# Mining Cytochrome b561 Proteins from Plant Genomes

**S. O. Opiyo** and
School of Biological Sciences, University of Nebraska-Lincoln, NE 68588

**E. N. Moriyama**
School of Biological Sciences and Center for Plant Science Innovation, University of Nebraska-Lincoln, 403 Manter Hall, Lincoln, NE 68588-0118, USA

S. O. Opiyo: sopiyo@unlserve.unl.edu; E. N. Moriyama: emoriyama2@unl.edu

## Abstract

Cytochrome b561 (Cyt-b561) proteins are important for plant growth, development, and prevention of damage to plants from excess light under drought condition. Because of their low sequence similarity, thorough mining of Cyt-b561 proteins from plants are not easy. Currently there is only one Cyt-b561 gene in the maize, none in the soybean and twenty two in the *Arabidopsis thaliana* genomes, respectively. We used alignment-free protein classifiers based on partial least squares and support vector machines to identify Cyt-b561 from plant genomes. These classifiers performed better than profile hidden Markov models and PSI-BLAST in identifying Cyt-b561 related proteins from plant genomes.

## Keywords

## 1 Introduction

Cytochrome b561 (Cyt-b561) proteins are important in plants because they transfer electrons from a soluble cytoplasmic donor (ascorbate) across a membrane bilayer to a soluble intravesicular acceptor (semidehydroascorbate). In plants, ascorbate plays a role in antioxidative defense reactions (Asard et al., 2001) [3], and the regeneration of ascorbate can be made possible by the transportation of electrons by Cyt-b561. Verelst and Asard (2001) [26] also reported that dopamine-β-hydroxylase Cyt-b561 is involved in catecholamine action in plants, which is important for plant growth and development. Furthermore, Cyt-b561 proteins prevent plants from damage due to excess light under drought condition (Nanasato et al., 2005) [19].

Cyt-b561 is a protein family with four well-conserved histidine (His) residues and six transmembrane helices. The four conserved His residues possibly coordinate two heme molecules. Furthermore, substrate-binding sites for ascorbate and monodehydro-ascorbate (MDHA) have been predicted (Asard et al., 2001) [3]. The average length of Cyt-b561 proteins is 237 amino acids (aa). In addition to these single-domain Cyt-b561s (those without any additional functional domains), some multiple-domain proteins have sequence regions similar to Cyt-b561 (Cyt-b561 domains) linked to other domains such as dopamine-hydroxylase (DOH; Asard et al., 2001) [3], cellulose dehydrogenase, carbohydrate binding protein, protein

Correspondence to: E. N. Moriyama, emoriyama2@unl.edu.

of unknown function (DUF569), and cobalamin domains. These different domains are considered to add functions by cooperatively working with the Cyt-b561 domain. The length of multiple-domain Cyt-b561 proteins varies and their average length is 511 aa. The sequence identities among these Cyt-b561 protein/domain sequences are as low as 30% (Verelst and Asard, 2003) [27] reflecting their functional diversity.

Because of their low sequence similarity, these protein sequences are not easy to be aligned, hence making them more difficult to be identified from diverse plant genomes. In the current release of InterPro database (17.0; Mulder et al., 2005) [18], 193 Cyt-b561 sequences (including both single and multiple-domain proteins) are identified from 22 plant species. The majority of plants are represented by very small numbers of Cyt-b561 proteins: e.g., one single-domain Cyt-b561 for maize (*Zea mays*) and none for soybean (*Glycine max*). On the other hand, twenty two (seven single-domain and 15 multiple-domain) Cyt-b561 protein sequences are known from the *A. thaliana* genome and twenty three (eight single-domain and 15 multiple-domain) are identified from the grapevine (*Vitis vinifer*a) genome. While there is a possibility that those plants have fewer Cyt-b561 related proteins compared to *Arabidopsis* and grapevine, it is more likely that currently often used methods are not capable of efficiently identifying these proteins from newly sequenced plant genomes because of low sequence similarities among these protein sequences.

The majority of currently used methods for protein classification require building alignments. Using multivariate analysis methods on various amino acid properties avoids this alignment-building process. Such alignment-free methods can be sensitive for remotely similar protein identification such as for Cyt-b561 related proteins where reliable alignments are difficult. Another advantage of using multivariate analysis methods is that they use both positive (protein of interest) and negative (unrelated protein) samples in their training. Alignment-based methods like PSI-BLAST and profile hidden Markov models (profile HMMs; used in, e.g., Pfam database; Bateman et al., 2004) [4], on the other hand, are trained using only positive samples since unrelated proteins cannot be included in the model alignments.

