

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

Stephen O. Opiyo and Etsuko N. Moriyama*

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 more cyclophilins 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 fewer 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.

Index Terms— Cyclophilins; partial least squares; profile hidden Markov model

I. INTRODUCTION

Cyclophilins possess the peptidyl-prolyl isomerase (PPIase; EC 5.2.1.8) activity and involved in diverse cellular processes including cell cycle control,

receptor signaling, protein folding, as well as cellular targets of immunosuppressant drugs [1]. In the presence of their drug ligand, cyclosporine A, cyclophilins gain their immunosuppressing function by forming a complex with cyclosporine A. This complex blocks T-cell activation by binding to 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 [2]. *Agrobacterium* recruits plant cyclophilins for transferring and integrating T-DNA (transferred DNA) into a plant cell. AtCYP18 has been identified to activate *Pseudomonas syringae* effector protein (AvrRpt2) by its PPIase activity [3]. 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 [4]. On the other hand, cyclophilins have been purified from seeds of cow pea, mung bean, and chickpea [5]. They possess antifungal activity against several fungi including *Mycosphaerella arachidichola*. The chickpea protein is also known to inhibit human immunodeficiency virus-1 reverse transcriptase [5].

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 “tetratricopeptide (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 [6]. 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

This work was in part supported by Grant Number R01LM009219 from the National Library of Medicine to ENM.

S. O. Opiyo was with the Department of Agronomy and Horticulture, University of Nebraska-Lincoln, 68583. He is now with the School of Biological Sciences, University of Nebraska-Lincoln, 68588; (e-mail:sopiyo@unlserve.unl.edu).

E. N. Moriyama is with the School of Biological Sciences and Plant Science Initiative, University of Nebraska-Lincoln, N107 Beadle Center, Lincoln, NE 68588-0660; phone: 402-472-4979; fax: 402-472-3139; (e-mail: emoriyama2@unl.edu).

*Corresponding author.

phosphatases [6].

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) [1]. 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, there are 302 protein sequences from plants, 595 from animals, 321 from fungi, and 1319 from bacteria. This clearly shows 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; [7]), Position Specific Iterative-BLAST (PSI-BLAST; [7]), and profile hidden Markov models (profile HMMs; [8]). 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 [9, 10]. 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.

II. MATERIALS AND METHODS

A. Datasets

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; [11]), and divided to prepare training and test datasets (Table I). Although 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 protein regions including both of cyclophilin and TPR domains of TPR multiple-domain cyclophilins 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 I).

TABLE I
NUMBERS OF SAMPLES INCLUDED IN CYCLOPHILIN DATASETS

Datasets	Cyclophilin	Non-cyclophilin	Total
Single-domain training	140	140	280
Single-domain test	140	1000	1400
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

B. 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 analysis

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 I). 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 analysis

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 I. 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.

C. Descriptors

Physico-chemical properties of amino acids

Opiyo and Moriyama [9] developed five descriptors (PC1 - PC5) using the 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-sheet, 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 [9] was used to transform each amino acid sequence using the five descriptor set. ACC with the maximum lag of 30 residues yielded 775 descriptors for each sequence.

Selection of important descriptors

In Opiyo and Moriyama [9], we observed that the PLS classifier using descriptors transformed by ACC had high false positive rates. The hypothesis is 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 is represented by 775 descriptors. We used the t-test and a non-parametric rank test to choose descriptors that showed significant difference between cyclophilins and non-cyclophilins 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 test, respectively. For the TPR multiple-domain cyclophilins, 647 and 665 descriptors were selected by the t-test and the rank test, respectively.

D. Classifiers

Partial least squares

Partial least squares (PLS; [12]) is a projection method similar to PCA where the independent variables, represented as the matrix \mathbf{X} , are projected onto a low dimensional space. PLS uses both independent variables \mathbf{X} and dependent variables \mathbf{Y} . PLS using descriptors transformed by ACC (PLS-ACC) was discussed in Opiyo and Moriyama [9]. In this study, we included PLS with descriptors selected by the t-test (PLS_T-ACC) and PLS with descriptors selected by the rank 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; [13]). 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.

PSI-BLAST

In a general use of PSI-BLAST [7], 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 ClustalX version 1.83 [14] 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 HMMs are the full probabilistic representation of sequence profiles [8]. The 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; [15]). Cut-off E-values of 1.02 and 1.23 were obtained for single-domain cyclophilins and TPR multiple-domain cyclophilins, respectively, using MEP.

E. Performance analysis

Predictions are grouped as follows:

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

Performance statistics are 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 = $TP/(TP + FN)$.

III. 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 the upper half of Table II, both of PLS-T_ACC and PLS-R_ACC showed higher accuracy rates than others including 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. Similar results were obtained from cross-validation tests for the TPR multi-domain dataset as shown in the lower half of Table II.

TABLE II
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 cross-validation 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

Between-family classification

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 III). 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*. Again, SAM and PSI-BLAST showed very high false negative rates.

TABLE III
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

Selection of significant and reduced numbers of descriptors appeared 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 IV summarizes the results of cyclophilin mining from the *Arabidopsis thaliana* and rice genomes. Currently only 21 and 8 *A. thaliana* sequences are experimentally confirmed as single-domain and multiple-domain cyclophilins. Two separate predictions were performed for each genome using classifiers trained with a single-domain dataset and those trained with a TPR multiple-domain dataset (Table I). The final prediction results were obtained by merging these results from the two predictions.

