PROTEIN FAMILY CLASSIFICATION USING MULTIVARIATE METHODS

by

Stephen O. Opiyo

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Agronomy

Under the Supervision of Professors Etsuko Moriyama and George L. Graef

Lincoln, Nebraska

June, 2007

PROTEIN FAMILY CLASSIFICATION USING MULTIVARIATE METHODS

Stephen O. Opiyo, PhD.

University of Nebraska, 2007

Advisors: Etsuko Moriyama and George L. Graef

The number of protein sequences from agriculturally important crops is rapidly increasing in databases. In order to identify their functions efficiently and accurately, good computational methods are needed. Commonly used methods search databases using alignments. Some proteins may lack enough sequence similarities even though they share similar structures and biochemical functions. In such cases, alignment-based methods fail to identify proteins correctly. In order to classify these difficult proteins, alignment-free methods based on, e.g., multivariate methods are required. I examined application of two multivariate methods; principal component analysis (PCA) and partial least squares (PLS). Their performances were compared against profile hidden Markov models (HMMs) and PSI-BLAST. G-protein coupled receptors (GPCRs), cyclophilins, cytochrome b561 (Cyt b561), and immunoglobulin protein families were included in this study. Using physico-chemical properties as descriptors, I examined how the training dataset affects performance of the methods, how the methods can identify short fragmented sequences, and how the methods can identify proteins when only remotely similar samples are included in the training sets. The PLS methods outperformed profile HMM and PSI-BLAST when only a small number of positive samples (5 or 10) were included in the training dataset. PLS methods performed also better than profile HMM and PSI-BLAST in the identification of short fragmented sequences, and Cyt b561 expressed sequence tags from the *Arabidopsis* genome. Combining the results of PLS with other alignment-free methods, 342 proteins were identified as GPCR candidates, including 20 of the known 22 *Arabidopsis* GPCRs. Profile HMM identified only 15 of them. PLS method with descriptors selected by the t-test outperformed PLS method with descriptors from auto and cross-covariance in identifying cyclophilins from *Arabidopsis* and rice genomes. Finally, I developed a simple statistics method (ST-method) that is sensitive to protein with weak sequence similarities and generates low false positives. The ST-method outperformed PLS methods, profile HMMs, and PSI-BLAST in the classification of GPCRs and immunoglobulin superfamily. It identified 579, 717, and 382 GPCR candidates from *Arabidopsis*, rice, and maize genomes.

DEDICATION

This dissertation is dedicated to the memory of my wife the late Rukia Dion Opiyo whose special love, support, encouragement and patience helped me to attain this degree.

This work is also dedicated to Vicky Oroma, Peace Acan, Gladys Acaa, Olive Lanyero, Galia Aciro, Isaac Ayela, Shamim Abu, and Apollo Obol. You are my inspiration.

Last but not least, the work is also dedicated in the memory of: Jennifer Abalo, Fred Okumu, Norbert Okello, Rose Acen, Florence Aringo, Kelementina Okutu, and Marino Okutu.

ACKNOWLEDGEMENTS

My heartfelt gratitude goes to my parents the late Mrs. Kelementina Okuto and the late Mr. Marino Okuto. You instilled in me the desire, work ethic, and confidence to accomplish any goal that I set.

I would like to express my deep and sincere thank to my advisor Dr. Etsuko Moriyama, for her excellent encouragement, support and guidance throughout my PhD studies. She took me in as a PhD student with no background in any computer programming and allowed me to pursue a specialization in Bioinformatics. It has been a pleasure and a privilege to work with Dr. Moriyama. I cannot express how appreciative I am for the opportunities, training, and experience she has given me. Appreciation and gratitude is also extended to Dr. George L. Graef for being my co-supervisor and guiding me during my study. I am very grateful to Drs. Han Asard and Paul Staswick for being dependable committee members, and giving useful comments. I would like to thank Dr. Stephen Kachman for his exceptional guidance in the statistical analysis, and in providing constructive suggestions on my research.

My thanks go to graduate and undergraduate students in Moriyama lab who made me laugh and helped me finish my research. Thank you for the useful comments you gave me during the lab meeting presentations. I would like to thank Dr. Daniel and Dee Ann Davidson who inspired and supported me throughout my academic career in the past years. My appreciation goes to John Karuta for all the support he has given me during the past years. My thank goes to my special friend Katherine Porter who always encourages me to do better. My appreciation goes to Anna Akot for the useful comments she gave me while working on this dissertation.

I am very grateful to my siblings Vincent Ochaya, Jimmy Obol, Mosses Oryem, Paul Otim, Richard Okwera, Jessica Aol, Milly Amony, and Monica Ajok. Your unconditional love and prayers made attaining this degree possible. I thank you for all the support you have given me in the past years.

My deep and sincere thanks go to Santa Adong, Christine Abalo, Bosco Onyuk, Grace Menya, Evelyne Nalwandu, and Geoffrey Kinyera for the support and encouragement you have given me in the past years. It was not easy for you to take all the responsibilities that you had but you did exceptional jobs.

Last but not least, I would like to thank the Davidson's family, especially Jim and June Davidson for being good grandparents. I appreciate the help that the Davidson's family gave me during my study in the United States.

TABLE OF CONTENTS

TABLE

FIGURES

LIST OF TABLES

CHAPTER 1.

CHAPTER 2.

Section 2.2

Section 2.3

Section 2.4

CHAPTER 3.

CHAPTER 4.

LIST OF FIGURES

CHAPTER 1.

CHAPTER 2.

SECTION 2.2

SECTION 2.3

CHAPTER 3.

CHAPTER 4.

CHAPTER 1

INTRODUCTION

1.1 BIOINFORMATICS IN AGRONOMY

Genome projects on model plants and many agriculturally important crops are resulting in rapid accumulation of genomic and expressed sequence tag (EST) sequences in many databases. These databases provide rich information sources for genes involved in agronomicaly important traits such as virus and insect resistance, bacterial resistance, abiotic stress tolerance, and also for novel genetic markers that can be used for crop improvements. The rate at which these sequence data for crop species are accumulating has lead to a rapid development of databases and analysis tools. Comparative genomics is one of the focal areas, for example, in both method development and application. Bioinformatics has surely become an integral part of conducting plant and crop science research.

1.1.1 Some examples of bioinformatics applications in crops

One example where bioinformatics is useful to crop scientists has been in the use of ESTs. ESTs are short cDNA sequences that serve to “tag” the genes from which the messenger RNAs (mRNAs) originated. EST sequences can be used to search DNA and protein databases using various bioinformatics tools for similar genes. Information obtained from searches can then be used to determine if the specific gene or sequence was already found in the same or other organisms and if its function has been determined. Most of these searches are done by Basic Local Alignment Search Tools (BLAST) [1].

Recently the rice Chromosomes 11 and 12 Sequencing Consortia[2] group annotated rice chromosomes 11 and 12 using profile hidden Markov models (HMMs)[3]. A total of 5,993 non-transposable element related genes were found on these chromosomes. Among them were 289 disease resistance-like and 28 defense response genes. Access to these genes will facilitate research on disease resistance-like genes and defense response genes with larger, partially sequenced genomes such as maize and sorghum. This will also be an essential resource for the engineering of rice with tolerance or resistant to diseases.

The production of low linolenic acid soybean improves the stability and flavor in soybean oil and eliminates the use of hydrogenation[4,5]. Consumption of trans fatty acid found in hydrogenated oils has been linked to an increased risk of coronary heart disease [6]. Three independent genetic loci are associated with seed linolenic acid levels, with mutant alleles identified as *fan*, *fan2*, *fan3*, and *fanx*[7]. In low linolenic acid line A5, the *fan*(A5) locus was shown to be associated with a deletion of omega-3-fatty acid desaturase gene (FAD3)[8]. Bilyeu et al.[9] searched the soybean genome using BLAST similarity search and a gene involved in biosynthesis of low linolenic acid (mutated *FAD3*)[10] from *Arabidopsis* as a query and identified a new FAD3 gene from the soybean genome. Bilyeu et al.[9] developed a molecular marker for low linolenic acid from the identified gene. The molecular marker developed can be used in soybean breeding programs for breeding low linolenic acid soybean.

Comparative data analysis within and between genomes is another method used in bioinformatics. For example, various bioinformatics searching methods have been applied to mine simple sequence repeats (SSRs, microsatellites) from genomic and EST data, which have considerable utility as selectable markers in genetics and crop breeding. Within the cereal and rice genomes a large number of ESTs have been screened for the presence of SSRs[11].

The development and use of T-DNA knockouts has become a central tool of plant functional genomics. A range of bioinformatics tools have been made available to identify relevant plant lines and to analyze and map integration sites. GABI-Kat [12] is a database of *Arabidopsis* lines with flanking T-DNA sequence tags. BLASTN[1] and TBLASTN[1] are used to search the database for sequences with flanking T-DNA sequence tags.

The ability of *Pseudomonas syringae* pv. to cause halo blight of bean is dependent on its ability to translocate effector proteins into the host's cells via the hypersensitive response and pathogenecity (Hrp) type III secretion system. Monica et al.[13] used profile HMMs approach to identify genes coding for type III effectors and other virulence factors that are regulated by HrpL alternative sigma factor from the genome of *P. syringae pv. phaseolicola* 148A. The bioinformatics approach used in their study was robust enough to predict accurately most of the effectors in *P. syringae pv. phaseolicola* 148A. The results of the prediction were experimentally confirmed using real-time PCR analysis. In a similar study, Silverstein et al.[14] searched the *Arabidopsis* genome using profile HMMs and BLAST similarity search to identify defensin-like sequences (DEFLs) in the *Arabidopsis* genome. They identified 317 DEFLs in *Arabidopsis* including 15 known defensins.

A bioinformatics approach has helped in identifying residues that are involved in substrate specificity in plants. Plant acyl-acyl carrier protein (ACP) thioesterases hydrolyze acyl-ACP thioester bonds releasing free fatty acids and ACP. There are two functional classes of plant acyl-ACP thioesterases (Fat A and Fat B). Fat A catalyzes unsaturated fatty acid-recognizing, and Fat B catalyzes saturated fatty acid-recognizing. Mayer and Shanklin[15] used a profile HMMs and TBLASTX[1] to identify amino acid residues involved in substrate specificity (saturated or unsaturated fatty acid). The approach allowed the identification of specificity determining residues that differ between Fat A and Fat B.

1.1.2 Plants genomics and Expressed Sequence Taqs

In plant genomics, complete genomic sequences are currently available only from four model plants (rice, *A. thaliana, Medicago truncatula*, and *Populus trichocarpa*). Most of the plant genomic information is found in EST sequences (See Appendix Table 1). Major crop genome projects that are mainly based on the EST strategy include *Zea mays* (maize) [16], *Glycine ma*x (soybean)[17], and *Triticum aestivum* (wheat)[18]. In the Expressed Sequence Tags Database at National Center for Biotechnology Information (dbEST)[19], the total number of ESTs released in March, 2007 was 42,050,137. The numbers of ESTs for major crops are: 1,211,154 for rice, 1,161,193 for maize, 855,098 for wheat, 371,897 for soybean, 249,794 for tomato, 227,351 for potato, 204,308 for sorghum, and 177, 113 for cotton.

1.2 OVERVIEW OF THIS STUDY

1.2.1 Objectives

Evidently bioinformatics has become a critical component of plant and crop science research as indicated in the examples given in Section 1.1.1. All the bioinformatics methods (BLAST and profile HMMs) used in the examples given in Section 1.1.1 rely on sequence alignments. Alignment-based methods have limitations because alignments are known to become unreliable when sequence similarity drops below forty percent[20]. In one of the examples given in Section 1.1.1, BLAST method was used to search for omega-3-fatty acid desaturase gene from soybean genome using a mutated *FAD3* gene from *Arabidopsis* as a query. The BLAST search identified a FAD3 gene from the soybean genome because the sequence similarity between the mutated *FAD3* from *Arabidopsis* and FAD3 from soybean is 85%. However, some homologous proteins such as G-protein coupled receptors (GPCRs) are highly divergent and lack enough sequence similarities even though they still share similar structures, biochemical properties, and functions. In such cases, obtaining reliable alignments among these protein sequences is extremely difficult, and methods that rely on alignments such as BLAST, PSI-BLAST[21] (position specific iterative BLAST), and profile HMMs would fail to identify these proteins from databases. Another disadvantage of these alignment-based methods is that they are trained using only “positive” samples (proteins of interest). “Negative” samples (unrelated proteins) cannot be included in the alignments their models are built from, and these methods cannot be optimized directly for discriminating between positive and negative samples.   In order to overcome the disadvantages of alignment-based methods, methods that do not rely on alignments are needed. Multivariate methods do not rely on alignments and they can be applied to identify proteins with weak sequence similarities from plant sequence databases.

As mentioned in Section 1.1.2, currently there are only four plants whose genomes have been completed. Whereas many more are in progress, the vast majority of genomic information for agriculturally important crops is found in rapidly increasing EST sequences (See Appendix Table 1 for a list). For example, because of the size and complexity of the soybean genome, it is unlikely that, given the current technology, its entire genome can be completely sequenced in the near future. Currently used alignment-based search methods do not perform well against these short EST sequences. Therefore, in order to mine sequences from short EST sequences efficiently and effectively from non-model plant and crop genomes, it is important to develop protein mining methods that are sensitive to such fragmented sequences. Multivariate methods that do not rely on alignments can be useful in this situation too.

In this dissertation, I will develop methods that can effectively identify proteins with weak similarities and short sequences from diverse plant genomes. Instead of relying on commonly used alignment-based methods, I will apply multivariate analysis methods using physico-chemical properties and compositions of amino acids. The underlying hypothesis is that the physico-chemical properties of proteins have enough specific information, so that they can be used to identify proteins that share similar functions even from short and diverged sequences where alignments cannot be reliable. The protein classification methods developed in this study will facilitate in the future analyzing proteins with weak similarities and short EST sequences available from many crop genomes.

My first objective of this dissertation was, therefore, to develop sequence descriptors from physico-chemical properties of amino acids. These descriptors were developed using principal component analysis. The second objective was to determine the number of samples required in a training dataset. The third objective was to examine the performance of the methods on short subsequences. The subsequences were obtained by taking 50, 75, and 100 amino acids from N and C-terminal of each sequence. Finally, the methods were applied to the *Arabidopsis*, rice and maize genomes, and several protein families were actually searched.

1.2.2 Relevance of this study in crop science

The focus of this study was to develop protein family classification methods for identifying protein sequences with weak similarities and methods that can fully utilize the information available in EST sequences. EST sequence information is largely underutilized especially for gene and/or protein mining due to the lack of sufficiently sensitive methods for such partial sequences. Beyond a few model plants as described earlier, many crop genome projects concentrate on EST sequencing based on the economical reason. Therefore, developing good EST mining methods will contribute to advancing genome research in many crops. The methods developed in this study will also allow us to perform thorough mining of protein families especially with weak sequence similarities, and currently, not many of these proteins are identified from diverse crop genomes. The results obtained from through mining protein families (e.g., GPCRs and cyclophilins) will significantly advance our knowledge on these protein families.

1.2.3 Organization of the dissertation

The dissertation is organized as follows. Chapter 1 discusses some of the examples of how bioinformatics has been applied to agronomy, the current status of genomic and EST databases in plants, and the overview of this study. This chapter also includes the commonly used methods of protein family classifications, new approaches to protein family classifications, and the protein families used in the study.

Chapter 2 discusses the application of partial least squares (PLS) in the protein family classifications. The chapter is divided into four sections as follows. Section 2.1 is an overview of Chapter 2. Section 2.2 is based on a manuscript that was published in the Journal of Proteome Research **(Opiyo and Moriyama, 2007. J. Proteome Res. 6:846-853).** The objectives of this section are to examine how the size of training datasets affects the classifier performance, and how different methods could identify sequence fragments with different lengths. Section 2.3 is about remote similarities classification of GPCRs. The objective of this section is to determine how different methods can identify GPCR classes that are not included in the training dataset. Section 2.4 is about the mining of highly-divergent seven transmembrane receptors from the *Arabidopsis* genome. This section is based on the manuscript **(Moriyama, Strope, Opiyo, Chen, and Jones, 2006. Genome Biology 7:R96).** The objective of this section was to mine seven transmembrane receptors from the *Arabidopsis* genome.

In Chapter 3 descriptor selection is explored, and how it affects the performance of PLS method is examined. Student t-test and rank test were used to reduce the number of descriptors after auto and cross-covariance transformation (explained in Chapter 2). These PLS methods are then used to mine cyclophilins from *Arabidopsis* and rice genomes.

From Chapters 2 and 3, I found that while PLS methods are sensitive to protein families with low sequence similarities, they also have high false positives. Therefore, a method that has fewer false positives as well as being sensitive to sequences with low similarities is needed. In Chapter 4, I developed a simple statistics method for protein family classification. This method was developed from descriptors selected by self-organizing map (explained in Chapter 4) and t-test statistics. This method was used to mine seven transmembrane receptors from *Arabidopsis,* rice, and maize genomes.

Chapter 5 concludes the dissertation and presents some suggestions for the future work.

1.3 COMMONLY USED METHODS FOR PROTEIN FAMILY CLASSIFICATION

1.3.1 Pairwise similarity search methods

The number of new proteins in databases is rapidly increasing. It has created a need of automated methods of protein classification. The most commonly used methods developed to meet this demand are based on pairwise alignments. Similarity search methods compare sequences in a database against the query sequence. Classification of the query sequence is done by assigning that of the most similar sequence in the database. SSEARCH[22] similarity search method uses the Smith-Waterman local alignment dynamic programming algorithm[23] for pairwise alignments. BLAST[21] and FASTA[24] use heuristics algorithms to search sequences in the databases. Heuristic algorithms are faster, but they do not guarantee that the truly best match can be found.

1.3.2 Motifs

Another category of alignment-based methods used in PROSITE[25] and PRINTS[26] searches a database for the presence of motifs. A motif is a short subsequence that is highly conserved across a group of proteins. The use of multiple alignments increases the sensitivity of searching for motifs as compared to pairwise alignment-based similarity searches. Figure 1 shows a motif in a regular expression pattern, this motif can be shown as Y-x(2)-G-x(2)-L. It starts with amino acid Y followed by any two amino acids, then G, then followed by any two amino acids, and with L.

1.3.3 PSI-BLAST

PSI-BLAST builds position-specific scoring matrices (PSSMs) from a multiple alignment. Figure 2 shows how PSI-BLAST performs protein similarly search. First, PSI-BLAST performs a regular protein similarity search (BLASTP) against a protein database using a single protein query. It then generates a multiple alignment of the sequences found by the BLASTP run above a certain preset score or e-value threshold and calculates a profile or a PSSM from the multiple. The PSSM is generated by calculating position-specific scores for each of these amino acids at each position in the alignment. This profile (PSSM) is used in place of the original substitution matrix for a further search of the database to detect sequences that match the conservation pattern. The newly detected sequences from this second round of the search that are above the specified score (e-value) threshold are again added to the alignment to refine the profile for another round of searching. This process is iteratively continued until a desired iteration or to convergence, i.e., the state where no new sequence is detected above the defined threshold. The iterative profile generation process makes PSI-BLAST far more capable of detecting distant sequence similarities than a single query alone in BLASTP.

1.3.4 Hidden Markov Models

A hidden Markov model[27] is a finite set of states, each of which is associated with a probability distribution. In a particular state, an output can be generated according to the associated probability distribution. It is only the outputs, not the states, which are visible to the observer, therefore states are hidden. A simple example of a hidden Markov model is given in Figure 3. It consists of a set of two states and transitions between these states. Each state emits a signal, based upon a state-specific emission probability distribution and then transitions to some other state, based upon a transition probability distribution.

A hidden Markov model is defined by:

Q = the set of states, (q_1 and q_2 in Figure 3)

V= the output alphabets, (V_1 and V_2 in Figure 3)

$\pi(_i)$ = the probability of being in a state (q_i) at time t = 0 ($\pi(_1)$ and $\pi(_2)$ in Figure 3)

A = transition probabilities = $\{a_{ij}\}$, where $\{a_{ii}\}$ is the probability of entering state q_i from

q_i, and $\{a_{ij}\}$ is the probability of entering state q_j from q_i at a time t + 1 (a_{11}, a_{22}, a_{12} and a_{21} in Figure 3)

B = emission probabilities = $\{b_i(k)\}$, where $b_i(k)$ is the probability of producing output V_k in state q_i at time t ($b_1(1)$ and $b_2(2)$ in Figure 3)

1.3.5 Profile hidden Markov models

A profile hidden Markov model[3] (profile HMM) is a probabilistic representation of a protein family. Profile HMM search consists of the following three steps: 1) a multiple sequence alignment is made from known members of a given protein family; 2) a profile HMM is built from the family; 3) and a query sequence is compared with all the HMMs in a database. A log-odds score is assigned with respect to the model to the query sequence. If the score is significant at (alpha level = 0.05 or selected p-value), there is a high chance that the query sequence is a member of the protein family represented by the model. An architecture of a profile HMM is shown in Figure 4. It has a simple left-to-right structure in which there is a repetitive set of three states, designated as match, delete, and insert (M, D, and I). The match state represents an amino acid for this position in the protein family. The delete state is a non-emitting state, and represents skipping this position in the multiple alignment. Finally, the insert state models the insertion of any number of residues after this consensus position. Match and insert states emit 20 amino acids with certain probabilities (emission probabilities).

Figure 5 is an example showing how a protein sequence from a multiple alignment can follow a path in the model. The path of Protein_X1 is highlighted with blue arrows. The first column corresponds to the match state M_1 where all four sequences have amino acids. In the second column, amino acid “I” is inserted in Protein_X1, corresponding to the insert state I_1. In the third column, an amino acid is deleted from Protein_X1 (following to the delete state D_2 in the model), and then the final column corresponds to the match state M_3. On the model, there is one more state "End" to end the sequence.

The two most commonly used methods for profile HMMs are HMMER used in the Pfam[28] and SMART[29] databases, and Sequence Alignment and Modeling Software System (SAM) [30] used in the Superfamiliy[31] database. SAM is used in this study.

1.3.6 Problems with currently used alignment-based methods

As mentioned in the section 1.2.1, building sequence models (e.g., patterns, motifs, profiles, profile HMMs) for these currently used methods requires reliable alignments. If homologous proteins lack enough similarities, generating alignments among them with enough confidence becomes difficult. Another problem with alignment-based methods which was also mentioned in the section 1.2.1 is that the sequence models are built using only positive samples. The lack of using negative information limits the discrimination power of these alignment-based methods.

1.4 NEW APPROACHES TO PROTEIN FAMILY CLASSIFICATION

1.4.1 Data used in alignment-free methods

Protein classification methods described so far rely on alignments. A new approach is to avoid using alignments, so called alignment-free methods for protein family classification. They also include both positive and negative samples in building their models. Since alignment-free methods do not rely on alignments, their protein sequences are transformed into a uniform matrix before any analysis is performed. The transformation of protein sequences is done by either converting each sequence into their composition of amino acids, dipeptides, etc, or representing it with mean phyisco-chemical properties or transforming it using auto and cross-covariance method (explained in Chapter 2).

(a) Amino acid composition

From each protein sequence, frequencies of 20 amino acids are calculated using

$$aa_i = \frac{total_{(aa)i}}{total_{aa}} \tag{1}$$

where aa_i is the composition of amino acid of i in a protein sequence, $total_{(aa)i}$ is the total number of amino acid i, and $total_{aa}$ is the total number of amino acids in the protein. Figure 6 shows an example of protein sequences represented as amino acid sequences as well as amino acid composition.

(b) Dipeptide composition

Dipeptide composition represents the 400 frequencies of all consecutive amino acid pairs in a protein sequence and corresponds to a (20 X 20) matrix. It encapsulates information on composition of amino acids as well as their local orders. The frequency of amino acid pair i is given using the following equation:

$$\text{dipeptide}_{\text{i}} = \frac{\text{total(dipeptide)}_{\text{i}}}{\text{total}_{\text{dipeptide}}} \qquad (2)$$

where total(dipeptide)$_{\text{i}}$ is the total number of dipeptide$_{\text{i}}$ out of 400 dipeptides, and total$_{\text{dipeptide}}$ is the total number of all possible dipeptides.

1.4.2 Unsupervised vs. supervised learning

In unsupervised learning, there is no prior knowledge of the group or class to which the samples belong. Examples of methods used in unsupervised learning are principal component analysis and self-organizing maps. Supervised learning requires a training set sample whose groupings are known. The task of supervised learning is to predict group of memberships of new samples. Discriminant analysis[32], partial least squares[33], and support vector machines [34] are some of the methods that are used in supervised learning.

1.4.3 Principal component analysis

Principal components analysis (PCA) is a dimensional reduction method where a set of independent variables $\boldsymbol{X}$ data matrix are presented by fewer number of variables $\boldsymbol{t}$. For example, let $\boldsymbol{X}$ be an m by n data matrix in which m rows are sample sequences and n columns are variables (e.g., amino acid compositions). We assume that $\boldsymbol{X}$ is mean centered such that its columns variables all have the zero mean. In PCA, $\boldsymbol{X}$ is decomposed into the sum of the product of n pairs of vectors. Each pair consists of an n by 1 vector called the *eigenvectors*, $\boldsymbol{p}_i$, and an m by 1 vector called the *scores*, $\boldsymbol{t}_i$. Thus $\boldsymbol{X}$ can be written as

$$\boldsymbol{X} = \boldsymbol{t_1 p_1}^T + \boldsymbol{t_2 p_2}^T + \ldots + \boldsymbol{t_n p_n}^T \qquad (3)$$

The matrix of loading vectors, ***P,*** forms a new orthogonal basis for the space spanned by ***X*** and the individual $\boldsymbol{p}_i$ are the eigenvectors of the covariance matrix of the mean-centered data matrix ***X***. When PCA is done on a data set, it is often found (and it is generally the objective) that only the first few eigenvectors are associated with systematic variation in the data, and the remaining eigenvectors are associated with "noise". PCA models are formed by retaining only the eigenvectors that are descriptive of systematic variation in the data, thereby decreasing the dimensions of the original data space.

PCA can be explained using a geometrical approach. For example, consider a dataset of ten protein sequences (Protein_X1, Protein_X2, …, Protein_X10), each represented with three variables: mass, volume, and isoelectric point (see Figure 7A). For each protein, the variables are transformed by standardization, a process known as data pre-treatment, using the following equation:

$$\mathrm{StnX}_{(\mathrm{i,j})} = \frac{\mathrm{X}_{(\mathrm{i,j})} - \overline{\mathrm{X}}_{(\mathrm{j})}}{\sqrt{\sum_{\mathrm{i=1}}^{\mathrm{N}} (\mathrm{X}_{(\mathrm{i,j})} - \overline{\mathrm{X}}_{(\mathrm{j})})^2 / (\mathrm{N}-1)}} \qquad (4)$$

where $\mathrm{StnX}_{(\mathrm{i,j})}$ is the standardized variable *j* for the protein *i*, $X_{(\mathrm{i,j})}$ is the original variable *j* for the protein *i*, $\overline{\mathrm{X}}_{(\mathrm{j})}$ the mean of variable *j* in the dataset, and N is the number of the samples in the dataset. After the transformation, the data have the zero mean and unit standard deviation (Figure 7B). Each observation of the ***X*** matrix is represented in K-dimensional variable space K=3 in the example (Figure 8A). The first principal component (PC1) is the line in the K-dimensional space (K = 3 in our example) that best approximates the data (Figure 8B). Each observation is projected onto this line in order to obtain a coordinate along the PC1-axis. This coordinate is known as a score (Figure 8B). The second principal component (PC2) is orthogonal to the first (Figure 8B), and observations are also projected on this PC2-axis to obtain the second set of scores. This process is repeated up to K times. As mentioned before, usually fewer than K principal components are used.

One advantage of PCA is that it reduces dimensions of the data space, which may be easily viewed and used. Each principal component can be displayed graphically and may often be interpreted according to pattern, knowledge or trends in the data. Another advantage of PCA is that it is applicable to almost any type of data. The data do not have to be normally distributed, and PCA can be applied to data with more variables than observations. PCA have been used in protein classification problems because of its advantages mentioned above. Lapnish et al.[35] used PCA to separate Class A G-protein coupled receptors (GPCRs) into amine, olfactory, glycoprotein hormone, and opsin receptor families. Gunnarsson et al.[36] used PCA to investigate whether there was a particular GPCR transmembrane region that was responsible for separation between amine and rhodopsin families. Using physico-chemical properties of amino acids of GPCR sequences, PCA score plots showed that all seven transmembrane regions were equally important in separation of the two families.

1.4.4 Partial least squares

Partial least squares (PLS)[33] is an extension of PCA that finds a relationship between a matrix of predictor variables, $\boldsymbol{X}$, and a matrix of dependent variables, $\boldsymbol{Y}$. PLS has two objectives: to approximate the $\boldsymbol{X}$ and $\boldsymbol{Y}$ data matrices, and to maximize the correlation between them. The extraction of PLS components is performed stepwise and a regression relating each $\boldsymbol{Y}$ variable and the $\boldsymbol{X}$ matrix is created. PLS analysis can be performed on a predictor of $\boldsymbol{X}$ (m x n) and a dependent Y (m x l), where m represents observations, n represents predictor variables, and l represent dependent variables. As in equation 3 for PCA, $\boldsymbol{X}$ can be decomposed as $\boldsymbol{X = TW}^{\mathrm{T}}$, where $\boldsymbol{T = (t_1, t_2, \ldots)}$, the score vectors, and $\boldsymbol{W}$ is the eigenvector of the matrix $\boldsymbol{X}^{\mathrm{T}}\mathbf{YY}^{\mathrm{T}}\mathbf{X}$. First, the eigenvectors for $\boldsymbol{X}$ and $\boldsymbol{Y}$, $\boldsymbol{w_i}$ and $\boldsymbol{q_i}$, respectively, and their corresponding score vectors are derived as $\boldsymbol{t_i} = \boldsymbol{Xw_i}$ and $\boldsymbol{u_i} = \boldsymbol{Yq_i}$. Next, rank-one reductions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are performed such that $\boldsymbol{X}_{\mathrm{i+1}} = \boldsymbol{X}_{\mathrm{i}} - \boldsymbol{t_i p_i}^{\mathrm{T}}$ and $\boldsymbol{Y}_{\mathrm{i+1}} = \boldsymbol{Y}_{\mathrm{i}} - \boldsymbol{t_i q_i}^{\mathrm{T}}$, where $\boldsymbol{p_i} = \boldsymbol{X}_{\mathrm{i}}^{\mathrm{T}}\boldsymbol{t_i}/(\boldsymbol{t_i}^{\mathrm{T}}\boldsymbol{t_i})$, and $\boldsymbol{q_i} = \boldsymbol{Y_i}^{\mathrm{T}}\boldsymbol{t_i}/(\boldsymbol{t_i}^{\mathrm{T}}\boldsymbol{t_i})$. These steps are repeated until the desired number (K) of score variables has been extracted. The regression coefficients $\boldsymbol{b}_{\mathrm{pls}}$ given in equation 5 are useful for predictions of new independent variables ($\boldsymbol{X}_{\mathrm{new}}$) using equation 6.

$$\mathbf{b}_{\mathrm{pls}} = \boldsymbol{W}(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{W})^{-1}\mathbf{q} \qquad (5)$$

$$\mathbf{y} = \boldsymbol{X}_{\mathrm{new}}\boldsymbol{b}_{\mathrm{pls}} \qquad (6)$$

As with PCA, PLS can be explained using a geometrical approach. For illustration purposes, I will use the same dataset used for PCA for our independent variables ($\boldsymbol{X}$) and a new dataset for dependent variables (***Y***) with l =3. These datasets are presented in Figures 7A and 9A respectively. After data transformation (Figures 7B and 9B), each observation from the independent dataset is plotted on the X-space, and each observation from the dependent dataset is plotted on the Y-space (represented in Figures 10A and 10B respectively). The first PLS component is a line in the X-space and another line in the Y-space (Figures 10 C and 10D respectively). These lines are calculated such that they approximate the points in ***X*** and ***Y***. Scores $\boldsymbol{t_1}$ and $\boldsymbol{u_1}$ for X and Y are obtained by projecting the observations onto the lines (Figures 10 C and 10D).

Because PLS reduces the dimension of the variables to a fewer components called scores, it performs well in the classification of data that have more variables than observations. This is a big advantage of PLS compared to other multivariate methods like multivariate regressions. It can deal with multicollinearity because many variables correlated to each other are reduced to a fewer scores that are not correlated. It does not rely on assumption of normality, either. The advantage found in PLS for analyzing high dimensional data is useful in my protein classification application because my approach involves dealing with high dimensional data.

There have been studies that used PLS in bioinformatics. Lapnish et al.[35] developed method based on PLS for classifying GPCRs into their different subfamilies. As the descriptors, they used five principal components derived from 26 physico-chemical properties of amino acids originally developed by Sanberg et al.[37]. Their PLS method was able to classify a dataset of 929 Class A GPCRs into their subfamilies except for a few orphan receptors, which would not be separated from peptide receptors. When the method was tested on a dataset of 535 GPCRs not included in a training dataset, only 14

GPCR sequences were misclassified. Nguyen and Rocke[38] used PCA and PLS to reduce the dimensions of five different datasets involving various tumor samples based on microarray gene expression experiments. The five datasets were then classified using logistic (LOG) and quadratic discriminant analyses (QDA). They found that LOG and QDA classifiers based on PLS data reduction performed better in the classifications of human tumor cells than those classifiers based on PCA data reduction.

1.4.5 Discriminant function analysis

The goal of discriminant function analysis (DA) is to predict group membership from a set of predictors. The purpose is to determine the class of an observation based on a set of variables known as predictors or input variables. The model is built based on a set of observations (training set) for which the classes are known. Based on the training set, linear discriminant analysis (LDA) constructs a set of linear functions of the predictors, known as discriminant functions, such that $L = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$, where the b's are discriminant coefficients, the x's are the input variables or predictors and c is a constant. These discriminant functions are used to predict the class of a new observation with unknown class. For a k class problem, k discriminant functions are constructed. Given a new observation, all the k discriminant functions are evaluated and the observation is assigned to class i if the i^{th} discriminant function has the highest value.

LDA is one of the methods known as parametric method. In parametric methods, the variables should be normally distributed within groups. On the other hand, the dependent variables in non-parametric methods are not necessarily normally distributed. In LDA, the variances and covariances in the different groups or classes are identical; this assumption is called the homogeneity of variances or covariances assumption. LDA is illustrated graphically in Figure 11. A dataset of ten proteins is grouped into two classes A and B. Using two variables (mass and volume), these samples are plotted and a discriminating line is plotted to separate the two classes (Figure 11B). We can observe that there is a separation between the two classes, but there is an overlap in the distributions (volume variables between) the two classes. A discriminating line can be chosen so that it maximizes the differences between the means of the two classes on the projection line that is perpendicular (Figure 11C). It can be shown that on the projected line, there is no overlap between the two normal distributions of the variables in the two classes (Figure11D).

Other types of DA methods used in protein classifications are quadratic discriminant analysis (QDA), logistic discrmininant analysis (LOG), and K Nearest Neighbors discriminant analysis (KNN). QDA is similar to LDA, where the variables follow the assumption of normality. Unlike LDA however, in QDA there is no assumption that the covariance of each of the classes is identical. As with LDA, the disadvantage of QDA is that it works well when variables are normally distributed. In LOG, the dependent variables take interger values of ordered catergories of groups. The method assumes that the posterior probabilities of group memebrship follow a logistic model.

KNN is a non-parametric DA method. The method is implemented as follows.

1. Assign training to known classes.
2. Calculate the distance of the unknown to all members of the training set using ‘Euclidean’ distances.
3. Rank these in order from the smallest distance.

4. Pick K smallest distances.
5. Take the majority vote for classification of the unknown.

KNN is a simple approach, and it has some limitations. The method takes no account of the spread of variance in a class, each variable assumes equal significance, and the numbers in each class of the training set should be approximately equal.

DA methods have been used for protein classification problems. Kim et al.[39] described a new protein classification method based on a nonparametric variant of linear discriminant analysis for classifying GPCRs from other proteins. Their method performed better than alignment-based methods (e.g., Pfam, PROSITE) by identifying 97% of the GPCRs with 2.1% false positives from the training dataset. On a test dataset, the method identified 99% of GPCRs with 0% false positives. The new method performed better when tested on short-sequences. In Moriyama and Kim[40], they further compared the performance among parametric and nonparametric discrimination methods (LDA, QDA, LOG, and KNN). Both parametric and nonparametric methods performed similar to or better than their previous method[39] with true positive rates at 98.0-98.7%, and false positive rates at 2.8-3.6%. Chou and Elrod[41] used covariant-discriminant analysis to classify 566 class A GPCR sequences into its subfamilies. Covariant-discriminant analysis uses Mahalanobis distance for discriminating groups. Chou and Elrod[42] used covariant-discriminant analysis in predicting enzyme family classes.

1.4.6 Support vector machines

Support vector machines (SVMs)[34] 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 points involves determining on which side of the hyperplane that points lie. SVMs have wide applications because of their use of kernel functions to represent data. The kernel function defines similarities between remapped data points. Some commonly used kernel functions are: linear, polynomial, radial basis, and sigmoid functions represented by equations 7, 8, 9, and 10, respectively.

$$\text{Linear Kernel: } K(x, y) = (x \cdot y + 1) \tag{7}$$

$$\text{Polynomial Kernel: } K(x, y) = (x \cdot y + c)^{p} \tag{8}$$

$$\text{Sigmoid Kernel: } K(x, y) = \tanh(x \cdot y + c) \tag{9}$$

$$\text{Radial Basis Kernel: } K(x, y) = e^{-\gamma||x-y||2} \tag{10}$$

where $K(x, y)$ is a kernel function, x and y are input vectors, p is the degree of polynomial, c and γ are parameters.

As an example, we use a dataset of ten proteins from two classes as shown in Figure 12A. Figure 12B is a 2-dimensional plot of the ten proteins before projecting into a higher dimension. We can observe that the two classes are not separate. After projecting the input data into a 3-dimensional space, they can be separated using a hyperplane (Figure 12C). A new unknown protein can be classified by determining which side of the hyperplane that protein falls into. In Figure 12C, the new protein indicated by a green diamond fell on the side of the class 1 protein (red).

One advantage of SVMs is that it can be used to classify linearly separable data as well as nonlinearly separable data. SVMs employ kernel functions. These kernel functions transform the data to a higher dimensional space where they can be linearly separable. One disadvantage of SVMs is that their performance is closely tied to the choice of optimal kernel functions. Currently, the methods of choosing kernels are based on the knowledge of the input data and there has not been any standardized method to obtain the best kernel function. The choice of the optimal kernel function is largely a trial and error procedure.

Karchin et al.[43] used SVMs with a kernel function obtained on profile HMMs. Their results showed that SVMs could classify GPCR subfamilies within the superfamily better than a profile HMM method. Liao and Noble[44] used pairwise similarity scores as input vectors, and their method performed better than profile HMMs and SVMs used by Karchin et al.[43] for discriminating SCOP protein families[45]. Bhasin and Raghava[46] used SVMs with amino acid and dipeptide compositions to classify GPCRs from non-GPCRs. Strope and Moriyama[47] applied SVMs with amino acid compositions for GPCR classification problems, and showed that such classifiers outperformed profile HMMs and decision trees for discriminating GPCRs from non-GPCRs.

1.4.7 Problems with the current alignment-free methods

Alignment-free methods have good sensitivity in detecting sequences with weak similarities. However, they have in general high false positives. Moreover, some of the alignment-free methods are computationally expensive. For example, if we use 5 descriptors for each amino acid in a protein sequence and we apply auto and cross-covariance with the maximum lag of 30 to transform the data (used in Lapnish et al[35]; explained in Chapter 2, section 2.1), each sequence will have 775 descriptors. With that many descriptors, it may take several hours or even days to mine a protein family from a complete genome of an organism.

1.5 PROTEIN FAMILIES USED IN THIS STUDY

1.5.1 Protein families

As mentioned before, the complete genomes are available only from four plants. Most of the sequences found in the databases are mainly from animals and other organisms rather than plants. When unknown plant sequences are used to search databases using alignment-based methods, their functions are inferred from the proteins that have highest scores from the databases. Most of the plant proteins are therefore annotated based on sequence similarities from animals and other organisms other than plants. Some proteins may have similar structure and biochemical functions but have low similarities between plants and other organisms (e.g., G-protein coupled receptors). Such proteins may be missed by the alignment-based search methods, and they are failed to be annotated. A second problem is that, the numbers of sequences from certain protein families found in databases from plants are very few. For example, currently there are only four known cytochrome b561 sequences from *Arabidopsis* found in the databases.

Protein families used in this study were selected to address such problems. These proteins are either directly involved with biotic and abiotic plant stress responses, or are involved with other proteins in plants stress related responses. G-protein coupled receptors, cytochrome b561, and cyclophilin protein families were chosen for this study. These protein families are described below.

1.5.2 G-protein coupled receptors

G-protein coupled receptors (GPCRs), also known as seven transmembrane receptors (7TMRs), are transmembrane proteins that via heterotrimeric guanine nucleotide-binding proteins (G-proteins) initiate some of the most important signaling pathways in the cell. They have seven transmembrane regions connected by three intracellular and three extracellular loops as presented in Figure 13. Members of the GPCR superfamily includes receptors of e.g., light, hormones, neurotransmitter, odorants, and taste molecules[48; 49]. GPCRDB (Information System for G Protein-Coupled Receptors)[50] divides the superfamily into six major classes A, B, C, D, E, and Frizzled/Smoothened (Table 1). Class A is by far the largest GPCR class with more than 4,000 entries in the database. Other minor families of GPCRs include: "Vomeronasal receptors”, “Plant Mlo receptors”, and "Taste receptors" (Table 1). The GPCR sequences of different classes/families are highly diverged from each other. Their lengths are also varied especially in the N and C-terminal as well as loop regions. Such high variation makes reconstructing reliable multiple alignments across families or from the entire GPCR superfamily very difficult.

(a) Ligand binding and signal transduction

G-proteins are composed of alpha (Gα), beta (Gβ), and gamma (Gγ) subunits[51]. Before a ligand binds to a GPCR, the three subunits of G-proteins are bound together. The ligand binding to the GPCR results in the conformational changes that stimulate exchange of guanosine triphosphate (GTP) for guanosine diphosphate (GDP) at the guanine-nucleotide binding site at the Gα subunit. GTP binding disrupts Gα interaction with Gβ, thereby freeing both Gα subunit and Gβγ pair to interact with variety of downstream effectors [52]. GTPase activity of the Gα subunit, accelerated by RGS (regulator of G-protein signaling) returns the G-protein to the inactive state[51;52]. However, an increasing number of alternative G protein-independent signaling mechanisms, especially in plants have been associated with groups of 7TMR proteins. Therefore, in this dissertation, GPCRs as well as 7TMRs will be used.

(b) G-protein coupled receptors in plants

A large number of GPCRs have been identified in animals. The human genome has 800 or more GPCRs[53], 557 GPCRs are found in chicken[54], about 300 more are found in the *Drosophila melanogaster*[53]*,* and more than 1000 are found in the *Caenorhabditis elegans*[55]. Compared to animal genomes, very few GPCRs have been identified in plants and fungi. For example, only twenty two GPCRs have been described in the *Arabidopsis thaliana* genome[56;57;58]. Fifteen of them constitute the “mildew resistance O” (MLO) family. The numbers of MLO proteins found in major crops are: nine in maize and eleven in rice. It shows that currently we do not have sufficient information from GPCRs from plants.

(c) Functions of GPCRs in plants

G-protein signaling components have been found in several plant species[59] and shown that G-protein-mediated signaling plays important roles in a wide range of plant processes including seed germination and seedling growth responses to light, phytohormones, ozone, sugars, pathogen resistance, etc, reviewed in Jones and Assmann[60]. *Arabidopsis* GPCR protein (GCR1) positively regulates seed germination. It is also possible that *Arabidopsis* RGS1 is a D-glucose receptor because *Arabidopsis* seedlings lacking AtRGS1 have reduced sensitivity to D-glucose [61; 62].

In summary, the GPCR superfamily will be useful for this specific study because of the weak sequence similarities among families, weak similarities between sequences of plants and animals, and currently, there are a very few numbers of GPCRs found in plants. On the agronomic and crop science ground, G-protein signaling components play important roles in a wide range of plant processes including pathogen resistance. Thus, this protein is significant in the field of agronomy and crop science.

1.5.3 Cyclophilins

Cyclophilins (CYPs) are the family of proteins that possess the peptidyl-prolyl isomerase (PPIase) activity. They are present in both eukaryotes and prokaryotes. They are involved in a wide variety of functions in cellular processes including cell-cycle control, protein trafficking, receptor signaling, as well as cellular targets of immunosuppressant drugs. There are 19 cyclophilins found in the human genome. On the contrary, 29 cyclophilins have been found in the *Arabidopsis thaliana* genome, which is currently the largest family of cyclophilins found in completed genomes (Table 2). However, the number of cyclophilins from plants found in Integrated Documentation Resource of Protein Families, Domains and Functional Sites (InterPro[63]; Release 14.1, dated, February 19th, 2007) is small compared to that of animals and other organisms (Table 3). All cyclophilin related proteins share a common domain of 160 amino acids, the cyclophilin domain. Although amino acid sequence similarities within each group of cyclophilins are highly conserved from bacteria to plants to animals, sequence similarities vary widely between different cyclophilin groups (10 to 90%). Heterogeneous domain architecture among its members increases the complexity (Figure 14). Some exist as single domain cyclophilins and others have multiple different domains[64; 65; 66]. Because of having multiple-domains, their sequences are not easily aligned. This makes it more difficult for researchers to mine complete sets of cyclophilin-related proteins from diverse plant genomes.

Cyclophilins (CYPs) are originally identified as cellular targets of cyclosporine A (CsA), a fungal metabolite with potent immunosuppressive activity[67]. With the FK506-binding proteins (FKBPs), they form a family of immunosuppressant receptors, immunophilins. Both groups have peptidyl-prolyl isomerase (PPIase, EC 5.2.1.8) activity that catalyzes the rapid *cis* to *trans* isomerization of peptide bonds N-terminal to proline residues in the polypeptide chains[65]. This *cis-trans* isomerazation is an important step in protein folding, and a critical determinant of protein structures.

While amino acid sequences of single-domain cyclophilins are in general divergent, their secondary structures remain well conserved[68]. The structure of cyclophilin proteins consists of eight stranded anti-parallel β-sheets capped at both ends by two helices. Numerous insertions/deletion are observed in the loop regions (Figure 15). However, the amino acid residues crucial for the PPIase activity and CsA-binding are well conserved even in the long loop regions (Figure 15)[69; 70]. Most of the amino acids involved with the PPIase activity are also known to be important for CsA-binding.

(a) Cyclophilin functions

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. While the peptidyl-prolyl *cis-trans* isomerase (PPIase) activity is important for protein folding, it has formed the basis for many complex interactions between cyclophilins and other proteins. Some of the cyclophilin functions include protein trafficking and maturation[71], receptor complex stabilization[72], apoptosis[73], receptor signaling[64], and plant-pathogen defense mechanism[74].

In plants, six of the *Arabidopsis* cyclophilins are localized in the chloroplast (reviewed in Romano et al.[75]). Five of them are single-domain cyclophilins (domain architecture is explained in the next section). AtCYP38 has multiple multiple-domains and is located in the thylakoid membrane. In spinach homologue TLP40 has been shown to play an important role in the protection of photosynthetic membranes by inducing dephosphorylation of damaged photosystem II D1 and thus inducing proteolysis[76;77].

Cyclophilins have been shown to play roles in both plant and animal pathogen recognition. Deng et al.[78] reported the interaction of *Agrobacterium tumefaciens* virulence protein (VirD2) with *Arabidopsis* cyclophilin AtCYP19. *Agrobacterium* recruits plant cyclophilins for transferring and integrating T-DNA (DNA fragment) into a plant cell. A recent study by Coaker et al.[74] identified another *Arabidopsis* cyclophilin AtCYP18 by its PPIase activity activating *Pseudomonas syringae* effector protein (AvrRpt2) after it is delivered into a plant cell. 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[79].

Antifungal proteins serve a protective function against fungal invasion. Cyclophilins have been purified from seeds of cow pea, mung bean and chickpea[80; 81], and they possess antifungal activity against several fungi including *Mycosphaerella arachidichola*. The chickpea protein also inhibits human immunodeficiency virus-1 reverse transcriptase.

(b) Cyclophilin protein family

Cyclophilin proteins are named according to organisms and their molecular weight. For example, AtCYP18 is an *Arabidopsis* cyclophilin protein with a molecular weight of 18kDa. As illustrated in Figure 14, cyclophilin proteins are classified into single-domain and multiple-domain groups. The single-domain cyclophilins contain only the cyclophilin catalytic domain, and their average length is 172 amino acids (aa) ranging from 147 aa to 210 aa. On the other hand, the multiple domain cyclophilins have other functional domains in addition to the cyclophilin catalytic domain. Their lengths range from 360 aa to 970 aa with the average of 550 aa. The other domains are expected to play roles in determining specific cyclophilin functions. For example, the “tetratricopetide (TPR) domain” is involved in protein-protein interactions.

Cyclophilin protein families were selected for this study because plants have the largest numbers of cyclophilins, but at the numbers of cyclophilins from plants in databases are small compared to that of animals and other organisms. Secondly, cyclophilin proteins have heterogeneous domains that make them difficult to mine by the alignment-based methods. And lastly, cyclophilins are involved in many cellular processing including plant-pathogen defense mechanisms which are significant in the field of agronomy and crop science.

1.5.4 Cytochrome b561

Cytochrome b561 (Cyt b561) is a transmembrane protein family with well conserved heme-binding histidine residues[82] (Figure 16). They are highly hydrophobic proteins with six transmembrane helices, four conserved His residues, possibly coordinating two heme molecules, and predicted substrate-binding sites for ascorbate and monodehydro-ascorbate (MDHA). The protein sequence similarities of Cyt b561 is low within and among species 34-45% within species, and around 30% among species) [83]. Cytochrome b561 is also present in the membranes of chromaffin granules and other neuroendocrine secretory vesicles. This cytochrome functions as a unique transmembrane electron transfer protein by mediating the transfer of electrons from a soluble cytoplasmic donor (ascorbate) across a membrane bilayer to a soluble intravesicular acceptor (semidehydroascorbate). Some of the Cytb561s are linked to other domains such as dopamine-hydroxylase (DOH) domain[84] (Figure 17). A possible link of Cyt b561 to neurodegenerative diseases, such as Parkinson's, Alzheimer's and Huntington's diseases has been postulated[84].

Recently Tsubaki et al.[85] classified Cyt b561 into seven subfamilies based on motifs. Group A (animals/neuroendocrine) subfamily is found in animals and group B found in plants. Both groups have motif 1 {FN(X)HP(X)$_2$M(X)$_2$G(X)$_5$G(X)ALLVYR} and motif 2 {YSLHSW(X)G}. Group C is found in insects while group D is found in fungi. There are no significant features characteristic to groups C and D. Group E is a multiple-domain Cyt b561 found in animals and group F is a multiple domain found in plants. Group E has modified version of motif 1 LFSWHP(X)$_2$M(X)$_3$F(X)$_3$M(X)EAIL(X)SP(X)2SS}and group F has motif 3 {DP(X)WFY(L)H(X)$_3$Q} and motif 4 {K(X)R(X)YWN(X)YHH(X)$_2$G(R/Y)}. Group G is a multiple domain Cyt b561 with the longest length on average. The longer size is due to multiple DOH domains.

As mentioned before, the protein sequence similarity of Cyt b561 is very low. Because of low sequence identity and the presence of multiple domains, these proteins are not easy to align, hence making it more difficult to mine from diverse plant genomes. For example, there are only four known Cyt b561s found in the *Arabidopsis* genome. The low sequence similarity combined with multiple domains make cyt b561 a good protein family for this study. Secondly, Cyt b561 transports electrons to MDAH which plays a role in the regeneration of ascorbate. Ascorbate in plants plays a role in antioxidative defense reactions[82]. This is important in the field of Agronomy and Crop science.

1.6 REFFERENCES

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol. 1990; 215:403-410.
2. The rice chromosomes 11 and 12 Sequencing Consortia. BMC Biology 2005;3:20.
3. Durbin R, Eddy S, Krogh A, and Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press: Cambridge; 1998.
4. Dutton HJ, Lancaster CJ, Evans CD, and Cowan JC. The flavor problem of soybean oil. VIII. Linolenic acid. J. Am. Oil Chem. Soc. 1951;28:115-118.
5. Lui HR, and White PJ. Oxidative stability of soybean oils with altered fatty-acid compositions. J. Am. Oil Chem. Soc. 1992; 69:528-532.
6. Hu FB, Stampfer JE, Manson JE, Rimm E, Colditz GA, Rosner BA, Hennekens CH, and Willett WC. Dietary fat intake and risk of coronary heart disease in women. N. Engl. J. Med. 1997;337:1491-1499.
7. Fehr WR, Welke GA, Hammond EG, Duvick DN, and Cianzio SR. Inheritance of reduced linolenic acid content in Cianzio. Soybean genotypes A16 and A17. Crop Sci. 1992;32:903-906.
8. Byrum JR, Kinney AJ, Stecca KL, Grace DJ, and Diers BW. Alteration of the omega-3 fatty-acid desaturase gene is associated with reduced linolenic acid in the A5 soybean genotype. Theor. Appl. Genet. 1997; 94:356-359.
9. Bilyeu KD, Palavalli L, Sleper DA, and Beuselinck PR. Three Microsomal omega-3 fatty-acid desaturase genes contributeto soybean linolenic acid levels. Crop Sci. 2003;

43:1833-1838.

10. Yadav NS, Wierzbicki A, Aegerter M, Caster CS, Perez-Grau L, Kinney AJ, Hitz WD, Booth JR, Schweiger B Jr ,and Stecca KL. Cloning of higher plant omega-3 fatty-acid desaturases. Plant Physiol. 1993;103:467-476.
11. Kantety RV, La Rota M, Matthews DE, Sorrells ME. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol. 2002;48:501-10.
12. Li Y, Rosso MG, Strizhov N, Viehoever P, Weisshaar B. GABIKat Simple Search: a flanking sequence tag (FST) database for the identification of T-DNA insertion mutants in *Arabidopsis thaliana*. Bioinformatics 2003;19:1441-1442.
13. Monica V, Fang T, James RA, Robin B et al. Bioinformatics-Enabled Identification of the HrpL Regulon and Type III Secretion System Effector Proteins of *Pseudomonas syringae* pv. *phaseolicola* 1448A. MPMI. 2006;19:1193-1206.
14. Kevin AT, Silverstein KAT, Graham MA, Paape TP, and VandenBosch KA. Genome organization of more than 300 defensin-Like genes in Arabidopsis. Plant Physiol. 2005;138:600-610
15. Mayer KM and Shanklin J. Identification of amino acid residues involved in substrate specificity of plant acyl-ACP thioesterases using a bioinformatics-guided approach. BMC Plant Biology 2007;7:1.
16. Gai X, Lai S, Xing L, Brendel V, and Walbot V. Gene discovery using maize genome database ZmDB. Nucl Acids Res. 2000;28:94-96.

17. Shoemaker R, Keim P, Vodkin L, Retzel E, Clifton FW, Waterston R, Smoller D, Coryell V, Khana A, and Erpelding J. A compilation of soybean ESTs: Generation and analysis. Genome. 2002;45:329-338.

18. Lazo GRS, Chao SDD, Hummel H, Edwards CC, Crossman N, Lui DE, Matthews V L, Carollo DL, Hane FM, You GE, Butler RE, Miller TJ, Close JH, Peng NLV, Lapitan JP, Gustafson LL, Qi B, Echalier BS, Gill M, Dilbirligi D, Sandhu D, Gill KS, et al. Development of an expressed sequence tag (EST) resource for wheat (Triticum aestivum L.): EST generation, unigene analysis, probe selection, and bioinformatics for a 16,000-locus bin-delineated map.Genetics 2004;168:585-593.

19. Boguski MS, Lowe TM, and Tolstoshev CM. dbEST-database for "expressed sequence tags". Nat Genet. 1993;4:332-333 (http://www.ncbi.nlm.nih.gov/dbEST/index.html).

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

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

22. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 1991;11: 635-650.

23. Smith TF and Waterman MS. Identification of common molecular subsequences. J. Mol. Biol. 1981;147:195-197.

24. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 1991:11; 635-650.

25. Hulo N, Sigrist CJA, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, and Bairoch A. Recent improvements to the PROSITE database. Nucleic Acids Res. 2004:32;D134-D137.

26. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine K, Taylor P, Uddin A, and Zygouri C. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 2003;31:400-402.

27. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech. Proceeding of the IEEE. 1989:77; 257-286

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

29. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, and Bork P. SMART 4.0: towards genomic data integration. Nucleic Acids Res 2004;32: D142-144.

30. Hughey R and Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. CABIOS 1996;12:95-107.

31. Gough J, Karplus K, Hughey R, and Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known

structure. J Mol Biol. 2001;313:903-19.

32. Tabachnik GB and Fidell SF. Using multivariate statistics. 4th edition; 2001, Allan and Bacon; Needham Heights, MA.

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

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

35. Lapnish M, Gutcaits A, Prusis P, Post C, Lundstedt T, and Wikberg JES. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci. 2002;11: 795-805.

36. Gunnarsson I, Andersson P, Wikberg J and Lundstedt T. Multivariate analysis of G protein-coupled receptors. J Chemometrics. 2003;17:82-92.

37. Sanberg M, Eriksson L, Jonsson J, and Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J. Med. Chem. 1998;41:2481-2491.

38. Nguyen DV and Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002;18:39-50.

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

40. Moriyama EN, and Kim J. Protein family classification with discriminant function analysis. In Genome Exploitation: Data Mining the Genome, J.P. Gustafson, ed, Springer; New York: 2005.

41. Chou KC and Elrod DW. Bioinformatical analysis of G-protein coupled receptors. J. Proteome Res. 2002;1:429-433.

42. Chou KC and Elrod DW. Prediction of enzymes family classes. J Proteome Res. 2003;2:183-190.

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

44. Liao L and 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.

45. Murzin A, Brenner SE, Hubbard TJP, and Chothia C. SCOP: a Structural classification of Proteins database for investigation of sequences and structures. J Mol Biol. 1995;247:536-540.

46. Bhasin M and Raghava GPS. GPCRsclass: a web tool for the classification of amine type of G-protein-coupled receptors. Nucleic Acids Res. 2005;33:W143-W147.

47. Strope PK and Moriyama EN. Simple alignment-free methods for protein classification: a case study from G-protein coupled receptors. Genomics 2007;89:602-612.

48. Bockaert J and Pin JP. Molecular tinkering of G-protein coupled receptors: an evolutionary success. EMBO J. 1999;18:1723-1729.

49. Pierce KL, Premont RT, and Lefkowitz RJ. Seven-transmembrane receptors. Nat Rev Mol Cell Biol. 2002;3:639-650.

50. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, and Vriend G. GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res. 2003;31:

294297.

51. Ross EM and Wilkie TM. GTPase-activating proteins for heterotrimeric G proteins: regulators of G protein signaling (RGS) and RGS-like proteins. Annu Rev Biochem. 2000;69:795-798.

52. Chen JG, Pandey S, Huang J, Alonso JM, Ecker JR, Assmann SM, and Jones AM. GCR1 can act independently of heterotrimeric G-protein in response to brassinosteroids and gibberellins in Arabidopsis seed germination. Plant Physiol 2004;135:907-915.

53. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G: GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res. 2003;31:294-297.

54. Lagerström MC, Hellstro¨m AR, Gloriam DE, Larsson TP, Schiöth HB, and Fredriksson R. The G protein–coupled receptor subset of the chicken genome. PLoS Comput Biol.2006;2:e54.

55. Bargmann CI: Neurobiology of the *Caenorhabditis elegans* genome. Science 1998;282:2028-2033.

56. Devoto A, Piffanelli P, Nilsson I, Wallin E, Panstruga R, von Heijne G, Schulze-Lefert P: Topology, subcellular localization, and sequence diversity of the Mlo family in plants.J Biol Chem 1999, 274:34993-35004.

57. Devoto A, Hartmann HA, Piffanelli P, Elliott C, Simmons C, Taramino G, Goh CS, Cohen FE, Emerson BC, Schulze-Lefert P, et al.: Molecular phylogeny and evolution of the plant-specific seven-transmembrane MLO family.J Mol Evol. 2003;56:77-88.

58. Josefsson LG, Rask L: Cloning of a putative G-protein-coupled receptor from

Arabidopsis thaliana. Eur J Biochem. 1997;249:415-420.

59. Assmann SM. G-protein Go Green. Science 2005;301:71-73.

60. Jones AM and Assmann SM. Plants: the latest model system for G-protein research. EMBO Rep 2004;5:572-578.

61. Chen JG, Willard FS, Huang J, Liang J, Chasse SA, Jones AM, and Siderovski DP A seven-transmembrane RGS protein that modulates plant cell proliferation. Science 2003;301:1728-1731.

62. Ullah H, Chen JG, Wang S, and Jones AM. Role of a heterotrimeric G protein in regulation of Arabidopsis seed germination. Plant Physiol. 2002;129: 897-907.

63. Mulder et al. InterPro, progress and status in 2005. Nucleic Acids Res. 2005;33: D201-D205.

64. Brazin KN, Mallis RJ, and Fulton DB. Regulation of the tyrosine kinase Itk by peptidyl-prolyl isomerase cyclophilin A. Proc Natl Acad Sci USA. 2002;99:1899-1904.

65. Miguel A, Xiaoyun W, Steven DH, and Joseph H. Prolyl isomerases in yeast. Frontiers in Bioscience 2004;9:2420-2446.

66. He Z, Li L, and Luan S. Immunophilins and parvulins. Superfamily of peptidyl propyl isomerases in *Arabidopsis*. Plant Cell Physiology 2004;134:1248-1267.

67. Handschumacher RE, Harding MW, Rice J, Drugge R J, and Spelcher DW. Cyclophilin: A specific cytosolic binding protein for cyclosporine A. Science 1984;226:544-547.

68. Galat A. Peptidylproplyl cis/trans isomerases (immunophilins): Biological diversity targets-functions. Curr Topic Med Chem. 2003;3:1315-1347.

69. Galat A. Variations of sequences and amino acid compositions of proteins that sustain thier biological functions: an analyis of the cyclophilin family of proteins. Arch Biochem Biophys. 1999;371:149-162.

70. Dornan J, Taylor P, and Wilkinshaw MD. Structures of immunophilins and their ligand complexes. Curr Topic Med Chem. 2003;3:1392-1409.
4:1413-1418.

71. Shieh BH, Stamens MA, Seavello S, Hariss GL, and Zuka CS. The NinaA gene required for visual transduction in *Drosophila* encodes a homolog of cyclosporin A-binding protein. Nature 1989;338:67-70.

72. Leverson JD and Ness SA. Point mutations in v-Myb disrupts cyclophilin-catalysed negative regulatory mechanism. Mol Cell. 1998;1:203-211.

73. Lin DT and Lechleiter JD. Mitochondria targeted cyclophilin D protects cell from death by petidyl prolyl isomerization. J Biol Chem. 2002;277:31134-31141.

74. Coaker G, Falick A, and Staskawicz B. Activation of a phytopathogenetic bacteria effector Protein by eukaryotic cyclophilin. Science 2005;308:548-550.

75. Romano PG, Horton P, and Gray JE. The Arabidopsis cyclophilin gene family. Plant Physiol. 2004b;134:1268-1282.

76. Fulgosie H, Vener AW, Altshmied L, Hermann RG, and Andersson B. A novel multi-functional chloroplast protein: Identification of a 40 kDa immunophilin-like protein located in the thylakoid lumen. EMBO J. 1998:17:1577-1587.

77. Vener AV, Rokka A, Fulgosi H, Andersson B, and Herrmann RG. A cyclophilin-regulated PP2-A like protein phosphatase in thylakoid membrane of plant chloroplasts. Biochemistry 1999;38:14955-14965.

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

79. Muriel CV, Pascale VB, and Nicholas, JT. A *Magnaporthe grisea* cyclophilin acts as a virulence determinant during plant infection. The Plant Cell 2002;14:917-930.

80. Ye XY and Ng TB A novel cyclophilin-like antifungal protein from the mung bean. Biochem.Biophys.Res.Commun. 2000;273:1111-1115.

81. Ye XY and Ng TB. Isolation of new cyclophilin-like protein from chicken peas with mitogenic, antifungal and anti-HIV reverse transcriptase activity. Life Sci. 2002;70:1128-1138.

82. Asard H, Kappila J, Verelst W, Bérczi A. Higher-plant plasma membrane cytochrome b561: A protein in search of a function. Protoplasma 2001;217:77-93.

83. Verelst W and Asard H. A phylogenetic study of cytochrome b561 proteins. Genome Biology 2003;4:R38.

84. Ponting CP. Domain homologues of dopamine b-hydroxylase and ferric reductase: roles for iron metabolism in neurodegenerative disorders. Hum. Mol. Genet. 2001;10:1853-1858.

85. Tsubaki M, Takeuchi F, Nakanishi N. Cytochrome b561 protein family: Expanding roles and versatile transmembrane electron transfer abilities as predicted by a new classification system and protein sequence motif analyses. Biochim Biophys Acta. 2005;1753:174-200.

1.7 TABLES

Table 1. The major GPCR classes from GPCRDB database (June 2006, release 10.0)

Classes	Families
Class A Rhodopsin like	Amine
	Peptide
	Hormone protein
	(Rhod)opsin
	Olfactory
	Prostanoid
	Nucleotide-like
	Cannabinoid
	Platelet activating factor
	Gonadotropin-releasing hormone
	Thyrotropin-releasing hormone and Secretagogue
	Melatonin
	Viral
	Lysosphingolipid
	Leukotriene B4 receptor
	Class A Orphan/other
Class B Secretin like	Calcitonin
	Corticotropin releasing factor
	Gastric inhibitory peptide
	Glucagon
	Growth hormone-releasing hormone
	Parathyroid hormone
	PACAP
	Secretin

Table 1 (continued)

Classes	Families
Class B Secretin like	Vasoactive intestinal polypeptide
	Diuretic hormone
	EMR1
	Latrophilin
	Brain-specific angiogenesis inhibitor (BAI)
	Methuselah-like proteins (MTH)
	Cadherin EGF LAG (CELSR)
	Very large G-protein coupled receptor
Class C Metabotropic glutamate / pheromone	Metabotropic glutamate
	Calcium-sensing like
	Putative pheromone receptors
	GABA-B
	Orphan GPRC5
	Orphan GPCR6
	Bride of sevenless proteins (BOSS)
	Taste receptors (T1R)
Class D Fungal pheromone	Fungal pheromone A-Factor like (STE2,STE3)
	Fungal pheromone B like (BAR,BBR,RCB,PRA)
	Fungal pheromone M- and P-Factor
Class E cAMP receptors	

Table 1 (continued)

Classes	Families
Frizzled/Smoothened family	frizzled Smoothened
Putative families	Ocular albinism proteins Insect odorant receptors Plant Mlo receptors Nematode chemoreceptors Vomeronasal receptors Taste receptors T2R
Orphans	Putative / unclassified GPCRs

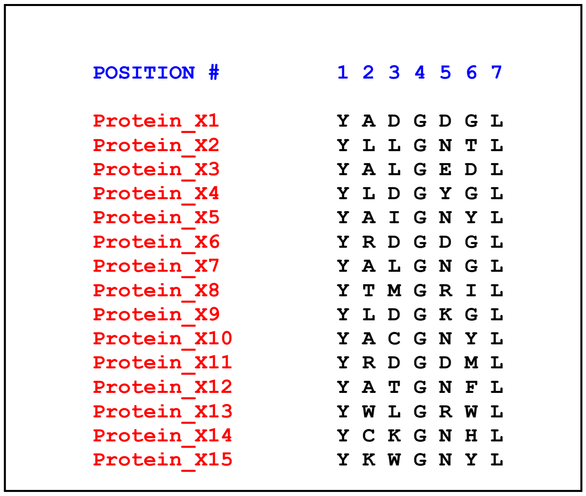
Table 2. The number of cyclophilin proteins in different genomes.

Organisms	Cyclophilins
Homo sapiens	19
Arabidopsis thaliana	29
Drosophila melanogaster	14
Saccharomyces cerevisiae	8
Clamydomonas reinharditi	28
Cyanidioschyzon merolae	4
Thalassiosira pseudonana	8
Escherichia coli	3

Table 3. The number of cyclophilin protein sequences found in the InterPro database (as of February, 2007).