Several studies have previously compared the performance of profile HMMs with alignment free methods for protein classifications. Karchin et al., (2002) [12] used support vector machines (SVMs) with the Fisher kernel for classifying G-protein coupled receptor (GPCR) subfamilies and reported better performance than profile HMMs. Liao and Noble (2003) [16] used SVMs with descriptors based on pairwise similarity scores, which also performed better than profile HMMs for discriminating SCOP protein families (Andreeva et al., 2004) [2]. Better performance than profile HMMs was further shown by Kim et al. (2000) [14] and Moriyama and Kim (2005) [17] with GPCR classifiers based on parametric and non-parametric discriminant functions. Various alignment-free descriptors and kernels have been successfully used with superior performance especially for identifying remotely similar protein families. Examples include: mismatch string kernel used by Leslie et al., (2003) [15]; dipeptide composition used with SVMs by Bhasin and Raghava, (2005) [5]; n-gram frequencies used with decision tree and nave Bayes classifiers by Cheng et al. (2005) [7]; and self-organizing maps used by Otaki et al., (2006) [22].

In Opiyo and Moriyama (2007) [20], we tested partial least squares regression (PLS) methods using phyisco-chemical properties of amino acids for identifying Cyt-b561 proteins from short expressed sequence tags (ESTs) derived from the *A. thaliana* genome. The PLS methods successfully identified Cyt-b561 EST sequences that profile HMMs and PSI-BLAST could not identify. In our more recent study (Opiyo and Moriyama, 2009) [21], we used PLS classifiers for identification of single-domain and multiple-domain cyclophilins from the *Arabidopsis* and rice genomes. PLS classifiers again performed better than profile HMMs and PSI-BLAST. In this study, our objective is to further examine PLS multivariate classifiers for

identifying Cyt-b561 related proteins and to mine these proteins from diverse plant genomes (*A. thaliana*, grapevine, soybean, and maize).

## 2 Materials and Methods

### 2.1 Dataset

Three hundred Cyt-b561 sequences (140 from plants, 130 from animals, 30 from fungi) were downloaded from InterPro (Release 17.0; Mulder et al., 2005 [18]). Table 1 shows the number of sequences from plants, animals, and fungi included in our Cyt-b561 datasets. Two hundred sequences were single-domain Cyt-b561 (70 from plants, 110 from animals, and 20 from fungi), and one hundred sequences were multiple-domain Cyt-b561 (70 from plants, 20 from animals, and 10 from fungi). Plant sequence data included all Cyt-b561 proteins from the *A. thaliana* and grapevine genomes. Other plant species included maize, rice, watermelon, banana, sugar beet, etc.

In order to search sequences similar to Cyt-b561 both from single and multiple domain proteins, we used all 300 Cyt-b561 protein sequences in the positive dataset including both of single-domain Cyt-b561 proteins as well as Cyt-b561-domain regions from multiple-domain proteins. Three hundred Non-Cyt-b561 proteins longer than 100 amino acids were randomly sampled from Swiss-prot database (Boeckmann et al., 2003) [6] and used as the negative samples.

Cyt-b561 related proteins were mined from four plant genomes. Their entire protein sequences were downloaded from the following sources:

- *Arabidopsis thaliana:* 32756 proteins from the release 8 (April, 2008) of The Arabidopsis Information Resource (TAIR) database (ftp://ftp.arabidopsis.org/home/tair/Sequences/datasets)

- *Vitis vinifera*: 30,434 proteins from the release 1 (October, 2007) of the Genoscope database (http://www.genoscope.cns.fr/externe/Download/Projets/).

- *Glycine max*: 62,877 proteins from the release 3 (September, 2007) of the Phytozome database (ftp://ftp.jgipsf.org/pub/JGIdata/Glycinemax/).

- *Zea mays*: 137,000 proteins from the release October, 2007 of the MaizeSequence database (http://ftp.maizesequence.org/20071003/).

### 2.2 Training classifiers

For alignment-based classifiers (profile HMMs and PSI-BLAST), only positive (Cyt-b561) samples were included in the training datasets. For alignment-free classifiers, both of positive and negative samples were used for training. For mining the plant genomes, the entire 600 sequences were used for training classifiers. As mentioned above, only 300 positive sequences were used for training alignment-based classifiers.

### 2.3 Sequence descriptors used for alignment-free classifiers

**Amino acid composition—**From each protein sequence, frequencies of 20 amino acids were calculated. Strope and Moriyama (2007) [24] applied SVMs with amino acid composition for GPCR classification problems, and showed that such classifiers outperformed profile HMMs and decision trees classifiers for discriminating GPCRs from non-GPCRs. In this study, amino acid composition was used as descriptors for an SVM classifier (SVM-AA).

**Dipeptide composition—**Dipeptide composition represents all 400 frequencies of consecutive amino acid pairs in a protein sequence and corresponds to a 400 ($20 \times 20$) feature

vector. It can encapsulate information on composition of amino acids as well as their local order. We used dipeptide composition as descriptors for an SVM classifier (SVM-DIP).

**Physico-chemical properties of amino acids—**In Opiyo and Moriyama (2007) [20], we developed five descriptors (PC1-PC5) using the principal component analysis (PCA) of 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). The five principal components (PCs) selected explained 93.2% of the total variance. The first principal component (PC1) covered 40.4% of the total variance of the original 12 physico-chemical properties. It represented all properties except isoelectric point and non-polar surface. PC2 (28% of the total variance) had negative relationships with the hydrophobicity properties. PC3 and PC5 had 15% of the combined total variance, and represented secondary structure properties. PC4 (8.9% of the total variance) represented isoelectric point and volume. We used the same five descriptors in (Opiyo and Moriyama, 2009) [21] for classifying cyclophilin proteins. The same descriptors were used in this study with PLS and SVM classifiers.

