academicJournals

Vol. 12(39), pp. 5804-5808, 25 September, 2013
DOI: 10.5897/AJB12.692
ISSN 1684-5315 ©2013 Academic Journals
http://www.academicjournals.org/AJB

African Journal of Biotechnology

Full Length Research Paper

Identification of proteins similar to AvrE type III effector proteins from *Arabidopsidis thaliana* genome with partial least squares

Stephen O. Opiyo[1*] and David Mackey[2]

[1]Molecular and Cellular Imaging Center-Columbus, Ohio Agricultural Research and Development Center, Columbus, OH 43210, U. S. A.
[2]Department of Horticulture and Crop Science, Ohio State University, Columbus, OH 43210.

Accepted 2 August, 2013

Type III effector proteins are injected into host cells through type III secretion systems. Some effectors are similar to host proteins to promote pathogenicity, while others lead to the activation of disease resistance. We used partial least squares alignment-free bioinformatics methods to identify proteins similar to AvrE proteins from *Arabidopsidis thaliana* genome and identified 61 protein candidates. Using information from Genevestigator, Arabidopsidis GEB, KEGG, (GEO: accession number GSE22274), and AraCyc databases, we highlighted 16 protein candidates from Arabidopsidis genome for further investigation.

Key words: Partial least squares, Type III effectors, AvrE, and Arabidopsis.

INTRODUCTION

Plant pathogens deliver small molecules referred to as effectors, by type III secreting systems (T3SS) directly into plants (Abramovitch et al., 2006; Block et al., 2008; Zhou and Chai, 2008). The injected effectors target different cellular compartments and subvert numerous signaling pathways for the benefit of the bacteria. Through the resistance (R) proteins, plants evolved to gain the ability to recognize directly or indirectly effectors. Several T3SS effectors contribute to virulence by sup-pressing Pathogen-Associated Molecular Patterns (PAMPs) (Hauck et al., 2003), and other effectors suppress hypersensitive cell death elicited by various Avr proteins (Abramovitch et al., 2003). Some effectors mimic plant proteins (Bender et al., 1999; Weiler et al., 1994), while others mimic plant molecules (Janjusevic et al., 2006; Rosebrock et al., 2007). AvrE are type III effectors proteins with very low sequence identity (Ham et al., 2009).

Plant genomics and many agriculturally important crops are resulting in a rapidly increasing database of genomic and sequences. These databases have proved to be rich resources for several genes of importance agronomic traits, such as, virus and insect resistance, bacterial resistance, abiotic stress tolerance, and novel genetic markers for crop improvements. Silverstein et al. (2005) searched Arabidopsidis genome using profile hidden Markov model (HMM) (Durbin et al., 1998) and Basic Alignment Search Tools (BLAST) (Altschul et al., 1990) to identify defensin-like sequences (DEFLs) in *Arabidopsidis* genome. They identified 317 DEFLs in *Arabidopsidis* including 15 known defensins. Thus, bioinformatics has become an integral aspect of plant and crop science research. The objective of the study was to identify proteins that are similar to AvrE-family effector proteins from Arabidopsidis genome with partial least squares (PLS) alignment-free methods (Opiyo and Moriyama, 2007).

Alignment-based methods have limitations because alignments are known to become unreliable when sequence similarity drops below 40% (Petsko and Ringe,

*Corresponding author. E-mail: opiyo.1@osu.edu. Tel: 614-292-7717. Fax: 614-292-4455.

2003). Some proteins such as AvrE are highly divergent and have low sequence identity (WtsE and AvrE have 27.1% amino acid identity) even though they still share similar structures, biochemical properties, and functions. In such cases, obtaining reliable alignments among these protein sequences is extremely difficult, and alignment-based methods such as BLAST, position specific iterative BLAST (PSI-BLAST) (Altschul et al., 1997), and profile HMMs would fail to identify these proteins from databases. Using PLS alignment-free methods, we predicted 61 protein candidates from Arabidopsidis genome as similar to AvrE effectors. Using information from Genevestigator v3 (Hruz et al., 2008), Arabidopsidis Gene Expression Browser (GEB) (Zhang et al., 2010), KEGG: Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000), (GEO: Gene Expression Omnibus (Edgar et al., 2002); accession number GSE22274 (Wang et al., 2011), and AraCyc (Mueller et al., 2003), we highlighted 16 protein candidates for further investigation.