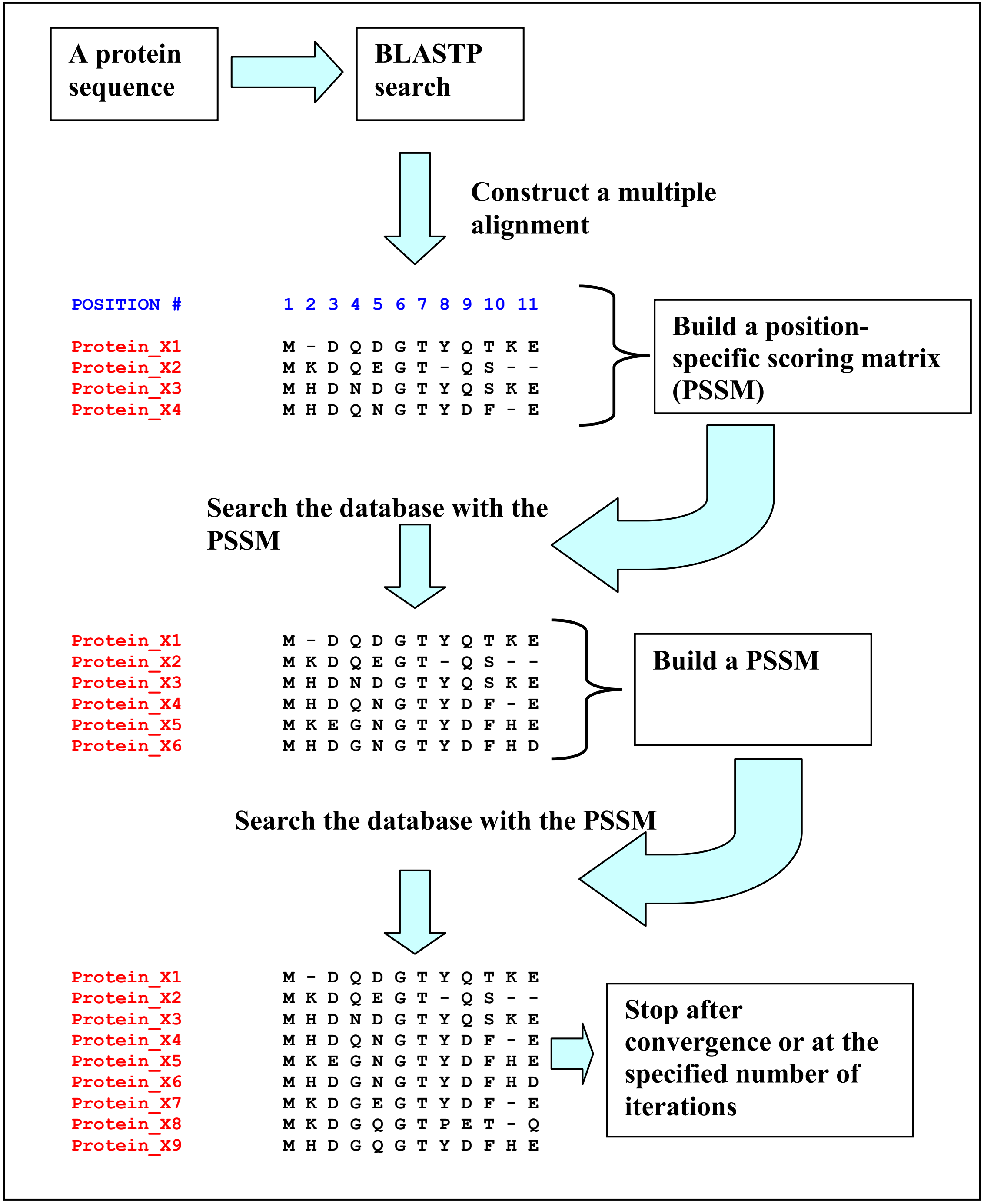
Organisms	Cyclophilins
Animals	738
Plants	202
Fungi	275
Bacteria	1106
Archaea	29
Virus	1

1.8 FIGURES



POSITION #	1	2	3	4	5	6	7
Protein_X1	Y	A	D	G	D	G	L
Protein_X2	Y	L	L	G	N	T	L
Protein_X3	Y	A	L	G	E	D	L
Protein_X4	Y	L	D	G	Y	G	L
Protein_X5	Y	A	I	G	N	Y	L
Protein_X6	Y	R	D	G	D	G	L
Protein_X7	Y	A	L	G	N	G	L
Protein_X8	Y	T	M	G	R	I	L
Protein_X9	Y	L	D	G	K	G	L
Protein_X10	Y	A	C	G	N	Y	L
Protein_X11	Y	R	D	G	D	M	L
Protein_X12	Y	A	T	G	N	F	L
Protein_X13	Y	W	L	G	R	W	L
Protein_X14	Y	C	K	G	N	H	L
Protein_X15	Y	K	W	G	N	Y	L

Figure 1. Fifteen sequences representing a motif from a protein family


A protein sequence
BLASTP search
Construct a multiple alignment
Build a position-specific scoring matrix (PSSM)
Search the database with the PSSM
Build a PSSM
Search the database with the PSSM
Stop after convergence or at the specified number of iterations

```
POSITION #          1 2 3 4 5 6 7 8 9 10 11

Protein_X1          M - D Q D G T Y Q T K E
Protein_X2          M K D Q E G T - Q S - -
Protein_X3          M H D N D G T Y Q S K E
Protein_X4          M H D Q N G T Y D F - E
```

```
Protein_X1          M - D Q D G T Y Q T K E
Protein_X2          M K D Q E G T - Q S - -
Protein_X3          M H D N D G T Y Q S K E
Protein_X4          M H D Q N G T Y D F - E
Protein_X5          M K E G N G T Y D F H E
Protein_X6          M H D G N G T Y D F H D
```

```
Protein_X1          M - D Q D G T Y Q T K E
Protein_X2          M K D Q E G T - Q S - -
Protein_X3          M H D N D G T Y Q S K E
Protein_X4          M H D Q N G T Y D F - E
Protein_X5          M K E G N G T Y D F H E
Protein_X6          M H D G N G T Y D F H D
Protein_X7          M K D G E G T Y D F - E
Protein_X8          M K D G Q G T P E T - Q
Protein_X9          M H D G Q G T Y D F H E
```

Figure 2. A flow chart of PSI-BLAST sequence similarity search using position-specific scoring matrices

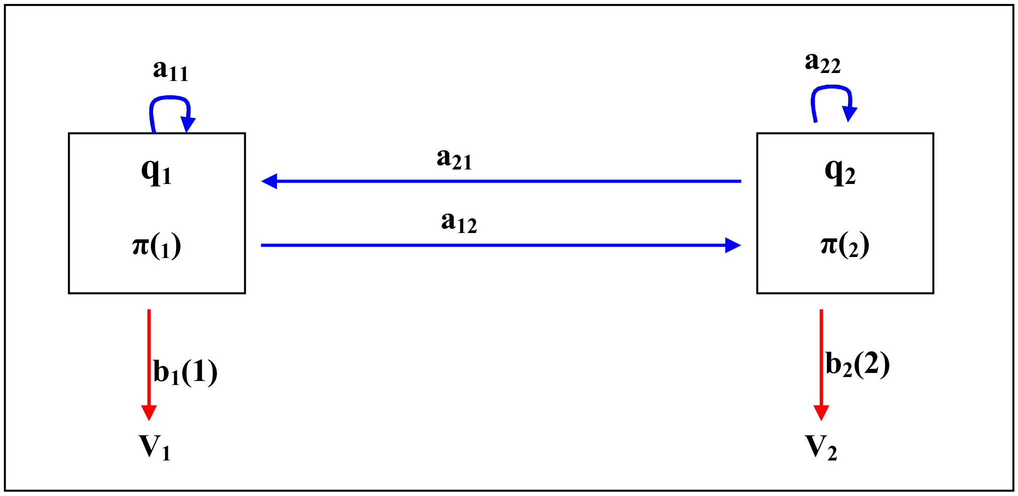


Figure 3. A simple two-state hidden Markov model. Each state has an associated emission probability distribution that determines which observation is emitted and transition probability distribution that determines which state will be visited next.

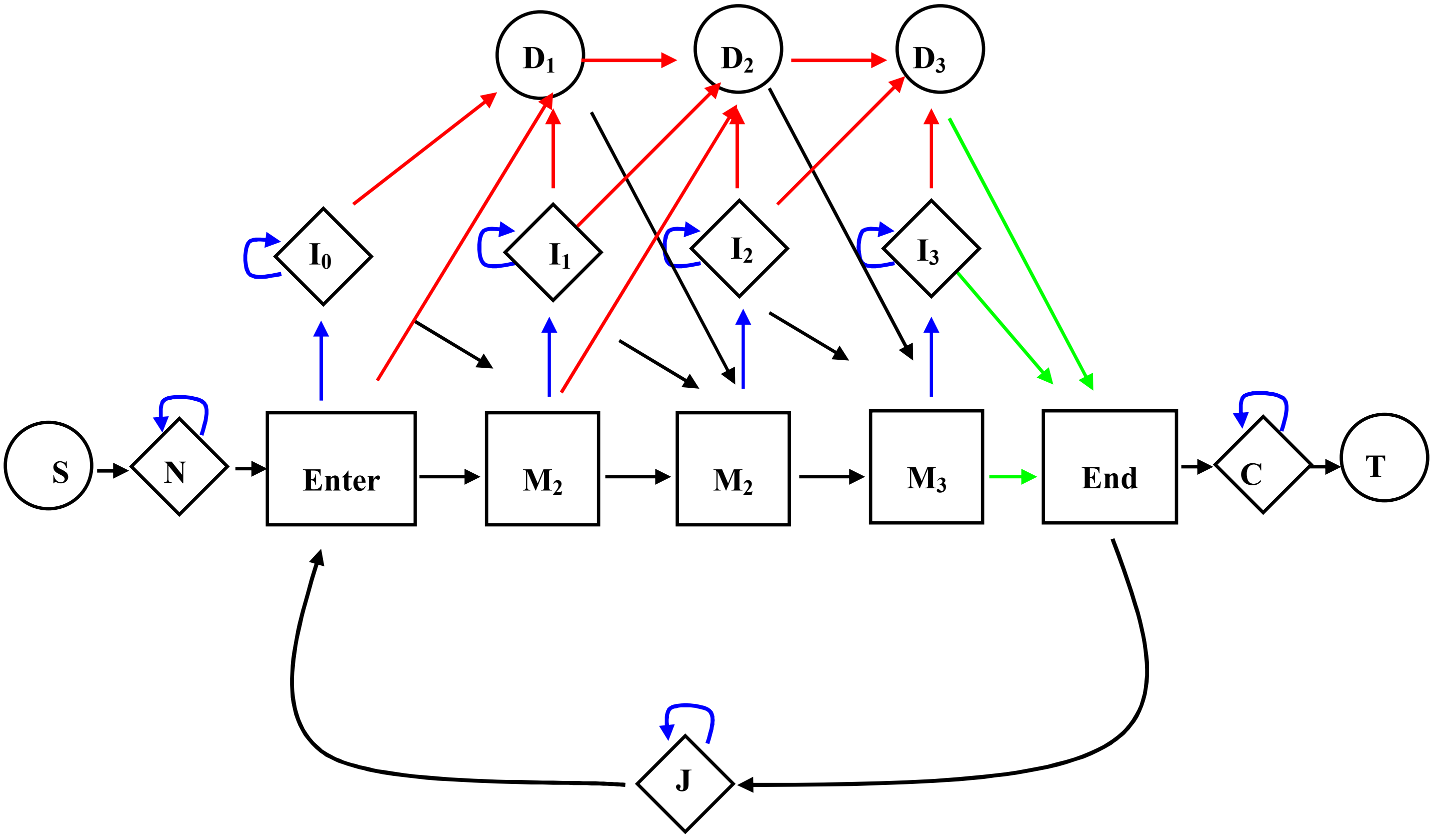


Figure 4. A hidden Markov model. Arrows indicate transition from state to state. Blue arrows = any state to insert state; black arrows = any state to match state; red arrows = any state to deletion state; green arrows = moving out of the model; M = match state; I = insert state; D = delete state; N = N-terminal unaligned sequence state; C = C-terminal unaligned sequence state; J = Joining segment unaligned sequence state; S = start, non emitter; and T stop non-emitter.

```
Protein_X1      NI-D
Protein_X2      N-GE
Protein_X3      N-GE
Protein_X4      N-GD
```

Alignment of four sequences

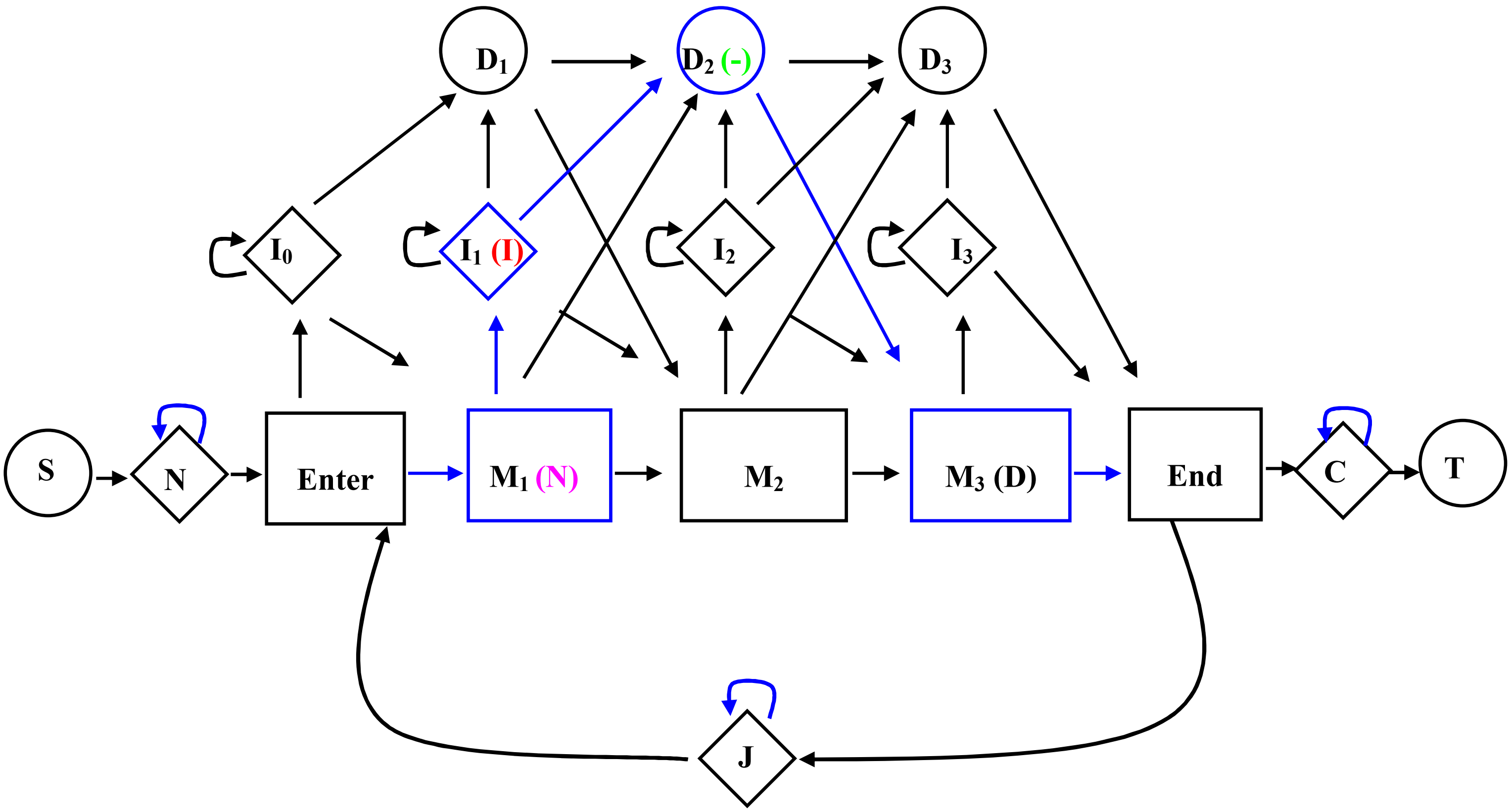
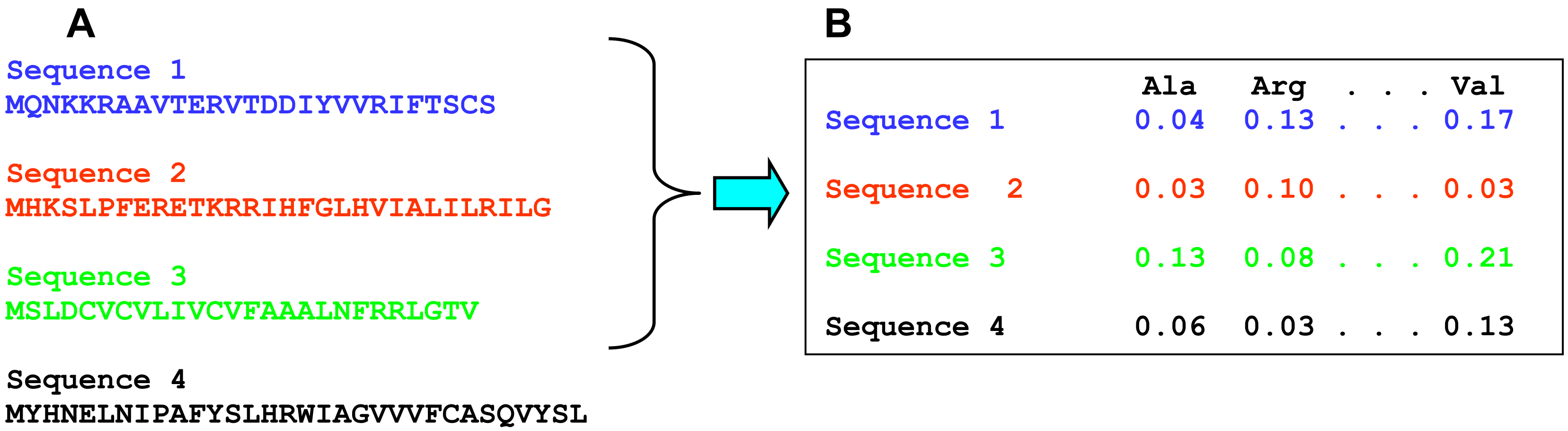


Figure 5. A hidden Markov model showing the path for Protein_X1 in the above alignment. The corresponding path and states are highlighted with blue arrows, blue boxes, a blue diamond, and a blue circle. Each amino acid from Protein_X1 is placed in each corresponding state.



Sequence 1
MQNKKRAAVTERVTDDIYVVRIFTSCS

Sequence 2
MHKSLPFERETKRRIHFGLHVIALILRILG

Sequence 3
MSLDCVCVLIVCVFAAALNFRRLGTV

Sequence 4
MYHNELNIPAFYSLHRWIAGVVVFCASQVYSL

	Ala	Arg	. . .	Val
Sequence 1	0.04	0.13	. . .	0.17
Sequence 2	0.03	0.10	. . .	0.03
Sequence 3	0.13	0.08	. . .	0.21
Sequence 4	0.06	0.03	. . .	0.13

Figure 6. An example of how protein sequences can be represented as amino acid sequence (A) and composition (B).

A

Protein	Mass	Volume	Isoelectric point (pI)
Protein_X1	124	99	4.3
Protein_X2	106	67	7.5
Protein_X3	111	90	9.2
Protein_X4	109	90	7.5
Protein_X5	113	112	9.6
Protein_X6	89	72	10.1
Protein_X7	78	112	7.7
Protein_X8	190	87	6.8
Protein_X9	123	68	7.6
Protein_X10	123	116	7.8

B

Protein	Mass	Volume	Isoelectric point (pI)
Protein_X1	0.248355	0.418024	-0.44814
Protein_X2	-0.35575	-1.31922	-0.03841
Protein_X3	-0.18794	-0.07058	0.179257
Protein_X4	-0.25507	-0.07058	-0.03841
Protein_X5	-0.12082	1.123779	0.230474
Protein_X6	-0.9263	-1.04777	0.294494
Protein_X7	-1.29548	1.123779	-0.0128
Protein_X8	2.463418	-0.23344	-0.12804
Protein_X9	0.214794	-1.26493	-0.02561
Protein_X10	0.214794	1.340934	0.00000

Figure 7. Pre-treatment of data for principal component analysis. The raw data set is shown in A, and the transformed data set by standardization is shown in B.

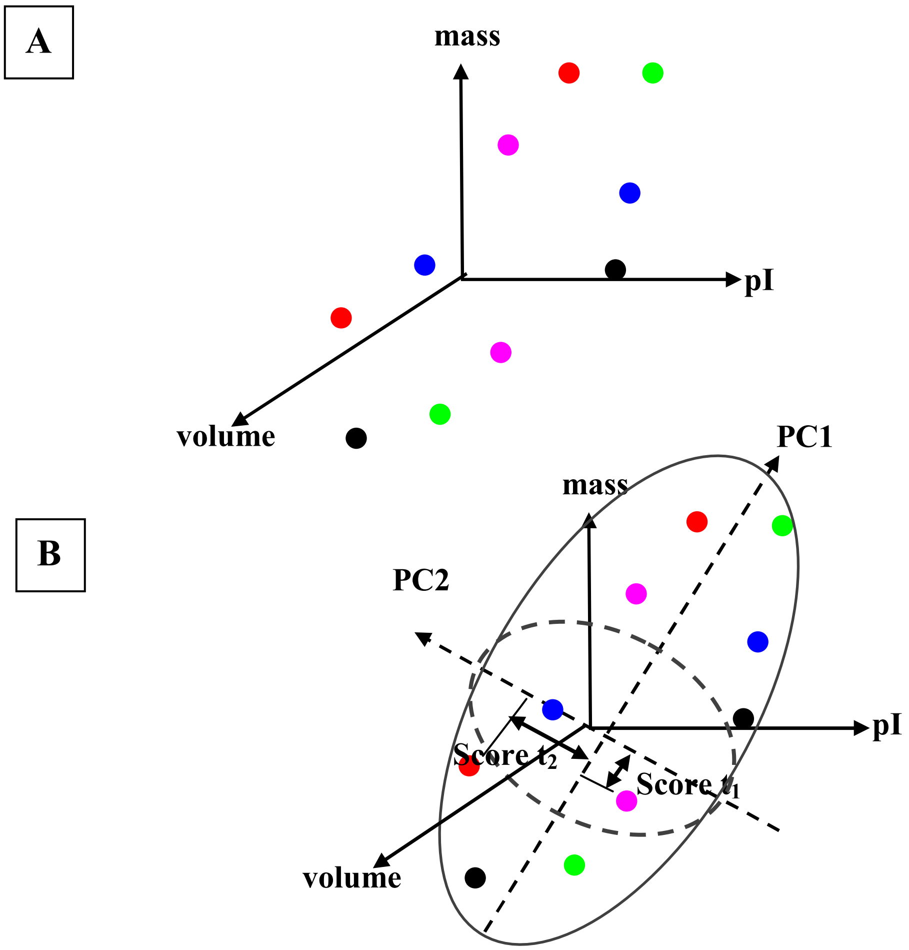


Figure 8. Geometrical representation of principal component analysis. Transformed data are plotted in the 3-dimensional space (A). Two principal components (PC1 and PC2) are shown in B. A short line perpendicular to PC1-axis shows the projection from each data point.

A

Protein	Y_1	Y_2	Y_3
Protein_X1	2.5	2.1	13.9
Protein_X2	-3.5	-4.2	35.8
Protein_X3	-3.5	10.0	17.3
Protein_X4	-0.4	7.0	17.6
Protein_X5	-3.2	1.4	21.8
Protein_X6	4.5	2.1	19.1
Protein_X7	3.8	-8.0	18.8
Protein_X8	-3.9	-4.2	21.3
Protein_X9	1.9	5.7	21.6
Protein_X10	2.9	-9.2	29.4

B

Protein	Y_1	Y_2	Y_3
Protein_X1	0.73082	0.28593	-1.20766
Protein_X2	-1.06628	-0.69844	2.2064
Protein_X3	-1.06628	1.5203	-0.68678
Protein_X4	-0.13778	1.05156	-0.63891
Protein_X5	-0.97643	0.17656	0.02516
Protein_X6	1.32298	0.28593	-0.40453
Protein_X7	1.12019	-1.29219	-0.44828
Protein_X8	-1.12019	-0.69844	-0.05609
Protein_X9	0.55111	0.84843	-0.00141
Protein_X10	0.70080	-1.47969	1.21109

Figure 9. Pre-treatment of dependent variables (Y) for partial least squares analysis. The raw data set is shown in A, and the transformed data set by standardization is shown in B.

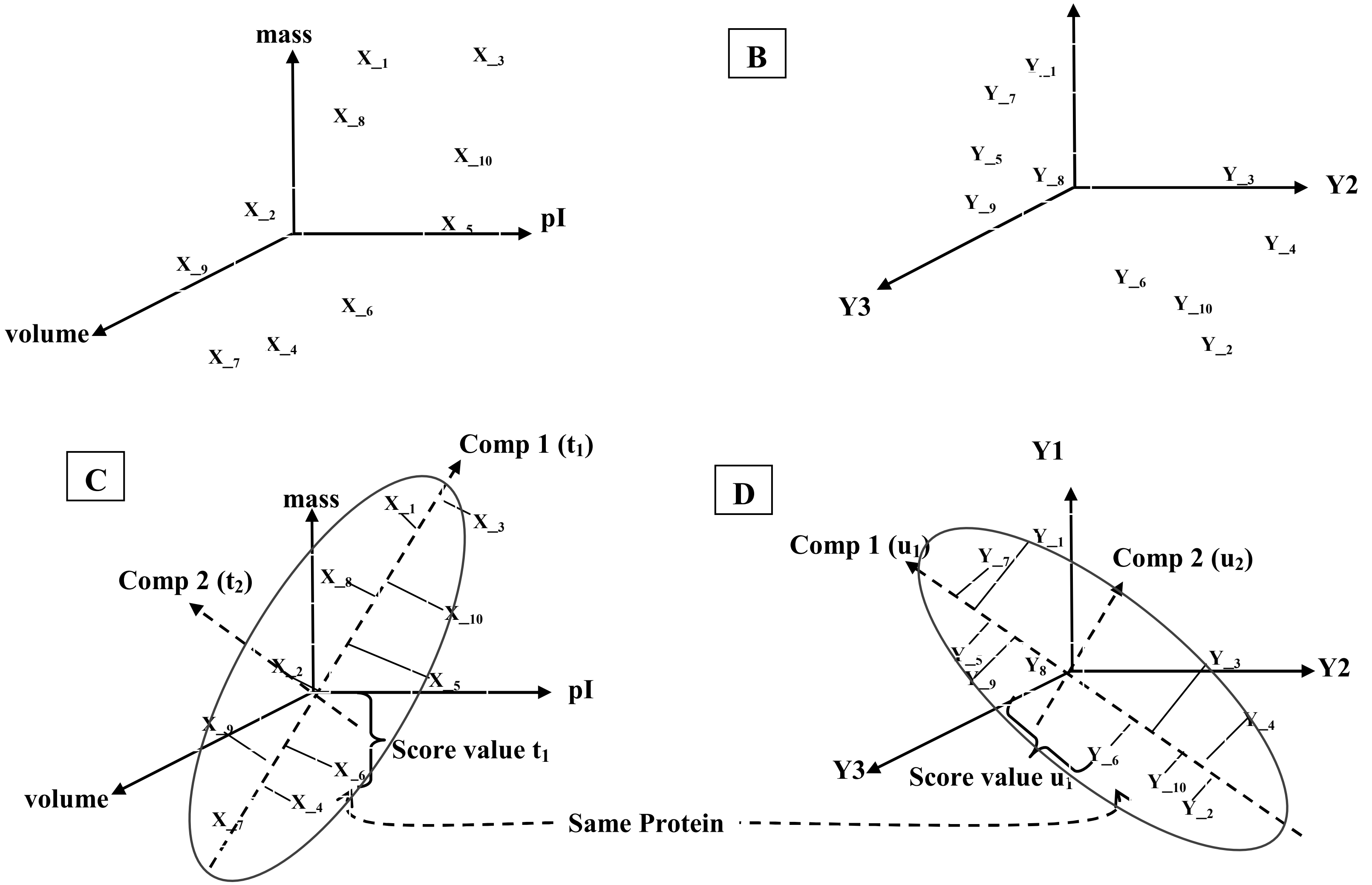


Figure 10. Geometrical representation of partial least squares. Transformed dependent and independent variables are plotted on each 3-D space (A and B). Each variable is projected onto the axis 1 (C and D)

Protein	Mass	Volume	Class
Protein_X1	0.148355	0.318024	A
Protein_X2	0.35575	2.31922	A
Protein_X3	2.18794	1.07058	A
Protein_X4	2.25507	1.07058	A
Protein_X5	0.12082	1.123779	A
Protein_X6	1..9263	1.04777	B
Protein_X7	2.29548	1.123779	B
Protein_X8	3.463418	0.23344	B
Protein_X9	2.214794	1.26493	B
Protein_X10	3.214794	1.340934	B

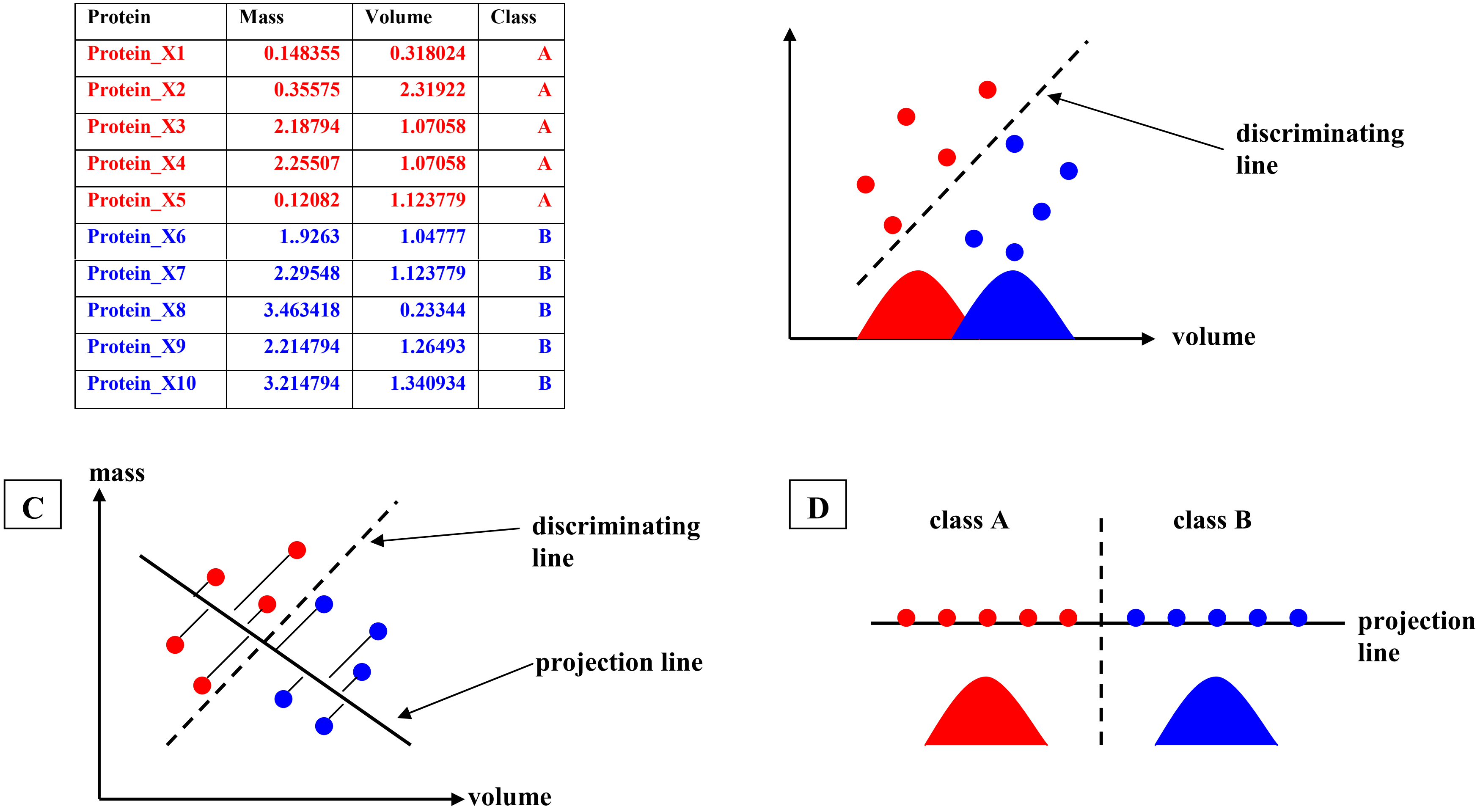


Figure 11. Discrimination between two protein classes A and B. Two classes of proteins and their variables are listed in A. A discriminating line is drawn to separate the classes in B. Volume variables have overlap in the normal distribution between classes A and B. Points projected to a line perpendicular to the discriminating line in C. Projected variables can be separated with no overlap in D.

Protein	Mass	Volume	Class
Protein_X1	0.248355	0.418024	1
Protein_X2	0.35575	1.31922	1
Protein_X3	0.18794	0.07058	1
Protein_X4	0.25507	0.07058	1
Protein_X5	0.12082	1.123779	1
Protein_X6	0.9263	1.04777	-1
Protein_X7	1.29548	1.123779	-1
Protein_X8	2.463418	0.23344	-1
Protein_X9	0.214794	1.26493	-1
Protein_X10	0.214794	1.340934	-1

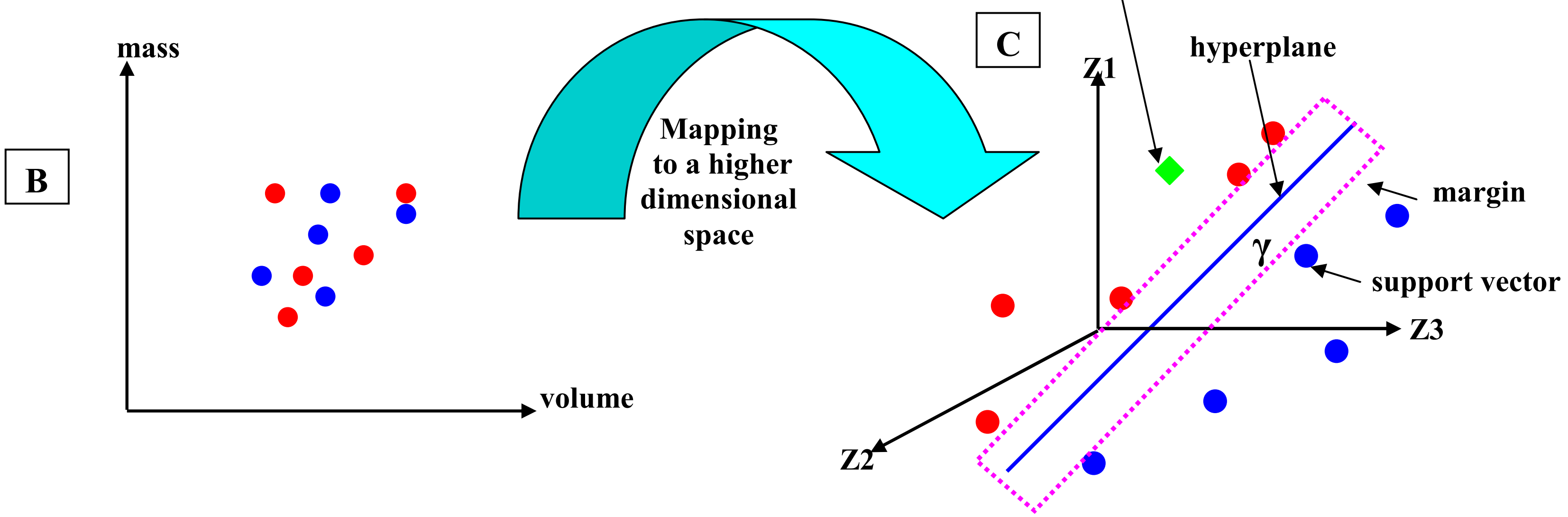


Figure 12. Remapping samples (B) from a 2-dimensinal input variable space to 3-dimensional future space. Ten sample proteins and their variables are listed in A. There is no separation between the two classes in the input 2-dimensional space (B). After remapping the variable space into is a higher dimensional space, samples can be separated by the hyperplane shown by the blue line with the margin **γ** to the closest training vectors (support vectors) in C.

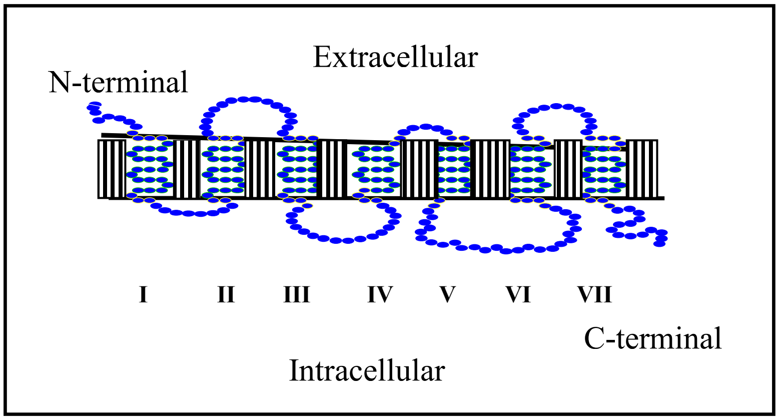


Figure 13. A model of G-protein coupled receptors. Seven transmembrane regions are shown from I to VII.

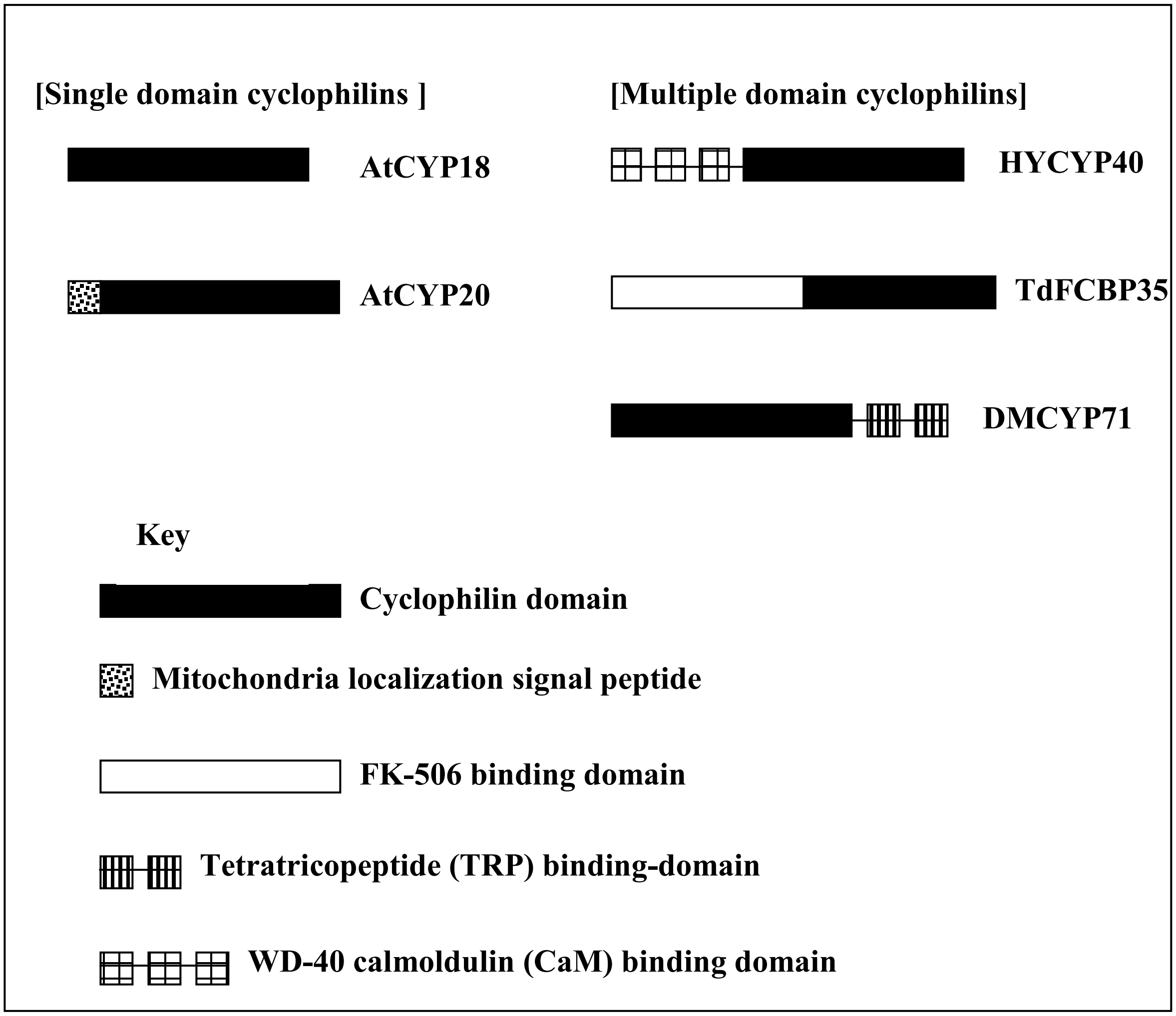

[Single domain cyclophilins]
AtCYP18
AtCYP20
[Multiple domain cyclophilins]
HYCYP40
TdFCBP35
DMCYP71
Key
Cyclophilin domain
Mitochondria localization signal peptide
FK-506 binding domain
Tetratricopeptide (TRP) binding-domain
WD-40 calmoldulin (CaM) binding domain

Figure 14. Domain architecture of the cyclophilin-related protein families.

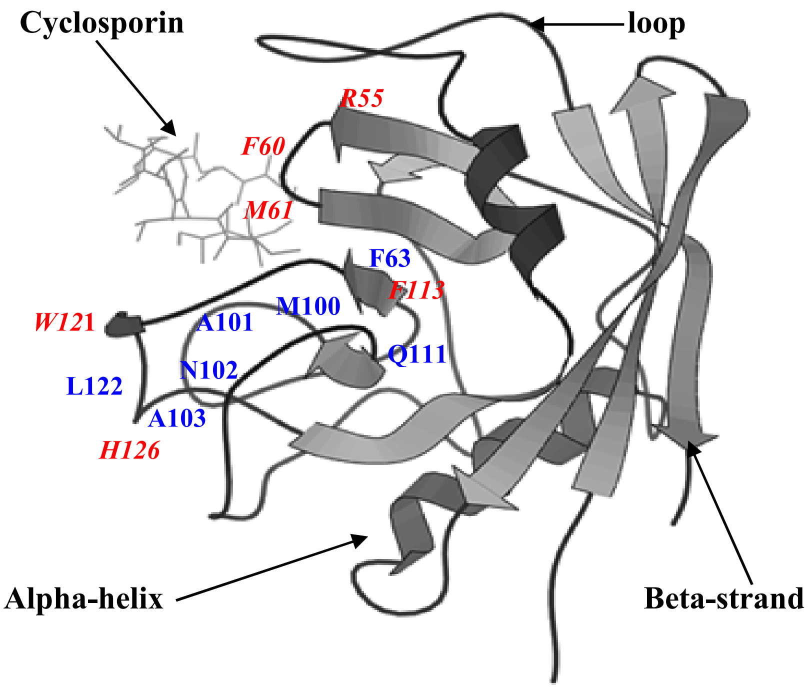


Figure 15. The 3D-structure of human cyclophilin A (from PDB entry 1BCK). Amino acid labels show these important for the PPIase activity and cyclosporin-binding. Amino acid labels for PPIase active sites are shown in red. All these residues except for H126 are also overlapped with cyclosporine-binding sites.

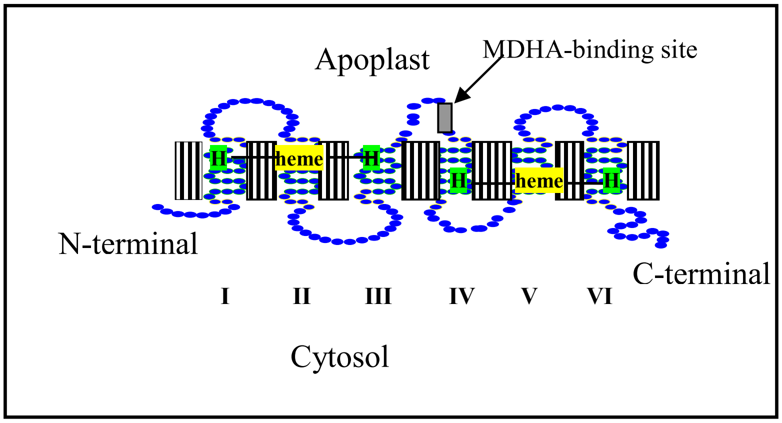


Figure 16. A model of Cytochrome b561. Six transmembrane regions are shown from I to VI. The conserved heme-binding histidine residues are represented by H. MDHA = monodehydroascobate

.

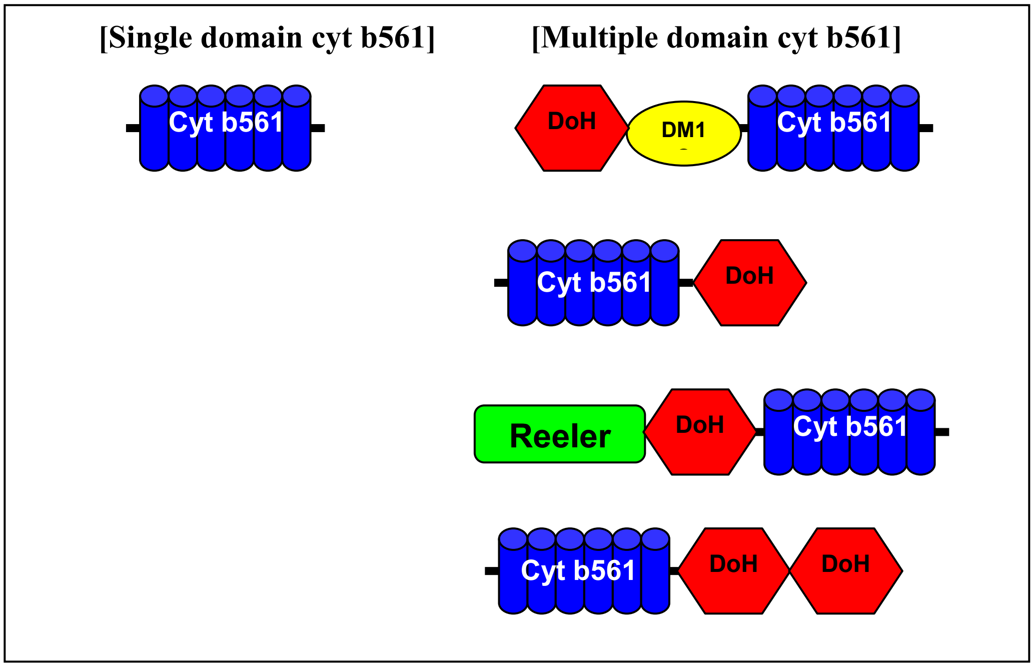


Figure 17 Domain architecture of the cyt b561 related protein family.

CHAPTER 2

APPLICATION OF PARTIAL LEAST SQUARES IN THE CLASSIFICATION OF PROTEIN FAMILIES

2.1 OVERVIEW

Chapter 2 discusses the application of partial least squares (PLS) in the classification of protein families using descriptors derived from composition and physico-chemical properties of amino acids. This chapter is divided into three sections as follows: Section 2.2 is based on the manuscript published in Opiyo and Moriyama (2007, J. Proteome Res. 6:846-853). It discusses how the size of training datasets affects the classifier performance, and how different methods could identify sequence fragments with different lengths. Remote similarity classification by PLS classifiers in section 2.3. In section 2.4, PLS classifiers are applied for mining of highly-divergent seven transmembrane receptors from *Arabidopsis* genome. This section is part of the study published in Genome Biology (Moriyama, Strope, Opiyo, Chen, and Jones. (2006, Genome Biology 7:R96).

2.2 PROTEIN FAMILY CLASSIFICATION WITH PARTIAL LEAST SQUARES

2.2.1 Abstract

The quality of protein function predictions rely on appropriate training of protein classification methods. We studied the use of partial least squares (PLS) for protein classification problems using the G-protein coupled receptor (GPCR) superfamily as an example. Physico-chemical properties and composition of amino acids were used as sequence descriptors. Four PLS classifiers were compared against profile hidden Markov models (HMMs) and PSI-BLAST. In this study, we focused on the effects of training dataset sizes and the test sequence lengths. The size of training datasets affected the performance of profile HMMs and PSI-BLAST significantly, but had little effect on the performance of the four PLS classifiers. Our PLS classifiers were also found to perform better for identifying short partial sequences of GPCRs. PLS classifiers using physico-chemical properties and composition of amino acids could successfully identified Cytochrome b561 protein sequences from *Arabidopsis thaliana* Expressed Sequence Tag sequences. None of them was identified by profile HMMs and PSI-BLAST. This study showed that my PLS classifiers have an advantage over profile HMMs and PSI-BLAST when only a limited number of training samples are available or when applied for short sequences. PLS-based protein classifiers are expected to be very effective in mining new or highly diverged protein sequences.

2.2.2 Introduction

Predicting protein functions is one of the most important problems in the post-genomic analyses. However, a larger number of proteins in many genomes are still not annotated for their functions. The majority of protein classification methods currently used rely on building alignments: e.g., PSI-BLAST[1], Pfam[2], and PROSITE[3]. As mentioned in Chapter 1, some homologous proteins are highly diverged and lack enough sequence similarities even though they still share similar structures, biochemical properties, and functions. Obtaining reliable alignments among these protein sequences is difficult. Another disadvantage of these classifiers is that they are trained using only "positive" samples (proteins of interest). "Negative" samples (unrelated proteins) cannot be included in the alignments their models are built from. Recently efforts have been put to increase sensitivities against such non-alignable similarities[4;5;6;7;8;9;10;11]. These methods (support vector machines, discriminant function analysis, and partial least squares) were discussed in Chapter 1.

In order to perform whole proteome prediction of protein function effectively and accurately, two often related problems need to be overcome: remote similarity detection and model-building based on limited samples. Such examples are found in extremely diverged protein families such as G-protein coupled receptors (GPCRs). As mentioned in Chapter 1, there are only 22 known GPCRs found in the *Arabidopsis thaliana* genome, in contrast to 1000 or more GPCRs found in human and mouse. It is possible that plants do not require this protein superfamily as much as animals. However, it is also possible that classifiers used to identify these proteins (mostly profile hidden Markov models) are affected by insufficiently represented training datasets. It is thus important for the users to recognize how different classifiers perform in such difficult situations.

In this study, we examined how the size of training datasets affects the classifier performance, and how different classifiers identify sequence fragments with different lengths. A large amount of Expressed Sequence Tag (EST) sequences are now available, and their numbers are rapidly increasing. These ESTs usually contain only fragments of protein coding sequences. Since many recent genome projects, especially for plant genomes, concentrate on sequencing ESTs rather than complete genomes, it is important to utilize this information source effectively. We therefore need to understand not only the sensitivity of various classifiers against non-alignable remote similarities but also against fragmented similarities.

The GPCR superfamily has been used to test classifier performance in many previous studies. As described in Chapter 1, GPCR proteins share little sequence similarities except a structural feature of having seven transmembrane regions. No reliable multiple alignment can be generated when different GPCR families are included, and attempting to find new GPCRs from new genomic or EST sequences is often hindered due to such extreme diversity. This provides a good test case for examining classifier performance in difficult situations. We thus used GPCR sequences as one example and compared the performance among profile hidden Markov models (HMMs), PSI-BLAST, and classifiers based on PLS with alignment-free descriptors: physico-chemical properties and composition of amino acids. As another example, we used Cytochrome b561 proteins for examining EST mining performance. The results of this comparative study will be useful for choosing appropriate methods for identifying protein functions in various protein-mining situations.

2.2.3 Materials and methods

(a) Data sources

GPCR data

All GPCR sequences were retrieved from the Swiss-Prot protein database[12].

Non-GPCR data

Non-GPCR sequences (negative samples) longer than 100 amino acids were randomly sampled from Swiss-Prot. The sequence identities, GPCR or not, were confirmed based on the Swiss-Prot annotations.

Cytochrome b561 data

Forty eight Cytochrome b561 (Cyt b561) protein sequences were retrieved from GenBank[13]. They included eleven sequences from plants (four from *A. thaliana*, one from maize, and six from rice) and 37 sequences from animals.

(b) Dataset preparation

Training and test datasets

Five sets of training data each including both positive (GPCR) and negative (non-GPCR) samples were generated as shown in Table 1. Three non-overlapping samples were prepared for each training dataset. For the smallest training set, Training10, five replications were prepared. The test dataset included 200 positive and 2000 negative samples. Note that for SAM and PSI-BLAST, only positive samples were used for training.

Subsequence test datasets

From each sequence of the test dataset, 50, 75, or 100 amino acid subsequences were taken from its N- and C-terminals. These six subsequence test sets including 2,200 sequences of a given length were used to examine classifier performance against short sequences.

Cyt b561 training sets

Three sets of training data were created from the Cyt b561 protein family: "Arabidopsis only" including four *A. thaliana* Cyt b561s, "Plants" including 11 plant Cyt b561s (including four from *A. thaliana*), and "Plants and animals" including 48 Cyt b561s including 11 from plants and 37 from animals. All these datasets included the same number of negative (non-Cyt b561) sequences.

Arabidopsis Expressed Sequence Tag (EST) sequences

362, 202 *A. thaliana* EST sequences were downloaded from The Arabidopsis Information Resource database (ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/AtEST.Z derived from the dbEST release for November, 2004). Each EST sequence was translated into three reading frames. ESTs containing the four known *A. thaliana* Ctyb-561 coding sequences (At1g14730, At1g26100, At4g25570, and At5g38630) were identified using the Smith-Waterman local alignment algorithm[14] implemented in SSearch[15]. Similarly ESTs containing multi-domain protein sequences with Cyt b561 domains were identified. Table 2 lists the accession numbers of these ESTs. These EST sequences were considered to be true positives.

(c) Descriptor development

In Lapinsh et al.[9], five principal components (z-scales) originally derived by Sanberg et al.[16] were used as descriptors for each amino acid. However, these z-scales gave poor results for our classification problem. Lapinsh et al.[9] classified Class A GPCRs into subfamilies, but our problem is to classify GPCRs from non-GPCRs. Therefore, we chose our own twelve physico-chemical properties considering the methods used, protein family used, etc. They include: mass[17], volume[18], surface area[18], hydrophilicity[19], hydrophobicity[20], isoelectric point[21], transfer of energy solvent to water[22], refractivity[23], and non-polar surface area[24], frequencies of alpha-helix[25], beta-sheet[25], and reverse turn[25]. These physico-chemical property values were first scaled to the unit variance. Principal component analysis (PCA) was performed on these twelve properties for 20 amino acids. The first five principal components (PCs) explaining 93.2% of the total variance were selected (Tables 3 and 4).

The first principal component (PC1) covered 40.4% of the total variance of the original 12 physico-chemical properties. It represents all properties except isoelectric point and non-polar surface. PC2 (28% of the total variance) has negative relationships with the hydrophobicity properties. PC3 and PC5, although their contributions are only 15% of the total variance, represent secondary structure properties. Using these five PC scores, each amino acid sequence was converted to a sequence of 5L descriptors, where L is the number of amino acids in the sequence. Since unaligned sequences have varied lengths, in order to transform them to a uniform matrix, four transformation methods were used: mean, auto and cross-covariance, amino acid composition, and amino acid composition with principal component analysis. Transformation of an unaligned set of amino acid sequences into a uniform matrix makes PLS and other multivariate methods applicable without aligning sequences

Transformation of amino acid sequences

Mean transformation

After each amino acid sequence was transformed to a set of five PC scores, the average was taken for each of the five PC scores. Regardless of the length, each amino acid sequence can be represented by an array of five average PC scores as follows:

$$\left(\frac{1}{L}\sum_{i=1}^{L}PC1_i,\ \ \frac{1}{L}\sum_{i=1}^{L}PC2_i,\ \ \frac{1}{L}\sum_{i=1}^{L}PC3_i,\ \ \frac{1}{L}\sum_{i=1}^{L}PC4_i,\ \ \frac{1}{L}\sum_{i=1}^{L}PC5_i\right) \tag{1}$$

where L is the length of the amino acid sequence and $PC1_i$, ..., $PC5_i$ are five PC values for the i-th amino acid.

Auto and cross-covariance (ACC) transformation

Auto/cross covariance (ACC) transformation of amino acid sequences was proposed by Wold et al.[26] and used in Lapinsh et al.[9]. 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:

$$AC_j(d) = \frac{1}{L - d}\sum_{i=1}^{L-d}(PCj_i)(PCj_{i+d}) \tag{2}$$

$$CC_{jk}(d) = \frac{1}{L - d}\sum_{i=1}^{L-d}(PCj_i)(PCk_{i+d}) \tag{3}$$

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, respectively ($j \neq k$; j, k = 1, 2, 3, 4, or 5), and L is the length. Note that Lapinsh et al.[9] used slightly modified and normalized version of ACC and reported improvement in classification performance compared to using the original Wold et al's ACC.

We used the R function, acf with the "covariance" option, for the ACC implementation (R version 2.2.0[27]). It uses the following equations:

$$AC_j(d) = \frac{1}{L}\sum_{i=1}^{L-d}(PCj_i - \overline{PCj})(PCj_{i+d} - \overline{PCj}) \tag{4}$$

$$CC_{jk}(d) = \frac{1}{L}\sum_{i=1}^{L-d}(PCj_i - \overline{PCj})(PCk_{i+d} - \overline{PCk}) \tag{5}$$

where $\overline{PCj}$ and $\overline{PCk}$ are the means of PCj_i and PCk_i, respectively. Use of the auto/cross-correlation (with normalization) did not show any difference in classification performance.

While the auto-covariances emphasize the interactions between amino acids, interactions between different amino acid properties are incorporated into the cross-covariances. Note that the ACC transformation incorporates positional information from sequences.

Amino acid composition transformation

From each amino acid sequence, frequencies of 20 amino acids were simply calculated. We used 19 frequencies as a set of descriptors since the 20-th amino acid frequency can be explained completely by the other 19.

Amino acid composition/PCA transformation

After each sequence is transformed to a 20 amino acid composition array, its dimension was further reduced by using PCA. As before, five top principal components were selected. Each sequence is represented with an array of five PC scores.

Classifiers

Partial least squares (PLS)

Partial least squares (PLS)[28] is an extension of PCA that finds a relationship between a matrix of predictor variables, ***X***, and a matrix of dependent variables, ***Y***. PLS was discussed in detail in Chapter 1. We used an R implementation, the pls package developed by Wehrens and Mevik[29], with the SIMPLS method and the cross-validation option. In this study, for each of the training samples, a response variable is assigned as 1 for the positive (GPCR) label, and 0 for the negative (non-GPCR) label. Therefore, the training dataset including N protein sequences is represented with two matrices ***X*** and ***Y*** with dimensions N x K (for K descriptors) and N x 1, respectively. The group membership, GPCR or non-GPCR, of a new sequence is predicted based on its K descriptors and calculating the ***X***-score, ***X***-residual, and y-value. Group assignment is done based on the y-value. In this study, four PLS classifiers were examined. They are based on different descriptor sets described before: "PLS-ACC" using ACC descriptors, "PLS-mean" based on mean PC scores, "PLS-AA" based on simple amino acid composition, and "PLS-AA_PCA" using amino acid composition transformed with PCA. The results obtained from the goodness of prediction are presented in Tables 5-7.

Goodness of fit of PLS model

The goodness of fit, R^2, describes how well the dataset can be mathematically reproduced by the fitted model:

$$R^2 = 1\text{-}RSS/SS_Y \qquad (6)$$

where RSS is the residual (error) sum of squares, SS_Y is the total sum of squares of observed Y. R^2 varies between 0 and 1. The closer to 1, the better the model is.

Goodness of prediction of PLS model

A model is not good enough if it has a high goodness of fit to the training data. More important is the predictive ability of the model. The goodness of prediction, Q^2, describes how well the model can predict a data. It is calculated similar to R^2, but using the cross-validation procedure:

$$Q^2 = 1\text{-PRESS}/SS_Y \tag{7}$$

where PRESS is the predictive residual sum of squares, which is calculated from the difference between observed and predicted Y values. $Q^2 > 0.5$ is considered good. In this study, the leave-one-out cross-validation procedure was used for the Q^2 calculation.

PSI-BLAST

PSI-BLAST[1] builds position-specific scoring matrices (PSSMs) from a multiple alignment. In the regular implementation, PSI-BLAST first performs a regular protein similarity search (BLASTP) against a protein database using a single protein query. The first PSSM is built from a set of highly similar sequences obtained from this search. Subsequent searches are done based on the PSSM iteratively built from high-score hits. In this study, we used pre-aligned positive sequences as the first input. Multiple alignments were generated using ClustalX version 1.83[30] with the default parameters. Ten iterations with E-value = 10 as the threshold for building PSSM were performed against the test dataset.

Profile hidden Markov models (HMMs)

Profile HMMs are the full probabilistic representation of sequence profiles[31]. 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 version 3.5)[32]. Multiple alignments used with the w0.5 script were built using the builmodel and align2model programs. The profile HMMs were built from multiple alignments using the modelfromalign program with Dirichlet mixture priors (recode3.2comp)[33] to improve the models by assigning prior probabilities of amino acids to the models, and the weight option 0.5 was used to save 0.5 bits of information per column of the multiple alignment. The test sequences were scored against models using hmmscore with the 'calibrate' option (for more accurate E-value calculation) and with the option -sw 2 for local scoring.

(f) EST mining

Classifiers were trained using the three datasets described earlier ("Arabidopsis only", "Plants", "Plants and animals"). The trained classifiers were applied to identify Cyt b561 containing ESTs. The E-value thresholds for PSI-BLAST and SAM, 1.01 and 1.24, respectively, were chosen based on the minimum error points (described later) obtained from GPCR classification using the training sets of 200 samples. For the PLS classifiers the cutoff of 0.4665 was chosen based on the minimum error points.

(g) Performance analysis

Statistics

Predictions are grouped as follows:

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

- False positives (FP): the number of actual negatives predicted as positives.

- True negatives (TN): the number of actual negatives predicted as negatives.

- False negatives (FN): the number of actual positives predicted as negatives.

The performance statistics are calculated as follows:

- Accuracy = (TP + TN)/ (TP +TN + FP + FN)
- False positive rate = FP/(FP + TN) = 1 – specificity
- False negative rate = FN/(FN + TP)
- True positive rate = TP/(TP + FN) = sensitivity

Receiver operating characteristics graph

Receiver operating characteristics (ROC) graph is useful for visualizing the performance of various methods[34]. An ROC curve is a graphical representation of the trade off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). It is illustrated as a plot with false positive rates (1-specificity) against true positive rates (sensitivity). The closer the curve follows the left-hand border and then the top border of the ROC space, the better classifiers perform. We can quantify the classifier performance by measuring the area under the curve. The closer the area to 1.0, the better the classifier is, and the closer the area to 0.5, the worse the classifier is (it is no better than random assignment).

Minimum error point

Minimum error point (MEP) was used by Karchin et al.[4] It is the threshold score where the method produces the minimum number of errors (false positives + false negatives).

2.2.4 Results

(a) Training set size and classifier performance

Datasets of five different sizes were used for training classifiers as shown in Table 1. PLS classifiers were compared against SAM and PSI-BLAST for their GPCR identification performance. Figure 1 summarizes the performance of the six classifiers. All classifiers were affected by the training dataset size but at varied degrees. The accuracy rates at the MEP among classifiers were affected by the size of training datasets similarly (Figure 1A top and Table 8). With the Training200 dataset, all classifiers showed 97% or higher accuracy. With smaller datasets, the accuracy rates decreased as low as 88-92% (Training10). When small datasets were used for training, PLS classifiers maintained low false negative rates (1-sensitivity; 4-8.5% with Training10 and 2-6% with Training20) as shown in Figure 1B. SAM and PSI-BLAST, however, had much more false negatives (66-78% with Training10 or Training20). Performance of SAM and PSI-BLAST also fluctuated much more than PLS classifiers among replicated trainings as indicated by the long error bars in Figure 1. Note that even though SAM and PSI-BLAST had very high false negative rates, they had very low false positive rates (1-specificity) even when very small training datasets were used (3-4% with Training10 or Training20; Figure 1A bottom and Table 8). PLS classifiers, on the other hand, had false positive rates ranging from 7 to 11% when trained with such small datasets.

The ROC plots in Figure 2 and Appendix Figure 1 show the difference in performance between SAM/PSI-BLAST and PLS classifiers more clearly. When trained on small training datasets, SAM and PSI-BLAST performed almost as a random classifier. They seemed to require more than 50 positive samples in the training sets to produce performance equivalent to PLS classifiers.

(b) Identification of short subsequences

In the second experiment, we examined how different classifiers can identify fragments of protein sequences with various lengths. The classifiers were trained with the dataset including 200 samples (Training200; see Table 1). Figure 3 and Table 9 show the results of N-terminal subsequence tests. Although the difference in accuracy was not large, PLS-ACC and PLS-AA performed slightly better than SAM and PSI-BLAST when sequence lengths were 75 amino acids (aa) or shorter (88-94% by PLS-ACC or PLS-AA *vs.* 86-89% by SAM or PSI-BLAST). The accuracy rate of PLS-ACC was 91.4% even against 50-aa sequences. For short sequences, the false negative rates of SAM and PSI-BLAST were very high (72-84% for 50aa sequences and 56-66% for 75aa sequences). False negatives given by PLS classifiers, especially PLS-ACC and PLS-AA, were much fewer (12% or fewer). Consistent with the results obtained against full-length sequences, SAM and PSI-BLAST had low false positive rates (below 5.5%). PLS classifiers had slightly higher false positive rates (9-15% for 50aa sequences and 7-10% for 75aa sequences). Results against C-terminal subsequences were also very similar (Figure 4 and Table 10).

(c) Identification of Cytochrome b561 from *A. thaliana* EST sequences

In order to examine the performance of these classifiers against the actual short partial sequences, we applied the classifiers for identifying *A. thaliana* ESTs. ESTs are usually 5' or 3' fragments of cDNA sequences, and in addition to 5' or 3' non-coding transcribed region, they could include short coding regions at the start or end of genes. For this test, we used another protein family, Cytochrome b561 (Cyt b561), as the second example. Cyt b561 is an integral membrane protein that is found in various organisms from humans to plants. It has six transmembrane regions as well as a pair of hemes. Cyt b561 sequences also exist in some multi-domain proteins, linked to other domains such as the Dopamine β-hydroxylase (DOH) domain[35].

When we used simply BLASTP[1] for similarity search using the four Arabidopsis Cyt b561 sequences (At5g38630, At1g26100, At4g25570, and At1g14730) as queries and using the default E-value threshold of 10, we could identify only six ESTs (one derived from At5g38630, two derived from At1g26100, and three derived from At5g25570) of 18 found by SSEARCH. With three multi-domain proteins containing the Cyt b561 domain (At5g47530, At5g47530, and At361750) as queries, we could identify only four ESTs (three derived from At5g47530 and one from At3g07570) of 11 found by SSEARCH.

Next, we trained classifiers for Cyt b561 identification and used these classifiers to find Arabidopsis ESTs that include Cyt b561 fragments. As shown in Table 11, PLS-ACC and PLS-AA were able to identify EST sequences as short as 359 bp including 225 bp of a 5' Cyt b561 coding region. Such short ESTs could not be correctly identified by SAM and PSI-BLAST. When classifiers were trained with more Cyt b561 including other plant and animal sequences, PLS-ACC and PLS-AA only missed two ESTs derived from At4g25570, but SAM and PSI-BLAST missed five. All eleven ESTs from multi-domain Cyt b561 sequences were identified by PLS-ACC and PLS-AA regardless of the size of the training sets, whereas SAM and PSI-BLAST could identify only eight when they were trained using the dataset including only four Arabidopsis Cyt-b516 samples (Table 12).

2.2.5 Discussion

The number of sequences used in the training dataset did not affect the performance of alignment-free PLS classifiers. Small samples from diverged sequences such as GPCRs must have caused both over-fitting as well as unreliable alignments to decrease sensitivities of the alignment-based methods. The consistent results were shown with Cyt b561 protein identification from Arabidopsis EST sequences. PLS classifiers did not require a large training set to gain optimal performance on these EST sequences. It confirmed that PLS classifiers are more sensitive (fewer false negatives) than SAM and PSI-BLAST regardless of the training data sizes, sequence lengths, or protein families. Note, however, that PLS classifiers are in general less specific (with more false positives) than SAM and PSI-BLAST. More work is required to reduce the number of false positives by alignment-free PLS-classifiers.

Among PLS classifiers PLS-ACC and PLS-AA performed slightly better than others. Although PLS-ACC can incorporate the interaction between amino acid positions, the performance gain with using ACC over simpler amino acid composition was not very significant. It is reasonable to assume that amino acid residues in a protein sequence are not independent and such relationships are specific to each protein group considering the existence of functional domains or sites. Such correlations among sites depending on functional domains can be identified and utilized by classifiers using ACC transformation. Therefore, it is surprising to see much simpler descriptors like amino acid composition indeed worked as well or better than ACC descriptors. One drawback of ACC is that calculating it is computationally expensive. With lag = 30, the input vector for each sequence includes 775 descriptors. Using much simpler descriptors as amino acid composition appears to be more attractive, for example, for the first-step classification of the whole proteome analyses.

As our results showed, the alignment-free PLS classifiers can be used in a situation where there are not enough example sequences. For example, currently only 12 members of Class D and 4 members of Class E GPCRs are available. Those few sequences can be effectively used to train PLS methods to search new GPCR sequences in those classes from whole proteome data as well as EST sequences.

2.2.6 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.
2. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, and Eddy SR. The Pfam protein families database. Nucleic Acids Res. 2004;32:D138-141.
3. Hulo N, Sigrist CJA, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, and Bairoch A. Recent improvements to the PROSITE database. Nucleic Acids Res. 2004;32:D134-D137.
4. Karchin R, Karplus K, Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002;18:147-159.
5. 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.
6. 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.
7. Moriyama EN and Kim J. Protein family classification with discriminant function analysis. In *Genome Exploitation: Data Mining the Genome*; Gustafson JP, Shoemaker R, Snape JW, Eds, Springer: New York, 2005. pp 121-132.