**Auto/cross covariance (ACC) transformation—**Auto/cross covariance (ACC) transformation method discussed in (Opiyo and Moriyama, 2007) [20] 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 package (version 2.60; R Development Core Team, 2007 [23]).

**Selection of important descriptors—**In Opiyo and Moriyama (2007) [20], we reported that the PLS classifier using descriptors transformed by ACC had high false positive rates. Our subsequent study showed that reducing the number of descriptors (690 for single-domain and 647 for multiple-domain cyclophilins) by selecting significant descriptors by the t-test decreased the number of false positives (Opiyo and Moriyama, 2009) [21]. In this study, similarly, the t-test was used to select descriptors that showed significant difference between Cytb-561 and Non-Cyt-b561 in training datasets at the alpha level of 0.01. From the 775 ACC descriptors, 459 were selected. These descriptors are available in Supplementary Table 1 at: http://bioinfolab.unl.edu/emlab/Cyt-b561/

### 2.4 Classifiers

**Partial least squares—**Partial least squares (PLS; Geladi and Kowalski, 1986) [10] is a projection method similar to 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 [20]) for GPCR classification and in (Opiyo and Moriyama, 2009) [21] for cyclophilin classification. As in (Opiyo and Moriyama, 2009) [21], we also included PLS with ACC descriptors selected by the t-test (PLS-T_ ACC). The cut-off point for PLS-T_ACC classification was chosen as 0.562 based on the minimum error point (MEP; Karchin et al., 2002 [12]). PLS analysis was performed using an R implementation, the PLS package developed by Wehrens and Mevik version 1.2.1 (Wehrens and Mevik, 2006 [28]).

**Support vector machines—**Support vector machines (SVMs; Vapnik, 1999 [25]) learn to separate a set of labeled training data by remapping them in a high-dimensional space and by discovering a hyperplane that separates the two classes in this space. The hyperplane is optimized in such a way that the distance, called margin, between the hyperplane and the closest training example is maximized. Support vectors are those data points that define the margin.

Once the hyperplane is found, predicting the label of a new, unlabeled data point involves determining on which side of the hyperplane that point lies. SVMs use kernel functions to represent data. The kernel function defines similarities between remapped data points. We used the radial basis kernel represented by the equation (1) (Cristianini and Shawe-Taylor, 2000 [8]):

$$K(x, y) = \exp^{-\gamma \left\| x - y \right\|^2}$$

(1)

where $x$ and $y$ are input vectors, and $\gamma$ is a parameter. In this study, SVMs were used with the following descriptors: five physico-chemical property based descriptors selected by t-test after ACC transformation (SVM-T_ACC), amino acid compositions (SVM-AA), and dipeptide compositions (SVM-DIP). One advantage of SVMs is that it can be used to classify linearly separable data as well as nonlinearly separable data. This is because SVMs employ kernel functions. One disadvantage of SVMs, on the other hand, is that their performance is closely tied to the choice of optimal kernel function parameters. We performed the grid-search of the optimal set of parameters for the radial basis kernel function with 10-fold cross-validation analysis using the training dataset (including 300 Cyt-b561 and 300 Non-Cyt-b561). Table 2 shows the results of the selected optimal parameters for the radial basis kernel functions used with SVM-T_ACC, SVM-AA, and SVM-DIP. SVM-T_ACC had the lowest error rate (3.8%). All SVM analyses were performed using R implementation (version 2.60: http//www.R-project.org).

**PSI-BLAST**—In the regular use of PSI-BLAST (Altschul et al., 1997 [1]), position-specific scoring matrices (PSSMs) are built from multiple alignments of significantly similar sequences obtained by similarity search. In this study, multiple alignments of positive (Cyt-b561) sequences were generated using Mafft version 6.5 (Kazutaka and Hiroyuki, 2008 [13]) with the default parameters. This multiple alignments were used to build the PSSMs and for the search. The cut-off E-value of 1.12 was obtained using MEP. For the E-value calculations, regardless of the actual dataset size, a constant sample size of 137,000 (based on the number of predicted proteins in the maize genome) was used for performance comparisons.

**Profile hidden Markov model**—Profile HMMs are full probabilistic representation of sequence profiles (Durbin et al., 1998 [9]). As mentioned before, profile HMMs are built using only positive samples as is the case with PSI-BLAST. In this study, profiles HMMs were built using the w0.5 script of the Sequence Alignment and Modeling Software System (SAM version 3.5; Hughey and Krogh, 1996 [11]). The cut-off E-value of 1.80 was obtained based on the MEP. The E-values were also calculated using the constant sample size of 137,000.

