Support Vector Machine Based Prediction of Glutathione S-Transferase Proteins

Nitish Kumar Mishra, Manish Kumar and G.P.S. Raghava*

Bioinformatics Center, Institute of Microbial Technology, Chandigarh, India

Abstract: Glutathione S-transferase (GST) proteins play vital role in living organism that includes detoxification of exogenous and endogenous chemicals, survivability during stress condition. This paper describes a method developed for predicting GST proteins. We have used a dataset of 107 GST and 107 non-GST proteins for training and the performance of the method was evaluated with five-fold cross-validation technique. First a SVM based method has been developed using amino acid and dipeptide composition and achieved the maximum accuracy of 91.59% and 95.79% respectively. In addition we developed a SVM based method using tripeptide composition and achieved maximum accuracy 97.66% which is better than accuracy achieved by HMM based searching (96.26%). Based on above study a web-server GSTPred has been developed (http://www.imtech.res.in/raghava/gstpred/).

Keywords: GST protein, Support vector machine, artificial intelligence, sensitivity, specificity, correlation.

INTRODUCTION

The functional annotation of genome involves the assignment of function to protein product of a gene. Function assignment is currently very hot topic in computational biology due to large number of protein sequences without function assignment. With the introduction of several high-throughput sequencing methods, the number of proteins whose sequence is known but function is not known is continuously increasing.

Glutathione S-transferases (GSTs) are a group of ubiquitous and multifunctional enzymes found in both prokaryotes and eukaryotes. The GSTs comprise a complex and widespread enzyme super-family that has been subdivided further into different classes on the basis of amino acid/nucleotide sequence, immunological, kinetic and tertiary/quaternary structural properties. These proteins participate in phase II of detoxification cycle by catalyzing the conjugation of glutathione (GSH) to a wide variety of xenobiotics. It protects cells against toxic compounds by conjugating them to GSH, thereby neutralizing their electrophilic sites, and rendering the product more water-soluble [1]. Besides this, they are also involved in several other functions, such as removal of reactive oxygen, regeneration of S-thiolated proteins, catalysis of conjugation with endogenous ligands, and catalysis of reactions in metabolic pathways not associated with detoxification [2]. The detoxification ability of GSTs plays a key role in imparting protection from environmental and oxidative stress.

Another important function of GSTs are making a cell drug-resistant by avoidance of apoptotic cells death, altered expression of multi-drug resistance-associated proteins or drug metabolism or uptake, and/or over-expression of GSTs [3]. GSTs are involved in drug resistance by either i) partici pation in detoxification process with GSH or ii) increasing the pumping out of drug molecule from the cell or iii) inhibition of MAP kinase pathways [4-8]. Overexpression of specific GSTs in mammalians cells cause anti-cancer drug (alkylating agent used in cancer chemotherapy) resistance [9-12]. Cell lines selected *in vitro* for resistance to anticancer drugs frequently over-express GST π; although overexpression of GST μ and GST α are reported but in lesser quantity [10-14]. Hence GST π become a prominent marker for cancer. GSTs are also involved in cell signaling and apoptosis. Hence it determines the fate of cells under stress condition [15]. GST indirectly regulate the cellular immune response by controlling GSH level. Since it is the level of GSH along with antigen-presenting cells, which determine whether a Th1 or Th2 patterns of response predominates [16]. Hence its level control survival levels in case of HIV disease [17]. Due to central importance of GSTs in drug metabolism, they are therapeutics target for asthma [18, 19], cancer, HIV and several other diseases.

In the past, a number of strategies have been used to search novel GSTs from protein sequence data. Similarity search-based techniques, BLAST [20-21] and FASTA [22], are best in which the query sequence is searched against experimentally annotated proteins. If query protein has significant sequence similarity with any GSTs protein then it is predicted as GSTs proteins. Other methods include hidden Markov Model based profile searching. But these methods fail to predict novel/new protein when query protein does not have significant similarity with database proteins [23]. In order to overcome this problem we developed a Support Vector Machine (SVM) based prediction method, GSTpred, for annotating GST proteins on the basis of amino acid sequence. In present study first we carried out systematic analysis of amino acid composition of both GST and non-GST proteins, and then on the basis of conclusion drawn we developed the SVM based prediction method.