8. Bhasin M, Raghava GPS, GPCRsclass: a web tool fro the classification of amine type of G-protein-coupled receptors. Nucleic Acids Res. 2005;33:W143-W147.

9. Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JES. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. Protein Sci. 2002;11:795-805.

10. Chou KC and Elrod DW. Bioinformatical analysis of G-protein coupled receptors. J. Proteome Res. 2002;1:429-433.

11. Chou KC. Prediction of G-protein-coupled receptors classes. J. Proteome Res. 2005;4:1413-1418.

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

13. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2006;34:D16-20.

14. Smith TF, Waterman MS. Identification of common molecular subsequences. J. Mol. Biol. 1981;147:195-197.

15. Pearson WR. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. Genomics 1991, 11, 635-650.

16. Sanberg M, Eriksson L, Jonsson J, and Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J. Med. Chem. 1998;41:2481-2491.

17. Biemann K. Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation. Methods Enzymol. 1990;193:455-479.

18. Chothia C. The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 1976;105:1-12.

19. Parker JM, Guo D, Hodges RS. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry 1986; 25:5425-5432.

20. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. 1984;179: 125142.

21. Zimmerman JM, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. J.Theor. Biol. 1968;21:170-201.

22. Nozaki Y Tanford C. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. J. Biol. Chem. 1971;246:2211-2217.

23. Jones DD. Amino acid properties and side-chain orientation in proteins: a cross correlation approach. J. Theor. Biol. 1975;50:167-183.

24. Bordo D, Argos P. Suggestions for "safe" residue substitutions in site-directed mutagenesis. J. Mol. Biol. 1991;217:721-729.

25. Levitt M. Conformational preferences of amino acids in globular proteins. Biochemistry 1978, 17, 4277-4285.

26 Wold S, Jonsson J, Sjostrom M, Sandberg M, and Rannar S. DNA and Peptide

Sequences and Chemical Processes Multivariately Modeled by Principal Component Analysis and Partial Least-Squares Projections to Latent Structures. Anal. Chim. Acta. 1993;277:239-253.

27. 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.
28. Geladi P and Kowalski BR. Partial least squares regression: A tutorial. Anal. Chim. Acta 1986;185:1-17.
29. Wehrens R, Mevik B. pls: Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR). R package version 1.2-1. 2006. http://mevik.net/work/software/pls.html
30. 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;22:4673-4680.
31. Durbin R, Eddy S, Krogh A, and Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press; Cambridge, 1998.
32. Hughey R and Krogh A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. Comput. Appl. Biosci. 1996;12:95-107.
33. Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, and Haussler D. Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. Comput. Appl. Biosci. 1996;12:327-345.

34. Hanley JA and McNeil BJ The meaning and use of the area under the receiver operating characteristics (roc) curve. Radiology 1982;143:29-36.

35. Ponting CP. Domain homologues of dopamine b-hydroxylase and ferric reductase: roles for iron metabolism in neurodegenerative disorders. Hum. Mol. Genet. 2001;10: 1853-1858.

2.2.6 Tables

Table 1. Numbers of samples included in GPCR datasets

Datasets	GPCRs					Non-GPCRs	Total
	Class A	Class B	Class C	Class D	Class E		
Training10	1	1	1	1	1	5	10
Training20	3	2	2	2	1	10	20
Training50	8	7	7	2	1	25	50
Training100	23	17	7	2	1	50	100
Training200	73	17	7	2	1	100	200
Test	136	50	10	4	0	2000	2200

Table 2. The list of *A. thaliana* ESTs containing known Cyt-b561 coding sequences

Cyt-b561 coding sequences	No. of ESTs	EST accession numbers
[Single-domain Cyt-b561 proteins]		
At1g14730	1	AI996604
At1g26100	3	BP587463, CF652328, BE522695
At425570	9	AV819134, BP571487, AV788828, BP621624, BX835859, BP611349, BE844714, BP571951, T0480
At5g38630	5	AU238085, BP623885, BP632139, BP663600, BP573945
[Multiple-domain proteins containing the Cyt-b561 domain]		
At3g07570	3	BP576524, BP585760, AV786182
At3g61750	1	CF773415
At5g47530	7	AV567140, AV565093, AV565001, AV562405, AV566417, AV567065, BP573319

Table 3. Five principal component scores for the 20 amino acids

	Principal component scores (% variance)[a]				
Amino acids	PC1 (40.4)	PC2 (28.7)	PC3 (10.9)	PC4 (8.9)	PC5 (4.2)
Alanine (A)	-1.74	-2.24	-1.88	0.41	-0.06
Arginine (R)	0.68	3.56	0.79	2.28	1.08
Aspartate (D)	-2.24	0.84	0.23	-1.99	0.51
Asparagine (N)	-1.82	0.98	0.53	-0.53	0.27
Cystine (C)	-0.22	0.30	-0.72	-0.67	-0.81
Glutamate (E)	-0.61	1.55	-1.14	-2.07	0.77
Glutamine (Q)	-0.69	1.65	-0.75	-0.55	0.65
Glycine (G)	-4.02	-2.56	1.12	0.58	-0.42
Histidine (H)	0.04	2.36	-1.18	-0.05	-1.67
Isoleucine (I)	2.16	-2.24	0.02	0.51	0.34
Leucine (L)	1.93	-1.85	-1.21	0.26	-0.60
Lysine (K)	-1.31	2.26	-0.91	2.51	-0.41
Methionine(M)	1.83	0.07	-1.76	-0.37	0.00
Phenylalanine (F)	3.56	-0.97	0.28	-0.23	0.00
Proline (P)	-2.07	-0.11	2.45	-0.22	-1.39
Serine (S)	-2.63	-1.01	0.47	0.13	0.24
Threonine (T)	-1.36	-0.44	0.62	0.03	0.91
Tryptophan (W)	4.96	0.41	1.22	-0.46	-0.56
Tyrosine (Y)	2.59	0.50	2.12	-0.46	0.29
Valine (V)	0.98	-3.05	-0.28	0.91	0.87

[a] Percent total variance of the original 12 physico-chemical properties. See Appendix Table 4 for the loadings for each principal component.

Table 4. Loadings of the physico-chemical properties of amino acids for the five principal components[a]

	Loadings for each principal component (% variance)				
Amino acid properties	PC1	PC2	PC3	PC4	PC5
	(40.4)	(28.7)	(10.9)	(8.9)	(4.2)
Mass	0.34	0.34	0.13	-0.05	0.07
Volume	0.33	0.13	0.17	-0.33	0.19
Surface area	0.39	0.24	0.09	0.08	0.05
Hydrophobicity	0.19	-0.47	-0.03	-0.18	-0.34
Hydrophilicity	-0.38	0.24	0.01	-0.11	0.30
Refractivity	0.36	0.23	0.06	-0.04	-0.20
Isoelectric point	0.03	0.20	0.03	0.85	-0.18
Energy of water to ethanol	-0.39	0.21	-0.17	0.04	0.22
Non-polar surface	0.03	-0.51	-0.12	0.19	0.02
Frequency of alpha-helix	0.12	0.15	-0.76	-0.14	0.07
Frequency of beta-sheet	0.21	0.32	0.20	0.19	0.77
Frequency of reverse-turn	-0.32	0.06	0.52	-0.16	-0.19

[a] Loadings reflect the relative contribution of each property to the five principal components.

Table 5. The number of PLS components and the predictive ability of PLS-ACC from the leave-one-out cross validation procedure.

Training sets[a]	Number of PLS components	Q^2
10	1	0.58
20	4	0.74
50	3	0.65
100	5	0.70
200	7	0.74

[a] See Table 1 for the details on each training set.

Table 6. The number of PLS components and the predictive ability of PLS-AA from the leave-one-out cross validation procedure.

Training sets[a]	Number of PLS components	Q^2
10	3	0.74
20	4	0.76
50	4	0.77
100	4	0.91
200	6	0.88

[a] See Table 1 for the details on each training set.

Table 7. The number of PLS components and the predictive ability of PLS-ACC and PLS-AA from the leave- one-out cross validation procedure obtained from Cyt-b561 training.

Training datasets	Number of PLS components	Q^2
[PLS-ACC]		
Arabidopsis only	4	0.65
Plants	4	0.62
Plants and animals	6	0.57
[PLS-AA]		
Arabidopsis only	3	0.62
Plants	4	0.82
Plants and animals	4	0.91

Table 8. Classifier performance using different training datasets.

Classifiers	% Accuracy	% False positive	% False negative
[Training10]			
PLS-ACC	91.9 ± 1.2	7.0 ± 1.2	4.0 ± 0.7
PLS-Mean	87.5 ± 1.0	11.0 ± 1.8	8.5 ± 1.2
PLS-AA	89.6 ± 0.6	10.5 ± 1.2	6.5 ± 0.7
PLS-AA_PCA	89.0 ± 1.5	11.2 ± 2.3	8.0 ± 0.6
SAM	89.2 ± 6.2	4.2 ± 0.2	76.0 ± 12.0
PSI-BLAST	89.9 ± 7.7	3.6 ± 0.6	77.5 ±13.1
[Training20]			
PLS-ACC	94.8 ± 0.8	6.9 ± 0.9	2.0 ± 0.6
PLS-Mean	89.2 ± 0.7	11.2 ± 0.8	6.0 ± 0.9
PLS-AA	91.3 ± 1.1	9.2 ± 0.7	3.0 ± 0.5
PLS-AA_PCA	91.6 ± 0.5	8.6 ± 0.6	3.0 ± 0.7
SAM	90.8 ± 5.3	3.2 ± 0.1	65.8 ± 13.0
PSI-BLAST	90.9 ± 5.4	3.0 ± 0.3	70.0 ± 10.5
[Training50]			
PLS-ACC	94.9 ± 0.9	5.5 ± 1.2	1.0 ± 0.1
PLS-Mean	91.4 ± 0.8	9.0 ± 0.9	4.0 ± 0.9
PLS-AA	95.4 ± 0.8	4.8 ± 0.2	2.5 ± 0.1
PLS-AA_PCA	93.1 ± 0.9	7.4 ± 1.2	4.0 ± 0.8
SAM	93.2 ± 5.6	2.1 ± 0.2	42.5 ± 11.0
PSI-BLAST	92.3 ± 2.7	2.0 ± 0.2	54.5 ± 7.5
[Training100]			
PLS-ACC	93.5 ± 0.3	2.2 ± 0.3	1.0 ± 0.3

Table 8 (continued).

Classifiers	% Accuracy	% False positive	% False negative
PLS-Mean	94.8 ± 0.5	6.2 ± 0.7	3.0 ± 0.1
PLS-AA	98.0 ± 0.2	2.0 ± 0.3	2.0 ± 0.1
PLS-AA_PCA	95.1 ± 0.3	5.0 ± 0.9	3.0 ± 0.2
SAM	96.5 ± 0.9	1.0 ± 0.2	28.5 ± 2.5
PSI-BLAST	96.3 ± 0.7	1.1 ± 0.2	29.5 ± 6.2
[Training200]			
PLS-ACC	99.1 ± 0.1	1.2 ± 0.2	1.0 ± 0.3
PLS-Mean	97.0 ± 0.2	3.2 ± 0.1	1.0 ± 0.5
PLS-AA	99.0 ± 0.1	1.3 ± 0.1	1.0 ± 0.3
PLS-AA_PCA	96.7 ± 0.1	1.4 ± 0 .1	1.5 ± 0.5
SAM	99.0 ± 0 .2	0.5 ± 0.1	6.0 ± 2.8
PSI-BLAST	98.5 ± 0.5	0.6 ± 0.1	10.5 ± 3.8

Table 9. Classifier performance on N-terminal sub-sequences

Methods	% Accuracy	% False positive	% False negative
[Length = 50aa]			
PLS-ACC	91.4	9.2	8.0
PLS-Mean	84.3	15.0	22.5
PLS-AA	88.2	11.7	12.0
PLS-AA_PCA	84.2	15.3	20.0
SAM	86.3	5.4	72.0
PSI-BLAST	86.2	5.2	84.0
[Length = 75aa]			
PLS-ACC	93.5	7.2	5.0
PLS-Mean	89.2	10.0	18.0
PLS-AA	93.1	6.6	8.5
PLS-AA_PCA	91.0	8.4	14.0
SAM	88.7	5.0	56.0
PSI-BLAST	87.2	5.1	65.5
[Length = 100aa]			
PLS-ACC	97.1	2.0	2.0
PLS-Mean	91.6	8.2	10.0
PLS-AA	95.8	4.2	4.0
PLS-AA_PCA	95.4	4.3	8.0
SAM	96.0	2.1	10.0
PSI-BLAST	95.6	3.2	12.0

Table 10. Classifier performance on C-terminal sub-sequences

Methods	% Accuracy	% False positive	% False negative
[Length = 50aa]			
PLS-ACC	91.4	7.2	7.0
PLS-Mean	84.3	17.0	25.0
PLS-AA	91.4	9.4	10.0
PLS-AA_PCA	86.2	13.3	22.0
SAM	84.2	6.4	68.0
PSI-BLAST	84.3	4.2	83.0
[Length = 75aa]			
PLS-ACC	93.5	6.2	5.5
PLS-Mean	89.2	13.0	10.0
PLS-AA	92.2	8.6	7.5
PLS-AA_PCA	92.2	7.4	12.0
SAM	91.0	5.0	55.0
PSI-BLAST	86.7	5.1	63.5
[Length = 100aa]			
PLS-ACC	97.1	2.0	3.0
PLS-Mean	91.6	8.2	12.0
PLS-AA	94.8	5.2	4.0
PLS-AA_PCA	96.8	3.3	10.0
SAM	96.4	2.1	12.0
PSI-BLAST	94.0	3.2	13.0

Table 11. Identification of Cyt-b561 containing Arabidopsis ESTs

Training datasets	Methods	Numbers of ESTs identified for each Cyt-b561[a]				Average lengths[b]
		At5g38630 (5)	At1g26100 (3)	At4g25570 (9)	At1g14730 (1)	
Arabidopsis only	PLS-ACC	3	3	5	1	435 (359)
	PLS-AA	2	3	5	1	432 (359)
	SAM	1	1	5	0	509 (411)
	PSI-BLAST	1	1	5	0	509 (411)
Plants	PLS-ACC	5	3	7	1	517 (359)
	PLS-AA	5	3	7	1	517 (359)
	SAM	2	1	4	0	534 (406)
	PSI-BLAST	2	1	4	1	528 (406)
Plants and animals	PLS-ACC	5	3	7	1	536 (359)
	PLS-AA	5	3	7	1	536 (359)
	SAM	5	3	4	1	522 (401)
	PSI-BLAST	5	3	4	1	522 (401)

[a] The numbers of ESTs identified by SSEARCH are shown in the parentheses.

[b] The average lengths (bp) of ESTs correctly identified. The minimum lengths are shown in the parentheses.

Table 12. Identification of multi-domain Cyt-b561 proteins from Arabidopsis ESTs[a]

Training datasets	Methods	Numbers of ESTs identified for each Cyt-b561			Average lengths
		At5g47530 (7)	At3g07570 (3)	At3g61750 (1)	
Arabidopsis only	PLS-ACC	7	3	1	530 (430)
	PLS-AA	7	3	1	532 (430)
	SAM	5	3	0	535 (438)
	PSI-BLAST	5	3	0	535 (438)
Plants	PLS-ACC	7	3	1	530 (430)
	PLS-AA	7	3	1	530 (430)
	SAM	7	3	1	530 (430)
	PSI-BLAST	7	3	1	530 (430)
Plants and animals	PLS-ACC	7	3	1	530 (430)
	PLS-AA	7	3	1	530 (430)
	SAM	7	3	1	530 (430)
	PSI-BLAST	7	3	1	530 (430)

[a] See the footnotes for Table 11.

2.2.8 Figures

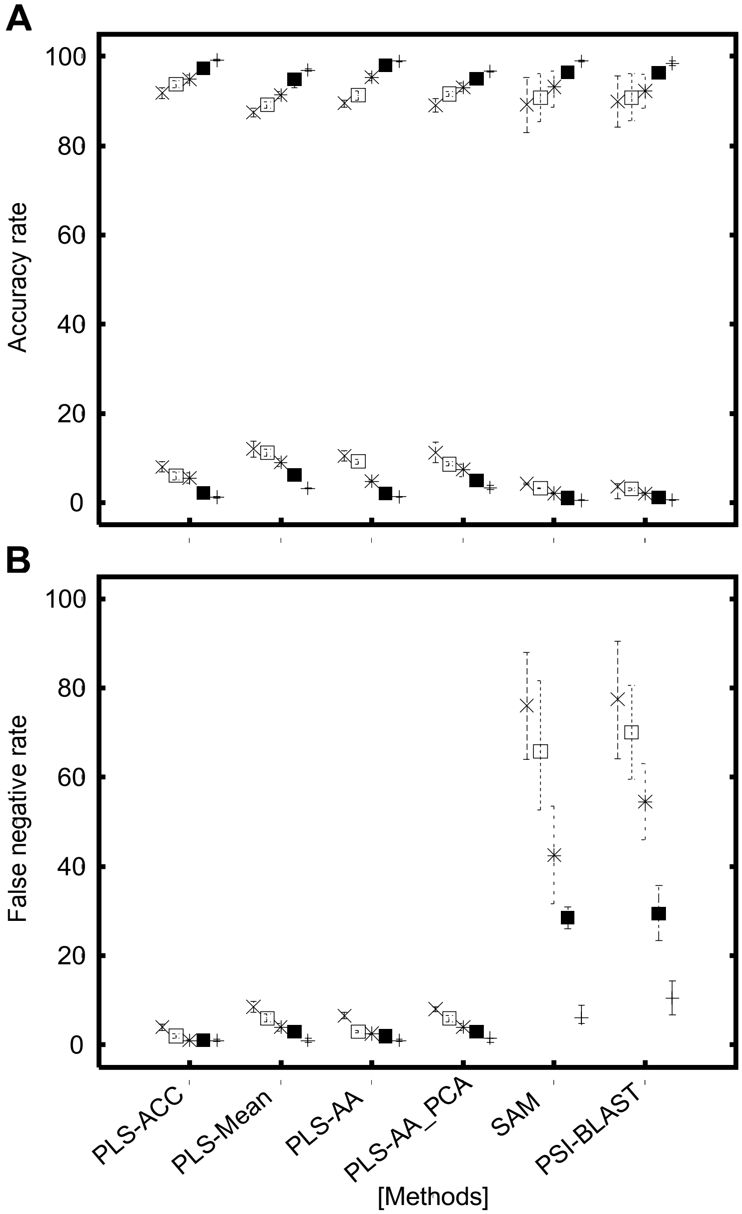


Figure 1. Training dataset size and classifier performance. The average values from three (or five) replications of training are plotted with error bars. A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%). For each classifier, the results are shown from left to right using the following training sets: Training10 (**X**), Training20 (open square), Training50 (*), Training100 (filled square), and Training200 (+).

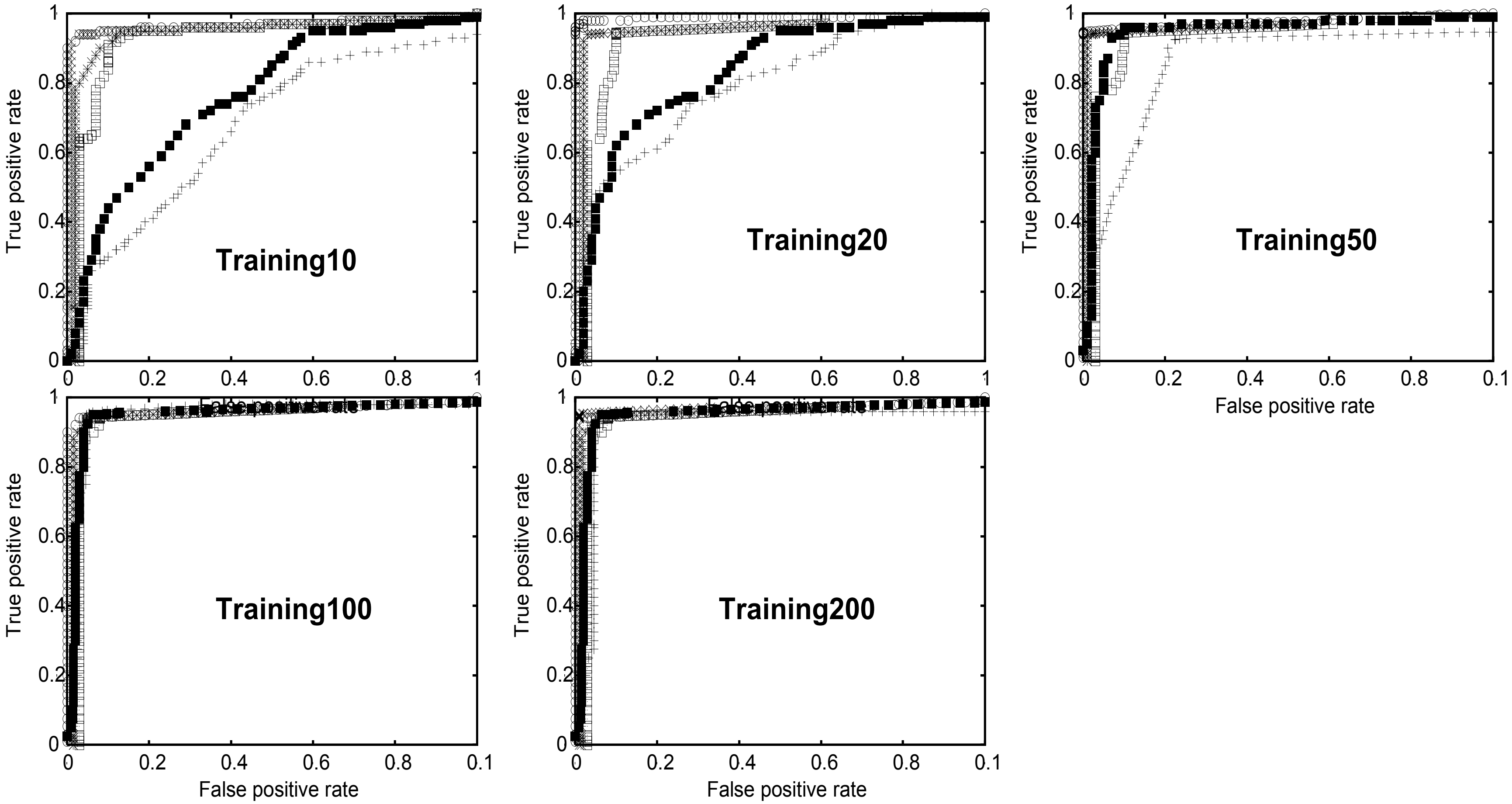


Figure 2. ROC graphs for classifiers using different sizes of training datasets. Plots are based on the mean of five replications for the Training10, and three replications for the Training20, Training50, Training100 and Training200, respectively. Classifier are shown as follows: PLS-ACC (open circle), PLS-Mean (**X**), PLS-AA (*****), PLS-AA_PAC (open square), SAM (filled square), and PSI-BLAST (+).

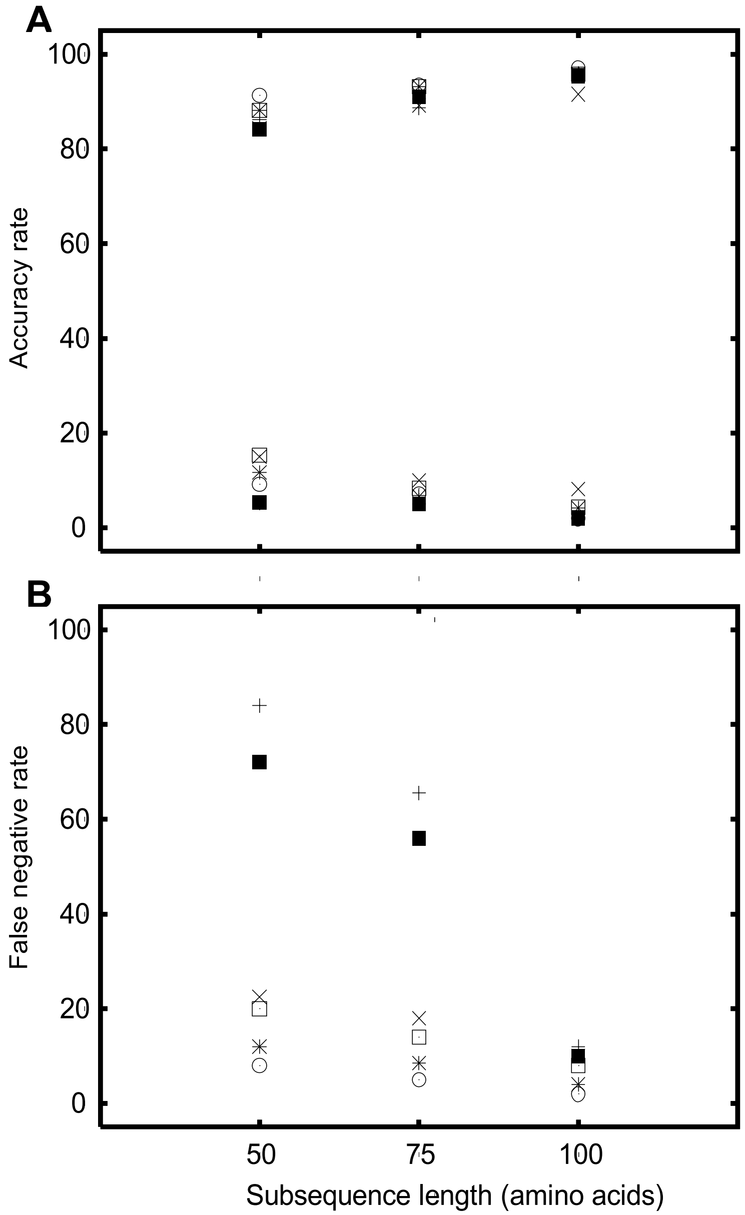


Figure 3. Subsequence lengths and classifier performance. Results of N-terminal subsequence tests are shown. A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%). Classifier symbols are described in Figure 2 legend.

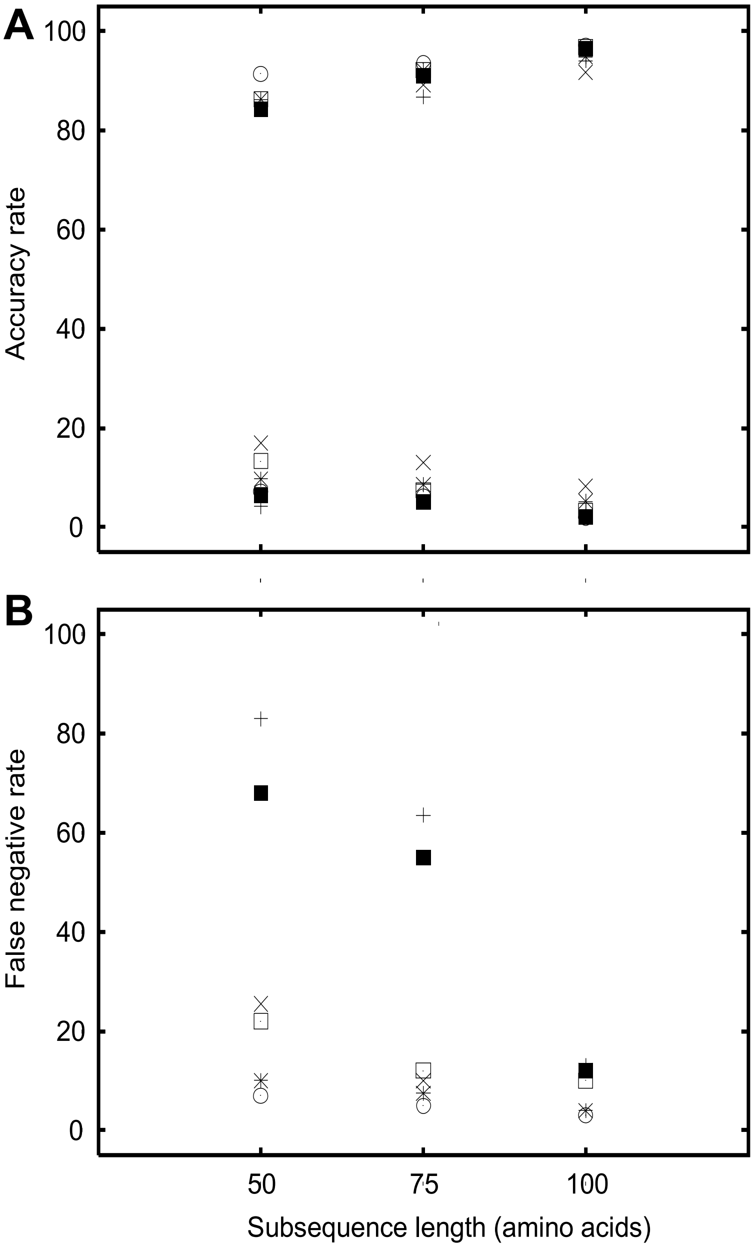


Figure 4. Subsequence lengths and classifier performance. Results of C-terminal subsequence tests are shown. A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%). Classifier symbols are described in Figure 2 legend.

2.3 REMOTE SIMILARITY CLASSIFICATION BY PLS CLASSIFIERS

2.3.1 Abstract

The sequence of different classes of GPCR are highly diverged from each other, except that they share one common structural feature, that is, they all have seven transmembrane regions. Due to extreme diversity among different GPCR families, no reliable multiple alignment can be generated from the entire GPCR superfamily. In this study, I examined how classifiers based on partial least square regression (PLS) could identify G-protein coupled receptors (GPCRs) classes that were not included in training datasets. Four PLS classifiers: PLS-ACC, PLS-Mean, PLS-AA, and PLS-AA_PCA described in section 2.3.3 (e) were compared against alignment-based classifiers: profile hidden Markov models (HMMs) and PSI-BLAST. These classifiers were trained on datasets containing sequences sampled from one of the four major GPCR classes A, B, C, and D. The classifiers were then tested for identification performance against GPCRs derived from the class that were not included in the training sets. PLS classifiers performed with 90% or higher accuracy rates, whereas profile HMMs and PSI-BLAST showed only 70% or lower accuracy rates. This study shows that PLS classifiers can be applied in the discovery of unknown or novel GPCR classes they have not been directly trained on.

2.3.2 Introduction

G-protein coupled receptors are transmembrane proteins that via guanine nucleotide-binding proteins, G-proteins, initiate some of the most important signaling pathways in the cell. The intracellular ligands that activate receptors are numerous and diverse, and include amino acids, ions, lipids, and polypeptides functioning as, for example, neurotransmitters or hormones (described in Chapter 1). The receptors are also involved in sensory system and are activated by light, odorants, and others. As mentioned in Chapter 1, the GPCR superfamily is divided into six major classes A, B, C, D, E, and Frizzled/Smoothened family according to the GPCRDB system[1]. These receptors share very little or no sequence similarity or length, but have a common structural feature in the transmembrane region where the sequence is arranged in seven alpha-helical domains that span the membrane. Each class is further divided into subclasses, subfamilies and so forth, depending upon the common ligands they bind to and sequence similarities. Class A is by far the most abundant class with more than 1,800 members in the Swiss-prot database. But the number of GPCRs in other classes are much less compared to class A. For example, there are only 12 GPCR sequences from Class D and 4 from Class E found in Swiss-prot. It is, therefore, important to have a classifier that can identify those from unidentified new GPCR groups. The objective of this study, therefore, is to examine the performances of classifiers on how they can identify GPCR classes not included in the training datasets.

2.3.3 Materials and methods

(a) Data sources

GPCR data

All GPCR sequences were retrieved from Swiss-Prot release 44, July 2004[2].

Non-GPCR data

Non-GPCR sequences (negative samples) longer than 100 amino acids were randomly sampled from Swiss-Prot. The sequence identities, GPCR or not, were confirmed based on Swiss-Prot annotations.

(b) Datasets preparation

Three training datasets consisting of 50 Class A GPCRs, 50 Class B GPCRs or 40 Class C plus 10 Class D GPCRs were generated as shown in Table 1. Three test sets were created using combinations of positive sets listed above A and B, A and D, and B and D as also shown in Table 1. Equal numbers of negative samples were included for each training and test dataset. Non-GPCR sequences were mutually exclusive between the training and test datasets.

(c) Classifiers

Partial least squares classifiers

PLS[3] classifiers based on different descriptors were described in section 2.2.3. They are "PLS-ACC" based on auto and cross-covariance descriptors, "PLS-mean" based on the mean PC scores, "PLS-AA" based on simple amino acid composition, and "PLS-AA_PCA" using amino acid composition transformed with PCA. For analysis of the sequences, the same procedures used in section 2.2.3 were applied in this study.

Profile hidden Markov model

Profile HMM[4] implemented in Sequence Alignment and Modeling Software System (SAM version 3.5)[5] was described in Chapter 1 and section 2.2.3. The same procedures used in the analysis in section 2.2.3 were applied for this study.

PSI-BLAST

PSI-BLAST [6] was also described in section 2.2.3. I followed the same procedures used in section 2.2.3 to do the analysis by PSI-BLAST in this study.

2.3.4 Results

I investigated how classifiers could identify GPCRs derived from the class that were not included in the training sets. These classifiers were trained on datasets containing sequences sampled from one of the four major GPCR Classes A, B, C, and D. The classifiers were then tested against GPCRs derived from the classes that were not included in the training sets. For example, when the dataset of Class A was used for training, tests were done against the test datasets Class (B +D) , and so on (See Table 1).

(a) Classifiers trained on Class A

Figure 1 summarizes the performance of the six classifiers on the test dataset from classes B and D. All PLS classifiers had 90% or higher accuracy rates for identifying GPCR Classes B and D, while SAM and PSI-BLAST had 70% or lower accuracy rates (Figure 1A top and Table 2). However, PLS classifiers had false positive rates of 8% to 18%, and SAM and PSI-BLAST had false positive rates of 0 and 4%, respectively (Figure 1A and Table 2). SAM and PSI-BLAST could not identify most of Class (B + D) GPCRs as indicated by high false negative rates (62 and 60%, respectively). However, PLS classifiers had low false negative rates of 4% (Figure 1B and Table 2).

(b) Classifiers trained on Class B

Consistent with the results of the classifiers trained on Class A, PLS classifiers had 90% or higher accuracy rates for identifying Class (A + D), while SAM and PSI-BLAST had low accuracy rate of 70% or lower (Figure 2A and Table 3). PLS classifiers had high false positive rates of 8 to 20%, and SAM and PSI-BLAST had no false positive (Figure 2A and Table 3). PLS classifiers had false negative rates of 0 to 6% whereas, SAM and PSI-BLAST could not identify most of the GPCRs found in Class (A + D) as indicated by high false negative rates of 62 and 60% (Figure 2B and Table 3)

(c) Classifiers trained on Class D

Figure 3 shows the results obtained from the classifiers when they were trained on the Class D and tested on Class (A + B). PLS classifiers performed better than SAM and PSI-BLAST with accuracy rates of 90% or higher, and SAM and PSI-BLAST had lower accuracy rates lower than 70% (Figure 3A and Table 4). These results are consistent with the results obtained when the classifiers were trained on Class A and tested on Class (B + D), and when trained on Class B and tested on Class (A + D). PLS classifiers had high false positive rates of 8 to 20%, and SAM and PSI-BLAST had no false positive (Figure 3A bottom; Table 4). SAM and PSI-BLAST had false negative rates of 68 and 66% respectively, while PLS classifiers had false negative rates of 0 to 6% (Figure 3B; Table 4).

2.3.5 Discussion

This study was to investigate the performance of the classifiers in identifying GPCR sequences that belong to different classes and were only remotely similar to the sequences included in the training datasets. PLS classifiers had 90% or higher accuracy rates in identifying such GPCR sequences that were extremely diverged from those included in the training datasets, whereas SAM and PSI-BLAST showed only 70% or lower accuracy rates in such cases. Both alignment-based methods (SAM) and PSI-BLAST classifiers performed poorly in interclass identification, and they misidentified many GPCRs as false negatives (higher than 50%). However, SAM and PSI-BLAST had higher specificity as shown by having low false positives. Such high specificity may have prevented SAM and PSI-BLAST from identifying sequences from classes that they were not trained on. Instead of relying on alignments, PLS used alignment-free sequence descriptors. This was very effective in identifying GPCRs sequences that were not included in the training datasets. However, PLS classifiers had high false positives. More work is needed to reduce the number of false positives from PLS classifiers.

Sequence similarities among GPCR classes are very low, and the length of sequences among GPCR classes also varies (Table 5). SAM and PSI-BLAST alignment-based classifiers could not build reliable models from such divergent sequences to identify GPCR classes that they were not trained on. On the other hand descriptors used by PLS classifiers had enough information that allowed PLS classifiers to identify GPCR classes that were not included in the training datasets. This study shows that PLS classifiers can be applied in the discovery of unknown or novel GPCR classes they have not been directly trained on. Our results clearly showed that this strategy is effective in classifying distantly related protein sequences.

2.3.6 References

1. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G. GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res. 2003;31: 294297.
2. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, and Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31:365-370.
3. Geladi P and Kowalski BR. Partial least squares regression: A tutorial. Anal. Chim. Acta 1986;185:1-17.
4. Durbin R, Eddy S, Krogh A, and Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press; Cambridge, 1998.
5. Hughey R and Krogh A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. Comput. Appl. Biosci. 1996;12:95-107.
6. 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.

2.3.7 Tables

Table 1. Numbers of samples included in the datasets including different GPCR classes

	GPCRs				Non-GPCRs	Total
Datasets	Class A	Class B	Class C	Class D		
[Training]						
Class A	50	0	0	0	50	100
Class B	0	50	0	0	50	100
Class D = (C + D)	0	0	40	10	50	100
[Test]						
Class (A + B)	50	50	0	0	100	200
Class (A+ D)	50	0	40	10	100	200
Class (B + D)	0	50	40	10	100	200

Table 2. Classifier performance when trained with GPCR Class A and tested against (B + D).

Classifiers	%Accuracy	%False positive	%False negative
PLS-ACC	93.0	8.0	4.0
PLS-Mean	91.0	16.0	4.0
PLS-AA	90.0	18.0	4.0
PLS-AA_PCA	91.0	16.0	4.0
SAM	68.0	4.0	62.0
PSI-BLAST	70.0	0.0	60.0

Table 3. Classifier performance when trained with GPCR Class B and tested against Class (A + D).

Classifiers	%Accuracy	%False positive	%False negative
PLS-ACC	92.0	14.0	4.0
PLS-Mean	90.0	20.0	0.0
PLS-AA	96.0	8.0	0.0
PLS-AA_PCA	90.0	16.0	6.0
SAM	69.0	0.0	62.0
PSI-BLAST	70.0	0.0	60.0

Table 4. Classifier performance when trained with GPCR Class D tested against Class (A + B).

Classifiers	% Accuracy	%False positive	%False negative
PLS-ACC	96.0	8.0	0.0
PLS-Mean	90.0	20.0	0.0
PLS-AA	93.0	12.0	4.0
PLS-AA_PCA	90.0	16.0	6.0
SAM	66.0	0.0	68.0
PSI-BLAST	68.0	0.0	66.0

Table 5. The average length of sequences found in each of the five GPCR classes

Class	Average length (amino acids)
A: Rhodopsin like	380 ± 106
B: Secretin like	735 ± 466
C: Metabotropic receptors	977 ± 205
D: Fungal pheromone	480 ± 109
E: cAMP receptors	425 ± 52

2.3.8 Figures

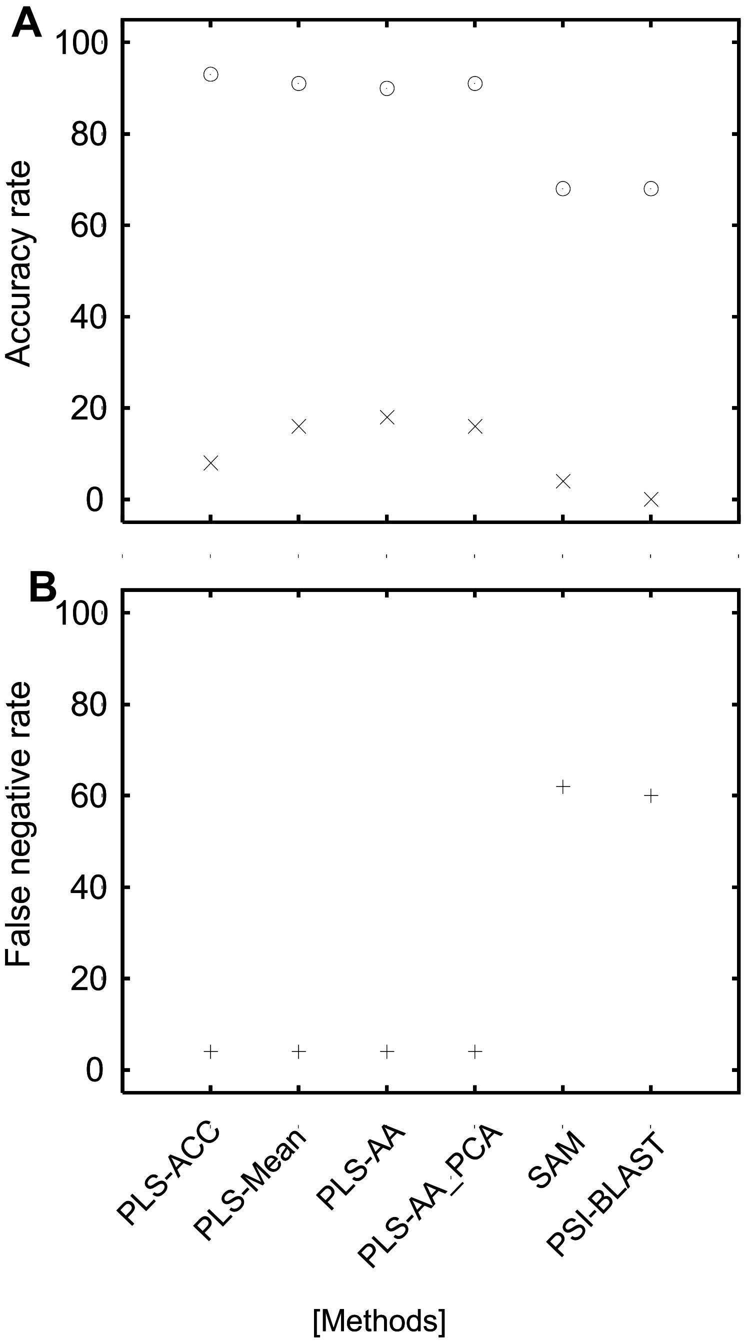


Figure 1. The performance of classifiers trained on Class A and tested on (B + D). A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

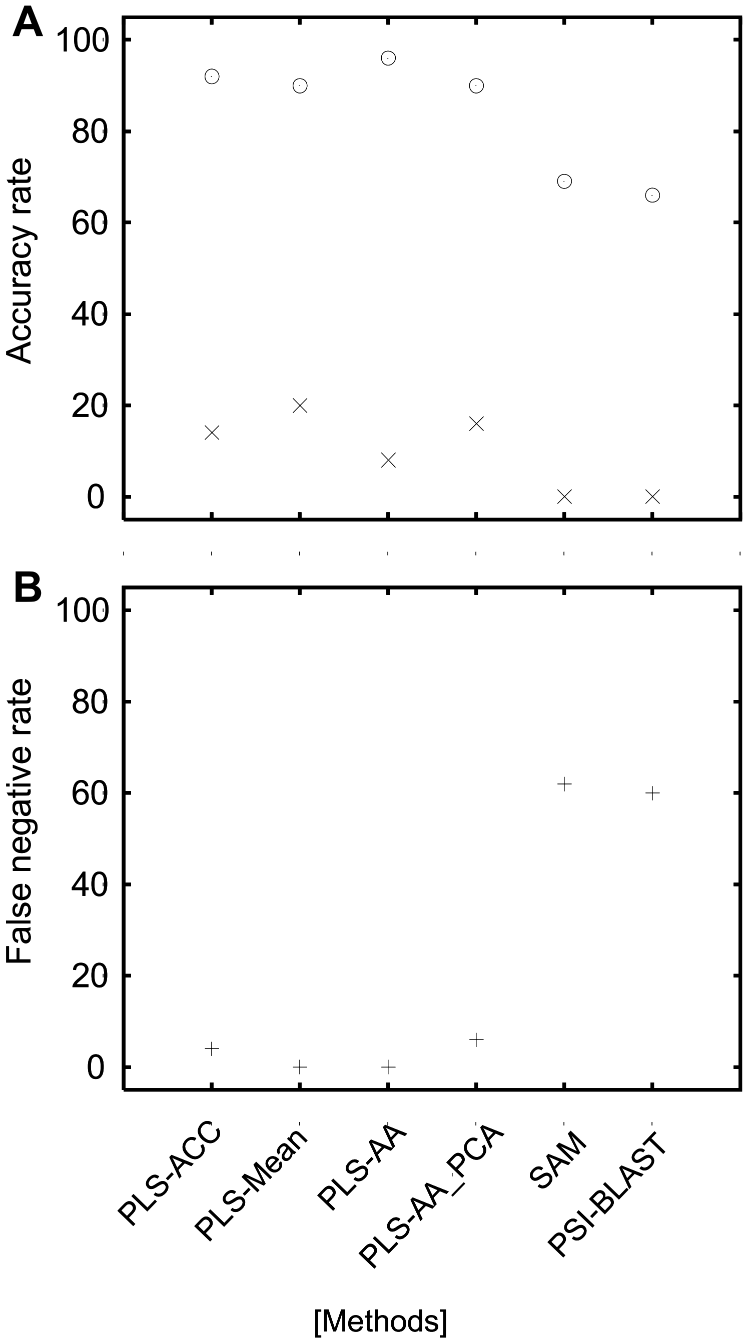


Figure 2. The performance of classifiers trained on Class B and tested on Class (A + D). A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

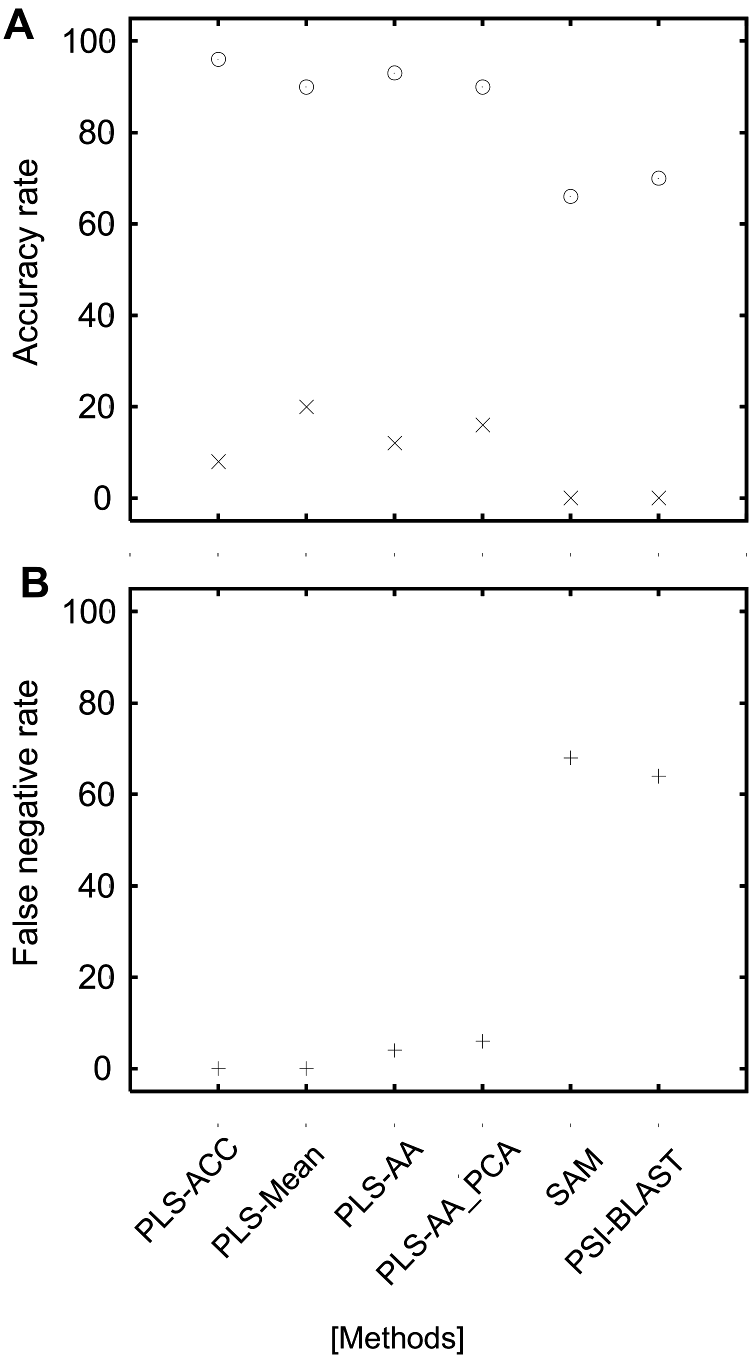


Figure 3. The performance of classifiers trained on Class D and tested on Class (A + B). A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

2.4 MINING THE *ARABIDOPSIS THALIANA* GENOME FOR HIGHLY-DIVERGENT SEVEN TRANSMEMBRANE RECEPTORS

2.4.1 Abstract

As mentioned in the Chapter 1, section 1.2.3, section 2.4 is part of the study that was published in Moriyama et al.[1]. The part I contributed to the study was the application of Partial least squares (PLS) in the mining of 7TMRs. Therefore, the description of this section is focused on my own part. In this study partial least squares method was combined with other alignment-free methods (linear discriminant analysis, quadratic discriminant analysis, support vector machines, and K-nearest neighbor) were used to search *Arabidopsis thaliana* genome for seven transmembrane receptors (7MRs). An alignment-based method profile hidden Markov models (HMMs), was also included in the study. Alignment-free methods identified 342 proteins as 7TMRs from the *Arabidopsis* genome, whereas the profile HMM classifier predicted only 15 proteins as 7TMRs. From 342 proteins predicted by alignment-free methods, 54 protein candidates were predicted to have seven transmembrane spans as well as the N-terminal region located outside of the membrane, a canonical topology of known G-protein coupled receptors (GPCRs). This study showed that the alignment-based profile HHMs appear to be too specific (conservative) when applied to mine the extremely diverged 7TMR protein family.

2.4.2 Introduction

Seven transmembrane receptors (7TMRs) also know as G-protein coupled receptors (GPCRs) are the largest superfamily of proteins found in eukaryotes, especially in the metazoan. As mentioned in Chapter 1, increasing numbers of alternative G protein-independent signaling mechanisms especially in plants have been associated with groups of 7TMRs. Therefore, in this section, I will use 7TMRs instead of GPCRs. 7TMRs are also one of the most diverse protein families. As discussed in detail in Chapter 1, the human genome includes 800 or more GPCRs, 557 are found in chicken, about 300 are found in the *Drosophila Melanogaster* genome, and more than 1000 are found in the *Caenorhabditis elegans* genome. Compared to the numbers found in other organisms, only 22 GPCRs are found in *Arabidopsis* genome. As mentioned in section 2.3, it may be possible that plants do not require as many GPCRs as animals. However, it is also possible that classifiers used to identify these proteins, e.g., profile hidden Markov models (HMMs)[2] and PSI-BLAST[3] can not identify some GPCRs from plants due to their low sequence similarities.

Currently used alignment-based protein search methods do not perform well against highly diverged proteins such as 7TMRs. As mentioned in Chapters 1 and section 2.3, these methods rely on multiple alignments for generating their models. The advantages and disadvantages of the alignment-free methods were discussed in Chapter 1.

In this study, PLS methods in combination with other alignment-free methods were used to mine 7TMRs from the *Arabidopsis thaliana* genome.

2.4.3 Materials and methods

(a) Arabidopsis genome data

28,952 protein sequences were downloaded from The Institute for Genomic Research (TIGR) *Arabidopsis thaliana* Database ftp site (ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/ATH1.pep.gz; Release 5, dated on June 10, 2004).

(b) Training data preparation for protein classification

Positive training samples (known 7TMR sequences) were obtained from GPCRDB (Information System for G Protein-Coupled Receptors, Release 9.0, last updated on June 28, 2005)[4]. In the GPCRDB, 2,030 7TMRs (originally collected from the Swiss-Prot protein database) were grouped into six major classes (Classes A - E plus the Frizzled/Smoothened family) and six putative families (ocular albinism proteins, insect odorant receptors, plant MLO receptors, nematode chemoreceptors, vomeronasal receptors, and taste receptors). Five hundred 7TMR sequences were randomly sampled and used as the positive samples. Note that "putative/unclassified" (orphan) 7TMRs and bacteriorhodopsins were not included in this dataset. These 500 7TMRs included six of the 15 known *Arabidopsis* MLO proteins.

For negative samples, 500 non-7TMR sequences longer than 100 amino acids were randomly sampled from the Swiss-Prot section of the UniProt Knowledgebase[5;6]. Positive and negative samples were combined to create a training dataset. Note that only positive samples were used to train the profile HMM classifier, SAM.

(c) Protein classification methods used

Partial least squares classifier

PLS[7] classifiers based on different descriptors were described in section 2.2.3. They are "PLS-ACC" based on auto and cross-covariance descriptor was used in this study.

Discriminant function analysis classifier

Discriminant function analysis classifiers (LDA, QDA and KNN20)[8:9] were described in Chapter 1. Amino acid index and three periodicity statistics described in[8:9] were used as descriptors in this study.

Support vector machines with amino acid composition

Support vector machines (SVMs)[10] was described in Chapter 1. SVM using amino acid compositions (SVM-aa) as descriptors was used in this study.

Support vector machines with dipeptide composition

SVM using dipeptide compositions (SVM-di) as descriptors was used in this study.

Profile hidden Markov model

Profile HMM[2] implemented in Sequence Alignment and Modeling Software System (SAM version 3.5)[11] was described in Chapter 1 and section 2.2.3. The same procedures used in the analysis in section 2.2.3 were applied for this study.

(d) Transmembrane region prediction

HMMTOP is a hidden Markov model method for topology prediction of helical transmembrane (TM) proteins[12]. It is currently one of the best TM prediction methods[13] HMMTOP 2.09[12;14] and TMHMM[15] implemented as S-TMHMM[16] were used for predicting transmembrane (TM) regions. TMHHM is also a hidden Markov model, however, in TMHMM, the model comprises seven sets of states, with each set corresponding to a type of regions in the protein sequence. Each set of states has an associated probability distribution over the 20 amino acids characterizing the compositional bias in the corresponding regions. To assess the performance HMMTOP and TMHMM, they were used to predict TM regions from 500 known 7TMR sequences used for classifier training. HMMTOP predicted 433 proteins as 7TMRs (86.6%), while only 165 (33%) were predicted to have seven TMs by TMHMM. HMMTOP predicted 97% or more to have 6-7 TMs, and with 5-9 TMs, more than 99% were included. Using TMHMM, in order to include 97% of the 500 TMR proteins, the range of predicted TM numbers needs to be from 4 to 10. Therefore, we decided to use HMMTOP in our further analysis. With HMMTOP using the range of 5-9 TMs, we should be able to cover almost all possible 7TM proteins.

2.4.4 Results and discussion

In this study, PLS was combined with five alignment-free methods LDA, QDA, KNN20, SVM-aa, and SVM-di (SVM) to mine 7TMRs from the *Arabidopsis thaliana* genome. Table 1 shows the results obtained from the alignment-free methods and alignment-based profile HMM (SAM). Each of the alignment-free methods predicted 2000-3000 proteins as 7TMR candidates and SAM predicted 15 proteins as 7TMR candidates. Taking the intersection of their results, the six alignment-free methods predicted 652 proteins as 7TMR candidates. Using the number of predicted TM regions of 5-10 as the next filter, 342 proteins were identified as 7TMR candidates (Table 1). The accession numbers for these proteins are presented in Appendix Table 2.

A canonical GPCR protein has seven transmembrane regions with its N-terminal in the outside of the membrane. HMMTOP predicted 125 proteins as having this specific topology from a total of the 28,952 predicted proteins from the *Arabidopsis* genome. From the 342 candidate proteins predicted by all of the six methods, 54 proteins were among these 7TMR candidates. It included 20 of the 22 known *Arabidopsis* 7TMRs (Table 2 for list the 54 proteins). SAM missed 7 of the 22 known 7TMRs from the *Arabidopsis* genome

From this study, we showed that the profile HMM protein classification method appears to be too specific (conservative) when applied to the extremely diverged 7TM protein family. Our premise is that there are more 7TMRs yet to be identified in the *A. thaliana* genome. Alignment-free methods are more sensitive, but they also have higher rates of false positives. More work is required to reduce the number of false positives in the alignment-free methods.

2.4.5 References

1. Moriyama EN, Strope PK, Opiyo SO, Chen Z, and Jones AM. Genome Biology 2006;7:R96.
2. Durbin R, Eddy S, Krogh A, and Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press; Cambridge, 1998.
3. 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.
4. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, Vriend G. GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res. 2003;31: 294297.
5. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al.: The Universal Protein Resource (UniProt). Nucleic Acids Res 2005;33:D154-159.
6. UniProt: The Universal Protein Resource [http://www.uni prot.org]
7. Geladi P and Kowalski BR. Partial least squares regression: A tutorial. Anal. Chim. Acta 1986;185:1-17.
8. Kim J, Moriyama EN, Warr CG, Clyne PJ, and Carlson JR. Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. Bioinformatics 2000;16:767-775.

9. Moriyama EN, and Kim J. Protein family classification with discriminant function analysis. In Genome Exploitation: Data Mining the Genome, J.P. Gustafson, ed, Springer; New York: 2005.

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

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

12. Tusnady GE, Simon I: The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849-850.

13. Cuthbertson JM, Doyle DA, and Sansom MSP. Transmembrane helix prediction: a comparative evaluation and analysis. Peds. 2005;18:295-308.

14. HMMTOP [http://www.enzim.hu/hmmtop]

15. Sonnhammer EL, von Heijne G, Krogh A: A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 1998; 6:175-182.

16. Viklund H, Elofsson A: Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci. 2004;13:1908-1917.

Table 1. Numbers of 7TMR candidates identified by various methods from the *A. thaliana* genome

	Arabidopsis thaliana genome		
Methods	Numbers of 7TMR candidates out of 28,952	Numbers of 7TMR proteins with TM region 5-10	(7TMs N-terminal)[a] out of 139 (125)[b]
SAM	16 (15)	16 (15)	8 (7)
LDA	3,211 (2,935)	909 (801)	110 (97)
QDA	2,006 (1,820)	733 (645)	100 (87)
KNN (K=20)	3,347 (3,043)	866 (767)	110 (97)
SVM-AA	2,263 (2,043)	869 (772)	114 (101)
PLS-ACC	2,004 (1,807)	616 (552)	107 (67)
SVM-di	2,671 (2,466)	843 (750)	107 (95)
Intersection[c]	652 (595)	394 (342)	64 (54)

[a] The numbers of seven transmembrane region predicted by HMMTOP with N-terminal outside.

[b] The number in parenthesis after removing sequences from splice variants.

[c] Intersection: Prediction by all six alignment-free methods (LDA, QDA, KNN (K=20), SVM-AA, SVM-di, and PLS-ACC).

Table 2. The 54 *Arabidopsis thaliana* 7 TMR candidates with seven transmembrane regions with N-terminal outside predicted by six alignment-free methods.

Group	TAIR Locus	Length	*Description*
[Multiple members from a big gene family (> 5 members)]			
Nodulin MtN3 family	At1g21460.1	247	68414.m02683 nodulin MtN3 family protein contains similarity to MTN3 (nodule development protein) GB:Y08726 GI:1619601 from [Medicago truncatula]
	At3g16690.1	230	68416.m02132 nodulin MtN3 family protein contains Pfam PF03083 MtN3/saliva family
	At3g28007.1	251	68416.m03496 nodulin MtN3 family protein contains Pfam PF03083 MtN3/saliva family; similar to LIM7 GI:431154 (induced in meiotic prophase in lily microsporocytes) from [Lilium longiflorum]
	At3g48740.1	289	68416.m05322 nodulin MtN3 family protein similar to MtN3 GI:1619602 (root nodule development) from [Medicago truncatula]
	At4g25010.1	281	68417.m03588 nodulin MtN3 family protein similar to MtN3 GI:1619602 (root nodule development) from [Medicago truncatula]
	At5g13170.1	292	68418.m01508 nodulin MtN3 family protein similar to MtN3 GI:1619602 (root nodule development) from [Medicago truncatula]; identical to cDNA senescence-associated protein (SAG29) mRNA, partial cds GI:4426938
	At5g23660.1	285	68418.m02774 nodulin MtN3 family protein similar to MtN3 GI:1619602 (root nodule development) from [Medicago truncatula]
	At5g50800.1	294	68418.m06293 nodulin MtN3 family protein similar to MtN3 GI:1619602 (root nodule development) from [Medicago truncatula]
MLO family	At1g11000.1	573	68414.m01263 seven transmembrane MLO family protein / MLO-like protein 4 (MLO4) identical to membrane protein Mlo4 [Arabidopsis thaliana] gi|14091578|gb|AAK53797; similar to MLO protein SWISS-PROT:P93766, NCBI_gi:1877221 [Hordeum vulgare][Barley]
	At1g26700.1	554	68414.m03252 seven transmembrane MLO family protein / MLO-like protein 14 (MLO14) identical to membrane protein Mlo14 [Arabidopsis thaliana] gi|14091598|gb|AAK53807; similar to MLO protein SWISS-PROT:P93766, NCBI_gi:1877221 [Hordeum vulgare][Barley]
	At1g42560.1	467	68414.m04907 seven transmembrane MLO family protein / MLO-like protein 9 (MLO9) nearly identical to membrane protein Mlo9 [Arabidopsis thaliana] GI:14091588; similar to MLO protein SWISS-PROT:P93766, NCBI_gi:1877221 [Hordeum vulgare][Barley]
	At2g33670.1	501	68415.m04126 seven transmembrane MLO family protein / MLO-like protein 5 (MLO5) identical to MLO-like protein 5 (AtMlo5) [Arabidopsis thaliana] SWISS-PROT:O22815; similar to MLO protein SWISS-PROT:P93766, NCBI_gi:1877221 [Hordeum vulgare][Barley]

Table 2 (continued).

Group	TAIR Locus	Length	*Description*
	At2g44110.1	496	68415.m05485 seven transmembrane MLO family protein / MLO-like protein 15 (MLO15) identical to MLO-like protein 15 (AtMlo15) SP:O80580 from [Arabidopsis thaliana]; similar to MLO protein SWISS-PROT:P93766, NCBI_gi:1877221 [Hordeum vulgare][Barley]
	At4g24250.1	478	68417.m03480 seven transmembrane MLO family protein / MLO-like protein 13 (MLO13) identical to membrane protein Mlo13 [Arabidopsis thaliana] gi|14091596|gb|AAK53806; similar to MLO protein SWISS-PROT:P93766, NCBI_gi:1877221 [Hordeum vulgare][Barley]
	At5g53760.1	573	68418.m06680 seven transmembrane MLO family protein / MLO-like protein 11 (MLO11) identical to membrane protein Mlo11 [Arabidopsis thaliana] gi|14091592|gb|AAK53804; similar to MLO protein SWISS-PROT:P93766, NCBI_gi:1877221 [Hordeum vulgare][Barley]
Expressed protein family 1	At1g77220.1	484	68414.m08994 expressed protein contains Pfam profile PF03619: Domain of unknown function
	At4g21570.1	294	68417.m03120 expressed protein contains Pfam profile PF03619: Domain of unknown function
[Multiple members from a small gene family]			
TOM3 family	At1g14530.1	293	68414.m01723 tobamovirus multiplication protein 3, putative / TOM3, putative (THH1) identical to THH1 (GI:15706301) [Arabidopsis thaliana]; supporting cDNA gi|15706300|dbj|AB057678.1|
	At2g02180.1	303	68415.m00154 tobamovirus multiplication protein 3 (TOM3) identical to tobamovirus multiplication protein (TOM3) GI:15425641 from [Arabidopsis thaliana]
	At4g21790.1	291	68417.m03152 transmembrane protein-related (TOM1) contains some similarity to transmembrane protein TOM3 GI:15425641 from [Arabidopsis thaliana]; identical to cDNA TOM1 GI:9967414

Table 2 (continued).

Group	TAIR Locus	Length	*Description*
GNS1/SUR4 membrane family	At1g75000.1	281	68414.m08707 GNS1/SUR4 membrane family protein contains Pfam profile PF01151: GNS1/SUR4 family
	At3g06470.1	278	68416.m00749 GNS1/SUR4 membrane family protein similar to SP|P39540 Elongation of fatty acids protein 1 {Saccharomyces cerevisiae}; contains Pfam profile PF01151: GNS1/SUR4 family
	At4g36830.1	289	68417.m05223 GNS1/SUR4 membrane family protein weak similarity to long chain polyunsaturated fatty acid elongation enzyme [Isochrysis galbana] GI:17226123; contains Pfam profile PF01151: GNS1/SUR4 family
Expressed protein family 2	At1g10660.1	320	68414.m01208 expressed protein
	At2g47115.1	300	68415.m05884 expressed protein
	At5g62960.1	347	68418.m07899 expressed protein
Perl1-like family	At1g16560.1	342	68414.m01983 Per1-like family protein contains Pfam profile PF04080: Per1-like
	At5g62130.1	343	68418.m07798 Per1-like protein-related
Expressed protein family 3	At3g09570.1	439	68416.m01137 expressed protein
	At5g42090.1	439	68418.m05124 expressed protein
Expressed protein family 4	At1g49470.1	302	68414.m05544 expressed protein contains Pfam profile PF04819: Family of unknown function (DUF716) (Plant viral-response family)
	At5g19870.1	276	68418.m02363 expressed protein contains Pfam profile PF04819: Family of unknown function (DUF716) (Plant viral-response family)

Table 2 (continued).

Group	TAIR Locus	Length	*Description*
Expressed protein family 5	At3g63310.1	239	68416.m07121 expressed protein low similarity to N-methyl-D-aspartate receptor-associated protein [Drosophila melanogaster] GI:567104; contains Pfam profile PF01027: Uncharacterized protein family UPF0005
	At4g02690.1	248	68417.m00364 hypothetical protein low similarity to N-methyl-D-aspartate receptor-associated protein [Drosophila melanogaster] GI:567104, NMDA receptor glutamate-binding subunit [Rattus sp.] GI:8248741; contains Pfam profile PF01027: Uncharacterized protein
[Single copy genes]			
GCR1	At1g48270.1	326	68414.m05392 G protein coupled receptor-related identical to putative G protein coupled receptor GI:2104224 from [Arabidopsis thaliana]
	At1g57680.1	362	68414.m06545 expressed protein
	At2g41610.1	310	68415.m05142 expressed protein
	At2g31440.1	250	68415.m03841 expressed protein identical to cDNA endonuclease III homologue (nth1 gene) GI:11181951
	At3g04970.1	397	68416.m00540 zinc finger (DHHC type) family protein similar to Golgi-specific DHHC zinc figer protein [Mus musculus] GI:21728103; contains Pfam profile PF01529: DHHC zinc finger domain
RGS1	At3g26090.1	459	68416.m03249 expressed protein
	At3g59090.1	367	68416.m06587 expressed protein
	At4g20310.1	513	68417.m02966 peptidase M50 family protein / sterol-regulatory element binding protein (SREBP) site 2 protease family protein weak similarity to SP|O54862 Membrane-bound transcription factor site 2 protease (EC 3.4.24.-) (Sterol-regulatory element-binding

Table 2 (continued).