### 2.5 Performance analysis

Ten-fold cross-validation analysis was done for examining classifier performance. The 300 positive (Cyt-b561) sequences were randomly partitioned into 10 subsamples. The same procedure was applied to 300 negative (Non-Cyt-b561) sequences. Combining positive and negative data, each subsample dataset included 60 sequences. Of the 10 subsamples, a single subsample was retained as the validation dataset for testing the classifiers, and the remaining 9 subsamples were used for training. The cross-validation process was then repeated, with each of the 10 subsamples being used exactly once as the validation data. The ten testing results were combined to calculate performance statistics. The advantage of this method is that all samples were used for both training and validation.

Predictions were grouped as follows:

- True positives (TP): the number of actual Cyt-b561 predicted as Cyt-b561 proteins.

- False positives (FP): the number of actual non-Cyt-b561 predicted as Cyt-b561 proteins.

- True negatives (TN): the number of actual non-Cyt-b561 predicted as non-Cyt-b561 proteins.

- False negatives (FN): the number of actual Cyt-b561 predicted as non-Cyt-b561 proteins.

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)

- Mathews correlation coefficient (MCC) = (TP x TN FP x FN)/ $\{$(TP + FN) (TP + FP) (TN + FP) (TN +FN)$\}^{1/2}$

## 3 Results and Discussion

### Classifier performance

Table 3 shows the classification test results of Cyt-b561 proteins from ten-fold cross validation analysis. PLS-T_ACC and SVM-T_ACC outperformed other classifiers as indicated by their higher accuracy rates and higher MCC values. The false positive rate of PLS-T_ACC was the lowest among the four alignment-free classifiers. While SAM and PSI-BLAST had lower false positive rates than alignment-free classifiers, these two classifiers showed significantly higher false negative rates. Tables 4 and 5 show the classification test results for single-domain and multiple-domain Cyt-b561 proteins separately. The results show that the classifiers performed consistently for both single-and multiple-domain Cyt-b561 classifications. Difference in performance is more pronounced for multiple-domain Cyt-b561 classification with significantly higher false negative rates with SAM and PSI-BLAST, likely due to higher sequence divergence among multiple-domain Cyt-b561 sequences.

Figure 1 shows the numbers of true positives identified by the two best alignment-free classifiers (PLS-T_ACC and SVM-T_ACC) and the two alignment-based classifiers (PSI-BLAST and SAM). The majority of the positive sequences (260 of 300) were correctly identified by all four classifiers. Most of these 260 sequences were animal Cyt-b561 proteins. Thirty other Cyt-b561 proteins, mostly from plants, were identified only by PLS-T_ACC and/ or SVM-T_ACC. PLS-T_ACC and SVM-T_ACC had higher false positive rates than SAM and PSI-BLAST. As shown in Figure 2, four Non-Cyt-b561 sequences were commonly identified by all four classifiers as false positives. Three of these four sequences in fact included DOH (or DOMON) domains. It would be interesting to see if these sequences contain unannotated weakly similar to Cyt-b561 sequences.

SVM-T_ACC had the highest number of false positives that were not misidentified by any other classifiers. As shown in Figure 3, PLS-T_ACC and SVM-T_ACC missed the same ten Cyt-b561 sequences (Six from plants, two from animals and two from fungi). All four classifiers misidentified five of these ten Cyt-b561 proteins. Four of these five were multiple-domain Cyt-b561 (three from plants and one from fungi), and one was a plant single-domain Cyt-b561. PSI-BLAST and SAM missed commonly 24 other Cyt-b561 sequences. Most of them were multiple-domain Cyt-b561 from plants. Selection of significant and reduced

numbers of descriptors after ACC transformation appeared to have contributed to higher accuracy and higher MCC values observed in PLS-T_ACC and SVM-T_ACC classifiers. It did not affect the sensitivity of the two classifiers as shown by low % false negative in classifying Cyt-b561 proteins. SVMs trained using amino acid and dipeptide compositions missed some of the Cyt-b561 proteins as indicated by higher false negative rates compared to PLS-T_ACC and SVM-T_ACC. These results indicate that reduced ACC descriptors are better for classifying Cyt-b561 proteins compared to amino acid and dipeptide compositions.

All the classifiers performed better in identifying Cyt-b561 sequences from animals. SAM and PSI-BLAST missed most of the multiple-domain Cyt-b561 proteins from plants. This can be explained by higher divergence among plant Cyt-b561 proteins especially among multiple-domain Cyt-b561 proteins. Sequence identity of Cty-b561 proteins among plants compared to that among animals is lower (22% among plant proteins compared to 35% among animal proteins). This observation is consistent with Cyt-b561 proteins in plants being mostly in multiple-domain forms while animal Cyt-b561 being single-domain proteins. For example, in the mouse genome, there are eleven single-domain Cyt-b561 proteins and a single multiple-domain protein, and in the human genome, there are six single-domain Cyt-b561 proteins and a single multiple-domain Cyt-b561 (based on InterPro release 17.0). On the other hand, there are seven single-domain and fifteen multiple-domain Cyt-b561s in *Arabidopsis*, and eight single-domain and fifteen multiple-domain Cyt-b561s found in the grapevine genome. It indicates that the alignment-based classifiers, SAM and PSI-BLAST, would miss many Cyt-b561 proteins from plants.

