SUBSEQUENCE FEATURE MAPS FOR PROTEIN FUNCTION ANNOTATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖMER SİNAN SARAÇ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

JULY 2008

Approval of the thesis

SUBSEQUENCE FEATURE MAPS FOR PROTEIN FUNCTION ANNOTATION

submitted by **ÖMER SİNAN SARAÇ** in partial fullfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen ________________
Dean, **Graduate School of Natural and Applied Sciences**

Prof. Dr. Volkan Atalay ________________
Head of Department, **Computer Engineering**

Prof. Dr. Volkan Atalay ________________
Supervisor, **Computer Engineering Dept., METU**

Examining Committee Members:

Prof. Dr. İsmail Hakkı Toroslu ________________
Computer Engineering Dept., METU

Prof. Dr. Volkan Atalay ________________
Computer Engineering Dept., METU

Prof. Dr. Gerhard Wilhelm Weber ________________
Institute of Applied Mathematics, METU

Prof. Dr. Faruk Polat ________________
Computer Engineering Dept., METU

Assoc. Prof. Dr. Işık Yuluğ ________________
Molecular Biology and Genetics Dept., Bilkent University

Date: ________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Ömer Sinan Saraç

Signature :

ABSTRACT

SUBSEQUENCE FEATURE MAPS FOR PROTEIN FUNCTION ANNOTATION

Saraç, Ömer Sinan

Ph.D., Department of Computer Engineering

Supervisor : Prof. Dr. Volkan Atalay

July 2008, 100 pages

With the advances in sequencing technologies, the number of protein sequences with unknown function increases rapidly. Hence, computational methods for functional annotation of these protein sequences become of the upmost importance. In this thesis, we first defined a feature space mapping of protein primary sequences to fixed dimensional numerical vectors. This mapping, which is called the Subsequence Profile Map (SPMap), takes into account the models of the subsequences of protein sequences. The resulting vectors were used as an input to support vector machines (SVM) for functional classification of proteins. Second, we defined the protein functional annotation problem as a classification problem and construct a classification framework defined on Gene Ontology (GO) terms. Different classification methods as well as their combinations are assessed on this framework which is based on 300 GO molecular function terms. The results showed that combination enhances the classification accuracy. The resultant system is made publicly available as an online function annotation tool.

Keywords: Functional Annotation, Subsequence Feature Maps, Protein Classification

ÖZ

PROTEİN FONKSİYON AÇIKLAMASI İÇİN ALTDİZİ ÖZELLİK HARİTALARI

Saraç, Ömer Sinan

Doktora, Bilgisayar Mühendisliği

Tez Yöneticisi : Prof. Dr. Volkan Atalay

Temmuz 2008, 100 sayfa

Sekans belirleme teknolojlerindeki gelişmelerle birlikte, işlevi bilinmeyen protein dizilerinin sayısı hızla artmaktadır. Bunun sonucunda proteinlerin işlevsel olarak etiketlenmesi için kullanılabilecek hesaplamalı metodlar çok büyük önem kazanmıştır. Bu tezde, ilk olarak protein birincil dizilerini sabit boyutlu sayısal vektörlere eşleyen bir öznitelik uzayı eşleme sistemi tanımladık. Altdizi profili eşlemesi adını verdiğimiz bu eşleme protein dizilerinin altdizi modellerini hesaba katmaktadır. Oluşan vektörler proteinleri işlevsel olarak sınıflandırmak için desktek vectör makinalarına girdi olarak kullanılmıştır. İkinci kısımda, proteinlerin işlevsel etiketlenme işini bir ilevsel sınıflandırma problemi olarak tanımladık ve Gen Ontoloji (GO) terimleri üzerinde tanımlanmış bir sınıflandırma çatısı bina ettik. Farklı sınıflandırma metodları ve bunların farklı birleşimleri 300 GO terimi üzerine kurulan bu sınıflandırma çatısında değerlendirildi. Sonuçlar gösterdi ki birleşim sınıflandırma doğruluğunu arttırmaktadır. Ortaya çıkan sistem internet üzerinde herkese açık bir işlevsel etiketleme uygulaması haline getirilmiştir.

Anahtar Kelimeler: İşlevsel Etiketleme, Altdizi Özellik Haritaları, Protein Sınıflandırması

ACKNOWLEDGMENTS

In one of his comic strips Jorge Cham rightfully states the second law of graduation as: "..a student in procrastination tends to stay in procrastination unless an external force is applied to it". My supervisor Volkan Atalay and his always positive attitude was the main driving force that made this thesis possible. It is always informative, fun and fulfilling to discuss almost any subject with him. I shall never thank him enough. I also would like to thank to his dearest wife, Rengul Cetin-Atalay, for her wonderful comments and suggestions about the biological aspect of this study.

I am grateful to the people of Bioinformatics Lab. They are: Dr. Tolga Can, Oral Dalay, Zerrin Sökmen, Gökçen Alay-Çilingir, Ayşe Gül Yaman, Mehmet Ersan Topaloğlu, Kırçiçeği Korkmaz, Perit Bezek, Alper Söyler. Not only they were always willing to help, but also they showed me how it is to do a PhD with real friends. I should also thank to my roommates Oral, Zerrin, Gökçen, Alper and Ayşe Gül for not throwing me and my violin out of the window.

Doing PhD is one thing and doing PhD in Dept. of Computer Engineering of METU is another thing. I would like to thank to every single person in this department who helps to keep this department as a big supportive family. One can never feel truly alone and stuck in this place. I partially worked in Dept. of Molecular Biology and Genetics in Bilkent University during last 6 months of study. I'd like to thank to the people there for helping me to find my way as a computer engineer in the wonderful and mysterious world of biology.

I'd like to thank to The Scientific and Technological Research Council of Turkey for their partial financial support under EEEAG-E105035 project.

Finally, I'd like to thank my dearest family. First of all my wife, Seda, for her love and understanding and sharing this life with me in hard times and fun times. Of course,

I’m always grateful to my parents and my brothers for their never ending support. They were always there for me and made me the person I am now. “For every action towards graduation, there is an equal and opposite distraction”, is the third law of graduation. I just want to thank my dear son, Ahmet Berat, for being the sweetest distraction during this thesis. I hope, one day you will be able to read this: I love you.

To my family

TABLE OF CONTENTS

CHAPTERS

LIST OF FIGURES

FIGURES

LIST OF TABLES

TABLES

CHAPTER 1

INTRODUCTION

Proteins are the macromolecules that are responsible for a wide range of essential functions of life and participate in every process within cells. Thus, discovering functions of proteins is crucial for understanding, and furthermore manipulating, the mechanisms of life. Designing more effective drugs with less side effects, engineering genetic codes for copious crops may only be possible with a deeper understanding of functions of proteins.

Amino acids are the basic structural building units of proteins. They form short polymer chains called peptides or longer chains called either polypeptides or proteins. An mRNA template is generated from a protein coding region of DNA through a process called transcription. This mRNA template is used in the process of building the protein combining the amino acids in the specified order known as translation, which is part of protein biosynthesis. Twenty amino acids are encoded by the standard genetic code and are called proteinogenic or standard amino acids.Each protein is formed as a linear chain of 20 different amino acids in a specific order. This linear sequence is called the primary structure of the protein. Thus proteins can be represented as strings of varying length formed by a 20 letter alphabet. Typically, the length of proteins may vary from several hundred to several thousand amino acids. Each amino acid has certain physiochemical properties such polarity, acidity, hydropathy, size, etc. These properties along with the specific sequence of amino acids of a protein determine its 3D structure and function.

