Published in final edited form as:
Int J Bioinform Res Appl. 2009 ; 5(3): 295–309.

Mining the *Arabidopsis* and Rice Genomes for Cyclophilin Protein Families

S.O. Opiyo and
Department of Agronomy and Horticulture, University of Nebraska-Lincoln 68583. E-mail: sopiyo@unlserve.unl.edu

E.N. Moriyama
School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln, 403 Manter Hall, Lincoln, NE 68588-0118; E-mail: emoriyama2@unl.edu

Abstract

Cyclophilins are a family of proteins that possess peptidyl-prolyl isomerase activity. They are present in both eukaryotes and prokaryotes. They are cellular targets of immunosuppressant drugs and involved in a wide variety of functions. The *Arabidopsis thaliana* genome contains the largest number of cyclophilins. However, the total number of plant cyclophilins available in sequence databases is small compared to that of other organisms. This implies that many cyclophilins are not yet identified in plants. In order to identify cyclophilin candidates from available plant sequence data, we examined alignment-free methods based on partial least squares (PLS) using physico-chemical properties for the mining of single and multiple-domain cyclophilins. PLS with selected descriptors after auto and cross-covariance (ACC) transformation had low false positives compared to PLS with all ACC descriptors. The former PLS classifier also performed better than profile hidden Markov models and PSI-BLAST in identifying cyclophilins from the *Arabidopsis* and rice genomes.

Keywords

Cyclophilins; partial least squares; profile hidden Markov model

1 Introduction

Cyclophilins possess the peptidyl-prolyl isomerase (PPIase; EC 5.2.1.8) activity and are involved in diverse cellular processes including cell cycle control, receptor signaling, protein folding as well as being cellular targets of immunosuppressant drugs (Romano et al., 2004). In the presence of their drug ligand, cyclosporine A (CsA), cyclophilins gain their immunosuppressing function by forming a complex with cyclosporine A. This complex blocks T-cell activation by binding to the T-cells and inhibiting the activity of calcineurin.

In the absence of immunosuppressive drugs, on the other hand, cyclophilins are involved in a variety of cellular processes. For example, cyclophilins have been shown to play roles in both plant and animal pathogen recognition. The interaction of *Agrobacterium tumefaciens* virulence protein (VirD2) with *Arabidopsis* cyclophilin AtCYP19 has been reported (Deng et

Correspondence to: E.N. Moriyama.

Reference to this paper should be made as follows: Opiyo, S.O and Moriyama, E.N. (200x) `Mining the *Arabidopsis* and Rice genomes for Cyclophilin Protein Families', *Int. J. Bioinformatics Research and Applications*, Vol. x, No. x, pp.xx.

al., 1998). *Agrobacterium* recruits plant cyclophilins for transferring and integrating T-DNA (transferred DNA) into a plant cell. AtCYP18 has been identified to activate a *Pseudomonas syringae* effector protein (AvrRpt2) by its PPIase activity (Coaker et al., 2005). In the case of the fungus *Magnaporthe grisea* infection in rice plants, which causes rice blast disease, a fungal cyclophilin (CYP1) acts as a virulence determinant (Muriel et al., 2002). Cyclophilins have been purified from seeds of cow pea, mung bean, and chickpea (Ye and Ng, 2000). These proteins possess antifungal activity against several fungi including *Mycosphaerella arachidichola*. The chickpea protein is also known to inhibit the activity of human immunodeficiency virus-1 reverse transcriptase (Ye and Ng, 2000)

Cyclophilins are classified into single-domain and multiple-domain families. Single-domain cyclophilins contain only the cyclophilin catalytic domain, and their average length is 172 amino acids (aa). Multiple-domain cyclophilins have other functional domains in addition to the cyclophilin catalytic domain. Their average length is 550 aa. The other domains are expected to play roles in determining specific functions. For example, the “tetratricopetide (TPR) domain” is involved in protein-protein interactions. The TPR domain is a 34-amino-acid motif. It exists usually as multiple tandem repeats in proteins with many cellular functions, including mitosis, transcription, protein transport, and development (Lamb et al., 1995). Proteins that contain TPR motifs include members of the FK506- and rapamycin-binding proteins, organelle-targeting proteins, TPR multiple-domain cyclophilins that facilitate assembling of protein complexes and protein phosphatases (Lamb et al., 1995).

While amino acid sequences of single-domain cyclophilins are in general divergent, their secondary structures remain well conserved (Galat, 2003). The structure of cyclophilin proteins consists of eight stranded anti-parallel β-sheets capped at both ends by two helices. Numerous insertions/deletions are observed in the loop regions. However, the amino acid residues crucial for the PPIase activity and CsA-binding are well conserved even in the long loop regions (Galat, 2003). Most of the amino acids involved with the PPIase activity are also known to be important for CsA-binding.

The *Arabidopsis thaliana* genome, in spite of its relatively small genome size, contains the largest number of known (experimentally confirmed) cyclophilin proteins, 29 of them in total (21 single-domain and 8 multiple-domain proteins; Romano et al., 2004). On the contrary, metazoa are known to have a fewer number of cyclophilins. There are 19 human cyclophilins and 14 found in *Drosophila melanogaster*. However, surprisingly, the number of cyclophilin sequences available from plants found in sequence databases is much smaller compared to those from animals and other higher eukaryotes. For example, in InterPro (Release 16.0; Mulder et al., 2005), there are 302 cyclophilin sequences from plants, 595 from animals, 321 from fungi, and 1319 from bacteria. This indicates that currently we do not have sufficient information on cyclophilin proteins from plants, even though they could provide the largest amount of information on these protein functions. In order to learn more about these cyclophilin proteins, more thorough searches are needed from available sequence data.

The most popularly used methods for protein family classification include Basic Local Alignment Search Tool (BLAST; Altschul et al., 1997), Position Specific Iterative-BLAST (PSI-BLAST; Altschul et al., 1997), and profile hidden Markov models (profile HMMs; Durbin et al., 1998). Because these methods require reliable alignments to compare sequences, they do not perform well on extremely diverged sequences and those with multiple-domains such as cyclophilin proteins. Another problem with these methods is that the models are built using only “positive” samples (proteins of interest). Previously we have shown that physico-chemical properties of amino acids can be used for mining proteins (Opiyo and Moriyama, 2007; Strope and Moriyama, 2007). This approach does not require aligning sequences and are known to be more sensitive to remote similarities.

