The final publication is available at IOS Press through http://content.iospress.com/articles/in-silico-biology/isb00353

# Vector – G: Multi-Modular SVM-Based Heterotrimeric G-Protein Prediction

Preti Jain[1], Puneet Wadhwa[2], Ramazan Aygun*[2], Gopi Podila[1]

[1]Department of Biological Sciences, University of Alabama in Huntsville, Huntsville AL, 35899, USA

[2]Computer Science Department, University of Alabama in Huntsville, Huntsville AL, 35899, USA

*To whom correspondence should be addressed:
Department of Computer Sciences,
Technology Hall N360
University of Alabama in Huntsville,
Huntsville AL, 35899, USA
E-mail: RAYGUN@CS.UAH.EDU

**ABSTRACT**

Heterotrimeric G proteins interact with G protein-coupled receptors in response to stimulation by hormones, neurotransmitters, chemokines, and sensory signals to intracellular signaling cascades. Recently reported studies indicate G protein subunits play a significant role in different eukaryotic diseases including inflammation, neurological diseases, cardiovascular diseases, endocrine disorders as well as plant pathogen response, infectious hyphae growth, differentiation and virulence of pathogenic fungi. Thus a study of their functions, signaling pathways, and protein interactions may lead to the development of various preventive approaches. The diversity of alpha, beta and gamma subunits of G proteins necessitates a need for the prediction algorithm that helps in the identification of new proteins such as G beta where WD-40 repeats are not well characterized. The currently available techniques for finding G proteins are homology based search analyses and wet lab experiments, which are not very effective in finding new classes of proteins. We present here a robust computational method for finding new G proteins and their homologs using SVM based pattern recognition algorithm. Several physicochemical and compositional properties including dipeptide, tripeptide and hydrophobicity composition are used for generating the SVM classifiers. This method has 96.17%, 95.38%, 97.6% sensitivity and 99.45%, 100%, 100% specificity on test sets for G protein alpha, beta, and gamma subunits, respectively. This algorithm correctly predicts the known alpha, beta and gamma subunits reported in literature. One important contribution of this algorithm is, it helps in improving genome annotation of several proteins as G proteins and serves as a useful tool for comparative genomic analysis of G proteins. Using this method, novel G protein subunits are predicted in 31 genomes covering plant, fungi and animal kingdom.

**Availability:** The software is available at the website
http://biomine.cs.uah.edu/bioinformatics/svm_prog/scripts/GProteins/vectorg.html

**Contact: raygun@cs.uah.edu**

**Supplementary files:** The supplementary files are available at In Silico Biology online.

Keywords: Heterotrimeric G proteins, SVM, compositional properties, signal transduction.

## INTRODUCTION

GTP-binding proteins (G-proteins) are key mediators of many important cellular functions such as signal transduction, cell cycle control, vesicle trafficking pathways, nucleo-cytoplasmic transport, phagocytosis, and cell migration. G protein subunits are involved in the coupling of a variety of cell surface receptors to different intracellular signaling pathways. G proteins may possibly exert cellular functions other than acting as signaling transducers (Melien, 2006). Heterotrimeric G proteins have three subunits: G alpha, G beta and G gamma. Alpha subunits possess an intrinsic GTPase activity, which enables them to act as time switches. Hydrolysis of the bound GTP to GDP promotes the re-association of alpha subunit with beta-gamma dimer and renders the G protein in an inactive form.

Investigations into the G protein signaling have revealed their association with eukaryotic diseases (Hauge, *et. al.,* 2006), growth (Wang *et. al.,* 2006; Chen *et. al.*, 2006), development (Wang, *et. al* 2006), pathogenesis (Prados-Rosales *et. al*., 2006; Yamagishi *et. al*., 2006; Sarah and Assmann, 2005; Trusov, *et. al* 2006.), fungal mating (Kawasaki, *et. al*., 2005), and nod factor signaling in legume hosts by Rhizobium bacteria (Kelly, and Irving, 2003). Many complex human diseases have roots in the malfunction of G proteins. For example, Pseudohypoparathyroidism Type I is a disease that results from a defective G alpha (Carter, *et. al*., 1987), which results Albright's hereditary osteodystrophy (AHO) (Ahrens and Hiort, 2006). G alpha has oncogenic potentials, leading to the development of human pituitary, endocrine, Leydig cell, Ovarian and adrenocortical tumors (Farfel *et. al*., 1999). G beta3 subunit is related to susceptibility to essential hypertension and inter-individual variation in blood pressure (Hegele *et. al* 1998). G alpha3 in *Botrytis cinerea* is required for proper plant host surface recognition and penetration ability of germinated conidia (Doehlemann *et. al*., 2006). New approaches to drug development focus on targeting these G proteins to abrogate associated diseases. For example, Suramin drug disrupts receptor-G protein coupling by blocking association of G protein alpha and beta gamma subunits (Chung and Kermode, 2005). Anticancer activity of BIM-46174, an inhibitor of the heterotrimeric G alpha/G beta-gamma protein complex (Prevost, *et. al*., 2006), is another example for the need to study G protein and identification of new classes.

Experimental identification of G proteins is an arduous task. Methods like expression profiling, are used to identify genes involved in cellular processes, helps in observing only transcriptional events. Computational algorithms based on homology search can aid in identification, but this procedure has limitations when there is no characterized homologue available. For example, WD-40 repeats are present

in more than 30 protein families and are poorly characterized. G Beta also has 7 WD-40 repeats, this makes this class of heterotrimeric G proteins difficult to characterize as G beta subunits using homology based methods. Sequence similarity methods have limitations to classify these proteins. In order to overcome these limitations, machine learning methods have been used. Support Vector Machines classification method is better in some aspects than simple BLAST (Altschul *et. al*., 1997) or Hidden Markov Model (HMM)-based methods (Karchin *et. al*., 2002). and Artificial neural networks (ANN) as it effectively handles noise, large datasets/input spaces (Zavaljevski et al., 2002) and structural risk minimization principle (Zhao *et al* 2006). Since structure and function of a protein are determined by its preference for certain amino acids [Ofran and Margalit, 2006], compositional properties of protein sequences may provide an enhanced way of functional analysis. Several computational methods have been developed over the past decade using compositional properties like pTARGET enables prediction of nine distinct protein subcellular localizations in eukaryotic non-plant species (Guda and Subramaniam, 2005), POPPs for clustering proteins into families based on peptide composition (Wise 2002), SPAAN for prediction of adhesin proteins in bacteria (Sachdeva *et. al*., 2005), PROFEAT for computing commonly-used structural and physicochemical features of proteins and peptides from amino acid sequence (Li *et. al*., 2006), pSLIP uses multiple physicochemical properties of amino acids to predict protein subcellular localization in eukaryotes across six different locations, namely, chloroplast, cytoplasmic, extracellular, mitochondrial, nuclear and plasma membrane (Sarda *et. al*., 2005) and many more for the prediction of protein function, structure, and localization.