1.0.1 Gene Ontology

Function of a protein is a vague term where the exact meaning depends on the context it is used. A protein may be involved in a cell signaling activity by catalyzing a metabolic reaction. Thus, same protein may be annotated by its enzymatic activity or by the cell signaling activity which it is involved in. Furthermore there is no unique vocabulary among the biologist to refer to same protein function. Phosphotransferase activity or kinase activity might be used interchangeably to refer to the same function. The size of the vocabulary that is required to cover all possible functions of proteins in all organisms is large. Some proteins are enzymes that catalyze numerous biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signaling, immune responses, cell adhesion, and the cell cycle.

Gene Ontology (GO) is the well-known and most widely used approach for formalization of protein functional terms and their relations ([4]). The GO project has developed three structured controlled vocabularies (ontology) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. Building blocks of GO are terms. Each term in GO has a unique numerical identifier of the form **GO:nnnnnnn** and a term name, e.g. *cell, fibroblast growth factor receptor binding or signal transduction*. Each term is also assigned to one of the three ontologies, molecular function, cellular component or biological process.

The ontologies are structured as directed acyclic graphs, which are similar to hierarchies but differ in that a more specialized term (child) can be related to more than one less specialized term (parent). For example, the biological process term **hexose biosynthetic process** has two parents, **hexose metabolic process** and **monosaccharide biosynthetic process**. This is because biosynthetic process is a type of metabolic process and a hexose is a type of monosaccharide. When any gene involved in hexose biosynthetic process is annotated to this term, it is automatically annotated to both hexose metabolic process and monosaccharide biosynthetic process. There are two main term-term relationships in ontologies, *is_a* relations and *part_of* relations. Both of these relations are are transitive, which means that the relationships are propagated from children terms to parent terms.

In this thesis, we focused on annotation of proteins with *molecular function* terms so here we will only give information about *molecular function* aspect of GO. Molecular function describes activities, such as catalytic or binding activities, that occur at the molecular level and they are mainly connected with *is_a* links. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where or when, or in what context, the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products, but some activities are performed by assembled complexes of gene products. Examples of broad functional terms are *catalytic activity, transporter activity, or binding*; examples of narrower functional terms are *adenylate cyclase activity or Toll receptor binding*. It is easy to confuse a gene product name with its molecular function, and for that reason many GO molecular functions are appended with the word *activity*. Currently, there are more than 8000 GO terms in *molecular function* ontology.

1.0.2 Problem Definition

Along with the recent advances in genome sequencing technologies, the number of protein sequences with missing functional annotations increases rapidly. The are more than one million proteins with identified primary sequence. Only around forty thousand of these proteins have functional annotations. Experimental methods for determining the structure and function of a protein are very expensive and typically takes about a year. Thus, computational annotation methods become indispensable for providing a road map for the biologist for further investigation of the excessive number of sequences with unknown functions *in vivo*. The functional annotation of a protein in the context of this thesis is thus defined as assigning a function to a given protein in the form of its primary sequence which is mainly a string of 20 letter alphabet. Since definition of the function of a protein is a vague term and requires formalization, we confine ourselves to annotating proteins with GO molecular function terms.

1.1 State of Function Annotation in the Literature

General *in silico* course of action for the annotation of a new sequence is to find similar sequences whose functions are experimentally determined. This is usually performed by searching public databases using local alignment search tools such as BLAST or PSI-BLAST and annotations for the highest scoring hits are transferred onto the new sequence ([1, 2]). We call this first track as the *transfer approach*. Although this simple method performs well in many cases and have some advantages such as being fast and easy to implement, it has some important drawbacks ([17, 24, 45, 23]). One of such drawbacks is the excessive transfer of annotations. In some cases proteins have multiple domains related to different functions. While transferring annotations, one should consider only those functions that are related to the region of similarity otherwise unrelated functions may be assigned to the new protein. Second drawback is the propagation of annotation errors in the source database. Some of the databases employ computational methods for annotation of proteins. Errors tends to propagate to similar sequences through highest scoring hits with erroneous annotations. Deciding the similarity threshold above which the annotations would be transferred is a painstaking work. Certain level of similarity to infer functional homology for a family of proteins may not be enough for a different family hence it poses an important problem for an automated annotation system. *Transfer approach* also suffers from low sensitivity/specificity. Low sensitivity results from remote homology cases, where pairwise sequence similarity is below 40%. One may choose a low similarity threshold to detect remote homologs. In this case specificity drops drastically due to excessive transfer of unrelated functional annotations. It has been shown recently that although inferring homology through sequence similarity generally holds for the 3D structure, it is far less justified for the function. Additional information than just pairwise similarity is needed to find more accurate annotations ([17]).

In the second track, annotation of proteins is formulated as a classification problem where the annotations are classes and proteins are samples to be classified. This *classification approach* allows scientists to use sophisticated and powerful classification algorithms such as support vector machines (SVM) and artificial neural networks (ANN). These methods explicitly form a boundary between the negative and positive training samples and are shown to be more accurate in many cases ([37]). Actually *classification approach* inherently solves most of the problems stated about the *transfer approach*. Yet, they are not as popular among biologist as one would expect. One reason is that, classification approach requires well defined classes and positive and negative training data for each class. But protein function is a vague term where the exact meaning depends on the context in which it is used ([23]). Furthermore, similar functions can be referred to with different terms having different levels of specificity. Thus, one would first need a controlled vocabulary for functional terms to train classifiers. Gene Ontology (GO) is the well-known and most widely used approach for formalization of protein functional terms and their relations ([4]) and we used GO as the functional ontology in this thesis. Second, positive and negative training data must be collected for each of these terms/classes. Data preparation is not straightforward since functional terms are related to each other and proteins may have more than one annotation. If one can establish a classification framework with rich number of important functions and high quality training data, methods in classification approach will receive more attention.

There is a wide range of classification approaches to automated functional annotation in the literature. They can be grouped into three categories depending on the employed features:

1. homology-based approaches,
2. subsequence-based approaches,
3. feature-based approaches.