The objectives of this study are 1) to develop alignment-free protein classification methods using physico-chemical properties of amino acids that can effectively identify cyclophilin protein families, and 2) to mine cyclophilins from *Arabidopsis* and rice genomes.

2 Materials and Methods

2.1 Dataset

Two hundred and eighty single-domain cyclophilin sequences (100 from animals, 60 from plants, 40 from fungi, and 80 from bacteria), were downloaded from InterPro (Release 13.1; Mulder et al., 2005), and divided to prepare training and test datasets (Table 1). Although the TPR multiple-domain cyclophilins are the largest multiple-domain cyclophilins found in InterPro, only 36 sequences (21 from animals, 5 from plants, and 10 from fungi) were available. Only one dataset was thus generated for TPR multiple-domain cyclophilins. The entire proteins of TPR multiple-domain cyclophilins including both of cyclophilin and TPR domains were used for training classifiers. Negative data (non-cyclophilin proteins) were obtained from Swiss-prot database.

The entire protein sequence sets for *Arabidopsis thaliana* (28,952 proteins; release 5, dated June 2004), and the rice, *Oryza sativa* (62,877 proteins; release 5, dated December 2006), were downloaded from The Institute for Genomic Research (TIGR). The two hundred eighty single-domain cyclophilins and the 36 TPR multiple-domain cyclophilins were used to train the methods for the mining of the genomes (Table 1). These training datasets are available at: http://bioinfolab.unl.edu/emlab/cyclophilin/.

2.2 Experimental Design

The following computational experiments were designed to identify the advantage and disadvantage of each classifier for detecting various types of similarities for cyclophilin proteins.

Within-family classification—In this experiment, classifiers were trained and tested using the datasets generated from the same cyclophilin group (e.g., single-domain training and single-domain test datasets as shown in Table 1). For single-domain cyclophilins, training and testing were done using two independent datasets. For TPR multiple-domain cyclophilins, the leave-one-out cross-validation analysis was performed using the single “TPR multiple-domain training” dataset.

Between-family classification—Classifiers were trained on a dataset generated from one group of cyclophilins (single-domain or multiple-domain) and tested against a dataset generated from another group of cyclophilins (multiple-domain or single-domain) as shown in Table 1. This is to evaluate how classifiers are sensitive to identify cyclophilin-related sequences even when they are trained on distantly related sequences belonging to other cyclophilin families. Sensitive classifiers should be able to identify new cyclophilins even if they were not directly trained on those sequences.

2.3 Selection of Descriptors

Physico-chemical properties of amino acids—Opiyo and Moriyama (2007) developed five descriptors (PC1 - PC5) using principal component analysis (PCA) from 12 physico-chemical 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-strand, and reverse turn). We used the same five descriptors for this study.

Auto/cross covariance (ACC) transformation—A set of amino acid sequences needs to be transformed to a uniform matrix before partial least squares can be applied. Auto/cross covariance (ACC) transformation method discussed in Opiyo and Moriyama (2007) was used to transform each amino acid sequence using the five descriptor set. Briefly, the ACC describes the average correlations between residues a certain lag apart. After each amino acid sequence was transformed to a set of five PC scores, auto-covariances (AC) and cross-covariances (CC) for each sequence were calculated as follows. The auto-covariance of PC1 at the amino acid position i with the lag size 1, $AC_{1,i}(1)$, is calculated with $PC1_i$ multiplied by $PC1_{i+1,}$ where $PC1_i$ is the PC1 value of the i-th amino acid. The auto-covariance of PC1 for a sequence with the lag size 1, $AC_1(1)$, is the average of these products from the position 1 to the position L-1 (L is the length of the sequence). The cross-covariance of PC1 and PC2 at the amino acid position i with the lag size 1, $CC_{12,i}(1)$, is calculated by multiplying $PC1_i$ with $PC2_{i+1}$. The cross-covariance of PC1 and PC2 for a sequence with the lag size 1, $CC_{12}(1)$, is the average of these products from the position 1 to the position L-1. The following equations summarize these calculations:

$$AC_j(d) = \frac{1}{L}\sum_{i=1}^{L-d}\left(PCj_i - \bar{PCj}\right)\left(PCj_{i+d} - \bar{PCj}\right) \quad (1)$$

$$CC_{jk}(d) = \frac{1}{L}\sum_{i=1}^{L-d}\left(PCj_i - \bar{PCj}\right)\left(PCk_{i+d} - \bar{PCk}\right) \quad (2)$$

where d is the lag size, PCj_i and PCk_i are the j and k-th PC value for the i-th amino acid, $\bar{PC}\,j$ and $\bar{PC}\,k$ are the means of PCj_i and PCk_i, respectively ($j \neq k$; j, k = 1, 2, 3, 4, or 5), and L is the length. While the auto-covariances emphasize the interactions between amino acids, interactions between different amino acid properties are incorporated into the cross-covariances. ACC transformation with the maximum lag of 30 residues yielded 775 descriptors for each sequence. The calculation of ACC was performed using the R implementation (version 2.60; http://www.R-project.org).

T-test—The t-test is the most commonly used method to evaluate the differences in means between two groups (e.g., cyclophilin proteins and non-cyclophilin proteins). The equation for the statistics is a ratio as shown in the equation (3). The top part of the ratio is simply the difference between the two means. The bottom part is a measure of the variability or dispersion of the groups.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{VarX_1}{n_1} + \frac{VarX_2}{n_2}}} \quad (3)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of the descriptor values (X), $VarX_1$ and $VarX_2$ are their variances, and n_1 and n_2 are the number of samples from the groups 1 and 2, respectively. In this study, the significance level (α) of 0.01 was used to examine if each descriptor can discriminate two groups of sequences (cyclophilins vs. non-cyclophilins) significantly based on the t statistics. The t-test was performed using the implementation in R (version 2.60; http://www.R-project.org).

