International Journal of Agriculture Innovations and Research
Volume 4, Issue 5, ISSN (Online) 2319-1473
Manuscript Processing Details (dd/mm/yyyy) :
Received : 22/04/2016 | Accepted on : 06/05/2016 | Published : 12/05/2016

Multiple Sequence Alignment of Model Plants Using Bioinformatics Approach

Nivedita Yadav[1], Apoorv Tiwari[1,2], Vijay Kumar Garg[1]*
[1]Department of Computational Biology and Bioinformatics, Jacob JSBB, Sam Higginbottom Institute of Agriculture, Technology and Sciences, Allahabad-211007, Uttar Pradesh, Bharat (India)
[2]Department of Molecular Biology & Genetic Engineering, CBSH, GBPUA &T, Pantnagar-263145, Uttarakhand, Bharat (India)
cnivi.2010@gmail.com[1], apoorvbio@gmail.com[1,2], bioinformatics.vijay@gmail.com[1]*, *corresponding author

***Abstract* – Bioinformatics is an interdisciplinary area of research, which also plays a vital role in the field of agriculture based studies. Tools of bioinformatics provides significant role in agriculture research. Present paper is also focusing on agriculture informatics. As we know using bioinformatics tool we can explore many more hidden information from agriculture data. In this paper we had applied CLUSTAL O tool for multiple sequence alignment of nine different model plants have same protein glycogen synthase. We had constructed phylogenetic tree for investigating relationship between model plants using neighbor -joining tree without distance corrections method by CLUSTAL O tool.**

***Keywords* – Multiple Sequence Alignment, Phylogenetic Tree, CLUSTAL O, Bioinformatics.**

I. INTRODUCTION

Bioinformatics is a fresh field of science but it is making evolution in each field of biotechnology incredibly. As it has its diligence in the drug by providing the genome information of different organisms, likewise the field of agriculture has also taken benefit of this field because microorganisms play significant function in agriculture and bioinformatics provides complete genomic information of these organisms **[1-9]**. The genome sequencing of the plants and animals has too provided benefits to agriculture **[10]**. We care about the sequence alignments in the computational biology because it gives biologists functional information about diverse aspects **[11]**. For example, it can tell us about the evolution of the organisms, we can see which realms of a gene (or its derived protein) are vulnerable to mutation and which can have one rest replaced by another without altering function, we can analyse homologous genes and can reveal paralogs and orthologs genes that are evolutionary connected. In problems such as the building of an evolutionary tree relates on sequence data, or in protein engineering, where a multiple alignment of related sequences may often give way the good number helpful information on the design of a new protein, a molecular biologist must evaluate more than two sequences concurrently **[12-19]**. A multiple sequence alignment (MSA) arranges protein sequences into a rectangular array with the objective that residues in a given column are homologous (derived from a single position in an ancestral sequence), identical (in a rigid local structural alignment) or participate a common functional role **[20]**. Although these three criterions are fundamentally corresponding for closely related proteins, sequence, structure and function deviate over evolutionary time and dissimilar criteria may result in unlike alignments **[21]**. Manually sophisticated alignments continue to be superior to solely automated methods; there is therefore an unremitting effort to improve the biological correctness of MSA tools. Moreover, the high computational cost of most naive algorithms motivates improvements in speed and memory usage to contain the rapid increase in accessible sequence data **[22-25]**.

In this paper we will perform multiple sequence alignment for glycogen synthase protein of nine different plants **[26]**.

Glycogen synthase is an enzyme concerned in converting glucose to glycogen. It takes petite polymers of glucose and converts them into long polymers of glycogen **[27]**. In other words, this enzyme converts surplus glucose residues one by one into a polymeric chain for storage as glycogen. Glycogen synthase concentration is highest in the bloodstream 30 to 60 minute following intense exercise. It is a key enzyme in glycogenesis **[28-31]**.

There are various tools available for multiple sequence alignment. Some frequently used tools are listed below in Table 1.