MATERIALS AND METHODS

Dataset sources

Training dataset

Twelve (12) AvrE proteins (positives) from study by Ham and associates (Ham et al., 2009) and non-AvrE proteins (negatives) were downloaded from National Center for Biotechnology Information (NCBI) websites (http://www.ncbi.nlm.nih.gov/), and were used for training the PLS methods.

Databases

Arabidopsidis thaliana: 35 386 proteins from the release 10 (November, 2010) of The Arabidopsidis Information Resource (TAIR) database (http://www.arabidopsidis.org/).

Sequence descriptors used for PLS alignment-free methods

Amino acid composition: From each protein sequence, frequencies of 20 amino acids were calculated. In this study, amino acid composition was used as descriptors for a PLS classifier (PLS-AA).

Dipeptide composition: Dipeptide composition represents all 400 frequencies of consecutive amino acid pairs in a protein sequence and corresponds to a 400 (20 × 20) feature vector. It can encapsulate information on composition of amino acids, as well as, their local order. We used dipeptide composition as descriptors for a PLS classifier (PLS-DIP).

Physicochemical properties of amino acids: We developed five descriptors (PC1- PC5) using the principal component analysis (PCA) of 12 physicochemical properties of amino acids (mass, volume, surface area, hydrophilicity, hydrophobicity, isoelectric point, transfer of energy solvent to water, refractivity, non-polar surface area, and frequencies of alpha-helix, beta-sheet, and reverse turn) (Opiyo and Moriyama, 2007). The five descriptors were used in this study.

Auto/cross covariance transformation: Auto/cross covariance (ACC) transformation method discussed in Opiyo and Moriyama (2007) was used to transform each amino acid sequence using the five physicochemical property based descriptor set (PC1-PC5). ACC with the maximum lag of 30 residues yielded 775 descriptors for each sequence. The calculation of ACC was performed using the R implementation (version 2.12.0; http//www.R-project.org; 2010).

Partial least squares

Partial least squares [PLS; (Geladi and Kowalski, 1986)] is a projection method similar to principal component analysis (PCA) where the independent variables, represented as the matrix X, are projected onto a low dimensional space. PLS uses both independent variables X (sequence descriptors such as amino acid composition) and dependent variables Y (positive or negative label). PLS using descriptors transformed by ACC (PLS-ACC) was used in (Opiyo and Moriyama, 2007). PLS discriminate analysis is performed to separate groups of observations. It consists of a classical PLS where the response variable is a categorical one (replaced by the set of dummy variables describing the categories, e.g., 0 and 1) expressing the class membership of the statistical units. In this study, each of a training sample, a response variable was a signed 1 for the positive sample (AvrE) and 0 for a negative sample (non-AvrE). The group membership, AvrE or non-AvrE of a new sequence was predicted based on descriptors and y-value. Predicted y-value closer to 1 was considered to be AvrE candidate and closer to 0 WAS considered to be non-AvrE candidate. PLS analysis was performed using an R implementation; the PLS package was developed by Wehrens and Mevik (version 1.2.1) (Wehrens and Mevik, 2007).

Performance analysis

Cross-validation analysis (leave-one-out) was performed for all the 24 sequences used for training the methods. One sequence in the training dataset was left out and the learning algorithm was trained on the rest of the sequences. The trained model was used to predict the class (AvrE or non-AvrE) of the earlier left out. For the 24 sequences, the process was repeated 24 times leaving each of the 24 sequences out and creating a model from the remaining 23 sequences.

Predictions were grouped as follows: i) True Positives (TP): the number of actual AvrEs that were predicted as AvrEs; ii) False Positive (FP): the number of actual non-AvrEs that were predicted as AvrEs; iii) True Negative (TN): the number of actual non-AvrEs that were predicted as non-AvrEs and iv) False Negative (FN): the number of actual AvrEs that were predicted as non-AvrEs.

Minimum error point

The minimum error point (Karchin et al., 2002) was used to determine threshold values of PLS methods. The sequences are ranked based on the values. The threshold value where the minimum number of errors (FN + FP) occurs is the minimum error point (MEP) and the number of false positives and false negatives are assessed at this point. The minimum error point tells us the best case accuracy of a method. The minimum error points for PLS-AA, PLS-DIP and PLS-ACC were 0.94, 0.96 and 0.94, respectively. The upper cut-off point for all methods was set at 1.00 to further reduce the number of false positives. To be selected as a candidate, a protein has to be identified by all the three methods (PLS-AA, PLS-DIP and PLS-ACC) as positive.