Non-parametric Wilcoxon rank-sum test—The t-test is a parametric test assuming normal distributions of the variables. However, there is no guarantee that the descriptors we use to classify protein sequences have a normal distribution, and most likely they do not. Therefore, we also used one of the non-parametric tests, Wilcoxon rank-sum test. Wilcoxon rank-sum test involves in calculating a statistics called *U*. In this test, each descriptor is ranked first by ignoring the group membership. Then the rank values, instead of the descriptor values, are added up for each group. Finally the *U* statistics is calculated by the equation (4):

$$U = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1 \tag{4}$$

where n_1 and n_2 are the number of samples in groups 1 and 2, respectively, and R_1 is the sum of the rank in the group 1. In this study, the significant level (α) of 0.01 was used to examine if the *U* statistics shows significant difference in the distribution of each descriptor values between cyclophilins and non-cyclophilins. R implementation (version 2.60; http://www.R-project.org) of Wilcoxon rank-sum test was also used for this study.

Selecting of the ACC descriptors—In Opiyo and Moriyama (2007), we observed that the PLS classifier using descriptors transformed by ACC had high false positive rates. We hypothesized that the number of false positives by PLS classifiers can be reduced if we select only descriptors that are important in discriminating cyclophilins from non-cyclophilins. As mentioned above, after the ACC transformation, each sequence was represented by 775 descriptors. We used the t-test and Wilcoxon rank-sum test to choose descriptors that showed significant difference between cyclophilins and non-cyclophilins included in training datasets at the alpha level of 0.01. From the 775 descriptors, 690 and 702 descriptors were selected for the single-domain cyclophilins by the t-test and by the rank-sum test, respectively. For the TPR multiple-domain cyclophilins, 647 and 665 descriptors were selected by the t-test and the rank-sum test, respectively. These reduced numbers of descriptors, as well as all of the 775 descriptors, were used with partial least square methods as described in the next section.

2.4 Classifiers

Partial least squares—Partial least squares (PLS; Geladi and Kowalski, 1986) is a projection method similar to PCA where the independent variables, represented as the matrix **X**, are projected onto a low dimensional space. In PLS, the goal is not only to explain the maximum amount of variance observed in independent variables **X**, but also to explain the correlation with dependent variables **Y**. PLS using descriptors transformed by ACC (PLS-ACC) was discussed in Opiyo and Moriyama (2007). In this study, we included PLS with descriptors selected by the t-test (PLS-T_ACC) and PLS with descriptors selected by the rank-sum test (PLS-R_ACC). For the single-domain cyclophilin classification, the cut-off points for PLS-ACC, PLS-T_ACC, and PLS-R_ACC were 0.446, 0.470, and 0.467, respectively, based on the minimum error point (MEP; Karchin et al., 2002). Similarly, the cut-off points for the TPR multiple-domain cyclophilin classification were 0.452, 0.477, and 0.482 for PLS-ACC, PLS-T_ACC, and PLS-R_ACC, respectively. We used an R implementation (version 2.60) with the PLS package (version 1.2-1) developed by Wehrens and Mevik (http://www.R-project.org; http://mevik.net/work/software/pls.html)

PSI-BLAST—In a general use of PSI-BLAST (Altschul et al., 1997), position-specific scoring matrices (PSSMs) are built from multiple alignments of significantly similar sequences obtained by similarity search. In this study, we used pre-aligned positive (cyclophilin) sequences as the first input. Multiple alignments were generated using Clustal-W version 1.83 (Thompson et al., 1994) with the default parameters. Ten iterations with E-value = 10 as the threshold for building PSSM were performed against the test dataset. Cut-off E-values of 2.3

and 2.6 were obtained for single-domain cyclophilins and TPR multiple-domain cyclophilins, respectively, using MEP.

Profile hidden Markov model—Profile hidden Markov models (HMMs) are the full probabilistic representation of sequence profiles (Durbin et al., 1998). Profile HMMs are built using only positive samples. In this study, profile HMMs were built using the w0.5 script of the Sequence Alignment and Modeling Software System (SAM; Hughey and Krogh, 1996). Cut-off E-values of 1.02 and 1.23 were obtained for single-domain cyclophilins and TPR multiple-domain cyclophilins, respectively, using MEP.

2.5 Performance analysis

Predictions were grouped as follows:

- True positives (TP): the number of actual cyclophilins predicted as cyclophilins.
- False positives (FP): the number of actual non-cyclophilins predicted as cyclophilins.
- True negatives (TN): the number of actual non-cyclophilins predicted as non-cyclophilins.
- False negatives (FN): the number of actual cyclophilins predicted as non-cyclophilins.

Performance statistics were calculated as follows

- Accuracy = (TP + TN)/(TP + TN + FP + FN)
- False positive rate = FP/(FP + TN)
- False negative rate = FN/(FN + TP)
- True positive rate = TP/(TP + FN)

3 Results and Discussion

Within-family classification

Classifiers were trained and tested using the datasets generated from the same family (single or multi-domain). This is to evaluate how well a method trained on a family can identify sequences from the same family. As shown in upper half of Table 2, both of PLS-T_ACC and PLS-R_ACC showed higher accuracy rates than others including the original PLS-ACC, although the difference was small. The false positive rates were also lower with PLS-T_ACC and PLS-R_ACC compared to PLS-ACC. While SAM and PSI-BLAST had lower false positive rates than PLS classifiers, these classifiers showed extremely high false negative rates. High false negative rates mean that SAM and PSI-BLAST often misidentify positives (cyclophilins) as negatives, even though they rarely misidentify non-cyclophilins as positives. Similar results were obtained from cross-validation tests for the TPR multi-domain dataset as shown in the lower half of Table 2.