[Single member from a small gene family]		
At2g01070.1	496	68415.m00013 expressed protein similar to membrane protein PTM1 precursor isolog GB:AAB65479
At3g19260.1	296	68416.m02443 longevity-assurance (LAG1) family protein similar to Alternaria stem canker resistance protein (ASC1) [Lycopersicon esculentum] GI:7688742; contains Pfam profile PF03798: Longevity-assurance protein (LAG1)
At2g35710.1	497	68415.m04380 glycogenin glucosyltransferase (glycogenin)-related low similarity to glycogenin-2 from Homo sapiens [SP|O15488]
At2g16970.1	414	68415.m01955 expressed protein; expression supported by MPSS
At1g15620.1	343	68414.m01877 expressed protein ; expression supported by MPSS
At1g63110.2	397	68414.m07132 cell division cycle protein-related contains 9 transmembrane domains; similar to PIG-U (GI:27372215) [Rattus norvegicus]; similar to Cell division cycle protein 91-like 1 (CDC91-like 1 protein) (PIG-U) (Swiss-Prot:Q9H490) [Hom sapiens]
At4g36850.1	374	68417.m05225 PQ-loop repeat family protein / transmembrane family protein similar to SP|Q10482 Seven transmembrane protein 1 {Schizosaccharomyces pombe}; contains Pfam profile PF04193: PQ loop repeat
At5g27210.1	297	68418.m03246 expressed protein weak similarity to seven transmembrane domain orphan receptor [Mus musculus] GI:4321619
[Single member from a large gene family]		
At1g71960.1	662	68414.m08318 ABC transporter family protein similar to breast cancer resistance protein GB:AAC97367 from [Homo sapiens]
At3g01550.1	383	68416.m00085 triose phosphate/phosphate translocator, putative similar to SWISS-PROT:P52178 triose phosphate/phosphate translocator [Cauliflower]{Brassica oleracea}
At5g23990.1	657	68418.m02819 ferric-chelate reductase, putative similar to ferric-chelate reductase (FRO1) [Pisum sativum] GI:15341529; contains Pfam profile PF01794: Ferric reductase like transmembrane component
At5g37310.1	*564*	68418.m04481 endomembrane protein 70, putative multispanning membrane protein, Homo sapiens, EMBL:HSU94831

CHAPTER 3

CLASSIFICATION OF CYCLOPHILIN PROTEIN FAMILIES

3.1 ABSTRACT

Cyclophilins comprise a highly conserved, ubiquitous family of proteins first identified as the intracellular receptors for the immunosuppressant drug cyclosporin A. The *Arabidopsis* genome contains the largest number of cyclophilins 29, compared to 19 in human, and 14 in *Drosophila melanogaster*. However, total numbers of plant cyclophilins found in sequence databases are small compared to that of other organisms. This implies that many cyclophilins are not yet identified. In order to identify more cyclophilins from available plant sequence data, I examined alignment-free classifiers based on partial least squares (PLS) using physico-chemical properties for the identification of single-domain cyclophilins and tetratricopeptide (TPR) multiple-domain cyclophilins. PLS with descriptors selected by t-test or rank test after auto and cross covariance (ACC) transformation (PLS_T-ACC and PLS_R-ACC) had low false positives compared to PLS with all ACC descriptors (PLS-ACC). PLS_T-ACC performed better than SAM and PSI-BLAST in the identification of cyclophilins from the *Arabidopsis* genome. PLS_T-ACC identified 290 proteins as single-domain cyclophilins

and 110 proteins as TPR multiple-domain cyclophilins. Both SAM and PSI-BLAST identified 31 proteins as single-domain cyclophilins and 91 and 481 proteins as TPR multiple-domain cyclophilins respectively from the *Arabidopsis* genome. When the methods were used to search rice genome, PLS_T-ACC identified 1259 and 304 proteins as single-domain and TPR multiple-domain cyclophilins respectively. SAM and PSI-BLAST identified 48 and 29 proteins as single-domain cyclophilins and 122 and 96 as TPR multiple-domain cyclophilins, respectively. This study shows that reducing the number of descriptors by selecting only those that are important for discriminating cyclophilins from non-cyclophilins can reduce the number of false positives. It also shows that alignment-based SAM and PSI-BLAST are too conservative when used to search proteins with heterogeneous domain. PLS_T-ACC will be useful for identifying new/unknown cyclophilins from plant genomic databases as they become available.

3.2 INTRODUCTION

Cyclophilins (CYPs) are originally identified as the cellular targets of cyclosporin A (CsA), a fungal metabolite with potent immunosuppressive activity[1]. As described in Chapter 1, with the FK506-binding proteins (FKBPs), they form a family of immunosuppressant receptors, called immunophilins. Although these two groups of proteins share no sequence similarity, both have peptidyl-prolyl isomerase (PPIase, EC 5.2.1.8) activity that catalyzes the rapid *cis*-to-*trans* isomerization of peptide bonds N-side of proline residues in polypeptide chains[2]. This *cis-trans* isomerazation is an important step in protein folding, and a critical determinant of protein structures.

As discussed in Chapter 1, the *Arabidopsis thaliana* genome contains the largest number of cyclophilin proteins, 29 of them in total, in spite of its relatively smaller genome[3;4]. However, the number of cyclophilin sequences available from plants found in Integrated Documentation Resource of Protein Families, Domains and Functional Sites (InterPro[5] Release 14.1, dated, February 19th, 2007) is much smaller compared to those from animals and other organisms. This is particularly surprising since green plants appear to have the largest cyclophilin family. Table 1 provides the current reality of plant cyclophilin data available in the databases. Very few cyclophilins are known from non-model plant species. 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 sequence data. More efficient and sensitive mining methods are required.

Detecting protein sequence similarities is often the base of predicting protein functions. If sequences or structures of proteins are sufficiently similar, characteristics known to be true for one protein or one family of proteins can be transferred to others. Therefore, function prediction problems are often referred to as protein classification problems. The most popularly used method for protein family classification is Basic Local Alignment Search Tool (BLAST)[6]. It searches sequence databases for local similarities to the query sequence. The more the sequences have diverged, the harder it becomes to recognize true sequence similarities separated from the similarity by random chance. For weaker similarities, most frequently sequence patterns or profiles are searched. Such more sensitive search methods include: e.g., PSI-BLAST [7], PROSITE [8], and Pfam[9]. As discussed in Chapters 1 and 2, building sequence models (e.g., patterns, motifs, profiles) for these currently used methods requires reliable alignments. Another problem with the currently available methods is that the sequence models are built using only "positive" samples (proteins of interest).

In order to identify sequences that do not have enough similarities, methods that do not rely on multiple alignments must be used. As Galat[10] has done, hydrophobicity, bulkiness, and other physicochemical properties of proteins can potentially be used for their classification, and this approach does not require aligning sequences.

The objectives for this study are 1) to develop protein classification methods based on partial least squares (PLS) that can effectively identify cyclophilin protein families, and 2) to mine cyclophilin proteins from the *Arabidopsis* and rice genomes.

3.3 MATERIALS AND METHODS

3.3.1 Datasets preparation

Cyclophilin data were downloaded from (InterPro[5], Release 13.1, October 17th, 2006). The cyclophilin families (InterPro: IPR002130 Peptidyl-prolyl cis-trans isomerase, cyclophilin-type)

(a) Positive samples.

Two hundred and eighty single-domain cyclophilin sequences were randomly divided into two. One was used for training, and the second set used for the test. There are very few sequences available from multiple-domain cyclophilins in the InterPro database. Tetratricopeptide (TPR) multiple-domain is the largest multiple-domain cyclophilins found in InterPro, and only 36 are available. The number of TPR multiple-domain cyclophilin sequences available was not enough to generate two independent datasets. Only one dataset generated, and instead of using an independent test dataset, the cross-validation test was performed (described in section 3.4.1(b)).

(b) Negative samples.

Non-cyclophilin protein sequences were randomly sampled from the SWISS-PROT[11] protein database.

(c) Training and test

Different types of datasets were produced as listed in Table 2. Note that profile hidden Markov models (HMMs)[12] implemented using Sequence Alignment and Modeling System (SAM)[13] and PSI-BLAST do not use negative samples for training; only positive samples were included for their training datasets. For the single-domain cyclophilin, two independent datasets were prepared: one for training and the other for testing.

(d) Arabidopsis genome data

28,952 *A. thaliana* protein sequences were downloaded from The Institute for Genomic Research (TIGR) *Arabidopsis thaliana* Database ftp site (ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/ATH1.pep.gz; Release 5, dated on June 10, 2004).

(e) Rice genome data

62,877 rice protein sequences were downloaded from The Institute for Genomic Research (TIGR) Rice Database site http://www.tigr.org/tdb/e2k1/osa1/data_download.shtml; Protein sequences .pep) with the option “Select All chromosome”; Release 5, updated on December 14, 2006.)

3.3.2 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.

(a) Within-group analysis

In this experiment, classifiers were trained and tested using the datasets generated from the same cyclophilin group (single-domain or multiple-domain cyclophilins) as shown in Table 2. For single-domain cyclophilins, training and testing was done using independent datasets. For TPR multiple-domain cyclophilins, classifiers were trained and cross-validation analysis was performed on the same dataset as explained in section 3.4.1 (b).

(b) Between-group 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 2. This is to test how a classifier could identify cyclophilin groups not included in the training sets.

3.3.3 Descriptors

(a) Physico-chemical properties of amino acids

Opiyo and Moriyama[14] developed 5 descriptors (PC1 to PC5) from 12 physico-chemical properties of amino acids. These descriptors were also discussed in Chapter 2. The same set of these five descriptors were also used in this study.

(b) Auto/cross covariance (ACC) transformation

As mentioned in Chapters 1 and 2, 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 Chapter 2 was used to transform datasets of single-domain cyclophilin, and TPR multiple-domain cyclophilin sequences. Using the five descriptor set, ACC with a maximum lag = 30 were performed on each sequence giving a total of 775 descriptors.

3.3.4 Selection of important descriptors after auto and cross-covariance transformation

In Chapter 2, it was observed that PLS classifier using descriptors transformed by ACC had high false positives. The hypothesis is that the number of false positives of PLS classifiers can be reduced if we select only descriptors that are important in discriminating cyclophilins from non-cyclophilins after the ACC transformation. To select the important descriptors, we can either use the t-test or rank test.

(a) 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 statistic is a ratio as shown in (equation 1). The top part of the ratio is simply the difference between the two means or averages. The bottom part is a measure of the variability or dispersion of the groups.

$$\mathrm{t} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s^2{}_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (1)$$

where $\overline{X}_1$, and $\overline{X}_2$ are the means of groups X_1 and X_2, respectively, $s^2{}_p$ is the pooled sample variance of groups X_1 and X_2, and n_1 and n_2 are the number of samples for the groups X_1 and $X_{2,}$ respectively. To test the significance of the t-test, a risk level (called the alpha level) is set. For example if the alpha level is set at .05, it means that five times out of a hundred one would find a statistically significant difference between the means even if there is no difference. The degree of freedom (df) of a t-test is the total number of observations in both groups minus 2. Given the alpha level, df, and the statistics, the significance difference is obtained, one can look the statistics up in a standard table of significance to determine whether the statistics is large enough to be significant. If it is, one can conclude that there is a difference between the means for the two groups. In this study, the t-test was used for selecting descriptors that can discriminate two groups of sequences significantly, for example, from 755 ACC descriptors.

(b) Non-parametric Wilcoxon rank-sum test

The t-test is a parametric test assuming a normal distribution. However, protein sequences found in the databases are not normally distribution. Wilcoxon rank-sum test parametric was used in this study. Wilcoxon rank-sum test is used in instances where there are independent data for which one wants to compare data for two different groups. Wilcoxon rank-sum test involves in calculating a statistics called *U*. The approach that is used to calculate the *U* statistics consists of the following steps:

1) Rank all observations in increasing order of magnitude, ignoring which group they come from. If two observations have the same magnitude, regardless of the group, then they are given an average ranking.
2) Add up the ranks in the smaller of the two groups. Call this "group 1" and call the larger group "group 2". If the two groups are of an equal size then either one can be chosen.
3) *U* is then calculated using equation 2 below.

$$U = \mathrm{n}_1\mathrm{n}_2 + \frac{n_1(n_1+1)}{2} - \mathrm{R}_1 \tag{2}$$

where n_1 and n_2 are two group sizes, and R_1 is the sum of the rank in the group 1.

4) For a small sample less than or equal to 20, p-value is obtained from a *U* table.
5) For a large sample, p-value is calculated using a normal distribution using equations below.

$$Z = (U - m_u) / \sigma_u \tag{3}$$

$$m_\mathrm{u} = (n_1 n_2)/2 \tag{4}$$

$$\sigma_u = \frac{\sqrt{n_1 n_2 (n_1 + n_2 + 1)}}{12} \tag{5}$$

where Z is a standard normal whose significance can be checked in a normal distribution table, m_u and σ_u are the mean and the standard deviation of U if the null hypothesis is true, and n_1 and n_2 are two group sizes. In this study, Wilcoxon rank-sum test was used in addition to the t-test described above for selecting descriptors.

(c) Selecting the ACC descriptors

The t-test and rank test were performed on the training datasets of single-domain cyclophilin and TPR multiple-domain cyclophilins after ACC transformation with five descriptors (PC1-PC5). Significant descriptors were selected at an alpha level of 0.01. From the 775 descriptors, 690 and 702 descriptors were selected for the single-domain cyclophilins by t-test and by rank test respectively. For the TPR multiple-domain cyclophilins, 647 and 665 descriptors were selected by t-test and rank test, respectively.

3.3.5 Classifiers

(a) Partial least squares

Partial least squares using descriptors transformed by ACC (PLS-ACC) was discussed in Chapter 2. In this study, besides using PLS-ACC, partial least squares with descriptors selected by t-test after ACC transformation (PLS_T-ACC) and partial least squares with descriptors selected by rank test after ACC transformation (PLS_R-ACC) were included for cyclophilins classification. For single-domain cyclophilins, the cut off points for PLS-ACC, PLS_T-ACC, and PLS_R-ACC were 0.446, 0.470, and 0.467, respectively. These cut off points were obtained by minimum error point (MEP) used by Karchin et al.[15] using the training dataset. MEP was also used to find the cut off points for the TPR multiple-domain cyclophilins using training datasets. They were 0.452, 0.477, and 0.482 for PLS-ACC, PLS_T-ACC, and PLS_R-ACC, respectively. MEP was explained previously in Chapter 2 section 2.2.3 (g).

(b) Profile hidden Markov model

Profile hidden Markov model (HMM) was discussed in Chapters 1 and 2. A cut-off E-value of 1.02, and 1.23 were obtained for single-domain cyclophilin and TPR multiple-domain cyclophilins, respectively, using MEP.

(c) PSI-BLAST

PSI-BLAST is an alignment-based method and it was discussed in Chapters 1 and 2. A cut-off E-value of 2.3 and 2.6 were obtained for single-domain cyclophilins and TPR multiple-domain cyclophilins respectively. They were obtained from training datasets using MEP.

3.4 RESULTS

3.4.1 Within-group classification

(a) Single-domain cyclophilins

Figure 1 shows the performance of classifiers on a test dataset of 140 positive (cyclophilins) and 1000 negative (non-cyclophilins) sequences. The accuracy rates for PLS-ACC, PLS_T-ACC, PLS_R-ACC, SAM, and PSI-BLAST were 97.2, 99.1, 98.7, 97.3, and 95.8%, respectively (Figure 1A; top). PLS-ACC had the highest false positive rate of 3%, lower than PLS_T-ACC and PLS_R-ACC, which were 0.8 and 1%, respectively. SAM and PSI-BLAST had low false positive rates of 0.2 and 0.3%, respectively (Figure 1A; bottom). SAM and PSI-BLAST had high false negatives of 15 and 22% respectively (Figure 1 and Table 3). PLS-ACC, PLS_T-ACC, and PLS_R-ACC had low false negative rates of 3, 1.5, and 1.5%, respectively.

(b) TPR multiple-domain cyclophilins

Since there were not enough sequences to create an independent test dataset, the cross-validation procedure was done for TPR multiple-domain cyclophilins. The cross-validation procedure was performed as follows. There were 72 sequences in the training dataset. The first sample was removed from the dataset. The classifier was trained with the remaining 71 samples, and first left-out sample was used for classification. The first sample was put back. The second sample was removed, the classifier was trained with the remaining 71 samples, and the classification was done against the second sample. This process was repeated for each of the 72 samples present in the dataset until all were classified. The 72 classification results were combined in the form of a confusion matrix, and the statistics discussed in Chapter 2 (accuracy rate, false positive rate, and false negative rate) were calculated. A confusion matrix contains information about actual and predicted classifications done by a classification system. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class (Table 4).

Figure 2 and Table 5 show the cross-validation test results for the five classifiers. The accuracy rates were 91.6, 94.4, 94.4, 91.6, and 83.3% for PLS-ACC, PLS_T-ACC, PLS_R-ACC, SAM, and PSI-BLAST, respectively. And the false positive rates were 13.8, 8, 8, 0, and 5% for PLS-ACC, PLS_T-ACC, PLS_R-ACC, SAM, and PSI-BLAST, respectively. As observed in single-domain cyclophilins, SAM and PSI-BLAST had high false negative rates of 16.5 and 25%, respectively. PLS classifiers had low false negative rates of 3.3, 2.7, and 2.7% for PLS-ACC, PLS_T-ACC, and PLS_R-ACC, respectively (Figure 2B).

3.4.2 Between-group classification

(a) Single-domain cyclophilins

To test how classifiers can identify TPR multiple-domain cyclophilins while trained using single-domain cyclophilins, a training dataset of 280 sequences including equal number of positive and negative samples were used to train the classifiers. The classifiers were tested on a dataset that included 36 TPR multiple-domain cyclophilins and 200 noncylophilins (Table 1). Figure 3 and Table 6 summarize the results. The accuracy rates of PLS-ACC, PLS_T-ACC, PLS_R-ACC, SAM, and PSI-BLAST were 92.3, 94.9, 93.2, 90.6, and 89.8%, respectively (Figure 3A top; Appendix Table 17). False positive rates were 6, 5, 6, 2.5, and 6% for PLS-ACC, PLS_T-ACC, PLS_R-ACC, SAM, and PSI-BLAST respectively (Figure 3A bottom; Appendix Table 17). SAM and PSI-BLAST had high false negative rates of 33%, and PLS-ACC, PLS_T-ACC, and PLS_R-ACC had 16.7, 11.1, and 11.1%, respectively.

(b) TPR multiple-domain cyclophilins

The classifiers were trained with a dataset of 36 TPR multiple-domain cyclophilins and 36 non-cyclophilins and tested on a test dataset of 140 single-domain cyclophilins and 1000 non-cyclophilins. Figure 4 and Table 7 show the performances of the classifiers on the test dataset. PLS-ACC, PLS_T-ACC, PLS_R-ACC, SAM, and PSI-BLAST had accuracy rates of 90.3, 93.4, 92.5, 92.5 and 89.4%, respectively. The false positive rates for the classifiers were 10, 6.5, 7.5, 3, and 6% for PLS-ACC, PLS_T-ACC, PLS_R-ACC, SAM, and PSI-BLAST, respectively. SAM and PSI-BLAST had very high false negative rates of 39 and 42.9%, respectively. PLS-ACC, PLS_T-ACC, and PLS_R-ACC had low false negative rates of 7.1, 6.7 and 7.1%, respectively.

3.4.3 Searching the *Arabidopsis* genome for cyclophilins

Only PLS-ACC and PLS_T-ACC were included in searching the genome, because no significant differences were observed in the performances of PLS_T-ACC and PLS_R-ACC. Table 8 summarizes the results of searching the *Arabidopsis* genome using PLS methods and other methods. When all the 775 descriptors obtained from ACC with the maximum lag = 30 were used, PLS-ACC predicted 980 sequences as single-domain cyclophilin proteins, but when the analysis was performed using PLS_T-ACC, the number of sequences predicted as single-domain cyclophilin proteins were reduced to 290. Twenty out of the known 21 *Arabidopsis* single-domain cyclophilins and 8 known multiple-domain cyclophilins including TPR (At2g15790) multiple-domain cyclophilin (reported by Romano et al.[16]) were among the 290 proteins predicted. A single-domain cyclophilin At5g35100 was not identified by both PLS-ACC and PLS_T-ACC methods. The lists of the 290 proteins are presented in Appendix Table 3. Both SAM and PSI-BLAST predicted 31 proteins as cyclophilin proteins (Appendix Table 4). These include all the known 21 single-domain cyclophilins and the known 8 multiple-domain cyclophilin proteins reported by Romano et al.[16]. PLS-ACC predicted 218 sequences, and PLS_T-ACC predicted 110 sequences as TPR multiple-domain cyclophilin proteins. The accession numbers of TPR multiple-domain cyclophilin proteins predicted by PLS_T-ACC are presented in Appendix Table 5. SAM predicted 91 and PSI-BLAST predicted 484 sequences as TPR multiple-domain proteins respectively (Table 8). These sequences are presented in Appendix Tables 6 and 7 for SAM and PSI-BLAST, respectively. Out of 484 sequences predicted by PSI-BLAST, 432 sequences have either TPR domain or are TPR proteins. PLS_T-ACC and SAM did not predict TPR domain or TPR proteins.

3.4.4 Searching the rice genome for cyclophilins

Since PLS-ACC seems to have high false positives compared to PLS_T-ACC, only PLS_T-ACC was used to search the rice genome. PLS_T-ACC predicted 1259 sequences excluding duplicates due to the splice variants as single-domain cyclophilin proteins (Table 8). Out of the 1259 predicted proteins, 747 were hypothetical proteins, 207 were expressed proteins, 84 were retrotransposon, 61 were transposon, 15 were known cyclophilins, and the rest were other proteins. Accession numbers for these sequences are included in Appendix Table 8. SAM and PSI-BLAST predicted 48 and 29 sequences as single-domain cyclophilin proteins respectively (Tables 8, Appendix Tables 9 and 10). PLS_T-ACC predicted 304 sequences as TPR multiple-domain proteins (Table 8, Appendix Table 11). Of the 304 predicted proteins, 134 were hypothetical proteins, 38 were expressed proteins, 14 were retrotransposon, 15 were transposon, and 12 were known cyclophilins proteins, respectively. SAM and PSI-BLST predicted 122 and 96 sequences as TPR multiple-domain cyclophilin proteins, respectively (Table 8, Appendix Tables 12 and 13).

3.5 DISCUSION

PLS classifiers using descriptors developed from physico-chemical properties of amino acids were used in the classification of cyclophilins. PLS_T-ACC and PLS_R-ACC had lower false positives compared to PLS-ACC. The decrease in the number of false positives observed in PLS_T-ACC and PLS_R-ACC might have been due to the selection of significant descriptors. No significant differences were observed in the number of false negatives between PLS-ACC and PLS_T_ACC or PLS_R-ACC. This indicates that reducing the number of descriptors from 775 does not affect the sensitivity of PLS classifiers in classification of cyclophilins.

PLS_T-ACC and PLS_R-ACC performed better than SAM and PSI-BLAST both within-group and between-group experiments. Such classifiers can be able to identify both single-domain and multiple-domain cyclophilins regardless of which cyclophilin sequences are included in the training dataset. Such classifiers are also expected to be useful for identifying new/unknown cyclophilins. SAM and PSI-BLAST performed poorly because they used alignments to build their models. Low sequence similarities found in cyclophilins, (for example, in *Arabidopsis*, the sequence similarity within the cyclophilins ranges from 10-90%), made SAM and PSI-BLAST miss some cyclophilins from the test datasets.

PLS_T-ACC predicted 747 hypothetical proteins as single-domain cyclophilin proteins and 134 hypothetical proteins as TPR multiple-domain cyclophilin proteins. Some of these proteins may be false positives; however some cyclophilin proteins in InterPro are annotated as hypothetical protein. For example, single-domain cyclophilin (Q9C9C7) and multiple-domain cyclophilin protein (Q9LXM7) are named hypothetical proteins. There were also large numbers of retrotransposon as well as transposon proteins predicted as cyclophilins. Again, some of the cyclophilins found in InterPro are named transposon, e.g., (Q7G6E3). Q7G6E3 is an archetypal cystolic cyclophilin similar to human cyclophilins A, B. It should be noted that more evidence is needed to confirm that the protein candidates predicted by PLS_T-ACC classifier from the *Arabidopsis* and rice genomes are cyclophilins. One possibility would be to check the secondary structures of the predicted proteins.

The number of sequences predicted by SAM and PSI-BLAST as single-domain cyclophilins were small compared to that predicted by PLS_T-ACC. However, PSI-BLAST predicted the largest number of TPR multiple-domain in the *Arabidopsis* genome. Most of the predicted sequences are either TPR proteins or they have TPR domains. It is most likely that some of these sequences predicted by PSI-BLAST are false positives. A large number of false positive of false positive by PSI-BLAST might have occurred due to the few number of sequences included in the training dataset. Using only 36 sequences for the training might have caused PSI-BLAST to build the model for searching TPR proteins but not TPR multiple-domain cyclophilin proteins.

TPR domain consists of a 34 amino-acid motif, usually as multiple tandem repeats in proteins with many cellular functions, including mitosis, transcription, protein transport, and development[17]. The Arabidopsis TPR multiple-domain contains four copies of the 34 amino-acid TPR motif. Other proteins that contain TPR motifs include members of the FKBP binding proteins, organelle-targeting proteins, and a protein phosphatase[17]. Because TPR-containing proteins in general are diverse, and TPR motif is more conserved than cyclophilin domain, it is most likely that PSI-BLAST model for searching TPR proteins but not TPR multiple-domain cyclophilin proteins.

3.6 REFERENCES

1. Handschumacher RE, Harding MW, Rice J, Drugge R J, and Spelcher DW. Cyclophilin: A specific cytosolic binding protein for cyclosporine A. Science 1984;226:544-547.
2. Miguel A, Xiaoyun W, Steven DH, and Joseph H. Prolyl isomerases in yeast. Frontiers in Bioscience 2004;9:2420-2446.
3. Adams B, Musiyenko A, Kumar R, and Batick S. A novel class of dual family Immunophilins. J. Biol. Chem. 2005;280: 24308-24314.
4. Brazin KN, Mallis RJ, and Fulton DB. Regulation of the tyrosine kinase Itk by peptidyl-prolyl isomerase cyclophilin A. Proc Natl Acad Sci USA. 2002;99:1899-1904.
5. Mulder et al. InterPro, progress and status in 2005. Nucleic Acids Res. 2005;33: D201-D205.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol. 1990; 215:403-410.
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389-3402.
8. Hulo N, Sigrist CJA, Le Saux V., Langendijk-Genevaux PS, Bordoli L, Gattiker A, Des Castro E, Bucher P, and Bairoch A. 2004. Recent improvements to the PROSITE database. Nucleic Acids Res. 2004;32: D134-D137. (http://www.ebi.ac.uk/interpro/).

9. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Maxon S, Sonnhammer ELL, Studholme DJ, Yeats C, and Eddy, SR The Pfam protein families database. *Nucleic Acids Res*. 2004;32:D138-141.http://pfam.wustl.edu/.

10. Galat A. Variations of sequences and amino acid compositions of proteins that sustain thier biological functions: an analyis of the cyclophilin family of proteins. Arch Biochem Biophys. 1999;371:149-162.

11. Bairoch A and Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. J. Mol. Med. 1997;75: 312-316.

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

13. Hughey R and Krogh A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. CABIOS 1996;12: 95-107.

14. Opiyo SO and Moriyama EN. Protein family classification with partial least squares. J Proteome Res. 2007; 6:846-853.

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

16. Romano PG, Horton P, and Gray JE. The Arabidopsis cyclophilin gene family. Plant Physiol. 2004b;134:1268-1282.

17. Lamb J R, Tugendreich S, and Hieter P. 1995. Tetratrico peptide repeat interactions: to TPR or not to TPR? Trends Biochem. Sci. 1995;20:257–259.

3.7 TABLES

Table 1. The number of cyclophilin protein sequences available from plants.

Organisms	Single-domain	Multiple-domain
Arabidopsis	21	8
Rice	12	7
Wheat	4	0
Maize	2	0
Tomato	2	0
Soybean	1	0

Table 2. The number of samples included in cyclophilin datasets

datasets	cyclophilin	non-cyclophilin	total
	Within-group classification		
Single-domain training	140	140	280
Single-domain test	140	1000	1140
TPR multiple-domain training[a]	36	36	72
	Among-group classification		
Single-domain training	140	140	280
TPR multiple-domain test	36	200	236
TPR multiple-domain training	36	36	72
Single-domain test	140	1000	1140

[a] Cross-validation test was performed

Table 3. Classifiers performance on an independent test dataset. Classifiers were trained with single domain cyclophilins.

Methods	% Accuracy	% False positive	% False negative
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
PSI-BLAST	95.8	0.3	22

Table 4. A confusion matrix showing true positive, false positive, true negative, and false negative.

		Predicted	
		Cyclophilin	**Non-cyclophilin**
Actual	**Cyclophilin**	True positive	False positive
	Non-cyclophilin	False positive	True negative

True positive: The number of correct predictions that an instance is cyclophilin.

False positive: The number of an incorrect prediction that an instance is cyclophilin.

True Negative: The number of correct predictions that an instance is non-cyclophilin.

False Negative: The number of an incorrect prediction that an instance is non-cyclophilin.

Table 5. Classifiers performance on TPR multiple domain cyclophilins after cross-validation test.

Methods	% Accuracy	% False positive	% False negative
PLS-ACC	91.6	13.8	3.7
PLS-T_ACC	94.4	8.0	2.7
PLS-R_ACC	94.4	8.0	2.7
SAM	91.6	0.0	16.5
PSI-BLAST	83.3	5.0	25.0

Table 6. Classifiers performance on a TPR multiple domain cyclophilins test dataset. Classifiers were trained with single domain cyclophilins.

Methods	% Accuracy	% False positive	% False negative
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 7. Classifiers performance on single domain cyclophilins. Classifiers were train on TPR multiple domain cyclophilins.

Methods	% Accuracy	% False positive	% False negative
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

Table 8. The number of cyclophilins identified from *Arabidopsis thaliana* and rice genomes.

Methods	Numbers of predicted proteins
	Arabidopsis thaliana genome
Single-domain cyclophilins	
PLS-ACC	1210 (980)[a]
PLS_T-ACC	304 (290)
SAM	34 (31)
PSI-BLAST	34 (31)
TPR multiple-domain	
PLS-ACC	325 (218)
PLS_T-ACC	117 (110)
SAM	105 (91)
PSI-BLAST	492 (484)
	Rice genome
Single-domain cyclophilins	
PLS_T-ACC	1336 (1259)
SAM	69 (48)
PSI-BLAST	50 (29)
TPR multiple-domain	
PLS_T-ACC	314 (304)
SAM	161 (122)
PSI-BLAST	132 (96)

[a] The number in parenthesis after removing splice variants

3.8 FIGURES

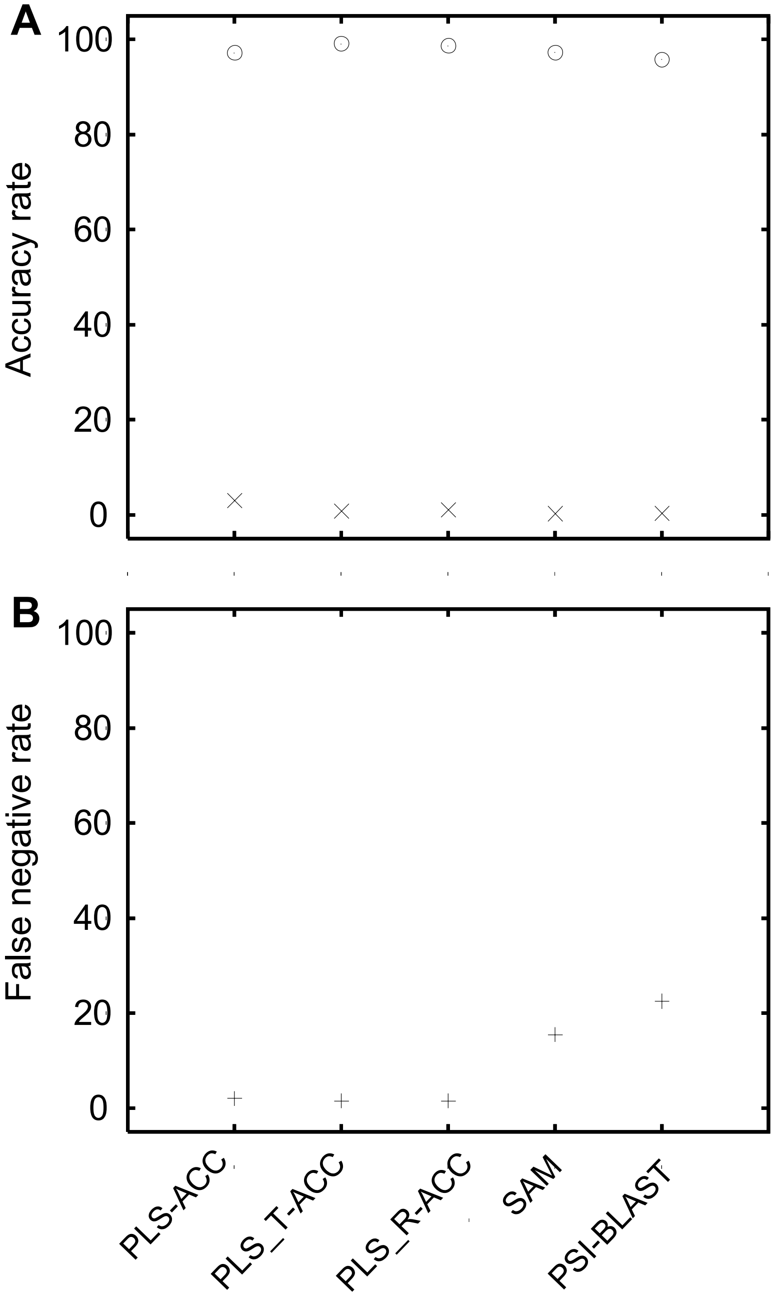


Figure 1. Classifier performance on single-domain cyclophilins test dataset. A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

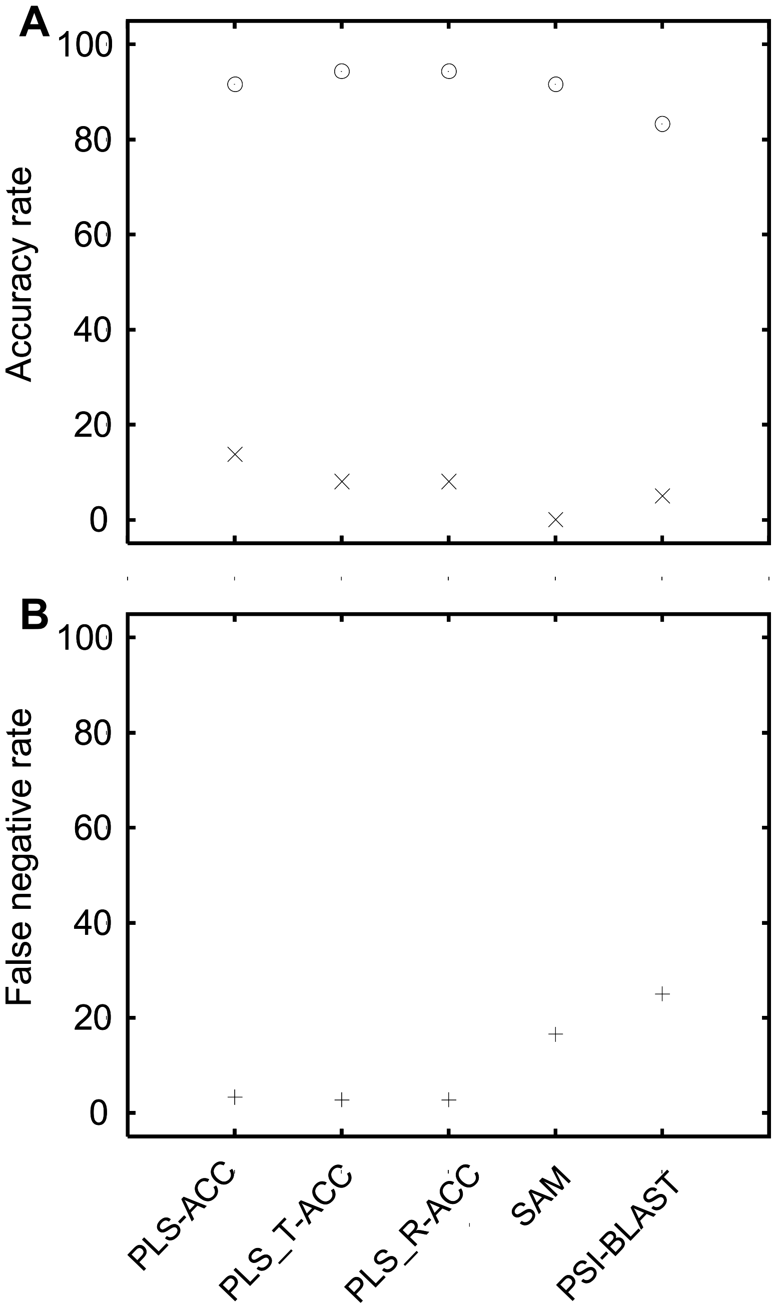


Figure 2. Classifier performance on TPR single-domain cross-validation test. A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

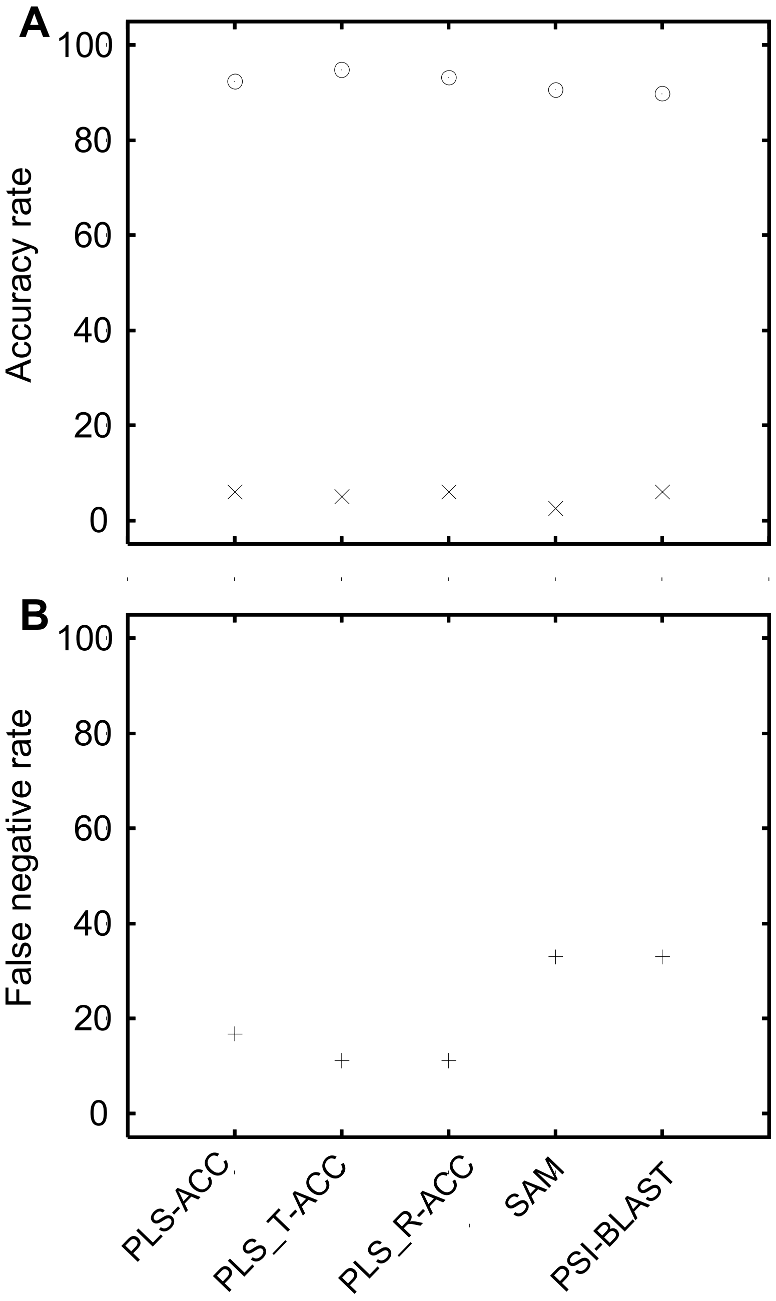


Figure 3. Classifier performance on TPR single-domain test dataset (between-group classification). A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

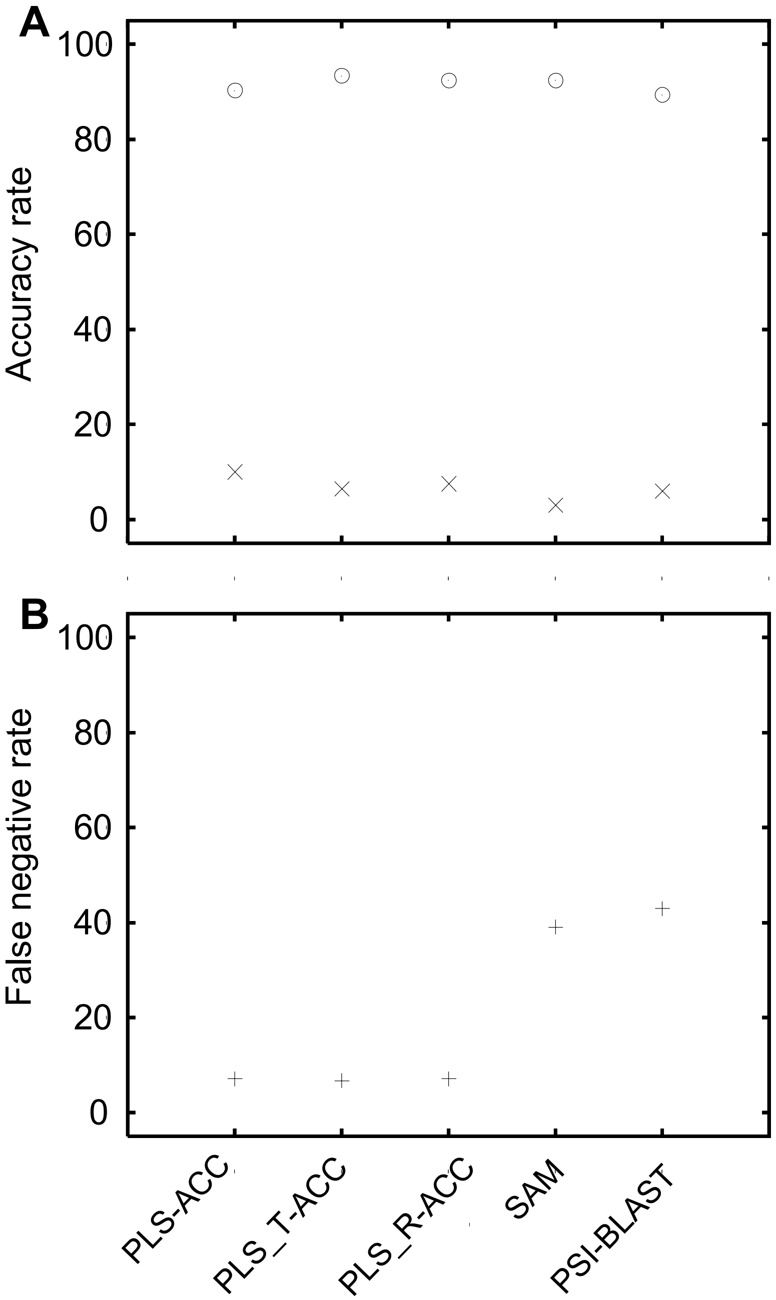


Figure 4. Classifier performance on single-domain test dataset (between-group classification). A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

CHAPTER 4

A SIMPLE STATISTICS METHOD FOR PROTEIN FAMILY CLASSIFICATION

4.1 ABSTRACT

With the rapid accumulation of proteomic data, computational prediction of functions of new protein candidates is becoming more important than ever. The most commonly used computational methods used for protein functional predictions rely on alignments. However, these methods do not perform well on protein sequences with weak sequence similarities. Alignment-free methods are sensitive in predicting proteins with weak similarities, but they have high false positive rates. The objectives of this study were to develop alignment-free method that was sensitive to protein with weak sequence similarities and have low false positive rates; and to use the method developed to mine G-protein coupled receptors (GPCRs) from the *Arabidopsis*, rice, and maize genomes. We used self-organizing map (SOM) and t-test statistics to select amino acid compositions that could discriminate GPCRs from non-GPCRs. Ten amino acids (Lys, Ser, Leu, Glu, Asp, Gly, Val, Gln, Cys, and Phe) were identified by SOM and t-test. The composition of these ten amino acids combined with isoelectric point, and transmmebrane topology were used to develop a new simple statistics method (ST-method) for GPCR

classification. We also included immunoglobulin superfamily from SCOP family database for a benchmark test. T-test analysis identified seven amino acids (Ala, His, Asn, Arg, Pro, Ser, and Thr) as ones that could discriminate immunoglobulins from non-immunoglobulins. These seven amino acids combined with mass, surface area, and frequencies of alpha-helices and beta-strands were used to develop ST-method for the immunoglobulin superfamily. ST-method outperformed partial least squares, profile hidden Markov models (HMMs) and PSI-BLAST on the test dataset of GPCRs, and a cross-validation test of immunoglobulin superfamily. ST-method identified all the known 22 GPCRs from the *Arabidopsis* genome, but SAM missed 7, PSI-BLAST missed 21, and PLS-ACC missed 2. Therefore, ST-method can be used to mine protein families with weak sequence similarities effectively from genomic data.

4.2 INTRODUCTION

Advancement in sequencing techniques have drastically increased the rate of uncovering new protein sequences. As mentioned in Chapter 1, computational methods developed for protein family classification can be divided into alignment-based and alignment-free methods. The most commonly used pairwise sequence similarity search method is the Basic Local Alignment Search Tool (BLAST)[1]. For more distantly related proteins where only certain sequence features or structural motifs are conserved, they cannot easily be recognized by pairwise alignment methods. In such cases, multiple alignments of related sequences properly translated into position-specific score matrices (used in PSI-BLAST[1]) or profile hidden Markov models (HMMs)[2] are used. As discussed in previous chapters, these alignment-based protein sequence search methods do not perform well against highly diverged proteins, because generating reliable models is difficult or impossible for extremely diverged protein families such as G-protein coupled receptors (GPCRs).

Problems of alignment-based methods can be overcome by using alignment-free methods. Advantages of alignment-free methods were discussed previously. In Chapters 2 and 3, alignments-free methods were found to be sensitive in identifying sequences with weak similarities such as GPCRs or sequences with heterogeneous domains such as cyclophilins. However, these methods were found to have high false positive rates. For example, in the remote similarity classification

(Chapter 2), the false positive rates of PLS methods in identifying GPCRs belonging to the classes not included in the training datasets ranged from 6 to 20%. In mining GPCRs from the *Arabidopsis* genome, the numbers of sequences predicted by alignment-free methods as GPCRs were 2,000 to 3,400, which is about 10% of the entire predicted *Arabidopsis* proteins[3]. Clearly, it includes many possible false positives.

The specific objectives of this study, therefore, are to develop an alignment-free method that is sensitive to sequences with weak similarities and has low false positives, and to use the method developed to mine 7TMRs from the *Arabidopsis*, rice, and maize genomes.

4.3 MATERIALS AND METHODS

4.3.1 Data sources

(a) GPCR data

I used 500 GPCR sequences from Moriyama et al.[3] study. These sequences were discussed in Chapter 2 section 2.4. In addition to these 500 GPCR sequences, more GPCRs sequences were randomly sampled from Swiss-Prot[4] database for a test dataset. GPCRs sequences that overlapped with the 500 GPCR sequences were removed and this new 500 GPCRs were selected for a test dataset.

(b) Non-GPCR data

500 non-GPCR sequences (negative samples) longer than 100 amino acids from Moriyama et al.[3] study were used for training. Other 2000 non-GPCR sequences were randomly sampled from Swiss-Prot[4] protein database. 1000 non-GPCR sequences were included in a test dataset, and 1000 non-GPCR sequences were used for calculating parameters for extreme value distribution[5;6]. As before, these non-GPCR sequences in training and test datasets are mutually exclusive.

(c) SCOP immunoglobulin superfamily data

Structural Classification of Proteins (SCOP)[7] is a database of proteins of known structures. It provides a hierarchical classification of a manually curated set of proteins derived from the Protein Data Bank (PDB)[8]. The PDB is the database that contains information on experimentally determined three-dimensional structures of proteins and other biological macromolecules. It contains the atomic information, secondary structure information, general information required for all deposited structures and information specific to the method of structure determination.

At the lowest level of the hierarchy, proteins clustered in a SCOP family have clear evolutionary relationships, indicated by their very similar functions and structures. Proteins in SCOP superfamilies show low degrees of sequence identities of 30% or less, but structural and functional features in the proteins give them a probable common evolutionary origin, meaning that proteins clustered in superfamilies are likely to be homologues. At the folds level, proteins have the same common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. These proteins may or may not be related evolutionary. In this study, immunoglobulin superfamily sequences were downloaded from SCOP version 1.71, release October 2006. SCOP data has been used for the benchmark test for comparing the performances of various protein family classification methods. In this study, we used immunoglobulin superfamily from SCOP for the benchmark test.

The immunoglobulin superfamily (IgSF) is believed to have arisen from a single primordial Ig-like domain during evolution of multicellular metazoans[9]. They have a similar tertiary structure, the immunoglobulin fold, which is a beta-sandwich that can have a varying number of strands. Sequence similarities among IgSF members are very low (26-28%), thus making them difficult to align. IgSF largely carries out functions in the immune system, in cell-cell recognition or in structural organization of muscle.

Members of IgSF include cell surface antigent receptors, co-receptors molecules of immune system, cell adhesion molecules and cytokine receptors.

GPCRs have low similarities among different groups, but they are transmembrane proteins. Since SIgSF proteins also have low sequence similarities among members and they are soluble proteins, we chose this superfamily as an alternative test case to the GPCR dataset. We expect a good classifier to be sensitive to both soluble and transmembrane proteins.

(d) *Arabidopsis thaliana* genome data

28,952 protein sequences were downloaded from The Institute for Genomic Research (TIGR) *Arabidopsis thaliana* Database ftp site (ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/SEQUENCES/ATH1.pep.gz; Release 5, dated on June 10, 2004).

(e) Rice genome data

62, 877 rice (*Oryza sativa*) protein sequences were downloaded from The Institute for Genomic Research (TIGR) Rice Database site http://www.tigr.org/tdb/e2k1/osa1/data_download.shtml using Protein sequences (*.pep) and “Select All chromosome” options. Release 5, updated on December 14, 2006.)

(f) Maize genome data

275, 904 genomic DNA sequences of maize (*Zea mays*) were downloaded from The Institute for Genomic Research (TIGR) Maize Database ftp site (ftp://ftp.tigr.org/pub/data/MAIZE/AZMs/; release_5.0, dated on June 13, 2005).

4.3.2 Dataset preparation

(a) G-protein coupled receptors

A dataset of 500 GPCRs and 500 non-GPCRs from Moriyama et al.[3] was used for training the methods. The same dataset was used for developing scores. The test dataset included 500 GPCRs and 1000 non-GPCRs.

(b) Immunoglobulins

A training dataset including 90 sequences of the immunoglobulin superfamily and 90 non-immunoglobulin sequences were created for training the methods. There were not enough sequences to create an independent test dataset; therefore, the cross-validation (explained in Chapter 3) was done for the immunoglobulin dataset.

4.3.3 Methods used

(a) Self-organizing maps

The self-organizing map (SOM)[10] is an unsupervised method that maps a set of high dimensional data onto a low dimensional array of cells (called neurons) in such a way that data projected onto adjacent cells are more similar than data projected on distant cells.

The SOM consists of two layers: the input layer and competitive layer (output layer) which is mainly a two dimensional grid. Both of these layers are fully interconnected. For example, let *m* be the number of neurons in the input layer (e.g., *m* = 7 in Figure 1) and *n* the number of neurons in the output layer which are arranged in hexagonal patterns. Each neuron in the input layer is connected to each neuron in the output layer. Thus, each neuron in the output layer has *m* connections to the input layer. Each one of these connections has a synaptic weight associated with it. Let W_j be the weight vector associated with the connection between *m* input neurons i = 1,…,*m* and one of the output neurons j(j = 1,..., *n*). The neurons of the map are connected to adjacent neurons by a neighborhood relation.

Components of SOM are described next.

Sample data

The first part of SOM is the data. For example, the data can be compositions or physico-chemical properties of amino acids. The data are presented as sample vectors.

Weights

The second component of SOM is the weight vectors, W_j. Each weight vector has two components to them. The first part of the weight is for its data. This has the same dimensions as the sample vectors. The second part of the weight is in reduced dimensions, (e.g., 1 or 2 dimensions).

SOM main algorithm

The way SOMs go about by representing themselves is by competing for representation of the samples. Neurons are allowed to change themselves by learning to become more like samples in hope of winning the next competition. It is this selection and learning process that makes the weights organize themselves into a map representing similarities. With these two components (the input and weight vectors), the weight vectors are ordered in such a way that they will represent similarities.

The first step of SOM calculation is the initialization of the weights vectors. This is done by giving each weight vector random values between 0 and 1 for its data. The second step is to get the best matching unit (BMU). Getting the best match is done by calculating the distance from each weight to the chosen sample vector. The weight with the shortest distance is the winner. If there are more than one best matches with the same distance, then the winning weight is chosen randomly among the weights with the shortest distance. The most common method of determining the distance is Euclidian distance. The last step of SOM algorithm is scaling of neighbors. Neighboring weights are scaled by concentric squares or hexagons (an example of hexagon is presented in Figure 1). The weights of all the neighbors that are enclosed by the hexagon will be adjusted. The second part of the scaling is learning function. The winning weight is rewarded by becoming more like the sample vector. The neighbors also become more like the sample vector. An attribute is that, the further away the neighbor is from the winning vector, the less it learns.

In a supervised self-organizing map introduced by Kohenen[10], the input data contains both the descriptors and the class information. Thus, class information influences topological ordering of the map during training. One key application area of self-organizing maps is data visualization. Using SOM for data visualization is useful in selecting descriptors that are important for differentiating proteins of interest from other proteins. During the training, the input vector contains descriptors from protein sequences as well as class information represented by a binary code. In this study SOM was used for selecting descriptors using RapAnalyst software[11]. This software was provided by the Breaking Through Management Group for trial.

(b) Student t-test and Wilcoxon sum-rank test

These tests were discussed in Chapter 3. In this study, they were also used for selecting descriptors using R package stats[12].

(c) HMMTOP

HMMTOP[13] method was discussed in Chapter 3. In this study it was used for predicting transmembrane topology of protein sequences.

(d) Isoelectric point determination

Isoelectric point of each protein sequence was determined using Compute pI/Mw Bioperl program from Bioperl organization. (http://www.bioperl.org).

(e) Profile hidden Markov models (profile HMMs)

Profile (HMMs)[2] method was discussed previously. The cut-off E-value of 1.2 and 0.9 were obtained for GPCRs and immunoglobulin from the training datasets using minimum error point (MEP; explained in Section 3.4.5). Profile HMM was included for comparative analysis.

(f) PSI-BLAST

PSI-BLAST[1] was also discussed previously. As with profile HMM, PSI-BLAST was included for comparative analysis. The cut-off E-value of 1.8 and 1.2 were obtained for GPCRs and immunoglobulin from the training datasets using MEP.

(g) TBLASTN

TBLASTN[1] compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands). In this study, sequences predicted by a ST-method (explained in Section 4.3.4) with seven transmembrane with N-external from rice and *Arabidopsis* were used as queries by TBLSTN to search maize genome for GPCRs.

(h) Partial least squares

In this study, partial least squares (PLS) with descriptors transformed by auto and cross covariance (PLS-ACC), and PLS with descriptors from amino acid compositions (PLS-AA) described in Chapter 2 and Opiyo and Moriyama[14] were included for comparative analysis. The cut-off points of 0.494 and 0.422 for PLS-ACC, and 0.498 and 0.457 for PLS-AA were obtained for GPCRs,

and immunoglobulin from the training datasets using MEP.

4.3.4 Development of Simple Statistics Method

As mentioned in Section 4.2, alignment-free methods are sensitive to identifying sequences with low similarities but they have high false positives. My motivation was to develop alignment-free method that is sensitive to protein sequences with weak similarities with low false positives. In Chapter 3, I found that reducing the number of descriptors by selecting only those that are important for discriminating cyclophilins from non-cyclophilins can reduce the number of false positives. A simple statistics method (ST-method) using SOM and t-test was developed for GPCRs and Immunoglobulin superfamily as follows.

(a) Simple statistics method for identifying G-protein coupled receptors

Amino acid composition transformation

From each amino acid of the 500 GPCR and 500 non-GPCR sequences from Moriyama et al[3], frequencies of 20 amino acids were calculated as explained in Chapter 1 Section 1.3.3(a).

Self organizing map analysis

SOM analysis was applied to the amino acid compositions of 500 GPCR and 500 non-GPCR sequences. The input data were represented with 1000 rows by 21 columns. Rows were made of 500 GPCRs and 500 non-GPCRs, and columns were made of 20 amino acids plus a label column of ones and zeros representing GPCRs and non-GPCRs, respectively.

Figure 2 shows the examples of the SOM. Figure 2A represents a class map of GPCRs and non-GPCRs (red for GPCRs, and blue for non-GPCRs). Figure 2B represents a map of composition of amino acid cysteine (from 0.00 to 0.12; see the scale Figure 2B). Each cell in the figure represents a cluster of protein sequences with the same composition of the amino acid cysteine. The map of composition of amino acid cysteine is similar to that of GPCR and non-GPCR classes. From visual inspection, we can conclude that on average the compositions of amino acid cysteine found in GPCRs are higher than those found in non-GPCRs. On the other hand, Figure 2C represents how compositions of the amino acid histidine were clustered. No distinct pattern is observed in the clustering of composition of amino acid histidine with that of GPCRs and non-GPCRs. Thus we conclude that there is no difference between the composition of the amino acid histidine in GPCRs and non-GPCRs. After visually comparing the color patterns between GPCR and non-GPCR classes map and each amino acid composition map, seven amino acids amino acids; (Glycine, Glutamine, Cysteine, Leucine, Phenylalanine, Serine, and Valine) were selected as those discriminating GPCRs from non-GPCRs by visual inspection (Figure 3). For all the seven amino acids selected, the absolute difference between the mean frequencies of each amino acid in 500 GPCRs and 500 non-GPCRs was equal to or greater than 0.01.

T-test analysis

Selection of the seven amino acids by SOM was done by visual inspection. By using only visual inspection, we may miss some amino acids that could discriminate GPCRs from non-GPCRs even if they may not show such clear difference visually. In order to make this selection step more quantitative and with statistical confidence, the t-test analysis with an alpha level of 0.01 was performed on the same data used for SOM analysis. This test was described in Chapter 3 and used for selecting descriptors for discriminating cyclophilins from non-cyclophilins. In order for an amino acid to be selected, it should have a significant difference of alpha ($p \leq 0.01$) and the absolute difference between the mean frequencies of in 500 GPCRs and 500 non-GPCRs of equal to or greater than 0.01. Ten amino acids, lysine, serine, leucine, glycine, glutamic acid, aspartic acid, valine, glutamine, cysteine, and phenylalanine were selected. The results of the t-test are presented in Appendix Table 14, and the differences in the distributions between GPCRs and non-GPCRs are presented in Appendix Figure 2.

Wilcoxon rank-sum test analysis

Wilcoxon rank-sum test was also performed on the same data used for SOM analysis. The results obtained from the rank-test were similar to that of the t-test (Appendix Table 14).

(b) Development of a score system

Given a protein sequence, we would like to know which group (GPCRs or non-GPCRs) a protein belongs to. Scores are needed to determine the group. To develop scores for the ST-method the same dataset from Moriyama et al[3] used for the training (500 GPCRs and 500 non-GPCRs) was used.

Amino acid composition score determination

For each of the ten selected amino acids, we calculated the frequency distributions from the 500 GPCR and 500 non-GPCR sequences (Appendix Figure 2 and Table 15). Based on these frequencies, the log-odds score for a protein that has an x was calculated using equation 1.

$$\text{Log-odds score}(x) = \log\left(F(x)_{GPCR}/F(x)_{non\text{-}GPCR}\right) \qquad (1)$$

where $F(x)_{GPCR}$ and $F(x)_{non\text{-}GPCR}$ are frequencies of GPCRs and non-GPCRs that have the frequency of x, respectively. The odds are the ratio of the likelihood of two events or outcomes. For our example, the odds are the ratio of the frequency of finding particular amino acid compositions in GPCRs to that in non-GPCRs. Odds range from 0 to infinity. Odds are converted to the logarithm the log-odds. Table 1 shows the log-odds score. Log-odd scores range from negative infinity to positive infinity. The amino acid composition score of a protein sequence can be calculated by summing up the log-odds for each of the 10 amino acids as shown in equation 2.

$$AA_{(i)} = K_{(i)} + S_{(i)} + L_{(i)} + G_{(i)} + E_{(i)} + D_{(i)} + V_{(i)} + Q_{(i)} + C_{(i)} + F_{(i)} \qquad (2)$$

where $AA_{(i)}$ are the amino acid composition score.

Isoelectric point scores determination

Similar to amino acid compositions, frequency distributions of pI values were obtained from the 500 GPCR and 500 non-GPCR sequences. Log-odds score were calculated for a given range of pI as summarized in Table 2.

Transmembrane topology score determination

Transmembrane (TM) regions were predicted from each protein using HMMTOP[13]. Frequency distributions for the number of predicted TM helices were calculated from the 500 GPCR and 500 non-GPCR sequences, and the log-odds score as shown in Table 3.

Calculation of the final scores

After calculating the amino acid composition score based on the 10 chosen amino acids, pI, and the score based on the number of predicted TM regions, the final score for a protein i, $S_{GPCR(i)}$, is computed as shown in equation 3.

$$S_{GPCR(i)} = AA_{(i)} + pI_{(i)} + TM_{(i)} \tag{3}$$

where $AA_{(i)}$ are the amino score calculated by the equation 2, and the $pI_{(i)}$ and $Tm_{(i)}$ are the pI and TM scores, respectively.

Determining a cut-off score to identify GPCRs

A cut-off score is the minimum acceptable score to identify a protein as GPCRs from non-GPCRs. It was determined by the minimum error point (explained later in section 4.3.5) using the same training dataset. The score 0.9 was selected as the cut-off value for identifying GPCR proteins. Figure 4 summarizes the entire process of the ST-method for identifying GPCR proteins.

(c) P-value calculation

In a database search, we generally wish to determine the statistical significance of a score. It is important to know the probability that we would see a given score by chance even from a database of unrelated sequences. Karlin and Altschul[5] showed that the scores of optimal alignments of random sequence pairs approach an extreme value distribution (EVD). The EVD is defined by two parameters λ and μ, where μ is the location parameter (the mode) and λ is the scale parameter. The EVD of a random alignment score *s* has the probability density of

$$\rho\,(\mathrm{s}) = \frac{1}{\lambda}(\exp(-\frac{s-\mu}{\lambda}))\exp\,(-\exp\,(-\frac{s-\mu}{\lambda})) \qquad (4)$$

The parameters can be determined either by the mean method[5] or by the maximum likelihood method[6]. After determining the parameters μ and λ, equation 5 can be used to estimate the probabilities of scores. P(S ≥ x) given in equation 5 is the probability of seeing the score S greater than or equal to a given score x from one random sample. In a database search samples are tested, and the P-value needs to be adjusted to accommodate multiple testing problem. In multiple testing, multiple hypotheses are tested simultaneously, and this increases the chance of false positives. The equation 6 incorporates this adjustment, and P(x, n) is the probability of observing the score x or greater at least one from a search against a database containing n sequences. An E-value is the expected number of sequences with scores greater than or equal to x in a database by chance, and it can be given by equation 7 below.

$$P(S \geq x) = 1\text{-}\exp(-\exp^{-\lambda(x-\mu)}) \tag{5}$$

$$P(x, n) = 1 - \exp^{-nP(S \geq x)} \tag{6}$$

$$E(x, n) = nP(S \geq x) \tag{7}$$

In order to use these equations to calculate the p-value for the score, first we need to confirm the ST-method scores follow the EVD. Figure 5 shows the distribution of the scores obtained from the 1000 non-GPCR sequences. From this distribution, we used the mean method to estimate μ and λ using 1000 non-GPCRs. For the EVD, the expectation value or mean $\overline{X}$ and the variance σ^2 are given as follows:

$$\overline{X} = \mu + \gamma\,\lambda \tag{8}$$

$$\sigma^2 = \lambda^2\pi^2/6 \tag{9}$$

where γ is the Euler-Mascheroni constant, 0.57722. Using equations 8 and 9, and estimating $\overline{X}$ and σ from the 1000 non-GPCRs, we can estimate μ and λ. The mean and standard deviation obtained from 1000 Non-GPCR scores were -10.8966 and 4.629, respectively. Note that we used the distribution obtained from only non-GPCRs because we would like to know the probability of finding a given score by chance even if the sequences are not GPCRs.

From equations 10 and 11 μ and λ are derived as follows:

$$\lambda = \pi / (\sigma \sqrt{6}) = 1.2825 / \sigma \qquad (10)$$

$$\mu = \overline{X} - \gamma / \lambda = \overline{X} - (0.4500 * \sigma) \qquad (11)$$

Using these equations, we estimated $\lambda = 0.2732$ and $\mu = -12.9906$. The expected EVD using these parameters are shown in Figure 5. A statistical analysis between the observed and expected distributions showed no significant difference ($p = 0.348$ by Chi square test). Therefore, the p-value for a score x can be calculated by equation 6 shown before.

The ST-method was also applied for the immunoglobulin superfamily classification.

Amino acid composition transformation and t-test analysis

Amino acid composition was calculated from each protein sequence and the t-test with the alpha level of 0.01 was performed to identify the amino acids that differentiate the immunoglobulin superfamily proteins from other proteins. Seven amino acids, alanine, histidine, asparagine, proline, arganine, serine, and threonine were selected. All statistics are presented in Appendix Table 16. The rank-test was also performed on amino acid composition (Appendix Table 16). The results obtained from rank-test were not different from the results obtained from t-test. The distributions of amino acids showing the difference between immunoglobulins and non-immunoglobulins are presented in Appendix Table 17 and Figures 3.

Physico-chemical property transformation and t-test analysis

Twelve physico-chemical properties (mass, volume, surface area, hydrophilicity, hydrophobicity, isoelectric point, transfer of energy solvent to water, refractivity, and non-polar surface area, frequencies of alpha-helix, beta-sheet, and reverse turn) used in Chapter 2 and also in Opiyo and Moriyama[12] was assigned to each amino acid in each sequence (see Table 4). The mean transformation method described in Chapter 2 and also in Opiyo and Moriyam,[14] was performed on each sequence. A t-test with the alpha level of 0.01 was performed on the transformed protein sequences. Four properties: mass, surface area, frequencies of alpha-helix and beta-sheet, were selected (Appendix Table 18). Similar procedure used for t-test analysis was performed with rank-test analysis (Appendix Table 18).

(d) Score development

The scores to identify the immunoglobulin superfamily proteins from the selected amino acid composition and physico-chemical properties were developed similarly as described in the Section 4.3.4 (b) for GPCRs. Their log-odds scores are presented in Tables 5, 6, 7 and 8.

The final score for a protein i, $S_{IMN(i)}$, is calculated by equation 12:

$$S_{IMN(i)} = A_{(i)} + H_{(i)} + N_{(i)} + P_{(i)} + R_{(i)} + S_{(i)} + Mass_{(i)} + Sur_{(i)} + Alph_{(i)} + Beta_{(i)} \qquad (9)$$

where $A_{(i)}$, $H_{(i)}$, …, $T_{(i)}$ are the scores of the 7 amino acids obtained for the protein i based on Table 5; $Mass_{(i)}$, $Sur_{(i)}$, Alpha $_{(i)}$, and $Beta_{(i)}$ are the scores based on the mean values of mass, surface area, frequencies of alpha-helix and beta-strands, respectively, for the protein i based on Tables 6, 7, and 8. The cut-off score, 0.1, was determined by the minimum error point using the training dataset (the minimum error point is explained in Section 4.3.5 (b)). Figure 6 summarizes the overall process of identifying the immunoglobulin proteins using ST-method.

4.3.5 Performance analysis

4.3.5 (a) Statistics

Predictions are grouped as follows:

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

The 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)

• Mathews correlation coefficient (MCC)

$= (TP \times TN - FP \times FN) / \{(TP + FN) (TP + FP) (TN + FP) (TN + FN)\}^{1/2}$

(b) Minimum error point (MEP)

The accuracy rates, false positive rates, and false negative rates were calculated at the minimum error point (MEP). MEP was used by Karchin et al.[15]. It is the threshold score where the method produces the minimum number of errors (false positives + false negatives).

4.4 RESULTS

4.4.1 Simple statistic method

In this study, my objective was to develop an alignment-free method that is sensitive to sequences with low similarities and has low false positive rates. A simple statistics method (ST-method) for GPCRs was developed using the compositions of ten chosen amino acids, pI, and transmembrane topology of sequences and tested for identifying GPCR sequencers. The scores developed for the ST-method were confirmed to the extreme value distribution (EVD) based on the frequency distribution obtained from (non-GPCRs) random sequences. The EVD parameters were obtained: $\lambda = 0.2732$ and $\mu = -12.9906$. These parameters can be use to calculate P and E values of a given score from ST-method. For example, the cut-off score for a GPCR protein was found to be 0.9.

We can calculate the probability of seeing the score greater than or equal to 0.9 in a database search of 1000 sequences. We can also find the numbers of Non-GPCRs whose scores are greater 0.9 in a database of 1000 sequences. The calculations can be done as follows:

$P(S \geq 0.9) = 1\text{-}\exp(-\exp^{-0.2733(0.9+12.9906)})$

9.28×10^{-7}

$E(x, 1000) = 9.28 \times 10^{-7} * 1000$

9.28×10^{-4}

From the p-value, we can observe that the probability of seeing the score greater than or equal to 0.9 in a database search is much smaller than p = 0.001. The chosen cut-off value gives highly significant results.

4.4.2 Comparison of ST-method with PLS-ACC, PLS-AA, SAM, and PSI-BLAST for identifying GPCR sequences

In order to compare the performance of ST-method, the same training dataset consisting of 500 GPCRs and 500 Non-GPCRs from Moriyama et al.[3] was used to train ST-method as well as PLS-ACC, PLS-AA, SAM, and PSI-BLAST, to was used to train all these methods. The performance of these methods was examined using a test dataset of 500 GPCRs and 1000 non-GPCRs obtained from Swiss-Prot database. Figure 7 and Appendix Table 19 illustrate their performance. It can be seen that ST-method had the highest accuracy rate (98.4%), lowest false negative rate (0.4%), and the best MCC value of 0.96. PLS-ACC and PLS-AA had slightly lower accuracy rates (95.2 and 94.9%), much higher false positive rates (6.5% and 8.5%), slightly higher false negative rates (0.6% and 0.8%), and lower MCC values (0.90 and 0.89). SAM and PSI-BLAST had the lowest accuracy rates (94.8 and 92.5%), much higher false negative rates (15.0% and 23.0%) and the lowest MCC values (0.88 and 0.83). It should be noted that although SAM and PSI-BLAST had high negative rates, both had lowest false positive rates among the classifiers compared.