Goodness of Prediction of PLS methods

The goodness of prediction, Q^2 equation 2, describes how well the

Table 1. The number of PLS components and the predictive abilities of PLS-AA, PLS-DIP, and PLS-ACC, respectively from the leave-one-out cross validation procedures.

Method	Number of PLS components	Q^2
PLS-AA	4	0.72
PLS-DIP	3	0.67
PLS-ACC	4	0.78

method can predict a data.

$$Q^2 = 1\text{-}PRESS/SS_Y \quad (2)$$

Where SS_Y is the total sum of squares, PRESS is the predictive residual sum of squares, which is calculated from the difference between observed and predicted Y values. $Q^2 > 0.50$ is considered good. In this study, the leave-one-out cross-validation procedure was used for the Q^2 calculation. Detailed results of PLS analyses are given in Table 1 for PLS-AA, PLS-DIP, and PLS-ACC, respectively.

RESULTS AND DISCUSSION

Mining *A. thaliana* proteome using three PLS methods

Our objective was to identify proteins similar to AvrEeffector proteins from Arabidopsidis genome. PLS methods trained using 12 AvrE-family effector proteins predicted 61 protein candidates from Arabidopsidis genome. Thirty-eight proteins (62%) were enzymes, and they included kinases, hydrolases, and proteases. Other proteins predicted were F-box family protein, unknown protein, transcription factor, auxin-responsive family protein, and other proteins. In order to further study the predicted proteins, we used Genevestigator, Arabidopsidis GEB, KEGG, (GEO: accession number GSE22274), and AraCyc databases to analyze the proteins.

Expression patterns of the predicted proteins in Genevestigator and Arabidopsidis Gene Expression Browser databases

We utilized the server of Genevestigator and Arabidopsidis GEB databases to study expression patterns of the predicted proteins. Of the 61 proteins predicted, three (AT2G44280, AT3G59590, and ATMG00140) had no expression data in the Genevestigator database. In this study, only responses with expression levels altered by more than two-fold under the biotic stress are presented. Protease inhibitor/seed storage/lipid transfer protein (LTP) family protein (AT2G13820), unknown protein (AT2G17850), and receptor protein kinase-related (AT3G46270) were down-regulated by both *Cryphonectria arabidopsidis,* and *Pseudomonas syringae*. Mean while, phosphoglycerate kinase (AT1G79550), cyclin-dependent kinase B1; 2 (AT2G38620), and exopolygalacturonase / galacturan 1, 4-alpha-galacturonidase / pectinase (AT3G07850) proteins were specifically down-regulated by *P. syringae*. In addition, invertase/pectin methylesterase inhibitor family protein (AT2G47340) was mainly up-regulated by *P. syringae.* From Arabidopsidis GEB, glycosyl hydrolase family 17 proteins / beta-1, 3-glucanase (AT3G55430), lipase class 3 family protein (AT3G62590), Senescence-Associated Gene 101 (AT5G14930), and glutathione transferase (AT5G62480) were up-regulated by *P. syringae, Golovinomyces orontii, and Botrytis cinerea*. These data reveal that AvrE proteins might mimic *Arabidopsidis* proteins that are up-regulated or/and down-regulated by both fungi and bacteria.