*Address correspondence to this author at the Bioinformatics Center, Institute of Microbial Technology, Chandigarh, India; Tel: +91-172-2690557; Fax: +91-172-2690632; E-mail: raghava@imtech.res.in

0929-8665/07 $50.00+.00

MATERIAL AND METHODS

Datasets

All sequences used in this study were downloaded from Swissprot database [24]. We got total 208 proteins in response to keyword Glutathione-S-transferase. All proteins were manually examined to retain only sequences, which have high quality annotation. For this we removed all sequences that was labeled as 'fragment' or annotated as putative or by similarity. We got total 137 proteins which were experimentally annotated as 'GST protein'. The sequence redundancy of dataset was further removed by using CD-HIT [25] such that no two proteins have sequence identity more than 90%. The final dataset contains total 107 GST protein sequences. Negative dataset was compiled by randomly selecting 107 proteins keeping in mind that they were experimentally annotated as non-GST protein and they didn't have sequence identity more that 90%. Here we are trying to developed broad-spectrum method for GSTs prediction hence we used both prokaryotes and eukaryotes (plant, fungi, animals) proteins in our study.

Evaluation Procedure

The performance of the modules constructed in this method was evaluated using five-fold cross-validation techniques. In five fold cross-validation, the non-redundant dataset was randomly divided into five sets. The training and testing were carried out for five times, each time using one set in testing and the remaining four sets in training. To evaluate the performance of methods we used several parameters routinely used in similar studies [26-27]. Brief description of these parameters is as follows (i) the sensitivity or percentage coverage of GSTs is the percentage of GSTs protein correctly predicted as GSTs proteins. (ii) The specificity or percentage coverage of non-GSTs proteins is the percentage of non-GSTs proteins correctly predicted as non-GSTs proteins. (iii) The accuracy is the percentage of correctly predicted proteins. This parameter can be calculated using equations-

$$Senstivity = \frac{TP}{TP + FN} \times 100 \quad \text{(I)}$$

$$Specifity = \frac{TN}{TN + FP} \times 100 \quad \text{(II)}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad \text{(III)}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \times 100 \quad \text{(IV)}$$

Where TP is correctly predicted positive (GSTs) proteins, TN is correctly predicted negative (non-GSTs) proteins, FP is number of non-GST proteins wrongly predicted as GSTs proteins and FN is number of GST proteins wrongly predicted non-GSTs proteins. Fig. **1** diagrammatically depicts the way of determining a prediction in above-mentioned four classes. Matthew's correlation coefficient (MCC) equal to 1 is regarded as a perfect prediction, whereas 0 is for a completely random prediction.

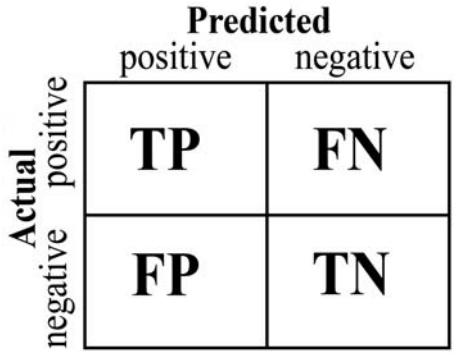

Figure 1. Criteria of classification of a prediction into true positive (TP), true negative (TN), false positive (FP), or false negative (FN).

All the measures described here above have a common drawback that they give the performance at a given threshold. A known threshold independent parameter is receiver operating characteristic (ROC), which is a plot between true positive proportion (TP/TP+FP) and false positive proportion (FP/FP+TN) [28]. The area under the curve give a single value to evaluate the performance of method. Area of 1 shows perfect prediction while 0.5 shows random prediction.

Support Vector Machine

An excellent machine learning technique SVM has been used for the prediction of GSTs proteins. In the present study, a freely downloadable package of SVM, SVM_light, has been used to predict the GSTs proteins. The software is downloaded from http://www.cs.cornell.edu/People/tj/svm_light/. SVM is a universal approximators based on statistical and optimisation theory. The SVM is particularly attractive to biological sequence analysis due to its ability to handle noise, dataset and large input space [29]. Further details about SVM can be obtained from Vapinik's papers [30] or http://www.imtech.res.in/raghava/gstpred/algorithm.html.
The software enables the user to define number of parameters as well as to select from some inbuilt kernel functions, including a Radial Basis Function (RBF), polynomial, and linear kernel. Preliminary test shows that the RBF kernel gives results better than the other kernels. Therefore, in this work we used the RBF kernel for all the experiments.