Between-family classification

In this experiment, classifiers were trained with sequences from one family and tests were done on another family. As mentioned before, this is to evaluate how classifiers are sensitive to identify cyclophilin-related sequences even when they are trained for distantly related sequences belonging to other cyclophilin families. Sensitive classifiers should be able to identify new cyclophilins even if they were not directly trained on those sequences. The results obtained for the between-family analyses were consistent to those we observed for the within-family analyses with more pronounced difference in performance (Table 3). PLS-T_ACC and PLS-R_ACC showed the highest accuracy rates and lower false positive rates compared to PLS-ACC. Similar results were obtained whether the classifiers were trained with single-

domain cyclophilins and tested on TPR multiple-domain cyclophilins or *vice versa*. SAM showed the lowest false positive rates, and again, both of SAM and PSI-BLAST showed very high false negative rates

Selection of significant and reduced numbers of descriptors appears to have contributed to lowering the numbers of false positives. On the other hand, it did not affect the sensitivity of PLS classifiers as shown in low or even lower than PLS-ACC % false negative in classifying cyclophilins. PLS-T_ACC and PLS-R_ACC can identify both single-domain and multiple-domain cyclophilins regardless of which cyclophilin sequences are included in the training dataset. Such classifiers are expected to be useful for identifying new/unknown cyclophilins. SAM and PSI-BLAST performed poorly because they require alignable sequences to build their models and to identify new sequences. In *Arabidopsis*, for example, the similarities between cyclophilin sequences range from 10 to 90%. Such varied and low sequence similarities made currently often used profile methods (SAM and PSI-BLAST) misidentify some cyclophilins.

Arabidopsis and rice genome mining

Table 4 and Figure 1 summarize the results of cyclophilin mining from the *Arabidopsis thaliana* and *Oryza sativa* (rice) genomes. Currently only 21 and 8 experimentally confirmed sequences are annotated as single-domain and multiple-domain cyclophilins in the *A. thaliana* genome, respectively. Not much is known on cyclophilins from the rice genome. Two separate predictions were performed for each of the *A. thaliana* and rice genomes. The first prediction was performed using classifiers trained with the single-domain dataset, and the second prediction was performed using those trained with the TPR multiple-domain dataset (Table 1). The final prediction results were obtained by merging the results from the two predictions. The predicted proteins by PLS-T_ACC and their scores, and the proteins identified by PSI-BLAST and SAM and their E-values are listed in the Supplementary Tables 1 (from single-domain trained) and 2 (from TPR-multiple-domain trained) available from http://bioinfolab.unl.edu/emlab/cyclophilin/.

PLS-T_ACC identified 302 proteins (after excluding alternative transcripts) from the *A. thaliana* genome (Table 4). PLS-T_ACC missed one known *Arabidopsis* single-domain protein out of 29 when it was trained using the single-domain dataset. All the twenty nine known *Arabidopsis* cyclophilin proteins were correctly identified when the PLS-T_ACC was trained using the TPR multiple-domain cyclophilin dataset (Table 5). Of these 302 proteins, forty six are multiple-domain cyclophilins including six TPR multiple-domain proteins. Other proteins include domains such as nucleotide-binding, WD40 repeat, RNA recognition, zinc finger, and U-box domains. Of the 302 proteins predicted by PLS-T_ACC, 34 proteins were also predicted by both PSI-BLAST and SAM as positives (Figure 1a). These 34 proteins include five new (yet to be confirmed) cyclophilin candidates. PSI-BLAST and SAM predicted in total 39 and 126 proteins as cyclophilins, respectively. Both classifiers predicted the same 31 proteins as cyclophilins when trained with single-domain cyclophilins. They included all the known 29 cyclophilins. When trained with TPR multiple-domain training dataset, they missed eleven (by PSI-BLAST) and nine (by SAM) of known *Arabidopsis* cyclophilins (Table 5). When trained with TPR multiple-domain cyclophilins, PSI-BLAST predicted 484 sequences as positive. However, 432 of them were identified based on similarities only against TPR domain sequences (Pfam: PF01535; INTERPRO: IP002885 PPR repeats). Since PSI-BLAST trained with the single-domain dataset did not identify them as positives, these proteins are most likely false positives. These 432 proteins were excluded from Table 4 and Figure 1a. Consequently, as Figure 1a shows, all but one of 39 proteins identified by PSI-BLAST were also identified by SAM. PLS-T_ACC and SAM predicted none of these 432 proteins as positives.

In the *Arabidopsis* genome project, 30 proteins including the known 29 proteins are annotated as cyclophilins. This extra one protein (At3g25230.1) was also identified by PLS-T_ACC but missed by SAM and PSI-BLAST. The InterPro database (release 16.0) contains fifty five *Arabidopsis* cyclophilin proteins. Of the fifty five sequences, five are "putative uncharacterized" fragments. Excluding these five uncharacterized fragments as well as splicing variants, the number of cyclophilins identified in InterPro is 33 including three more cyclophilin candidates in addition to the 30 annotated (Table 4). All the 33 proteins were identified as positives by all the three classifiers. The accession numbers and the descriptions of the protein sequences predicted by the three classifiers from the *Arabidopsis thaliana* genome are presented in Supplementary Table 1 (available at: http://bioinfolab.unl.edu/emlab/cyclophilin/).

From the rice genome, PLS-T_ACC predicted 1360 sequences (excluding splicing variants) as cyclophilins (Table 4). Of them, one thousand two hundred and fifty nine proteins were predicted by the classifier trained using single-domain cyclophilins. PSI-BLAST and SAM predicted 118 and 165 proteins as cyclophilins, respectively, again much fewer than those predicted by PLS-T_ACC. Eighty six proteins were positively identified by all the three classifiers (Figure 1b). The total number of cyclophilins found in InterPro is 52 (32 single-domain and 20 multiple-domain) after excluding splicing variants (Table 4). Of these 52, 30 proteins are annotated as cyclophilins in the rice genome project. All these 30 proteins were predicted as positives by all the three classifiers. The other 22 proteins were predicted as positives by PLS-T_ACC. PSI-BLAST and SAM, however, missed the majority of them. The accession numbers and the descriptions of the protein sequences predicted from the rice genome are presented in Supplementary Table 2 (available at: http://bioinfolab.unl.edu/emlab/cyclophilin/). Figures 2 and 3 show that most of the sequences that were identified when classifiers were trained by single-domain cyclophilins were also identified when classifiers were trained using TPR multiple-domain proteins.