Pattern recognition algorithm based on statistical features is an important tool in assigning biological function of proteins. The fraction of experimentally analyzed proteins and the information on the function is limited for most of the proteins. A study for the identification of G protein subunits using a pattern search tool and BLAST by gpDB (G proteins/GPCRs relational database) (Elefsinioti *et. al*., 2004) is limited to amino acid sequence. In Vector-G, we adopted multi modular approach using dipeptide, tripeptide and hydrophobicity composition. In the present work, SVM classifier based on the sequence features has been trained to predict G proteins subunits in higher and lower eukaryotes. The protein data sets used for training in the present study are G alpha, beta and gamma as reported in the literature and it covered a wide representation of eukaryotic organisms. The sequence attributes like dipeptide, tripeptide and hydrophobicity composition are used for G protein subclass prediction in a wide spectrum of species. Genomes of very diverse organisms including human, plants, and newly sequenced fungal genomes having different habitats like saprophytic (*Uncinocarpus reesii*), pathogenic (*C. neoformans, M. grisea,*

*B. cinerea*) and symbiotic (*L. bicolor)* are used for testing and prediction of G protein subunits. These diverse organisms show conserved principle of cell signaling and pathogenesis and properties that make them unique. For example, *C. neoformans* virulence is controlled by a nutrient sensing pathway related to *S. cerevisiae* filamentation cascade. This pathway involves a heterotrimeric G protein which signals via adenyl cyclase (Pan et al., 2000; Wang et al., 2000). The SVM-based algorithm developed in the present study will serve as an essential tool for the large-scale genomics initiatives where high-throughput methods are needed to annotate and identify protein functions based on sequence information especially for genomes that are not annotated or poorly annotated.

## SYSTEMS AND METHODS

### Extraction of Features

The SVM classifiers are trained based on three different physicochemical properties: dipeptide frequencies, tripeptide frequencies, and hydrophobic composition. The advantage of dipeptide and tripeptide composition over single amino acid composition is that it encapsulates information about the fraction of amino acids as well as their local order (Kumar *et. al*., 2006).

*Dipeptide frequencies (DF)*

The total number of amino-acids is 20 and the total possible dipeptides are 400. A matrix of 400 dipeptides of each protein sequence is generated and fed as an input to SVM. If any dipeptide is not present in a sequence, its feature value is not used for training. The frequency of each dipeptide is calculated by the following formula:

$$DF_{ij} = \frac{N_{ij}}{N}$$

$N_{ij}$ = total number of *ij*-th dipeptide

N = Total number of possible dipeptides,

where *i, j* = 1–20.

*Tripeptide frequencies (TF)*

The total number of possible tripeptides is 8000. The training method is similar to dipeptide module. The frequency of each tripeptide is calculated by the following formula:

$$TF_{ijk} = \frac{N_{ijk}}{N}$$

$N_{ijk}$ = Total number of *ijk*-th tripeptide

N = Total number of possible tripeptides,

where *i, j, k* = 1–20.

*Hydrophobic composition (HC)*

The arrangement of hydrophobic and hydrophilic amino acid residues in a protein plays an important role in protein folding, interaction with other molecules and catalytic mechanisms. Hydrophobic and hydrophilic amino acids positions in a protein sequence create different type of proteins with different amphiphilic features (Chou, 2005). To compute hydrophobicity, the amino acids were classified into five

groups based on their hydrophobicity scores: (–8 for K, E, D and R), (–4 for S, T, N and Q), (–2 for P and H), (+1 for A, G, Y, C and W) and (+2 for L, V, I, F and M) (Brendel *et al*., 1992). The final hydrophobicity composition is computed as:

$$HD_g = \frac{N_g}{N}$$

$N_g$ = number of amino acids belonging to group 'g' in a protein sequence, where 'g' ranges from 1–5.
N = total number of amino acids in the protein.

We use moments to determine the distribution of amino acids in the sequence. The position of an amino acid is determined by the distance from the beginning of the protein. Moment describes the shape of distribution. First order of moment is mean and describes the central value. Second order of moment is variance which describes the dispersion. Third moment is skewness which describes the asymmetry. Forth order of moment is kurtosis which describes peakedness (Press 1992). Higher orders of moment are not very well described. We empirically tested our results on different order of moments and we used 2-5 order of moment of position of amino acids.

The order of moment of position is calculated as follows:

$M_{gr}$ = *r*-th order moment of positions of amino acids in group *g*, where *r* = 2–5.

$$M_{gr} = \sum_{i \in g} \left| \frac{(P_{gi} - \mu_g)^r}{N_g} \right|,$$

Where $P_{gi}$ = the position of i-th amino acid belonging to group *g*; N*g* is the total number of amino acids in group *g*, $\mu_g$ is the mean of all positions of amino acids of *g* group and computed as:

$$\mu_g = \sum_{i=1}^{Ng} \frac{P_{gi}}{N_g},.$$

A total of 25 inputs representing the hydrophobic composition of a protein were fed to the SVM.

**Database Creation**

*Positive Database*

Protein sequences were downloaded from NCBI Genbank http://www.ncbi.nlm.nih.gov with the keyword G proteins alpha subunit, beta subunit, and gamma subunit. Manual curation was done and sequences having keywords putative, hypothetical, patent, unknown, partial sequence, and gene product were removed.

*Negative Database*

Protein sequences with a keyword actin binding FH2 protein, ribosomal protein, acyl Co-A, alcohol dehydrogenase, Autophagy-related protein, cartilage matrix proteins, clusterin, cpn60, cullin, decarboxylase, flagellin, Cell division protein FtsA, helix-loop-helix proteins, tyrosine aminotranferase, hydroxylase, isomerase, kinase, oxidoreductase, rad24, replicase, spore coat protein, spore germination protein, ZipA were downloaded. Proteins were chosen from a wide-spectrum of organisms and protein families to include diverse representatives.

*Second level database for G beta proteins:*

The traditional classifiers cannot distinguish G beta proteins from other WD repeat proteins since G beta proteins contain WD repeats. A positive data set of known G beta proteins and a negative data set of WD repeat proteins were created to distinguish G beta proteins from WD repeats. The sequences were downloaded from http://www.ncbi.nlm.nih.gov. Unknown (un-annotated) WD repeat proteins were removed from negative data set.

*Training datasets for subunits:*

The positive training dataset for G alpha is comprised of only G alpha proteins, and negative training dataset for G alpha is created as the combination of G beta, G gamma and negative database. Similarly the databases of G beta and G gamma were created. Those sequences that have similarity more than 90% were removed using ClustalW (Thompson *et. al*., 1994). The total number of G proteins in positive database was 1250 out of which G alpha were 716, G beta were 120 and G gamma were 414. The total number of proteins in negative database was 1978. For second level database, the total G beta proteins were 120, and negative proteins (WD repeat proteins other than G beta) were 89.

## ALGORITHM

*Support Vector Machine*

SVM tool, $SVM^{Light}$, (Joachims, 1998) was used for the classification of G-Proteins. SVM classifier was generated using cost factor, model complexity, biased hyperplane and a kernel function. The linear, sigmoid and radial basis kernel functions were used. SVM is allowed a tradeoff (SVM light parameter 'C') between the training error and the margin in order to avoid model over-fitting (Tan *et al.,* 2006). Allowing a training error may allow larger margin. Larger margin avoids model over-fitting.