Metabolic pathways identified from KEGG and AraCyc databases

Out of the 61 proteins predicted, 13 were linked to KEGG metabolic pathways. The KEGG pathways linked to the predicted proteins are Typtophan metabolism (ath00380), biosynthesis of secondary metabolites (ath01110), starch and sucrose metabolism (ath00500) and others. The AraCyc software (http:// www.arabidopsidis.org/tools/aracyc) provides a good starting point to paint expression data on metabolic pathways. AraCyc metabolic pathways linked to the predicted proteins were oxidative ethanol degradation and superoxide radicals degradation (AT1G20620), abscisic acid glucose ester biosynthesis (AT2G29740), gibberellin biosynthesis III (AT1G80330), choline and phosphatidylcholine biosynthesis (AT2G32260), trehalose biosynthesis (AT4G12430), and photorespiration (AT5G47760). AT2G29740 protein *syringae* as shown in Genevestigator involved in abscisic acid glucose ester biosynthesis was up-regulated by *P.* database. These results show that AvrE proteins might mimic proteins that involve metabolomics pathways related to biosynthesis of secondary metabolites, steroid and trahalose biosynthesis, abscisic and gibberellin biosynthesis as shown from both KEGG and AraCyc databases. Based on the information from Genevestigator, Arabidopsidis GEB, KEGG, and AraCyc, we highlighted 16 protein candidates as priorities for further investigation (Table 2). We predicted protein subcellular localizations by WoLF PSORT (Horton et al., 2007); and the predictions show that 50% (8 proteins) of the 16 proteins candidates are located in Cytosol. These protein candidates were identified by computational predictions; experiments are ultimately needed to determine if they are mimic by AvrE effector proteins.

Conclusions

In this study, we predicted 61 proteins from *Arabidopsidis*

Table 2. Sixteen protein sequences highlighted for further investigations.

Accession number	Length (aa)	TAIR Description
AT1G20620	485	SEN2, CAT3 CAT3 (CATALASE 3); catalase chr1:7143132-7146183 FORWARD
AT1G79550	401	PGK PGK (PHOSPHOGLYCERATE KINASE) chr1:29929240-29931188 REVERSE
AT1G80330	355	ATGA3OX4, GA3OX4 ATGA3OX4 (GIBBERELLIN 3-OXIDASE 4); gibberellin 3-beta-dioxygenase chr1:30202953-30204429 REVERSE
AT2G13820	129	protease inhibitor/seed storage/lipid transfer protein (LTP) family protein chr2:5782887-5783361 REVERSE
AT2G29740	474	UDP-glucoronosyl/UDP-glucosyl transferase family protein chr2:12713824-12715248 FORWARD
AT3G07850	444	exopolygalacturonase / galacturan 1,4-alpha-galacturonidase / pectinase chr3:2505819-2507444 REVERSE
AT3G25070	211	RIN4 RIN4 (RPM1 INTERACTING PROTEIN 4); protein binding chr3:9132465-9133754 FORWARD
AT3G45640	370	MPK3, ATMPK3 ATMPK3 (MITOGEN-ACTIVATED PROTEIN KINASE 3); MAP kinase/ kinase/ protein kinase chr3:16767903-16769461 FORWARD
AT3G53250	109	auxin-responsive family protein chr3:19753946-19754275 FORWARD
AT3G55430	449	glycosyl hydrolase family 17 protein / beta-1,3-glucanase, putative chr3:20560783-20562981 REVERSE
AT3G62590	649	lipase class 3 family protein chr3:23158949-23161145 REVERSE
AT4G04740	520	CPK23 CPK23 (calcium-dependent protein kinase 23); calmodulin-dependent protein kinase/ kinase chr4:2405404-2408491 REVERSE
AT4G12720	282	NUDT7, GFG1, AtNUDT7 AtNUDT7 (ARABIDOPSIDIS THALIANA NUDIX HYDROLASE HOMOLOG 7); hydrolase/ nucleoside-diphosphatase chr4:7487713-7489554 FORWARD
AT5G14930	239	GENE101, SAG101 SAG101 (SENESCENCE-ASSOCIATED GENE 101) chr5:4828757-4830168 FORWARD
AT5G48870	88	SAD1 SAD1 (SUPERSENSITIVE TO ABA AND DROUGHT 1) chr5:19830633-19831588 FORWARD
AT5G62480	240	GST14, GST14B, ATGSTU9 ATGSTU9 (GLUTATHIONE S-TRANSFERASE TAU 9); glutathione transferase chr5:25106001-25106792 REVERSE

genome as proteins that are similar to AvrE proteins using PLS alignment-free bioinformatics method. Furthermore, we used information from gene expression data, and metabolomics pathways to highlight 16 proteins for further investigations. This study suggests that using different PLS alignment-free bioinformatics methods com-

bined with information from available databases offers a promising approach to predict proteins that are similar to AvrE proteins. Such approaches may address a challenging issue of effector target discovery.

REFERENCES

Abramovitch RB, Janjusevic R, Stebbins CE, Martin GB (2006). Type III effector AvrPtoB requires intrinsic E3 ubiquitin ligase activity to suppress plant cell death and immunity. Proc. Natl. Acad. Sci. USA 103(8): 2851-2856.