4 Conclusion

In this study, we found that selecting only important descriptors after auto and cross covariance transformation reduces the number of false positives. We also found that alignment-based SAM and PSI-BLAST are too conservative when used to search highly divergent proteins and those with multiple-domains such as cyclophilins. PLS-T_ACC classifier can be used to identify new cyclophilin candidates from plant genomes as they become available. We should note that these predicted proteins most likely include false positives. Experimental confirmation will be ultimately required. However, based on our test results, SAM and PSI-BLAST in general predict fewer false positives. Therefore, 34 *Arabidopsis* and 86 rice proteins positively predicted by all three classifiers are more likely to be true positives. These candidate proteins should be prioritized for further analysis. The secondary list would include 93 (for *Arabidopsis*) and 83 (for rice) proteins identified by SAM only or by at least two classifiers (Figure 1). Finally, considering the high false negative rates by SAM and PSI-BLAST, those identified by neither of these methods but identified by PLS-T_ACC should be also examined to achieve thorough cyclophilin mining.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

We thank all anonymous reviewers for the BIOT-2007 conference for their useful comments on our manuscript. This work was in part supported by Grant Number R01LM009219 from the National Library of Medicine to ENM.

Biographies

Stephen O. Opiyo was a PhD student in the Department of Agronomy and Horticulture, University of Nebraska-Lincoln. He is currently a postdoctoral fellow in the School of Biological Sciences, University of Nebraska-Lincoln.

Etsuko N. Moriyama, PhD., is an Associate Professor at the School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402. [PubMed: 9254694]

Coaker G, Falik A, Staskawicz B. `Activation of a phytopathogenetic bacteria effector Protein by eukaryotic cyclophilin ote. Science 2005;308:548–550. [PubMed: 15746386]

Deng WY, Chen LS, Wood DW, Metclaf T, Liang XY, Gordon MP, Comai L, Nester EW. *Agrobacterium* VirD2 protein interacts with plant host cyclophilins. Proc Natl Acad Sci USA 1998;75:7040–7045. [PubMed: 9618535]

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

Galat A. Peptidylproplyl cis/trans isomerases (immunophilins): Biological diversity targets-functions. Curr Topic Med Chem 2003;3:1315–11347.

Geladi P, Kowalski BR. Partial least squares regression: A tutorial. Anal. Chim. Acta 1986;185:1–17.

Hughey R, Krogh A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. Comp. Appl. Biosci 1996;12:95–107. [PubMed: 8744772]id12140654

Karchin R, Karplus K, D. Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002;18:147–159. [PubMed: 11836223]

Lamb JR, Tugendreich S, Hieter P. Tetratrico peptide repeat interactions: to TPR or not to TPR? Trends Biochem. Sci 1995;20:257–259. [PubMed: 7667876]

Mulder NJ, et al. InterPro, progress and status in 2005. Nucleic Acids Res 2005;33:D201–205. [PubMed: 15608177]

Muriel CV, Pascale VB, Nicholas NJ. A *Magnaporthe grisea* cyclophilin acts as a virulence determinant during plant infection. The Plant Cell 2002;14:917–930. [PubMed: 11971145]

Opiyo SO, Moriyama EN. Protein family classification by partial least squares. J. Proteome Res 2007;6:846–853. [PubMed: 17269741]

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2006. http://www.R-project.org

Romano PG, Horton P, J. E. Gray JE. The *Arabidopsis* cyclophilin gene family. Plant Physiol 2004;134:1268–1282. [PubMed: 15051864]

Strope PK, Moriyama EN. Simple alignment-free methods for protein classification: a case study from G-protein coupled receptors. Genomics 2007;89:602–612. [PubMed: 17336495]

Thompson JD, Higgins DG, Gibson TJ. Clustal-W Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:673–4680.

Wehrens, R.; Mevik, B. pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR). 2006. R package version 1.2-1; http://mevik.net/work/software/pls.html

Ye XY, Ng TB. A novel cyclophilin-like antifungal protein from the mung bean. Biochem. Biophys. Res. Commun 2000;273:1111–1115. [PubMed: 10891380]

