MAP Kinase analyser: A tool for plant kinase and substrate analysis

Saranga Dhar Samantaray[1], Parvendra Singh[1], Bonthla Venkata Suresh[2], Sugandha Sharma[2], Gohar Taj[2]*, Anil Kumar[2]

[1]Department of Computer Science & Technology, College of Basic Sciences & Humanities, G.B. Pant University of Agriculture & Technology, Pantnagar, Uttrakhand-263145, India; [2]Department of Molecular Biology & Genetic Engineering, College of Basic Sciences & Humanities, G.B. Pant University of Agriculture & Technology, Pantnagar, Uttrakhand-263145, India; Gohar Taj - Email: gohartajkhan@rediffmail.com; *Corresponding author

Abstract:
MAPK (Mitogen Activated Protein Kinase) is a Ser/Thr kinase, which plays a crucial role in plant growth and development, transferring the extra cellular stimuli into intracellular response *etc*. Manual identification of these MAPK in the plant genome is tedious and time taking process. There are number of online servers which predict the P-site (phosphorylation site), find the motifs and domain but there is no specific tool which can identify all them together. In order to identify the P-Site, phosphorylation site consensus sequences and domain of the MAPK in plant genome, we developed a tool, MAP Kinase analyzer. MAP kinase analyzer take protein sequence as input in the fasta format and the output of tool includes: 1) The prediction of the phosphorylation site *viz.,* Serine (S), Threonine (T), and Tyrosine (Y), Contex, Position, Score and phosphorylating kinase as well as the graphical output; 2) Phosphorylation site consensus sequence pattern for different kinases and 3) Domain information about the MAPK's. The MAP kinase analyser tool and supplementary files can be downloaded from http://www.bioinfogbpuat/mapk_OWN_1/.

Background:
MAPK are special class of kinases which are activated by various growth factors, differentiating factors, M-phase phosphorylation cascade reactions and involved in biotic and abiotic stress signaling pathways [1]. They play a key role in the transmission of external signals such as mitogens, hormones and different stresses. A number of prediction servers are available over the World Wide Web. These servers facilitate the prediction like GPS: a comprehensive www server for phosphorylation sites prediction; PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory; PhoScan: Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach; MEME Suite: Motif-based sequence analysis tools; FANMOD: a tool for fast network Protein Consensus Sequence Motif detection; SMART (a Simple Modular Architecture Research Tool); d-Omix: a mixer of generic protein domain analysis tool; Phospho.ELM etc. Manual identification of MAPKs are tedious and time consuming, in order to identify MAPK there are no specific tools as yet, which predicts altogether the phosporylation site (P-Site), P-site consensus sequence or pattern and domain. Thus we developed a tool MAP kinase analyzer, which solves most of the above faced problems.

Methodology:
Datasets:
For training, the P-sites, positive datasets were obtained from Phospho.ELM database [2], which contains 2540 substrate proteins from different species covering 4799 S, 974 T and 1433 Y sites. To remove redundant fragments within the datasets, the initial datasets were filtered using a 30% sequence identity. The negative (*i.e.* non-phosphorylation sites, N_S, N_T and N_Y) were obtained from these 2540 protein sequences and represented all S, T and Y residues that were not reported as being phosphorylated in Phospho.ELM database. For the phosphorylation site consensus sequence and domain analysis, the primary data is retrieved from TAIR [3], NCBI [4] and UniProt databases [5], and after retrieval these data are manually as well as with the help of tool like Multalin [6], optimized.