4.4.3 Comparison of ST-method with PLS-ACC, PLS-AA, SAM and PSI-BLAST on SCOP Immunoglobulin benchmark dataset

As discussed in the materials and methods section, this family was chosen because it is a soluble protein and has low sequence similarities among groups. (Sequences were filtered to 30% or lower identities). Figure 8 and Appendix Table 20 show the results of cross-validation test for the methods. ST-method again outperformed the other four methods with the highest accuracy rate of 93.3%, and the lowest a false negative rate (4.4%). The accuracy rates of PLS-ACC, PLS-AA, SAM, and PSI-BLAST were 90.0% or lower. PLS-ACC and PLS-AA had very high false positive rates (15.6% and 18.3%), where as SAM and PSI-BLAST had very low false positive rate (5.5%). Note also that SAM and PSI-BLAST could not identify most of the Immunoglobulin superfamily as indicated with extremely high false negative rates (50% and 52.2%) and very low MCC (0.52 and 0.50). The false negative rates of ST-method, PLS-ACC, and PLS-AA were

4.4%, 4.6%, and 4.8%, respectively (Figure 8B; Appendix Table 20). Note again that the ST-method had relatively low false positive rate (8.0%) which contributed to the highest MCC.

4.4.4 Mining the *Arabidopsis* genome for seven transmembrane receptors

We applied ST-method, PSI-BLAST, and SAM to search for 7TMR candidates from the *Arabidopsis thaliana* genome. Table 9 shows our results as well as the results obtained from Moriyama et al.[3]. ST-method predicted 659 proteins to be 7TMR candidates. These proteins included all the 22 known *Arabidopsis* 7TMRs. SAM predicted only 15 proteins to be 7TMR candidates, missing 7 out of the 22 known *Arabidopsis* 7TMRs. PSI-BLAST predicted only one protein RGS1 (At3g26090.1) as a 7TMR candidate missing 21 of the 22 known *Arabidopsis* 7TMRs. Each of the six alignment-free methods LDA, QDA, SVM-AA, SVM-di, PLS-ACC, and KNN20) from Moriyama et al.[3] predicted 2000 - 3400 proteins as 7TMR candidates.

Moriyama et al.[3] combined the results of those six to select 7TMR candidates probably also reducing the number of false positives. Using the strict intersection, 652 proteins were predicted to be 7TMR candidates. Using the number of predicted TM regions to be 5-10, they were able to identify 342 proteins as 7TMR candidates (Table 9). In this study, using the number of TM regions to be 5-10, ST-method above predicted 579 proteins as 7TMR candidates. These 579 7TMR candidates, 250 were overlapped with those predicted by Moriyama et al.[3]. We can consider those 250 as the more refined candidate sets. The list of these proteins is in Appendix Table 21. A canonical GPCR protein has seven transmembrane regions with the external N-terminal region. HMMTOP predicted 125 protein candidates to have such transmembrane topology from a total of the 28,952 proteins from the *Arabidopsis* genome. ST-method predicted 92 out of 125 proteins predicted by HMMTOP as 7TMR candidates. This number includes all the 54 7TMR candidates predicted by Moriyama et al.[3] and are presented in Appendix Table 22.

4.4.5 Mining the rice genome for seven transmembrane receptors

ST-method and PLS-ACC predicted 717 and 702 sequences with 5-10 TM regions 114 and 72 sequences with 7TM regions with the external N-terminal regions as 7TMR candidates, respectively (Tables 10, Appendix Tables 23 and 24). All the 702 sequences predicted by PLS-ACC are included in the 717 sequences predicted by ST-method. SAM and PSI-BLAST predicted 5 and 3 sequences with TM regions of 5-10, and 5 and 2 sequences with 7TM regions and external N-terminal as 7TMR candidates, respectively (Table 10).

4.4.6 Mining the maize genome for seven transmembrane receptors

Maize does not have protein sequence databases as *Arabidopsis* and rice. We therefore, translated maize genomic DNA sequences into six frames. Protein sequences equal to or greater than two hundred amino acids were selected, giving a total of 53,613 sequences. The results obtained from searching the genome by ST-method, PLS-ACC, SAM and PSI-BLAST are presented in Table 11. ST-method and PLS-ACC predicted 464 and 1120 sequences as 7TMR proteins, and PSI-BLAST predicted only 7. SAM did not predict any sequence as a 7TMR protein. The number of sequences with 5-10 TM regions predicted as 7TMR proteins by ST-method, PLS-ACC, and PSI-BLAST were 382, 362, and 1, respectively (Appendix Table 25). ST-method predicted 48 and PLS-ACC predicted 31 sequences with 7 TM regions with external N-terminal as 7TMR candidates, and PSI-BLAST predicted none. 48 sequences predicted by ST-method include all the 31 sequences predicted by PLS-ACC. The list of these sequences is presented in Appendix Table 26.

Sequences with 7 TM regions with external N-terminal obtained by ST-method from *Arabidopsis* (92) and rice (114) were also used as queries to search the maize DNA database by TBLASTN with a cut-off value of E-value = 0.001. TBLASTN identified 194 similar sequences. Sixteen of them have 5-10 TM regions and 5 have 7 TM regions with external N-terminal (Table 9). These 5 sequences were among the 48 7TMR proteins predicted by ST-method and PLS-ACC.

4.5 DISCUSION

ST-method using compositions of ten chosen amino acids combined with pI and transmembrane topology was sensitive enough to discriminate 7TMRs from other proteins with very high accuracy. The scores developed follow the EVD distribution that as it has been reported for sequence alignments scores[5;6]. In similarity search, the focus is to know how high a value will be obtained next time another score of random sequences is obtained. Thus, the distribution of alignment scores between random sequences follows the EVD but not the normal distribution. Use of the EVD enables to evaluate the probability that a score between random or unrelated sequences will reach the score found between two real sequences of interest. A low probability (e.g., 0.01) indicates that the alignment score between the real sequences is significant and the sequence similarity could be derived from related sequences. Non-GPCR scores from ST-method also followed the EVD. This is important because we can estimate the probabilities base on the ST-method scores and we can conclude that a low score indicates a high probability that the sequence is non-GPCR and a high score indicates that a sequence is a GPCR.

ST-method performed better than PSI-BLAST and SAM because it does not rely on alignments and it includes both positive and negative samples in building the model. 7TMRs have weak sequence similarities, and it makes not easy to align them. Using sequence descriptors and TM topology, we avoided the alignment process, but we were still able to identify 7TMRs that were not identified by SAM and PSI-BLAST.   SAM and PSI-BLAST missed 7 and 22 known *Arabidopsis* 7TMR proteins, respectively. ST-method predicted 717 sequences as 7TMR candidates with 5-10 TM regions, and 114 sequences with 7TM regions external N-terminal from rice genome. SAM and PSI-BLAST predicted fewer than 10 sequences as 7TMR candidates with 5-10 TM regions as well as 7TM with external N-terminal. Diverged 7TMR sequences must have caused unreliable alignment to decrease sensitivities of these alignment-based methods. It is also important to note that ST-method had lower false positive rate compared to PLS-ACC and PLS-AA. PLS-ACC used descriptors developed from 12 amino acids and these descriptors were transformed by Auto and cross covariance (ACC) discussed in Chapter 2 and Opiyo and Moriyama[12]. After ACC transformation, each sequence had 775 variables, and PLS-AA used 19 out of the 20 amino acids compositions as descriptors. It is most likely that not all 775 variables are needed for PLS-ACC to discriminate GPCRs from Non-GPCRs. Some of these variables that do not discriminate GPCRs from non-GPCRs may be contributing to false positives by introducing noise in the data. Also, for PLS-AA, it is most likely that not all the 19 amino acids are needed for discriminating GPCRs from Non-GPCRs.

ST-method also outperformed PLS-ACC, PLS-AA, SAM, and PSI-BLAST on the SCOP benchmark test on the immunoglobulin superfamily. The sequence similarity among IgSF members are very low (26-28%), thus making them difficult to align by the alignment-based methods. That should have been one of the reasons SAM and PSI-BLAST missed 50% and 52% of the Immunoglobulin sequences from the cross-validation test. Immunoglobulin superfamily includes both multigene and single-gene representatives. This superfamily represents an amazingly diverse array of functions from immune receptors to cartilage formation, reflecting the versatility of the shared common structure. Some examples of this superfamily include antigen receptors, co-receptors, adhesion molecules, receptors and natural killers' cells, Ig binding receptors. Consistent with the results obtained from GPCR analysis, PLS-ACC and PLS-AA had higher false positive rates than ST-method. Again, all the 775 variables were included for PLS-ACC and 19 amino acid compositions for PLS-AA.

In Moriyama et al.[3] study, they used an intersection method to reduce the number of possible false positives for alignment-free methods. The intersection method predicted 54 proteins with seven transmembrane regions with external N-terminal as 7TMR candidates.

ST-method predicted 92 proteins with seven transmembrane regions with external N-terminal including all 54 proteins identified by the intersection by Moriyama et al.[3]. The numbers of protein sequences predicted by each of the six alignment-free methods separately with seven TM helices with external N-terminal ranges from 67 for PLS-ACC to 87 for QDA, with SVM-di, SVM-AA, and KNN K=20 having 95, 101, 97 respectively (Table 9). It may be possible that some of the 92 sequences predicted by ST-method are false positives. There is no way to verify further unless experiments are done to confirm whether these sequences are 7TMRs.

The first objective of this study was to develop alignment-free method that was sensitive to sequences with low similarities and has low false positives. The number of false positives produced by PLS-ACC and PLS-AA were higher than that produced by ST-method on a test dataset of GPCRs and in a cross-validation test from immunoglobulin benchmark test. However, all methods had low false negative rate and high MCC in both tests. ST-method identified 659 sequences as 7TMR proteins from *Arabidopsis* genome, whereas PLS-ACC identified 2087 (Table 9). W cannot directly compare the performance of ST-method with other five alignment-free methods (LDA, QDA, SVM-AA, SVM-di, and KNN =20) from Moriyama et al.[3] study for the independent GPCR dataset, and immunoglobulin cross-validation test. It should be noted that ST-method identified all the known 22 *Arabidopsis* sequences that were missed by some of these alignment-free methods used in Moriyama et al.[3] even though exactly the same training dataset was used for all methods. ST-method also predicted fewer numbers of sequences as 7TMRs compared to alignment-free methods used in their study.

The second objective of the study was to mine 7TMRs from the *Arabidopsis*, rice and maize genomes. The results of mining the *Arabidopsis* genome was discussed in the earlier paragraph. From the maize and rice genomes, again ST-method predicted smaller number of sequences as 7TMRs from rice and maize genomes compared to PLS-ACC. SAM and PSI-BLAST predicted far fewer, probably too few sequences compared to ST-method and PLS-ACC from the rice genome. SAM did not identify any sequence from maize genome. TBLASTN identified 194 sequences similar to 7TMRs from the maize genome. It should be noted that maize sequences came from genomic DNA sequences, therefore, they have introns. These sequences need to be confirmed first if they are coding sequences.

The numbers of false positives by ST-method were shown to be much lower than those by PLS-ACC and PLS-AA. When using amino acid compositions as descriptors, most alignment-free methods include compositions of all amino acids in their models. In ST-method, we used only ten important amino acids that could differentiate 7TMRs from non-7TMRs significantly. By reducing the number of amino acids, ST-method was able to reduce the number of false positives that were observed in alignment-free methods (e.g., PLS-ACC and PLS-AA) and at the same time was sensitive enough to identify 7TMRs and Immunoglobulin superfamily. ST-method also identified all the known 22 *Arabidopsis* known GPCRs from the *Arabidopsis* genome, whereas PLS-ACC missed two. These results show that not all the compositions of the 20 amino acids are needed for GPCR classification. Furthermore it implies that there is a greater possibility that we will be able to discover unknown or novel 7TMRs using few important descriptors, even if the alignment-based methods (e.g., SAM, PSI-BLAST) do not work. Therefore, we could conclude that by using only a few important amino acids together with other attributes (e.g., pI and TM topology), the numbers of false positives can be reduced. Another possible explanation why ST-method produced fewer false positives than PLS-ACC and PLS-AA, especially for the GPCR test data is because of the inclusion of transmebrane topology as one of the descriptors. Most of non-GPCRs proteins are not transmembrane proteins, and those which are transmembrane proteins, very few has 7 TM regions. Including transmebrane topology as descriptors eliminated most of the non-GPCRs, hence the numbers of false positives were reduced. However, non-GPCR proteins that were predicted to have 7 TM regions were discriminated from GPCRs because their amino acid compositions and pI are different from those of GPCRs.

ST-method outperformed PLS-ACC, PLS-AA, SAM, and PSI-BLAST on a benchmark test of immunoglobulin superfamily. Immunoglobulin superfamily is a soluble protein and also divergent. However, using amino acid compositions, mass, surface area, and frequencies of alpha-helix and beta-sheet, ST-method was able to classify this divergent superfamily. This shows that ST-method is not only restricted to transmmebrane proteins, but it can also be used to classify soluble proteins.

4.6 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.
2. Durbin R, Eddy S, Krogh A, and Mitchison G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acid*s. Cambridge University Press,
3. Moriyama EN, Strope PK, Opiyo SO, Chen Z, and Jones AM. Mining the *Arabidopsis thaliana* genome for highly-divergent seven transmembrane receptors. Genome Biology 2006;7:R96.
4. Bairoch A and Apweiler R. The SWISS-PROT protein sequence database: its relevance to human molecular medical research. J. Mol. Med. 1997;75: 12-316.
5. Altschul S and Erickson BW. Optimal sequence alignment using affine gap costs. J.Mol. Biol. 1986;48:603-616.
6. Eddy SR. Maximum likelihood fitting of extreme value distributions. 1997: Unpublished, see ftp://ftp.genetics.wustl/edu/pub/eddy/papers/evd.pdf.
7. Murzin AG, Brenner SE, Hubbard T, Chothia C: SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol Biol. 1995;247:536-540.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. Nucleic Acids Res. 2000;28:235-242.

9. Doolittle RF. The multiplicity of domains in proteins. Annu. Rev. Biochem. 1995;64: 287-314.

10. Kohonen T. *Self-Organizing Maps*. 1st edition. 1995; Berlin-Heidelberg: Springer.

11. RapAnalyst. Breaking Through Management Group. Longmont, Colorado.

12. 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.

13.Tusnady GE. and Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics 2001;17:849-850.

14. Opiyo SO and Moriyama EN. Protein family classification with partial least squares. J Proteome Res. 2007;6:846-853.

15. Karchin R, Karplus K, and Haussler D. Classifying G-protein coupled receptors with support vector machines. Bioinformatics 2002;18:147-15

4.7 TABLES

Table 1. Log-odds scores assigned for the ten based on their composition range.

	Scores									
Amino acid composition (x)	Asp (D)	Cys (C)	Glu (E)	Phe (F)	Gly (G)	Leu (L)	Lys (K)	Gln (Q)	Ser (S)	Val (V)
x = 0	-3.301	-3.301	-2.004	-2.904	-3.000	0.000	-2.700	-1.114	0.000	-0.477
0 < x <0.01	-0.103	-4.602	0.476	-3.000	-3.000	0.000	-0.252	0.368	-0.237	-0.477
0.01 ≤ x < 0.02	0.860	0.000	0.813	-3.914	0.000	0.000	0.450	0.377	-0.269	-0.845
0.02 ≤ x < 0.03	0.834	0.288	0.854	-0.991	-0.133	-0.269	0.502	0.228	-1.014	-1.279
0.03 ≤ x < 0.04	0.349	0.929	0.541	-0.840	0.430	-0.237	0.492	0.012	-1.313	-1.724
0.04 ≤ x < 0.05	-0.378	0.896	0.131	-0.110	0.401	-0.921	0.167	-0.211	-0.625	-0.988
0.05 ≤ x < 0.06	-0.854	1.174	-0.010	0.308	0.192	-0.940	-0.082	-1.067	-0.343	-0.359
0.06 ≤ x < 0.07	-0.778	0.204	-0.893	0.460	0.025	-1.240	0.010	-0.940	-0.038	0.220
0.07 ≤ x < 0.08	-0.743	-3.000	-0.808	0.984	-0.389	-0.651	-0.240	-1.462	0.253	0.273
0.08 ≤ x < 0.09	-3.532	-2.303	-3.991	0.964	-0.529	-0.605	-0.984	-1.114	0.386	0.314
0.09 ≤ x < 0.1	-3.079	-2.303	-3.699	0.855	-0.830	-0.117	-3.778	-1.041	0.411	0.222
0.1 ≤ x < 0.11	-3.000	-2.303	-3.681	-3.000	-3.699	-0.140	-3.505	-0.237	0.266	0.235
0.11 ≤ x < 0.12	-2.603	-2.603	-3.623	-2.603	0.000	0.109	-3.826	-0.269	0.169	-0.109
0.120 ≤ x	-3.362	-2.004	-3.301	-3.000	-2.303	0.502	-3.301	-1.041	-0.503	-0.921

Table 2. Frequency distributions of pI values in the GPCRs and non-GPCRs and the log-odds scores

Isoelectric point (x)	Frequency of GPCRs	Frequency of non-GPCRs	Scores
0 ≤ x ≤ 4	0.00	0.20	-2.303
4.01≤ x≤ 4.55	0.00	2.80	-3.447
4.56≤ x≤ 5.00	0.40	8.60	-1.331
5.01≤ x≤ 5.00	0.40	16.20	-1.632
5.51≤ x≤ 6.00	1.00	11.40	-1.057
6.01≤ x≤ 6.50	4.80	10.40	-0.376
6.51≤ x≤ 7.00	4.60	9.20	-0.195
7.01≤ x≤ 7.50	2.00	4.60	-0.389
7.51≤ x≤ 8.00	6.00	4.40	0.135
8.01≤ x≤ 8.50	14.40	4.60	0.507
8.51≤ x≤ 9.00	24.40	8.60	0.487
9.01≤ x≤ 9.55	28.00	8.40	0.491
9.51≤ x≤ 10.00	12.40	6.60	0.287
10.01≤ x≤ 10.50	12.00	2.50	-0.336
10.51≤ x≤ 11.00	0.40	0.80	-0.300
11.01≤ x≤ 11.50	0.00	0.20	-2.303
11.51≤ x≤ 12.00	0.00	0.40	-2.603
>12.00	0.00	0.10	-2.004

Table 3. Frequency distributions of predicted TM numbers in GPCRs and non-GPCRs and the log-odds scores.

Number of Transmembrane regions	Frequency of GPCRs	Frequency of non-GPCRs	Scores
0	0.00	58.00	-4.763
1	0.00	14.00	-4.146
2	0.00	5.00	-3.699
3	0.00	2.40	-3.380
4	0.00	1.20	-3.079
5	0.40	7.20	-1.254
6	7.20	2.80	0.410
7	82.50	0.90	1.961
8	9.40	1.80	0.717
9	0.20	2.50	0.010
10	0.20	2.80	-1.144
> 10	0.10	1.40	-2.580

Table 4. Twelve physico-chemical properties of amino acids[a]

Amino acid	Mass	S_area	Volume	H_phob	H_phil	Refra	Ip	TFen	NP_surf	Alph	Beta	Turn
Ala	71.09	115.00	91.50	0.25	2.10	4.34	6.00	-0.50	182.00	1.29	0.90	0.78
Arg	156.19	225.00	202.00	-1.76	4.20	26.66	10.76	-0.10	9.00	0.96	0.99	0.88
Asp	114.11	150.00	124.50	-0.72	10.00	12.00	2.77	-0.20	28.00	1.04	0.72	1.41
Asn	115.09	160.00	135.20	-0.64	7.00	13.28	5.41	0.11	33.00	0.90	0.76	1.28
Cys	103.15	135.00	118.00	0.04	1.40	35.77	5.05	-0.50	38.00	1.11	0.74	0.80
Glu	129.12	190.00	155.10	-0.62	7.80	17.26	3.22	-0.30	11.00	1.44	0.75	1.00
Gln	128.14	180.00	161.10	-0.69	6.00	17.56	5.65	0.20	26.00	1.27	0.80	0.97
Gly	57.05	75.00	66.40	0.16	5.70	0.00	5.97	0.00	161.00	0.56	0.92	1.64
His	137.14	195.00	167.30	-0.40	2.10	21.81	7.59	-0.40	20.00	1.22	0.08	0.69
Ile	113.16	175.00	168.80	0.73	-8.00	19.06	6.02	-2.00	125.00	0.97	1.45	0.51
Leu	113.16	170.00	167.90	0.53	-9.20	18.78	5.98	-2.00	165.00	1.30	1.02	0.59
Lys	128.17	20.00	171.30	-1.10	5.70	21.29	9.74	-0.30	5.00	1.23	0.77	0.96
Met	131.19	185.00	170.80	0.26	-4.20	21.64	5.74	-1.30	49.00	1.47	0.97	0.39

[a] S_area: Surface area; H_phob: Hydrophobicity; H_phil: Hydrophilicity; Refra: Refractivity; Ip: Isoelectric point; TFen: Transfer free energy from water to ethanol; NP_surf: Non-polar surface; Alph: Frequency of alpha-helix with weight; Beta: Frequency of beta-sheet with weight Turn: Frequency of reverse turn with weight

Table 4 (continued)

Amino acid	Mass	S_area	Volume	H_phob	H_phil	Refra	Ip	TFen	NP_surf	Alph	Beta	Turn
Phe	147.18	210.00	203.40	0.61	-9.20	29.4	5.48	-2.50	89.00	1.07	1.32	0.58
Pro	97.12	145.00	129.30	-0.07	2.10	10.93	6.30	-1.00	34.00	0.52	0.64	1.91
Ser	87.08	115.00	99.10	-0.26	6.50	6.35	5.68	-0.20	108.00	0.82	0.95	1.33
Thr	101.11	140.00	122.10	-0.18	5.20	11.01	5.66	-0.40	38.00	0.82	1.21	1.03
Trp	186.12	255.00	237.60	0.37	-10.00	42.53	5.89	-3.00	79.00	0.99	1.14	0.75
Tyr	163.18	230.00	203.60	0.02	-1.90	31.53	5.66	-2.20	38.00	0.72	1.25	1.05
Val	99.14	155.00	141.70	0.54	-3.70	13.92	5.96	-1.50	206.00	0.91	1.49	0.47

Table 5. Log-odds scores assigned for the seven chosen amino acids for identifying the Immunoglobulin superfamily.

	Scores						
Amino acid composition (x)	Ala (A)	His (H)	Asn (N)	Pro (P)	Arg (R)	Ser (S)	Thr (T)
x = 0	-1.613	0.297	-0.392	-1.613	0.000	0.000	-1.613
0 < x <0.01	0.000	1.004	0.000	1.613	1.785	0.000	0.000
0.01 ≤ x < 0.02	0.612	0.097	-0.594	1.785	0.000	-1.613	-1.613
0.02 ≤ x < 0.03	1.303	-0.079	-0.203	-2.085	-0.470	-0.123	-0.219
0.03 ≤ x < 0.04	0.293	-0.124	-0.041	-0.253	0.175	-0.349	-0.145
0.04 ≤ x < 0.05	0.421	0.078	0.175	-0.378	-0.146	-0.421	-0.096
0.05 ≤ x < 0.06	-0.096	-1.785	0.297	0.00	-0.096	-0.210	0.058
0.06 ≤ x < 0.07	-0.078	0.000	0.096	0.297	0.124	0.145	0.145
0.07 ≤ x < 0.08	0.219	-1.785	0.000	0.219	1.004	-0.173	-0.058
0.08 ≤ x < 0.09	-0.690	0.000	0.613	0.297	0.123	1.004	-0.296
0.09 ≤ x < 0.1	-0.096	0.000	0.613	1.785	-1.613	0.123	1.004
0.1 ≤ x < 0.11	-1.613	0.000	0.000	1.785	1.908	2.083	0.908
0.11 ≤ x < 0.12	-0.392	0.000	0.000	0.000	0.000	0.000	0.000
0.120 ≤ x	0.00	0.00	0.00	0.00	-2.004	0.00	0.00

Table 6. Frequency distributions of the mean amino acid mass in immunoglobulin and non-immunoglobulin proteins, and log-odds scores.

Mean mass (x)	Frequency of immunoglobulin	Frequency of non-Immunoglobulin	Scores
< 100	0.000	0.000	0.000
100 ≤ x < 110	11.000	20.000	-0.258
110 ≤ x < 120	73.000	80.000	-0.040
120 ≤ x < 130	16.000	0.000	2.207
130 ≤ x < 140	0.000	0.000	0.000
140 ≤ x < 150	0.000	0.000	0.000
150 ≤ x < 160	0.000	0.000	0.000
160 ≤ x	0.000	0.000	0.000

Table 7. Frequency distributions of the mean amino acid surface area in immunoglobulin and non-immunoglobulin proteins, and log-odds scores.

Mean surface area (x)	Frequency of immunoglobulin	Frequency of non-Immunoglobulin	Scores
< 130	0.000	0.000	0.000
130 ≤ x < 140	5.000	38.500	-0.750
140 ≤ x < 150	48.000	32.500	0.247
150 ≤ x < 160	47.000	29.000	0.515
160 ≤ x	0.000	0.000	0.000

Table 8. Frequency distributions of the mean amino acid alpha and beta-strand in immunoglobulin and non-immunoglobulin proteins, and log-odds scores.

Mean frequency (x)	Alpha-helix			Beta-strands		
	Frequency of immunoglobulin	Frequency of non-immunoglobulin	Scores	Frequency immunoglobulin	Frequency of non-immunoglobulin	Scores
< 0.80	0.000	0.00	0.000	0.000	0.00	0.000
0.80 ≤ x < 0.90	0.000	2.00	-1.322	0.000	12.000	-2.083
0.9 ≤ x < 0.10	0.600	8.00	0.870	88.000	84.000	0.020
1.00 ≤ x < 1.10	0.400	80.00	-0.300	12.00	4.000	0.470
1.10 ≤ x < 1.20	0.000	10.00	-2.004	0.000	0.000	0.000
1.20 ≤ x	0.000	0.00	0.000	0.000	0.000	0.000

Table 9. The numbers of 7TMR candidates identified from the *Arabidopsis* genome by different methods

Methods	Numbers of 7TMR candidates	Numbers of 7TMR candidates predicted to have 5-10 TM regions	Numbers of 7TMR candidates with external N-terminal[a]
ST-method	659	579	92
SAM	15	15	7
PSI-BLAST	1	1	1
SAM[b]	15	15	7
LDA[b]	2,935	801	97
QDA[b]	2,020	645	87
KNN (K=20)[b]	3,043	767	97
SVM-AA[b]	2,043	772	101
PLS-ACC[b]	2,087	552	67
SVM-di[b]	2,466	750	95
Intersection[c]	595	342	54

[a] The numbers proteins predicted to have seven transmembrane region and external N-terminal by HMMTOP.

[b] Results from Moriyama et al.[3]

[c] Identified by all six alignment-free methods (LDA, QDA, KNN (K=20), SVM-AA, SVM-di, and PLS-ACC) in Moriyama et al.[3]

Table 10. The numbers of 7TMR candidates identified from the maize genome by different methods

	Numbers of 7TMR candidates	Numbers of 7TMR candidates with 5-10 TM regions	Numbers of 7TMR candidates with external N-terminal[a]
ST-method	1100[b]	717	114
PLS-ACC	3820	702	72
SAM	22	5	5
PSI-BLAST	8	3	2

[a] The numbers proteins predicted to have seven transmembrane region and external N-terminal by HMMTOP.

[b] These sequences came from genomic DNA sequences, therefore, they have introns. These sequences need to be confirmed first if they are coding sequences. They may include possible duplicates.

Table 11 The numbers of 7TMR candidates identified from the maize genome by different methods

Methods	Numbers of 7TMR candidates	Numbers of 7TMR candidates with 5-10 TM regions	Numbers of 7TMR candidates with external N-terminal[a]
ST-method	464[b]	382	48
PLS-ACC	1120	364	31
SAM	0	0	0
PSI-BLAST	7	1	0
TBLASTN	194	16	5

[a] The numbers of seven transmembrane region predicted by HMMTOP with N-terminal outside.

[b] These sequences came from genomic DNA sequences, therefore, they have introns. These sequences need to be confirmed first if they are coding sequences. They may include possible duplicates.

4.8 FIGURES

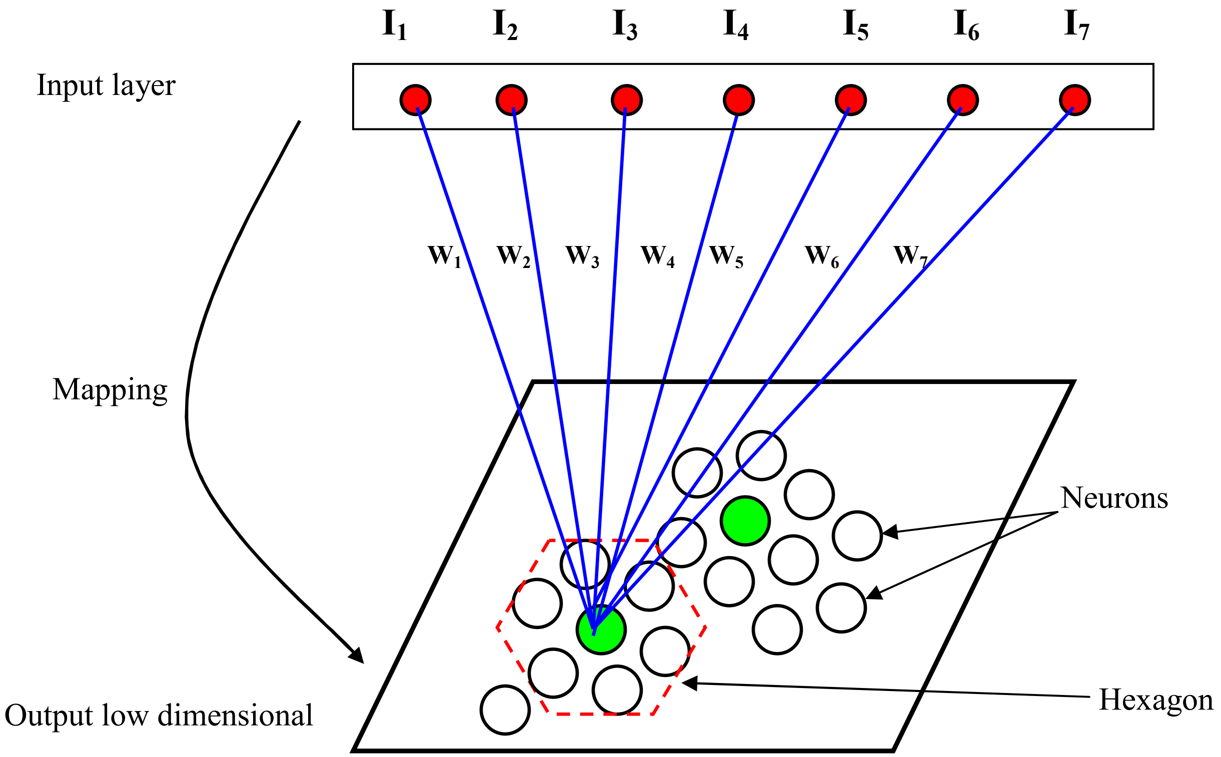


Figure 1. Graphical illustration of a self-organizing map. I_1,…,I_7: Input data high dimensional; W_1,…W_7; weights

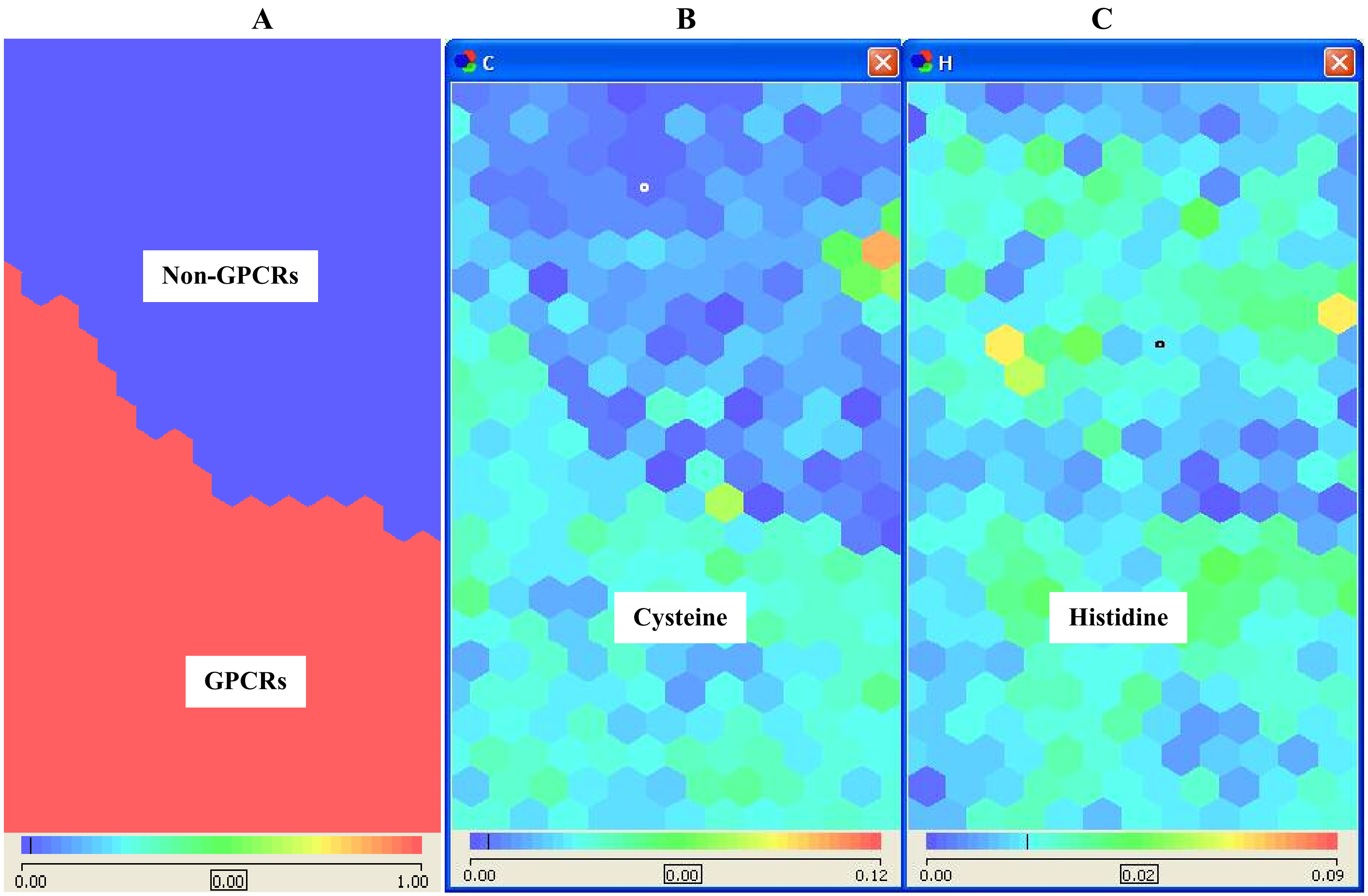


Figure 2. Examples of self-organizing maps for GPCR/non-GPCR (A), cysteine (B),and histidne (C). While cysteine was selected because the map is similar to GPCR/non-GPCR, histidine was not selected because the map is not similar to GPCR/non-GPCR.

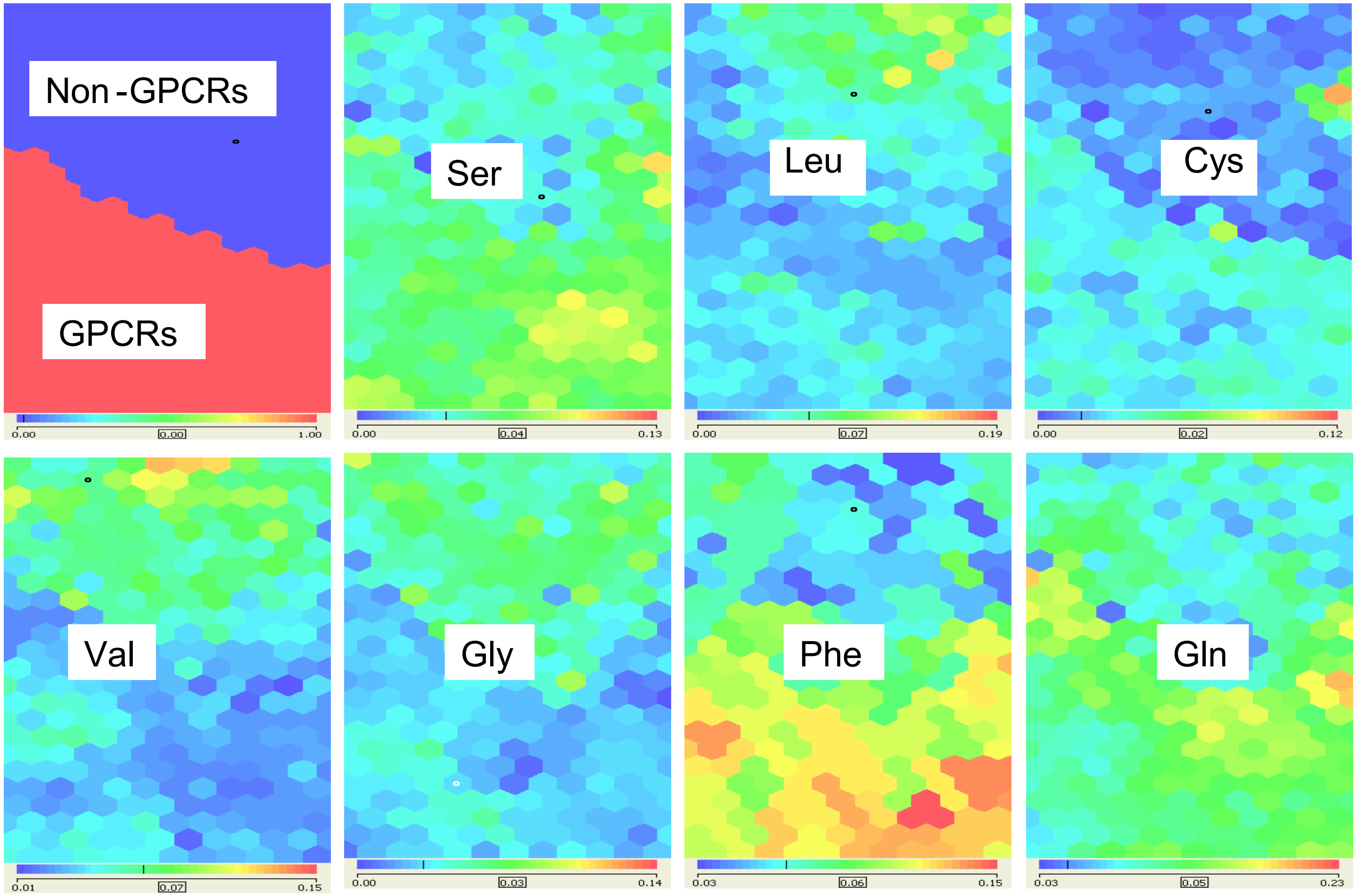


Figure 3. Self organizing maps of GPCRs/non-GPCRs together with seven amino acids that were selected by visual inspection. Seven amino acids are as follows: serine (Ser), leucine (Leu), cysteine (Cys), valine (Val), glycine (Gly), phenylalanine (Phe) and glutamine (Gln).

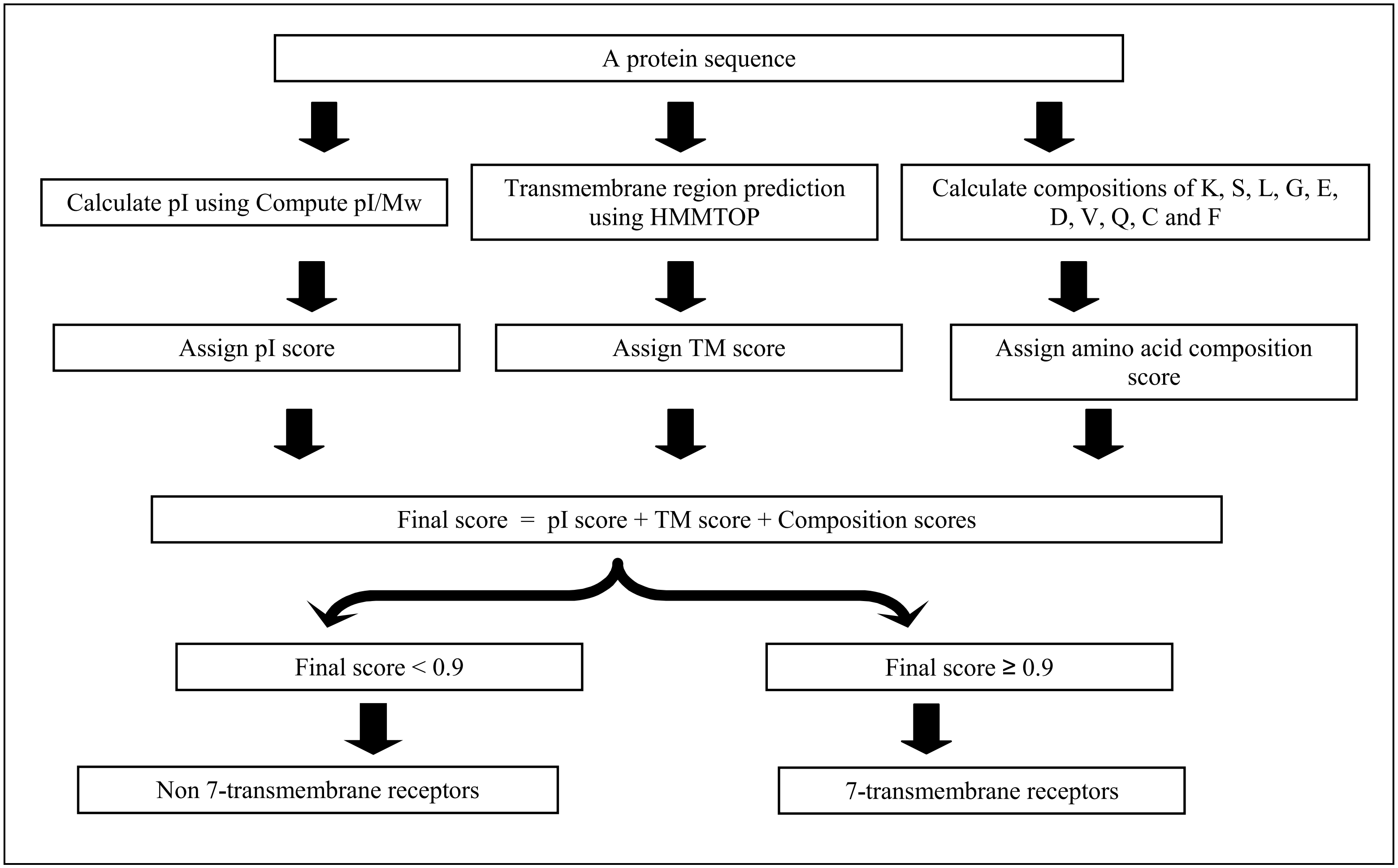

A protein sequence
Calculate pI using Compute pI/Mw
Transmembrane region prediction using HMMTOP
Calculate compositions of K, S, L, G, E, D, V, Q, C and F
Assign pI score
Assign TM score
Assign amino acid composition score
Final score = pI score + TM score + Composition scores
Final score < 0.9
Final score ≥ 0.9
Non 7-transmembrane receptors
7-transmembrane receptors

Figure 4. **A flowchart showing the process of the ST-method for identifying seven transmembrane receptors.**

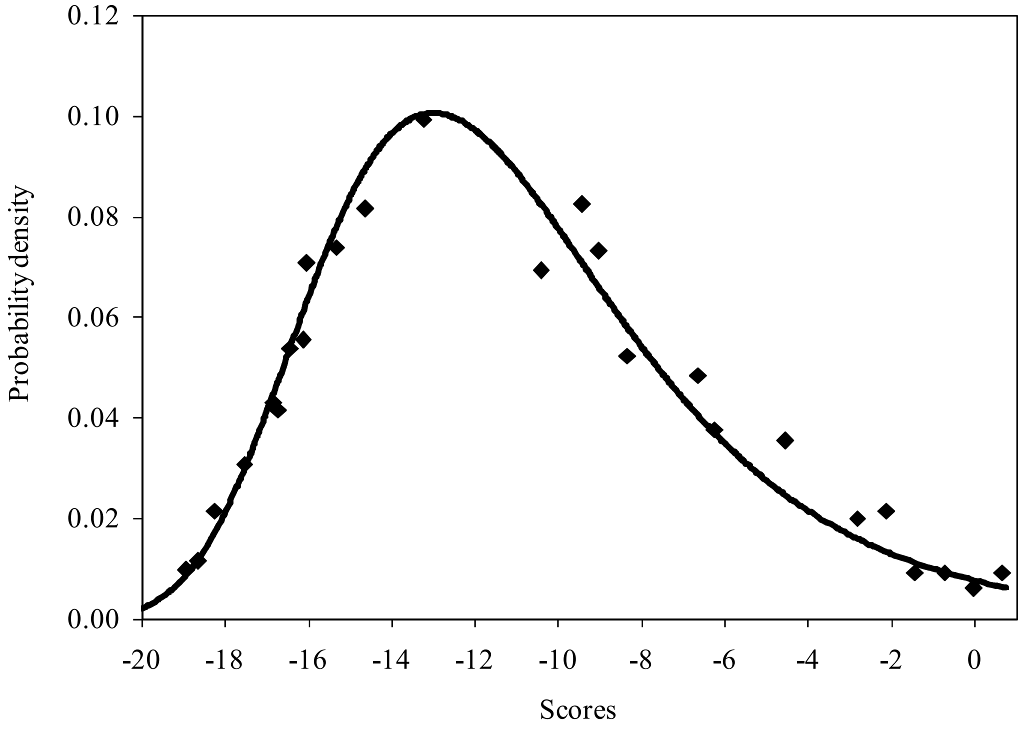


Figure 5. The observed distribution of ST-method scores obtained from 1000 non-GPCRs (mean = -10.8966 and standard deviation = 4.6929). The expected EVD using λ = 0.2733 and μ = -12.9906 calculated based on the observed mean and standard deviation shown in the solid line.

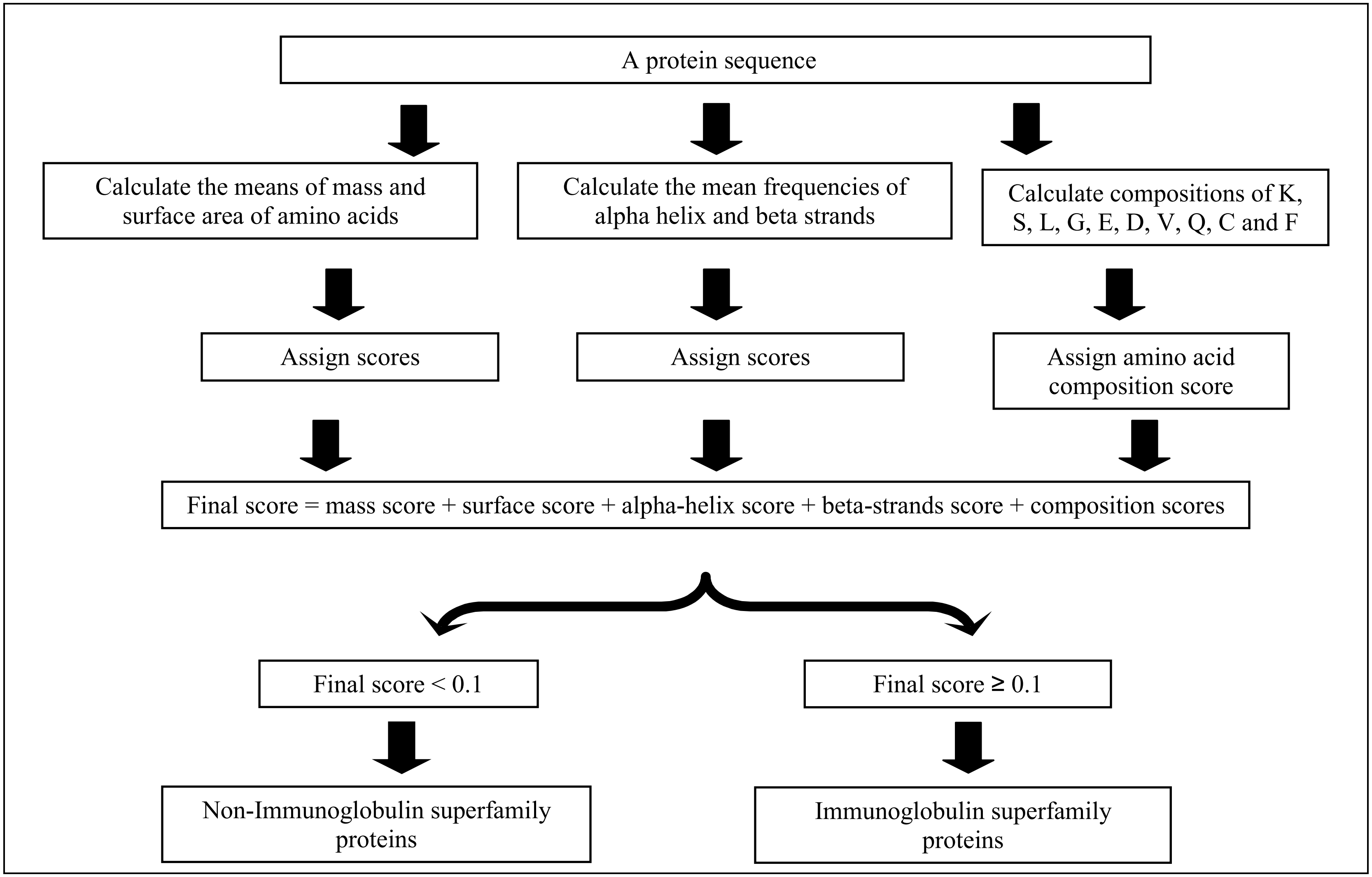

A protein sequence
Calculate the means of mass and surface area of amino acids
Calculate the mean frequencies of alpha helix and beta strands
Calculate compositions of K, S, L, G, E, D, V, Q, C and F
Assign scores
Assign scores
Assign amino acid composition score
Final score = mass score + surface score + alpha-helix score + beta-strands score + composition scores
Final score < 0.1
Final score ≥ 0.1
Non-Immunoglobulin superfamily proteins
Immunoglobulin superfamily proteins

Figure 6. A flowchart showing the process of the of the ST-method for identifying the Immunoglobulin superfamily protein.

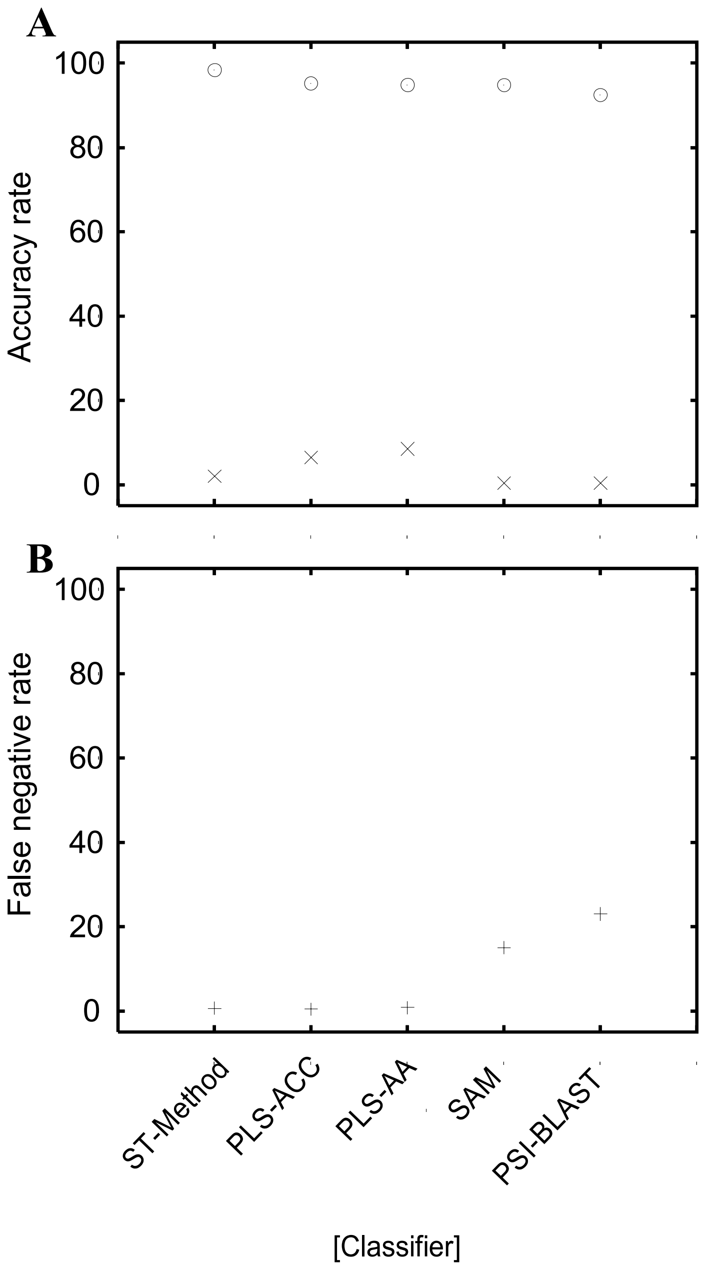


Figure 7. The classifier performance on the GPCR test dataset. A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

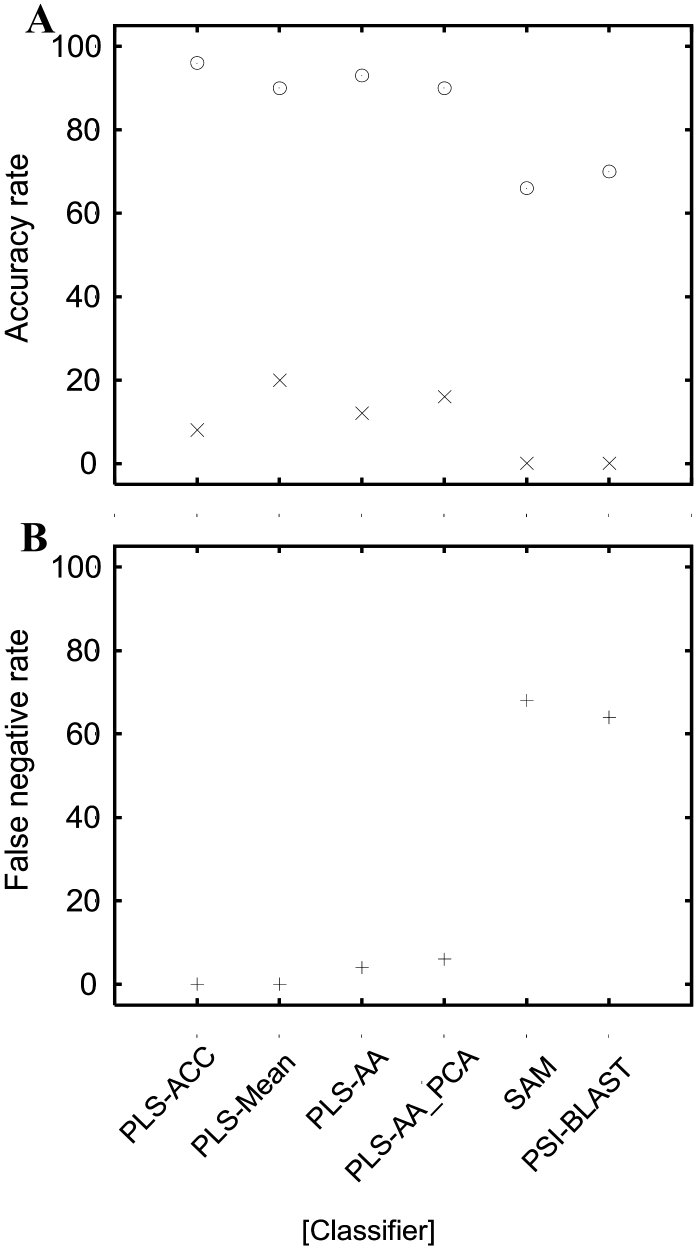


Figure 8. The classifier performance for the immunoglobulin data cross-validation test. A: accuracy rate (%) as well as false positive rate (%, at the bottom); B: false negative rate (%).

CHAPTER 5

CONCLUSION AND FUTURE WORK

In this dissertation, multivariate alignment-free methods were used for classification of protein families with weak sequence similarities. The two multivariate methods used in the dissertation were partial least squares (PLS) and principal component analysis (PCA). Self-organizing maps (SOM) and t-test were also included in the study. These alignment-free methods were compared against representative alignment-based methods: profile hidden Markov models (HMMs) and PSI-BLAST. Four protein families (G-protein coupled receptors: GPCRs, cyclophilins, cytochrome b561: Cyt b561, and immunoglobulins) were chosen to test the classifiers performance. Twelve physico-chemical properties of amino acids and amino acid compositions were used as descriptors. The dimensions of the twelve physico-chemical properties were reduced to five principal components by PCA.

Using the five descriptors developed by PCA and amino acid compositions, four PLS methods: “PLS-ACC” using auto and cross-covariance descriptors, “PLS-mean” based on mean PC scores, “PLS-AA” based on simple amino acid composition, and “PLS-AA_PCA” using amino acid composition transformed with PCA were developed.

Using GPCR datasets, first, how the size of training sets affects the performance of classifiers was examined. These PLS classification methods were not much affected by the number of sequences (e.g., 5 to 10 positive samples) in the training datasets. But alignment-based methods, SAM (profile HMM) and PSI-BLAST were affected. The PLS methods also could identify short GPCR subsequences as short as 150 base pairs that SAM and PSI-BLAST could not identify. When the methods were applied to mine Cyt b561 from *Arabidopsis* Expressed Sequence Tags (ESTs) database, PLS-ACC and PLS-AA identified Cyt b561 ESTs that could not be identified by the profile HMM and PSI-BLAST. We can therefore conclude that PLS methods can be used to identify EST sequences from plant and crop sequence databases that alignment-based methods cannot identify.

There are situations where a protein family may have only a very few sample sequences in a class. For example, currently only 12 members of Class D and 4 members of Class E GPCRs are available. In order to simulate such situations, the methods mentioned above were trained on GPCR class and tested on classes of GPCRs that they were not trained on. PLS methods outperformed alignment-based methods SAM and PSI-BLAST in identifying GPCR sequences that were not included in the training datasets. They can be used to discover unknown or novel GPCRs even if the classifiers may not be trained on such sequences before.

PLS-Methods in combination with other alignment-free methods (support vector machines, discriminant analysis, and K-nearest neighbor) were applied to mine GPCRs from the *Arabidopsis* genome. Using the intersection method, these alignment-free methods identified 342 protein sequences as GPCRs. Among the 342 sequences were 20 of the 22 known *Arabidopsis* GPCRs. But alignment-based method SAM missed 7 of the 22 GPCRs. However, we also found that these alignment-free methods have high false positives. In conclusion, alignment-based methods such as SAM are very specific, and cannot perform well in identifying protein families with weak sequence similarities. On the other hand, alignment-free methods are sensitive to these proteins but they have high false positives.

In order to reduce to reduce the number of false positives without losing the sensitivity, the t-test was used to select descriptors after ACC transformation, and PLS was used to classify cyclophilin protein family. This classifier, PLS_T-ACC outperformed the original PLS-ACC in classifying cyclophilins by producing low false positives. From this study, we can conclude that by selecting only the descriptors that significantly discriminate between cyclophilins and non-cyclophililns, we can reduce the number of false positives.

Finally, a new non-alignment method that is sensitive to low sequence similarities and at the same time has low false positives was developed. Using SOM and t-test, a simple statistics method (ST-method) was applied for the GPCR and immunoglobulin superfamilies. ST-method outperformed PLS-ACC, PLS-AA, SAM, and PSI-BLAST on both GPCR and identification. ST-method had lower false positive rates than PLS-ACC and PLS-AA in both cases. ST-method identified 597 GPCR candidates from the *Arabidopsis* genome including all the 22 known *Arabidopsis* GPCRs, but SAM identified 15, and PSI-BLAST identified 1. ST-method also identified GPCR sequences from rice and maize genomes that SAM, PSI-BLAST and TBLASTN could not identify. It should be noted that some of these sequences may not be GPCRs since they are not yet confirmed experimentally. When applied to mine GPCRs from crop genomes, ST-method also identified 717 from rice and 382 from maize, respectively. The majority of them were not identified by SAM, and PSI-BLAST. It should be noted that not all of these sequences are likely to be GPCRs. They need to be confirmed experimentally. Since the ST-method performed equally well against GPCRs, divergent transmembrane proteins, and immunoglobulin superfamily, soluble proteins, with even lower sequence similarities, we can conclude that ST-method is sensitive to sequences with weak similarities and at the same time has low false positives. It can be used in mining a novel or new proteins, both soluble and transmembrane protein types.

In this dissertation, the mining of Cyt b561 was done from ESTs. It would be interesting to go back and search for Cyt b561 from *Arabidopsis*, rice, and maize genomes. Likewise, in the future, GPCR and cyclophilin ESTs should be mined from plant and crop databases using the methods developed.

In the mining of cyclophilins using PLS-ACC, only one multiple-domain family was included in the study due to too few numbers of sequences from other multiple-domain cyclophilins. To learn more about the cyclophilin superfamily, more information is needed from other multiple-domain cyclophilins.

Three plant genomes *Arabidopsis*, rice, and maize, were mined in this study. Mining of GPCRs, cyclophilins, and Cyt b561 should be extended to other plant and crop genomes such as soybeans, wheat, and potato as their sequences or ESTs become available.

Lastly, we used only twelve physico-chemical properties and five descriptors were chosen by PCA in this study. Other physico-chemical properties of amino acids should be explored to developed new descriptors to identify protein families with weak sequence similarities.

APPENDIX TABLE

Table 1. The lists of plant projects that are based on ESTs

1	*Allium cepa* (onion)
2	*Amborella trichopoda*
3	*Ananas comosus* (pineapple)
4	*Aquilegia formosa x Aquilegia pubescens*
5	*Arabidopsis thaliana*
6	*Arachis hypogaea* (peanut)
7	*Asparagus officinalis* (garden asparagus)
8	*Avena sativa* (oat)
9	*Beta vulgaris subsp. vulgaris* (beet)
10	*Betula pendula* (European white birch)
11	*Brachypodium distachyon*
12	*Brassica napus* (rape)
13	*Brassica rapa L. spp pekinensis* (Chinese cabbage)
14	*Bruguiera gymnorrhiza* (Burma mangrove)
15	*Ceratadon purpureus* (moss)
16	*Cicer arietinum* (chickpea)
17	*Cichorium intybus* (chicory)
18	*Citrus aurantium* (Seville orange)
19	*Citrus clementina*
20	*Citrus jambhiri* (jambhiri orange)
21	*Citrus macrophylla* (colo)
22	*Citrus reticulata x Citrus temple*
23	*Citrus reticulata* (tangerine)
24	*Citrus sinensis* (apfelsine/navel orange)
25	*Citrus unshiu* (Satsuma orange)
26	*Citrus x paradisi* (grapefruit)
27	*Citrus sinensis x Poncirus trifoliata* (Carrizo citrange)
28	*Citrus x paradisi x Poncirus trifoliata*
29	*Coffea arabica L.* (coffee)
30	*Cucumis sativus* (cucumber)
31	*Cynodon dactylon* (Bermuda grass)
32	*Eleusine coracana* (finger millet)
33	*Eragrostis tef* (tef)
34	*Eschscholzia californica* (California poppy)
35	*Eucalyptus*
36	*Euphorbia esula* (leafy spurge)
37	*Euphorbia tirucalli*

Table 1 (continued).

38	*Festuca arundinacea* (tall fescue)
39	*Fragaria x ananassa* (strawberry)
40	*Gerbera hybrid cv.* 'Terra Regina'
41	*Glycine max* (domesticated soybean)
42	*Glycine soja* (wild soybean)
43	*Gossypium arboreum* (tree cotton)
44	*Gossypium hirsutum* (upland cotton)
45	*Gossypium raimondi* (cotton)
46	*Hedyotis centrathoides*
47	*Hedyotis terminalis*
48	*Hevea brasiliensis* (Para rubber tree)
49	*Helianthus annuus* (sunflower)
50	*Helianthus paradoxus*
51	*Hordeum vulgare* (barley)
52	*Hordeum vulgare subsp spontaneum* (wild barley)
53	*Hordeum vulgare subsp vulgare* (two-rowed barley)
54	*Juglans regia* (English walnut)
55	*Lactuca sativa* (lettuce)
56	*Lactuca serriola* (lettuce)
57	*Lilium longiflorum* (trumpet lily)
58	*Limonium bicolor*
59	*Linum usitatissimum* (flax)
60	*Liriodendron tulipifera*
61	*Lolium multiflorum* (Italian ryegrass)
62	*Lotus japonicus*
63	*Lupinus albus* (white lupine)
64	*Lycoris longituba*
65	*Malus sieboldii* (Toringo crab-apple)
66	*Malus x domestica* (domesticated apple)
67	*Malus x domestica x Malus sieversii*
68	*Manihot esculenta* (cassava)
69	*Marchantia polymorpha* (liverwort)
70	*Medicago sativa* (alfalfa)
71	*Medicago truncatula* (barrel medic)
72	*Melaleuca alternifolia* (tea tree)
73	*Mesembryanthemum crystallinum* (common ice plant)
74	*Musa acuminata* (Cavendish banana)
75	*Nicotiana benthamiana*
76	*Nuphar advena*
77	*Oryza sativa* (rice)

Table 1 (continued).

78	*Panax ginseng* (ginseng)
79	*Panicum virgatum* (switchgrass)
80	*Persea americana* (avocado)
81	*Phaseolus coccineus* (bean)
82	*Phaseolus vulgaris* (kidney bean)
83	*Picea glauca* (white spruce)
84	*Picea sitchensis* (Sitka spruce)
85	*Pinus radiata* (Monterey pine)
86	*Pinus taeda* (loblolly pine)
87	*Plumbago zeylanica*
88	*Poncirus trifoliata*
89	*Populus alba x Populus tremula* (gray poplar)
90	*Populus deltoides*
91	*Populus euphratica*
92	*Populus tremula* (European aspen)
93	*Populus tremula x Populus tremuloides* (poplar)
94	*Populus tremuloides* (quaking aspen)
95	*Populus trichocarpa* (western balsam poplar)
96	*Populus trichocarpa x Populus deltoides*
97	*Populus trichocarpa x Populus nigra*
98	*Prunus armeniaca* (apricot)
99	*Prunus dulcis* (almond)
100	*Prunus persica* (peach)
101	*Quercus petraea* (sessile oak)
102	*Quercus robur* (English oak)
103	*Ricinus communis* (castor bean)
104	*Robinia pseudoacacia* (black locust)
105	*Rosa chinensis* (China rose)
106	*Rosa hybrid cultivar*
107	*Saccharum sp.* (sugarcane)
108	*Salix viminalis* (basket willow)
109	*Saruma henryi*
110	*Secale cereale* (rye)
111	*Solanum habrochaites*
112	*Solanum lycopersicuum* (tomato)
113	*Solanum pennellii*
114	*Solanum tuberosum* {potato}
115	*Sorghum bicolor* (sorghum)
116	*Sorghum halepense* (Johnson grass)

Table 1 (continued).

117	*Sorghum propinquum*
118	*Stevia rebaudiana* (stevia)
119	Tamarix androssowii
120	*Taraxacum kok-saghyz*
121	*Theobroma cacao* (cacao)
122	*Triticum aestivum* (wheat)
123	*Vaccinium spp.* (blueberry)
124	*Vitis aestivalis*
125	*Vitis hybrid cultivar*
126	*Vitis riparia* (riverbank grape)
127	*Vitis shuttleworthii* (callose grape)
128	*Vitis vinifera* (wine grape)
129	*Yucca filamentosa* (spoon-leaf yucca)
130	*Zantedeschia aethiopica* (arum-lily)
131	*Zea mays* (corn)

Table 2. The 342 *Arabidopsis thaliana* 7TMR candidates with 5-10 transmembrane regions predicted by six alignment-free methods.