Protein Feature

The SVM modules were trained on the basis of following features of protein sequence-

Amino Acid Composition

The SVM was provided with 20 dimensional vectors encapsulating the amino acid composition of protein. Amino acid composition is the fraction of each amino acid in a protein. The fraction of each of 20 natural amino acids was calculated using the following formula:

$$\text{Fraction of amino acid (i)} = \frac{\text{Total number of amino acids (i)}}{\text{Total number of amino acid in protein}} \quad \text{(V)}$$

where (i) can be any amino acid.

Dipeptide Composition

Dipeptide composition was used to encapsulate the global information about each protein sequence, which gives

a fixed pattern length of 400 (20*20). This representation encompassed the information about amino acid composition along the local order of amino acids. The fraction of each dipeptide was calculated using the following equation:

$$\text{Fraction of dipeptide (i)} = \frac{\text{Total number of dipeptide (i)}}{\text{Total number of all possible dipeptides}} \quad \text{(VI)}$$

where dipeptide (i) is one out of 400 dipeptides.

Tripeptide Composition

Tripeptide composition was used to encapsulate the global information about each protein sequence, which gives fixed pattern length of 8000. This representation encompasses the information about amino acid composition along the local order of amino acids. The fraction of each tripeptides was calculated using the following equation:

$$\text{Fraction of tripeptide (i)} = \frac{\text{Total number of tripeptide (i)}}{\text{Total number of all possible tripeptide}} \quad \text{(VII)}$$

where tripeptide (i) is one out of 8000 tripeptides.

Hidden Markov Model

Hidden Markov models (HMM) are statistical models of the primary structure consensus of a sequence family. Initially HMM used for speech recognition [31]. HMM build profile, which capture important information about the degree of conservation at various positions in the multiple alignments, and the varying degree to which gaps and insertion are permitted. HMM (profile based) typically outperform pairwise method in both alignment accuracy and database search sensitivity and specificity. The advantage of HMMER over other methods is that it simply works on formal probabilistic basis. Further details about HMM can be obtained from Krogh's paper [32]. In the present study, we have done HMM based searching by using a freely downloadable implementation of HMM, HMMER (http://www.psc.edu/general/software/packages/hmmer/).

Building and Searching of HMM Profile

First of all, whole dataset was divided in 5 sets similar to 5 fold cross validation. Then four set of sequence was multiply aligned by using ClustalW [33] and alignment profiles were generated by 'hmmbuild' of HMMER. The profile was then calibrated by using 'hmmcalibrate' and searched for other set of sequences, which is similar to this profile by 'hmmpfam' module of HMMER.

3. RESULTS

Amino Acid Composition Analysis

In order to analyze the biasness in amino acid composition of GSTs proteins we calculate the average of all 20 natural amino acid (Fig. **2**). We found that Lys, Leu, Pro and Tyr are more abundant in GST proteins. On the other hand Ile, Asn and Ser are more abundant in non-GST proteins.

N-terminal of GSTs forms an important part of active site. This region contains catalytically essential Tyr, Ser or Cys residues that interact with G-site of GSH through thiol group. The Tyr residue participates in hydrogen binding [34-35]. When N-terminal 20 amino acid residues were analyzed (Fig. **3**) then we found abundance of Tyr, Ser, Gly, Arg, Leu, which was also reported by Mcllwain *et al*. [3].

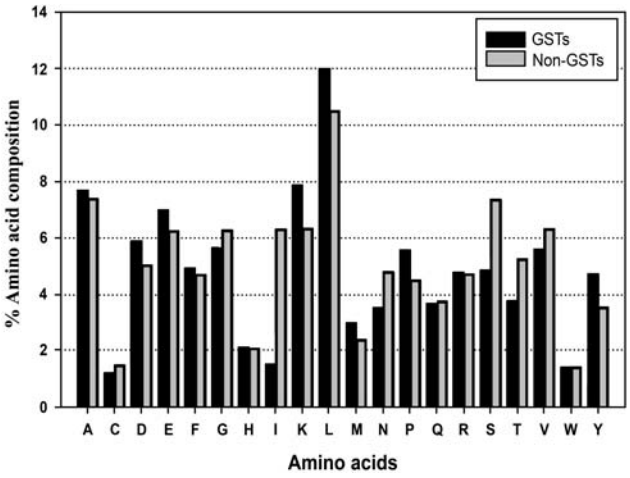

Figure 2. Average amino acid composition for GSTs and non-GSTs proteins.