## Cyt-b561 mining from plant genomes

We used the two best alignment-free classifiers (PLS-T_ACC and SVM-T_ACC) as well as the two alignment-based methods (SAM and PSI-BLAST) to mine Cyt-b561 proteins from four plant genomes. Table 6 summarizes the results. From the *A. thaliana* genome, PLS-T_ACC and SVM-T_ACC predicted 311 and 659 proteins from *A. thaliana* genome as Cyt-b561 proteins, respectively, while SAM and PSI-BLAST predicted 37 and 29 sequences, respectively, as Cyt-b561 proteins. Twenty four sequences were predicted as Cyt-b561 proteins from the *A. thaliana* genome by all the four classifiers. It includes all 22 Cyt-b561 proteins identified in InterPro and the five Cyt-b561 proteins annotated in the *Arabidopsis* genome project. All the proteins identified by the four classifiers from the *Arabidopsis* genome are presented in Supplementary Table 2 (available at: http://bioinfolab.unl.edu/emlab/Cyt-b561/).

From the grapevine, maize, and soybean genomes, the numbers of Cyt-b561 protein candidates identified were comparable to those from the *Arabidopsis* genome for all four classifiers (Table 6). This implies that even though very few number of Cyt-b561 proteins are currently known from many plants, more Cyt-b561 protein candidates are yet to be identified. Supplementary Tables 3, 4, and 5 present all Cyt-5561 protein candidates identified by the four classifiers from grapevine, maize and soybean genomes, respectively (available at: http://bioinfolab.unl.edu/emlab/Cyt-b561/).

As expected, SAM and PSI-BLAST identified fewer Cyt-b561 protein candidates from all genomes. PLS-T_ACC identified fewer Cyt-b561 protein candidates from all genomes compared to SVM-T_ACC. Based on our performance test results; it is likely that some of the Cyt-b561 protein candidates identified by SVM-T_ACC are false positives. Although PLS-T_ACC prediction would also include false positives, this classifier as well as SVM-T_ACC should miss much fewer true positives compared to SAM and PSI-BLAST.

## 4 Conclusion

In this study, we showed that alignment-based methods, SAM and PSI-BLAST, are too conservative when they are used to search highly divergent proteins as Cyt-b561. We also showed that reduced ACC descriptors are better than amino acid composition as well as dipeptide composition for identifying Cyt-b561 proteins. In conclusion, PLS-T_ACC classifier will be useful for identifying new Cyt-b561 candidates from diverse plant genomes as they become available.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

## References

1. 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]

2. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–D229. [PubMed: 14681400]

3. Asard H, Kappila J, Brczi W. Higher-plant plasma membrane cytochrome b561: A protein in search of a function. Protoplasma 2001;217:77–93. [PubMed: 11732342]

4. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones SN, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR. The Pfam protein families database. Nucleic Acids Res 2004;32:D138–141. [PubMed: 14681378]

5. Bhasin M, Raghava GPS. GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. Nucleic Acids Res 2005;33:W143–W147. [PubMed: 15980444]

6. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout SN, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370. [PubMed: 12520024]

7. Cheng BY, Carbonell JG, Klein-Seetharaman J. Protein classification based on text document classification techniques. Proteins 2005;58:955–970. [PubMed: 15645499]

8. Cristianini, N.; Shawe-Taylor, J. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, U.K: Cambridge Univ. Press; 2000.

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

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

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

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

13. Kazutaka K, Hiroyuki T. Recent development in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics. 200810.1093/bib/bbn013

14. Kim J, Moriyama EN, Warr CG, Clyne PJ, Carlson JR. Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. Bioinformatics 2000;16:767–775. [PubMed: 11108699]

15. Leslie CS, Eskin E, Cohen1 A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. Bioinformatics 2004;20:467–476. [PubMed: 14990442]

16. Liao L, Noble WS. Combining pairwise sequence similarity and Support vector machines for detecting remote protein evolutionary and structural relationships. J Comput Biol 2003;10:857–868. [PubMed: 14980014]

17. Moriyama, EN.; Kim, J. Protein family classification with discriminant function analysis. In: Gustafson, JP.; Shoemaker, R.; Snape, JW., editors. Genome Exploitation: Data Mining the Genome. Springer; New York: 2005. p. 121-132.

18. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L, Copley R, Courcelle E, Das U, Durbin R, Fleischmann W, Gough J, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lonsdale D, Lopez R, Letunic I, Madera M, Maslen J, McDowall J, Mitchell A, Nikolskaya NA, Orchard S, Pagni M, Ponting CP, Quevillon E, Selengut J, Sigrist CJA, Silventoinen V, Studholme DJ, Vaughan R, Wu CH. InterPro, progress and status in 2005. Nucleic Acids Res 2005;33:D201–205. [PubMed: 15608177]

19. Nanasato Y, Akasi K, Yokata A. Co-expression of Cytochromeb561 and Ascorbate oxidase in leaves of wild Watermelon under drought and high light conditions. Plant Cell Phisiol 2005;46:1515–1524.