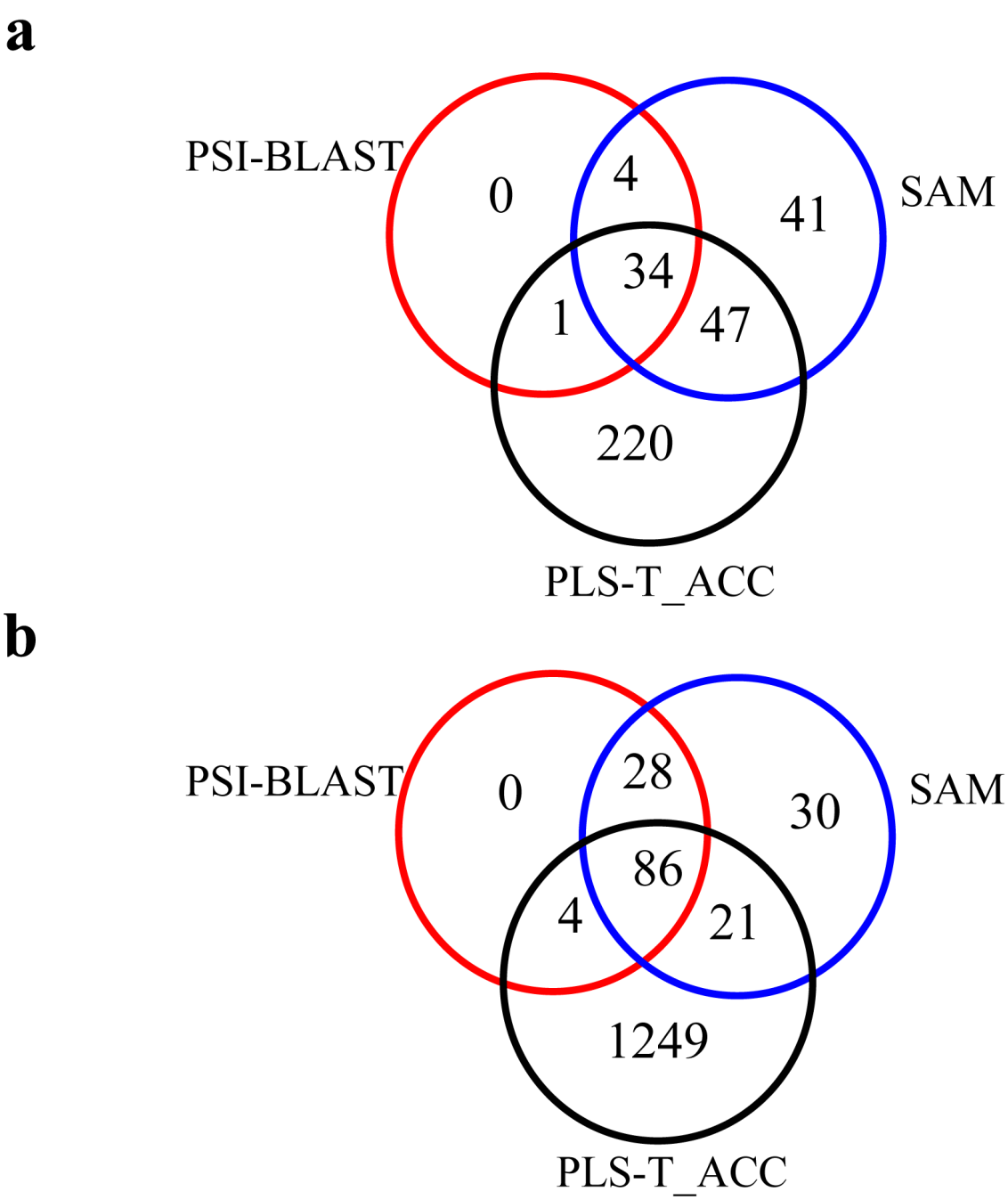

Figure 1.
The number of cyclophlin proteins predicted by the three classifiers from the *A. thaliana* (a) and rice (b) genomes.

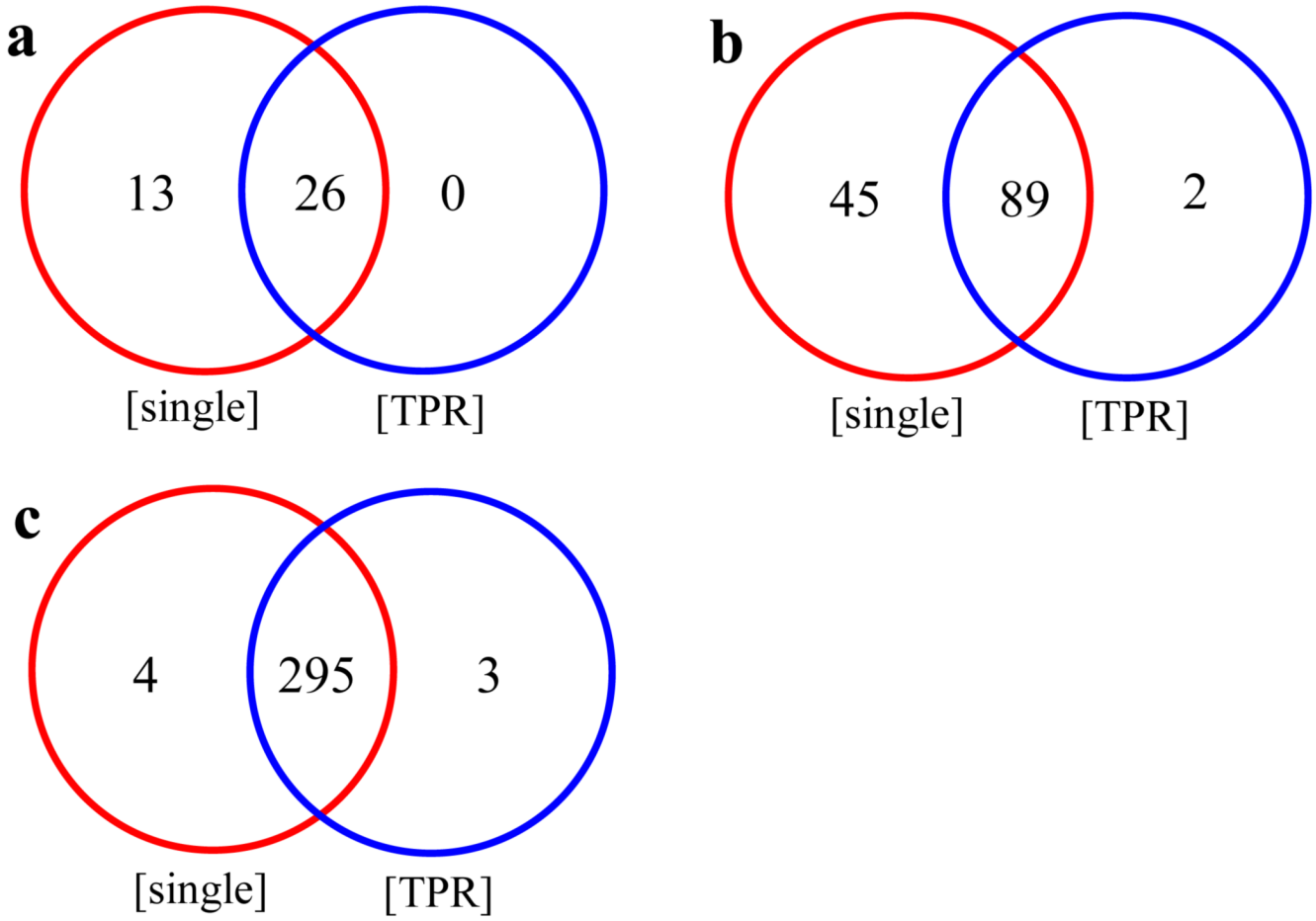

Figure 2.
The number of cyclophilin proteins identified using the classifier trained on single-domain [single] and TPR multiple-domain [TPR] cyclophlin proteins from the *A. thaliana* genome. The classifiers are PSI-BLAST (a), SAM (b), and PLS-T_ACC (c).