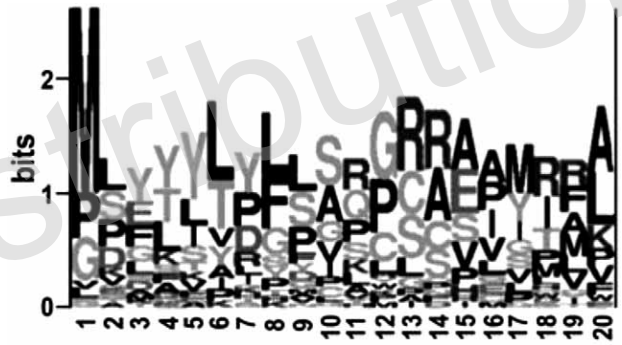

Figure 3. Sequence logo of N-terminus dataset depicting the positional propensity of amino acid at all 20 positions

HMM Based Profile Searching

We have build HMM profile by ClustalW and then searched for similar sequences. As shown in Table **1**, at e-value 0.1 profile based searching had shown 96.26% accuracy.

Table 1. The Performance of HMMER at Different E-value Threshold

Threshold E-value	Accuracy (%)
0.1	96.26
0.001	96.26
1e-4	94.39
1e-8	91.59
1e-10	90.65

When searching condition was made stricter then accuracy went down.

SVM Model Using Different form of Composition

The amino acid composition based SVM module has been able to achieve 91.59% accuracy. In order to incorporated both information about frequency of amino and their local order we have also developed different SVM models (Table **2**). SVM model based on *i* +1 and *i* +1+2 residue have substantially increased the accuracy. With dipeptide input (*i* + 1) based SVM model, MCC increased to 0.92. With tripeptide based SVM model, MCC further increased to 0.95. Fig. **4** shows the relationship between specificity and sensitivity using a ROC plot.

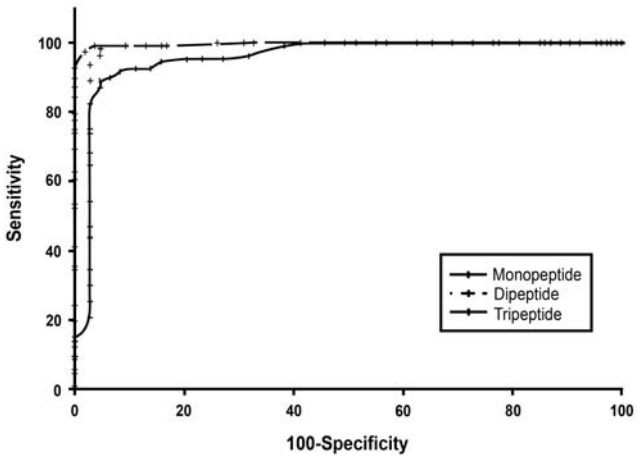

Figure 4. Comparison of ROC plot from mono-peptide, di-peptide and tri-peptide composition SVM module.

4. DESCRIPTION OF WEB SERVER

Based on our study, we have developed a web server, GSTPred, which allow the users to predict GSTs proteins from its amino acid sequence. GSTPred is freely available at http://www.imtech.res.in/raghava/gstpred. The common gateway interface (CGI) script for GSTPred was written using PERL 5.005_03. This server is installed on a Sun Server (420E) under a UNIX (Solaris 7) environment. User can enter the primary amino acids sequence for the prediction using file uploading or cut and paste option. The SVM was implemented by obtaining the SVM_light obtained from http://www.cs.cornell.edu/People/tj/svm_light/. It is a user-friendly web server and allows users to submit their protein sequence in one of the standard formats like FASTA, GenBank, EMBL, GCG or plain format (See Fig. **5a**).

