TP--AAYRAVTWPEYMGVRKKAFTTGASALK   350
 |sp|Q3I410.1|G3O22_ LTNGRFHSVYHRAVVNRDSDRISLGYFLG---PPAHVKVAPLREALAG-TP--AAYRAVTWPEYMGVRKKAFTTGASALK   350
 |sp|Q3I409.1|G3O23_ LTNGRFHSVYHRAVVNRESDRISLGYFLG---PPAHVKVAPLREALAG-TP--AAYRAVTWPEYMGVRKKAFTTGASALK   350
 |sp|Q9S818.1|FL3H_A LSNGRFKNADHQAVVNSNSSRLSIATFQN---PAPDATVYPLKVREGEKAI---LEEPITFAEMYKRKMGRDLELARLKK   338
p|P28038.1|FL3H_HOR MSNGRFKNADHQAVVNGESSRLSIATFQN---PAPDARVWPLAVREGEEPI---LEEPITFTEMYRRKMERDLDLAKRKK   341
sp|Q96330.1|FLS1_AR  LSNGRYKNVLHRTTVDKEKTRMSWPVFLE---PPREKIVGPLPELTGDDNP--PKFKPFAFKDYSYRKLNKLPLD-----   336
5|sp|A2Z1W9.1|ACCO1  ITNGRYKSVMHRVVAQTDGNRMSIASFYN---PGSDAVISPAPALVKEEEA-VVAYPKFVFEDYMKLYVRHKFEAKEPRF   305
 |sp|Q9C6I4.1|G2OX7_ LSNGVYQSVRHRVISPANIERMSIAFFVC-----------PYLETEIDCFGYPKKYRRFSFREYKEQSEHDVKETGDKVG   330


p|P10967.1|ACCH3_SO  RYKI--------------------------------   363
 |sp|Q84MB3.1|ACCH1_ YLKI--------------------------------   365
 |sp|Q39110.2|GAOX1_ AFSDWLTK---PI-----------------------   377
 |sp|Q39111.1|GAOX2_ SFSNWVITNNNPI-----------------------   378
 |sp|Q39112.1|GAOX3_ EFSIWLKNRRSF------------------------   380
 |sp|O04705.1|GAO1D_ VFSSWIVQQQQP----QPART---------------   361
 |sp|O04707.1|GAO1A_ VFSSWIVQQQQGQLALQPAMT---------------   365
 |sp|O04706.1|GAO1B_ VFSSWIVQQQQGQLLPPLASH---------------   365
 |sp|P93771.2|GAOX1_ AFSDWLNHHRHL----QPTIYS--------------   372
7|sp|P0C5H5.1|GAOX2  AFTRWLAPPAADAAATAQVEAAS-------------   389
 |sp|Q39103.2|G3OX1_ NHREE-------------------------------   358
 |sp|Q9ZT84.2|G3OX2_ N-----------------------------------   347
 |sp|Q9C971.1|G3OX4_ VVNPTN------------------------------   355
 |sp|Q3I411.1|G3O21_ MVAISTDNDAANDTDDLISS----------------   370
 |sp|Q3I410.1|G3O22_ MVAISTDNDAANHTDDLISS----------------   370
 |sp|Q3I409.1|G3O23_ MVAISTD-DAANDTDDLILS----------------   369
 |sp|Q9S818.1|FL3H_A LAKEERDHKEVD----------------KPVDQIFA   358
p|P28038.1|FL3H_HOR QAKDQLMQQQLQLQQQQAVAAAPMPTATKPLNEILA   377
sp|Q96330.1|FLS1_AR  ------------------------------------   336
5|sp|A2Z1W9.1|ACCO1  EAFKSMETETSNRIAIA-------------------   322
 |sp|Q9C6I4.1|G2OX7_ LSRFLI------------------------------   336
```

Fig. 1 : ClustalX2 output format for BLAST hits representing conserved residues(*), conserved substitutions(:) and semi-conserved substitutions(.)