Homology-based approaches utilize overall sequence similarity of the target protein to the positive and negative training data to decide which functional class it belongs. It is generally accepted that high level of sequence similarity is a strong indicator of functional homology. It is important to note that although homology-based methods in *classification approach* also utilize sequence similarity, they are fundamentally different than *transfer approach* in that classification methods considers the similarities of the query sequence to all of the sequences in the positive and negative training data corresponding functional term for deciding a single annotation. The main drawback of the homology-based approach is the remote homology situations where the sequence similarity to the already annotated proteins is low. Subsequence-based approaches focus on highly conserved subregions such as motifs or domains that are critical for a protein to perform a specific function. These methods are especially effective when function to be assigned requires a specific motif or domain. Existence of these highly conserved regions in a protein enables us to infer a specific annotation even in remote homology situations ([27, 53, 38, 7, 54, 36, 8, 44]. The main problem with the subsequence-based approach is the identification of these motifs. Finding motifs is an NP-hard problem and furthermore motifs do not exist for all classification tasks. In the feature-based approach, biologically meaningful properties of a protein such as frequency of residues, molecular weight, secondary structure, extinction coefficients are extracted from the primary sequence. These properties are then arranged as feature vectors and used as input to classification techniques such as artificial neural networks (ANN) or support vector machines (SVM) ([18, 34, 42, 31, 11, 33, 12]). Each of these approaches may have different strengths and weaknesses on the classification of different functional terms. For example, their specific 3D structure is a good discriminatory feature of immunoglobulins, thus a homology-based approach that considers overall sequence similarity would be effective in identifying immunoglobulins. On the other hand, G-proteins may have different 3D structures but they share common motifs. A subsequence-based approach would be more appealing in classifying G-proteins. Hydrophobic core is a hallmark of transmembrane proteins. A method that considers hydrophobicity of residues will be a better classifier of transmembrane proteins. As a result, combining methods from different approaches will be more successful on classification of a wide range of protein functions.

1.2 Contributions

In this thesis, we first described a feature space mapping, called subsequence profile map (SPMap) ([44]) which maps protein sequences to fixed dimensional numerical vectors and enables the use of classical machine learning techniques. Our approach incorporates the information coming from important subregions that are conserved over a family of proteins as well as the overall sequence similarity. Furthermore, SPMap avoids explicit identification of motifs. As a result, SPMap combines strength of both homology-based and subsequence-based methods while avoiding the main problems of these approaches. Then, we present a method to prepare accurate training data for the terms defined in Gene Ontology (GO) framework ([4]). Finally, we investigated the effect of combining different methods for protein function classification. We focused on annotation of proteins with 300 GO molecular function terms. We formulated this problem as a classification problem with 300 classes where proteins can be assigned to more than one class. We developed an online GO annotation tool named *GOPred*.

The main contributions of this thesis are:

1. development of a subsequence-based feature map (SPMap) for protein classification;
2. an accurate training dataset preparation method for GO terms that takes into account the Directed Acyclic Structure (DAG) of GO and evidence codes provided for each available annotation;
3. a GO annotation tool (*GOPred*) that combines 3 different classification methods and covers 300 GO terms.

First chapter of this thesis defines the functional annotation problem and introduces to available methods in the literature. Chapter 2 gives some computational and biological background information. Chapter 3 and Chapter 4 correspond to two papers published in the course the PhD study and can be read independently after one has read the introduction. Chapter 3 describes details of Subsequence Profile Map (SPMap) and presents test results and comparisons to state of the art methods for functional classification ([44]). Chapter 4 presents an automated annotation tool, *GOPred*, initially covering 300 GO terms and describes dataset preparation and presents results of combining different methods for a more accurate annotation system. Chapter 5 concludes the thesis and gives some future directions in functional annotation of proteins.

CHAPTER 2

BACKGROUND INFORMATION

Bioinformatics is a multidisciplinary study that requires knowledge of both biological and computational concepts. In this chapter, we summarize some necessary computational and biological background information. In the computational background section, we describe classification and Support Vector Machines (SVM). In the biological background section, proteins and fundamentals of molecular biology of cell is described. Reader may safely skip this chapter if he/she is already familiar with the concepts.

2.1 Computational Background

Many real world problems can be formulated as a classification problem. A two-class classification problem may be defined as assigning labels $y_i \in \{-1, 1\}$ to the samples to be classified. One way two achieve this is to find a separating decision boundary between the samples from the distinct classes. SVM achieves this by finding a separating hyperplane, thus, SVM is a linear classifier. Later in this chapter, we'll show how SVMs are extended to non-linear boundaries using the *kernel trick*. The equation of a general hyperplane is $\mathbf{w} \cdot \mathbf{x} + b = 0$. The hyperplane should separate the data, so that $\mathbf{w} \cdot \mathbf{x}_k + b > 0$ for all the samples $\mathbf{x}_k$ of positive class labeled by 1, and $\mathbf{w} \cdot \mathbf{x}_j + b < 0$ for all the samples $\mathbf{x}_j$ of the negative class labeled by -1. If the positive and negative data are in fact separable in this way, there is possibly more than one way to do it. Among the possible hyperplanes, SVMs select the one where the distance of the hyperplane from the closest data points (the "margin") is as large as possible. By maximizing the margin, SVMs try to reduce the structural risk ([10]). This maximization approach can be formulated as follows. We have our training data $\{\mathbf{x}_i, y_i\}, i = 1, .., l, \quad y_i \in \{-1, 1\}, \quad \mathbf{x}_i \in \mathbf{R}^d$. If the positive and negative data are linearly separable, we can define two hyperplanes $H_1 : \mathbf{w} \cdot \mathbf{x} + b = 1$ and $H_2 : \mathbf{w} \cdot \mathbf{x} + b = -1$ such that the training data satisfies the following constraints:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 \quad for\, y_i = +1 \tag{2.1}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 \quad for\, y_i = -1 \tag{2.2}$$

These two set of inequalities can be combined as

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \;\; \forall i \tag{2.3}$$

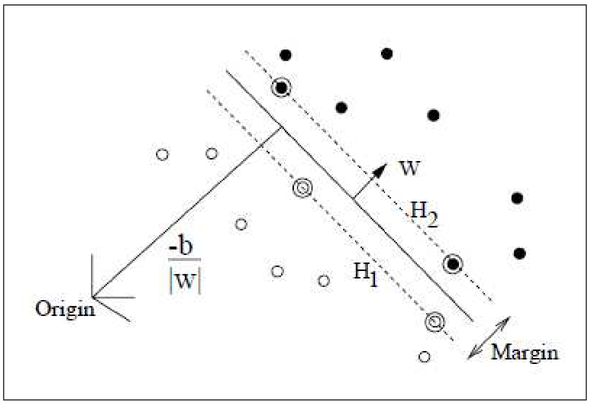


Figure 2.1: Linear separating hyperplane. The support vectors are circled ([10]).

Note that H_1 and H_2 are parallel and the are no data points between them. This means that closest data points to the separating hyperplane (which is $H : \mathbf{w} \cdot \mathbf{x} + b = 0$) lie on H_1 and H_2. Thus, *"margin"* is the perpendicular distance between these hyperplanes. The distance of H_1 to the origin is $|1 - b|/||\mathbf{w}||$ where $||w||$ denotes the norm of $\mathbf{w}$. Similarly, the distance of H_2 to the origin is $|-1-b|/||\mathbf{w}||$ and the margin is simply $2/||w||$. It is easy to see that, in order to maximize the margin it is enough to minimize $||w||^2$ with respect to the constraints given in Equation 2.3. Note that only those data points that satisfies the equality constraints given in Equation 2.3 effects the solution. They are called the support vectors (see Figure 2.1).