At1g01070.1	At1g23020.1	At1g63010.1	At2g21080.1	At3g06550.1	At4g01430.1	At4g25750.1	At5g22130.1
At1g01070.2	At1g23830.1	At1g63010.2	At2g23680.1	At3g07080.1	At4g01430.2	At4g26580.1	At5g23660.1
At1g01580.1	At1g23840.1	At1g63110.1	At2g24150.1	At3g07330.1	At4g01440.1	At4g26770.1	At5g23980.1
At1g01590.1	At1g23850.1	At1g63110.2	At2g24170.1	At3g08930.2	At4g02600.1	At4g27970.1	At5g23990.1
At1g02260.1	At1g24400.1	At1g63110.3	At2g25270.1	At3g09320.1	At4g02690.1	At4g28040.1	At5g24030.1
At1g02570.1	At1g25500.1	At1g63120.1	At2g28170.1	At3g09340.1	At4g02900.1	At4g28040.2	At5g24600.1
At1g05360.1	At1g26180.1	At1g63690.2	At2g28315.1	At3g09570.1	At4g03410.2	At4g28210.1	At5g24790.1
At1g05820.1	At1g26440.2	At1g66760.1	At2g29050.1	At3g10290.1	At4g03820.1	At4g28370.1	At5g25100.1
At1g06100.1	At1g26440.3	At1g67060.1	At2g29650.2	At3g11320.1	At4g03820.2	At4g29200.1	At5g25420.1
At1g06350.1	At1g26650.1	At1g67570.1	At2g29900.1	At3g11810.1	At4g03950.1	At4g30850.1	At5g27210.1
At1g06470.1	At1g26700.1	At1g67640.1	At2g29980.2	At3g13420.1	At4g04340.1	At4g30850.2	At5g27730.1
At1g06470.2	At1g26730.1	At1g68000.1	At2g31440.1	At3g13772.1	At4g04340.2	At4g35180.1	At5g33320.1
At1g08350.1	At1g28220.1	At1g68070.1	At2g32295.1	At3g15380.1	At4g04340.3	At4g35870.1	At5g35160.1
At1g08960.1	At1g28760.1	At1g68170.1	At2g32530.1	At3g16090.1	At4g08290.2	At4g36830.1	At5g35460.1
At1g09380.1	At1g29330.1	At1g68820.1	At2g32610.1	At3g16690.1	At4g08700.1	At4g36850.1	At5g35730.1
At1g09860.1	At1g29390.1	At1g69430.1	At2g33205.1	At3g17430.1	At4g08878.1	At4g38640.1	At5g35810.1
At1g10090.1	At1g29390.2	At1g69450.1	At2g33670.1	At3g18215.1	At4g10310.1	At4g39030.1	At5g37310.1
At1g10660.1	At1g29395.1	At1g70260.1	At2g33750.1	At3g19260.1	At4g10360.1	At4g39390.1	At5g40780.1
At1g10660.2	At1g30360.1	At1g70505.1	At2g33750.2	At3g20300.1	At4g11680.1	At4g39390.2	At5g40780.2
At1g10660.3	At1g30840.1	At1g71680.1	At2g34390.1	At3g21620.1	At4g12000.1	At5g01460.1	At5g41160.1
At1g10660.4	At1g31130.1	At1g71960.1	At2g34390.2	At3g24460.1	At4g12650.1	At5g02630.1	At5g41800.1
At1g10950.1	At1g31300.1	At1g72130.2	At2g34980.1	At3g25040.1	At4g13345.1	At5g04160.1	At5g42090.1
At1g11000.1	At1g32400.1	At1g73950.1	At2g35650.1	At3g25540.1	At4g13345.2	At5g05310.1	At5g42420.1
At1g11310.1	At1g32400.2	At1g74810.1	At2g35710.1	At3g25805.1	At4g13800.1	At5g05310.2	At5g44860.1
At1g11460.1	At1g34490.1	At1g75000.1	At2g35710.2	At3g25950.1	At4g14730.1	At5g05310.3	At5g45095.1
At1g11540.1	At1g34500.1	At1g75470.1	At2g36300.1	At3g26090.1	At4g15430.1	At5g05350.1	At5g45105.1
At1g11880.1	At1g34520.1	At1g76530.1	At2g36305.1	At3g27270.1	At4g15470.1	At5g05820.1	At5g47900.1
At1g12450.1	At1g42560.1	At1g77220.1	At2g36590.1	At3g28007.1	At4g16600.1	At5g07050.1	At5g49630.1
At1g12480.1	At1g43580.1	At1g77860.1	At2g37940.1	At3g28050.1	At4g16850.1	At5g07250.1	At5g50800.1
At1g12500.1	At1g44010.1	At2g01070.1	At2g38120.1	At3g28060.1	At4g17250.1	At5g07630.1	At5g52180.1
At1g12730.1	At1g44750.2	At2g01735.1	At2g39200.1	At3g30340.1	At4g17790.1	At5g10840.1	At5g53760.1

Table 2 (continued).

At1g12750.1
At1g13580.1
At1g14530.1
At1g14530.2
At1g14670.1
At1g15110.1
At1g15600.1
At1g15610.1
At1g15620.1
At1g15640.1
At1g15960.1
At1g16560.1
At1g16560.2
At1g16560.3
At1g18470.1
At1g19770.1
At1g20050.1
At1g21460.1
At1g22750.1
At1g44960.1
At1g48230.1
At1g48270.1
At1g48460.1
At1g48640.1
At1g49470.1
At1g50630.1
At1g52615.1
At1g55130.1
At1g55240.1
At1g57680.1
At1g57680.2
At1g57943.1
At1g57980.1
At1g57990.1
At1g58520.1
At1g60050.1
At1g61560.1
At1g62320.1
At2g01970.1
At2g02180.1
At2g02370.1
At2g03330.1
At2g03530.1
At2g03590.1
At2g03600.1
At2g04360.1
At2g05755.1
At2g07739.1
At2g12400.1
At2g15240.1
At2g16970.1
At2g17430.1
At2g17480.1
At2g18680.1
At2g18950.1
At2g19450.1
At2g20725.1
At2g39805.1
At2g41610.1
At2g41705.1
At2g44110.1
At2g44110.2
At2g45150.3
At2g46440.1
At2g46450.1
At2g46890.1
At2g47115.1
At2g47360.1
At2g47600.1
At2g47760.1
At3g01550.1
At3g03700.1
At3g04440.1
At3g04970.1
At3g04970.2
At3g06470.1
At3g45290.1
At3g48200.1
At3g48740.1
At3g51970.1
At3g52760.1
At3g53780.2
At3g54020.1
At3g54450.1
At3g54510.1
At3g57170.1
At3g58460.1
At3g59090.1
At3g59090.2
At3g60590.1
At3g60590.2
At3g60590.3
At3g60620.1
At3g61750.1
At3g63310.1
At4g18190.1
At4g18210.1
At4g18220.1
At4g18230.1
At4g18540.1
At4g19645.1
At4g19645.2
At4g19950.1
At4g20310.1
At4g21260.1
At4g21570.1
At4g21790.1
At4g22270.1
At4g22330.1
At4g22340.1
At4g22990.1
At4g24250.1
At4g25010.1
At4g25350.1
At5g11870.1
At5g12170.1
At5g13170.1
At5g13670.1
At5g13750.2
At5g13760.1
At5g13890.1
At5g13890.2
At5g13890.3
At5g14570.2
At5g15240.1
At5g15410.1
At5g15410.2
At5g17630.1
At5g17830.1
At5g18520.1
At5g19380.1
At5g19870.1
At5g20270.1
At5g55320.1
At5g55370.1
At5g55950.1
At5g58560.1
At5g60220.1
At5g60750.1
At5g62130.1
At5g62960.1
At5g63040.1
At5g63040.2
At5g64700.1
At5g65970.1
At5g66450.1

Table 3. The 290 *Arabidopsis thaliana* single domain cyclophilin candidates predicted by PLS_T-ACC.

At1g01940.1	At1g01940.1	At1g62510.1	At2g14460.1	At2g33775.1	At3g09735.1	At3g42310.1
At1g02710.1	At1g02710.1	At1g63390.1	At2g16200.1	At2g35300.1	At3g10116.1	At3g44940.1
At1g03020.1	At1g03020.1	At1g65350.1	At2g16600.1	At2g36130.1	At3g13570.1	At3g45370.1
At1g05920.1	At1g05920.1	At1g66590.1	At2g16940.1	At2g36170.1	At3g15353.1	At3g45730.1
At1g07600.1	At1g07600.1	At1g66590.2	At2g17870.1	At2g37120.1	At3g15520.1	At3g47965.1
At1g08020.1	At1g08020.1	At1g68430.1	At2g18340.1	At2g38140.1	At3g15670.1	At3g48120.1
At1g10460.1	At1g10460.1	At1g68845.1	At2g18810.1	At2g38730.1	At3g15790.1	At3g48450.1
At1g12090.1	At1g12090.1	At1g69750.1	At2g19030.1	At2g39320.1	At3g17520.1	At3g49270.1
At1g12725.1	At1g12725.1	At1g72260.1	At2g19040.1	At2g39540.1	At3g17580.1	At3g49307.1
At1g14060.1	At1g14060.1	At1g72720.1	At2g19750.1	At2g40650.1	At3g18550.1	At3g50540.1
At1g15400.2	At1g15400.2	At1g73885.1	At2g20595.1	At2g41420.1	At3g20900.1	At3g50670.1
At1g16000.1	At1g16000.1	At1g74070.1	At2g21130.1	At2g41730.1	At3g22120.1	At3g52590.1
At1g16610.1	At1g16610.1	At1g76010.1	At2g22425.1	At2g42000.1	At3g22820.1	At3g53250.1
At1g16610.2	At1g16610.2	At1g76300.1	At2g22470.1	At2g42310.1	At3g22920.1	At3g53370.1
At1g16850.1	At1g16850.1	At1g78810.1	At2g23240.1	At2g47110.1	At3g23900.1	At3g53530.1
At1g16950.1	At1g16950.1	At1g78810.2	At2g23240.2	At2g47200.1	At3g24640.1	At3g55920.1
At1g17490.1	At1g17490.1	At2g01200.1	At2g23270.1	At2g47320.1	At3g24710.1	At3g56010.1
At1g20580.1	At1g20580.1	At2g01310.1	At2g23940.1	At3g01170.1	At3g25165.1	At3g56070.1
At1g21520.1	At1g21520.1	At2g02100.1	At2g24590.1	At3g04270.1	At3g25170.1	At3g56240.1
At1g22140.1	At1g22140.1	At2g02120.1	At2g25890.1	At3g05220.1	At3g27410.1	At3g56720.1
At1g22140.2	At1g22140.2	At2g02130.1	At2g26120.1	At3g05220.2	At3g28790.1	At3g57785.1
At1g22990.1	At1g22990.1	At2g02440.1	At2g26520.1	At3g05460.1	At3g29090.1	At3g59900.1
At1g23410.1	At1g23410.1	At2g03720.1	At2g27830.1	At3g05730.1	At3g29280.1	At3g61980.1
At1g25275.1	At1g25275.1	At2g04600.1	At2g29210.1	At3g05860.2	At3g29600.1	At3g62030.1
At1g26550.1	At1g26550.1	At2g05185.1	At2g29960.1	At3g06110.1	At3g30350.1	At3g62250.1
At1g26940.1	At1g26940.1	At2g07505.1	At2g29995.1	At3g06895.1	At3g30530.1	At3g63100.1
At1g27330.1	At1g27330.1	At2g07785.1	At2g33130.1	At3g07230.1	At3g42130.1	At3g66654.1

Table 3 (continued).

At4g00140.1
At4g00180.1
At4g02840.1
At4g02890.1
At4g02890.2
At4g02890.3
At4g03205.2
At4g03750.1
At4g05010.1
At4g05050.1
At4g05320.1
At4g05320.2
At4g05320.3
At4g05320.4
At4g05320.5
At4g06728.1
At4g09530.1
At4g09610.1
At4g09840.1
At4g11510.1
At4g11760.1
At4g12190.1
At4g12510.1
At4g12520.1
At4g13195.1
At4g15030.1
At4g15735.1
At4g16380.1
At4g18580.1
At4g18580.2
At4g20280.1
At4g20690.1
At4g21020.1
At4g26965.1
At4g28180.1
At4g28360.1
At4g29280.1
At4g29390.1
At4g30500.1
At4g30670.1
At4g31580.1
At4g32080.1
At4g32420.1
At4g33060.1
At4g33610.1
At4g34870.1
At4g34960.1
At4g35785.1
At4g35785.2
At4g36690.1
At4g36690.2
At4g36690.3
At4g38740.1
At5g03240.1
At5g03240.2
At5g03370.1
At5g03580.1
At5g06250.1
At5g09880.1
At5g11760.1
At5g12990.1
At5g13120.1
At5g13340.1
At5g14330.1
At5g14690.1
At5g15260.1
At5g17450.2
At5g17650.1
At5g18180.1
At5g18810.1
At5g19480.1
At5g20620.1
At5g22280.1
At5g23760.1
At5g24590.1
At5g26350.1
At5g27860.1
At5g28463.1
At5g33390.1
At5g37640.1
At5g40420.1
At5g40570.1
At5g41440.1
At5g42635.1
At5g43260.1
At5g44310.2
At5g48657.1
At5g49050.1
At5g49400.1
At5g51210.1
At5g52730.1
At5g53800.1
At5g53880.1
At5g56670.1
At5g57370.1
At5g58710.1
At5g59845.1
At5g60800.1
At5g61590.1
At5g61610.1
At5g62750.1
At5g63460.1
At5g63460.2
At5g33230.1
At5g64140.1
At5g64200.1
At5g64200.2
At4g15460.1
At4g36780.1
At5g28800.1
At5g33210.1
At4g15160.1
At4g36730.1
At5g28720.1
At5g02650.1

Table 4. The 31 *Arabidopsis thaliana* single domain cyclophilin candidates predicted by SAM and PSI-BLAST

At1g01940.1
At1g26940.1
At1g44478.1
At1g53720.1
At1g74070.1
At2g15790.1
At2g16600.1
At2g21130.1
At2g29960.1
At2g36130.1
At2g38730.1
At2g47320.1
At3g01480.1
At3g15520.1
At3g22920.1
At3g44600.1
At3g55920.1
At3g56070.1
At3g62030.1
At3g63400.1
At3g63400.2
At3g66654.1
At3g66654.2
At3g66654.3
At4g17070.1
At4g32420.1
At4g33060.1
At4g34870.1
At4g34960.1
At4g38740.1
At5g13120.1
At5g35100.1
At5g58710.1
At5g67530.1

Table 5. The 110 *Arabidopsis thaliana* multiple domain cyclophilin candidates predicted by PLS_T-ACC.

At1g03020.1
At1g08180.1
At1g10370.1
At1g10420.1
At1g27435.1
At1g27710.1
At1g32290.1
At1g36030.1
At1g43415.1
At1g47395.1
At1g47400.1
At1g53680.1
At1g55060.1
At1g55675.1
At1g60640.1
At1g63270.1
At1g64560.1
At1g65350.1
At1g70830.1
At1g70830.2
At1g70850.1
At1g78110.1
At2g03180.1
At2g05440.2
At2g05520.1
At2g10020.1
At2g14340.1
At2g15170.1
At2g15790.1
At2g16600.1
At2g16700.1
At2g18540.1
At2g21130.1
At2g21660.1
At2g24410.1
At2g26120.1
At2g29960.1
At2g31200.1
At2g31490.1
At2g34120.1
At2g36030.1
At2g36130.1
At2g37950.1
At2g38730.1
At2g40475.1
At2g47200.1
At3g03405.1
At3g05460.1
At3g28940.1
At3g29075.1
At3g29780.1
At3g33230.1
At3g42130.1
At3g42310.1
At3g42560.1
At3g43110.1
At3g45180.1
At3g47965.1
At3g50540.1
At3g52040.1
At3g52550.1
At3g53232.1
At3g53740.2
At3g55920.1
At3g56070.1
At3g62030.1
At3g62990.1
At4g01895.1
At4g02170.1
At4g02890.1
At4g02890.2
At4g02890.3
At4g04398.1
At4g05320.1
At4g05320.2
At4g05320.3
At4g05320.4
At4g09840.1
At4g15460.1
At4g19095.1
At4g20350.1
At4g23870.1
At4g24972.1
At4g25433.1
At4g26290.1
At4g27360.1
At4g28460.1
At4g34620.1
At4g34870.1
At4g34960.1
At4g38740.1
At5g02020.2
At5g03240.1
At5g03240.2
At5g03580.1
At5g07100.2
At5g08185.1
At5g12990.1
At5g13120.1
At5g19270.1
At5g20620.1
At5g25240.1
At5g26350.1
At5g27810.1
At5g28010.1
At5g33390.1
At5g39240.1
At5g41440.1
At5g42300.1
At5g42635.1
At5g49260.1
At5g49590.1
At5g52200.1
At5g55430.1
At5g57570.1
At5g58710.1
At5g63030.1

Table 6. The 91 *Arabidopsis thaliana* multiple domain cyclophilin candidates predicted by SAM.

At1g01940.1	At2g06210.1	At3g17970.1	At4g33060.1
At1g02910.1	At2g06210.2	At3g21640.1	At4g34870.1
At1g04130.1	At2g15790.1	At3g22920.1	At4g34960.1
At1g04190.1	At2g16600.1	At3g25230.1	At4g35230.1
At1g05150.1	At2g20000.1	At3g44600.1	At4g37460.1
At1g12270.1	At2g21130.1	At3g50030.1	At4g38740.1
At1g18660.1	At2g25290.1	At3g54010.1	At4g39820.1
At1g18660.2	At2g29960.1	At3g54010.2	At5g03160.1
At1g18660.3	At2g32450.1	At3g55920.1	At5g09420.1
At1g18660.4	At2g36130.1	At3g56070.1	At5g10090.1
At1g22700.1	At2g38730.1	At3g58620.1	At5g10200.1
At1g22700.2	At2g41520.1	At3g62030.1	At5g10940.1
At1g26760.1	At2g41520.2	At3g63400.1	At5g12430.1
At1g26940.1	At2g42580.1	At3g63400.2	At5g13120.1
At1g33400.1	At2g42810.1	At3g66654.1	At5g20360.1
At1g44478.1	At2g47320.1	At3g66654.2	At5g21990.1
At1g50990.1	At3g01480.1	At3g66654.3	At5g35100.1
At1g53300.1	At3g04240.1	At4g08320.1	At5g43120.1
At1g53720.1	At3g04710.1	At4g11260.1	At5g48570.1
At1g56090.1	At3g07370.1	At4g12400.1	At5g48850.1
At1g56440.1	At3g11540.1	At4g22670.1	At5g58710.1
At1g58450.1	At3g11540.2	At4g23570.1	At5g59010.1
At1g62390.1	At3g14950.1	At4g2[illegible]	At5g65160.1
At1g62740.1	At3g16320.1	At4g30480.1	At5g67530.1
At1g74070.1	At3g16760.1	At4g30480.2	
At1g77230.1	At3g17670.1	At4g32070.1	
At1g78120.1	At3g17880.1	At4g32420.1	

Table 7. The 432 *Arabidopsis thaliana* multiple domain cyclophilin candidates predicted by PSI-BLAST.

At1g01940.1
At1g01970.1
At1g02060.1
At1g02150.1
At1g02370.1
At1g02420.1
At1g03100.1
At1g03510.1
At1g03540.1
At1g03560.1
At1g04590.2
At1g04840.1
At1g05150.1
At1g05600.1
At1g05670.1
At1g05750.1
At1g06140.1
At1g06150.1
At1g06270.1
At1g06580.1
At1g06710.1
At1g07590.1
At1g07740.1
At1g08070.1
At1g08610.1
At1g09190.1
At1g09220.1
At1g09410.1
At1g09680.1
At1g09820.1
At1g09900.1
At1g10270.1
At1g10330.1
At1g10910.1
At1g11630.1
At1g11710.1
At1g11900.1
At1g12270.1
At1g12300.1
At1g12620.1
At1g12700.1
At1g12770.1
At1g13040.1
At1g13410.1
At1g13630.1
At1g13800.1
At1g14470.1
At1g15480.1
At1g15510.1
At1g16480.1
At1g16830.1
At1g17630.1
At1g18900.1
At1g18900.2
At1g19290.1
At1g19525.1
At1g19720.1
At1g20230.1
At1g20300.1
At1g22830.1
At1g22960.1
At1g23450.1
At1g25360.1
At1g26460.1
At1g26500.1
At1g26900.1
At1g28000.1
At1g28020.1
At1g28690.1
At1g29710.1
At1g30290.1
At1g30610.1
At1g31430.1
At1g31790.1
At1g31840.1
At1g31920.1
At1g32415.1
At1g33350.1
At1g34160.1
At1g43010.1
At1g43980.1
At1g50270.1
At1g51965.1
At1g52620.1
At1g52640.1
At1g53330.1
At1g53600.1
At1g55630.1
At1g55890.1
At1g56570.1
At1g56690.1
At1g59720.1
At1g60770.1
At1g61870.1
At1g62260.1
At1g62590.1
At1g62670.1
At1g62680.1
At1g62720.1
At1g62860.1
At1g62910.1
At1g62930.1
At1g63070.1
At1g63080.1
At1g63130.1
At1g63150.1
At1g63230.1
At1g63320.1
At1g63330.1
At1g63400.1
At1g63630.1
At1g64100.1
At1g64310.1
At1g64580.1
At1g66345.1
At1g68930.1
At1g68980.1
At1g69290.1
At1g69350.1
At1g71060.1
At1g71210.1
At1g71420.1
At1g71460.1
At1g71490.1
At1g73400.1
At1g73710.1
At1g74400.1
At1g74580.1
At1g74600.1
At1g74630.1
At1g74750.1
At1g74850.1
At1g74900.1
At1g76280.1
At1g77010.1
At1g77170.1
At1g77340.1
At1g77360.1
At1g77405.1
At1g79080.1
At1g79490.1
At1g79540.1
At1g80150.1
At1g80270.1
At1g80270.2
At1g80550.1
At1g80880.1
At2g01360.1
At2g01390.1
At2g01510.1
At2g01740.1
At2g01860.1
At2g02150.1
At2g02750.1
At2g02980.1
At2g03380.1
At2g03880.1
At2g04860.1
At2g06000.1
At2g06000.2
At2g13600.1
At2g15630.1
At2g15690.1
At2g15790.1
At2g15820.1
At2g15980.1
At2g16600.1
At2g16880.1
At2g17140.1
At2g17210.1
At2g17525.1
At2g17670.1
At2g17670.2
At2g18520.1
At2g18940.1
At2g19280.1
At2g20000.1
At2g20540.1
At2g20710.1
At2g20710.2
At2g21090.1
At2g21130.1
At2g22410.1
At2g25580.1
At2g26790.1
At2g27800.1
At2g28050.1
At2g29760.1
At2g29960.1

Table 7 (continued).

At2g30100.1	At2g41720.1	At3g11380.1	At3g22670.1	At3g49170.1	At3g62470.1	At4g17910.1
At2g30780.1	At2g41720.2	At3g11460.1	At3g22690.1	At3g49240.1	At3g62540.1	At4g18520.1
At2g31400.1	At2g42580.1	At3g12770.1	At3g22920.1	At3g49710.1	At3g62890.1	At4g18750.1
At2g32230.1	At2g42920.1	At3g13150.1	At3g23020.1	At3g49730.1	At3g63370.1	At4g18840.1
At2g32450.1	At2g44880.1	At3g13160.1	At3g23330.1	At3g49740.1	At3g63400.1	At4g19220.1
At2g32630.1	At2g45350.1	At3g13770.1	At3g24000.1	At3g50420.1	At3g63400.2	At4g19440.1
At2g33680.1	At2g46050.1	At3g13880.1	At3g25060.1	At3g51320.1	At4g01030.1	At4g19900.1
At2g33760.1	At2g48000.1	At3g14330.1	At3g25210.1	At3g53170.1	At4g01400.1	At4g20090.1
At2g34370.1	At3g01580.1	At3g14580.1	At3g25230.1	At3g53360.1	At4g01570.1	At4g20740.1
At2g34400.1	At3g02010.1	At3g14730.1	At3g25970.1	At3g53700.1	At4g01990.1	At4g20770.1
At2g35030.1	At3g02330.1	At3g15130.1	At3g26540.1	At3g55920.1	At4g02750.1	At4g21070.1
At2g35130.1	At3g02490.1	At3g15200.1	At3g26630.1	At3g56030.1	At4g02820.1	At4g21170.1
At2g36130.1	At3g02650.1	At3g15590.1	At3g26780.1	At3g56070.1	At4g04370.1	At4g21190.1
At2g36240.1	At3g03580.1	At3g15930.1	At3g28640.1	At3g56550.1	At4g04790.1	At4g21300.1
At2g36730.1	At3g04130.1	At3g16010.1	At3g28660.1	At3g57430.1	At4g08210.1	At4g21705.1
At2g36980.1	At3g04240.1	At3g16610.1	At3g29230.1	At3g58590.1	At4g11690.1	At4g21880.1
At2g37230.1	At3g04260.1	At3g16710.1	At3g29290.1	At3g58620.1	At4g12400.1	At4g22760.1
At2g37310.1	At3g04750.1	At3g16890.1	At3g42630.1	At3g59040.1	At4g13650.1	At4g25270.1
At2g37320.1	At3g04760.1	At3g17370.1	At3g44600.1	At3g59040.2	At4g14050.1	At4g26680.1
At2g37400.1	At3g05240.1	At3g18020.1	At3g46610.1	At3g60040.1	At4g14170.1	At4g26800.1
At2g38420.1	At3g05340.1	At3g18110.1	At3g46790.1	At3g60050.1	At4g14190.1	At4g30700.1
At2g38730.1	At3g06430.1	At3g18840.1	At3g46870.1	At3g60960.1	At4g14820.1	At4g30825.1
At2g39230.1	At3g06920.1	At3g18970.1	At3g47530.1	At3g60980.1	At4g14850.1	At4g31070.1
At2g39620.1	At3g07290.1	At3g20730.1	At3g47840.1	At3g61170.1	At4g15720.1	At4g31850.1
At2g40240.1	At3g09040.1	At3g21470.1	At3g48250.1	At3g61360.1	At4g16390.1	At4g32420.1
At2g40720.1	At3g09060.1	At3g22150.1	At3g48810.1	At3g61520.1	At4g16470.1	At4g32430.1
At2g41080.1	At3g09650.1	At3g22470.1	At3g49140.1	At3g62030.1	At4g16835.1	At4g32450.1

Table 7 (continued).

At4g33170.1
At4g33990.1
At4g34830.1
At4g34870.1
At4g34960.1
At4g35130.1
At4g35850.1
At4g36680.1
At4g37170.1
At4g37380.1
At4g38010.1
At4g38150.1
At4g38740.1
At4g39530.1
At4g39620.1
At4g39952.1
At5g01110.1
At5g02830.1
At5g02860.1
At5g03800.1
At5g04780.1
At5g04810.1
At5g06400.1
At5g06540.1
At5g08310.1
At5g08490.1
At5g08510.1
At5g09450.1
At5g09950.1
At5g10690.1
At5g11310.1
At5g12100.1
At5g13120.1
At5g13230.1
At5g13270.1
At5g13770.1
At5g14080.1
At5g14350.1
At5g14770.1
At5g14820.1
At5g15010.1
At5g15280.1
At5g15300.1
At5g15340.1
At5g15980.1
At5g16420.1
At5g16640.1
At5g16860.1
At5g17270.1
At5g18390.1
At5g18475.1
At5g18950.1
At5g19020.1
At5g21222.1
At5g24830.1
At5g25630.1
At5g27110.1
At5g27270.1
At5g27300.1
At5g27460.1
At5g28340.1
At5g28370.1
At5g28380.1
At5g28460.1
At5g37130.1
At5g37570.1
At5g38730.1
At5g39350.1
At5g39680.1
At5g39710.1
At5g39980.1
At5g40400.1
At5g40410.1
At5g41170.1
At5g42310.1
At5g42450.1
At5g43820.1
At5g44230.1
At5g45990.1
At5g46100.1
At5g46460.1
At5g46580.1
At5g46680.1
At5g47360.1
At5g47460.1
At5g48570.1
At5g48730.1
At5g48910.1
At5g50280.1
At5g50390.1
At5g50990.1
At5g52630.1
At5g52850.1
At5g55740.1
At5g55840.1
At5g56310.1
At5g57260.1
At5g58710.1
At5g59200.1
At5g59600.1
At5g59900.1
At5g60960.1
At5g61370.1
At5g61400.1
At5g61800.1
At5g61990.1
At5g62370.1
At5g64320.1
At5g65560.1
At5g65570.1
At5g65820.1
At5g66500.1
At5g66520.1
At5g67570.1

Table 8. The 1259 rice single domain cyclophilin candidates predicted by PLS_T-ACC.

LOC_Os01g01150.1
LOC_Os01g01150.2
LOC_Os01g01590.1
LOC_Os01g01680.1
LOC_Os01g01730.1
LOC_Os01g02110.1
LOC_Os01g03060.1
LOC_Os01g03060.2
LOC_Os01g03060.3
LOC_Os01g03860.1
LOC_Os01g04850.1
LOC_Os01g05420.1
LOC_Os01g05550.1
LOC_Os01g05570.1
LOC_Os01g06290.1
LOC_Os01g06290.2
LOC_Os01g06290.3
LOC_Os01g06620.1
LOC_Os01g09090.1
LOC_Os01g09270.1
LOC_Os01g09400.1
LOC_Os01g09480.1
LOC_Os01g10160.1
LOC_Os01g10170.1
LOC_Os01g10250.3
LOC_Os01g10560.1
LOC_Os01g11470.1
LOC_Os01g11640.1
LOC_Os01g12060.1
LOC_Os01g12374.1
LOC_Os01g12510.1
LOC_Os01g13780.1
LOC_Os01g13840.1
LOC_Os01g14710.1
LOC_Os01g14910.1
LOC_Os01g15270.1
LOC_Os01g15320.1
LOC_Os01g15370.1
LOC_Os01g17090.1
LOC_Os01g18210.1
LOC_Os01g18790.1
LOC_Os01g19010.1
LOC_Os01g19080.1
LOC_Os01g19340.1
LOC_Os01g19940.1
LOC_Os01g19970.1
LOC_Os01g20000.1
LOC_Os01g20710.1
LOC_Os01g20870.1
LOC_Os01g20894.1
LOC_Os01g21542.1
LOC_Os01g22490.1
LOC_Os01g22990.2
LOC_Os01g24150.1
LOC_Os01g24190.1
LOC_Os01g24230.1
LOC_Os01g24590.1
LOC_Os01g24700.1
LOC_Os01g25130.1
LOC_Os01g25910.1
LOC_Os01g26776.1
LOC_Os01g26816.1
LOC_Os01g26880.1
LOC_Os01g26900.1
LOC_Os01g26940.1
LOC_Os01g27070.1
LOC_Os01g27570.1
LOC_Os01g28500.2
LOC_Os01g29030.1
LOC_Os01g29060.1
LOC_Os01g29770.1
LOC_Os01g31330.1
LOC_Os01g31910.1
LOC_Os01g31950.1
LOC_Os01g32460.1
LOC_Os01g32540.1
LOC_Os01g33150.1
LOC_Os01g33560.1
LOC_Os01g34090.1
LOC_Os01g34110.1
LOC_Os01g34260.1
LOC_Os01g35129.1
LOC_Os01g35149.1
LOC_Os01g35530.1
LOC_Os01g35550.1
LOC_Os01g37280.4
LOC_Os01g37280.5
LOC_Os01g37660.1
LOC_Os01g38590.1
LOC_Os01g38860.1
LOC_Os01g39280.1
LOC_Os01g39510.1
LOC_Os01g41170.1
LOC_Os01g42130.1
LOC_Os01g42310.1
LOC_Os01g42670.1
LOC_Os01g43060.1
LOC_Os01g43630.3
LOC_Os01g43660.1
LOC_Os01g43880.1
LOC_Os01g44180.1
LOC_Os01g44290.1
LOC_Os01g45310.1
LOC_Os01g46460.1
LOC_Os01g47390.1
LOC_Os01g48780.1
LOC_Os01g49340.1
LOC_Os01g49710.1
LOC_Os01g50000.1
LOC_Os01g50670.1
LOC_Os01g50950.1
LOC_Os01g51930.1
LOC_Os01g52850.1
LOC_Os01g53310.1
LOC_Os01g53820.1
LOC_Os01g54750.1
LOC_Os01g55060.1
LOC_Os01g55840.1
LOC_Os01g56949.1
LOC_Os01g56969.1
LOC_Os01g58170.1
LOC_Os01g58310.1
LOC_Os01g58590.1
LOC_Os01g58930.1
LOC_Os01g59030.1
LOC_Os01g59060.1
LOC_Os01g59310.1
LOC_Os01g60460.1
LOC_Os01g60630.1
LOC_Os01g61750.1

Table 8 (continued).

LOC_Os01g62730.1
LOC_Os01g63444.1
LOC_Os01g63444.2
LOC_Os01g63500.1
LOC_Os01g63570.1
LOC_Os01g65440.1
LOC_Os01g65640.1
LOC_Os01g65692.1
LOC_Os01g65770.1
LOC_Os01g68179.1
LOC_Os01g71549.1
LOC_Os01g71599.1
LOC_Os01g71640.1
LOC_Os01g72890.1
LOC_Os01g72890.2
LOC_Os01g73090.1
LOC_Os01g73280.1
LOC_Os01g74100.1
LOC_Os01g74300.1
LOC_Os02g01120.1
LOC_Os02g01290.1
LOC_Os02g01370.1
LOC_Os02g01780.1
LOC_Os02g02870.1
LOC_Os02g02890.1
LOC_Os02g03250.1
LOC_Os02g03520.1
LOC_Os02g04220.1
LOC_Os02g04620.1
LOC_Os02g05290.1
LOC_Os02g06610.1
LOC_Os02g06640.1
LOC_Os02g06640.2
LOC_Os02g06780.1
LOC_Os02g06850.1
LOC_Os02g07320.1
LOC_Os02g08050.1
LOC_Os02g08430.1
LOC_Os02g10500.1
LOC_Os02g10970.1
LOC_Os02g10970.2
LOC_Os02g10970.3
LOC_Os02g11030.1
LOC_Os02g11030.2
LOC_Os02g11970.2
LOC_Os02g12080.1
LOC_Os02g12410.1
LOC_Os02g12610.1
LOC_Os02g13940.1
LOC_Os02g14970.1
LOC_Os02g15460.1
LOC_Os02g15704.1
LOC_Os02g15720.1
LOC_Os02g16230.1
LOC_Os02g16860.1
LOC_Os02g17020.1
LOC_Os02g18050.1
LOC_Os02g18300.1
LOC_Os02g18920.1
LOC_Os02g19980.1
LOC_Os02g20729.1
LOC_Os02g20880.1
LOC_Os02g20940.1
LOC_Os02g21860.1
LOC_Os02g21990.1
LOC_Os02g23829.1
LOC_Os02g23920.1
LOC_Os02g24460.1
LOC_Os02g25200.1
LOC_Os02g25240.1
LOC_Os02g26349.1
LOC_Os02g26349.2
LOC_Os02g27910.1
LOC_Os02g27990.1
LOC_Os02g28460.1
LOC_Os02g28570.1
LOC_Os02g29290.1
LOC_Os02g29400.1
LOC_Os02g29540.1
LOC_Os02g30330.1
LOC_Os02g30560.1
LOC_Os02g30760.1
LOC_Os02g31950.1
LOC_Os02g31990.1
LOC_Os02g32330.1
LOC_Os02g32630.1
LOC_Os02g33220.1
LOC_Os02g33670.1
LOC_Os02g34070.1
LOC_Os02g34090.1
LOC_Os02g34140.1
LOC_Os02g34290.1
LOC_Os02g34780.1
LOC_Os02g34820.1
LOC_Os02g35260.1
LOC_Os02g35650.1
LOC_Os02g35710.1
LOC_Os02g35810.1
LOC_Os02g36370.1
LOC_Os02g36910.1
LOC_Os02g37100.1
LOC_Os02g37110.1
LOC_Os02g37260.1
LOC_Os02g37320.1
LOC_Os02g37430.1
LOC_Os02g37550.1
LOC_Os02g37560.1
LOC_Os02g38150.1
LOC_Os02g38330.1
LOC_Os02g38490.1
LOC_Os02g38750.1
LOC_Os02g38990.1
LOC_Os02g39450.1
LOC_Os02g39720.1
LOC_Os02g39720.2
LOC_Os02g39720.3
LOC_Os02g39720.4
LOC_Os02g39720.5
LOC_Os02g39720.6
LOC_Os02g39720.7
LOC_Os02g41600.1
LOC_Os02g42610.1
LOC_Os02g44210.1
LOC_Os02g44690.1
LOC_Os02g44710.1
LOC_Os02g44950.1
LOC_Os02g45150.1
LOC_Os02g46020.1
LOC_Os02g46190.1
LOC_Os02g46800.1
LOC_Os02g46830.1
LOC_Os02g48020.1
LOC_Os02g48510.1
LOC_Os02g48530.1
LOC_Os02g49130.1

Table 8 (continued).

LOC_Os02g50080.1
LOC_Os02g50380.1
LOC_Os02g50470.1
LOC_Os02g50720.1
LOC_Os02g50770.1
LOC_Os02g50890.1
LOC_Os02g52360.1
LOC_Os02g52360.2
LOC_Os02g52620.1
LOC_Os02g53110.1
LOC_Os02g53110.2
LOC_Os02g53290.1
LOC_Os02g54770.1
LOC_Os02g54770.2
LOC_Os02g55360.1
LOC_Os02g55850.1
LOC_Os02g56020.1
LOC_Os02g56020.2
LOC_Os02g56030.1
LOC_Os02g56440.1
LOC_Os02g56590.1
LOC_Os02g57350.1
LOC_Os02g57550.1
LOC_Os02g58060.1
LOC_Os02g58139.3
LOC_Os03g01840.1
LOC_Os03g02070.1
LOC_Os03g03080.1
LOC_Os03g03110.1
LOC_Os03g03190.1
LOC_Os03g03810.1
LOC_Os03g04590.2
LOC_Os03g05090.1
LOC_Os03g05650.1
LOC_Os03g05830.1
LOC_Os03g05890.1
LOC_Os03g05990.1
LOC_Os03g06770.1
LOC_Os03g06780.1
LOC_Os03g06790.1
LOC_Os03g06800.1
LOC_Os03g06810.1
LOC_Os03g07290.1
LOC_Os03g07560.1
LOC_Os03g10334.1
LOC_Os03g10590.1
LOC_Os03g10670.1
LOC_Os03g10740.1
LOC_Os03g12170.1
LOC_Os03g12400.1
LOC_Os03g12750.1
LOC_Os03g12780.1
LOC_Os03g12879.1
LOC_Os03g13410.1
LOC_Os03g14070.1
LOC_Os03g14390.1
LOC_Os03g14440.1
LOC_Os03g14810.1
LOC_Os03g15150.1
LOC_Os03g15280.1
LOC_Os03g16369.1
LOC_Os03g16369.2
LOC_Os03g16660.1
LOC_Os03g17490.1
LOC_Os03g17490.2
LOC_Os03g17670.1
LOC_Os03g17870.1
LOC_Os03g18090.1
LOC_Os03g18770.1
LOC_Os03g19490.1
LOC_Os03g19640.1
LOC_Os03g19840.1
LOC_Os03g20220.1
LOC_Os03g21500.1
LOC_Os03g21570.1
LOC_Os03g21670.1
LOC_Os03g22240.1
LOC_Os03g24140.1
LOC_Os03g24560.1
LOC_Os03g25060.1
LOC_Os03g25090.1
LOC_Os03g25304.1
LOC_Os03g25314.1
LOC_Os03g25770.2
LOC_Os03g25869.1
LOC_Os03g26060.1
LOC_Os03g26280.1
LOC_Os03g26520.1
LOC_Os03g27030.3
LOC_Os03g27050.1
LOC_Os03g27160.1
LOC_Os03g28280.1
LOC_Os03g29000.1
LOC_Os03g29910.1
LOC_Os03g30600.1
LOC_Os03g30850.1
LOC_Os03g31500.1
LOC_Os03g31934.1
LOC_Os03g32240.1
LOC_Os03g32636.1
LOC_Os03g33384.1
LOC_Os03g33540.1
LOC_Os03g35420.1
LOC_Os03g37020.1
LOC_Os03g37440.1
LOC_Os03g37620.1
LOC_Os03g37940.1
LOC_Os03g38300.1
LOC_Os03g38440.1
LOC_Os03g39460.1
LOC_Os03g40030.1
LOC_Os03g40060.1
LOC_Os03g40380.1
LOC_Os03g40390.1
LOC_Os03g40490.1
LOC_Os03g41060.1
LOC_Os03g41260.1
LOC_Os03g41320.1
LOC_Os03g41339.1
LOC_Os03g41500.1
LOC_Os03g41910.1
LOC_Os03g42060.1
LOC_Os03g42590.1
LOC_Os03g42720.1
LOC_Os03g42870.1
LOC_Os03g43130.1
LOC_Os03g43540.1
LOC_Os03g44010.1
LOC_Os03g45350.1
LOC_Os03g45830.1
LOC_Os03g46440.3
LOC_Os03g48070.1
LOC_Os03g49810.1
LOC_Os03g50260.1
LOC_Os03g51170.1

Table 8 (continued).

LOC_Os03g52890.1
LOC_Os03g53340.4
LOC_Os03g54050.1
LOC_Os03g54050.2
LOC_Os03g54150.1
LOC_Os03g54150.2
LOC_Os03g55290.1
LOC_Os03g56290.1
LOC_Os03g57230.1
LOC_Os03g57770.1
LOC_Os03g58020.1
LOC_Os03g59700.1
LOC_Os03g60449.1
LOC_Os03g60500.1
LOC_Os03g60509.1
LOC_Os03g60600.1
LOC_Os03g60939.2
LOC_Os03g63530.1
LOC_Os03g63840.1
LOC_Os03g64340.1
LOC_Os04g02330.1
LOC_Os04g03510.1
LOC_Os04g03740.1
LOC_Os04g03910.1
LOC_Os04g03970.1
LOC_Os04g04380.1
LOC_Os04g05750.1
LOC_Os04g06460.1
LOC_Os04g07310.1
LOC_Os04g07610.1
LOC_Os04g07808.1
LOC_Os04g07910.1
LOC_Os04g08360.1
LOC_Os04g08610.1
LOC_Os04g09300.1
LOC_Os04g09790.1
LOC_Os04g09820.1
LOC_Os04g09870.1
LOC_Os04g10040.1
LOC_Os04g10150.1
LOC_Os04g10724.1
LOC_Os04g10830.1
LOC_Os04g11100.1
LOC_Os04g13940.1
LOC_Os04g14630.1
LOC_Os04g14970.1
LOC_Os04g15680.1
LOC_Os04g15780.1
LOC_Os04g16870.1
LOC_Os04g17000.1
LOC_Os04g17310.1
LOC_Os04g18559.1
LOC_Os04g19250.1
LOC_Os04g20000.1
LOC_Os04g20570.1
LOC_Os04g20600.1
LOC_Os04g21060.1
LOC_Os04g21190.1
LOC_Os04g21330.1
LOC_Os04g21470.1
LOC_Os04g21950.1
LOC_Os04g22040.1
LOC_Os04g22150.1
LOC_Os04g22250.1
LOC_Os04g24380.1
LOC_Os04g24620.1
LOC_Os04g24670.1
LOC_Os04g24770.1
LOC_Os04g25499.1
LOC_Os04g25740.1
LOC_Os04g25860.1
LOC_Os04g26150.1
LOC_Os04g26450.1
LOC_Os04g26930.1
LOC_Os04g28360.1
LOC_Os04g28530.1
LOC_Os04g28660.1
LOC_Os04g28710.1
LOC_Os04g29240.1
LOC_Os04g30070.1
LOC_Os04g30140.1
LOC_Os04g31100.1
LOC_Os04g31310.1
LOC_Os04g31460.1
LOC_Os04g32004.1
LOC_Os04g32240.1
LOC_Os04g32430.1
LOC_Os04g33430.1
LOC_Os04g33440.1
LOC_Os04g33440.2
LOC_Os04g33440.3
LOC_Os04g33610.1
LOC_Os04g34170.1
LOC_Os04g34170.2
LOC_Os04g34170.3
LOC_Os04g34690.1
LOC_Os04g34830.1
LOC_Os04g34910.1
LOC_Os04g34940.1
LOC_Os04g35550.1
LOC_Os04g36570.1
LOC_Os04g36570.2
LOC_Os04g36650.1
LOC_Os04g39890.1
LOC_Os04g40270.1
LOC_Os04g41010.1
LOC_Os04g41060.1
LOC_Os04g41720.1
LOC_Os04g41790.1
LOC_Os04g41930.1
LOC_Os04g42170.1
LOC_Os04g42560.1
LOC_Os04g43960.1
LOC_Os04g44120.1
LOC_Os04g44170.1
LOC_Os04g44590.1
LOC_Os04g44590.2
LOC_Os04g44590.3
LOC_Os04g44800.1
LOC_Os04g45130.1
LOC_Os04g45300.1
LOC_Os04g45350.1
LOC_Os04g45870.1
LOC_Os04g46040.1
LOC_Os04g46430.1
LOC_Os04g48440.1
LOC_Os04g48720.1
LOC_Os04g49240.1
LOC_Os04g50080.1
LOC_Os04g51960.1
LOC_Os04g53620.1
LOC_Os04g53620.2
LOC_Os04g53950.1
LOC_Os04g54030.1
LOC_Os04g54720.1

Table 8 (continued).

LOC_Os04g55950.1	LOC_Os05g07840.1	LOC_Os05g24090.1	LOC_Os05g33000.2	LOC_Os05g46480.2
LOC_Os04g56370.1	LOC_Os05g08340.1	LOC_Os05g24180.1	LOC_Os05g33070.1	LOC_Os05g46670.1
LOC_Os04g56610.1	LOC_Os05g08470.1	LOC_Os05g24390.1	LOC_Os05g35080.1	LOC_Os05g46770.1
LOC_Os04g56720.1	LOC_Os05g09400.2	LOC_Os05g24870.1	LOC_Os05g35380.1	LOC_Os05g46770.2
LOC_Os04g56960.1	LOC_Os05g09400.3	LOC_Os05g25050.1	LOC_Os05g35900.1	LOC_Os05g46780.1
LOC_Os04g57230.3	LOC_Os05g09420.1	LOC_Os05g25440.1	LOC_Os05g36340.1	LOC_Os05g49940.1
LOC_Os04g57250.1	LOC_Os05g09580.1	LOC_Os05g25690.1	LOC_Os05g37420.1	LOC_Os05g50110.1
LOC_Os04g57260.1	LOC_Os05g10590.2	LOC_Os05g26150.1	LOC_Os05g37490.1	LOC_Os05g50150.1
LOC_Os04g57280.1	LOC_Os05g11250.1	LOC_Os05g26670.1	LOC_Os05g38300.1	LOC_Os05g50240.1
LOC_Os04g58770.1	LOC_Os05g11490.1	LOC_Os05g27000.1	LOC_Os05g38310.1	LOC_Os05g51380.1
LOC_Os05g01270.1	LOC_Os05g11832.1	LOC_Os05g27180.1	LOC_Os05g38860.1	LOC_Os05g51380.2
LOC_Os05g01540.1	LOC_Os05g12160.1	LOC_Os05g27200.1	LOC_Os05g38870.1	LOC_Os05g51900.1
LOC_Os05g01540.2	LOC_Os05g12280.1	LOC_Os05g27250.1	LOC_Os05g38910.1	LOC_Os06g01600.1
LOC_Os05g02070.3	LOC_Os05g12500.1	LOC_Os05g28020.1	LOC_Os05g40400.1	LOC_Os06g02010.1
LOC_Os05g02620.1	LOC_Os05g12670.1	LOC_Os05g28450.1	LOC_Os05g41820.1	LOC_Os06g02040.1
LOC_Os05g03250.1	LOC_Os05g13760.1	LOC_Os05g28590.1	LOC_Os05g42170.1	LOC_Os06g02690.1
LOC_Os05g04150.2	LOC_Os05g13870.1	LOC_Os05g29170.1	LOC_Os05g42424.1	LOC_Os06g02970.1
LOC_Os05g04380.2	LOC_Os05g14140.1	LOC_Os05g29730.1	LOC_Os05g42436.1	LOC_Os06g03350.1
LOC_Os05g04760.1	LOC_Os05g16380.1	LOC_Os05g29829.1	LOC_Os05g43700.1	LOC_Os06g03370.1
LOC_Os05g05820.1	LOC_Os05g16610.1	LOC_Os05g30000.1	LOC_Os05g43720.1	LOC_Os06g04000.1
LOC_Os05g06220.1	LOC_Os05g18640.1	LOC_Os05g30650.1	LOC_Os05g44120.1	LOC_Os06g05450.2
LOC_Os05g06520.1	LOC_Os05g19230.1	LOC_Os05g30780.1	LOC_Os05g44640.1	LOC_Os06g07170.1
LOC_Os05g06770.1	LOC_Os05g20650.1	LOC_Os05g30880.1	LOC_Os05g44720.1	LOC_Os06g07270.1
LOC_Os05g07200.1	LOC_Os05g21060.1	LOC_Os05g31000.1	LOC_Os05g44950.1	LOC_Os06g07410.1
LOC_Os05g07240.1	LOC_Os05g22780.1	LOC_Os05g31580.1	LOC_Os05g45940.1	LOC_Os06g07580.1
LOC_Os05g07360.1	LOC_Os05g23240.1	LOC_Os05g32160.1	LOC_Os05g46340.2	LOC_Os06g08780.1
LOC_Os05g07610.1	LOC_Os05g23630.1	LOC_Os05g33000.1	LOC_Os05g46470.1	LOC_Os06g08980.1

Table 8 (continued).

LOC_Os06g09010.1
LOC_Os06g09380.1
LOC_Os06g09410.1
LOC_Os06g10730.1
LOC_Os06g11320.1
LOC_Os06g11320.2
LOC_Os06g11750.1
LOC_Os06g13610.1
LOC_Os06g14210.1
LOC_Os06g14500.1
LOC_Os06g14660.1
LOC_Os06g14690.1
LOC_Os06g14870.1
LOC_Os06g14930.1
LOC_Os06g15240.1
LOC_Os06g16020.1
LOC_Os06g16060.1
LOC_Os06g16610.1
LOC_Os06g16750.1
LOC_Os06g17760.1
LOC_Os06g18020.1
LOC_Os06g20530.1
LOC_Os06g21100.1
LOC_Os06g21120.1
LOC_Os06g21540.1
LOC_Os06g21580.1
LOC_Os06g22120.1
LOC_Os06g22400.1
LOC_Os06g22410.1
LOC_Os06g22540.1
LOC_Os06g22700.1
LOC_Os06g22950.1
LOC_Os06g23850.1
LOC_Os06g24020.1
LOC_Os06g24390.1
LOC_Os06g24650.1
LOC_Os06g24870.1
LOC_Os06g24870.2
LOC_Os06g25090.1
LOC_Os06g25620.1
LOC_Os06g27730.1
LOC_Os06g27960.1
LOC_Os06g28110.1
LOC_Os06g28220.1
LOC_Os06g28610.1
LOC_Os06g28700.1
LOC_Os06g29880.1
LOC_Os06g29940.1
LOC_Os06g30280.1
LOC_Os06g30530.1
LOC_Os06g31110.1
LOC_Os06g31556.1
LOC_Os06g31940.1
LOC_Os06g32310.1
LOC_Os06g32430.1
LOC_Os06g32640.1
LOC_Os06g33380.1
LOC_Os06g34180.1
LOC_Os06g34190.1
LOC_Os06g34630.1
LOC_Os06g34900.1
LOC_Os06g35230.1
LOC_Os06g35500.1
LOC_Os06g35790.1
LOC_Os06g35890.1
LOC_Os06g36240.1
LOC_Os06g36250.1
LOC_Os06g36380.1
LOC_Os06g36600.1
LOC_Os06g36690.1
LOC_Os06g36810.1
LOC_Os06g36940.1
LOC_Os06g37320.1
LOC_Os06g37490.1
LOC_Os06g38170.1
LOC_Os06g39410.1
LOC_Os06g39560.1
LOC_Os06g39912.1
LOC_Os06g40610.1
LOC_Os06g41290.1
LOC_Os06g42990.1
LOC_Os06g44390.1
LOC_Os06g44870.1
LOC_Os06g45610.1
LOC_Os06g45900.1
LOC_Os06g45910.1
LOC_Os06g46090.1
LOC_Os06g46230.1
LOC_Os06g46300.1
LOC_Os06g46470.1
LOC_Os06g46770.1
LOC_Os06g47500.1
LOC_Os06g48190.1
LOC_Os06g48500.1
LOC_Os06g48550.1
LOC_Os06g49330.1
LOC_Os06g49470.1
LOC_Os06g49480.1
LOC_Os06g49480.2
LOC_Os06g49480.3
LOC_Os06g49590.1
LOC_Os06g49610.1
LOC_Os06g49710.1
LOC_Os06g50890.2
LOC_Os06g50890.3
LOC_Os06g50890.4
LOC_Os06g50890.5
LOC_Os06g51140.1
LOC_Os06g51320.1
LOC_Os06g51320.2
LOC_Os07g01350.1
LOC_Os07g01470.1
LOC_Os07g01570.1
LOC_Os07g01720.1
LOC_Os07g03540.1
LOC_Os07g03640.1
LOC_Os07g05970.1
LOC_Os07g06240.1
LOC_Os07g06480.1
LOC_Os07g07040.1
LOC_Os07g07290.1
LOC_Os07g07370.1
LOC_Os07g08190.1
LOC_Os07g08250.1
LOC_Os07g08380.1
LOC_Os07g09360.1
LOC_Os07g10930.1
LOC_Os07g11290.1
LOC_Os07g11450.1
LOC_Os07g12430.1
LOC_Os07g13440.1
LOC_Os07g14200.1
LOC_Os07g14480.1
LOC_Os07g14910.1
LOC_Os07g14930.1

Table 8 (continued).

LOC_Os07g15020.1
LOC_Os07g15530.1
LOC_Os07g16310.1
LOC_Os07g17070.1
LOC_Os07g17410.1
LOC_Os07g17560.1
LOC_Os07g17940.1
LOC_Os07g18830.1
LOC_Os07g18980.1
LOC_Os07g20110.1
LOC_Os07g20180.1
LOC_Os07g20410.1
LOC_Os07g22140.1
LOC_Os07g22830.1
LOC_Os07g23640.1
LOC_Os07g23660.1
LOC_Os07g23780.1
LOC_Os07g23910.1
LOC_Os07g23970.1
LOC_Os07g24170.1
LOC_Os07g24200.1
LOC_Os07g24710.1
LOC_Os07g24730.1
LOC_Os07g24930.1
LOC_Os07g25140.1
LOC_Os07g25250.1
LOC_Os07g25960.1
LOC_Os07g26240.1
LOC_Os07g26740.1
LOC_Os07g26790.1
LOC_Os07g27050.1
LOC_Os07g28100.1
LOC_Os07g28500.1
LOC_Os07g29200.1
LOC_Os07g29290.1
LOC_Os07g29510.1
LOC_Os07g29580.1
LOC_Os07g29594.1
LOC_Os07g29970.1
LOC_Os07g30420.1
LOC_Os07g30640.1
LOC_Os07g32360.1
LOC_Os07g33180.1
LOC_Os07g33860.1
LOC_Os07g33870.1
LOC_Os07g33880.1
LOC_Os07g33898.1
LOC_Os07g33921.1
LOC_Os07g33943.1
LOC_Os07g33979.1
LOC_Os07g33997.1
LOC_Os07g34560.1
LOC_Os07g35090.1
LOC_Os07g35120.1
LOC_Os07g35550.1
LOC_Os07g35820.1
LOC_Os07g36580.1
LOC_Os07g36620.1
LOC_Os07g37020.1
LOC_Os07g37330.1
LOC_Os07g38300.4
LOC_Os07g38550.1
LOC_Os07g39450.1
LOC_Os07g39550.1
LOC_Os07g40510.1
LOC_Os07g41220.1
LOC_Os07g41840.1
LOC_Os07g41960.1
LOC_Os07g41990.1
LOC_Os07g42940.7
LOC_Os07g43050.1
LOC_Os07g43050.2
LOC_Os07g43316.1
LOC_Os07g43440.1
LOC_Os07g44010.1
LOC_Os07g44600.1
LOC_Os07g49130.1
LOC_Os08g01440.1
LOC_Os08g01690.1
LOC_Os08g02180.1
LOC_Os08g02660.1
LOC_Os08g02800.1
LOC_Os08g03250.1
LOC_Os08g05020.1
LOC_Os08g05090.1
LOC_Os08g05100.1
LOC_Os08g05160.1
LOC_Os08g05960.1
LOC_Os08g06210.1
LOC_Os08g06290.1
LOC_Os08g07150.1
LOC_Os08g07510.1
LOC_Os08g08090.1
LOC_Os08g09820.1
LOC_Os08g10620.1
LOC_Os08g12150.1
LOC_Os08g13200.1
LOC_Os08g14040.1
LOC_Os08g14410.1
LOC_Os08g15430.1
LOC_Os08g15610.1
LOC_Os08g15620.1
LOC_Os08g16300.1
LOC_Os08g16820.1
LOC_Os08g17800.1
LOC_Os08g19060.1
LOC_Os08g19110.1
LOC_Os08g19300.1
LOC_Os08g20820.1
LOC_Os08g21180.1
LOC_Os08g23350.1
LOC_Os08g23500.1
LOC_Os08g23754.1
LOC_Os08g24740.1
LOC_Os08g25040.1
LOC_Os08g25680.1
LOC_Os08g26210.1
LOC_Os08g26370.1
LOC_Os08g26580.1
LOC_Os08g27210.1
LOC_Os08g27880.1
LOC_Os08g27960.1
LOC_Os08g28120.1
LOC_Os08g28300.1
LOC_Os08g29440.1
LOC_Os08g29600.1
LOC_Os08g29650.1
LOC_Os08g30020.7
LOC_Os08g30430.1
LOC_Os08g32110.1
LOC_Os08g32490.1
LOC_Os08g32980.1
LOC_Os08g32980.2
LOC_Os08g33154.1
LOC_Os08g33290.1

Table 8 (continued).

LOC_Os08g34200.1
LOC_Os08g35070.1
LOC_Os08g35790.1
LOC_Os08g36179.1
LOC_Os08g36206.1
LOC_Os08g36470.1
LOC_Os08g36800.1
LOC_Os08g37410.1
LOC_Os08g37960.1
LOC_Os08g37960.2
LOC_Os08g38230.1
LOC_Os08g38300.2
LOC_Os08g38900.2
LOC_Os08g39480.1
LOC_Os08g39990.1
LOC_Os08g40220.1
LOC_Os08g41360.1
LOC_Os08g42800.1
LOC_Os08g42820.1
LOC_Os08g43620.1
LOC_Os08g44030.1
LOC_Os08g44520.1
LOC_Os08g44520.2
LOC_Os08g44520.3
LOC_Os09g01160.1
LOC_Os09g01630.1
LOC_Os09g01640.1
LOC_Os09g01730.1
LOC_Os09g01860.1
LOC_Os09g02450.1
LOC_Os09g02610.1
LOC_Os09g02770.1
LOC_Os09g03050.1
LOC_Os09g03120.1
LOC_Os09g03480.1
LOC_Os09g04100.1
LOC_Os09g04250.1
LOC_Os09g04390.1
LOC_Os09g06570.1
LOC_Os09g07030.1
LOC_Os09g07240.1
LOC_Os09g07650.1
LOC_Os09g08210.1
LOC_Os09g08330.1
LOC_Os09g09010.1
LOC_Os09g09030.1
LOC_Os09g09350.1
LOC_Os09g10010.1
LOC_Os09g10210.1
LOC_Os09g10860.1
LOC_Os09g11700.1
LOC_Os09g11830.1
LOC_Os09g12200.1
LOC_Os09g12450.1
LOC_Os09g12640.1
LOC_Os09g12790.1
LOC_Os09g13080.1
LOC_Os09g13130.1
LOC_Os09g13720.1
LOC_Os09g15150.1
LOC_Os09g15639.1
LOC_Os09g15690.1
LOC_Os09g16130.1
LOC_Os09g16270.1
LOC_Os09g16500.1
LOC_Os09g16610.1
LOC_Os09g17440.1
LOC_Os09g17640.1
LOC_Os09g17690.1
LOC_Os09g19580.1
LOC_Os09g19780.1
LOC_Os09g19840.1
LOC_Os09g20100.1
LOC_Os09g20920.1
LOC_Os09g20960.1
LOC_Os09g21460.1
LOC_Os09g23330.1
LOC_Os09g23450.1
LOC_Os09g23610.1
LOC_Os09g23630.1
LOC_Os09g24170.1
LOC_Os09g24550.1
LOC_Os09g25190.1
LOC_Os09g25700.1
LOC_Os09g25830.1
LOC_Os09g26740.1
LOC_Os09g27170.1
LOC_Os09g27730.1
LOC_Os09g28270.1
LOC_Os09g28380.1
LOC_Os09g29270.1
LOC_Os09g29780.1
LOC_Os09g29780.2
LOC_Os09g29850.1
LOC_Os09g29980.1
LOC_Os09g29980.2
LOC_Os09g30340.1
LOC_Os09g30498.1
LOC_Os09g30502.1
LOC_Os09g31019.4
LOC_Os09g32880.1
LOC_Os09g34310.2
LOC_Os09g34340.1
LOC_Os09g35560.1
LOC_Os09g36670.1
LOC_Os09g37290.1
LOC_Os09g37560.1
LOC_Os09g38500.1
LOC_Os09g38500.2
LOC_Os09g38680.1
LOC_Os09g38880.1
LOC_Os09g39780.1
LOC_Os09g39780.2
LOC_Os10g01740.1
LOC_Os10g02930.1
LOC_Os10g02930.2
LOC_Os10g03370.1
LOC_Os10g05150.1
LOC_Os10g06630.1
LOC_Os10g07520.1
LOC_Os10g08630.1
LOC_Os10g08950.1
LOC_Os10g09250.1
LOC_Os10g10310.1
LOC_Os10g10794.1
LOC_Os10g10849.1
LOC_Os10g11512.1
LOC_Os10g12320.1
LOC_Os10g12490.1
LOC_Os10g15000.1
LOC_Os10g15120.1
LOC_Os10g17810.1
LOC_Os10g18750.1
LOC_Os10g18920.1
LOC_Os10g19250.1

Table 8 (continued).

LOC_Os10g19930.1
LOC_Os10g20480.3
LOC_Os10g20590.1
LOC_Os10g20870.1
LOC_Os10g21440.1
LOC_Os10g22184.1
LOC_Os10g22356.1
LOC_Os10g22440.1
LOC_Os10g22750.1
LOC_Os10g24870.1
LOC_Os10g25160.1
LOC_Os10g25570.1
LOC_Os10g25570.3
LOC_Os10g26680.1
LOC_Os10g27000.1
LOC_Os10g28320.3
LOC_Os10g28420.1
LOC_Os10g28914.1
LOC_Os10g29060.1
LOC_Os10g29580.1
LOC_Os10g30450.1
LOC_Os10g32100.1
LOC_Os10g32509.1
LOC_Os10g32940.1
LOC_Os10g33954.1
LOC_Os10g34570.1
LOC_Os10g34840.1
LOC_Os10g34990.1
LOC_Os10g36200.1
LOC_Os10g36200.2
LOC_Os10g36200.3
LOC_Os10g36340.1
LOC_Os10g37459.1
LOC_Os10g37590.1
LOC_Os10g37830.2
LOC_Os10g38100.1
LOC_Os10g38150.1
LOC_Os10g38360.1
LOC_Os10g38870.1
LOC_Os10g39210.1
LOC_Os10g39610.1
LOC_Os10g39720.1
LOC_Os10g41670.1
LOC_Os10g41980.1
LOC_Os10g42020.2
LOC_Os10g42448.1
LOC_Os10g42520.1
LOC_Os11g01060.1
LOC_Os11g01230.1
LOC_Os11g01350.1
LOC_Os11g01650.1
LOC_Os11g01990.1
LOC_Os11g02330.1
LOC_Os11g02330.2
LOC_Os11g03660.1
LOC_Os11g04560.1
LOC_Os11g05180.3
LOC_Os11g05300.1
LOC_Os11g05500.1
LOC_Os11g05810.1
LOC_Os11g06030.1
LOC_Os11g07310.1
LOC_Os11g07830.1
LOC_Os11g08250.1
LOC_Os11g09100.1
LOC_Os11g09260.1
LOC_Os11g09270.1
LOC_Os11g09800.1
LOC_Os11g09870.1
LOC_Os11g10560.1
LOC_Os11g10960.1
LOC_Os11g11030.1
LOC_Os11g11080.1
LOC_Os11g11090.1
LOC_Os11g11740.1
LOC_Os11g12610.1
LOC_Os11g13580.1
LOC_Os11g13720.1
LOC_Os11g13830.1
LOC_Os11g14060.1
LOC_Os11g14120.1
LOC_Os11g14190.1
LOC_Os11g14190.2
LOC_Os11g14310.1
LOC_Os11g14500.1
LOC_Os11g15610.1
LOC_Os11g16340.1
LOC_Os11g16740.1
LOC_Os11g16780.1
LOC_Os11g16800.1
LOC_Os11g16890.1
LOC_Os11g17040.1
LOC_Os11g17450.1
LOC_Os11g17570.1
LOC_Os11g18020.1
LOC_Os11g18760.1
LOC_Os11g18940.1
LOC_Os11g19040.1
LOC_Os11g19170.1
LOC_Os11g19280.1
LOC_Os11g19680.1
LOC_Os11g20530.1
LOC_Os11g20600.1
LOC_Os11g22200.1
LOC_Os11g23080.2
LOC_Os11g24020.1
LOC_Os11g24080.1
LOC_Os11g24900.1
LOC_Os11g25540.1
LOC_Os11g26190.1
LOC_Os11g26780.1
LOC_Os11g26790.1
LOC_Os11g27640.1
LOC_Os11g27680.1
LOC_Os11g28010.1
LOC_Os11g28160.1
LOC_Os11g28420.1
LOC_Os11g28450.1
LOC_Os11g29810.1
LOC_Os11g29930.1
LOC_Os11g30100.1
LOC_Os11g31140.1
LOC_Os11g32310.1
LOC_Os11g32400.1
LOC_Os11g32740.1
LOC_Os11g34150.1
LOC_Os11g34150.2
LOC_Os11g34150.3
LOC_Os11g34560.1
LOC_Os11g34580.1
LOC_Os11g34870.1
LOC_Os11g35720.1
LOC_Os11g36620.1
LOC_Os11g36820.1
LOC_Os11g38280.1

Table 8 (continued).

LOC_Os11g38470.1
LOC_Os11g38510.1
LOC_Os11g38990.1
LOC_Os11g38990.2
LOC_Os11g38990.3
LOC_Os11g39079.1
LOC_Os11g39300.1
LOC_Os11g40660.1
LOC_Os11g41060.1
LOC_Os11g41090.1
LOC_Os11g41820.1
LOC_Os11g41820.2
LOC_Os11g42740.1
LOC_Os11g43570.1
LOC_Os11g44600.1
LOC_Os11g45170.1
LOC_Os11g45220.1
LOC_Os11g45270.1
LOC_Os11g45590.1
LOC_Os11g45630.1
LOC_Os11g45890.1
LOC_Os11g45960.1
LOC_Os11g47494.1
LOC_Os11g47830.1
LOC_Os11g47960.1
LOC_Os12g01050.1
LOC_Os12g02140.1
LOC_Os12g02480.1
LOC_Os12g02920.1
LOC_Os12g02990.1
LOC_Os12g04250.1
LOC_Os12g04360.1
LOC_Os12g04680.1
LOC_Os12g05070.1
LOC_Os12g07000.1
LOC_Os12g07400.1
LOC_Os12g07430.1
LOC_Os12g08250.1
LOC_Os12g08320.1
LOC_Os12g09050.1
LOC_Os12g09310.1
LOC_Os12g09940.1
LOC_Os12g10240.1
LOC_Os12g10750.1
LOC_Os12g11830.1
LOC_Os12g12080.3
LOC_Os12g12330.1
LOC_Os12g12640.1
LOC_Os12g13190.1
LOC_Os12g13350.1
LOC_Os12g13480.1
LOC_Os12g13500.1
LOC_Os12g13740.1
LOC_Os12g15930.1
LOC_Os12g16510.1
LOC_Os12g17730.1
LOC_Os12g18620.1
LOC_Os12g18700.1
LOC_Os12g19100.1
LOC_Os12g21750.1
LOC_Os12g22640.1
LOC_Os12g22790.1
LOC_Os12g23280.1
LOC_Os12g23280.2
LOC_Os12g23869.1
LOC_Os12g24090.1
LOC_Os12g25480.1
LOC_Os12g25950.1
LOC_Os12g26050.1
LOC_Os12g26930.1
LOC_Os12g27190.1
LOC_Os12g27670.1
LOC_Os12g29460.1
LOC_Os12g29510.1
LOC_Os12g30170.1
LOC_Os12g30460.1
LOC_Os12g30560.1
LOC_Os12g31310.1
LOC_Os12g31330.1
LOC_Os12g31710.1
LOC_Os12g31770.1
LOC_Os12g32300.1
LOC_Os12g32434.1
LOC_Os12g32890.1
LOC_Os12g33180.1
LOC_Os12g34190.1
LOC_Os12g34620.1
LOC_Os12g34660.1
LOC_Os12g34790.1
LOC_Os12g35390.1
LOC_Os12g35510.1
LOC_Os12g35670.1
LOC_Os12g35860.1
LOC_Os12g35980.1
LOC_Os12g36030.1
LOC_Os12g36030.2
LOC_Os12g36030.3
LOC_Os12g36060.1
LOC_Os12g36070.1
LOC_Os12g36740.1
LOC_Os12g38064.1
LOC_Os12g38064.2
LOC_Os12g38064.3
LOC_Os12g38190.1
LOC_Os12g38250.1
LOC_Os12g38290.1
LOC_Os12g38700.1
LOC_Os12g38990.1
LOC_Os12g40619.1
LOC_Os12g40840.1
LOC_Os12g40840.2
LOC_Os12g40840.3
LOC_Os12g41130.1
LOC_Os12g41634.1
LOC_Os12g41962.1
LOC_Os12g41962.2
LOC_Os12g42780.1
LOC_Os12g43400.1
LOC_Os12g43780.1
LOC_Os12g44040.1
LOC_Os12g44110.1

Table 9. The 48 rice single domain cyclophilin candidates predicted by SAM.