The server uses the ReadSeq program to read the input sequence. The server provides option to the user to use any SVM module by default the server uses the amino acid composition approach. The server presents the results of comprehensive analysis in user-friendly format (Fig. **5b**).

5. DISCUSSION

Functional annotation of proteins is one of the major challenges in post-genomic era. The most widely used methods for predicting the function of a new protein involved sequence alignment, similarity search, or profile search, like BLAST [20], PSI-BLAST [21], FASTA [22]. These methods failed in the absence of significant similarity between queried and annotated proteins. One another reasons of failure of the similarity-based methods is that the variation in the size of proteins. The problem with profiles is that they are complicated models with many free parameters. There are number of difficult problems like the best ways to set the position-specific residue scores, to score gaps and insertion, and to combine structural and multiple sequence information.

In general artificial intelligence (AI) base techniques such as SVMs and neural networks are elegant approaches for the extraction of complex patterns from biological sequence data. These techniques are highly successful for residue state prediction where fixed window/pattern length is used [37]. AI techniques in the classification of protein (e. g. subcellular localization prediction [38-39], fold recognition [40] because similar/homologous protein often has variable length. In order to overcome this problem a fixed length pattern must be generated for proteins, for AI techniques to be implemented.

The percentage amino acid composition, which gives a fixed length pattern of 20, is commonly used by AI techniques to classify the proteins. However, this approach provides information only about the amino acid frequency, but no information about local order of amino acids. To provide the information about frequency and local order of amino acids, dipeptide composition can be used as input unit to AI techniques. Dipeptide gives a fixed pattern of 400. More information about protein sequence can be encapsulated using tripeptide composition. Tripeptide composition gives a fixed pattern of 8000. In few cases of tripeptide composition, ANN and SVM unable to handle the noise due to large number of input units and number of missing tripeptide in protein. But in this paper, we have used a SVM module on the basis of tripeptide composition. This module is able to predict the GSTs protein with very good accuracy.

Table 2. The Performance of Various SVM Modules. (Where Thres = Threshold, Sen = sensitivity, Spec = specificity, Acc = accuracy and MCC = Mathew correlation coefficient)

Approach	Parameter	Thres	Sen	Spec	Acc	MCC
Mono-peptide	J=2, g=0.01, c=1	0.2	91.59	91.59	91.59	0.83
Di-peptide	J=1, g=0.01, c=2	0.0	96.26	95.33	95.79	0.92
Tri-peptide	J=2, g=0.001, c=9	-0.2	97.20	98.13	97.66	0.95

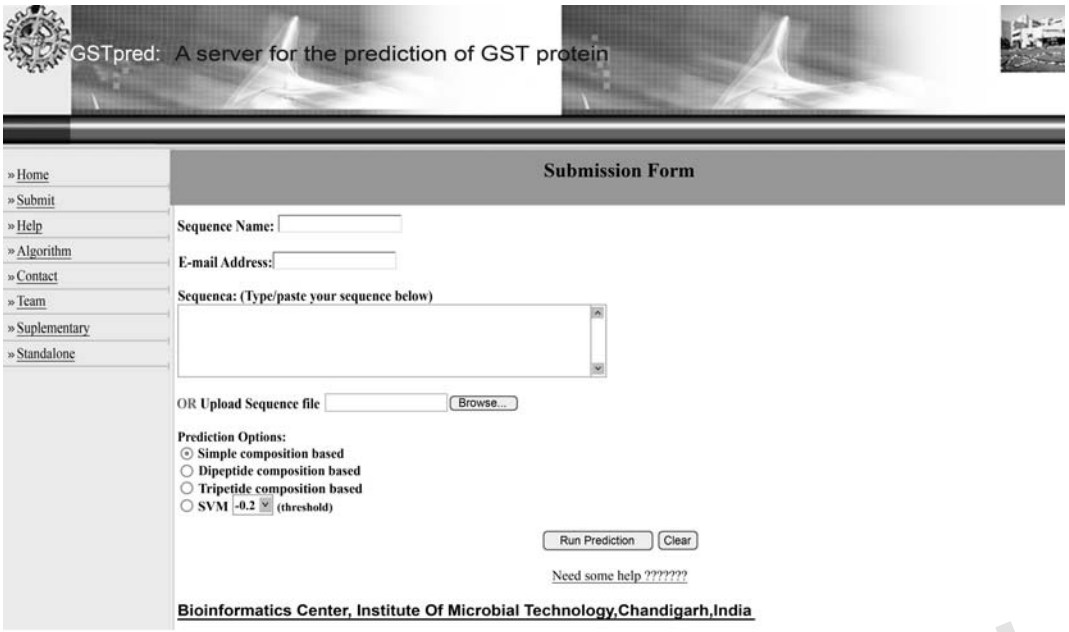

Figure 5a. A snapshot of sequence submission page of GSTPred server.