*Selection of Parameters*

In our experiments, biased and non-biased hyperplanes were used to provide flexibility to the kernel function. We have tested three kernel functions with the gamma parameter in the range [1, 3000], the cost factor ranges in [0.01, 2000] and tradeoff between training error and margin lies in [0.1, 51]. We have tested over 2000 combinations of these parameters.

*Evaluation of the Classifier*

Leave-One-Out (LOO) method was used to test the correctness of each model. In LOO, if there are N data samples, 1 data sample is separated and SVM is trained based on the remaining N-1 training data samples. A classifier is generated for each data sample. The classifier trained for item *i* is $SVM_i$. Each data sample *i* is classified with respect to its classifier $SVM_i$. The final SVM output is calculated based on precision and recall of each module and decision value for data samples by the classifier. LOO is equivalent to K-Cross validation where K becomes equal to N.

*Selection of the Classifier*

The best classifier is chosen based on its LOO performance. For some data sets, it is possible to get the same precision and recall values. In case of equality, we have considered the error of ζα-estimate. ζα-estimate gives a pessimistic error bound on the classifier. The risk of overfitting is balanced with the error of ζα-estimate.

*Performance Measures*

For each class, a SVM classifier is trained for each module, precision (positive predictive value; Pr) and sensitivity (recall; Sn) values are utilized to estimate the correctness and confidence on the result. $Pr_{DF}$, $Pr_{TF}$, and $Pr_{HC}$ represent the precision values of the selected classifiers for modules DF, TF, and HC, respectively. In the same way, $Sn_{DF}$, $Sn_{TF}$, and $Sn_{HC}$ represent the recall values of the selected classifiers for modules DF, TF, and HC, respectively. PrDF and SnDF were used to calculate F measure for DF module. In the same way PrTF, SnTF, and PrHC, SnHC were used for calculating F measure for TF and HC module respectively.

$$F_i = \frac{2 * \mathrm{Pr}_i * Sn_i}{\mathrm{Pr}_i + Sn_i}$$

where i ϵ {DF, TF, HC}

The Sensitivity, Specificity and Precision are calculated as follows:

Recall or Sensitivity $S_n = \frac{TP}{TP + FN}$

Specificity $S_p = \frac{TN}{TN + FP}$

Precision = $\frac{TP}{TP + FP}$

where TP, TN, FN and FP stands for true positive, true negative, false negative and false positive respectively.

For each protein, the SVM classifier returns a decision value. If the decision value is positive, the protein is classified as a positive prediction otherwise as a negative prediction. Therefore, high positive decision values or low negative decision values indicate better confidence of membership to the corresponding class. The notation $v_i$ denotes the decision value obtained from classifier *i* where $i \in$ {DF, TF, HC}. In our experiments, zero is used as threshold. Since, all classifiers yielded good precision and recall on a training set. We tested higher threshold and we almost got same results in our experiments. The modules that yield poor results on LOO are eliminated by threshold τ during classification. Based on this threshold τ, each classifier is assigned a weight (w), such that whenever the trust of a classifier goes below τ the weight is 0 and whenever the trust takes the maximum value, the weight is 1.0. The weight of a classifier is computed as follows:

If $(F_i - \tau) < 0$ then $w_i = 0$
Else $w_i = (F_i - \tau)/ (100 - \tau)$

where $i \in$ {DF, TF, HC}.
In addition to rational weight values ranging in [0,1.0], the binary weight values provide information whether a classifier classifies as positive or not. The binary weight value is determined as:

if module i returns as positive
$m_i = 1$
else $m_i = 0$

where i$\in$ {DF, TF, HC}.

Based on these parameters, we have defined one metric to analyze the results: *dscore*.

*Decision Score (dscore):* The decision score determines how good the classification is. For positive class, high score values are desirable. The score ranges between 0 and $\infty$. The score is calculated as

$$dscore = \sum_{i=1}^{M} v_i * w_i$$

*Normalized score (nscore): dscore* values indicate distance from the margin. Since the *dscore* value might be too high or too low, it doesn't give good indication of results. So we mapped *dscore* into [-1, 1] interval. We used 4 constants here: 'x' as a threshold for the highest *dscore* for positive classes, 'z' as a threshold for the lowest *dscore* for negative classes, 'a' as a threshold for *nscore* for a protein highly likely to be in positive class, 'c' as a threshold for *nscore* for a protein highly likely to be in negative class. The *dscore* values that are above 'x' show high likelihood of being a member of the class. In the same way, *dscore* values that are less than 'z' belong to a negative class. *dscore* values above 'x' are mapped to [a, 1], whereas *dscore* values between '0' and 'x' are mapped to [0, a]. Negative *dscore* values less than 'z' are mapped to [-1, c], whereas negative *dscore* between 'z' and '0' are mapped to [-c, 0]. The max score is chosen as highest *dscore* value among test set values (a and c) chosen as '0.9', '-0.9' respectively, since values close to '1' reliably indicate good membership to the positive class and values close to '-1' reliably indicate membership to negative class. 'x' and 'z' are chosen as 1 and -1, respectively.

If *dscore* >x

*nscore* = min(1, a+(((1-a)*(*dscore* –x))/(maxscore-x))

*dscore* in [0,x]

*nscore* = (*dscore**a)/x

*dscore* in (z,0)

*nscore* = (*dscore**c)/z

*dscore* <z

*nscore* = max (-1, c- (1+c) *(z-*dscore*))/(z-minscore))

*Genome wide prediction*

The whole genome sequences of different organisms were downloaded from JGI eukaryote genomes website http://genome.jgi-psf.org/ and from Broad institute genome sequence data http://www.broad.mit.edu/annotation/fgi/. We covered 31 eukaryotic organisms for prediction. The detailed list of organisms and prediction results for G proteins are shown in supplementary table 1.0, 1.1, 1.2, 1.3.

## RESULTS AND DISCUSSION

*Performance*

The performance of Vector-G is tested by LOO (leave one out) method while training. Three separate test sets of 182 G alpha, 65 G beta and 250 G gamma sequences and same number of sequences in negative test set were analyzed further to see the performance of SVM predictions. The Vector-G software predicted with 96.17%, 95.38%, 97.6% sensitivity and 99.45%, 100%, 100% specificity for alpha, beta and gamma subunits respectively on a test set. The performance of three classes is shown in Fig 1-a, b, c. Except 7 proteins all 175 G alpha are predicted with >0.9 *nscore*. In G beta and G gamma, 3 and 6 proteins are predicted as false negatives. The sensitivity and specificity of dipep, tripep and hydrophobicity modules for each class are shown in Table1. The HC module gave low sensitivity and specificity than TF and DF module but it is still a critical parameter because hydrophobic and hydrophilic residues play a very important role in protein folding and interaction with other proteins and help in prediction. The TF module provides better results than DF and HC modules. The TF module provided almost excellent results when the similarities among proteins increase. Alpha class shares good similarity among the proteins (with respect to features or structure). Tripep module works very well for this class. Please see Table 1. We considered values from all modules for final prediction because that provided the best result.