LOC_Os01g02080.1
LOC_Os01g18210.1
LOC_Os01g40050.1
LOC_Os01g40050.2
LOC_Os02g02090.1
LOC_Os02g02890.1
LOC_Os02g10970.1
LOC_Os02g10970.2
LOC_Os02g10970.3
LOC_Os02g14570.1
LOC_Os02g52360.1
LOC_Os02g52360.2
LOC_Os03g01090.1
LOC_Os03g10400.1
LOC_Os03g10400.2
LOC_Os03g10400.3
LOC_Os03g59700.1
LOC_Os04g22750.1
LOC_Os05g01270.1
LOC_Os05g05420.1
LOC_Os06g04000.1
LOC_Os06g11320.1
LOC_Os06g11320.2
LOC_Os06g18130.1
LOC_Os06g18140.1
LOC_Os06g20770.1
LOC_Os06g45900.1
LOC_Os06g45910.1
LOC_Os06g49470.1
LOC_Os06g49480.1
LOC_Os06g49480.2
LOC_Os06g49480.3
LOC_Os06g50910.1
LOC_Os07g07960.1
LOC_Os07g08190.1
LOC_Os07g10230.1
LOC_Os07g10240.1
LOC_Os07g29390.1
LOC_Os07g29390.2
LOC_Os07g37830.1
LOC_Os07g37830.2
LOC_Os08g02340.1
LOC_Os08g09270.1
LOC_Os08g19610.1
LOC_Os08g19610.2
LOC_Os08g19610.3
LOC_Os08g19610.4
LOC_Os08g29370.1
LOC_Os08g29370.2
LOC_Os08g44330.1
LOC_Os08g44330.2
LOC_Os08g44520.1
LOC_Os08g44520.2
LOC_Os08g44520.3
LOC_Os09g03970.1
LOC_Os09g36670.1
LOC_Os09g39780.1
LOC_Os09g39780.2
LOC_Os10g06630.1
LOC_Os10g06640.1
LOC_Os10g35090.1
LOC_Os10g35230.1
LOC_Os10g35240.1
LOC_Os10g35436.1
LOC_Os11g38860.1
LOC_Os11g38990.1
LOC_Os11g38990.2
LOC_Os11g38990.3
LOC_Os11g46820.1

Table 10. The 29 rice single domain cyclophilin candidates predicted by PSI-BLAST.

LOC_Os01g02080.1
LOC_Os01g18210.1
LOC_Os01g40050.1
LOC_Os01g40050.2
LOC_Os02g02090.1
LOC_Os02g02890.1
LOC_Os02g10970.1
LOC_Os02g10970.2
LOC_Os02g10970.3
LOC_Os02g52360.1
LOC_Os02g52360.2
LOC_Os03g01090.1
LOC_Os03g10400.1
LOC_Os03g10400.2
LOC_Os03g10400.3
LOC_Os03g59700.1
LOC_Os05g01270.1
LOC_Os06g04000.1
LOC_Os06g11320.1
LOC_Os06g11320.2
LOC_Os06g45900.1
LOC_Os06g45910.1
LOC_Os06g49470.1
LOC_Os06g49480.1
LOC_Os06g49480.2
LOC_Os06g49480.3
LOC_Os07g08190.1
LOC_Os07g29390.1
LOC_Os07g29390.2
LOC_Os07g37830.1
LOC_Os07g37830.2
LOC_Os08g19610.1
LOC_Os08g19610.2
LOC_Os08g19610.3
LOC_Os08g19610.4
LOC_Os08g29370.1
LOC_Os08g29370.2
LOC_Os08g44330.1
LOC_Os08g44330.2
LOC_Os08g44520.1
LOC_Os08g44520.2
LOC_Os08g44520.3
LOC_Os09g36670.1
LOC_Os09g39780.1
LOC_Os09g39780.2
LOC_Os10g06630.1
LOC_Os10g06640.1
LOC_Os11g38990.1
LOC_Os11g38990.2
LOC_Os11g38990.3

Table 11. The 304 rice multiple domain cyclophilin candidates predicted by PLS_T-ACC.

LOC_Os01g01540.1
LOC_Os01g02130.1
LOC_Os01g03040.1
LOC_Os01g03250.1
LOC_Os01g10250.3
LOC_Os01g10400.2
LOC_Os01g16550.1
LOC_Os01g18210.1
LOC_Os01g18980.1
LOC_Os01g20870.1
LOC_Os01g21260.1
LOC_Os01g21680.1
LOC_Os01g21980.1
LOC_Os01g23180.1
LOC_Os01g23340.1
LOC_Os01g24720.1
LOC_Os01g27620.1
LOC_Os01g28160.1
LOC_Os01g31450.1
LOC_Os01g35950.1
LOC_Os01g43660.1
LOC_Os01g44370.1
LOC_Os01g44380.1
LOC_Os01g45410.1
LOC_Os01g45700.1
LOC_Os01g48710.1
LOC_Os01g51930.1
LOC_Os01g58140.1
LOC_Os01g58370.1
LOC_Os01g61750.1
LOC_Os01g62244.3
LOC_Os01g66400.1
LOC_Os01g72440.1
LOC_Os02g01770.1
LOC_Os02g02890.1
LOC_Os02g03830.1
LOC_Os02g04220.1
LOC_Os02g06640.1
LOC_Os02g06640.2
LOC_Os02g07550.1
LOC_Os02g10970.1
LOC_Os02g10970.2
LOC_Os02g10970.3
LOC_Os02g12330.1
LOC_Os02g12410.1
LOC_Os02g13530.1
LOC_Os02g13700.1
LOC_Os02g17120.1
LOC_Os02g18940.1
LOC_Os02g19240.1
LOC_Os02g19960.1
LOC_Os02g20710.1
LOC_Os02g21990.1
LOC_Os02g25240.1
LOC_Os02g34400.1
LOC_Os02g35260.1
LOC_Os02g39030.1
LOC_Os02g39030.2
LOC_Os02g40800.1
LOC_Os02g42120.1
LOC_Os02g45030.1
LOC_Os02g47600.1
LOC_Os02g50390.1
LOC_Os02g50400.1
LOC_Os02g52360.1
LOC_Os02g52360.2
LOC_Os02g53370.1
LOC_Os02g56190.1
LOC_Os03g03080.1
LOC_Os03g03190.1
LOC_Os03g05650.1
LOC_Os03g06800.1
LOC_Os03g07660.1
LOC_Os03g09050.1
LOC_Os03g11640.1
LOC_Os03g13660.1
LOC_Os03g13950.1
LOC_Os03g19890.1
LOC_Os03g21200.1
LOC_Os03g21930.1
LOC_Os03g23000.1
LOC_Os03g23020.1
LOC_Os03g24210.1
LOC_Os03g27050.1
LOC_Os03g31500.1
LOC_Os03g40380.1
LOC_Os03g40490.1
LOC_Os03g40800.1
LOC_Os03g42170.1
LOC_Os03g43480.1
LOC_Os03g44040.1
LOC_Os03g45860.1
LOC_Os03g56350.1
LOC_Os03g56790.1
LOC_Os03g57010.1
LOC_Os03g59700.1
LOC_Os03g60590.1
LOC_Os03g63170.1
LOC_Os03g64060.1
LOC_Os04g04380.1
LOC_Os04g05600.1
LOC_Os04g05750.1
LOC_Os04g08792.1
LOC_Os04g10040.1
LOC_Os04g10150.1
LOC_Os04g13610.1
LOC_Os04g16480.1
LOC_Os04g16900.1
LOC_Os04g19220.1
LOC_Os04g20550.1
LOC_Os04g27010.1
LOC_Os04g27530.1
LOC_Os04g27990.1
LOC_Os04g30140.1
LOC_Os04g31250.1
LOC_Os04g35670.1
LOC_Os04g40160.1
LOC_Os04g41040.1
LOC_Os04g42444.1
LOC_Os04g42560.1
LOC_Os04g46430.1
LOC_Os04g47650.1
LOC_Os04g49080.1
LOC_Os04g53620.1
LOC_Os04g53620.2
LOC_Os04g54740.2
LOC_Os04g56600.1
LOC_Os05g01270.1
LOC_Os05g04140.1
LOC_Os05g04800.1
LOC_Os05g06019.1
LOC_Os05g06028.1
LOC_Os05g07480.1
LOC_Os05g08390.1
LOC_Os05g10590.1

Table 11 (continued).

LOC_Os05g10820.1
LOC_Os05g12474.2
LOC_Os05g13870.1
LOC_Os05g13900.1
LOC_Os05g13940.1
LOC_Os05g18130.1
LOC_Os05g20540.1
LOC_Os05g23100.1
LOC_Os05g23630.1
LOC_Os05g23760.1
LOC_Os05g25440.1
LOC_Os05g25510.1
LOC_Os05g25974.1
LOC_Os05g25974.2
LOC_Os05g28020.1
LOC_Os05g30230.1
LOC_Os05g30780.1
LOC_Os05g34060.1
LOC_Os05g35380.1
LOC_Os05g42170.1
LOC_Os05g42424.1
LOC_Os05g43174.1
LOC_Os05g44570.2
LOC_Os05g46540.1
LOC_Os05g50150.1
LOC_Os06g10840.1
LOC_Os06g11320.1
LOC_Os06g11320.2
LOC_Os06g12650.1
LOC_Os06g16650.1
LOC_Os06g16750.1
LOC_Os06g19790.1
LOC_Os06g20690.1
LOC_Os06g22950.1
LOC_Os06g31360.1
LOC_Os06g32570.1
LOC_Os06g40900.1
LOC_Os06g41290.1
LOC_Os06g41530.1
LOC_Os06g41550.1
LOC_Os06g41920.1
LOC_Os06g46090.1
LOC_Os06g46650.1
LOC_Os06g46754.1
LOC_Os06g46770.1
LOC_Os06g47690.1
LOC_Os06g49470.1
LOC_Os06g49480.1
LOC_Os06g49480.2
LOC_Os06g49480.3
LOC_Os06g51220.6
LOC_Os07g01910.1
LOC_Os07g03180.1
LOC_Os07g04460.1
LOC_Os07g05920.1
LOC_Os07g06480.1
LOC_Os07g10280.1
LOC_Os07g14880.1
LOC_Os07g14910.1
LOC_Os07g16310.1
LOC_Os07g17170.1
LOC_Os07g17690.1
LOC_Os07g20790.1
LOC_Os07g25660.1
LOC_Os07g28990.1
LOC_Os07g29720.1
LOC_Os07g30540.1
LOC_Os07g31570.1
LOC_Os07g32000.1
LOC_Os07g34560.1
LOC_Os07g35550.1
LOC_Os07g35990.1
LOC_Os07g36020.1
LOC_Os07g36040.1
LOC_Os07g38470.1
LOC_Os07g39200.1
LOC_Os07g39540.1
LOC_Os07g39550.1
LOC_Os07g40070.1
LOC_Os07g48390.1
LOC_Os08g01440.1
LOC_Os08g02080.2
LOC_Os08g05100.1
LOC_Os08g05160.1
LOC_Os08g14040.1
LOC_Os08g15430.1
LOC_Os08g17540.1
LOC_Os08g17800.1
LOC_Os08g19020.1
LOC_Os08g21720.1
LOC_Os08g23900.1
LOC_Os08g30030.1
LOC_Os08g30230.1
LOC_Os08g32490.1
LOC_Os08g33960.1
LOC_Os08g34940.1
LOC_Os08g38420.1
LOC_Os08g38500.1
LOC_Os08g38540.1
LOC_Os08g42360.1
LOC_Os08g44520.1
LOC_Os08g44520.3
LOC_Os09g01860.1
LOC_Os09g03540.1
LOC_Os09g10860.1
LOC_Os09g12260.1
LOC_Os09g13720.1
LOC_Os09g15400.1
LOC_Os09g15690.1
LOC_Os09g17050.1
LOC_Os09g20900.1
LOC_Os09g20940.1
LOC_Os09g23430.1
LOC_Os09g27530.1
LOC_Os09g27730.1
LOC_Os09g29860.1
LOC_Os09g31260.1
LOC_Os09g33464.1
LOC_Os09g36670.1
LOC_Os09g39780.1
LOC_Os09g39780.2
LOC_Os10g01710.1
LOC_Os10g02020.1
LOC_Os10g02680.1
LOC_Os10g06630.1
LOC_Os10g07390.1
LOC_Os10g16990.1
LOC_Os10g20820.1
LOC_Os10g27000.1
LOC_Os10g33886.1
LOC_Os10g37780.1
LOC_Os10g39044.1
LOC_Os11g01350.1
LOC_Os11g01650.1
LOC_Os11g04340.1

Table 11 (continued).

LOC_Os11g04880.1
LOC_Os11g06250.1
LOC_Os11g06530.1
LOC_Os11g07980.2
LOC_Os11g09150.1
LOC_Os11g11740.1
LOC_Os11g13580.1
LOC_Os11g14730.1
LOC_Os11g16740.1
LOC_Os11g17100.1
LOC_Os11g19270.1
LOC_Os11g19530.1
LOC_Os11g20070.1
LOC_Os11g20500.1
LOC_Os11g22380.1
LOC_Os11g24290.1
LOC_Os11g27720.1
LOC_Os11g31150.1
LOC_Os11g32400.1
LOC_Os11g34070.1
LOC_Os11g34150.4
LOC_Os11g39830.1
LOC_Os11g42740.1
LOC_Os11g43520.1
LOC_Os11g43570.1
LOC_Os11g44980.1
LOC_Os12g10240.1
LOC_Os12g11830.1
LOC_Os12g12830.1
LOC_Os12g14350.1
LOC_Os12g15490.1
LOC_Os12g15840.1
LOC_Os12g19820.1
LOC_Os12g24280.1
LOC_Os12g26680.1
LOC_Os12g28390.1
LOC_Os12g29240.1
LOC_Os12g29460.1
LOC_Os12g29510.1
LOC_Os12g30960.1
LOC_Os12g32700.1
LOC_Os12g38250.1
LOC_Os12g43340.1
LOC_Os12g43400.1

Table 12. The 122 rice multiple domain cyclophilin candidates predicted by SAM.

LOC_Os01g02080.1
LOC_Os01g07640.1
LOC_Os01g07980.1
LOC_Os01g09384.1
LOC_Os01g11920.1
LOC_Os01g16680.1
LOC_Os01g18210.1
LOC_Os01g38180.1
LOC_Os01g38229.1
LOC_Os01g38229.2
LOC_Os01g40050.1
LOC_Os01g40050.2
LOC_Os01g42960.1
LOC_Os01g43540.1
LOC_Os01g66690.1
LOC_Os01g68680.1
LOC_Os01g74580.1
LOC_Os01g74580.2
LOC_Os02g01030.1
LOC_Os02g01030.2
LOC_Os02g01960.1
LOC_Os02g02090.1
LOC_Os02g02890.1
LOC_Os02g10180.1
LOC_Os02g10180.2
LOC_Os02g10970.1
LOC_Os02g10970.2
LOC_Os02g10970.3
LOC_Os02g28830.1
LOC_Os02g28980.1
LOC_Os02g29030.1
LOC_Os02g29100.1
LOC_Os02g29130.1
LOC_Os02g29130.2
LOC_Os02g29140.1
LOC_Os02g29150.1
LOC_Os02g29150.2
LOC_Os02g29160.1
LOC_Os02g29190.1
LOC_Os02g34540.1
LOC_Os02g34830.1
LOC_Os02g37470.1
LOC_Os02g43020.1
LOC_Os02g47880.1
LOC_Os02g51730.1
LOC_Os02g51810.1
LOC_Os02g52360.1
LOC_Os02g52360.2
LOC_Os02g54280.1
LOC_Os03g04050.1
LOC_Os03g10400.1
LOC_Os03g10400.2
LOC_Os03g10400.3
LOC_Os03g10980.1
LOC_Os03g15740.1
LOC_Os03g22140.1
LOC_Os03g25140.1
LOC_Os03g42350.1
LOC_Os03g47620.1
LOC_Os03g47640.1
LOC_Os03g47650.1
LOC_Os03g47670.1
LOC_Os03g47686.1
LOC_Os03g47702.1
LOC_Os03g47720.1
LOC_Os03g50010.1
LOC_Os03g50010.2
LOC_Os03g53910.1
LOC_Os03g59700.1
LOC_Os03g61010.1
LOC_Os04g28420.1
LOC_Os04g35690.1
LOC_Os04g45480.1
LOC_Os04g45480.2
LOC_Os04g52580.1
LOC_Os04g55230.1
LOC_Os04g57780.1
LOC_Os04g58750.1
LOC_Os04g58750.2
LOC_Os05g01270.1
LOC_Os05g01310.1
LOC_Os05g01310.2
LOC_Os05g01310.3
LOC_Os05g03910.1
LOC_Os05g11550.1
LOC_Os05g11990.1
LOC_Os05g11990.2
LOC_Os05g50990.1
LOC_Os06g04000.1
LOC_Os06g06450.1
LOC_Os06g06470.1
LOC_Os06g06490.1
LOC_Os06g06760.1
LOC_Os06g06760.2
LOC_Os06g11320.1
LOC_Os06g11320.2
LOC_Os06g33250.1
LOC_Os06g41750.1
LOC_Os06g45900.1
LOC_Os06g45910.1
LOC_Os06g49470.1
LOC_Os06g49480.1
LOC_Os06g49480.2
LOC_Os06g49480.3
LOC_Os07g02300.1
LOC_Os07g02300.2
LOC_Os07g06710.1
LOC_Os07g07540.2
LOC_Os07g08190.1
LOC_Os07g29390.1
LOC_Os07g29390.2
LOC_Os08g02140.1
LOC_Os08g02140.2
LOC_Os08g13640.1
LOC_Os08g19610.1
LOC_Os08g19610.2
LOC_Os08g19610.3
LOC_Os08g19610.4
LOC_Os08g20020.1
LOC_Os08g20020.2
LOC_Os08g23580.1
LOC_Os08g29370.1
LOC_Os08g29370.2
LOC_Os08g41390.1
LOC_Os08g41390.2
LOC_Os08g44330.1
LOC_Os08g44330.2
LOC_Os08g44510.1
LOC_Os08g44520.1
LOC_Os08g44520.2
LOC_Os08g44520.3
LOC_Os09g03630.1
LOC_Os09g03680.1
LOC_Os09g03750.1
LOC_Os09g03890.1

Table 12 (continued).

LOC_Os09g04990.1
LOC_Os09g15670.1
LOC_Os09g23650.1
LOC_Os09g36670.1
LOC_Os09g38390.1
LOC_Os09g38690.1
LOC_Os09g39780.1
LOC_Os09g39780.2
LOC_Os10g06630.1
LOC_Os10g06640.1
LOC_Os10g32300.1
LOC_Os10g32300.2
LOC_Os10g34540.1
LOC_Os10g36250.1
LOC_Os10g36250.2
LOC_Os10g39670.1
LOC_Os11g05090.1
LOC_Os11g38990.1
LOC_Os11g38990.2
LOC_Os11g38990.3
LOC_Os12g05090.1
LOC_Os12g40770.1
LOC_Os12g40780.1
LOC_Os12g41190.1
LOC_Os12g43840.1
LOC_Os12g43940.1

Table 13. The 96 rice multiple domain cyclophilin candidates predicted by PSI-BLAST.

LOC_Os01g02080.1
LOC_Os01g07640.1
LOC_Os01g07980.1
LOC_Os01g11920.1
LOC_Os01g18210.1
LOC_Os01g40050.1
LOC_Os01g40050.2
LOC_Os01g42960.1
LOC_Os01g43540.1
LOC_Os01g74580.1
LOC_Os01g74580.2
LOC_Os02g01030.1
LOC_Os02g01030.2
LOC_Os02g01960.1
LOC_Os02g02090.1
LOC_Os02g02890.1
LOC_Os02g10180.1
LOC_Os02g10180.2
LOC_Os02g10970.1
LOC_Os02g10970.2
LOC_Os02g10970.3
LOC_Os02g28980.1
LOC_Os02g29030.1
LOC_Os02g29100.1
LOC_Os02g29130.1
LOC_Os02g29130.2
LOC_Os02g29150.1
LOC_Os02g29150.2
LOC_Os02g29160.1
LOC_Os02g29190.1
LOC_Os02g43020.1
LOC_Os02g47880.1
LOC_Os02g51730.1
LOC_Os02g51810.1
LOC_Os02g52360.1
LOC_Os02g52360.2
LOC_Os03g10400.1
LOC_Os03g10400.2
LOC_Os03g10400.3
LOC_Os03g10980.1
LOC_Os03g25140.1
LOC_Os03g42350.1
LOC_Os03g47620.1
LOC_Os03g47640.1
LOC_Os03g47650.1
LOC_Os03g47670.1
LOC_Os03g47686.1
LOC_Os03g47702.1
LOC_Os03g47720.1
LOC_Os03g50010.1
LOC_Os03g50010.2
LOC_Os03g59700.1
LOC_Os04g28420.1
LOC_Os04g35690.1
LOC_Os04g45480.1
LOC_Os04g45480.2
LOC_Os04g52580.1
LOC_Os05g01270.1
LOC_Os05g01310.1
LOC_Os05g01310.2
LOC_Os05g01310.3
LOC_Os05g03910.1
LOC_Os05g11550.1
LOC_Os05g11990.1
LOC_Os05g11990.2
LOC_Os05g31062.1
LOC_Os05g34820.1
LOC_Os05g50990.1
LOC_Os06g04000.1
LOC_Os06g06470.1
LOC_Os06g06760.1
LOC_Os06g06760.2
LOC_Os06g11320.1
LOC_Os06g11320.2
LOC_Os06g45900.1
LOC_Os06g45910.1
LOC_Os06g49470.1
LOC_Os06g49480.1
LOC_Os06g49480.2
LOC_Os06g49480.3
LOC_Os07g02300.1
LOC_Os07g02300.2
LOC_Os07g06710.1
LOC_Os07g08190.1
LOC_Os07g29360.1
LOC_Os07g29390.1
LOC_Os07g29390.2
LOC_Os08g02140.1
LOC_Os08g02140.2
LOC_Os08g13640.1
LOC_Os08g19610.1
LOC_Os08g19610.2
LOC_Os08g19610.3
LOC_Os08g19610.4
LOC_Os08g20020.1
LOC_Os08g20020.2
LOC_Os08g23580.1
LOC_Os08g41390.1
LOC_Os08g41390.2
LOC_Os08g44330.1
LOC_Os08g44330.2
LOC_Os08g44520.1
LOC_Os08g44520.2
LOC_Os08g44520.3
LOC_Os09g03630.1
LOC_Os09g03680.1
LOC_Os09g03750.1
LOC_Os09g03890.1
LOC_Os09g04990.1
LOC_Os09g23650.1
LOC_Os09g36670.1
LOC_Os09g38390.1
LOC_Os09g39780.1
LOC_Os09g39780.2
LOC_Os10g06630.1
LOC_Os10g06640.1
LOC_Os10g34540.1
LOC_Os10g36250.1
LOC_Os11g05090.1
LOC_Os11g05090.1
LOC_Os11g38990.1
LOC_Os11g38990.1
LOC_Os11g38990.2
LOC_Os11g38990.2
LOC_Os11g38990.3
LOC_Os11g38990.3
LOC_Os12g05090.1
LOC_Os12g40770.1
LOC_Os12g40780.1
LOC_Os12g41190.1
LOC_Os12g43840.1
LOC_Os12g43940.1

Table 14. Comparison of 20 amino acid compositions between GPCRs and non-GPCRs.

	Amino acids									
	Lys	Asn	Thr	Arg	Ser	Ile	Met	His	Pro	Leu
GPCRs[a]	0.0357	0.0375	0.0602	0.0455	0.0802	0.0699	0.0307	0.0229	0.0443	0.1267
Non-GPCRs[a]	0.0525	0.0539	0.0661	0.0600	0.0251	0.0224	0.0462	0.0966	0.0654	0.0522
Absolute mean difference	0.0258	0.0036	0.0077	0.0084	0.0141	0.0099	0.0056	0.0006	0.0018	0.0301
P-values[b]	<0.0010*	<0.0010	<0.0010	<0.0010	<0.0010*	<0.0010	<0.0010	0.4102	0.1554	<0.0010*
Wilcoxon's U-statistics	53842.0	115216.0	12458.0	98590.5	176901.5	453679.5	76549.0	131021.5	125328.0	54367.5
P-values[c]	<0.0010	0.0320	<0.0010	<0.0010	<0.0010	<0.0010	<0.0010	0.1873	0.9427	<0.0010

[a] The mean frequencies of each amino acid in 500 GPCRs and 500 non-GPCRs.

[b] P-values were obtained from the T-test.

*Chosen amino acids.

[c] P-values were obtained from the Wilcoxon's rank-sum test.

Table 14 (continued).

	Amino acids									
	Glu	Asp	Ala	Gly	Val	Gln	Tyr	Cys	Trp	Phe
GPCRs[a]	0.0325	0.0280	0.0729	0.0507	0.0798	0.0282	0.0391	0.0330	0.0173	0.0648
Non-GPCRs[a]	0.0806	0.0695	0.0694	0.0394	0.0305	0.0155	0.0117	0.0405	0.0117	0.0405
Mean difference	0.0329	0.0242	0.0077	0.0188	0.0105	0.0112	0.0087	0.0175	0.0056	0.0242
P-values	<0.0010*	<0.0010*	<0.0010	<0.0010*	<0.0010*	<0.0010*	<0.0010	<0.0010*	<0.0010	<0.0010*
Wilcoxon's U-statistics	61441.0	2393.5	13876.0	9734.5	55534.0	29657.5	54326.0	66451.0	90031.5	2015.0
P-values[c]	<0.0010	<0.0010	<0.0010	<0.0010	<0.0010	<0.0010	<0.0010	<0.0010	<0.0010	<0.0010

Table 15. Frequency distributions of GPCRs and non-GPCRs based on the compositions of the ten chosen amino acids.

	Asp		Cys		Glu		Phe		Gly	
Amino acid composition (x)	GPCR	non-GPCR	GPCR	non-GPCR	GPCR	non-GPCR	GPCR	non-GPCR	GPCR	non-GPCR
x = 0	0.00	2.00	0.00	2.00	0.00	0.10	0.00	0.80	0.00	1.00
0 < x <0.01	2.60	3.30	0.00	40.00	0.90	0.30	0.00	1.00	0.00	1.00
0.01 ≤ x < 0.02	17.40	2.40	18.00	30.10	18.20	2.80	0.00	8.20	1.00	1.00
0.02 ≤ x < 0.03	42.40	6.20	28.20	17.60	30.70	4.60	2.00	19.60	2.50	3.40
0.03 ≤ x < 0.04	22.40	12.80	38.90	5.40	23.70	7.40	3.50	24.20	19.90	7.40
0.04 ≤ x < 0.05	8.30	19.80	10.30	1.60	14.60	10.80	15.20	19.60	28.80	11.20
0.05 ≤ x < 0.06	3.30	23.60	3.00	0.20	7.30	15.40	25.20	12.40	21.10	14.20
0.06 ≤ x < 0.07	2.30	14.40	1.60	1.00	2.30	18.00	19.60	6.80	15.90	15.00
0.07 ≤ x < 0.08	1.30	7.20	0.00	1.00	2.30	14.80	17.30	3.00	6.20	15.20
0.08 ≤ x < 0.09	0.00	3.40	0.00	0.20	0.00	9.80	12.90	1.40	2.90	9.80
0.09 ≤ x < 0.1	0.00	1.20	0.00	0.20	0.00	5.00	4.30	0.60	1.30	8.80
0.1 ≤ x < 0.11	0.00	1.00	0.00	0.20	0.00	4.80	0.00	1.00	0.00	5.00
0.11 ≤ x < 0.12	0.00	0.40	0.00	0.40	0.00	4.20	0.00	0.40	0.00	6.80
0.120 ≤ x	0.00	2.30	0.00	0.10	0.00	2.00	0.00	1.00	0.00	0.20

Table 15 (continued).

	Leu		Lys		Gln		Ser		Val	
Amino acid composition (x)	GPCR	non-GPCR	GPCR	non-GPCR	GPCR	non-GPCR	GPCR	non-GPCR	GPCR	non-GPCR
x = 0	0.00	0.00	0.00	0.50	0.00	1.20	0.00	0.00	0.00	0.20
0 < x <0.01	0.00	0.00	1.40	2.50	2.00	0.80	0.00	0.20	0.00	0.20
0.01 ≤ x < 0.02	0.00	0.00	11.85	4.20	21.60	9.00	0.00	0.40	0.00	0.60
0.02 ≤ x < 0.03	0.00	0.40	21.60	6.80	35.60	21.00	0.20	3.00	0.00	1.80
0.03 ≤ x < 0.04	0.00	0.20	28.10	9.40	28.60	27.40	0.40	10.20	0.00	5.20
0.04 ≤ x < 0.05	0.20	2.40	20.30	13.90	10.80	17.60	3.20	13.80	1.00	10.60
0.05 ≤ x < 0.06	0.60	6.00	11.10	13.40	0.80	10.40	8.20	18.20	6.60	15.20
0.06 ≤ x < 0.07	0.40	8.60	3.60	13.00	0.60	6.00	15.20	16.60	16.80	22.00
0.07 ≤ x < 0.08	2.60	12.00	1.20	10.20	0.00	2.80	23.00	12.80	28.20	15.00
0.08 ≤ x < 0.09	3.80	15.60	0.85	8.20	0.00	1.20	23.00	9.40	23.20	11.20
0.09 ≤ x < 0.1	9.60	12.60	0.00	6.00	0.00	1.00	14.60	5.60	15.40	9.20
0.1 ≤ x < 0.11	9.40	13.00	0.00	3.20	0.00	0.20	8.20	4.40	6.60	3.80
0.11 ≤ x < 0.12	13.40	10.40	0.00	6.70	0.00	0.40	3.00	2.00	2.00	2.60
0.120 ≤ x	60.00	18.80	0.00	2.00	0.00	1.00	1.00	3.40	0.20	

Table 16. Comparison of the amino acid compositions between the immunoglobulin and other proteins.

	Amino acids									
	Lys	Asn	Thr	Arg	Ser	Ile	Met	His	Pro	Leu
Immunoglobulin[a]	0.0623	0.0446	0.0728	0.0407	0.0823	0.0462	0.0137	0.0331	0.0590	0.0821
Non-immunoglobulin[a]	0.0645	0.0358	0.0562	0.0526	0.0594	0.0556	0.0184	0.0219	0.0444	0.0815
Absolute mean difference	0.0022	0.0112	0.0166	0.0119	0.0229	0.0094	0.0047	0.0112	0.0146	0.0006
P-values [b]	0.6761	0.0020*	<0.0010*	0.0022*	<0.0010*	0.0139	0.0116	0.0001*	<0.0010*	0.8831
Wilcoxon's U-statistics	4034.5	5514.0	5614.0	2989.5	5637.0	4332.5	4216.5	4094.0	5523.5	4048.0
P-values[c]	0.7666	0.00108	<0.0010	0.0024	<0.0010	0.3729	0.07611	<0.0010	<0.0010	0.7834

[a] The mean frequencies of each amino acid in 90 immunoglobulin and 90 non-immunoglobulin.

[b] P-values were obtained from the T-test.

*Chosen amino acids.

[c] P-values were obtained from the Wilcoxon's rank-sum test.

Table 16 (continued).

	Glu	Asp	Ala	Gly	Val	Gln	Tyr	Cys	Trp	Phe
Immunoglobulin	0.0673	0.0510	0.0523	0.0647	0.0811	0.0451	0.0344	0.0247	0.0179	0.0336
Non-immunoglobulin	0.0672	0.0600	0.0824	0.0698	0.0719	0.0403	0.0324	0.0167	0.0131	0.0351
Absolute mean difference	0.0001	0.0110	0.0301	0.0051	0.0092	0.0048	0.0020	0.0080	0.0048	0.0015
	0.9890	<0.0010	<0.0010*	0.0202	0.0206	0.1567	0.4543	0.0205	0.0145	0.5473
Wilcoxon’s U-statistics	4121.5	4104.0	2049.0	3258.0	4849.0	4527.5	4402.5	5854.5	5120.0	3775.0
P-values[c]	0.5844	<0.0010	<0.0010	0.0811	0.04521	0.13461	0.1298	0.0771	0.0118	0.144

Table 17. Frequency distributions of immunoglobulin and non-immunoglobulin proteins based on the compositions of the chosen amino acids.

	Ala		His		Asn		Pro	
Amino acid composition (x)	Imm[a]	Non-Imm[b]	Imm[a]	Non-Imm[b]	Imm[a]	Non-Imm[b]	Imm[a]	Non-Imm[b]
x = 0	0.00	4.00	12.00	6.00	4.00	10.00	0.00	4.00
0 < x <0.01	0.00	0.00	10.50	0.00	0.00	0.00	4.50	0.00
0.01 ≤ x < 0.02	4.00	0.00	30.50	27.20	4.00	15.20	5.00	0.00
0.02 ≤ x < 0.03	20.00	0.00	20.30	25.00	10.90	16.80	0.00	12.00
0.03 ≤ x < 0.04	16.00	8.00	13.90	16.80	21.10	21.90	10.00	18.70
0.04 ≤ x < 0.05	16.00	6.80	12.80	26.00	24.00	16.10	12.50	24.00
0.05 ≤ x < 0.06	8.30	10.00	0.00	11.20	11.10	6.00	22.00	22.80
0.06 ≤ x < 0.07	9.40	12.20	0.00	7.10	10.90	8.00	12.00	6.50
0.07 ≤ x < 0.08	9.30	6.00	0.00	0.00	6.00	6.00	10.00	6.00
0.08 ≤ x < 0.09	4.00	21.00	0.00	6.00	4.00	0.00	12.00	6.00
0.09 ≤ x < 0.1	8.00	10.00	0.00	0.00	4.00	0.00	6.00	0.00
0.1 ≤ x < 0.11	0.00	4.00	0.00	0.00	0.00	0.00	6.00	0.00
0.11 ≤ x < 0.12	4.00	10.00	0.00	0.00	0.00	0.00	0.00	0.00
0.120 ≤ x	0.00	8.00	0.00	0.00	0.00	0.00	0.00	0.00

Imm[a] = Immunoglobulin; Non-Imm[b] = Non-immunoglobulin.

Table 17 (continued).

	Arg		Ser		Thr	
Amino acid composition (x)	Imm[a]	Non-Imm[b]	Imm[a]	Non-Imm[b]	Imm[a]	Non-Imm[b]
x = 0	0.00	0.00	0.00	0.00	0.00	4.00
0 < x <0.01	6.00	0.00	0.00	0.00	0.00	0.00
0.01 ≤ x < 0.02	0.00	0.00	0.00	6.50	0.00	4.00
0.02 ≤ x < 0.03	0.40	11.50	6.80	9.20	6.00	10.00
0.03 ≤ x < 0.04	19.00	12.50	11.80	18.00	10.00	14.00
0.04 ≤ x < 0.05	21.10	28.00	8.90	16.00	16.50	20.00
0.05 ≤ x < 0.06	7.90	10.00	16.00	26.00	15.50	14.00
0.06 ≤ x < 0.07	16.00	13.50	15.90	11.20	14.00	10.00
0.07 ≤ x < 0.08	10.00	0.00	7.20	7.10	14.00	16.00
0.08 ≤ x < 0.09	8.00	8.60	11.20	0.00	4.00	8.00
0.09 ≤ x < 0.1	0.00	5.90	9.50	6.00	11.00	0.00
0.1 ≤ x < 0.11	8.00	0.00	12.70	0.00	9.00	0.00
0.11 ≤ x < 0.12	0.00	0.00	0.00	0.00	0.00	0.00
0.120 ≤ x	0.00	10.00	0.00	0.00	0.00	0.00

Imm[a] = Immunoglobulin; Non-Imm[b] = Non-immunoglobulin.

Table 18. Statistical significance of 12 physico-chemical properties of amino acids between immunoglobulin and other proteins[a].

	Mass	S_area	Volume	H_phob	H_phil	Refra	Ip	TFen	NP_surf	Alph	Beta	Turn
P-values[b]	<0.001	0.011	0.384	0.559	0.870	0.025	0.304	0.501	0.423	<0.001	<0.001	0.105
P-values[c]	<0.001	0.001	0.087	0.112	0.045	0.018	0.781	0.254	0.721	<0.001	<0.001	0.115

[a] S_area = Surface area; H_phob = Hydrophobicity; H_phil =Hydrophilicity; Refra = Refractivity; Ip = Isoelectric point; TFen = Transfer free energy from water to ethanol; NP_surf = Non-polar surface; Alph = Frequency of alpha-helix with weight; Beta = Frequency of beta-sheet with weight Turn = Frequency of reverse turn with weight.

[b] P-value from t-test.

[c] P-value from rank test.

Table 19. Classifier performance on the GPCR test dataset.

Classifiers	% Accuracy	%False positive	%False negative	MCC[a]
ST-method	98.4	2.0	0.4	0.96
PLS-ACC	95.2	6.5	0.6	0.90
PLS-AA	93.4	8.5	0.8	0.89
SAM	94.8	0.4	15.0	0.88
PSI-BLAST	92.5	0.3	23.0	0.83

[a]Mathews correlation coefficient.

Table 20. Cross-validation test results of the classifiers on the immunoglobulin dataset.

Classifiers	%Accuracy	%False positive	%False negative	MCC[a]
ST-method	93.3	8.0	4.4	0.87
PLS-ACC	90.0	15.6	4.6	0.81
PLS-AA	87.0	18.3	4.8	0.78
SAM	72.2	5.5	50.0	0.52
PSI-BLAST	71.1	5.5	52.2	0.50

[a] Mathews correlation coefficient.

Table 21. The 579 *Arabidopsis* proteins with 5-10 transmembarane regions predicted by ST-method as 7TMR candidates. The 250 proteins that overlapped with proteins form Moriyama et al.[3] study are marked in bold with shaded background.

At1g01070.1	**At1g11000.1**	**At1g18470.2**	**At1g28760.1**	At1g47670.1	At1g62280.1	At1g71900.1
At1g01070.2	At1g11200.1	At1g19800.1	**At1g29330.1**	**At1g48230.1**	At1g62430.1	At1g71940.1
At1g01580.1	**At1g11310.1**	At1g19800.2	**At1g29390.1**	**At1g48270.1**	At1g63050.1	**At1g71960.1**
At1g01590.1	At1g11450.1	At1g19800.3	**At1g29390.2**	**At1g48460.1**	**At1g63110.1**	At1g72590.1
At1g01620.1	**At1g11460.1**	At1g19970.1	At1g29395.1	**At1g48640.1**	**At1g63110.2**	At1g73240.1
At1g01650.1	**At1g11540.1**	**At1g20050.1**	**At1g30840.1**	At1g49960.2	**At1g63110.3**	At1g74440.1
At1g02190.1	**At1g11880.1**	At1g20925.1	At1g31260.1	At1g50430.1	**At1g63120.1**	**At1g75000.1**
At1g02190.2	**At1g12450.1**	At1g21070.1	**At1g31300.1**	**At1g50630.1**	At1g64990.1	**At1g75470.1**
At1g02260.1	**At1g12480.1**	**At1g21460.1**	At1g31770.1	At1g51460.1	**At1g66760.1**	At1g75500.1
At1g03070.1	**At1g12500.1**	At1g21790.1	At1g31885.1	At1g51500.1	At1g66770.1	At1g75760.1
At1g04220.1	At1g12600.1	At1g21870.1	At1g32120.1	At1g52580.1	At1g67060.1	At1g76520.1
At1g05360.1	At1g12640.1	At1g21890.1	At1g34020.1	**At1g52615.1**	At1g67570.1	At1g76520.2
At1g06080.1	**At1g12730.1**	**At1g23020.1**	At1g34050.1	At1g52750.1	At1g67640.1	**At1g76530.1**
At1g06120.1	At1g12730.2	At1g23480.1	At1g34470.1	At1g53270.1	At1g67960.1	At1g76670.1
At1g06360.1	**At1g12750.1**	At1g23480.2	At1g34490.1	At1g53390.1	**At1g68000.1**	At1g77110.1
At1g06890.1	At1g13560.1	**At1g23830.1**	At1g34500.1	At1g54730.3	**At1g68170.1**	At1g77610.1
At1g08230.1	At1g13560.2	**At1g23850.1**	At1g34520.1	**At1g55130.1**	At1g68530.1	At1g77690.1
At1g08960.1	**At1g14530.1**	At1g24070.1	At1g34630.1	At1g55230.1	At1g68530.2	At1g78000.1
At1g09090.1	**At1g14530.2**	**At1g24400.1**	At1g35180.1	**At1g55240.1**	At1g68740.1	At1g78000.2
At1g09090.2	**At1g15110.1**	At1g25270.1	At1g42470.1	**At1g57680.1**	**At1g68820.1**	At1g78560.1
At1g09380.1	**At1g15960.1**	At1g25450.1	**At1g42560.1**	**At1g57680.2**	**At1g69420.1**	At1g79975.1
At1g09860.1	At1g16040.1	At1g25500.1	**At1g43580.1**	**At1g57943.1**	**At1g69420.2**	At1g79975.2
At1g10090.1	**At1g16560.1**	**At1g26180.1**	**At1g44010.1**	**At1g57980.1**	**At1g69430.1**	At1g80310.1
At1g10660.1	**At1g16560.2**	**At1g26650.1**	At1g44750.1	**At1g57990.1**	**At1g69450.1**	At1g80760.1
At1g10660.2	**At1g16560.3**	**At1g26700.1**	**At1g44750.2**	At1g58520.1	**At1g70260.1**	**At2g01070.1**
At1g10660.3	At1g16900.1	**At1g26730.1**	**At1g44960.1**	At1g60600.1	At1g70900.1	At2g01180.1

Table 21 (continued).

At1g10950.1	At1g18420.1	At1g28230.1	At1g47640.1	At1g61800.1	**At1g71680.1**	At2g02180.1
At2g02810.1	At2g20650.1	At2g29650.2	At2g37450.1	**At3g01550.1**	**At3g16090.1**	**At3g28050.1**
At2g03330.1	At2g20650.2	**At2g29980.1**	At2g37700.1	At3g01760.1	**At3g16690.1**	**At3g28060.1**
At2g03520.1	**At2g20725.1**	**At2g29980.2**	**At2g37940.1**	At3g03305.1	**At3g17430.1**	At3g28070.2
At2g03530.1	At2g21050.1	At2g30080.1	**At2g38120.1**	**At3g03700.1**	At3g17690.1	At3g28130.1
At2g03590.1	**At2g21080.1**	At2g30890.1	**At2g38170.2**	**At3g04970.1**	At3g17700.1	At3g28180.1
At2g03600.1	At2g21190.1	At2g31360.1	At2g39060.1	**At3g04970.2**	At3g18200.1	At3g29060.1
At2g04360.1	At2g21340.1	**At2g31440.1**	**At2g39200.1**	At3g05010.1	**At3g18215.1**	**At3g30340.1**
At2g04850.1	At2g21340.2	At2g31530.1	At2g39510.1	At3g05280.1	**At3g19260.1**	**At3g45290.1**
At2g05755.1	At2g22830.1	At2g32270.1	At2g41050.1	At3g05940.1	At3g20100.1	At3g45810.1
At2g07727.1	**At2g23680.1**	**At2g33205.1**	At2g41560.1	At3g06460.1	At3g20240.1	At3g45870.1
At2g12400.1	**At2g24150.1**	At2g33280.1	**At2g41610.1**	**At3g06470.1**	At3g21090.1	At3g46180.1
At2g13610.1	**At2g24170.1**	At2g33640.1	At2g43240.1	**At3g06550.1**	At3g21580.2	At3g46980.2
At2g13650.2	At2g24610.1	**At2g33670.1**	**At2g44110.1**	At3g06710.1	At3g21620.1	At3g47360.1
At2g15090.1	At2g24630.1	**At2g33750.1**	**At2g44110.2**	**At3g07330.1**	At3g23430.1	At3g47730.1
At2g15240.1	At2g24710.1	**At2g33750.2**	At2g44520.1	**At3g08930.1**	At3g23870.1	At3g47740.1
At2g15970.1	At2g24720.1	At2g33820.1	At2g44660.1	**At3g08930.2**	**At3g24460.1**	At3g47750.1
At2g16280.1	At2g25520.1	At2g34020.1	At2g45670.1	**At3g09320.1**	At3g25160.1	At3g47770.1
At2g16530.1	At2g26900.1	**At2g34390.1**	At2g45670.2	**At3g09570.1**	At3g25410.1	At3g47780.1
At2g16530.2	At2g27240.1	**At2g34390.2**	At2g45960.1	**At3g10290.1**	At3g25585.1	At3g47790.1
At2g16800.1	At2g28070.1	At2g34410.1	At2g46060.1	At3g10390.1	At3g25585.2	At3g47980.1
At2g16970.1	At2g28260.1	**At2g34980.1**	At2g46060.2	**At3g11320.1**	**At3g25805.1**	**At3g48740.1**
At2g17430.1	**At2g28315.1**	**At2g35650.1**	At2g46430.1	At3g11680.1	**At3g25950.1**	At3g50920.1
At2g17480.1	**At2g29050.1**	**At2g35710.1**	**At2g46440.1**	At3g13220.1	**At3g26090.1**	**At3g51970.1**
At2g18590.1	At2g29390.1	**At2g35710.2**	At2g46450.1	**At3g13772.1**	**At3g27270.1**	At3g52160.1
At2g18690.1	At2g29390.2	**At2g36300.1**	**At2g47115.1**	At3g14410.1	At3g27325.1	At3g52310.1
At2g18950.1	At2g29390.3	**At2g36305.1**	At2g47760.1	**At3g15380.1**	At3g27390.1	**At3g52760.1**

Table 21 (continued).

At2g19850.1	At2g29525.1	At2g37010.1	At2g48070.1	At3g15830.1	At3g27770.1	At3g53100.1
At2g19880.1	At2g29525.2	At2g37180.1	At2g48070.2	At3g15850.1	At3g28007.1	At3g53210.1
At3g53420.1	At4g01130.1	**At4g13800.1**	At4g20100.1	**At4g28370.1**	At5g01240.2	At5g15240.1
At3g53780.1	**At4g01430.1**	At4g14170.1	**At4g20310.1**	At4g28620.1	**At5g01460.1**	**At5g15410.1**
At3g53780.2	**At4g01430.2**	At4g14730.1	**At4g21260.1**	**At4g29200.1**	At5g01490.1	**At5g15410.2**
At3g54020.1	**At4g01440.1**	At4g14950.1	**At4g21570.1**	At4g30420.1	At5g01690.1	At5g16190.1
At3g54450.1	At4g01450.1	At4g14950.3	At4g21700.1	At4g30560.1	At5g01990.1	At5g16530.1
At3g54730.1	At4g01450.2	At4g15233.1	**At4g21790.1**	**At4g30850.1**	**At5g02630.1**	At5g16740.1
At3g55360.1	At4g01450.3	At4g15290.1	**At4g22270.1**	**At4g30850.2**	At5g03760.1	At5g17520.1
At3g56160.1	**At4g02600.1**	At4g15320.1	**At4g22330.1**	At4g31350.1	At5g03910.1	At5g17630.1
At3g56620.1	**At4g02690.1**	**At4g15430.1**	**At4g22340.1**	At4g31590.1	**At5g04160.1**	At5g18480.1
At3g57170.1	**At4g02900.1**	**At4g15470.1**	**At4g22340.2**	At4g32390.1	At5g04490.1	**At5g18520.1**
At3g57650.1	At4g03320.1	At4g16370.1	At4g22756.1	At4g33905.1	At5g04680.1	At5g19130.1
At3g58020.1	At4g03440.1	At4g16590.1	At4g23000.1	At4g34250.1	**At5g05350.1**	At5g19130.2
At3g58490.1	At4g03490.1	At4g16600.1	At4g23010.1	At4g34510.1	**At5g07050.1**	**At5g19380.1**
At3g59070.1	**At4g03820.2**	**At4g16850.1**	At4g23070.1	At4g34520.1	**At5g07250.1**	**At5g19870.1**
At3g59090.1	**At4g03950.1**	**At4g17250.1**	At4g23400.1	At4g35080.1	**At5g07630.1**	At5g19980.1
At3g59090.2	At4g07390.1	At4g17580.1	At4g23660.1	At4g35100.1	At5g08090.1	**At5g20270.1**
At3g59310.1	At4g07960.1	**At4g17790.1**	At4g23660.2	**At4g35180.1**	**At5g10840.1**	**At5g22130.1**
At3g59310.2	**At4g08290.1**	**At4g18210.1**	**At4g24250.1**	At4g35335.1	At5g11230.1	At5g22740.1
At3g59330.1	**At4g08878.1**	**At4g18220.1**	At4g24460.1	At4g35870.1	At5g11870.1	**At5g23660.1**
At3g59520.1	At4g09580.1	**At4g18230.1**	**At4g25010.1**	**At4g36830.1**	At5g11960.1	**At5g23980.1**
At3g60590.1	At4g09640.1	**At4g18540.1**	**At4g25350.1**	**At4g36850.1**	**At5g12170.1**	**At5g23990.1**
At3g60590.2	At4g09810.1	At4g18910.1	At4g25450.1	At4g37270.1	**At5g13170.1**	**At5g24600.1**
At3g60590.3	**At4g10310.1**	At4g19090.1	**At4g26580.1**	At4g37680.1	At5g13750.2	**At5g24790.1**
At3g60620.1	At4g10850.1	At4g19185.1	At4g27420.1	At4g37760.1	At5g13760.1	**At5g25100.1**
At3g61430.1	At4g11230.1	**At4g19645.1**	At4g27630.1	At4g38790.1	**At5g13890.1**	At5g25400.1

Table 21 (continued).

At3g63310.1
At4g00430.1
At4g01010.1
At5g27210.1
At5g27490.1
At5g27730.1
At5g35160.1
At5g35460.1
At5g35730.1
At5g37310.1
At5g38380.1
At5g38380.2
At5g40210.1
At5g40230.1
At5g40240.1
At5g40260.1
At5g40640.1
At5g40670.1
At5g40780.1
At5g40780.2
At5g41800.1
At5g42420.1
At5g45095.1
At5g45105.1
At5g45370.1
At5g45370.2
At5g45370.3
At4g12030.1
At4g12680.1
At4g13410.1
At5g47530.1
At5g47580.1
At5g47900.1
At5g49630.1
At5g50375.1
At5g50770.1
At5g50790.1
At5g50800.1
At5g52860.1
At5g53760.1
At5g55320.1
At5g55340.1
At5g55350.1
At5g55370.1
At5g55380.1
At5g57800.1
At5g57940.1
At5g57940.2
At5g57940.3
At5g59500.1
At5g59740.1
At5g60220.1
At5g60660.1
At5g60740.1
At4g19645.2
At4g19690.1
At4g19690.2
At5g62130.1
At5g62960.1
At5g64700.1
At5g65000.1
At5g65000.2
At5g65970.1
At4g27630.2
At4g28040.1
At4g28040.2
At4g39030.1
At4g39390.1
At4g39390.2
At5g13890.2
At5g13890.3
At5g14870.1
At5g25420.1
At5g26740.1
At5g26740.2
At5g46110.2
At5g46370.1
At5g47120.1
At5g47470.1
At5g60750.1
At5g61690.1
At5g61730.1
At5g61740.1

Table 22. The 92 *Arabidopsis* proteins with 7 transmembrane regions and external N-terminal predicted by ST-method as 7TMR candidates. The 54 proteins that overlapped with the proteins from Moriyama et.l[3] are in bolds with shaded background.

At1g01650.1
At1g03070.1
At1g10660.1
At1g10660.2
At1g10660.3
At1g10660.4
At1g11000.1
At1g11200.1
At1g11450.1
At1g14530.1
At1g14530.2
At1g15620.1
At1g16560.2
At1g16560.3
At1g18420.1
At1g21460.1
At1g26700.1
At1g32120.1
At1g42560.1
At1g47640.1
At1g48270.1
At1g49470.1
At1g52580.1
At1g53390.1
At1g55230.1
At1g57680.1
At1g57680.2
At1g63110.1
At1g71960.1
At1g75000.1
At1g77220.1
At2g01070.1
At2g02180.1
At2g16530.1
At2g16530.2
At2g16970.1
At2g31440.1
At2g33670.1
At2g35710.1
At2g35710.2
At2g39060.1
At2g41050.1
At2g41610.1
At2g44110.1
At2g44110.2
At2g46060.1
At2g47115.1
At3g01550.1
At3g04970.1
At3g05010.1
At3g05940.1
At3g06460.1
At3g06470.1
At3g09570.1
At3g10390.1
At3g15830.1
At3g16690.1
At3g17690.1
At3g19260.1
At3g20100.1
At3g26090.1
At3g27770.1
At3g28007.1
At3g28070.2
At3g28130.1
At3g28180.1
At3g48740.1
At3g50920.1
At3g59090.1
At3g59090.2
At3g59520.1
At3g63310.1
At4g02690.1
At4g07960.1
At4g10850.1
At4g17580.1
At4g20100.1
At4g20310.1
At4g21570.1
At4g21790.1
At4g24250.1
At4g25010.1
At4g36830.1
At4g36850.1
At5g03760.1
At5g13170.1
At5g19870.1
At5g23660.1
At5g23990.1
At5g26740.1
At5g26740.2
At5g27210.1
At5g37310.1
At5g38380.1
At5g40260.1
At5g42090.1
At5g50790.1
At5g50800.1
At5g53760.1
At5g62130.1
At5g62960.1
At1g63110.2
At3g09570.1
At4g25010.1

Table 23. The 717 rice proteins with 5-10 transmembarane regions predicted by ST-method as 7TMR candidates.

LOC_Os03g09060.1
LOC_Os04g43916.1
LOC_Os07g31140.1
LOC_Os02g26650.1
LOC_Os02g26650.2
LOC_Os02g26650.3
LOC_Os08g30780.1
LOC_Os02g11960.1
LOC_Os05g02870.1
LOC_Os03g64200.1
LOC_Os05g02890.1
LOC_Os06g08560.1
LOC_Os06g38950.1
LOC_Os06g51460.1
LOC_Os07g33780.1
LOC_Os11g22350.1
LOC_Os11g07600.1
LOC_Os01g08260.1
LOC_Os01g08260.2
LOC_Os06g40550.1
LOC_Os09g07670.1
LOC_Os12g22110.1
LOC_Os04g11820.1
LOC_Os04g44610.1
LOC_Os05g13520.1
LOC_Os06g30730.1
LOC_Os07g18874.1
LOC_Os01g57360.1
LOC_Os05g02490.1
LOC_Os06g39260.1
LOC_Os04g01510.1
LOC_Os02g46320.1
LOC_Os04g34010.1
LOC_Os05g34980.1
LOC_Os06g12330.1
LOC_Os06g36210.1
LOC_Os07g04180.1
LOC_Os07g04180.2
LOC_Os05g50920.1
LOC_Os01g40360.1
LOC_Os05g30150.1
LOC_Os11g19240.1
LOC_Os02g49060.1
LOC_Os04g39489.1
LOC_Os04g56470.1
LOC_Os06g16420.1
LOC_Os06g16420.2
LOC_Os06g16420.3
LOC_Os06g34830.1
LOC_Os06g36180.1
LOC_Os09g16550.1
LOC_Os07g34390.1
LOC_Os11g08020.1
LOC_Os05g37210.1
LOC_Os08g05590.1
LOC_Os05g11560.1
LOC_Os04g44060.1
LOC_Os07g26630.1
LOC_Os09g36930.1
LOC_Os02g57720.1
LOC_Os05g14240.1
LOC_Os09g20630.1
LOC_Os01g63770.2
LOC_Os05g37470.1
LOC_Os05g37470.2
LOC_Os05g50140.1
LOC_Os01g51780.1
LOC_Os07g20510.1
LOC_Os09g31478.1
LOC_Os09g31478.2
LOC_Os09g32770.1
LOC_Os03g09850.1
LOC_Os05g48270.1
LOC_Os02g03280.1
LOC_Os02g03280.2
LOC_Os11g37900.1
LOC_Os01g45750.1
LOC_Os01g45750.2
LOC_Os05g08430.1
LOC_Os01g12680.1
LOC_Os01g12680.2
LOC_Os05g13320.1
LOC_Os07g08350.1
LOC_Os01g14520.1
LOC_Os05g18670.1
LOC_Os05g28950.1
LOC_Os05g28950.2
LOC_Os06g30910.1
LOC_Os03g62430.1
LOC_Os09g12790.1
LOC_Os05g02940.2
LOC_Os01g52070.1
LOC_Os06g14310.1
LOC_Os05g42250.1
LOC_Os01g20160.1
LOC_Os02g07830.1
LOC_Os04g51820.1
LOC_Os06g48800.1
LOC_Os02g02750.1
LOC_Os02g02750.2
LOC_Os02g02750.3
LOC_Os06g10280.1
LOC_Os06g29650.1
LOC_Os02g46350.1
LOC_Os02g46350.2
LOC_Os07g24190.2
LOC_Os07g14850.1
LOC_Os04g35020.1
LOC_Os07g36750.1
LOC_Os06g22980.1
LOC_Os02g49332.1
LOC_Os02g49332.2
LOC_Os02g49332.3
LOC_Os06g02180.1
LOC_Os12g25200.1
LOC_Os05g28200.1
LOC_Os05g33230.1
LOC_Os01g68970.2
LOC_Os01g68970.3
LOC_Os06g33570.1
LOC_Os06g33600.1
LOC_Os09g38580.2
LOC_Os02g41710.1
LOC_Os03g55100.1
LOC_Os02g53340.1
LOC_Os06g10580.1
LOC_Os04g55080.1
LOC_Os03g44440.1
LOC_Os11g19700.1
LOC_Os02g42890.1
LOC_Os12g04270.2
LOC_Os11g14080.1
LOC_Os05g09550.1
LOC_Os03g16790.1
LOC_Os05g38360.1

Table 23 (continued).

LOC_Os12g22284.1	LOC_Os02g17280.1	LOC_Os01g12680.3	LOC_Os07g36630.1	LOC_Os05g38360.2
LOC_Os05g38360.3	LOC_Os01g07700.1	LOC_Os01g67870.1	LOC_Os04g31020.1	LOC_Os06g03380.1
LOC_Os06g20400.1	LOC_Os01g10100.1	LOC_Os01g70550.2	LOC_Os04g36630.1	LOC_Os05g09050.1
LOC_Os06g20400.2	LOC_Os01g10100.2	LOC_Os01g70550.3	LOC_Os04g37520.1	LOC_Os06g10070.1
LOC_Os05g23700.1	LOC_Os01g15770.1	LOC_Os01g72570.1	LOC_Os04g38850.1	LOC_Os06g10100.1
LOC_Os02g02110.1	LOC_Os01g23870.1	LOC_Os01g72570.2	LOC_Os04g39170.2	LOC_Os06g19260.1
LOC_Os04g41810.1	LOC_Os01g24340.1	LOC_Os02g03790.1	LOC_Os04g40700.1	LOC_Os06g19370.1
LOC_Os03g47070.1	LOC_Os01g24430.1	LOC_Os02g09440.1	LOC_Os04g40700.2	LOC_Os06g19370.2
LOC_Os05g32720.1	LOC_Os01g29220.1	LOC_Os02g19470.1	LOC_Os04g40700.3	LOC_Os06g38320.1
LOC_Os02g49050.1	LOC_Os01g42760.1	LOC_Os02g22060.1	LOC_Os04g45200.1	LOC_Os06g38320.2
LOC_Os06g37160.1	LOC_Os01g48620.1	LOC_Os02g31874.1	LOC_Os04g55110.1	LOC_Os06g42850.1
LOC_Os06g44140.1	LOC_Os01g48640.1	LOC_Os02g31874.2	LOC_Os04g58470.1	LOC_Os06g43780.1
LOC_Os08g44150.1	LOC_Os01g48660.1	LOC_Os02g32504.2	LOC_Os04g58504.1	LOC_Os06g46820.2
LOC_Os11g07910.1	LOC_Os01g48800.1	LOC_Os02g35830.1	LOC_Os05g02010.3	LOC_Os07g08290.1
LOC_Os11g07910.2	LOC_Os01g51040.1	LOC_Os02g36490.1	LOC_Os05g02750.1	LOC_Os07g18250.1
LOC_Os12g07670.1	LOC_Os01g51190.1	LOC_Os02g36490.2	LOC_Os05g24140.1	LOC_Os07g18250.2
LOC_Os02g02980.1	LOC_Os01g54784.1	LOC_Os02g38430.1	LOC_Os05g24690.1	LOC_Os07g36400.1
LOC_Os02g02980.2	LOC_Os01g56230.1	LOC_Os02g49790.1	LOC_Os05g25890.1	LOC_Os07g39280.1
LOC_Os07g37110.1	LOC_Os01g56230.2	LOC_Os02g55590.1	LOC_Os05g40700.1	LOC_Os07g40470.1
LOC_Os08g43470.1	LOC_Os01g57700.1	LOC_Os03g09090.1	LOC_Os05g43540.1	LOC_Os07g46030.1
LOC_Os09g36370.1	LOC_Os01g57710.1	LOC_Os03g10230.1	LOC_Os05g48370.1	LOC_Os07g46090.1
LOC_Os05g45310.1	LOC_Os01g57710.2	LOC_Os03g14880.1	LOC_Os05g51620.1	LOC_Os07g46430.1
LOC_Os01g25189.1	LOC_Os01g58500.1	LOC_Os03g44840.1	LOC_Os05g51620.2	LOC_Os07g46430.2
LOC_Os01g25189.2	LOC_Os01g58620.1	LOC_Os03g51650.1	LOC_Os06g02370.1	LOC_Os08g03430.1
LOC_Os01g25189.3	LOC_Os01g60120.1	LOC_Os03g54920.1	LOC_Os06g02370.2	LOC_Os08g03430.2
LOC_Os09g39220.1	LOC_Os01g63854.1	LOC_Os03g55730.1	LOC_Os06g02370.3	LOC_Os08g15650.1
LOC_Os01g03110.1	LOC_Os01g64930.1	LOC_Os03g55730.2	LOC_Os06g02830.1	LOC_Os08g28970.1

Table 23 (continued).

LOC_Os01g04250.1	LOC_Os01g66190.1	LOC_Os04g03920.1	LOC_Os06g02960.1	LOC_Os08g30020.4
LOC_Os01g07660.1	LOC_Os01g66190.2	LOC_Os04g07110.1	LOC_Os06g03050.1	LOC_Os08g41000.1
LOC_Os09g04310.1	LOC_Os03g08360.2	LOC_Os01g64990.1	LOC_Os02g44450.1	LOC_Os08g21710.1
LOC_Os09g04339.1	LOC_Os12g37530.3	LOC_Os01g64990.2	LOC_Os02g56730.1	LOC_Os08g25830.1
LOC_Os09g09360.1	LOC_Os07g24230.1	LOC_Os03g49700.1	LOC_Os04g05620.1	LOC_Os08g32310.1
LOC_Os09g16760.1	LOC_Os04g51180.1	LOC_Os02g08320.1	LOC_Os04g06560.1	LOC_Os08g32730.1
LOC_Os09g17329.2	LOC_Os04g51180.2	LOC_Os03g10300.1	LOC_Os05g14680.1	LOC_Os09g04430.1
LOC_Os09g27110.1	LOC_Os08g01610.1	LOC_Os03g13040.1	LOC_Os05g34190.1	LOC_Os09g12400.1
LOC_Os09g38660.1	LOC_Os02g08230.1	LOC_Os06g43620.1	LOC_Os05g35310.1	LOC_Os09g14750.1
LOC_Os09g39250.1	LOC_Os07g01560.2	LOC_Os06g44250.1	LOC_Os05g35840.1	LOC_Os09g15200.1
LOC_Os11g01030.1	LOC_Os08g08070.1	LOC_Os12g32640.1	LOC_Os06g03820.1	LOC_Os09g19990.1
LOC_Os11g06070.1	LOC_Os04g59550.2	LOC_Os01g04190.2	LOC_Os06g07110.1	LOC_Os09g21200.1
LOC_Os11g32470.1	LOC_Os04g59550.3	LOC_Os06g43880.1	LOC_Os06g18880.1	LOC_Os09g26290.1
LOC_Os11g33100.1	LOC_Os08g01410.1	LOC_Os03g48030.1	LOC_Os06g21430.1	LOC_Os10g02260.1
LOC_Os11g34110.1	LOC_Os04g46750.1	LOC_Os01g07670.1	LOC_Os06g40540.1	LOC_Os10g05410.1
LOC_Os11g39920.1	LOC_Os04g46750.2	LOC_Os01g14230.1	LOC_Os06g42900.1	LOC_Os10g07994.1
LOC_Os11g40320.1	LOC_Os07g03260.1	LOC_Os01g23070.1	LOC_Os06g48490.1	LOC_Os10g07998.1
LOC_Os11g42080.1	LOC_Os01g56130.1	LOC_Os01g23370.1	LOC_Os07g02520.1	LOC_Os10g08014.1
LOC_Os12g02100.2	LOC_Os05g43530.1	LOC_Os01g24240.1	LOC_Os07g11500.1	LOC_Os10g11310.1
LOC_Os12g21940.1	LOC_Os01g34930.1	LOC_Os01g42340.1	LOC_Os07g24280.1	LOC_Os10g12190.1
LOC_Os12g31480.1	LOC_Os02g41520.1	LOC_Os01g43110.1	LOC_Os07g24770.1	LOC_Os10g12400.1
LOC_Os12g32260.1	LOC_Os01g11260.1	LOC_Os01g46690.1	LOC_Os07g30250.1	LOC_Os10g12750.1
LOC_Os12g38810.1	LOC_Os07g46640.1	LOC_Os01g46890.1	LOC_Os07g37940.1	LOC_Os10g14920.1
LOC_Os12g44180.1	LOC_Os06g12460.1	LOC_Os01g61960.1	LOC_Os07g45550.1	LOC_Os10g14920.2
LOC_Os06g29790.1	LOC_Os07g43710.1	LOC_Os01g62530.1	LOC_Os07g46040.1	LOC_Os10g14920.4
LOC_Os12g40340.1	LOC_Os03g26044.1	LOC_Os01g62550.1	LOC_Os07g46080.1	LOC_Os10g14920.5
LOC_Os04g48930.1	LOC_Os03g49480.1	LOC_Os01g74130.1	LOC_Os07g47850.1	LOC_Os10g20090.1

Table 23 (continued).

LOC_Os04g48930.2
LOC_Os05g04120.1
LOC_Os03g08360.1
LOC_Os10g23180.1
LOC_Os10g25450.1
LOC_Os10g27220.1
LOC_Os10g28440.1
LOC_Os10g30910.1
LOC_Os10g31290.1
LOC_Os10g33820.1
LOC_Os10g33920.1
LOC_Os10g33920.2
LOC_Os10g34050.1
LOC_Os10g34110.1
LOC_Os10g38030.1
LOC_Os10g39220.1
LOC_Os10g40640.1
LOC_Os10g42180.1
LOC_Os10g42780.2
LOC_Os10g42780.3
LOC_Os10g42780.4
LOC_Os10g42780.5
LOC_Os11g06350.1
LOC_Os11g07530.1
LOC_Os11g07550.1
LOC_Os11g18070.1
LOC_Os11g24220.1
LOC_Os12g43890.1
LOC_Os03g61100.1
LOC_Os01g48980.1
LOC_Os12g33170.1
LOC_Os12g36160.1
LOC_Os12g43080.1
LOC_Os12g43300.1
LOC_Os06g09930.1
LOC_Os01g08290.1
LOC_Os01g19290.2
LOC_Os01g32280.1
LOC_Os02g06010.1
LOC_Os02g06010.2
LOC_Os03g11590.1
LOC_Os08g10350.1
LOC_Os09g14520.1
LOC_Os01g10970.1
LOC_Os01g19240.1
LOC_Os05g01580.1
LOC_Os11g04140.1
LOC_Os01g10980.1
LOC_Os01g10990.1
LOC_Os01g36580.1
LOC_Os02g22680.1
LOC_Os03g05530.1
LOC_Os03g49940.1
LOC_Os05g01570.1
LOC_Os02g19680.1
LOC_Os02g24480.1
LOC_Os02g38830.1
LOC_Os09g25784.1
LOC_Os09g25784.2
LOC_Os09g25810.2
LOC_Os11g37720.1
LOC_Os12g18960.1
LOC_Os12g18960.2
LOC_Os05g03000.1
LOC_Os05g03000.2
LOC_Os01g07310.1
LOC_Os01g07310.2
LOC_Os01g60780.1
LOC_Os02g41780.1
LOC_Os05g07670.1
LOC_Os05g38250.1
LOC_Os06g30950.1
LOC_Os06g30950.2
LOC_Os05g36150.1
LOC_Os01g40280.1
LOC_Os01g40280.2
LOC_Os03g03590.1
LOC_Os03g03590.2
LOC_Os07g39010.1
LOC_Os07g39010.2
LOC_Os07g39010.3
LOC_Os08g08530.1
LOC_Os08g13600.1
LOC_Os08g19000.1
LOC_Os09g13870.2
LOC_Os06g03760.1
LOC_Os06g03760.2
LOC_Os03g15750.1
LOC_Os03g21690.1
LOC_Os03g21690.2
LOC_Os05g38720.1
LOC_Os09g26830.1
LOC_Os05g14820.1
LOC_Os12g14100.1
LOC_Os07g08310.1
LOC_Os11g04060.1
LOC_Os11g04030.1
LOC_Os11g04030.3
LOC_Os11g04030.4
LOC_Os11g04150.1
LOC_Os12g03860.2
LOC_Os12g03860.3
LOC_Os12g03860.4
LOC_Os06g13200.1
LOC_Os06g21950.1
LOC_Os03g43720.5
LOC_Os04g51970.1
LOC_Os09g38690.2
LOC_Os10g20350.1
LOC_Os10g20390.1
LOC_Os10g20770.1
LOC_Os12g10280.1
LOC_Os07g29610.1
LOC_Os03g46750.1
LOC_Os02g44910.1
LOC_Os02g44910.2
LOC_Os03g61210.1
LOC_Os07g14090.1
LOC_Os07g32230.1
LOC_Os07g32230.2
LOC_Os01g69010.1
LOC_Os12g36660.1
LOC_Os02g58620.1
LOC_Os03g37470.1
LOC_Os05g48040.2
LOC_Os06g29844.1
LOC_Os06g29994.1
LOC_Os06g49310.1
LOC_Os06g49310.2
LOC_Os07g31884.1
LOC_Os03g11734.3
LOC_Os06g29950.1
LOC_Os03g08910.1
LOC_Os11g03484.1
LOC_Os12g03200.1

Table 23 (continued).