Abramovitch RB, Kim YJ, Chen S, Dickman MB, Martin GB (2003). Pseudomonas type III effector AvrPtoB induces plant disease susceptibility by inhibition of host programmed cell death. EMBO J. 22(1): 60-69.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17): 3389-3402.

Altschul SF, Gish W, Webb Miller, Eugene W. Myers, David J. Lipman (1990). Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Bender CL , Alarcón-Chaidez F, Gross DC (1999). Pseudomonas syringae phytotoxins: mode of action, regulation, and biosynthesis by peptide and polyketide synthetases. Microbiol. Mol. Biol. Rev. 63(2): 266-292.

Block A, Li G, Fu ZQ, Alfano JR (2008). Phytopathogen type III effector weaponry and their plant targets. Curr. Opin. Plant Biol. 11(4): 396-403.

Durbin R, Eddy SR, Krogh A, Mitchison G (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, Cambridge University Press.

Edgar R, Domrachev M, Lash AE (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30(1): 207-210.

Geladi P, Kowalski BR (1986). Partial least squares regression: A tutorial. Anal. Chim. Acta. 185: 1-7.

Ham JH, Majerczak DR, Nomura K, Mecey C, Uribe F, He SY, Mackey D, Coplin DL (2009). Multiple activities of the plant pathogen type III effector proteins WtsE and AvrE require WxxxE motifs. Mol. Plant Microbe Interact. 22(6): 703-712.

Paula Hauck, Roger Thilmony, Sheng Yang He (2003). A Pseudomonas syringae type III effector suppresses cell wall-based extracellular defense in susceptible Arabidopsis plants. Proc. Natl. Acad. Sci. USA 100(14): 8577-8582.

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007). WoLF PSORT: protein localization predictor. Nucleic Acids Res. 35(Web Server issue): W585-587.

Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P (2008). Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv. Bioinformatics 2008: 420747.

Janjusevic R, Abramovitch RB, Martin GB, Stebbins CE (2006). A bacterial inhibitor of host programmed cell death defenses is an E3 ubiquitin ligase. Science 311(5758): 222-226.

Kanehisa M, Goto S (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1): 27-30.

Karchin R, Karplus K, Haussler D (2002). Classifying G-protein coupled receptors with support vector machines. Bioinformatics 18(1):147-159.

Mueller LA, Zhang P, Rhee SY (2003). AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol. 132(2): 453-460.

Opiyo SO, Moriyama EN (2007). Protein family classification with partial least squares. J. Proteome Res. 6(2): 846-853.

Petsko GA, Ringe D (2003). Protein Structure and Function. Primers in Biology. Sunderland (Massachusetts), Science Press; London, in association with Blackwell Publishing; Oxford, and Sinauer Associates.

Rosebrock TR, Zeng L, Brady JJ, Abramovitch RB, Xiao F, Martin GB (2007). A bacterial E3 ubiquitin ligase targets a host protein kinase to disrupt plant immunity. Nature 448(7151): 370-374.

Silverstein KA, Graham MA, Paape TD, VandenBosch KA (2005). "Genome organization of more than 300 defensin-like genes in Arabidopsis. Plant Physiol. 138(2): 600-610.

Wang W , Barnaby JY, Tada Y, Li H, Tör M et al. (2011). Timing of plant immune responses by a central circadian regulator. Nature 470(7332): 110-114.

Wehrens R, Mevik B (2007). pls: Partial Least Squares Regression(PLSR) and Principal Component Regression (PCR). R package version 1.2-1.

Weiler EW, Kutchan TM, Gorba T, Brodschelm W, Niesel U, Bublitz F (1994). The Pseudomonas phytotoxin coronatine mimics octadecanoid signalling molecules of higher plants. FEBS Lett. 345(1): 9-13.

Zhang M, Zhang Y, Liu L, Yu L, Tsang S, Tan J, Yao W, Kang MS, An Y, Fan X (2010). Gene Expression Browser: large-scale and cross-experiment microarray data integration, management, search & visualization. BMC Bioinformatics 11: 433.

Zhou JM, Chai J (2008). Plant pathogenic bacterial type III effectors subdue host responses. Curr. Opin. Microbiol. 11(2): 179-185.