This optimization problem with inequality constraints can be represented as a Lagrangian. We introduce Lagrange multipliers $\alpha_i \geq 0$, $i = 1, .., l$, one for each inequality constraint (Equation 2.3). The Lagrangian is:

$$L_p = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i \tag{2.4}$$

We now have to minimize L_p with respect to $\mathbf{w}$, b subject to $\alpha_i \geq 0$. This is a convex optimization problem so it can be converted to its *dual* form. For convex problems, the partial derivatives of the function with respect the variables are equal to zero at the optimum point.

$$\frac{\partial L_p}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = 0 \tag{2.5}$$

$$\frac{\partial L_p}{\partial b} = - \sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.6}$$

If we solve Equation 2.5 for $\mathbf{w}$ and substitute it in Equation 2.4 we get the dual form

$$L_d = \sum_{i} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{2.7}$$

subject to equality constraints $\sum_{i=1}^{l} \alpha_i y_i = 0$. Now we have to maximize the *dual* formulation in order to minimize the L_p in Equation 2.4. In the *dual* form, data points $\mathbf{x}_i$ only appear as dot product form. We will see that this property of the dual problem is exploited as the *kernel trick* to extend the ideas described above to non-linear decision boundaries. Note that there is a Lagrange multiplier α_i for each training point. In the solution, α_i are non-zero only if the point $\mathbf{x}_i$ lies on H_1 or H_2. Otherwise $\alpha_i = 0$. Thus, *support vectors* are those points with $\alpha_i > 0$. The resulting α_i can be substituted in Equation 2.5 to solve for $\mathbf{w}$. Note that only support vectors effect the solution. If all other training points removed or moved (without crossing H_1 or H_2), the solution would not change.

The above formulation works only if the positive and negative training data are linearly separable, which is not the case in most of the real life problems. In order to allow misclassified samples we should relax the constraints given in Equation 2.1 and Equation 2.2. This can be achieved by introducing *slack variables*, $\xi_i \geq 0$, $i = 1, .., l$ ([15]).

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq +1 - \xi_i \quad for\, y_i = +1 \tag{2.8}$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \xi_i \quad for\, y_i = -1 \tag{2.9}$$

If a data point $\mathbf{x}_i$ is misclassified by the hyperplane, ξ_i becomes greater than 0. Thus, $\sum_i \xi_i$ is an upper bound on the number of training errors. Hence, objective function to be minimized can be changed from $||w||^2/2$ to $||w||^2/2 + C \sum_i \xi_i$, where C is a parameter to control how much penalty will be given to misclassified samples. If we convert this new optimization problem to dual form, it becomes:

Maximize:

$$L_d = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{2.10}$$

subject to:

$$0 \leq \alpha_i \leq C, \tag{2.11}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.12}$$

The only difference with the linearly separable case is that α_i now have an upper bound, C.

SVM finds the optimum hyperplane to separate the positive and negative examples. But for some problems the best decision boundary is not linear. The trick to use SVMs in non-linear cases is to map the data into a higher dimensional space where it is possible to separate the two classes with a hyperplane. Note that we still use the linear SVM but in a higher dimensional space. In fact it is not necessary to apply the mapping to the data points. Notice that the data $\mathbf{x}_i$ only appear as the dot product form in the *dual* Lagrangian. So, for a mapping $\Phi : \mathbf{R}^d \to H$ which maps the data into higher (possibly infinite) dimensional space H if we find a function $\mathbf{K}$ such that

$$\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \mathbf{x}_j \tag{2.13}$$

we do not have to do the mapping explicitly. Such a function $\mathbf{K}$ is called the kernel function. There are many kernel functions while the most well known ones being the *polynomial* kernel (Equation 2.14), Gaussian radial basis function (RBF) kernel (Equation 2.15) and the hyperbolic tangent kernel (Equation 2.16).

$$K(x, y) = (y \cdot x + 1)^p \tag{2.14}$$

$$K(x, y) = e^{(} - \|x - y\|^2/2\sigma^2) \tag{2.15}$$

$$K(x, y) = \tanh(\kappa \mathbf{x} \cdot \mathbf{y} - \gamma) \tag{2.16}$$

The Gaussian RBF kernel maps the data to an infinite dimensional space. With sufficiently small width of (i.e. small σ) Gaussian RBF is capable of classifying an arbitrarily large number of training points correctly. For a more detailed review on SVMs, one can refer to [10].

2.2 Biological Background

Proteins are essential molecules of life. They are not only the building blocks that constitute the structure of the cells and tissues, they also execute nearly all cell functions. Some proteins act as enzymes to catalyze chemical reactions. Others may work as transporters or carry messages from one cell to another. Yet, some others work as tiny molecular machines with moving parts. Which of these functions the protein performs is determined by its unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. Basically there are 20 types of amino acids encoded in DNA. All amino acids possess common structural features, including an α carbon to which an amino group, a carboxyl group, and a variable side chain are bonded. The side chains of the standard amino acids have different physical and chemical properties that produce three-dimensional protein structure and are therefore critical to protein function. Table 2.1 shows the list of amino acids and some important physiochemical features. Depending on the polarity of the side chain, amino acids vary in their hydrophilic or hydrophobic character. These properties are important in protein structure and protein-protein interactions. The importance of the physical properties of the side chains comes from the influence this has on the amino acid residues' interactions with other structures, both within a single protein and between proteins. The distribution of hydrophilic and hydrophobic amino acids determines the structure of the protein, and their physical location on the outside structure of the proteins influences their interaction with other proteins. Sometimes amino acids in a protein sequence may be replaced by other amino acids through genetic changes such as mutation. Some of these changes have detrimental effects on the function of a protein. But there are also biologically plausible replacements which causes proteins with different sequences performing the same function. Thus, sequence similarity is widely used to infer structural and functional similarity.

The genetic code that encodes the sequence information of the proteins is composed of a set of three consecutive nucleotides called *codons*. Each codon represents an amino acid. Since there are 4 different nucleotides in the DNA, there are 64 possible codons. Some amino acids are represented by more than one codon and some codons have special meanings such as stop codons. The gene encoded in DNA first *transcribed* into messenger RNA (mRNA). This mRNA is a template for protein synthesis in ribosome. The process of synthesizing the protein at the ribosome from an mRNA template is called the *translation*. mRNA is read 1 codon at a time and the amino acid carried by a transfer RNA (tRNA) which has the corresponding *anti-codon* is added to the growing polypeptide. Chain length of proteins may vary from a few hundred to a few thousand amino acids. Largest known proteins are the titins with a total length of almost 27000 amino acids.

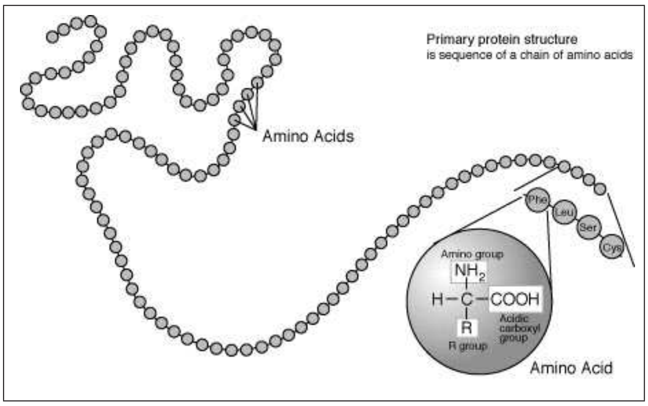


Figure 2.2: Protein primary structure is a sequence of amino acids ([55]).