Algorithm:
In the case of P-site prediction we designed a neural network for the prediction and genetic algorithm for training this neural network which can be called as genetic algorithm neural network (GANN) [7]. The GANN uses GA (Genetic Algorithm) to optimize the connection weights of the ANN (Artificial Neural Network) over the training dataset. In our GANN model, the number of input nodes is equal to the dimensionality of feature vector, *i.e* 24. The neural network uses a sigmoid function to provide a continuous activation function. GANN is used to construct a P-site predictor with the following configuration: (1) the maximum generation number–1000; (2) the threshold of the fitness value–0.7; (3) population size–100; (4) crossover probability –0.9. In case of P-site consensus pattern and domain search for the MAP kinases, firstly MSA (Multiple Sequence Alignment) has been done using Multalin tool. From MSA, we detected 11 conserved sub-domains which are present in all plant MAPKs. Among all sub-domains, 8 sub-domains contain the MAP kinase activation loop ((**TE/D) Y**) and the MAPK's signature sequence (**T(E/D) YxxTRWYRAPEL**) [8]. These sub-domains are optimized and included in the tool for prediction. This tool is developed on NetBeans suite with Java language [9].
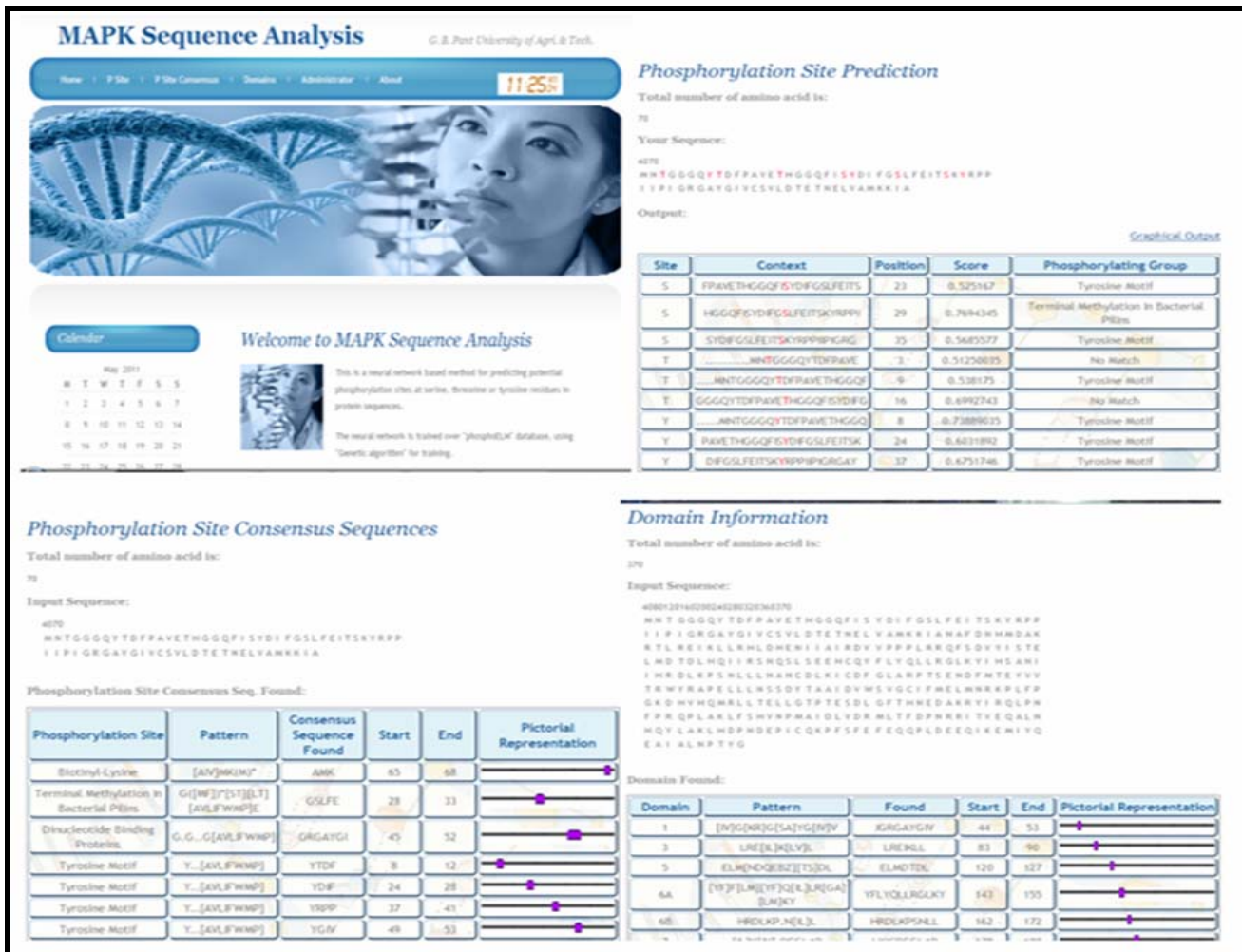
Figure 1: A Snapshots of the MAP kinase analyzer showing homepage, P-site prediction, P-site consensus sequence pattern and domain prediction of MAPK.

User interface:
In the case of P-site prediction, when user inputs a sequence in fasta format, the sequence is then processed by the neural network and result is displayed in the form of numeric scores corresponding to all S, T and Y present in the sequence and also highlight them. The user can opt for a threshold value at which user wants to see the P-sites. In the case of P-site consensus sequence pattern or motif and domain analysis for MAPK, a file is created containing all the patterns that have to be searched for in the given sequence. The sequence is then processed and the motif as well as domains present in the sequence is displayed to the user in text as well as graphical representation form.

Utility:
A novel tool for P-site, P-site consensus sequence pattern and for the domain prediction of MAPK has been developed (**Figure 1**). We have given a graphical and user friendly interface, so the tool is easy to use. Through this tool we can identify whether the given protein sequence is a MAP kinase or not on the basis of presence of the specific MAPK domain, in addition, we can also identify the possible kinase substrate by the analysis of P-site consensus sequence pattern, which consequently gives an idea about the functioning of the protein.

Conclusion:
Performance evaluation with dataset and database variants clearly indicates that MAP kinase analyzer has significantly high accuracy in terms of specificity and sensitivity. To the best of our knowledge MAP kinase analyzer is the first ever tool which identifies the P-Site, phosphorylation site consensus sequences and domain of the MAPK in plant genome altogether.

References:
[1] Jonak *et al. Cell Mol Life Sci*. 1999 **55**: 204 [PMID: 10188583]
[2] http://phospho.elm.eu.org/
[3] http://www.arabidopsis.org
[4] http://blast.ncbi.nlm.nih.gov
[5] http://www.uniprot.org
[6] http://www.multalin.toulouse.inra.fr/multalin/multalin.html
[7] http://www.cbs.dtu.dk/services/NetPhos/
[8] Nadarajah K & Sidek HM. *Asian J Plant Sci*. 2010 **9**: 1
[9] http://www.java.com