BLAST tries to find a similar protein based on the homology. In the database, it is possible that there are multiple hits per protein in the test set. In such a case, even if one of the proteins is removed, another match is available for BLAST. Therefore, LOO is not an effective method of evaluation of BLAST in presence of multiple hits in the database. Instead of leaving one protein out, we removed several sets (or clusters) of proteins having similar homology level from the training set (see Fig 2). In such a case, we are able to measure and observe the performance of BLAST and our system. So we performed 3 experiments. G-alpha training set has 280 G-alpha proteins as a positive set and 1587 proteins as a negative set. The negative set was prepared with the same categories of proteins as mentioned in methods. G-beta training set has 67 G-beta proteins as a positive set and 1800 proteins as a negative set. G-gamma training

set has 88 G-gamma proteins and 1779 proteins as a negative set. In case of G-beta and G-gamma proteins, vector-G is able to predict some proteins which blast cannot.

Blast and SVM performance was compared on 3 test sets individually.

*Beta Test Set*

The SVM was trained again on a subset of training database after removing proteins that fall in the same cluster. In case of G beta subclass, there is a considerable similarity between G beta and other WD-40 repeat containing proteins. BLAST is not powerful to distinguish G beta from other WD-40 proteins. SVM has overcome this feature since it does not only depend upon sequence similarity. Out of 40 test set proteins, 8 proteins are not correctly predicted by Blast. Vector-G has predicted 4 proteins incorrectly. This experiment clearly indicates the 10% higher performance of SVM than Blast. One protein gi-72129013 was not correctly predicted by SVM but correctly predicted by Blast. 5 proteins that are correctly predicted by Vector-G (but incorrectly predicted by Blast) are Gi – 88766385, 1001939, 5174447, 77745452 and 30024660. Blast finds incorrect hits for these proteins with gi – 148655047, 66846416. Gi – 148655047 has unusually high number of WD repeats (10) which predicts that it is not a G-beta protein which generally has 7 WD repeats. Gi- 66846416 has 8 WD-repeats as predicted by Interpro (Quevillon *et. al*. 2005). There are some variations in the number of WD repeats in G-beta but the number of WD repeats usually varies from 5 to 7 repeats. Blast is not effective enough to distinguish two sequences when two sequences have high similarity. For detailed results please see supplementary file: beta_blast_comparision.xls

*Gamma Test Set*

In case of G gamma test set (48 proteins), blast is not able to find hits for 17 proteins. Out of these 17 proteins vector-G is able to predict 7 proteins correctly. Protein having gi- 55957436 was not correctly predicted by SVM but blast predicted this correctly. This indicates the 12.49% higher performance of SVM over the blast. For detailed results please see supplementary file: gamma_blast_comparision.xls

*Alpha Test Set*

In case of G alpha test set, the number of hits for each protein with training database was very high. Determining and removing clusters was more difficult than the one in beta test set. Our results are similar to blast. Out of 112 total proteins in test set, Blast could not find hits for 8 proteins. Vector-G was also not able to predict 8 proteins. So the performance of both the methods in alpha test set was same

(92.85%). It was difficult to test non homology based method on this class of proteins. For detailed results please see supplementary file: alpha_blast_comparision.xls

*Prediction of G Beta Proteins*

Defining the function of a WD-repeat protein is the current challenge (Li and Roberts 2001), since G beta subunits have WD-40 repeats. Vector-G was able to predict correctly G beta with high precision and is highly useful in improving the annotation of existing WD-40 repeat containing proteins as well as finding the new G beta proteins. The top scoring G beta proteins include AGB1 of *Arabidopsis thaliana*, beta-1, beta-2, beta-3, beta-4, beta-5 in *Homo sapiens,* Cblp in *Chlamydomonas*, Pigpb1 in *Phytopthora,* sfaD in *Aspergillus nidulance*, MGB1 in *Magnaporthe grisea,* gnb1 in *Neurospora crassa* and Fgb1 in *Fusarium oxysporium*, Gib-2 protein in human fungal pathogen *Cryptococcus neoformans*. Gib-2 in *Cryptococcus*, Vps15 (Slessareva *et.al*. 2006), Gpb1 and Gpb2 in *S. cerevisiae* (Harashima and Heitman 2002) are predicted recently showing G beta like function. These proteins are also predicted as G beta like by Vector-G.

Despite multiple G alpha subunits functioning in fungi, only single G beta species has been identified, suggesting that non-conventional signaling exist in eukaryotic organisms. An interesting finding emerging from this analysis is the identification of several new G beta proteins namely, 5 G beta in newly sequenced genome of mycorrhizal symbiotic fungus *Laccaria bicolor*, 1 in *Coccidioides immitis*, 1 in *Coprinus cinereus*, with high *nscore* value indicating that these proteins could have G beta like characteristics. The reason that *L. bicolor* has high number of G alpha, and G beta proteins, is the nature of its habitat. Since, it is a symbiotic fungus that interacts with a large number of plants as well as can live as a saprophyte, which suggest that it could have complex signal transduction mechanisms possibly involving many G alpha and G beta subunits. The complete list of newly predicted G proteins is listed in Supplementary file 1.1, 1.2, 1.3.

*Application of vector-G*

Although a sample run of Vector-G on a well defined test set of G protein subunits and non- G proteins showed high sensitivity and specificity, yet, the test set was small. Therefore, to assess general applicability of Vector-G, we examined its prediction power by analyzing well-characterized G protein subunits and protein coding regions from 31 genomes from a wide range of organisms including human, pathogenic, non-pathogenic and symbiotic organisms having a variety of functions. The results of the well

characterized G protein subunits and genome scan results are displayed in Table 2 and Supplementary Table 1.1, 1.2, 1.3 respectively. The experimentally characterized G protein subunits from a wide range of organisms top the list in whole genome scans in our experiments. Subsequently, we restricted our analysis to proteins having *nscore* greater than 0.5. Several of the predicted G protein subunits are supported by complementary evidence such as Prosite pattern search (Gattiker, *et. al.* 2002) and BLASTP. A fraction (~ 30% - 78% depending on the organism) of the top scoring predicted G protein alpha subunits by Vector-G also contain P-loop or ATP/GTP-binding site motif. Many of the WD-repeat proteins are characterized as G beta like protein. Similarly a fraction of top scoring G gamma proteins contain CAAX motif. The CAAX box has been found to be associated with G gamma, Ras, lamins and Rhodopsin kinase also. These results taken together support a significant fraction of the G protein subunits predicted by Vector-G in a wide range of organisms. In addition, Vector-G guided the improved annotation of a number of G proteins by suggesting re-examination of these proteins using available evidences.

It is clear that the highly studied organisms show a considerable number of G protein subunits but in case of lower eukaryotes and plants, a lot more G protein subunits need to be analyzed. This method also aided in identifying and analyzing G proteins in new genomes for which not much annotation and experimental data is available. Vector-G adopts multi-modular approach which is robust and comprehensive. The currently available methods that can be used for G Protein subclass prediction include BLAST, WD-repeat prediction (Smith *et.al*. 1999), gpdb (Elefsinioti *et. al*., 2004). However, they require a characterized homolog whereas Vector-G is a non-homology based method. We have applied our prediction method to more than 31 completely sequenced genomes from across the three domains of life. Some proteins are predicted with high *nscore* but did not show any available evidence. These could serve as leads for experimental testing especially in organisms where G proteins have not yet been identified or functionally characterized.