Table 1. Multiple sequence alignment tools

Tool	URL
Jalview	www.jalview.org
SeaView	www.pbil.univ-lyon1.fr/software/seaview.html
CINEMA	www.bioinf.manchester.ac.uk/dbbrowser/CINEMA2.1/
Kalignvu	www.msa.cgb.ki.se/
GeneDoc	www.nrbsc.org/gfx/genedoc/
STRAP	www.charite.de/bioinf/starp/
ClustalX	www.clustal.org
BoxShade	www.ch.embnet.org/software/BOX_form.html
ALTAVIST	www.bibiserv.techfak.uni-bielefeld.de/altavist/

II. METHODS AND MATERIALS

For multiple alignment and tree construction NCBI and CLUSTAL O tool were used. First of all we had selected protein named Glycogen synthase for the study.From NCBI we had searched nine model plants carrying this protein. Glycogen synthase [Bathycoccus prasinos],

International Journal of Agriculture Innovations and Research
Volume 4, Issue 5, ISSN (Online) 2319-1473

Glycogen synthase [Morus notabilis], Glycogen synthase [Gossypium arboreum], Glycogen synthase [Auxenochlorella protothecoides], glycogen synthase family protein [Populus trichocarpa], Glycogen synthase [Auxenochlorella protothecoides], glycogen synthase [Arabidopsis thaliana], glycogen synthase kinase-3 [Glycine max] and glycogen (starch) synthase [Solanum tuberosum] were taken for multiple sequence alignment. MSA was carried out by CLUSTAL OMEGA program from EMBL-EBI (http://www.ebi.ac.uk/Tools/msa/clustalw2/). This program is freely available and also highly recommended for protein multiple sequence alignment **[32]**. The output of MSA was our desired result. Further this result can we used as input for phylogenetic analysis and we can use it as input for other bioinformatics analysis tool like PHYLIP **[33]**.