Table 2.1: Amino acids.

Amino Acid	1-Letter	Side chain polarity	Side chain acidity	Hydropathy index
Alanine	A	nonpolar	neutral	1.8
Arginine	R	polar	basic (strongly)	-4.5
Asparagine	N	polar	neutral	-3.5
Aspartic acid	D	polar	acidic	-3.5
Cysteine	C	nonpolar	neutral	2.5
Glutamic acid	E	polar	acidic	-3.5
Glutamine	Q	polar	neutral	-3.5
Glycine	G	nonpolar	neutral	-0.4
Histidine	H	polar	basic (weakly)	-3.2
Isoleucine	I	nonpolar	neutral	4.5
Leucine	L	nonpolar	neutral	3.8
Lysine	K	polar	basic	-3.9
Methionine	M	nonpolar	neutral	1.9
Phenylalanine	F	nonpolar	neutral	2.8
Proline	P	nonpolar	neutral	-1.6
Serine	S	polar	neutral	-0.8
Threonine	T	polar	neutral	-0.7
Tryptophan	W	nonpolar	neutral	-0.9
Tyrosine	Y	polar	neutral	-1.3
Valine	V	nonpolar	neutral	4.2

The linear amino acid chain of the protein is called the *primary structure* of the protein (see Figure 2.2). This primary structure then folds into some local repeating structures stabilized by *hydrogen bonds* which is called the *secondary structure* of proteins. Most common examples of secondary structures are the alpha helices and beta sheets. These secondary structures folds into the overall shape of the protein stabilized by non-local interactions. This overall shape of a single protein is called the *tertiary structure* (see Figure 2.3). There is also the *quaternary structure* which is actually a protein complex formed by the interaction of more than one protein.

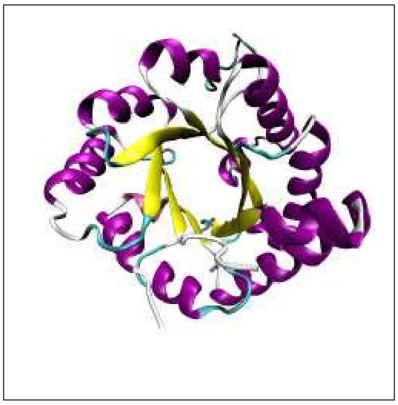


Figure 2.3: 3-dimensional structure of the protein triose phosphate isomerase. Alpha helices are colored by purple and beta sheets are colored with yellow ([55]).

The function of a protein is mainly determined by the 3 dimensional shape and the physiochemical properties of the amino acids at the specific positions of this 3 dimensional shape of the protein. A thorough understanding of biological systems is only possible through identifying the functions and interactions of these essential molecules of life.

CHAPTER 3

SUBSEQUENCE PROFILE MAP (SPMAP)

Subsequence-based methods employs conserved subsequences among a class of proteins. The main idea is that, conserved subsequences among different proteins are strong indicators of functional or structural similarity because functionally important regions (catalytic sites, binding sites, structural motifs) are conserved over much wider taxonomic distances than the sequences themselves. Thus, in subsequence-based approach feature vectors are constructed according to the existence of specific motifs or domains in the protein sequences. The critical step in this approach is the extraction and selection of motifs. One possibility is to use motif information from protein databases ([7, 54]) in which motifs are assumed to be already available for the family of proteins to be classified. In contrast, most of the methods of subsequence-based approach attempt to extract motifs explicitly for the given families ([27, 53, 38, 36, 8]). Although motifs are powerful discriminators even in low similarity (remote homology) situations, motif finding is a very difficult task, especially for protein sequences since there are 20 different amino acids and many biologically plausible mutations. Multiple sequence alignments and other computational pattern extraction algorithms are often employed for motif finding. Unfortunately, algorithms that can find optimal solutions in all of these methods have exponential time complexities, hence approximation or heuristic algorithms are used instead. As a consequence, there is always the risk of missing some relatively implicit motifs. Furthermore, classical motif finding algorithms find a specified number of motifs even if there are not that many biological motifs in the family. These insignificant additional motifs might reduce the accuracy of the classification. One other issue is that, depending on the classification task, proteins to be classified might not have a common motif at all. As an example, in the problem of subcellular localization, when discriminating cytosolic proteins, it is not possible to find motifs specific to this class. Methods that consider overall sequence similarity may perform better in such cases.

In the following sections, we describe a feature space mapping, called *subsequence profile map* (SPMap), that takes into account the information coming from the subsequences of a protein. Our approach incorporates the information coming from important subregions that are conserved over a family of proteins as well as the overall sequence similarity. Instead of focusing on function specific motifs, SPMap considers all of the subsequences as a distribution over a quantized space by discretizing and reducing the dimension of an otherwise huge space of all possible subsequences.

3.1 Systems and Methods

The system described in this study is based on a discriminative method which requires positive and negative examples to classify and annotate proteins whose functions are not known. Instead of looking for the overall similarity of protein sequences, we make use of the distribution of short subsequences of a given protein over a subsequence profile map. We generated the profiles using all possible fixed-length subsequences of the protein sequences in the positive training set. Similar subsequences were clustered together and clusters were represented as probabilistic profiles. The major reasoning behind this approach is that, subsequences extracted from the conserved regions are more frequent than any other subsequence extracted from the positive training data. If the frequent subsequences are represented as dimensions of feature vectors, discriminative methods can make use of this information. If there is a conserved motif or a domain in the given sequences or there is an overall similarity between sequences, they would produce similar distributions on the profile map. Classifiers such as Support Vector Machines (SVM) may then identify these similar distributions and hence improve the classification accuracy.