Since some false positives are also predicted because of small protein length for some genes in whole genome scans, it is judicious to consider length of the protein, long single amino acid repeats and available resources for further experimental work to validate these genes. Vector-G makes a valuable contribution in the group of computational approaches that use compositional properties for addressing biologically interesting issues such as identification of secretory proteins in bacteria (Schneider 1999) and apicoplast targeted proteins in the malarial parasite *Plasmodium falciparum* (Zuegge *et.al*. 2001). Application of Vector-G could rapidly aid in experimental characterization of several proteins with known and unknown predicted functional roles in investigating their role in the signal transduction process. Most

importantly Vector-G would be an invaluable tool for predicting G proteins from newly sequenced genomes where there is very little experimental data or annotation is available. It is also useful in the prediction of new classes of G proteins that are not easily identified by standard homology based methods. The modular nature of SVM methods developed here can be easily extended to other groups of proteins such as kinases, transcription factors etc.

**ACKNOWLEDGEMENTS**

Part of this work is supported by NSF grant MCB-0421326 to Dr. G. K. Podila.

**REFERENCES**

Ahrens,W. and Hiort,O. (2006) Determination of Gs alpha protein activity in Albright's hereditary osteodystrophy. *J Pediatr Endocrinol Metab*. 19, 647-651.

Alspaugh,J.A. et al. (1997) Cryptococcus neoformans mating and virulence are regulated by the G-protein alpha subunit GPA1 and cAMP. *Genes Dev*. 11, 3206-3217.

Altschul,S.F. et. al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389-3402.

Brendel,V. et. al. (1992) Methods and algorithms for statistical analysis of protein sequences. *Proc Natl Acad Sci U S A*. 89(6), 2002-6.

Carter,A. et. al. (1987) Reduced expression of multiple forms of the alpha subunit of the stimulatory GTP-binding protein in pseudohypoparathyroidism type Ia. *Proc. Natl. Acad. Sci. USA* 84, 7266– 7269.

Chen,J.G. et. al. (2006) Differential roles of Arabidopsis heterotrimeric G-protein subunits in modulating cell division in roots. *Plant Physiol*. 141, 887-897.

Chou,K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*. 21, 10-19.

Chung,W.C. and Kermode,J.C. (2005) Suramin disrupts receptor-G protein coupling by blocking association of G protein alpha and beta gamma subunits**.** *J Pharmacol Exp Ther*. 313, 191-198.

Doehlemann,G. et. al. (2006) Different signalling pathways involving a Galpha protein, cAMP and a MAP kinase control germination of Botrytis cinerea conidia. *Mol Microbiol*. 59, 821-835.

Dubeykovskiy,A. et al. (2006) Runx-dependent regulation of G-protein gamma3 expression in T-cells. *Cell. Immunol*. 240, 86-95.

Elefsinioti,A.L. et. al. (2004) A database for G proteins and their interaction with GPCRs. *BMC Bioinformatics*. 5, 208.

Farfel,Z. et. al. (1999) The expanding spectrum of G protein diseases. *N. Engl. J. Med.* 340, 1012–1020.

Ferris,J. et al. (2006) G(o) signaling is required for Drosophila associative learning. *Nat Neurosci*. 9, 1036-1040.

Fuse,N. et al. (2003) Heterotrimeric G proteins regulate daughter cell size asymmetry in Drosophila neuroblast divisions. *Curr Biol.* 13, 947-954.

Gattiker,A. et. al. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinformatics*. 1(2) 107-108.

Guda,C. and Subramaniam,S. (2005) pTARGET a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics*. 21, 3963-9.

Harashima,T. and Heitman,J. (2002). The Galpha protein Gpa2 controls yeast differentiation by interacting with kelch repeat proteins that mimic Gbeta subunits. *Mol Cell.* 10(1), 163-73.

Hauge Opdal,S. et. al. (2006) The G protein beta3 subunit 825C allele is associated with sudden infant death due to infection. *Acta Paediatr*. 95, 1129-1132.

Hegele,R.A. et. al. (1998) G protein beta3 subunit gene variant and blood pressure variation in Canadian Oji-Cree. *Hypertension*. 32, 688-692

Hossain,M.N. et al. (2006) G-protein gamma subunit GNG11 strongly regulates cellular senescence. *Biochem Biophys Res Commun*. 351, 645-650.

Ishimoto,H. et al. (2005) G-protein gamma subunit 1 is required for sugar reception in Drosophila. *EMBO J*. 24, 3259-3265.

Jansen,G. et al. (2002) The G-protein gamma subunit gpc-1 of the nematode C.elegans is involved intaste adaptation. *EMBO J*. 21, 986-994.

Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf, B., Burges,C. and Smola, A. (eds), *Advances in Kernal Methods—Support Vector Learning*. MIT Press, Cambridge, MA, pp. 42–56.

Karchin,R. et. al. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*. 18, 147-159.

Kawasaki,L. et. al. (2005) The Gbeta(KlSte4p) subunit of the heterotrimeric G protein has a positive and essential role in the induction of mating in the yeast Kluyveromyces lactis. *Yeast*. 22, 947-956.

Kays,A.M. et al. (2000) Regulation of conidiation and adenylyl cyclase levels by the Galpha protein GNA-3 in Neurospora crassa. *Mol Cell Biol.* 20, 7693-705.

Kelly,M.N. and Irving,HR. (2003) Nod factors activate both heterotrimeric and monomeric G-proteins in Vigna unguiculata (L.) Walp. *Planta*. 216, 674-685.

Krispel,C.M., et al. (2003) Prolonged photoresponses and defective adaptation in rods of Gbeta5-/- mice *J. Neurosci.* 23, 6965-6971.

Krystofova,S. and Borkovich,K.A. (2005) The heterotrimeric G-protein subunits GNG-1 and GNB-1 form a Gbetagamma dimer required for normal female fertility, asexual development, and galpha protein levels in Neurospora crassa. *Eukaryot Cell.* 4, 365-378.

Kumar,M. et. al. (2006) Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *J Biol Chem*. 281, 5357-5363.

Li,D. and Roberts,R. (2001) WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. *Cell Mol Life Sci*. 58, 2085-2097.

Li,ZR. et. al. (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* 34, W32-37.

Matsuki,M. et al. (2006) Goalpha regulates olfactory adaptation by antagonizing Gqalpha-DAG signaling in Caenorhabditis elegans. *Proc Natl Acad Sci U S A.* 103, 1112-1117.

Ray,K. and Robishaw,J.D. (1994) Cloning and sequencing of a rat heart cDNA encoding a G-protein beta subunit related to the human retinal beta 3 subunit. *Gene* 149 (2), 337-340.

Melien,O. (2006) Heterotrimeric g proteins and disease. *Methods Mol Biol*. 361, 119-144.