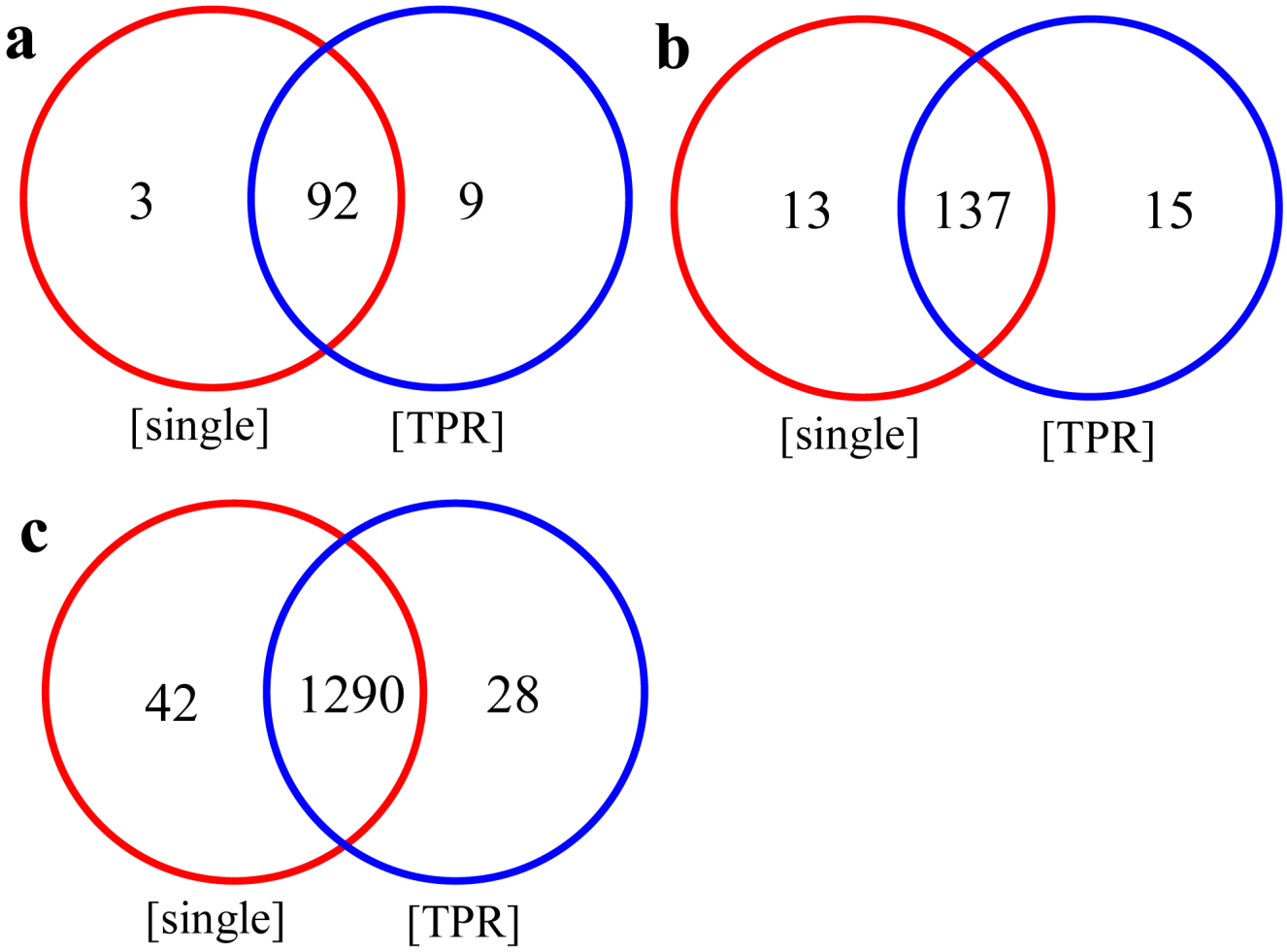

Figure 3.
The number of cyclophlin proteins identified using the classifier trained on single-domain [single] and TPR multiple-domain [TPR] cyclophilin proteins from the rice genome. The classifiers are PSI-BLAST (a), SAM (b), and PLS-T_ACC (c).

Table 1

Numbers of samples included in cyclophilin datasets

Datasets	*Cyclophilin*	*Non-cyclophilin*	*Total*
Single-domain training	140	140	280
Single-domain test	140	1000	1140
Single-domain training for mining	280	1140	1420
TPR Multiple-domain training	36	36	72
TPR multiple-domain test	36	200	236
TPR Multiple-domain training for mining	36	236	272

Table 2

Classifier performance for within-family classification

Classifiers	*%Accuracy*	*%False positive*	*%False negative*
	[Single-domain test]		
PLS-ACC	97.2	3.0	3.0
PLS-T_ACC	99.1	0.8	1.5
PLS-R_ACC	98.7	1.0	1.5
SAM	97.3	0.2	15.0
PSI-BLAST	95.8	0.3	22.0
	[TPR multiple-domain test]		
PLS-ACC	91.6	13.8	3.3
PLS-T_ACC	94.4	8.0	2.7
PLS-R_ACC	94.4	8.0	2.7
SAM	91.6	0.5	16.5
PSI-BLAST	83.3	5.0	25.0

Table 3

Classifier performance for between-family classification

Classifiers	*%Accuracy*	*%False positive*	*%False negative*
	[Single-domain test]		
PLS-ACC	90.3	10.0	7.1
PLS-T_ACC	93.4	6.5	6.7
PLS-R_ACC	92.5	7.5	7.1
SAM	92.5	3.0	39.0
PSI-BLAST	89.4	6.0	42.9
	[TPR multiple-domain test]		
PLS-ACC	92.3	6.0	16.7
PLS-T_ACC	94.4	5.0	11.1
PLS-R_ACC	93.2	6.0	11.1
SAM	90.6	2.5	33.0
PSI-BLAST	89.8	6.0	33.0

Table 4

The number of predicted cyclophilins from the *Arabidopsis* and rice genomes[a]

Genome database	*InterPro*[b]	*Genome*[c]	*Known*[d]	*PSI-BLAST*[e]	*SAM*[e]	*PLS-T_ACC*[e]
A. thaliana	50 (24/9)	30 (21/9)	29 (21/8)	41 (26/13)	134 (101/25)	321 (256/46)
O. sativa	94 (32/20)	30 (14/16)	-	151 (66/52)	215 (93/72)	1448 (1207/153)

[a]The numbers in parentheses are single/multiple-domain cyclophilins after excluding splicing variants.