alignment employing combined PAM 120 scoring matrices of FASTA and WU-Blast2 yielded gapped, linear and organized alignments with homologous sequences. The study represented selection of parameters and overlap residues for different scoring matrices that resulted in better explanation and understanding of evolutionary relationships with semidwarf protein of rice. The quality of alignments produced by BLAST reveals clear evolutionary relationship when compared with FASTA and WU-Blast2. The conserved residue pattern observed for homologous sequences scanned by BLAST (Blosum80) resulted in 21 conserved residues by ClustalX multiple sequence analysis in comparison with WU-Blast2 (PAM120) and FASTA (PAM120) that showed 12 and 11 conserved residues, respectively. The results suggests that selection of correct matrix for proper sequence analysis plays a major role in identifying an organized alignment pattern for predicting functional and evolutionary biology.

REFERENCES

Altschul, S.F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.*, **36**: 290-300.

Altschul, S.F., Gish, W., Miller, W., Eugene, W.M. and David, J.L. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**: 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**: 3389-3402.

Barton, G.J. and Stemberg, M.J.E. (1987). A strategy for the rapid multiple alignment of protein sequences. *J. Mol. Biol.*, **198**: 327-337.

Feng, D.F. and Doolittle, R.F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, **25**: 351-360.

Gonnet, G.H., Cohen, M.A. and Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. *Sci.*, **256**: 1443-1445.

Higgins, D., Thompson, J., Gibson, T., Thompson, J.D., Higgins, D.G. and Gibson T.J. (1994). CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**: 4673-4680.

International Rice Research Institute (1967) *Annual Report for 1966*, 59–82.

Ishii, T., Nakano, T. and Maeda, H. (1996). Phylogenetic relationships in A-genome species of rice as revealed by RAPD analysis. *Genes & genetic Systems*, **71**: 195-201.

Jia, Q., Zhang, J., Westcott, S., Zhang, X.Q., Bellgard, M., Lance, R. and Li, C.(2009). GA-20 oxidase as a candidate for the semidwarf gene sdw1/denso in barley. *Functional Integartive Genomics*, **9**(2): 255-262.

Karlin, S and Altschul, S.F.(1990). Methods for accessing the statistical significance of molecular features by using general scoring schemes. *Proceedings of National Academy of Sciences*, **87**: 2264-2268.

Monna, L., Kitazawa, N., Yoshino, R., Suzuki, J., Masuda,H., Maehara, Y., Tanji, M., Sato, M., Nasu, S. and Minobe, Y. (2002). Positional cloning of rice semidwarfing gene, sd-1 rice "green revolution gene" encodes a mutant enzyme involved in gibberellins synthesis. *DNA Res.*, **9**: 11-17.

Nagano, H., Onishi, K., Ogasawara, M., Horiuchi, Y. and Sano, Y. (2005). Genealogy of the "Green Revolution" gene in rice. *Genes Genetics Systems*, **80**: 351-356.

Ogi, Y., Kato, H., Marayuma, K. and Kilkuchi, F.(1993). The effects of culm length and other agronomic characters caused by semidwarfing gene at the *sd-1* locus in rice. *Japanese J. Breed.*, **43**: 267-275.

Pearson, W.R. and Lipman, D.J.(1988). Improved Tools for Biological Sequence Comparison. *Proceedings of National Academy of Sci.*, **85**: 2444-2448.

Pearson, W.R.(1991). Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**: 635-650.

Sasaki, A., Ashikari, M., Tanaka, M., Itoh, H., Nishimura, A., Swapan, D., Ishiyama, K., Saito, T., Kobayashi, M., Khush, G.S., Kitano, H. and Matsuoka, M. (2002). Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature*, **416**(6882):701-702.

Smith, T.F. and Waterman, M.Sc.(1981). Identification of common molecular subsequences. *J. Mol. Biol.*, **147**: 195-197.

Timo Lassmann and Erik L. L. Sonnhammer (2002). Quality assessment of multiple alignment programs. *FEBS Letters*, **529**: 126-130.

Wang, Y. and Li, J.(2005). The Plant Architecture of Rice (*Oryza sativa*). *Plant Mol. Biol.*, **59**(1): 75-84.

****
*****