Ofran Y, Margalit H. (2006) Proteins of the same fold and unrelated sequences have similar amino acid composition. Proteins. 64(1):275-9.

Palmer,D.A. et al. (2006) Gib2, a novel Gbeta-like/RACK1 homolog, functions as a Gbeta subunit in cAMP signaling and is essential in Cryptococcus neoformans. *J. Biol. Chem.* 281, 32596-32605.

Pan X et al. (2000) Signal transduction cascades regulating pseudohyphal differentiation of Saccharomyces cerevisiae. *Curr Opin Microbiol*. 3(6), 567-72.

Prados-Rosales,RC. et. al. (2006) Distinct signalling pathways coordinately contribute to virulence of Fusarium oxysporum on mammalian hosts. Microbes Infect. 8, 2825-2831.

Press, W. H. et. al. (1992) "Moments of a Distribution: Mean, Variance, Skewness, and So Forth." Numerical Recipes in FORTRAN 77: The Art of Scientific Computing, 2nd ed. Cambridge, England: Cambridge University Press, pp. 604-609.

Prevost,G.P. et. al. (2006) Anticancer activity of BIM-46174, a new inhibitor of the heterotrimeric Galpha/Gbetagamma protein complex. *Cancer Res*. 66, 9227-9234.

Quevillon E. et. al. (2005) InterProScan: protein domains identifier. *Nucleic Acids Research* 33: W116-120

Rieken,S. et al. (2006) G12/G13 family G proteins regulate marginal zone B cell maturation, migra tion, and polarization. *J Immunol.* 177, 2985-2993.

Ryba,N.J. and Tirindelli,R. (1995) A novel GTP-binding protein gamma-subunit, G gamma 8, is expressed during neurogenesis in the olfactory and vomeronasal neuroepithelia. *J Biol Chem.* 270, 6757-6767.

Sachdeva,G. et al. (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics*. 21, 483-491.

Sarah,M. and Assmann. (2005) G Protein Regulation of Disease Resistance During Infection of Rice with Rice Blast Fungus. *Sci STKE*. 2005, cm13.

Sarda,D. et. al. (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. *BMC Bioinformatics*. 6, 152.

Schneider,G. (1999) How many potentially secreted proteins are contained in a bacterial genome? *Gene* 237, 113-121

Slessareva JE, et. al. (2006) Activation of the phosphatidylinositol 3-kinase Vps34 by a G protein alpha subunit at the endosome. *Cell*. 126(1), 191-203.

Smith,T. F. et al. (1999) The WD repeat: a common architecture for diverse functions. *Trends Biochem. Sci*. 24, 181–185

Spiegel,A. M. (1996). Defects in G protein-coupled signal transduction in human disease. *Annu. Rev. Physiol*. 58, 143–170.

Suwazono,Y. et al. (2006) G-protein beta3 subunit variant C825T is a risk factor for hypertension in Japanese females--a prospective cohort study over 5 years. *Ann. Hum. Genet*. 70 , 767-777.

Thompson,J.D. et. al. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22, 4673–4680.

Trusov,Y. et. al. (2006) Heterotrimeric G proteins facilitate Arabidopsis resistance to necrotrophic pathogens and are involved in jasmonate signaling. *Plant Physiol*. 140, 210-20.

Tan,P.N. et. al. (2006) *Introduction to Datamining.* Pearson Education, Inc. Boston, M.A.

Ueda,T. et al. (2003 ) Functional interaction between T2R taste receptors and G-protein alpha subunits expressed in taste receptor cells. *J Neurosci*. 23, 7376-7380.

Van der Linden,A.M., et al. (2001) The G-protein beta-subunit GPB-2 in Caenorhabditis elegans regulates the G(o)alpha-G(q)alpha signaling network through interactions with the regulator of G-protein signaling proteins EGL-10 and EAT-16. *Genetics* 158, 221-235.

Wang,L. et. al. (2006) Heterotrimeric G protein alpha subunit is involved in rice brassinosteroid response. *Cell Res.* 16, 916-922.

Wang,H.X. et. al. (2006) A Golgi-localized hexose transporter is involved in heterotrimeric G protein-mediated early development in Arabidopsis. *Mol Biol Cell*. 17, 4257-4269.

Wang P et. al. (2000) The G-protein beta subunit GPB1 is required for mating and haploid fruiting in Cryptococcus neoformans. *Mol Cell Biol*. 20(1), 352-62.

Wise,M.J. (2002) The POPPs: clustering and searching using peptide probability profiles. *Bioinformatics.* 18, S38-45.

Yamagishi,D. et. al. (2006) G protein signaling mediates developmental processes and pathogenesis of Alternaria alternata. *Mol Plant Microbe Interact*. 19, 1280-1288.

Yang,Q. et al. (2002) A G-protein beta subunit required for sexual and vegetative development and maintenance of normal G alpha protein levels in Neurospora crassa. *Eukaryotic Cell*. 1, 378-390.

Zavaljevski,N. et. al. (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*. 18, 689-696.

Zhao CY et al. (2006). Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology* 217, 105-119.

Zuegge,J. et al. (2001) Deciphering apicoplast targeting signals--feature extraction from nuclear-encoded precursors of Plasmodium falciparum apicoplast proteins. *Gene* 280,19-26.

Table1. Performance of 3 modules of Vector G on a test set.

| Classifiers | Tripeptide composition | | Dipeptide composition | | Hydrophobicity composition | |
|---|---|---|---|---|---|---|
| | *Sn*[a] | *Sp*[b] | *Sn* | *Sp* | *Sn* | *Sp* |
| Alpha Subunit | 99.45 | 99.45 | 95.05 | 98.90 | 39.01 | 100 |
| Beta Subunit | 95.38 | 100 | 93.84 | 100 | 72.30 | 100 |
| Gamma Subunit | 97.6 | 100 | 97.6 | 100 | 91.2 | 100 |

[a] Sensitivity
[b] Specificity

Table2. Experimentally verified G protein subunits and Vector-G prediction

| *Organism* | G Protein Subunits | Associated Function | Vector-G prediction [a] | References |
|---|---|---|---|---|
| *Homo sapiens* | GNAS1 | Pseudohypoparathyroidism, McCune–Albright syndrome | 0.93669 | Spiegel, A. M. 1996. |
| | GNB3 | Hypertension | 0.97627 | Suwazono et.al. 2006 |
| | GNG11 | regulates cellular senescence | 0.938211 | Hossain et.al. 2006 |
| *Mus musculus* | G alpha 12 | regulate marginal zone B cell maturation, migration, and polarization | 0.937694 | Rieken et.al. 2006 |
| | Gnb5 | Prolonged photoresponses and defective adaptation in rods | 0.96541 | Krispel et.al. 2003 |
| | Gng3 | effective immune response | 0.938965 | Dubeykovskiy et.al. 2006 |
| *Rattus norvegicus* | gustducin | Bitter taste perception | 0.935633 | Ueda et.al 2003 |
| | G beta 3 | | 0.9838 | Ray and Robishaw 1994 |
| | Gng8 | development of olfactory and vomeronasal neurons | 0.946735 | Ryba and Tirindelli 2005 |
| *Drosophila melanogaster* | G alpha | associative learning | 0.960986 | Ferris et.al 2006 |
| | G beta 13F | spindle development | 0.97625 | Fuse et.al. 2003 |
| | G gamma1 | sugar reception | 0.936403 | Ishimoto et.al. 2005 |
| *Caenorhabditis elegans* | goa-1 | role in olfactory adaptation | 0.948741 | Matsuki et.al 2006 |
| | GPB-2 | behavioral defects | 0.95996 | van der Linden et.al 2001 |
| | Gpc-1 | involved in taste adaptation | 0.93628 | Jansen et.al. 2002 |
| *Neurospora crassa* | gna-3 | Regulation of conidiation and adenylyl cyclase levels | 0.94215 | Kays et.al. 2000 11 |
| | gnb-1 | Sexual and vegetative development and maintenance of normal G alpha | 0.97496 | Yang et.al 2002 |