III. Results

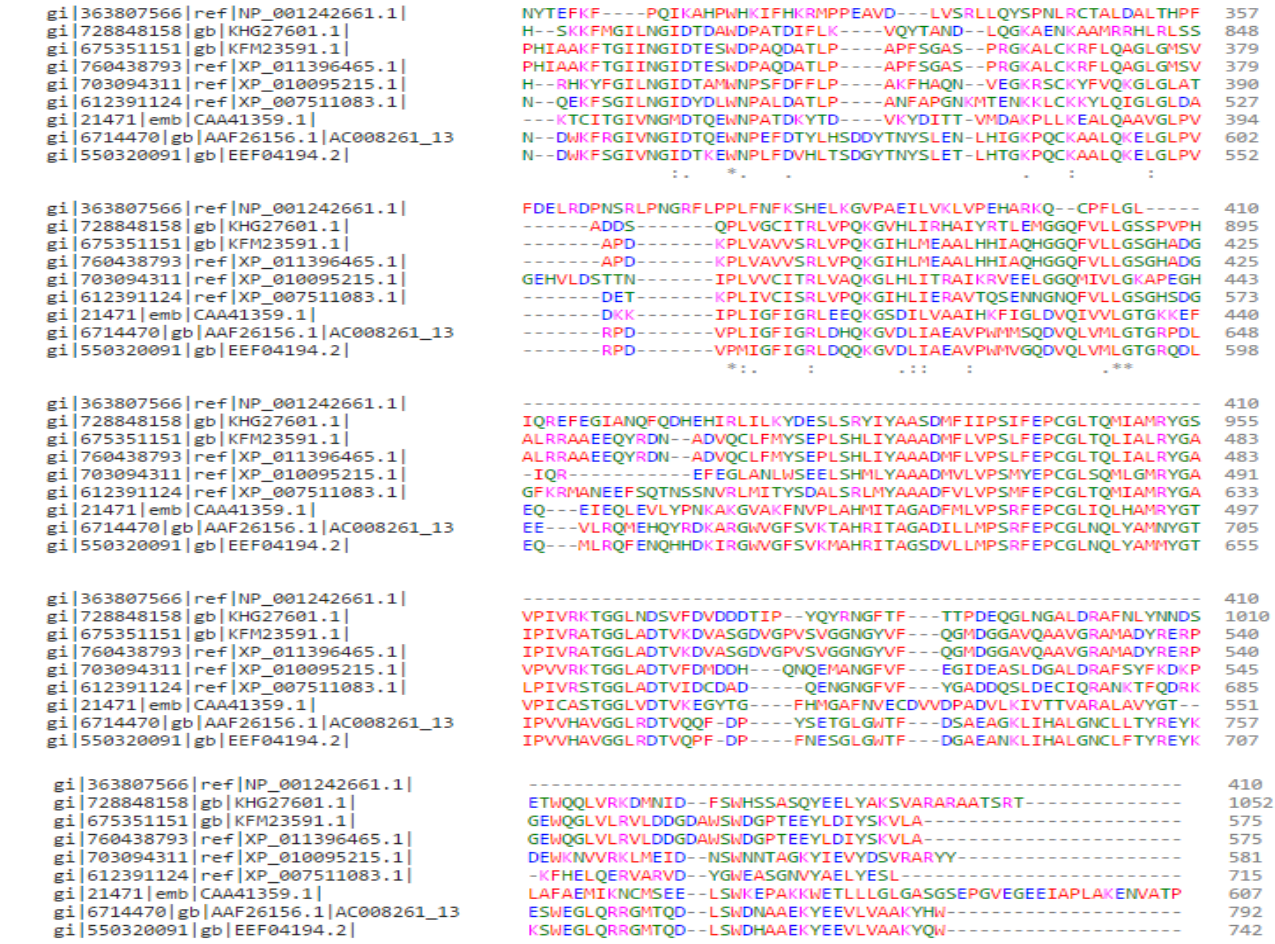
A. *Multiple Sequence Alignment*

A multiple sequence alignment (MSA) is a sequence conjunction of three or extra biological sequences, usually protein, DNA, or RNA. In loads of cases, the input set of query sequences are unspecified to have an evolutionary affiliation by which they contribute to a lineage and are descended from a universal ancestor **[34]**. From the consequential MSA, sequence homology can be incidental and phylogenetic study can be conducted to review the sequences alignment is frequently used to assess sequence preservation of protein domains, tertiary and secondary structures, and amino acid or nucleotides **[35]**.

Multiple sequence alignment also refer to the procedure of aligning such a sequences of biologically applicable length can be tricky and are almost always prolonged to align by hand, computational algorithms are used to fabricate and analyze the alignment. MSAs necessitate more sophisticated methodologies that pair wise alignment because they are more computationally complex.

The majority of multiple sequence alignment programs use heuristic methods rather than global optimization because distinguishing the most favorable alignment between more than a few sequences of reasonable length is prohibitively computationally expensive **[36]**.

International Journal of Agriculture Innovations and Research
Volume 4, Issue 5, ISSN (Online) 2319-1473



Sequence alignment produced by CLUSTAL O program, of above protein sequences is a key denoting conserved sequence **(*),** conservative mutations **(:),** semi-conservative mutation **(.),** and non-conservative mutations **().**

In biology, conserved sequences are analogous or indistinguishable sequences that place within nucleic acid sequences, protein sequences, protein structures or polymeric carbohydrates across species (orthologous sequences) or within dissimilar molecules formed by the similar organism (paralogous sequences).

In the case of cross species preservation, this indicates that a meticulous sequence may have been maintained by evolution despite speciation.

The further support the phylogenetic tree a particular conserved sequences may occur the more highly conserved it is said to be. Because sequence information is normally carried from parents to progeny by genes, a conserved sequence involves that there is a conserved gene; whereas conservative mutations are mutations that alter an amino acid to a diverse amino acid with alike biochemical properties (eg. charge, hydrophobicity and size). Conservative mutations in proteins often have a lesser consequence on function than non-conservative mutations. The compact outcome of conservative mutations on function can also be seen in the incidence of dissimilar mutations in nature. Non-conservative mutations between proteins are far more probable to be detached by natural selection due to their venomous effects **[37]**.