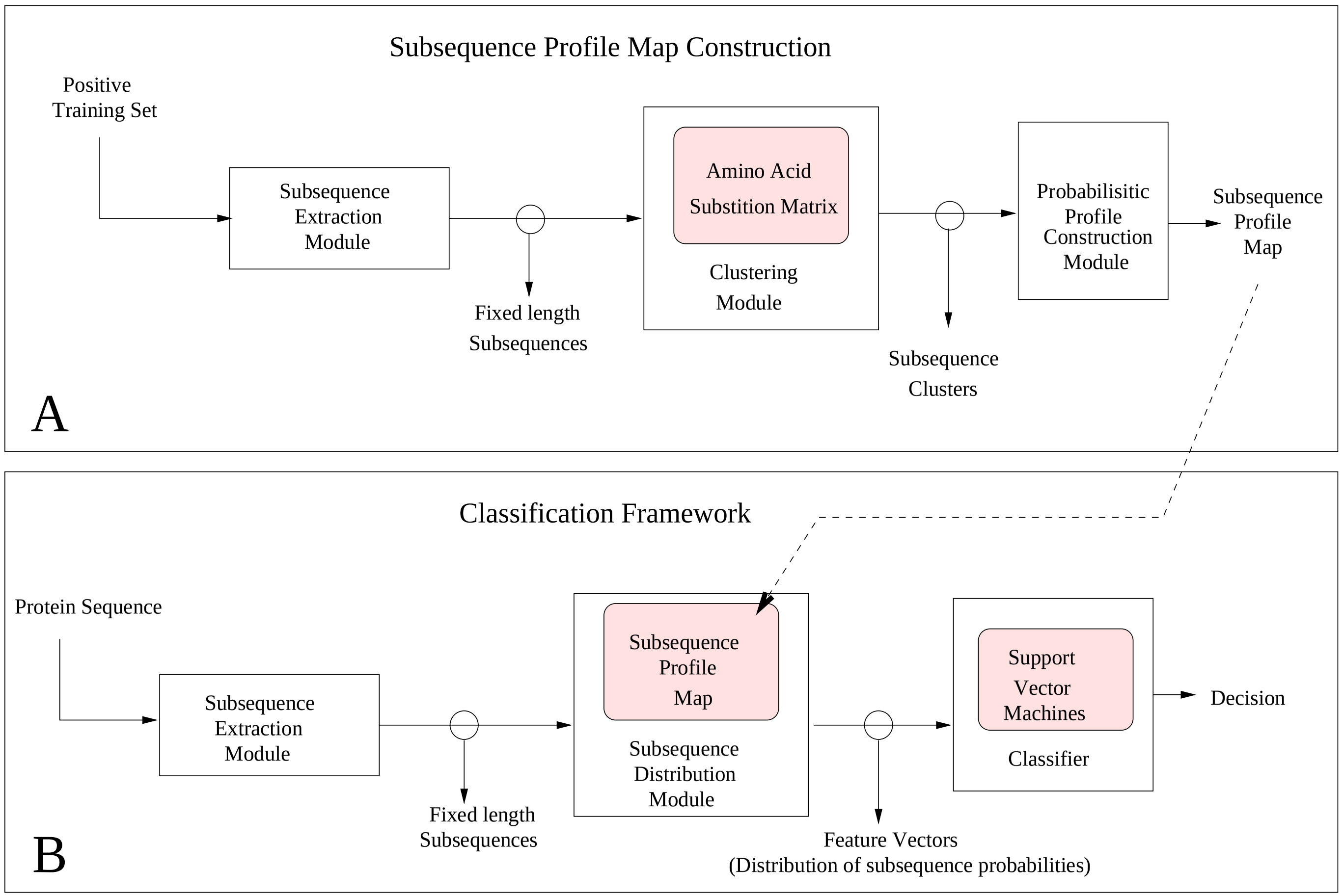

Subsequence Profile Map Construction
Positive Training Set
Subsequence Extraction Module
Fixed length Subsequences
Amino Acid Substition Matrix
Clustering Module
Subsequence Clusters
Probabilisitic Profile Construction Module
Subsequence Profile Map
A
Classification Framework
Protein Sequence
Subsequence Extraction Module
Fixed length Subsequences
Subsequence Profile Map
Subsequence Distribution Module
Feature Vectors (Distribution of subsequence probabilities)
Support Vector Machines
Classifier
Decision
B

Figure 3.1: SPMap flow diagram . (**A**) Subsequence profile map construction: subsequences of the proteins in positive training set are clustered to construct subsequence profile map. (**B**) Classification: constructed profile map is utilized to find the feature space representation of the protein sequence to be classified.

In order to perform the classification, SVMs were used. We constructed fixed dimensional vectors that represent the subsequence distribution information. There are 2 critical steps in SPMap as shown in Figure 3.1:

A. subsequence profile map construction,

B. feature vector generation and classification.

3.1.1 Subsequence Profile Map Construction

In SPMap, feature space representation of a protein sequence is the distribution of its subsequences over a map of generative models. General framework for finding this generative feature map is summarized as follows.

- **Subsequence Extraction Module**: Extract all possible subsequences of a given length from positive training sequences.
- **Clustering Module**: Cluster similar subsequences by an appropriate clustering method.
- **Profile Construction Module**: Build a model for each cluster.

The important step here is the clustering of subsequences. Note that the space of all possible subsequences of length l is of size 20^l, since there are 20 possible amino acids. Instead of working in this very high dimensional space, we quantized this space using the clusters of subsequences that are actually existing in the positive training examples. Using all possible subsequences would increase the number of resulting clusters hence the dimension of the feature space. Most of these feature dimensions would be representing subsequences irrelevant to the classification problem in hand. In turn, because of curse of dimensionality, the necessary number of training samples will increase exponentially. But training data is very few for some classification problems. As an alternative, one might think of using positive and negative training data for clustering. The problem here is that negative training data do not actually represent all possible negative examples of a functional class. Also, negative examples are coming from a very wide range of different functional classes each one having very different discriminative subregions. Thus, different negative examples will have higher signature in different feature dimensions. This will again increase the number of training examples necessary for training a successful classifier. Using only positive training data is a useful in terms of efficiency and in terms of increasing discriminative accuracy.

One should note that, as we clustered the subsequences, we were not actually looking for underlying groupings. The aim here was to generate a meaningful quantization of the subsequence space that especially represent groups of frequent and similar subsequences in the positive training data. These subsequences might have been conserved because of their importance for the function of that class of proteins and we wanted our feature space to take them into account. Thus, we do not try to increase the inter-cluster distances but reduce the intra-cluster distances. Clustering algorithm is given in Algorithm 1. It is similar to the average link hierarchical clustering, however it can be implemented very efficiently without calculating all the pairwise distances. Initially, the number of clusters is set to 0. Each subsequence is compared against all of the existing clusters and average similarity to the elements of each cluster is calculated. A subsequence is assigned to the cluster, C_{max}, which gives the maximum average similarity value. If the similarity to C_{max} is less than a threshold, δ, a new cluster is created and the subsequence is assigned to the new cluster. Similarity between two subsequences x and y was calculated by the formula

$$s(x,y) = \sum_{i=1}^{l} M(x(i), y(i)) \tag{3.1}$$

where l is the length of the subsequences and $M(x(i), y(i))$ is the value in the similarity matrix for the ith elements of x and y. For M, we used an amino acid similarity matrix, since it allows us to incorporate evolutionary information in finding and representing important conserved regions of a family of proteins. The final number of clusters depend on the threshold value δ. If it is set to a high value, clusters will be smaller only allowing very similar subsequences and the total number of clusters will be high. If it is set to a low value, biologically unrelated subsequences might end up in the same cluster.