| | | | | |
|---|---|---|---|---|
| | GNG-1 | Required for Normal Female Fertility, Asexual Development, and G{alpha} Protein Levels | 0.936258 | Krystofova and Borkovich 2005 |
| *Cryptococcus neoformans var. neoformans* | GPA1 | Regulation of mating and virulence | 0.944513 | Alspaugh et.al. 1997 |
| | Gib2 | cAMP signaling | 0.93907 | Palmer et.al. 2006 |
| | - | - | - | - |

[a]. *nscore* value

Fig1. (a)

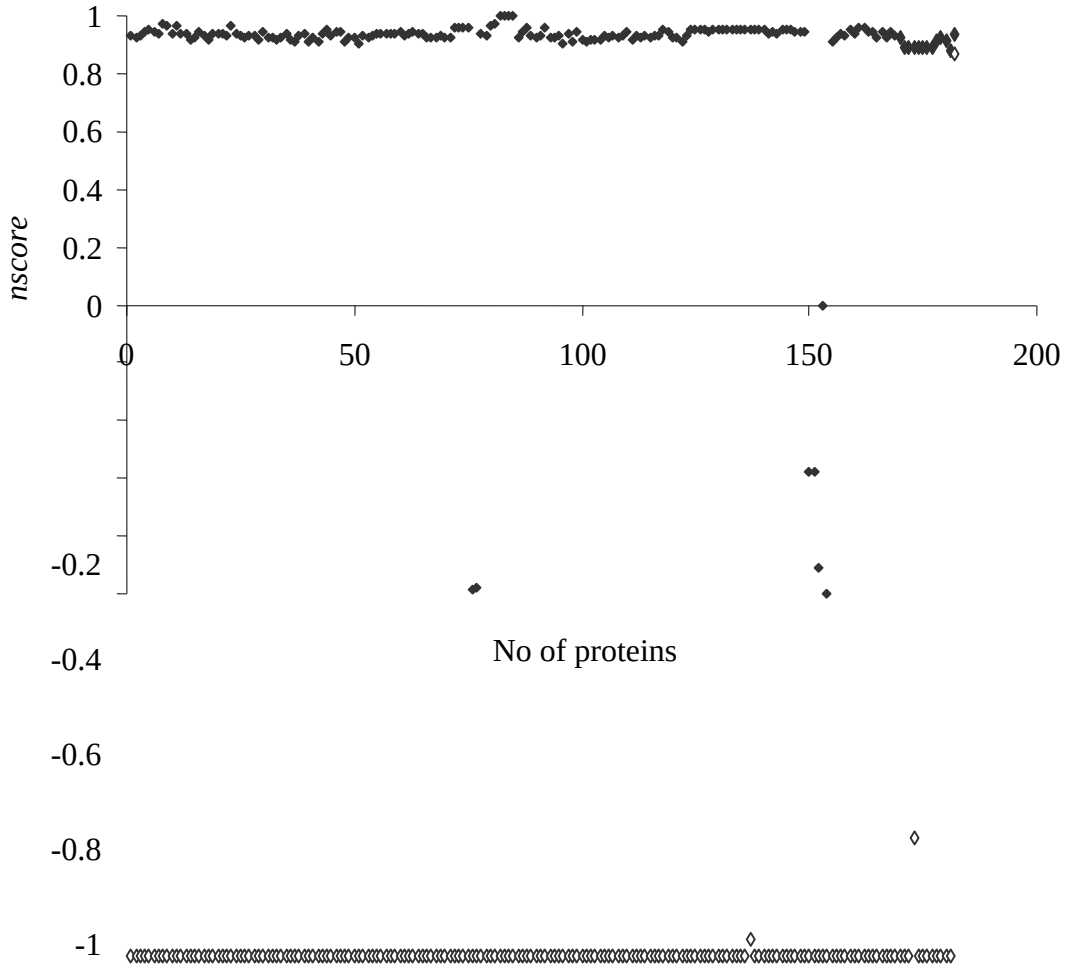
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
nscore
0
50
100
150
200
No of proteins

Fig1. (b)

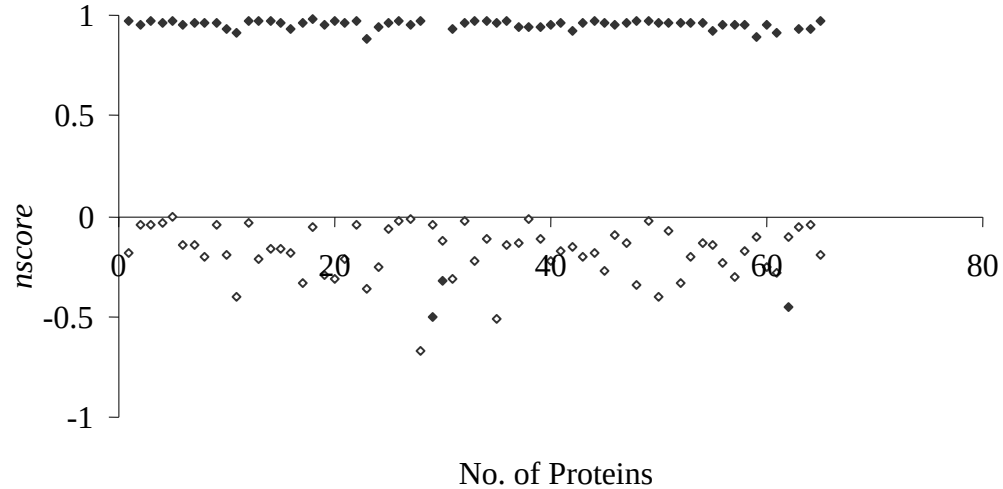
1
0.5
0
-0.5
-1
nscore
0
20
40
60
80
No. of Proteins

Fig1. (c)