A. Phylogenetic Tree

A phylogenetic tree or evolutionary tree is a furcating illustration or tree viewing the condition evolutionary association between diverse biological species or other entities.

Their phylogeny based on similarities and deviations in their physical or genetic uniqueness. The taxa connected mutually in the tree are indirect to have descended from a same root.

Phylogenetic trees are essential to the area of phylogenetics. This phylogenetic tree is constructed by Neighbour-joining tree without distance correction method by CLISTAL O program **[39]**.

International Journal of Agriculture Innovations and Research
Volume 4, Issue 5, ISSN (Online) 2319-1473

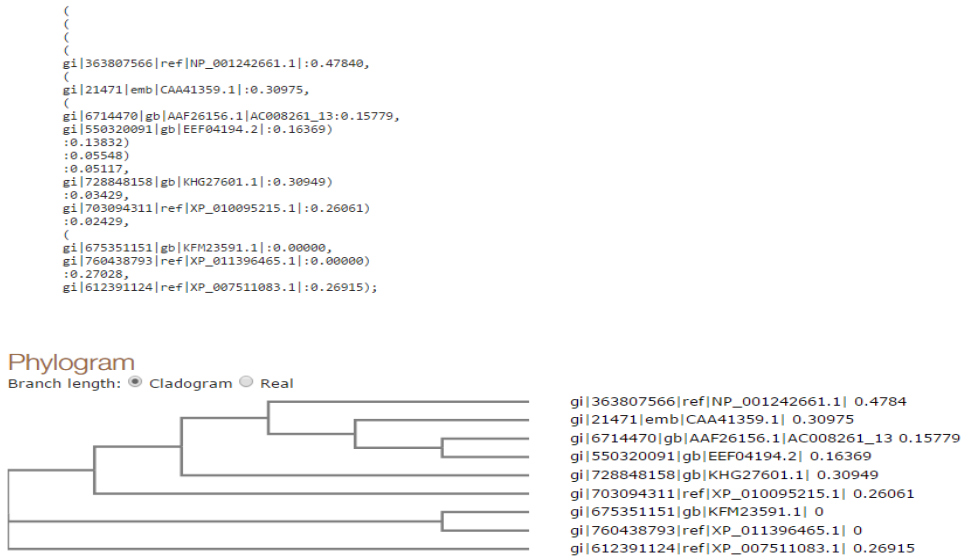


Fig. 1. A horizontal cladogram, with the root to the left

A cladogram derived from Greek clados "branch" and gramma "character", is a map used in cladistics analysis which shows associations between organisms. A cladogram is not; however, an evolutionary tree since it does not show how ancestors are related to offspring or how much they have distorted. Many evolutionary trees can be indirect from a particular cladogram. A cladogram uses lines that devise off in dissimilar directions ending at a clade, groups of organisms with a concluding common ancestor. There are many builds of cladograms but they all have lines that branch off from supplementary lines. The lines can be followed back to where they branch off. These branching off points symbolize a hypothetical ancestor (not a genuine entity) which is inferred to display the traits shared between the concluding taxa above it. This hypothetical ancestor might then supply clues about the arrangement of evolution of diverse features, alteration, and other evolutionary narratives about ancestors. Even if conventionally such cladograms were generated mostly on the basis of morphological typescript, DNA and RNA sequencing data and computational phylogenetics are nowadays extremely used in the generation of cladograms, either on their own or in amalgamation with morphology [40].

IV. Conclusion

We had concluded with above study that CLUSTAL O tool can be used for multiple sequence alignment for all nine model plants from different families and how we can generate phylogenetic tree from the same tool. With Fig 1 one can view the relations among model plants having same protein and from result of multiple sequence alignment it is apparently shown the conserved sequence, conservative mutations semi-conservative mutation, and non-conservative mutations among nine different sequences.

Hence, this multiple alignment tool is fast and accurate tool for agriculture research. The results can be further useful in various significant outcomes of agriculture research.