Threshold δ is critical to allow only biologically similar subsequences to appear in the same cluster. The expected value of similarity between two random subsequences of length l is

$$E[s(x,y)] = E[M] * l \tag{3.2}$$

Algorithm 1 Clustering Algorithm

```
$X \leftarrow$ all fixed length subsequences of the positive training set
$C \leftarrow \{\}$
for all $x_i \epsilon X$ do
    for all Clusters $C_k$ do
        $s_k(x_i) = \frac{\sum_{x_j \epsilon C_k} s(x_i,x_j)}{|C_k|}$
    end for
    $m = argmax_{k=1..|C|} s_k(x_i)$
    if $s_m > \delta$ then
        Add $x_i$ to $C_m$
    else
        Create a new cluster $C_{|C|+1}$ and add $x_i$ to $C_{|C|+1}$
    end if
end for
```

where $s(x, y)$ is the similarity between two random subsequences x and y and $E[M]$ is the expected similarity of two random amino acids using similarity matrix M. $E[M]$ can be calculated as

$$E[M] = \frac{\sum_{i=1}^{20} \sum_{j=1}^{20} M(i, j)}{20 \times 20} \tag{3.3}$$

since there are 20 different amino acids. For M we used amino acid similarity matrix Blosum62 and $E[M]$ for Blosum62 is -1.0650 (see Appendix A). For our 5 length subsequences, expected value of two random subsequences $[E(s(x, y))]$ is -5.325. The maximum diagonal entry of M is 11 and the minimum diagonal entry of M is 4. So the minimum possible similarity value for two equal subsequences of length 5 is 5×4=20. Threshold δ should be more than the similarity of a random alignment but less than the minimum possible similarity value of an exact match. If we require an exact match at least in 3 position and mismatches at 2 positions, threshold value must be around 9. We tested with integer threshold from 6 to 10 and 8 turned out to be a best choice for most of the tests in terms of classification accuracy. In addition, when 8 is chosen for δ the number of clusters stayed at an acceptable level. The number of clusters with different

Table 3.1: Number of clusters formed with different threshold δ values and different training data size.

Datasize	**Threshold**	**Number of Clusters formed**
77	6	612
	7	859
	8	1223
	9	1668
	10	2211
353	6	677
	7	1067
	8	1604
	9	2357
	10	3435
968	6	718
	7	1102
	8	1680
	9	2555
	10	3864

training data sizes and different threshold is given in Table 3.1. The cluster size, thus the feature space dimension is mostly dependent on the threshold value δ. The effect of size of the training data is negligible compared to the effect of δ. As the δ increases, the partitioning of the subsequence space becomes more fine-grained. Consequently, biologically related subsequences which might have been safely aligned may result in having effects on different dimensions of the feature space. In contrast, if the partitioning of the subsequence space is more coarse-grained, then biologically unrelated subsequences will have same effects on the same dimensions of the feature space. Also, unnecessary increase in the size of the feature space would result in poor SVM performance because of the curse of dimensionality. Experimental results showed that 8 is the best choice for the δ for subsequence length 5.

After the clustering step, we generated a probabilistic profile for each cluster. A probabilistic profile PP_k for cluster k, is an $l \times 20$ matrix, where l is the length of a subsequence. Entry $P_k(i, j)$ of this matrix represents the probability of amino acid j to occur at the i^{th} position of the subsequence. Given a cluster C_k, the profile for this cluster is calculated by Equation 3.4.

$$PP_k(i, j) = \log \frac{\phi_k(i, j) + \kappa}{|C_k|} \tag{3.4}$$

where $\phi_k(i, j)$ represents the count of the amino acid j at position i of the subsequences in C_k. We added a pseudo-count κ for amino acids at each position to avoid over-fitting and zero probabilities. Actually, we took the log of the profiles and worked with log-probabilities in the conversion step.

3.1.2 Feature Vector Generation

Proteins were represented in the feature space as the distribution of their subsequences over the generated subsequence profile map. All the subsequences of a protein were extracted to construct a feature vector. Each subsequence x was compared with each probabilistic profile PP_k and a probability was calculated as

$$P(x|PP_k) = \sum_{i=0}^{l} PP_k(i, x(i)). \tag{3.5}$$

The value for the k^{th} dimension of the feature vector V is set to

$$V(k) = \max_{x_i \in S} P(x_i|PP_k), \tag{3.6}$$

the probability of highest scoring subsequence of protein S on probabilistic profile PP_k. This algorithm is similar to the vector generation algorithm presented in [8] with the difference that we set $V(k)$ to 0 if the probability is very small.

3.1.3 Classification

Once the protein sequences are mapped onto the feature space, any numerical machine learning tool can be employed. Our choice was to use SVMs since they are experimentally proven to be successful for various problems ([14]). Radial basis function (RBF) was chosen as the kernel for SVM. In all of the experiments, SVM parameter C and RBF kernel parameter γ were fixed to be 2 and 0.05, respectively. SVM-light software was used for learning and classification steps ([32]).

3.1.4 Experimental Setup

In all of the experiments, Blosum62 matrix was employed to calculate the similarity between subsequences ([29]) although it is possible to use different similarity matrices depending on the sequence divergence or the taxonomic distance between the proteins to be classified ([5, 52]). Blosum62 is shown to be useful for a wide range of problems and is the default selection for most of the alignment tools ([1, 2]). Length of the subsequences was set to 5. Setting the subsequence length to 5 did not mean that we sought for motifs of 5 amino acid length. In SPMap, motifs were the overall distribution of the subsequences over the profiles constructed from resulting 5 length subsequence clusters. Hence subsequence length 5 allowed us to capture longer motifs as a distribution over more than one profile. Subsequences shorted than 5 have a larger probability of appearing by chance in functionally unrelated proteins. Thus, 5 is a generally accepted to be a minimum length to represent a motif and is used in other subsequence-based methods ([37]). We tested the performance of SPMap by changing the subsequence length in the interval [5,12] on selected sample sets of data. We observed that although there were differences in the performance with respect to the change in the subsequence length, 5 was the optimal in the sense of performance versus computational complexity. Threshold similarity score δ in Algorithm 1 was fixed to 8 where the expected similarity score of two random subsequences of length 5 using Blosum62 matrix is -5.325. Compared to the expected value, 8 is high enough to disallow random similarities. Extensive tests with different threshold values showed that 8 performed better in most of the test cases and it was set as default in all of the experiments.

3.2 Results

First, in order to see SPMap's ability to capture shared motifs and the overall similarity, we prepared a profile map using 8 G-Protein coupled receptor (GPCR) Kinase proteins

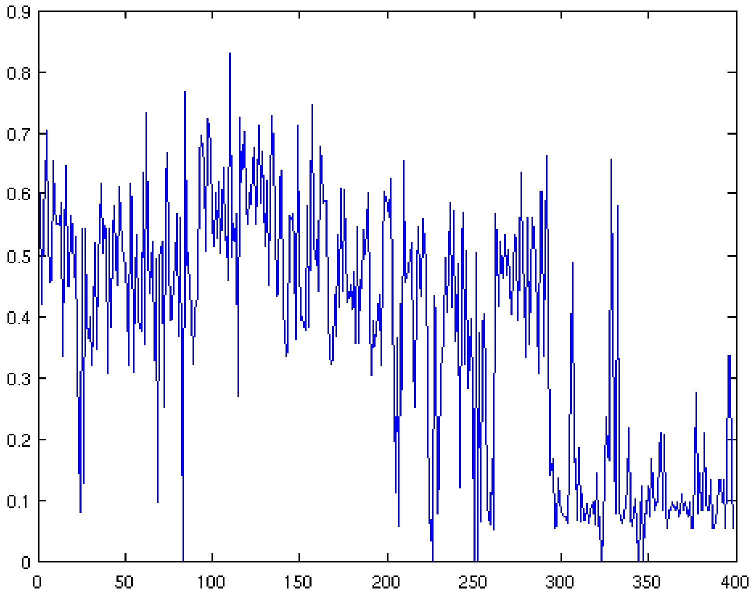


Figure 3.2: Feature vector representation of GRK4_HUMAN which is a GPRC Kinase protein. Profile map was generated by using 8 different GPCR Kinase proteins.