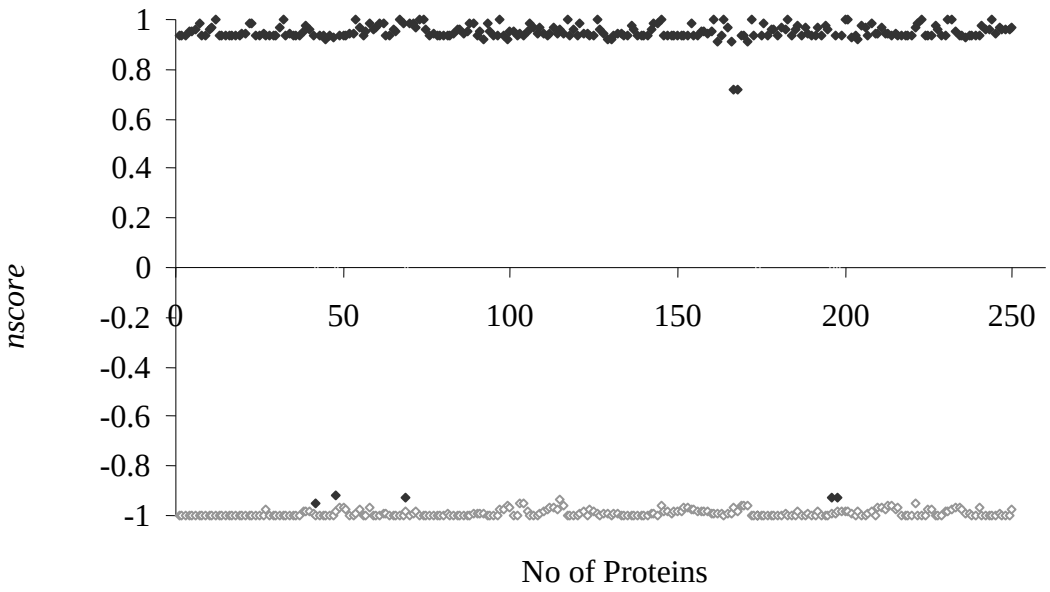
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
50
100
150
200
250
nscore
No of Proteins

Fig 2.

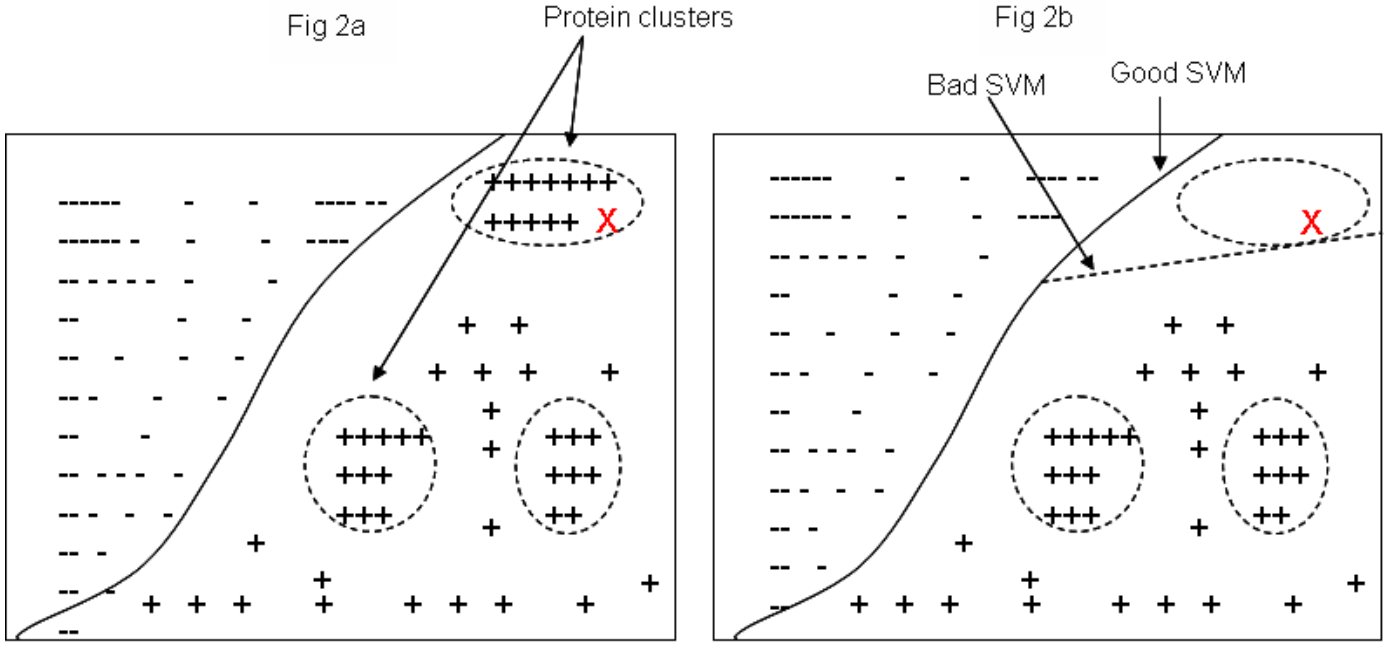
Fig 2a
Protein clusters
Fig 2b
Bad SVM
Good SVM

Figure 1: Performance of SVM classifier on a test set of G protein subclasses and corresponding *nscore*. The theoretical cut-off based on this SVM algorithm is 0. A test set of 182 alpha, 65 beta, 250 gamma subunits and similar number of negative set was taken for each positive set. (a) Scatter plot showing SVM output of alpha test set. The shaded squares show predicted value for alpha subunit and open squares show predicted value for negative test set. This method can efficiently segregate positive and negative set. Note that positive predictions are more concentrated near 1 and negative predictions are near -1. The middle zone comes under twilight zone. Two proteins (gi - 41351777, 41351779) from *M. musculus*, two proteins (gi-556256, 1169858) from *Leishmania donovani*, one protein (gi- 113638666, 3550264 and 68427275) each from *Oryza sativa*, *Stentor coeruleus*and *Danio rerio* respectively have *nscore* less than 0. Proteins from *L. donovani*, *D. rerio* and *S. coeruleus* shows half G alpha domain and 2 proteins of *M. musculus* shows no G alpha domain. In (b) and (c) similar analyses are done for beta and gamma subunits respectively. In G beta classification module, three proteins (gi - 1001939 and 33150694, and 33150742) from *Homo sapiens* have *nscore* less than 0. In G gamma classification module, two proteins (gi-66847902 and 70990842) from *Aspergillus. fumigatus*, one protein (gi – 55957436) from *Homo. sapiens*, one protein (gi- 6321317) from *S. cerevisiae*, one protein (gi- 52782800) from *Yarrowia lipolytica* and one protein (gi- 28201803) from *Schizosaccharomyces pombe* showed *nscore* less than 0.

Fig 2. This figure represents the working of Blast and vector-G. The (-) and (+) represents an example of negative and positive set respectively. The X in red represents a query protein. If one cluster of similar sequences was removed as shown in Fig 2b, blast will not be able to find a match for protein X. Vector-G performs classification by constructing an *N*-dimensional hyperplane that optimally separates the data into two categories ( - , + ). Here, upon removing a cluster of proteins SVM may work two ways. After training, good SVM will remain there, where it was initially. Bad SVM will try to divide two groups and shift the dividing line away from the original one. This way it is not able to predict the query protein X.