Results for given protein on peptide composition based

Total numbr of positive predicted protein: **21**
Total numbr of negative predicted protein:

Sequence name	Predicted value	Prediction	Score
>GST11ARATH	0.96953944	Positive	1.16953944
>GST16ARATH	1.058469	Positive	1.258469
>GST1ASCSU	2.1215976	Positive	2.3215976
>GST1BLAGE	1.3856834	Positive	1.5856834
>GST1DROME	0.9999366	Positive	1.1999366
>GST1SCHPO	0.012883519	Positive	0.212883519
>GST1YEAST	1	Positive	1.2
>GST26FASHE	1.7256841	Positive	1.9256841
>GST26SCHJA	1.8620683	Positive	2.0620683
>GST26SCHMA	1.6224547	Positive	1.8224547
>GST27FASHE	1.6432023	Positive	1.8432023
>GST27SCHMA	1.361689	Positive	1.561689
>GST28FASHE	1.3771392	Positive	1.5771392
>GST28SCHBO	0.54359259	Positive	0.74359259
>GST29FASHE	1.5931145	Positive	1.7931145
>GST2MANSE	1.2436087	Positive	1.4436087
>GST2SCHPO	1.2983949	Positive	1.4983949
>GST2YEAST	0.99983708	Positive	1.19983708
>GST3SCHPO	0.99990824	Positive	1.19990824
>GSTA1ANTST	0.99943966	Positive	1.19943966
>GSTA1BOVIN	2.2887946	Positive	2.4887946

Figure 5b. Results of the prediction after analysis of query protein using SVM based module.

We have used two approaches for GSTs proteins prediction (i) HMM based method by using HMMER software, (ii) amino acid composition, dipeptide composition and tripeptide composition based SVM module using SVM_light software. Although HMMER give a very good accuracy 95.66% at threshold E-value 0.1 but this E-value is too much. For further improve accuracy we have used amino acid composition, dipeptide and tripeptide based SVM models which give accuracy 91.59%, 95.79%, and 97.66% respectively.

Although GSTs are very important proteins and play great role in stress, survival, immunity, cancer and other diseases but no any pre-existing server available for its prediction. In this study, an attempt has been made to develop a direct method for predicting the GSTs proteins.

These method used here for classify GSTs and non-GSTs give very good accuracy but tripeptide composition based SVM method is more accurate than the others in this case. Our method will help to researchers in finding GSTs proteins in both prokaryotes and eukaryotes.

ACKNOWLEDGEMENT

The authors are thankful to the Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India for financial support. Nitish Kumar Mishra and Manish Kumar are financially supported by CSIR as Junior and Senior Research Fellow respectively.