[b]The numbers of cyclophilin proteins identified in InterPro (release 16, August 2007; IPR002130).

[c]The numbers of proteins annotated as cyclophilins in each genome project.

[d]The numbers of currently known experimentally confirmed cyclophilins (based on Romano et al. 2004).

[e]The numbers include all proteins predicted as cyclophilins including those based on alternative transcripts.

Int J Bioinform Res Appl. Author manuscript; available in PMC 2009 June 17.

Table 5

The twenty nine known *Arabidopsis* cyclophilins and their prediction results by PSI-BLAST, SAM, and PLS-T_ACC[a]

Locus name[b]	*PSI-BLAST [single]*	*PSI-PLAST [TPR]*	*SAM [single]*	*SAM [TPR]*	*PLS-T_AC [single]*	*PLS-T_ACC [TPR]*
At1g01940.1	**6.46E-13**	3.35E+00	**8.20E-20**	**2.34E-35**	**0.827**	**0.937**
At1g26940.1	**4.58E-08**	**7.37E-22**	**4.26E-15**	**1.10E-13**	**0.937**	**0.917**
At1g53720.1	**1.41E+00**	**2.11E-31**	**1.24E-23**	**1.69E-09**	**1.021**	**0.779**
At1g74070.1	**7.60E-08**	**1.38E-15**	**1.79E-02**	**3.58E-11**	**1.041**	**0.825**
At2g15790.1	**2.30E+00**	**4.33E-30**	**2.67E-18**	4.09E+02	**0.790**	**0.897**
At2g16600.1	**6.51E-04**	3.67E+00	**2.61E-01**	**9.10E-11**	**1.066**	**0.876**
At2g21130.1	**1.24E-26**	**7.82E-24**	**6.85E-04**	**1.23E-03**	**0.683**	**0.883**
At2g29960.1	**7.37E-22**	**4.89E-21**	**2.36E-38**	**9.33E-01**	**0.930**	**1.031**
At2g36130.1	**1.38E-15**	**8.20E-20**	**4.52E-01**	**1.16E-28**	**0.938**	**0.736**
At2g38730.1	**8.65E-08**	**4.26E-15**	**9.54E-38**	**2.36E-18**	**0.857**	**0.916**
At2g47320.1	**4.00E-06**	**1.79E-02**	**1.02E-32**	**4.37E-05**	**0.883**	**1.056**
At3g01480.1	**6.33E-05**	5.77E+00	**2.83E-22**	**2.92E-04**	**0.908**	**0.804**
At3g15520.1	**6.50E-05**	5.25E+00	**5.60E-30**	8.30E+03	**0.728**	**0.995**
At3g22920.1	**1.30E-04**	**2.36E-38**	**1.55E-29**	**2.02E-01**	**0.757**	**0.844**
At3g44600.1	**4.40E-04**	6.78E+00	**2.91E-22**	1.68E+04	**0.974**	**0.947**
At3g55920.1	**2.92E-04**	9.77E+00	**1.02E-28**	3.70E+02	**0.945**	**0.970**
At3g56070.1	**5.98E-03**	9.82E+00	**1.98E-25**	**7.60E-08**	**0.992**	**0.859**
At3g62030.1	**2.11E-31**	**9.17E-47**	**4.07E-25**	1.49E+03	**1.126**	**0.795**
At3g63400.1	**7.39E-04**	**2.92E-04**	**8.37E-21**	**1.35E-19**	**0.907**	**0.946**
At3g63400.2	**4.33E-30**	5.48E+00	**1.27E-22**	**1.83E-34**	**0.879**	**1.074**
At3g66654.1	**6.59E-24**	**6.50E-05**	**2.92E-16**	**1.89E-22**	**0.999**	**0.896**
At3g66654.2	**3.91E-23**	**1.30E-04**	**6.59E-14**	**1.19E-22**	**0.909**	**1.048**
At3g66654.3	**4.81E-23**	**2.92E-04**	**3.76E-10**	6.78E+01	**1.033**	**0.782**
At4g32420.1	**1.07E-02**	8.38E+00	**3.19E-15**	2.79E+03	**0.951**	**0.913**
At4g33060.1	**1.69E-09**	**6.59E-24**	**7.95E-31**	9.09E+02	**0.998**	**0.970**
At4g34870.1	**1.07E-19**	**2.36E-38**	**1.00E+00**	**7.47E-21**	**0.767**	**0.980**
At4g34960.1	**3.41E-18**	**2.92E-04**	**1.10E-06**	**4.40E-04**	**0.680**	**0.915**
At4g38740.1	**1.15E-17**	5.48E+00	**2.18E-29**	1.97E+02	**1.114**	**0.915**
At5g13120.1	**3.65E-16**	**2.36E-38**	**1.75E-27**	**6.51E-04**	**1.091**	**0.963**
At5g35100.1	**3.69E-16**	6.78E+00	**3.04E-25**	**1.24E-26**	**1.145**	**0.814**

Locus name[b]	*PSI-BLAST [single]*	*PSI-PLAST [TPR]*	*SAM [single]*	*SAM [TPR]*	*PLS-T_AC [single]*	*PLS-T_ACC [TPR]*
At5g58710.1	**8.85E-15**	7.25E+00	**1.59E-04**	**7.37E-22**	0.389	**1.060**
At5g67530.1	**6.62E-03**	**5.98E-03**	**2.70E-26**	**2.36E-38**	**0.946**	**0.782**

[a]Each classifier was trained using single-domain and TPR multiple-domain cylophilin training datasets. Prediction results for these two trainings are indicated as [single] and [TPR]. The E-values and scores for positive predictions are shown in bold faces.

[b]Locus names from the *Arabidopsis* Information Resource (TAIR) database (http://www.arabidopsis.org/). At3g63400 and At3g66654 include splicing variants indicated by '.2' and '.3' as the locus name suffix.