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

21. Opiyo SO, Moriyama EN. Mining the Arabidopsis and rice genomes for cyclophilin protein families. Int J Bioinformatics Research and Applications 2009;5:295–309.

22. Otaki JM, Mori A, Itoh Y, Nakayama T, Yamamoto H. Alignment free classification of G-protein-coupled receptors using self-organizing maps. J Chem Inf Model 2006;46:1479–1490. [PubMed: 16711767]

23. R Development Core Team. R Foundation for Statistical Computing. Vienna, Austria: 2007. R: A language and environment for statistical computing. http://www.R-project.org

24. 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]

25. Vapnik, VN. The nature of statistical learning theory. New York: Springer-Verlag; 1999.

26. Verelst, W.; Asard, H. Journal of plant physiology. Vol. 161. 2001. Analysis of an *Arabidopsis thaliana* protein family, structurally related to cytochrome b561 and potentially involved in catecholamine biochemistry; p. 175-181.

27. Verelst W, Asard H. A phylogenetic study of cytochrome b561 Proteins. Genome Biology 2003;4:R38. [PubMed: 12801412]

28. 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

## Biographies

Stephen O. Opiyo is 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.
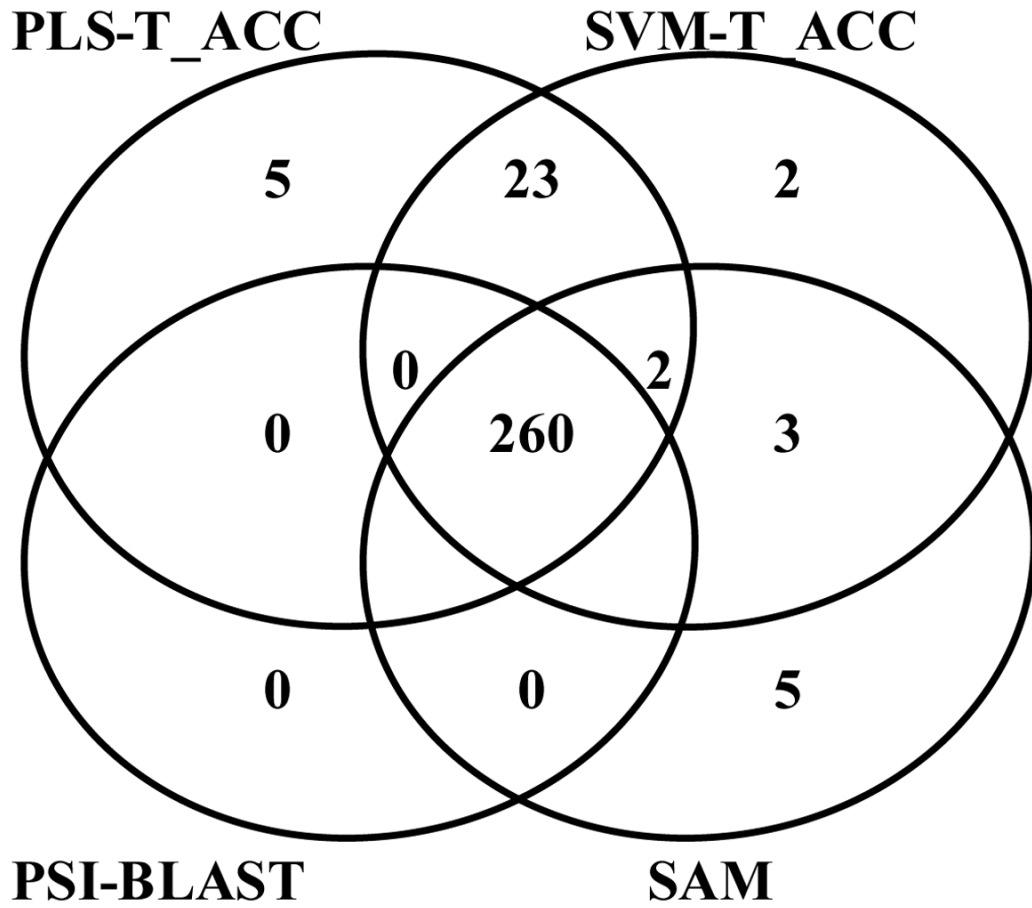
**Figure 1.**
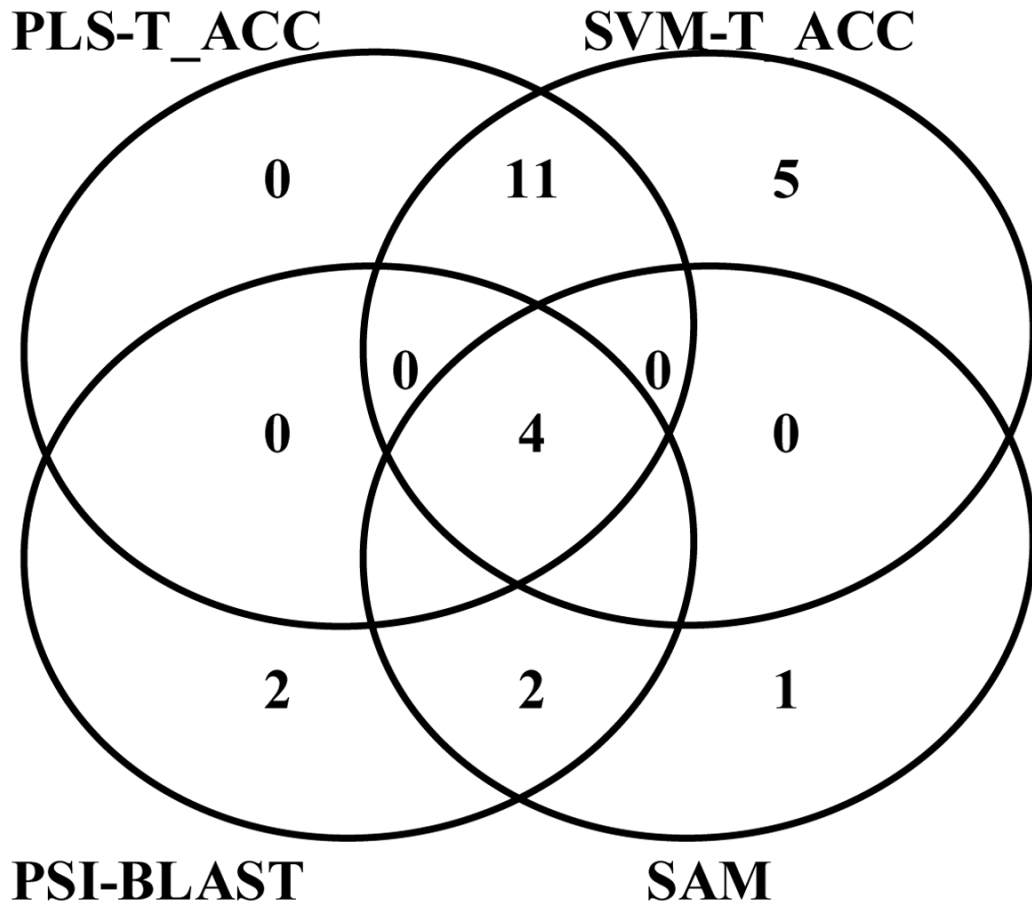The number of true positives by PLS-T_ACC, SVM-T_ACC, PSI-BLAST, and SAM