LOC_Os12g01020.1	LOC_Os05g37200.1	LOC_Os02g19820.1	LOC_Os09g38690.3	LOC_Os12g03230.1
LOC_Os12g03950.1	LOC_Os06g01972.1	LOC_Os12g30040.1	LOC_Os09g38690.4	LOC_Os02g26840.1
LOC_Os12g11430.1	LOC_Os07g30210.1	LOC_Os01g47580.1	LOC_Os11g01590.1	LOC_Os06g36800.1
LOC_Os12g19830.1	LOC_Os07g34110.1	LOC_Os08g27030.1	LOC_Os01g08660.1	LOC_Os02g25700.1
LOC_Os04g20880.1	LOC_Os11g03670.1	LOC_Os12g33300.1	LOC_Os12g32820.1	LOC_Os01g02890.1
LOC_Os05g05200.1	LOC_Os01g36070.1	LOC_Os09g36600.1	LOC_Os01g68524.1	LOC_Os01g49020.1
LOC_Os05g05200.2	LOC_Os01g50460.1	LOC_Os12g01570.1	LOC_Os07g19530.1	LOC_Os05g48060.1
LOC_Os05g10810.1	LOC_Os01g65880.1	LOC_Os01g65310.1	LOC_Os05g13330.1	LOC_Os05g48060.2
LOC_Os07g35570.1	LOC_Os05g12320.1	LOC_Os01g65986.1	LOC_Os05g13330.2	LOC_Os02g40870.1
LOC_Os02g45344.1	LOC_Os05g35140.1	LOC_Os02g29510.1	LOC_Os09g21340.1	LOC_Os02g40870.2
LOC_Os01g67030.1	LOC_Os08g42350.1	LOC_Os02g29510.2	LOC_Os08g32500.1	LOC_Os02g37050.1
LOC_Os02g12870.1	LOC_Os02g18700.1	LOC_Os02g29510.3	LOC_Os06g10810.1	LOC_Os03g03690.1
LOC_Os03g02530.1	LOC_Os12g37580.1	LOC_Os04g30450.1	LOC_Os01g58870.1	LOC_Os06g12500.1
LOC_Os04g48130.1	LOC_Os06g06440.1	LOC_Os05g35060.1	LOC_Os03g49570.1	LOC_Os06g44610.1
LOC_Os08g43320.1	LOC_Os12g29220.1	LOC_Os05g35570.1	LOC_Os05g03070.1	LOC_Os06g44610.2
LOC_Os09g35730.1	LOC_Os02g48460.1	LOC_Os05g43790.1	LOC_Os05g41480.1	LOC_Os08g36040.1
LOC_Os06g46310.1	LOC_Os01g31870.3	LOC_Os05g43790.2	LOC_Os05g41480.2	LOC_Os09g27250.1
LOC_Os06g46310.2	LOC_Os01g31870.4	LOC_Os11g09140.1	LOC_Os11g04380.1	LOC_Os04g46050.1
LOC_Os06g46310.3	LOC_Os01g31870.5	LOC_Os04g41320.1	LOC_Os12g04170.1	LOC_Os04g46050.2
LOC_Os07g15460.1	LOC_Os06g38294.1	LOC_Os01g67330.1	LOC_Os02g56510.1	LOC_Os06g49240.1
LOC_Os02g51110.1	LOC_Os07g09010.1	LOC_Os01g67330.2	LOC_Os08g04110.2	LOC_Os11g18110.1
LOC_Os06g12310.1	LOC_Os08g44750.1	LOC_Os02g36390.1	LOC_Os02g03460.1	LOC_Os01g55610.1
LOC_Os09g38100.1	LOC_Os01g12130.1	LOC_Os02g36390.2	LOC_Os02g05320.1	LOC_Os01g65100.3
LOC_Os04g36680.1	LOC_Os02g30910.1	LOC_Os02g39200.2	LOC_Os02g05320.2	LOC_Os01g65110.1
LOC_Os02g10350.1	LOC_Os11g31190.1	LOC_Os08g38400.2	LOC_Os11g01450.1	LOC_Os01g65140.1
LOC_Os03g03700.1	LOC_Os01g42110.1	LOC_Os06g03540.1	LOC_Os12g01480.1	LOC_Os01g65200.1
LOC_Os06g29110.1	LOC_Os01g42090.1	LOC_Os06g03560.1	LOC_Os02g05400.1	LOC_Os03g13274.3

Table 23 (continued).

LOC_Os01g66510.1	LOC_Os05g51090.1	LOC_Os06g03700.1	LOC_Os03g59070.1	LOC_Os05g27010.1
LOC_Os05g34550.1	LOC_Os01g36590.1	LOC_Os05g44360.2	LOC_Os01g63060.1	LOC_Os05g35650.1
LOC_Os10g39520.1	LOC_Os12g03920.1	LOC_Os07g23430.1	LOC_Os05g21180.3	LOC_Os06g49220.1
LOC_Os05g09450.1	LOC_Os02g47500.1	LOC_Os04g53930.2	LOC_Os05g37910.1	LOC_Os08g41590.1
LOC_Os06g22080.1	LOC_Os06g08110.1	LOC_Os02g47570.1	LOC_Os05g37910.2	LOC_Os01g55200.1
LOC_Os07g48130.1	LOC_Os12g35610.1	LOC_Os02g45520.2	LOC_Os03g05390.1	LOC_Os02g40030.1
LOC_Os07g48130.2	LOC_Os12g35610.2	LOC_Os06g03860.1	LOC_Os03g05390.2	LOC_Os09g02170.1
LOC_Os07g47350.1	LOC_Os05g44280.1	LOC_Os03g57840.1	LOC_Os03g05390.3	LOC_Os05g47530.1
LOC_Os07g47350.2	LOC_Os01g57974.1	LOC_Os03g57840.2	LOC_Os03g05390.4	LOC_Os01g53570.1
LOC_Os06g15910.1	LOC_Os04g20960.1	LOC_Os03g07480.1	LOC_Os01g16260.2	LOC_Os03g53400.1
LOC_Os08g39950.1	LOC_Os07g17270.1	LOC_Os03g07480.4	LOC_Os02g40090.1	LOC_Os03g58140.1
LOC_Os09g27580.2	LOC_Os09g37310.1	LOC_Os12g44380.3	LOC_Os02g40090.2	LOC_Os03g58150.1
LOC_Os01g16170.1	LOC_Os11g47840.1	LOC_Os07g05640.1	LOC_Os02g40090.3	LOC_Os05g33360.1
LOC_Os01g16170.2	LOC_Os03g24390.1	LOC_Os07g10590.1	LOC_Os05g12490.1	LOC_Os05g33360.2
LOC_Os01g16170.3	LOC_Os04g01300.1	LOC_Os11g38160.1	LOC_Os12g05780.2	LOC_Os07g08060.1
LOC_Os01g16170.4	LOC_Os04g01300.2	LOC_Os06g05160.1	LOC_Os12g05830.1	LOC_Os07g08070.1
LOC_Os06g44840.1	LOC_Os07g39400.1	LOC_Os07g38110.1	LOC_Os08g27980.1	LOC_Os11g37200.1
LOC_Os08g15460.2	LOC_Os06g19680.1	LOC_Os07g38110.2	LOC_Os02g02460.1	LOC_Os12g31850.1
LOC_Os12g41840.1	LOC_Os06g19680.2	LOC_Os02g54990.1	LOC_Os02g43410.1	LOC_Os12g31850.2
LOC_Os03g31570.1	LOC_Os07g17280.1	LOC_Os02g54990.2	LOC_Os02g43410.2	LOC_Os12g31890.1
LOC_Os05g27570.1	LOC_Os02g02530.1	LOC_Os03g34300.2	LOC_Os05g16280.1	LOC_Os05g51610.1
LOC_Os11g01842.1	LOC_Os05g36070.1	LOC_Os04g42720.1	LOC_Os01g13770.1	LOC_Os01g61780.1
LOC_Os01g61060.1	LOC_Os05g36070.2	LOC_Os02g45870.1	LOC_Os01g13770.2	LOC_Os03g14690.1
LOC_Os05g39730.1	LOC_Os04g31210.1	LOC_Os03g02850.1	LOC_Os05g15160.1	LOC_Os01g46270.1
LOC_Os07g36820.2	LOC_Os04g31210.2	LOC_Os03g02850.2	LOC_Os05g07870.1	LOC_Os09g15170.1
LOC_Os09g17830.2	LOC_Os12g07270.1	LOC_Os03g36790.1	LOC_Os08g25624.1	LOC_Os06g43200.1
LOC_Os09g33720.1	LOC_Os01g11414.1	LOC_Os03g36790.2	LOC_Os07g46780.1	LOC_Os06g43210.1

Table 23 (continued).

LOC_Os01g05720.1	LOC_Os02g04630.1	LOC_Os07g01250.1	LOC_Os07g46780.3	LOC_Os07g31850.1
LOC_Os01g34390.1	LOC_Os05g05590.1	LOC_Os12g27310.1	LOC_Os07g38850.1	LOC_Os07g30100.1
LOC_Os09g29210.1	LOC_Os07g47100.3	LOC_Os01g41420.1	LOC_Os06g33210.1	LOC_Os08g01040.1
LOC_Os09g29210.2	LOC_Os06g21360.1	LOC_Os06g12320.1	LOC_Os07g38610.1	LOC_Os01g74110.1
LOC_Os09g38510.1	LOC_Os09g11450.2	LOC_Os07g38380.1	LOC_Os03g02670.1	LOC_Os08g01030.1
LOC_Os01g61880.2	LOC_Os02g45520.1	LOC_Os03g01700.1	LOC_Os03g02670.2	
LOC_Os04g52310.1	LOC_Os05g10940.1	LOC_Os08g10630.1	LOC_Os03g46470.1	

Table 24. The 114 rice proteins with 7 transmembrane regions and external N-terminal predicted by ST-method as 7TMR candidates.

LOC_Os01g07310.2
LOC_Os01g07700.1
LOC_Os01g08260.1
LOC_Os01g08260.2
LOC_Os01g42090.1
LOC_Os01g42110.1
LOC_Os01g46690.1
LOC_Os01g49020.1
LOC_Os01g50460.1
LOC_Os01g54784.1
LOC_Os01g61060.1
LOC_Os01g61780.1
LOC_Os01g61960.1
LOC_Os01g64930.1
LOC_Os01g65880.1
LOC_Os01g66190.1
LOC_Os01g66190.2
LOC_Os01g66510.1
LOC_Os02g02750.2
LOC_Os02g03790.1
LOC_Os02g10350.1
LOC_Os02g17280.1
LOC_Os02g19470.1
LOC_Os02g19680.1
LOC_Os02g19820.1
LOC_Os02g29510.2
LOC_Os02g29510.3
LOC_Os02g36490.1
LOC_Os02g36490.2
LOC_Os02g39200.2
LOC_Os02g43410.1
LOC_Os02g43410.2
LOC_Os02g44450.1
LOC_Os02g44910.1
LOC_Os02g44910.2
LOC_Os02g45870.1
LOC_Os03g02850.1
LOC_Os03g02850.2
LOC_Os03g03590.1
LOC_Os03g03590.2
LOC_Os03g14690.1
LOC_Os03g14880.1
LOC_Os03g47070.1
LOC_Os03g49480.1
LOC_Os03g55730.1
LOC_Os05g02750.1
LOC_Os05g12320.1
LOC_Os05g13330.2
LOC_Os05g24690.1
LOC_Os05g27010.1
LOC_Os05g34550.1
LOC_Os05g35140.1
LOC_Os05g39730.1
LOC_Os05g51090.1
LOC_Os05g51620.2
LOC_Os06g02960.1
LOC_Os06g03820.1
LOC_Os06g05980.2
LOC_Os06g06130.1
LOC_Os06g09930.1
LOC_Os06g22600.1
LOC_Os06g29110.1
LOC_Os06g35930.1
LOC_Os06g43200.1
LOC_Os06g46284.1
LOC_Os06g46340.1
LOC_Os06g51100.1
LOC_Os06g51100.2
LOC_Os06g51460.1
LOC_Os07g01250.1
LOC_Os07g02430.1
LOC_Os07g14090.1
LOC_Os07g17270.1
LOC_Os07g30100.1
LOC_Os07g31140.1
LOC_Os07g32230.1
LOC_Os07g33780.1
LOC_Os07g43080.1
LOC_Os07g46430.2
LOC_Os08g05590.1
LOC_Os08g13600.1
LOC_Os08g15460.2
LOC_Os08g15650.1
LOC_Os08g25830.1
LOC_Os08g30020.4
LOC_Os08g36030.1
LOC_Os08g36040.1
LOC_Os08g41920.1
LOC_Os08g42350.1
LOC_Os08g44150.1
LOC_Os09g07670.1
LOC_Os09g16290.1
LOC_Os09g25784.1
LOC_Os09g26144.1
LOC_Os09g26660.1
LOC_Os09g27250.1
LOC_Os09g27260.1
LOC_Os09g32470.1
LOC_Os09g33470.1
LOC_Os09g34990.1
LOC_Os09g37310.1
LOC_Os09g38660.1
LOC_Os09g38690.4
LOC_Os10g05410.1
LOC_Os10g12750.1
LOC_Os10g14920.5
LOC_Os10g27220.1
LOC_Os10g34110.1
LOC_Os12g03230.1
LOC_Os12g03950.1
LOC_Os12g10280.1
LOC_Os12g13790.1
LOC_Os12g18110.1
LOC_Os12g18110.2
LOC_Os12g19830.1
LOC_Os12g23610.1
LOC_Os12g29220.1
LOC_Os12g32260.1
LOC_Os12g32640.1
LOC_Os12g33170.1
LOC_Os12g43890.1
LOC_Os02g30910.1
LOC_Os05g51620.1
LOC_Os08g15420.1
LOC_Os12g02589.1

Table 25. The 382 maize sequences with 5-10 transmembrane regions predicted by ST-method.

AZM5_21847_509
AZM5_20071_1020
AZM5_19331_353
AZM5_92883_1001
AZM5_101376_3
AZM5_22280_1020
AZM5_85827_1331
AZM5_89000_1596
AZM5_85328_1347
AZM5_2920_393
AZM5_101343_1
AZM5_2921_888
AZM5_23307_107
AZM5_104049_371
AZM5_99085_441
AZM5_4326_194
AZM5_60537_3
AZM5_5303_1211
AZM5_13121_2278
AZM5_7620_1137
AZM5_21488_455
AZM5_92740_617
AZM5_2624_4959
AZM5_5304_350
AZM5_87650_2010
AZM5_84788_565
AZM5_26815_584

AZM5_83999_2
AZM5_12563_2353
AZM5_3464_176
AZM5_5885_3083
AZM5_92279_1568
AZM5_25226_483
AZM5_1636_3632
AZM5_14381_2725
AZM5_15360_1765
AZM5_4856_3010
AZM5_27028_257
AZM5_29502_108
AZM5_5361_31
AZM5_19390_1311
AZM5_20013_3
AZM5_6268_557
AZM5_16199_2760
AZM5_11310_4656
AZM5_87258_2646
AZM5_28522_2
AZM5_25256_683
AZM5_14993_1286
AZM5_93971_3
AZM5_25772_61
AZM5_6154_145
AZM5_85129_670
AZM5_19557_2015

AZM5_3562_1610
AZM5_14541_2079
AZM5_4242_756
AZM5_17889_909
AZM5_5031_1338
AZM5_4476_1209
AZM5_17079_863
AZM5_4718_2082
AZM5_6602_2120
AZM5_17410_885
AZM5_14373_1817
AZM5_7331_3
AZM5_4243_291
AZM5_22359_735
AZM5_7790_372
AZM5_12563_4047
AZM5_5383_1861
AZM5_7768_1861
AZM5_3804_934
AZM5_31108_202
AZM5_85137_124
AZM5_12529_263
AZM5_6773_2482
OGTBE52TV_3
AZM5_16584_3
AZM5_19503_785
AZM5_34016_313

AZM5_12528_2055
AZM5_20897_1591
AZM5_6923_2088
AZM5_18966_2
AZM5_92120_754
AZM5_27164_857
AZM5_12167_1352
AZM5_86216_1119
AZM5_6155_1860
AZM5_20551_1281
AZM5_107468_612
AZM5_108454_651
AZM5_92642_3
AZM5_21945_1
AZM5_15626_2167
AZM5_16510_2456
AZM5_14902_1563
AZM5_86616_342
AZM5_91015_1431
AZM5_19332_2
AZM5_105945_420
AZM5_4480_3814
AZM5_19832_1144
AZM5_18427_261
AZM5_89680_142
AZM5_16068_1827
AZM5_21478_1179

AZM5_15445_125
AZM5_13841_1065
AZM5_1474_1408
AZM5_12385_632
AZM5_25277_1153
AZM5_10621_3540
AZM5_13416_385
AZM5_18894_1020
AZM5_102010_688
AZM5_13558_35
AZM5_86477_346
AZM5_21668_2179
AZM5_136700_1
AZM5_26182_3
AZM5_29331_245
AZM5_93253_1877
AZM5_3974_1355
PUIKX69TB_1
AZM5_27695_379
AZM5_13741_4394
AZM5_5305_391
AZM5_17726_1245
AZM5_5170_86
AZM5_637_3590
AZM5_94397_208
AZM5_1477_235
AZM5_13166_875

Table 25 (continued).

AZM5_50_1248	AZM5_7789_2	AZM5_14617_1	AZM5_20341_453	AZM5_97107_3
AZM5_1475_1908	AZM5_13946_3	AZM5_86682_3	AZM5_17212_1374	PUIGA58TD_2
AZM5_98494_506	AZM5_7545_2669	AZM5_11941_1009	AZM5_14246_597	AZM5_588_3743
AZM5_13241_2606	AZM5_6524_823	AZM5_16041_155	AZM5_26415_1432	AZM5_32700_187
AZM5_6414_1551	AZM5_10267_4301	AZM5_107833_206	AZM5_94597_1666	AZM5_21715_361
AZM5_85664_1056	AZM5_85906_991	AZM5_5633_360	AZM5_1636_2	AZM5_95474_461
AZM5_12392_885	AZM5_6227_1537	AZM5_44345_334	AZM5_18216_1942	AZM5_87683_164
AZM5_43239_282	AZM5_15522_1643	AZM5_98354_2	AZM5_6270_158	AZM5_85540_1
AZM5_129081_2	AZM5_16040_642	AZM5_30323_167	AZM5_22034_2	AZM5_105249_26
AZM5_23638_1263	AZM5_45093_281	AZM5_10733_1305	AZM5_1223_2734	AZM5_17531_2469
AZM5_2680_2498	AZM5_21667_639	AZM5_10731_1415	AZM5_87106_3	AZM5_95178_2
AZM5_588_725	AZM5_99430_1460	AZM5_104173_195	AZM5_10386_2825	AZM5_4500_3866
AZM5_91540_1929	AZM5_92745_1230	AZM5_3800_1608	PUIAN49TD_18	AZM5_11918_223
AZM5_17322_1475	AZM5_10560_1537	AZM5_10732_1363	AZM5_23750_779	AZM5_104842_3
AZM5_25628_1047	AZM5_94595_218	AZM5_10763_1294	AZM5_50151_271	AZM5_11769_270
AZM5_21904_754	AZM5_40694_494	AZM5_112524_395	AZM5_98981_491	AZM5_87682_349
AZM5_13997_338	AZM5_101987_1750	AZM5_94526_543	AZM5_86898_2333	AZM5_16890_244
AZM5_5458_1057	AZM5_92504_526	AZM5_3576_69	PUJDN06TD_3	AZM5_7371_1771
AZM5_21182_2	AZM5_12714_3222	AZM5_102847_99	AZM5_16577_1209	AZM5_4501_306
AZM5_3705_3	AZM5_4389_2119	AZM5_19994_1097	AZM5_84925_1741	AZM5_30370_173
AZM5_24606_522	AZM5_10560_3270	AZM5_6843_366	AZM5_19096_1616	AZM5_5322_3238
AZM5_87895_1311	AZM5_99466_996	AZM5_20200_1966	AZM5_91063_2	AZM5_85517_160
AZM5_94301_415	PUJCM86TD_2	AZM5_13097_2680	AZM5_5360_355	AZM5_17960_1013
AZM5_88324_467	AZM5_14129_3061	AZM5_20519_1322	AZM5_26645_74	AZM5_108467_2
AZM5_16816_1556	AZM5_4774_4705	AZM5_10085_3	AZM5_90047_3	AZM5_87895_2516
AZM5_11772_2876	AZM5_45678_231	AZM5_10082_154	AZM5_5758_1	AZM5_87597_202
AZM5_2902_1751	AZM5_31088_1113	AZM5_9533_1879	AZM5_1476_1573	AZM5_9283_1650

Table 25 (continued).

AZM5_13219_1096
AZM5_86073_3740
AZM5_21054_947
AZM5_11794_3915
AZM5_49472_3
AZM5_3610_598
AZM5_7189_2253
AZM5_1896_1678
OGTAE49TV_1
AZM5_19866_194
AZM5_9281_5209
AZM5_150790_321
AZM5_3148_4779
AZM5_19529_2087
AZM5_92320_839
AZM5_27679_1858
AZM5_61647_463
AZM5_25048_513
AZM5_18931_941
AZM5_136868_636
AZM5_86778_689
AZM5_95643_3
AZM5_126118_308
AZM5_86532_1448
AZM5_89954_1570
AZM5_92087_68
AZM5_97522_268
AZM5_96057_1378
AZM5_102299_82
AZM5_17148_161
AZM5_88913_3
AZM5_18484_1206
AZM5_86841_2449
AZM5_997_8487
AZM5_14216_1727
AZM5_108905_112
AZM5_3339_3642
AZM5_15998_3551
AZM5_31425_583
AZM5_93098_775
AZM5_15342_890
AZM5_19474_2230
AZM5_3945_3284
AZM5_15446_220
AZM5_90939_2
AZM5_100630_639
AZM5_90201_2493
AZM5_10372_2029
AZM5_90817_1177
AZM5_11452_3857
AZM5_19768_396
AZM5_34230_377
AZM5_89265_1569
AZM5_23045_1272
AZM5_99466_3
AZM5_12166_2
AZM5_16738_1178
AZM5_36807_1
AZM5_94301_1579
AZM5_13722_2672
AZM5_4056_2
AZM5_11207_5296
AZM5_28534_1
AZM5_98743_573
AZM5_11503_229
AZM5_1229_6582
AZM5_7599_2238
AZM5_84788_1952
AZM5_93677_1097
AZM5_15078_2226
AZM5_1230_3517
AZM5_89200_1843
AZM5_114158_37
AZM5_100765_187
AZM5_91061_588
AZM5_104148_446
AZM5_4390_1523
AZM5_20119_121
AZM5_4175_952
AZM5_9896_1837
AZM5_99965_915
AZM5_20191_1854
AZM5_32350_596
AZM5_6582_2502
AZM5_84555_1743
AZM5_3603_5952
AZM5_87597_1667
AZM5_661_1108
AZM5_7371_531
AZM5_10631_2947
AZM5_656_2948
AZM5_105507_115
AZM5_4918_638
AZM5_59564_468
AZM5_115404_19
AZM5_86355_1
AZM5_31997_18
AZM5_11674_4498
AZM5_3966_4400
AZM5_94372_1333
AZM5_19768_1619
AZM5_64752_319
AZM5_12347_2085
AZM5_10575_311
AZM5_94852_603
AZM5_7333_2423
AZM5_17149_131
AZM5_89174_639
AZM5_14520_907
AZM5_84610_1155
OGVCF63TV_29
AZM5_1896_637

Table 26. The 48 maize sequences with 7 TM regions and external N-terminal predicted by ST-method.

AZM5_92120_754
AZM5_27164_857
AZM5_12167_1352
AZM5_86216_1119
AZM5_6155_1860
AZM5_20551_1281
AZM5_107468_612
AZM5_108454_651
AZM5_92642_3
AZM5_21945_1
AZM5_15626_2167
AZM5_16510_2456
AZM5_14902_1563
AZM5_86616_342
AZM5_91015_1431
AZM5_19332_2
AZM5_105945_420
AZM5_4480_3814
AZM5_19832_1144
AZM5_18427_261
AZM5_89680_142
AZM5_16068_1827
AZM5_21478_1179
AZM5_20341_453
AZM5_15445_125
AZM5_13841_1065
AZM5_1474_1408
AZM5_12385_632
AZM5_25277_1153
AZM5_10621_3540
AZM5_13416_385
AZM5_18894_1020
AZM5_102010_688
AZM5_13558_35
AZM5_86477_346
AZM5_21668_2179
AZM5_136700_1
AZM5_26182_3
AZM5_29331_245
AZM5_93253_1877
AZM5_3974_1355
PUIKX69TB_1
AZM5_27695_379
AZM5_13741_4394
AZM5_5305_391
AZM5_17726_1245
AZM5_5170_86
AZM5_637_3590

APPENDIX FIGURES

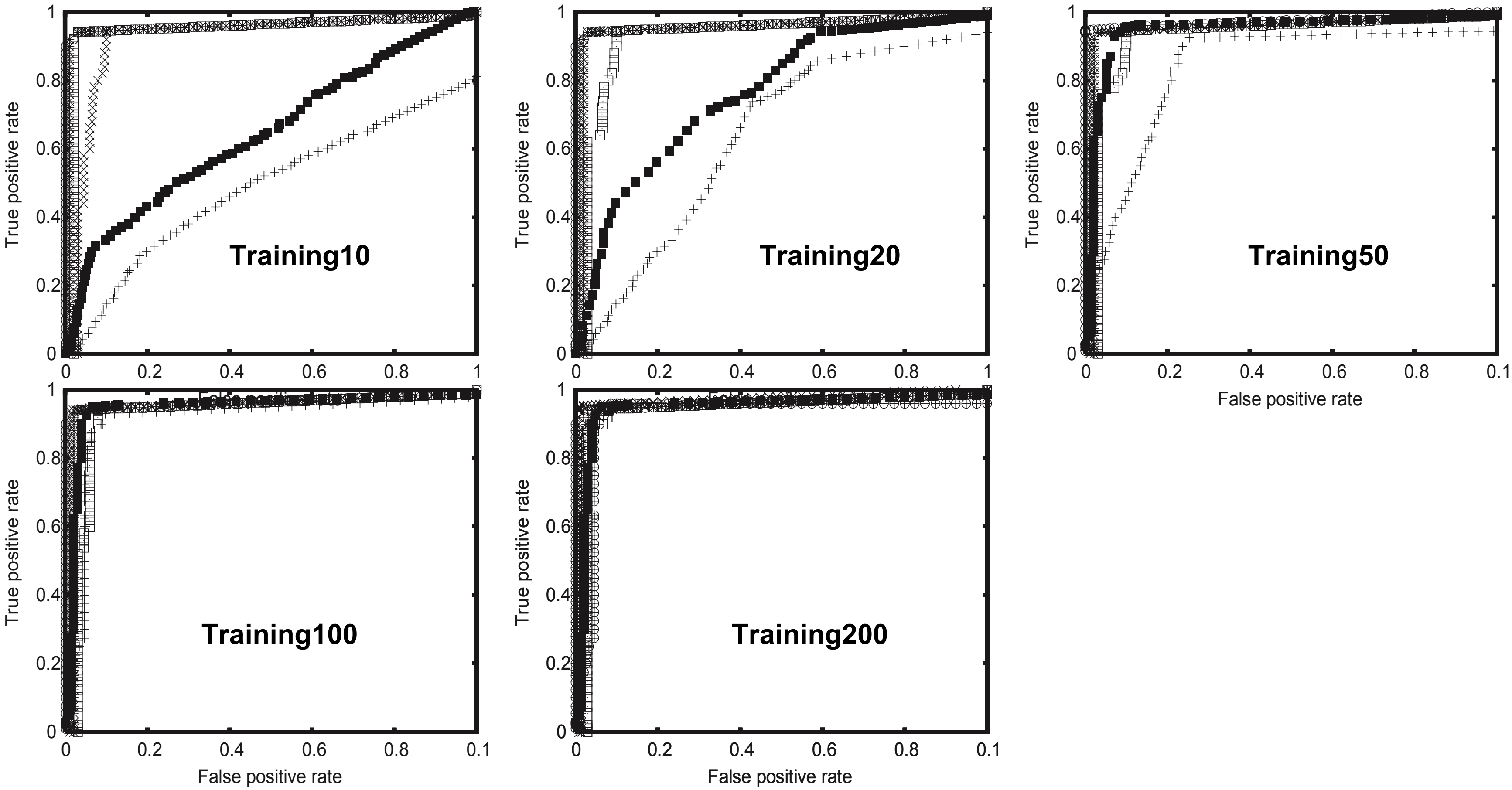


Figure 1 (continued). ROC graphs for classifiers using different sizes of training datasets for the first replication. Classifiers are shown as follows: PLS-ACC (open circle), PLS-Mean (**X**), PLS-AA (*****), PLS-AA_PAC (open square), SAM (filled square), and PSI-BLAST (+).

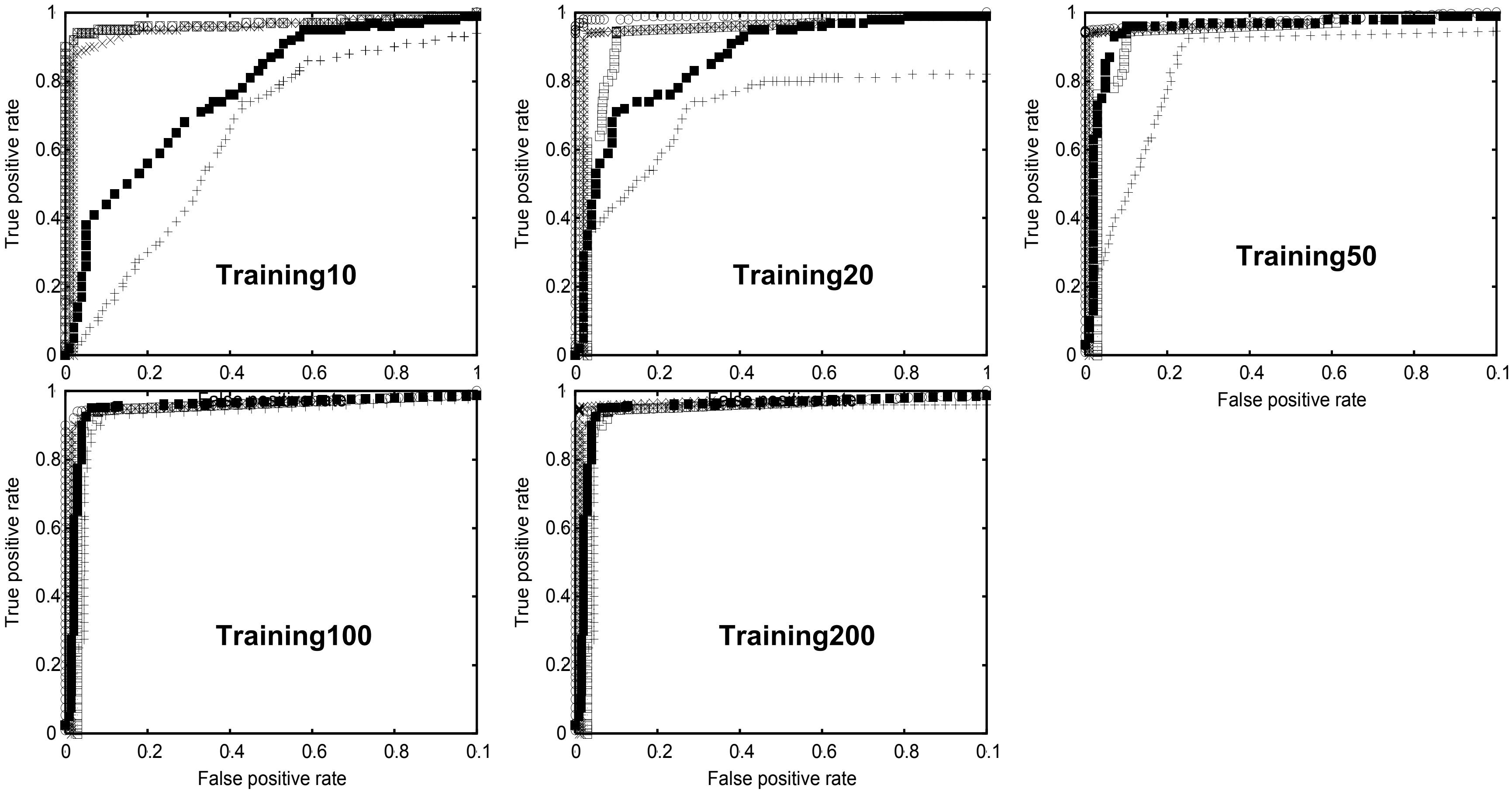


Figure 1 (continued). ROC graphs for classifiers using different sizes of training datasets for the second replication. Classifiers are shown as follows: PLS-ACC (open circle), PLS-Mean (**X**), PLS-AA (*****), PLS-AA_PAC (open square), SAM (filled square), and PSI-BLAST (+).

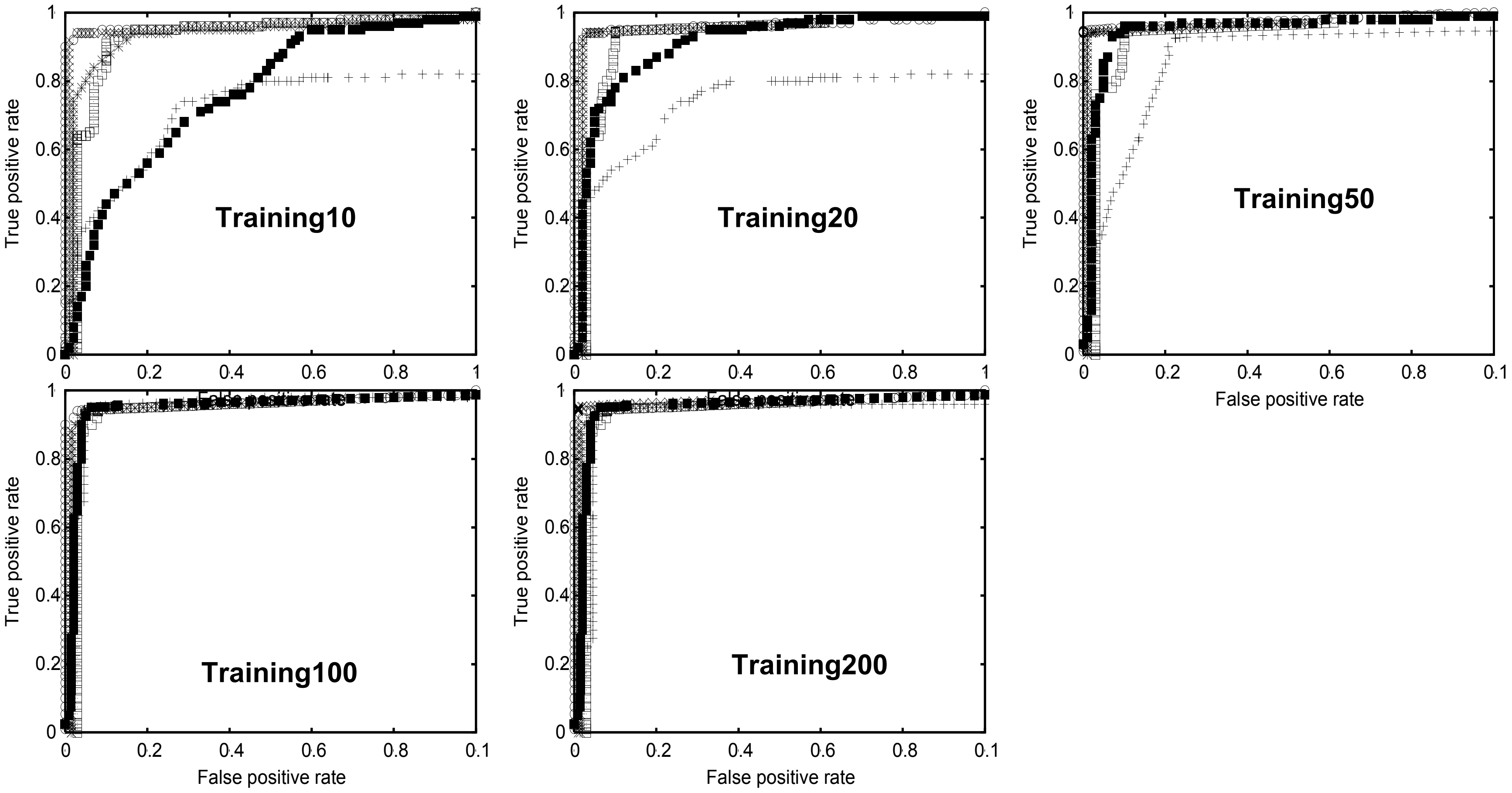


Figure 1 (continued). ROC graphs for classifiers using different sizes of training datasets for the third replication. Classifiers are shown as follows: PLS-ACC (open circle), PLS-Mean (**X**), PLS-AA (*****), PLS-AA_PAC (open square), SAM (filled square), and PSI-BLAST (+).

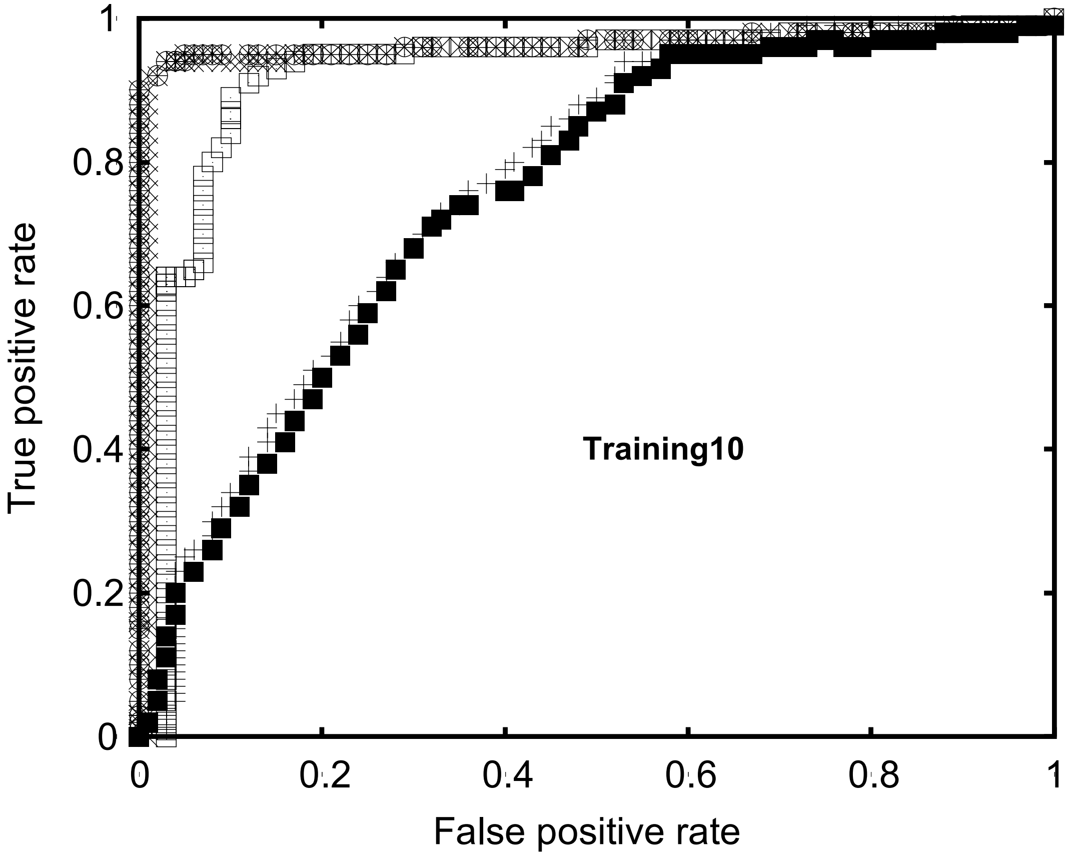


Figure 1 (continued). ROC graphs for classifiers using Training10 dataset for the fourth replication. Classifier are shown as follows: PLS-ACC (open circle), PLS-Mean (**X**), PLS-AA (*), PLS-AA_PAC (open square), SAM (filled square), and PSI-BLAST (+).

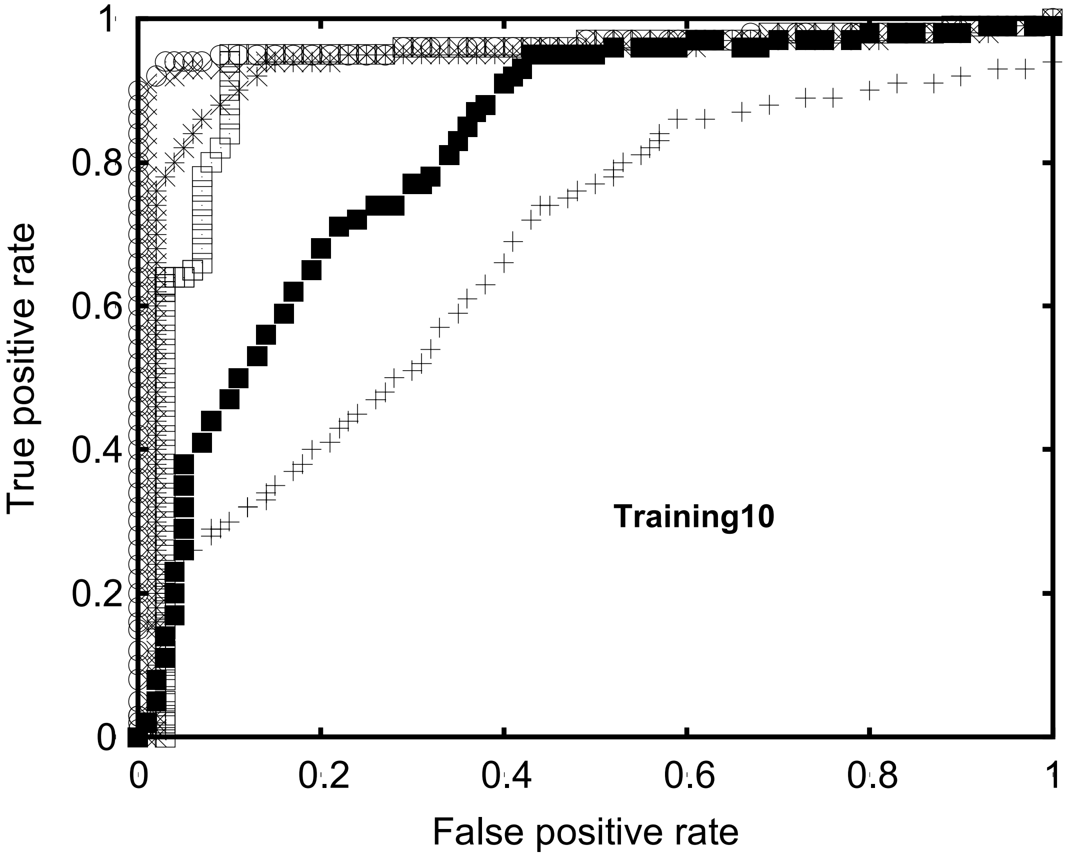


Figure 1 (continued). ROC graphs for classifiers using Training10 dataset for the fifth replication. Classifier are shown as follows: PLS-ACC (open circle), PLS-Mean (**X**), PLS-AA (*), PLS-AA_PAC (open square), SAM (filled square), and PSI-BLAST (+).

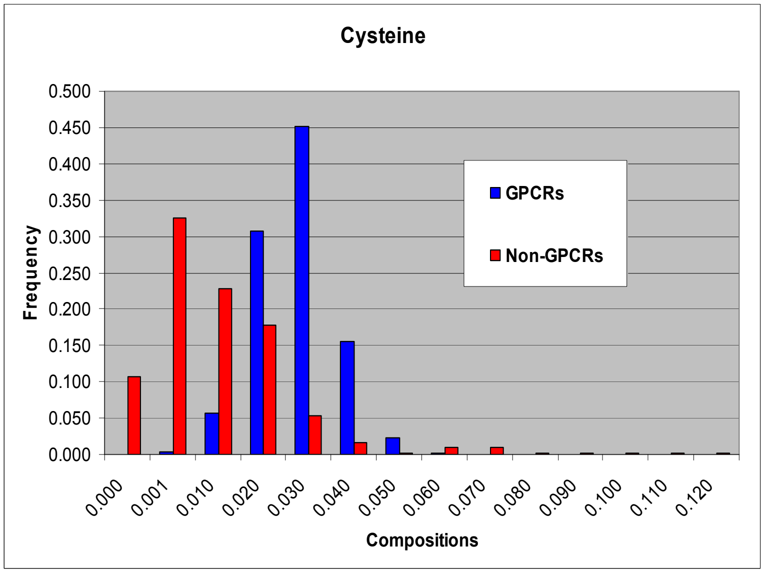


Figure 2. Histogram of ten amino acid compositions comparing between GPCRs and non-GPCRs.

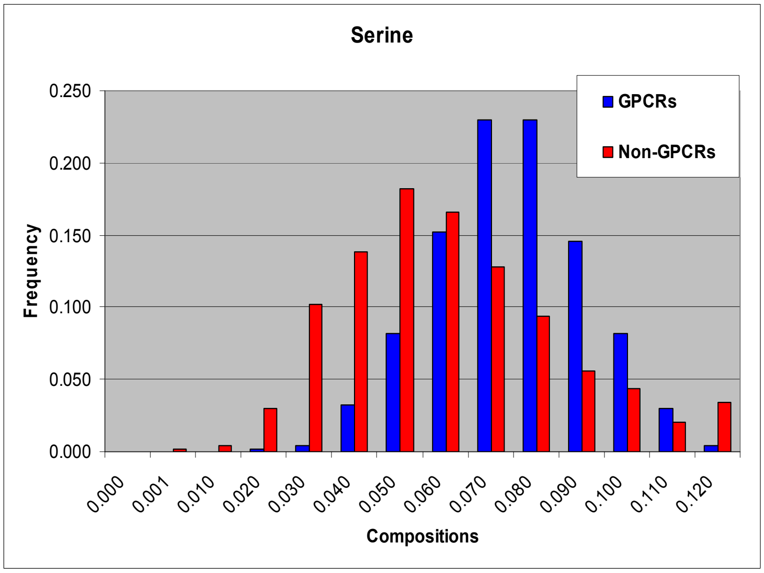


Figure 2 (continued)

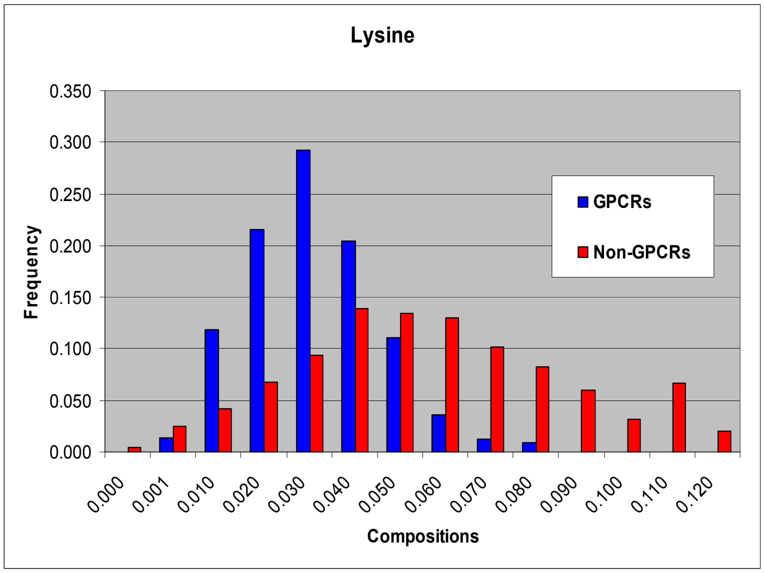


Figure 2 (continued).

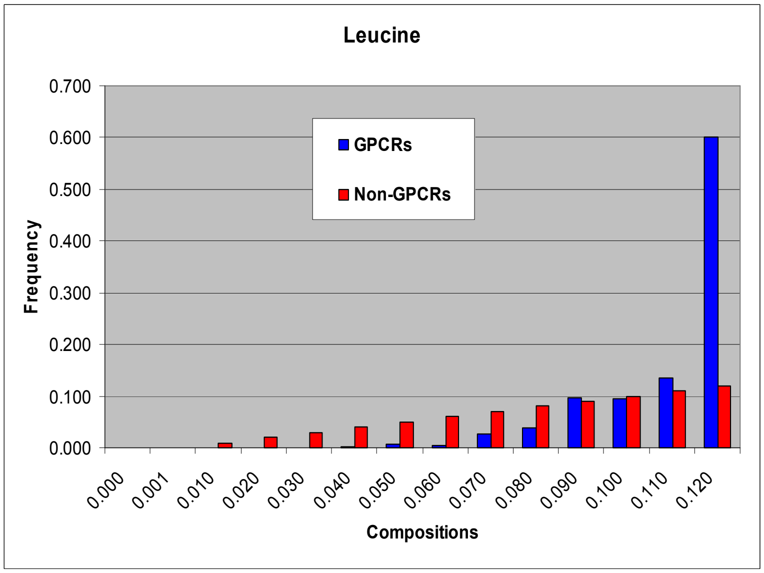


Figure 2 (continued).

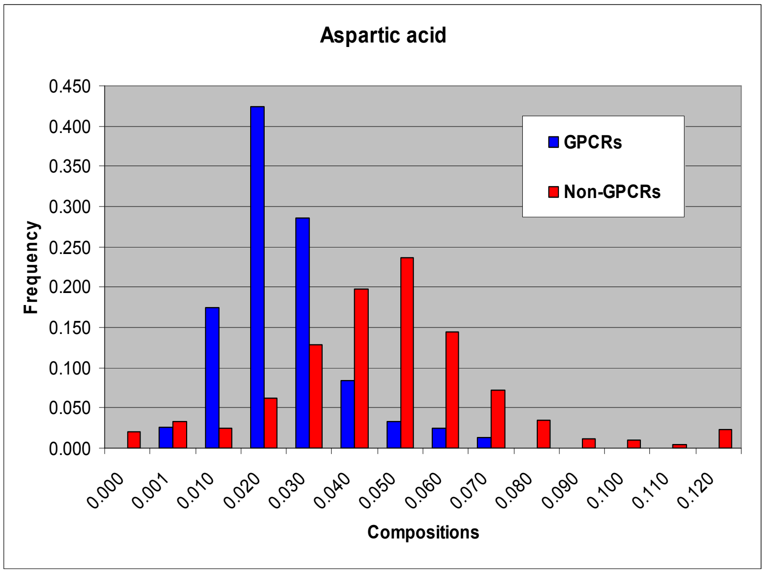


Figure 2 (continued).

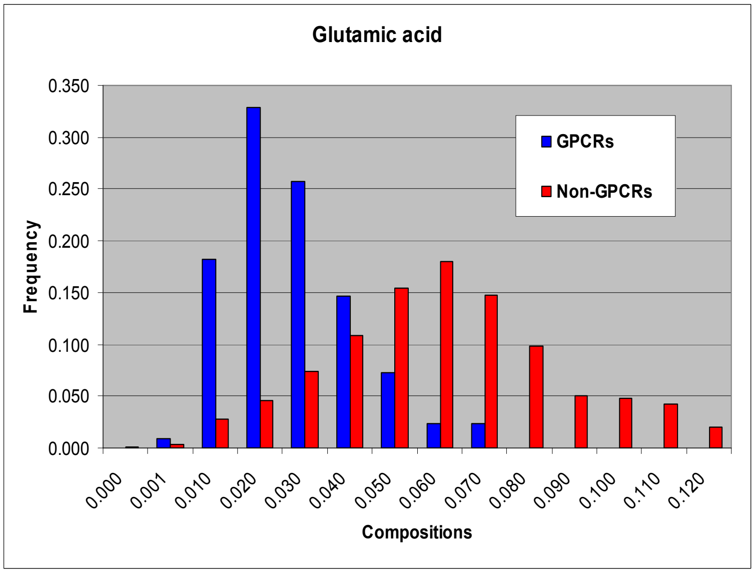


Figure 2 (continued).

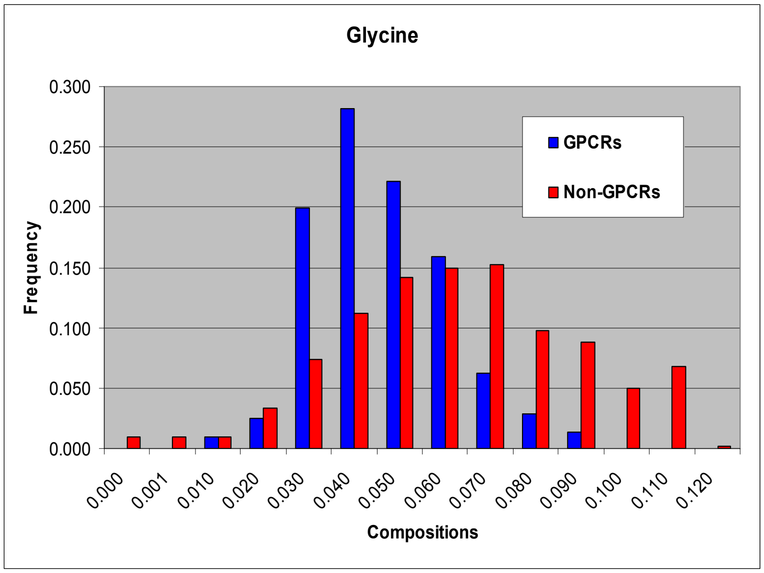


Figure 2 (continued).

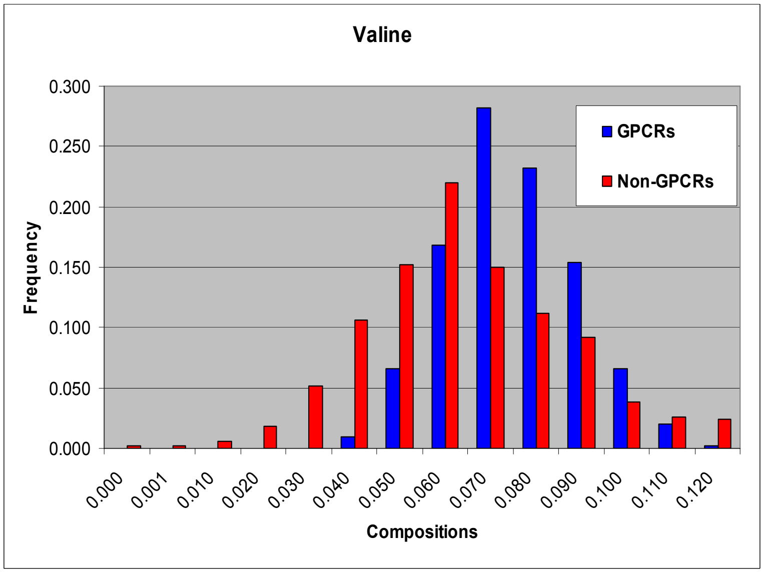


Figure 2 (continued).

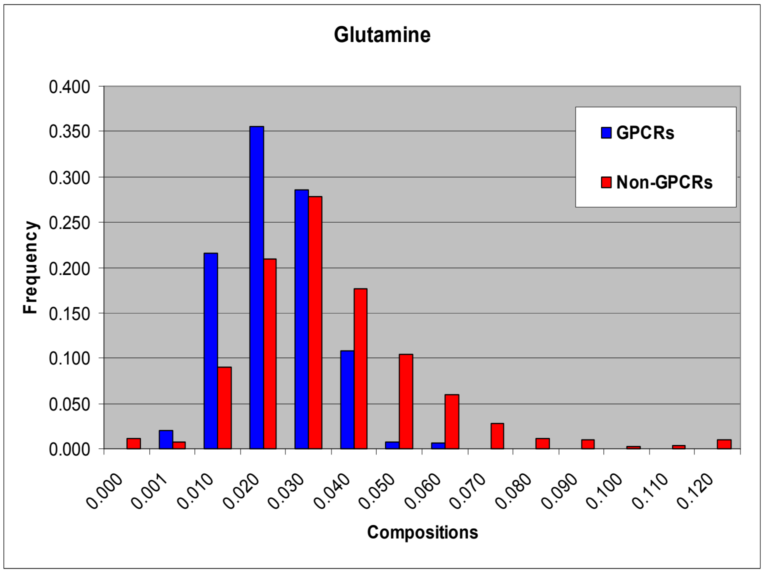


Figure 2 (continued).

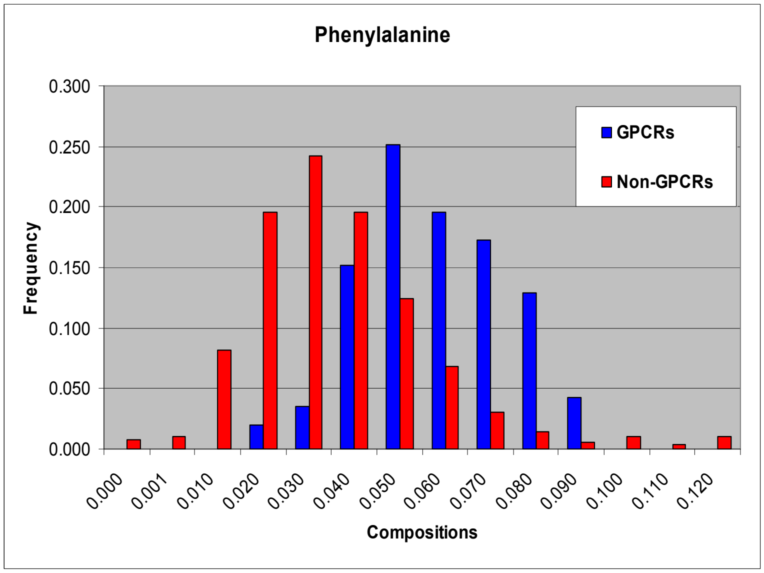


Figure 2 (continued).

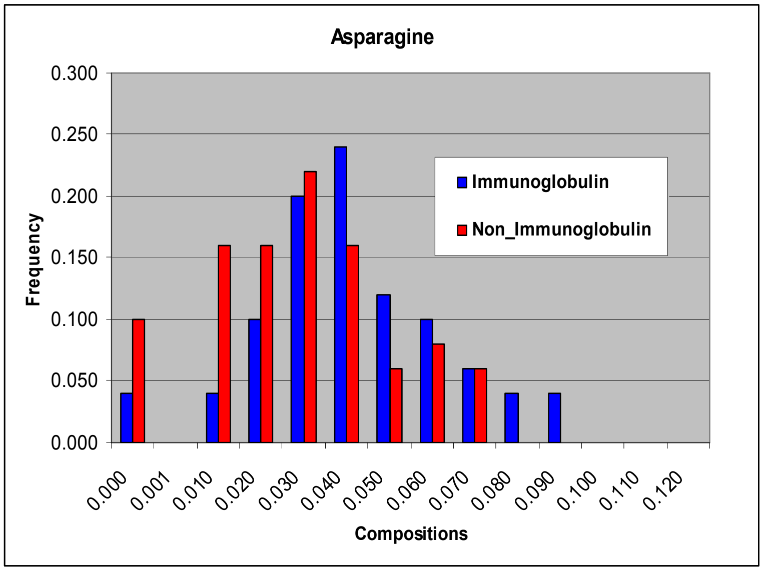


Figure 3. Histogram of 7 amino acids compositions comparing between 90 immunoglobulin and 90 non-immunoglobulin proteins.

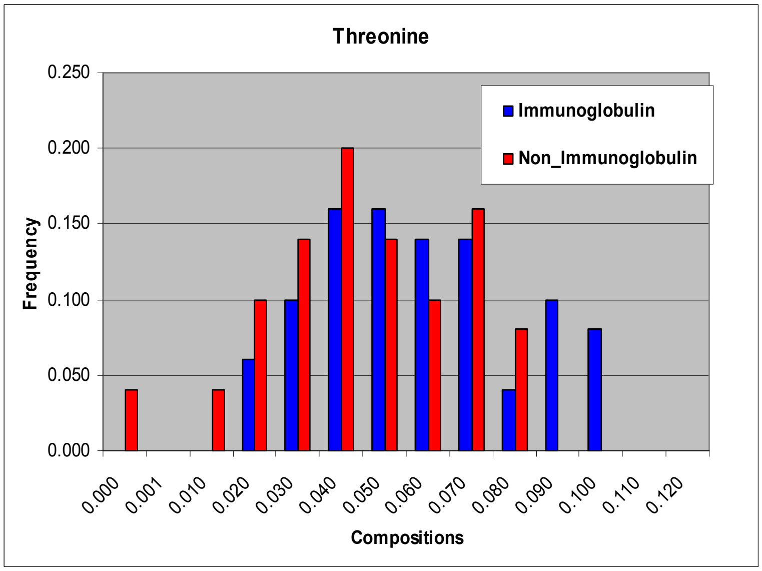


Figure 3 (continued).

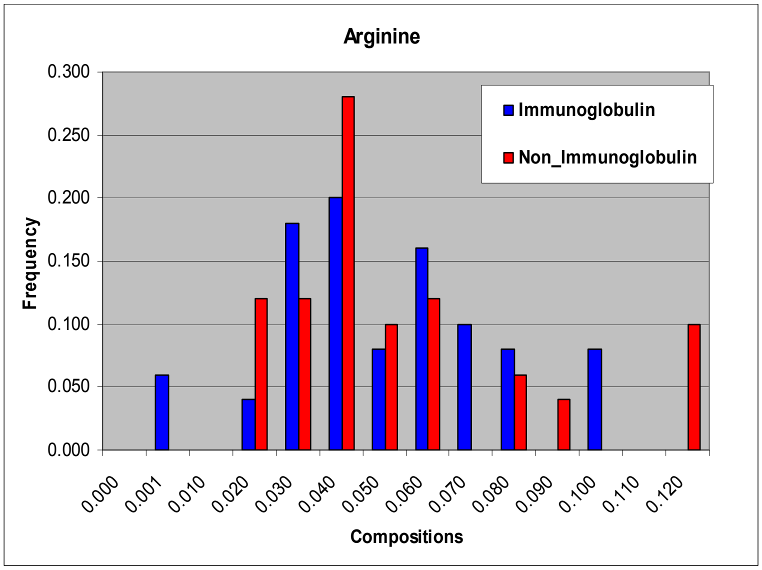


Figure 3 (continued).

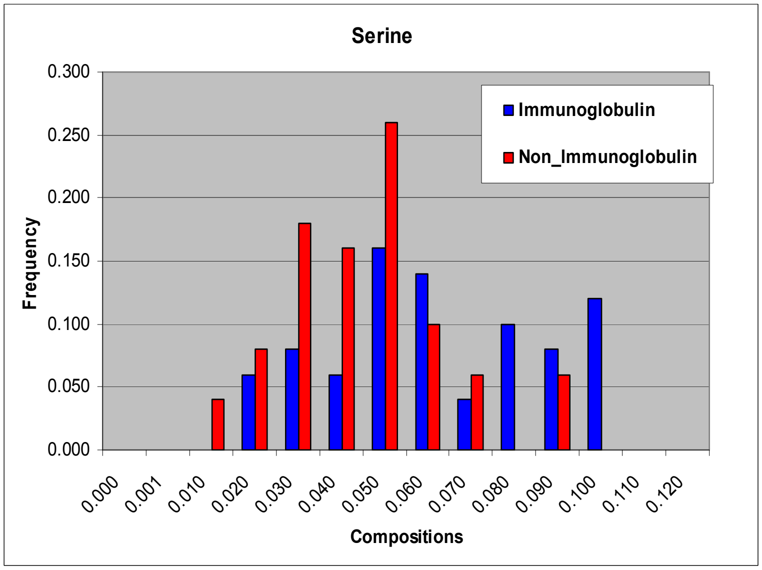


Figure 3 (continued).

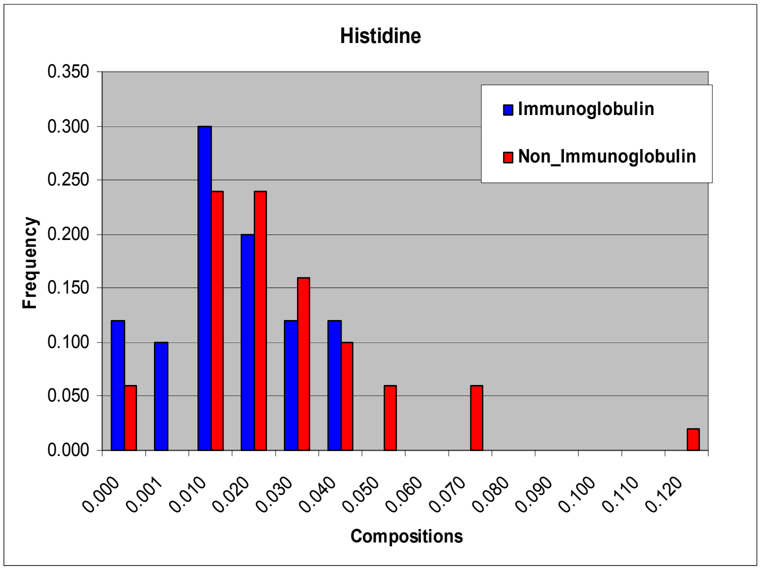


Figure 3 (continued).

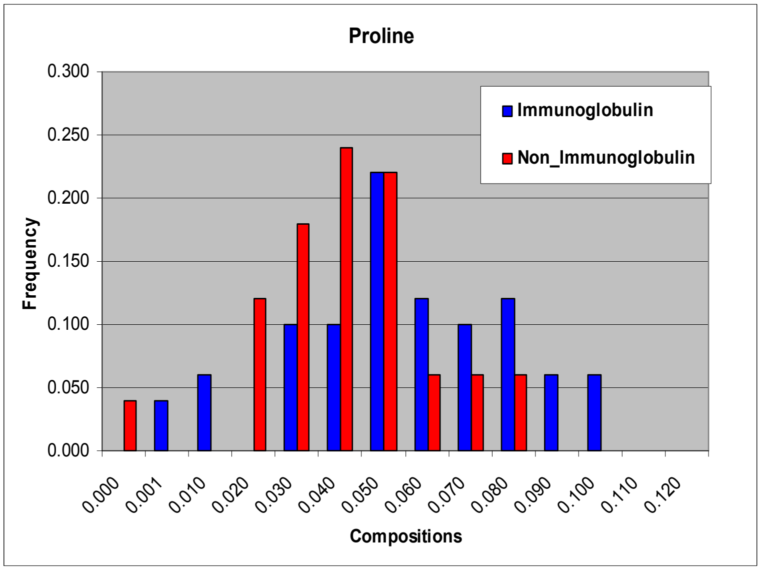


Figure 3 (continued).

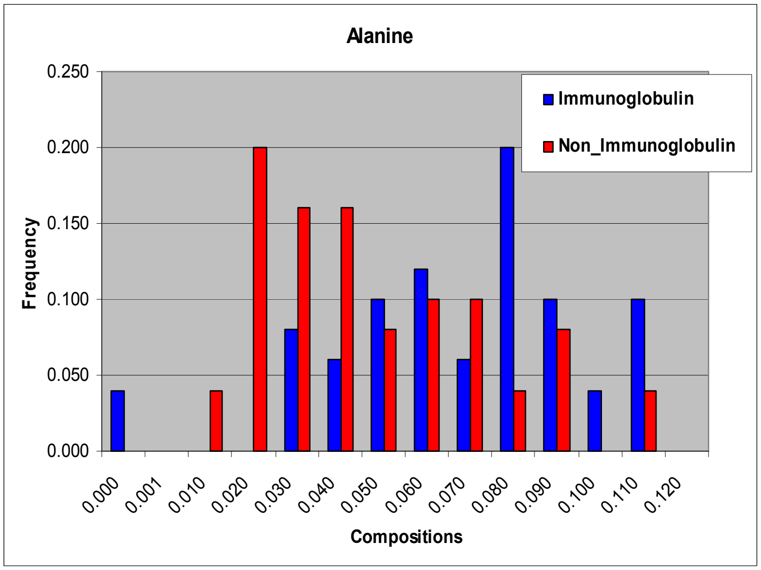


Figure 3 (continued).