From the *A. thaliana* genome, PLS-T_ACC identified 302 proteins (after excluding alternative transcripts).

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 PLS-T_ACC was trained using the TPR multiple-domain cyclophilin dataset. Of these 302 proteins, forty six are multiple-domain cyclophilins including six TPR multiple-domain proteins. Others include domains such as nucleotide-binding, WD40 repeat, RNA recognition, zinc finger, and U-box domains, in addition to cyclophilin domains. Of the 302 proteins predicted by PLS-T_ACC, 34 proteins were also predicted by both of PSI-BLAST and SAM as positives. 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. When trained with TPR multiple-domain cyclophilins, PSI-BLAST predicted 473 sequences as positives. However, 432 of them were identified based on similarities only against TPR domain sequences (Pfam: PF01535; INTERPRO: IPO02885 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 IV. 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 IV). 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 I (available at: <http://bioinfolab.unl.edu/emlab/cyclophilin/>).

From the rice genome, PLS-T-ACC predicted 1360 sequences (excluding splicing variants) as cyclophilins. 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. The total number of cyclophilins found in InterPro is 52 (32 single-domain and 20 multiple-domain) after excluding splicing variants (Table IV). 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 II (available at: <http://bioinfolab.unl.edu/emlab/cyclophilin/>).

Finally, we should note that these predicted proteins include false positives. Experimental confirmation will be ultimately required. However, based on our test results,

TABLE IV
THE NUMBER OF PREDICTED CYCLOPHILINS FROM THE *ARABIDOPSIS THALIANA* AND RICE GENOMES

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

^aThe numbers of cyclophilin proteins identified in InterPro (release 16, August 2007; IPR002130). The numbers in parenthesis are single/multiple-domain cyclophilins after excluding splicing variants.

^bThe numbers of proteins annotated as cyclophilins in each genome project. The numbers in parenthesis are single/multiple-domain cyclophilins after excluding splicing variants.

^cThe numbers of currently known experimentally confirmed cyclophilins. The numbers in parenthesis are single/multiple-domain cyclophilins.

^dThe numbers include all proteins predicted as cyclophilins including those based on alternative transcripts. Numbers excluding such alternative transcripts are shown in square brackets (single/multiple-domain cyclophilins).

SAM and PSI-BLAST in general predict fewer false positives. Therefore, 5 new *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. Other candidates predicted only by PLS classifier need to be investigated in the next step.

REFERENCES

- [1] P. G. Romano, P. Horton, and J. E. Gray "The *Arabidopsis* cyclophilin gene family", *Plant Physiol.* vol. 134, pp. 1268-1282, 2004b.
- [2] W. Y. Deng, L. S. Chen, D. W. Wood, T. Metclaf, X. Y. Liang, M. P. Gordon, L. Comai, and E. W. Nester. "Agrobacterium VirD2 protein interacts with plant host cyclophilins". *Proc. Natl. Acad. Sci. USA.* vol. 75, pp. 7040-7045, 1998.
- [3] G. Coaker, A. Falick, and B. Staskawicz. "Activation of a phytopathogenetic bacteria effector Protein by eukaryotic cyclophilin". *Science*, vol. 308, pp. 548-550, 2005.
- [4] C. V. Muriel, V. B. Pascale, and N. J. Nicholas. "A *Magnaporthe grisea* cyclophilin acts as a virulence determinant during plant infection". *The Plant Cell* vol. 14, pp. 917-930, 2002.
- [5] X. Y. Ye and T. B. Ng "A novel cyclophilin-like antifungal protein from the mung bean". *Biochem. Biophys. Res. Commun.* vol. 273, pp 1111-1115, 2000.
- [6] J. R. Lamb, S. Tugendreich, and P. Hieter "Tetratricopeptide repeat interactions: to TPR or not to TPR?" *Trends Biochem. Sci.* vol. 20, pp 257-259, 1995.
- [7] S. F. Altschul, T. L Madden, A. A Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Res.* vol. 25, pp. 3389-3402, 1997.
- [8] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge University Press, Cambridge, 1998.
- [9] S. O. Opiyo and E. N. Moriyama "Protein family classification by partial least squares", *J. Proteome Res.* vol. 6, pp. 846-853, 2007.
- [10] P. K. Strop and E. N. Moriyama. "Simple alignment-free methods for protein classification: a case study from G-protein coupled receptors" *Genomics* vol. 89, pp 602-612, 2007.
- [11] N. J. Mulder, et al. "InterPro, progress and status in 2005." *Nucleic Acids Res.* vol. 33, pp, D201-205, 2005.
- [12] P. Geladi and B. R. Kowalski. "Partial least squares regression: A tutorial". *Anal. Chim. Acta.* vol. 185, pp, 1-17, 1986.
- [13] R. Karchin, K. Karplus, and D. Haussler. "Classifying G-protein coupled receptors with support vector machines". *Bioinformatics* vol. 18, pp,147-159, 2002.
- [14] J. D. Thompson, D. G. Higgins, T.J. Gibson. "Clustal-W Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice". *Nucleic Acids Res.* vol. 22 pp, 673-680, 1994.
- [15] R. Hughey and A. Krogh. "Hidden Markov models for sequence analysis: Extension and analysis of the basic method". *Compu. Appl. Biosci.* vol.12, pp 95-107, 1996.