**Figure 2.**
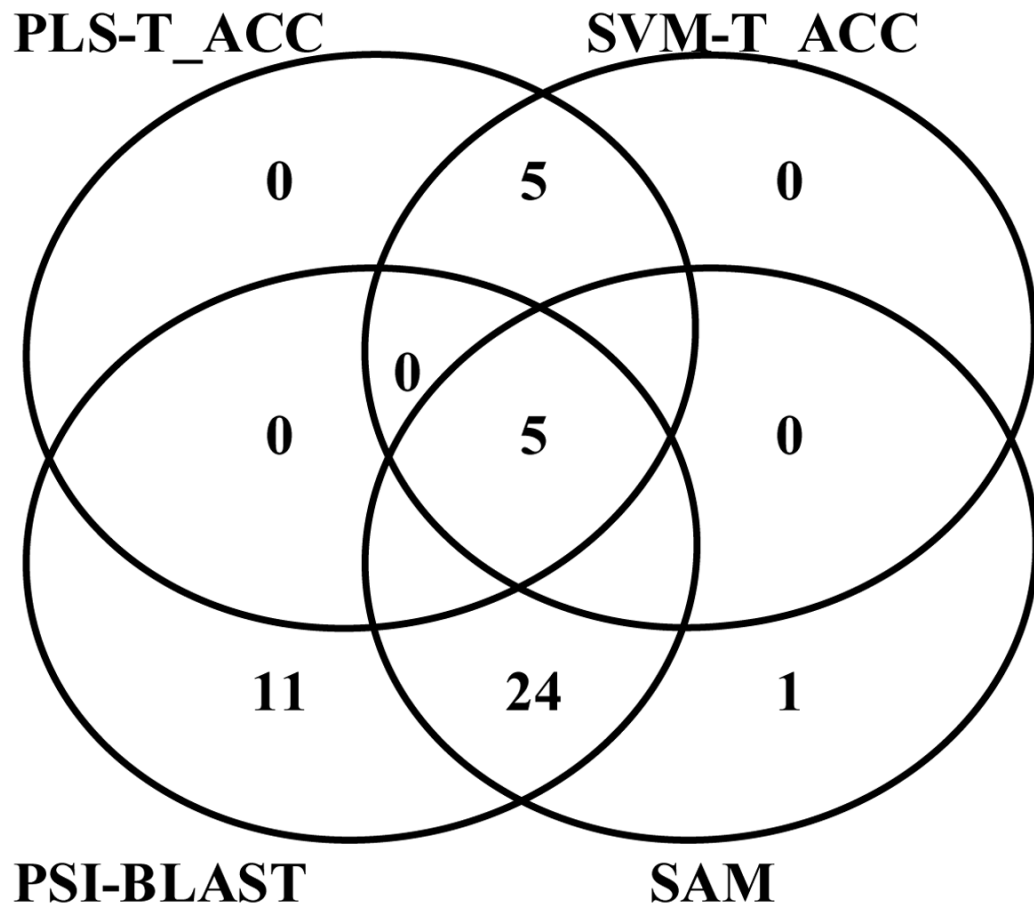The number of false positives by PLS-T_ACC, SVM-T_ACC, PSI-BLAST, and SAM