V. Acknowledgment

Authors are very grateful to Jacob School of Biotechnology and Bioengineering, Department of Computational Biology and Bioinformatics, Sam Higginbottom Institute of Agriculture, Technology and Sciences, Deemed University, Allahabad.

References

[1] Notredame, C. (2002) Recent progress in multiple sequence alignment: a survey. Pharmacogenomics 3, 131–144.
[2] Needleman, S. B. and Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.
[3] Smith, T. F. and Waterman, M. S. (1981) Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.
[4] Gotoh, O. (1982) An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705–708.
[5] Myers, E. W. and Miller, W. (1988) Optimal alignments in linear space. Comput. Appl. Biosci. 4, 11–17.
[6] Murata, M., Richardson, J. S., and Sussman, J. L. (1985) Simultaneous comparison of three protein sequences. Proc. Natl. Acad. Sci. USA 82, 3073–3077.
[7] Waterman, M. S. and Jones, R. (1990) Consensus methods for DNA and protein sequence alignment. Methods Enzymol. 183, 221–237.
[8] Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1999) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.
[9] Gonnet, G. H., Korostensky, C., and Benner, S. (2000) Evaluation measures of multiple sequence alignments. J. Comput. Biol. 7, 261–276.
[10] Wang, L. and Jiang, T. (1994) On the complexity of multiple sequence alignment. J. Comput. Biol. 1, 337–348.
[11] Bonizzoni, P. and Della Vedova, G. (2001) The complexity of multiple sequence alignment with SP-score that is a metric. Theor. Comput. Sci. 259, 63–79.
[12] Just, W. (2001) Computational complexity of multiple sequence alignment with SP-score. J. Comput. Biol. 8, 615–623.
[13] Elias, I. (2006) Settling the intractability of multiple alignment. J. Comput. Biol. 13, 1323–1339.