REFERENCES

[1] Boyland, E. and Chasseuad, L.F. (**1969**) *Adv. Enzymol. Rela. Areas Mol. Biol., 32*, 173.
[2] Sheehan, D. and Meade, G. (**2001**) *Biochem. J., 360*, 1.
[3] McIlwain, C.C., Townsend, D.M. and Tew, K.D. (**2006**) *Oncogene, 25*, 1639.
[4] Walters, M.J. and Baan, R.A. (**1998**) *Br. J. Cancer, 77*, 556.
[5] Muller, M., Meijer, C., Zaman, G.J.R., Borst, P., Scheper, R.J., Mulder, N.H., de Vries E.G.E. and Jansen, P.L.M. (**1994**) *Proc. Natl. Acad. Sci. USA, 91*, 13033.
[6] Townsend, D.M. (**2003**) *Oncogene, 22*, 7369.
[7] Kramer, R.A. and Kim, G. (**1988**) *Science, 241*, 694.
[8] Black, S.M. and Sakai, M. (**1990**) *Biochem. J., 268*, 309.
[9] Morrow, C.S. and Cowan, K.H. (**1990**) *Cancer Cells, 2*, 15.
[10] Bastin, G., Sinha, B. and Cowan, K.H. (**1986**) *J. Biol. Chem., 261*, 15544.
[11] Moscow, J.A. and Cowan, K.H. (**1989**) *Mol. Pharmacol., 36*, 22.
[12] Puchalski, R.B. and Fahal, W.E. (**1990**) *Proc. Natl. Acad. Sci., 87*, 2443.
[13] Hayes, J.D. and Pulford, D.J. (**1995**) *Crit. Rev. Biochem. Mol. Biol., 30*, 445.
[14] Tew, K.D. (**1994**) *Cancer Res., 54*, 4313.
[15] Liu, M., Pelling, J.C. and Brash, D.E. (**1998**) *Cancer Res., 58*, 1723.
[16] Peterson, J.D. and Herzenberg, L.A. (**1998**) *Proc. Natl. Acad. Sci. USA, 95*, 3071.
[17] Herzenberg, L.A. and DeRosa, S.C. (**1997**) *Proc. Natl. Acad. Sci. USA, 94*, 1967.
[18] Piker, P.J. (**1984**) *Physiol. Rev., 64*, 744.
[19] Ford-Hutchinson, A.W. (**1990**) *Crit. Rev. Immunol., 10*, 1.
[20] Alstchul, S.F. and Lipman, D.J. (**1990**) *J Mol Biol., 215*, 403.
[21] Alstchul, S.F. and Lipman, D.J. (**1997**) *Nucleic Acids Res., 25*, 3389.
[22] Pearson, W.R. and Lipman, D.J. (**1988**) *Proc. Natl. Acad. Sci. USA, 85*, 2444.
[23] Reinhardt, A. and Hubbard, T. (**1998**) *Nucleic Acids Res., 26*, 2230.
[24] Boeckmann, B. and Bairoch, A. (**2003**) *Nucleic Acids Res., 31*, 365.
[25] Weizhong, Li. (**2006**) *Bioinformatics, 22*, 1658.
[26] Kumar M., Bhasin, M. and Raghava, G.P.S. (**2005**) *Nucleic Acids Res., 33*, 154.
[27] Bhasin, M. and Raghava, G. P. S. (**2004**) *Nucleic Acids Res., 32*, W414.
[28] Sweats, J.A. (**1988**) *Science*, 240, 1285.
[29] Zvaljevaski, N. and Stevens, F.J. (**2002**) *Bioinformatics, 18*, 689.
[30] Joachims, T. (**1999**) in *Advanced in Kernel Methods-Support Vector Learning* (eds. Scholkopf B., *et al.*) pp. 42-56. MIT Press, Cambridge.
[31] Rabiner, L.R. (**1989**) *Proc. I.E.E.E., 77*, 257.
[32] Krogh, A. and Brown, M. (**1994**) *J. Mol. Biol., 235*, 1501.
[33] Thompson, J.D. and Gibson, T.J., (**1994**) *Nucleic Acids Res., 22*, 4673.
[34] Dirr, H.W. and Huber, R. (**1994**) *Eur. J. Biochem., 220*, 645.
[35] Liu, S., Johnson, W.W. and Armstrong, R.N. (**1992**) *J. Biol. Chem., 268*, 4296.
[36] Dixon, D.P., Lapthron, A. and Edward, R. (**2002**) *Genome Biol., 3*, 3004.
[37] Krogh, A. and Riis, S.K. (**1996**). in *Advanced in Neural Information Processing System 8* (Touretzky, D.S., Mozer, M.C. (eds.)) pp. 917-923. MIT Press, Cambridge.
[38] Bhasin, M., Garag, A. and Raghava, G.P.S. (**2005**) *Bioinformatics, 21*, 2522.
[39] Garg, A., Bhasin, M. and Raghava, G.P.S. (**2005**) *J Biol. Chem., 280*, 14427.
[40] Kaur, H. and Raghava, G.P.S. (**2004**) *Proteins, 55*, 83.

Received: April 09, 2007 **Accepted: April 23, 2007**