**Figure 3.**
The number of false negatives PLS-T_ACC, SVM-T_ACC, PSI-BLAST, and SAM

**Table 1**

Numbers of Cytochrome b561 sequences used in this study

| Organisms | Single-domain Cyt-b561 | Multiple-domain Cyt-b561 |
|---|---|---|
| Plants (22 species) | | |
| *A. thaliana* | 7 | 15 |
| Grapevine | 8 | 15 |
| Maize | 1 | 0 |
| Others | 54 | 40 |
| Animal (18 species) | 110 | 20 |
| Fungi (14 species) | 20 | 10 |

**Table 2**

The optimal parameters chosen for kernel functions used with SVM classifiers

| Classifiers | Parameters[a] | %Error[b] |
|---|---|---|
| SVM-T_ACC | $\gamma = 0.0032$, $C = 1.45$ | 3.80 |
| SVM-AA | $\gamma = 0.0640$, $C = 3.44$ | 5.20 |
| SVM-DIP | $\gamma = 0.0065$, $C = 2.12$ | 6.20 |

[a] $\gamma$ = gamma parameter; C = regularization cost parameter

[b] %Error based on 10-fold cross-validation analysis

**Table 3**

Classifier performance on Cytochrome b561 identification

| Classifiers | %Accuracy | %False positive | %False negative | %MCC |
|---|---|---|---|---|
| PLS-T_ACC | 96.2 | 4.5 | 3.0 | 0.92 |
| SVM-T_ACC | 95.5 | 6.1 | 3.0 | 0.90 |
| SVM-AA | 93.6 | 7.8 | 4.8 | 0.87 |
| SVM-DIP | 92.4 | 9.6 | 5.4 | 0.85 |
| SAM | 93.5 | 2.4 | 10.6 | 0.87 |
| PSI-BLAST | 92.7 | 2.4 | 12.1 | 0.86 |

**Table 4**

Classifier performance on single-domain Cytochrome b561 identification

| Classifiers | %Accuracy | %False positive | %False negative | %MCC |
|---|---|---|---|---|
| PLS-T_ACC | 97.3 | 2.6 | 2.6 | 0.94 |
| SVM-T_ACC | 96.9 | 4.3 | 1.7 | 0.93 |
| SVM-AA | 95.2 | 5.6 | 3.9 | 0.90 |
| SVM-DIP | 94.3 | 9.6 | 5.4 | 0.89 |
| SAM | 95.8 | 1.7 | 6.5 | 0.91 |
| PSI-BLAST | 95.8 | 1.7 | 6.5 | 0.91 |

**Table 5**

Classifier performance on multiple-domain Cytochrome b561 identification

| Classifiers | %Accuracy | %False positive | %False negative | %MCC |
|---|---|---|---|---|
| PLS-T_ACC | 94.0 | 6.0 | 3.6 | 0.88 |
| SVM-T_ACC | 93.0 | 10.0 | 4.0 | 0.86 |
| SVM-AA | 89.0 | 13.0 | 9.0 | 0.79 |
| SVM-DIP | 87.0 | 16.0 | 10.0 | 0.74 |
| SAM | 90.0 | 4.0 | 15.0 | 0.81 |
| PSI-BLAST | 88.0 | 4.0 | 20.0 | 0.77 |

**Table 6**

The number of predicted Cyt-b561 from four plant genomes rice genomes

| Genomes | InterPro[a] | PLS-T_ACC[b] | SVM-T_ACC[b] | SAM[b] | PSI-BLAST[b] |
|---|---|---|---|---|---|
| A. thaliana | 22 (5) | 311 (22) | 659 (22) | 37 (22) | 29 (22) |
| Grapevine | 23 (0) | 199 (23) | 372 (23) | 43 (23) | 32 (23) |
| Maize | 1 (0) | 200 (1) | 557 (1) | 54 (1) | 25 (1) |
| Soybean | 0 (0) | 312 (0) | 364 (0) | 25 (0) | 27 (0) |

[a]The numbers of Cyt-b561 proteins identified in InterPro (release 17.0, March 2008; IPR006593). The numbers of Cyt-b561 proteins annotated in each genome project are shown in parenthesis.

[b]The numbers of Cyt-b561 proteins predicted by each classifier; those also identified in InterPro and/or genome are shown in parenthesis.