[14] Lipman, D. J., Altschul, S. F., and Kececioglu, J. D. (1989) A tool for multiple sequence alignment. Proc. Natl. Acad. Sci. USA 86, 4412–4415.
[15] Gupta, S. K., Kececioglu, J. D., and Schaffer, A. A. (1995) Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. J. Comput. Biol. 2, 459–472.
[16] Carrillo, H. and Lipman, D. (1988) The multiple sequence alignment problem in biology. SIAM J. Appl. Math. 48, 1073–1082.
[17] Dress, A., Fullen, G., and Perrey, S. (1995) A divide and conquer approach to multiple alignment. Proc. Int. Conf. Intell. Syst. Mol. Biol. 3, 107–113.
[18] Stoye, J., Perrey, S. W., and Dress, A. W. M. (1997) Improving the divide-andconquer approach to sum-of-pairs multiple sequence alignment. Appl. Math. Lett. 10, 67–73.
[19] Stoye, J., Moulton, V., and Dress, A. W. (1997) DCA: an efficient implementation of the divide-and-conquer approach to simultaneous multiple sequence alignment. Comput. Appl. Biosci. 13, 625–626.
[20] Stoye, J. (1998) Multiple sequence alignment with the divide-and-conquer method. Gene 211, GC45–56.
[21] Reinert, K., Stoye, J., and Will, T. (2000) An iterative method for faster sum-ofpairs multiple sequence alignment. Bioinformatics 16, 808–814.
[22] Holland, J. H. (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor. Protein Multiple Sequence Alignment 401
[23] Zhang, C. and Wong, A. K. (1997) A genetic algorithm for multiple molecular sequence alignment. Comput. Appl. Biosci. 13, 565–581.
[24] Anbarasu, L. A., Narayanasamy, P., and Sundararajan, V. (1998) Multiple sequence alignment using parallel genetic algorithms. SEAL.
[25] Chellapilla, K. and Fogel, G. B. (1999) Multiple sequence alignment using evolutionary programming. Congress on Evolutionary Computation.
[26] Gonzalez, R. R., Izquierdo, C. M., and Seijas, J. (1999) Multiple protein sequence comparison by genetic algorithms. SPIE-98.
[27] Cai, L., Juedes, D., and Liakhovitch, E. (2000) Evolutionary computation techniques for multiple sequence alignment. Congress on Evolutionary Computation.
[28] Zhang, G.-Z. and Huang, D.-S. (2004) Aligning multiple protein sequence by an improved genetic algorithm. IEEE International Joint Conference on Neural Networks.
[29] Notredame, C. and Higgins, D. G. (1996) SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res. 24, 1515–1524.
[30] Isokawa, M., Takahashi, K., and Shimizu, T. (1996) Multiple sequence alignment using a genetic algorithm. Genome Inform. 7, 176–177.
[31] Harada, Y., Wayama, M., and Shimizu, T. (1997) An inspection of the multiple alignment methods with use of genetic algorithm. Genome Inform. 8, 272–273.
[32] Hanada, K., Yokoyama, T., and Shimizu, T. (2000) Multiple sequence alignment by genetic algorithm. Genome Inform. 11, 317–318.
[33] Yokoyama, T., Watanabe, T., Taneda, A., and Shimizu, T. (2001) A web server for multiple sequence alignment using genetic algorithm. Genome Inform. 12, 382–383.
[34] Nguyen, H. D., Yoshihara, I., Yamamori, K., and Yasunaga, M. (2002) A parallel hybrid genetic algorithm for multiple protein sequence alignment. Evol. Comput. 1, 309–314.
[35] Kirkpatrick, S., Gelatt, J., C. D., and Vecchi, M. P. (1983) Optimization by simulated annealing. Science 220, 671–680.
[36] Ishikawa, M., Toya, T., Hoshida, M., Nitta, K., Ogiwara, A., and Kanehisa, M. (1993) Multiple sequence alignment by parallel simulated annealing. Comput. Appl. Biosci. 9, 267–273.
[37] Kim, J., Pramanik, S., and Chung, M. J. (1994) Multiple sequence alignment using simulated annealing. Comput. Appl. Biosci. 10, 419–426.
[38] Eddy, S. R. (1995) Multiple alignment using hidden Markov models. Proc. Int. Conf. Intell. Syst. Mol. Biol. 3, 114–120.
[39] Ikeda, T. and Imai, H. (1999) Enhanced A* algorithms for multiple alignments: optimal alignments for several sequences and k-opt approximate alignments for large cases. Theor. Comput. Sci. 210, 341–374.
[40] Horton, P. (2001) Tsukuba BB: a branch and bound algorithm for local multiple alignments of DNA and protein sequences. J. Comput. Biol. 8, 283–303.

AUTHOR'S PROFILE

Ms. Nivedita Yadav, Junior Research Fellow, at Jacob School of Biotechnology and Bioengineering, Department of Computational Biology and Bioinformatics, Sam Higginbottom Institute of Agriculture, Technology and Sciences, Allahabad-211007, Uttar Pradesh, India, have completed her Engg. Diploma in Information Technology from BTEUP, Gorakhpur, UP. Bachelor of Technology in Biotechnology from AITH, Kanpur, UP. Master of Technology in Bioinformatics from SHIATS, Allahabad, UP.
She has developed a database named ToPDb: Tomato crop Pathogen Database which is a manually created database which dissects type of pathogen, pathogen name, genes responsible for disease, disease name, symptoms of disease, controls of disease, pathogen etiology. available at http://www.e-bioinformatics.net/db/topdb/
She is working on microarray data analysis using bioinformatics approaches.
She has published the following papers during her academic years **Gene Expression Profiling of Transcription Factors of *Arabidopsis thaliana* using Microarray Data Analysis**" in **International Journal of Advanced Research in Computer Science and Software Engineering.** This paper is available at www.ijarcsse.com. In 2015
Developed of new technique for using *E.coli* DNA polymerase in PCR in International Journal of Applied Biotechnology and Biochemistry. This paper is available at http://www.ripublication.com in 2014

Mr. Apoorv Tiwari, Research Scholar at Department of Computational Biology and Bioinformatics, SHIATS, Allahabad, pursuing Doctor of Philosophy in Bioinformatics From SHIATS, Allahabad, have completed his B.Sc Biology from CSJM University, Kanpur, UP, M.Sc. (Bioinformatics) from UIET, CSJM University, Kanpur, UP and M.Tech (Bioinformatics) from SHIATS, Allahabad, UP.
Mr. Tiwari carrying his research works in the Department of Molecular Biology and Genetic Engineering, GBPUA&T, Pantnagar. His major research areas are Genomics, GBS and Transcriptome Data analysis. Main objective of his research are genome wide association mapping, diversity analysis, markers identification for agriculturally important traits and database development. He also involved in the development of a web interface AkritiV.1.0 which calculates physico-chemical property for Multi-Fasta protein.
Mr. Tiwari is a member of International Society of Computational Biology and SILAE: The Scientific and Cultural Network. He published 4 research papers in the reputed journals and 6 abstracts in the national and international conferences. The research papers he published are as follows:
1. **Genotyping-by-Sequencing Analysis for Determining Population Structure of Finger Millet Germplasm of Diverse Origins**. Kumar et al., *The plant genome* 9(2): (2016). This paper is available online at http://dx.doi.org/10.3835/plantgenome2015.07.0058
2. **MFPPI- Multi Fasta Prot Param Interface, Garg et al., Bioinformation** 12(2): 74-77 (2016). This paper is available online at http://www.bioinformation.net/012/97320630012074.htm
3. **High-throughput Omics Data for mining of important genes/traits linked to Agricultural Productivity: A National Bioinformatics workshop report**. Anil Kumar et al., *Int J Comput Bioinfo In Silico Model* 4(6): 749-752 (2015). This paper is available online at http://bioinfo.aizeonpublishers.net/content/2015/6/749-752.html
4. **In silico identification of MAPK3/6 substrates in WRKY, bZIP, MYB, MYB- related, NAC and AP-2 transcription factor family in Arabidopsis thaliana**, Avashthi et al., *Int J Comput Bioinfo In Silico Model* 3(4): 454-459 (2014). This paper is available online at http://bioinfo.aizeonpublishers.net/content/2014/4/bioinfo454-459.pdf

Mr. Vijay Kumar Garg, Senior Research Fellow at Department of Computational Biology and Bioinformatics, SHIATS, Allahabad, pursuing Doctor of Philosophy in Bioinformatics From SHIATS, Allahabad, have completed his B.Sc (Hons) in Zoology from BHU, Varanasi, UP, M.Sc. (Bioinformatics) from Kashi Vidyapeeth, Varanasi, UP and M.Tech (Bioinformatics) from SHIATS, Allahabad, UP.

He has developed a web interface AkritiV.1.0 which calculates physico-chemical property for Multi-Fasta protein, currently he is working on Lipoxygenase protein family lipoxygenase gene family which is mainly responsible for inflammatory, neurodegenerative, tumorigenic and cancerous disease, his area of specialization Genomics & Proteomics and in-silico Genome analysis. He has published the following paper.

1. **MFPPI- Multi Fasta Prot Param Interface**, Garg et al., Bioinformation 12(2): 74-77 (2016). This paper is available online at http://www.bioinformation.net/012/97320630012074.htm Mr. Garg is the Life member of Asian PGPR Society and SILAE: The Scientific And Cultural Network also he is awarded with following awards in different national and international conferences and congress proceedings listed as below:

a). Best Poster Presentation Award in **6th World Congress on Biotechnology** Conference held at New Delhi October 5th-7th 2015.

b). Second Poster Award in **National Conference on Plant & Animal Molecular Biology (NCPAMB-2015)** conference held at Mody University Lakshamanarh, Sikar, Rajsthan September 25-26, 2015

c). Third Poster Presentation Award in **Third (3rd) Uttar Pradesh Agricultural Science Congress,** held at SHIATS, Allahabad 14th-16th June 2015