which share GPCR kinase motifs. 399 clusters were formed hence we had 399 profiles in our feature map. Using this profile map, we converted two held out GPCR Kinase proteins, namely *GRK4_HUMAN* and *GRK5_RAT*, to 399 dimensional feature vectors. Note that these proteins have more than 80% sequence identity to each other. Figure 3.2 and Figure 3.3 show feature mapping of *GRK4_HUMAN* and *GRK5_RAT* respectively. It can be seen that similar proteins have similar fingerprints on the same profile map. It is also important to note that values of the feature vector dimensions is generally high on most of the 399 profiles. This is especially true for first 200 profiles. We can conclude that the subsequences extracted from these two proteins are well represented by the profiles from 1 to 200. Having high values for most of the vector dimensions also means that these two proteins have high similarity to the proteins that are used to construct the profile map. Figure 3.4 shows the feature vector representation of an *androgen receptor protein* which has less than 30% sequence identity to the GPCR kinase proteins used to construct the profile map. It also has less than 30% sequence identity to *GRK4_HUMAN* and *GRK5_RAT*. It is clear that *androgen receptor protein* has a different distribution than both of the GPCR Kinase proteins and also if we look at the the vector dimensions, the

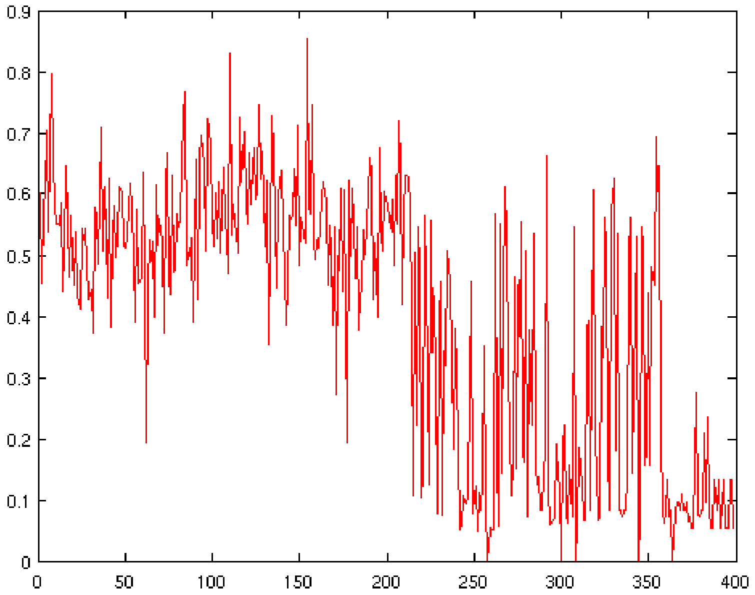


Figure 3.3: Feature vector representation of GRK5_RAT which is a GPRC Kinase protein. Profile map was generated by using 8 different GPCR Kinase proteins.

values are generally low. We also converted a known single GPCR Kinase motif, **KFSTGSVPIPWQNEMIET**, to feature vector representation. The resulting vector is shown on Figure 3.5. Note that on both *GRK4_HUMAN* and *GRK5_RAT* the corresponding dimensions that produce high probability with the given motif is also high. The main profiles representing this 18 amino acid long motif turned out to be profiles from 196 to 203. On these profiles, GPCR Kinase motif and the GPCR Kinase proteins produced an average value of 0.55. On the contrary, these dimensions are low with *androgen receptor protein* where the average value for the mentioned profiles is 0.09. Finally we inserted this motif into a random position of *androgen receptor protein* and converted this resulting hypothetical protein into feature vector representation (Figure 3.6). One can clearly see that SPMap successfully captures the motif in this hypothetical protein which still has very low sequence similarity to GPCR Kinase proteins. And we observed that SVM is capable of capturing these discriminative dimensions which are high on positive training data and low on negative training data. This property renders SPMap a powerful classifier even in remote homology situations.

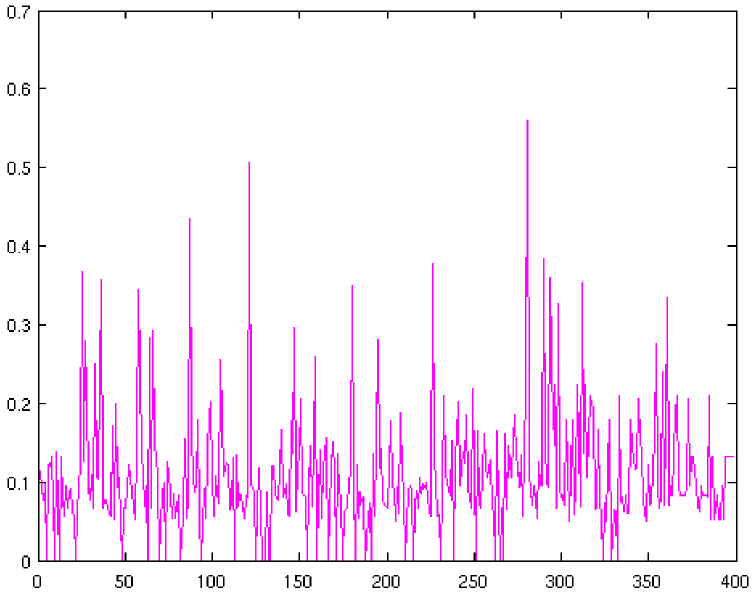


Figure 3.4: Feature vector representation of androgen receptor alpha. Profile map was generated by using 8 different GPCR Kinase proteins.

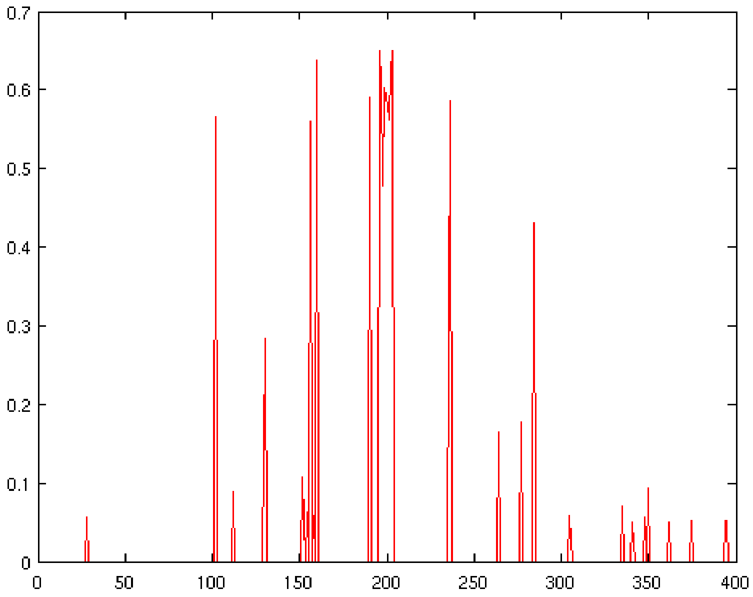


Figure 3.5: Feature vector representation of GPRC Kinase motif **KFSTGSVPIPWQNEMIET**. Profile map was generated by using 8 different GPCR Kinase proteins.

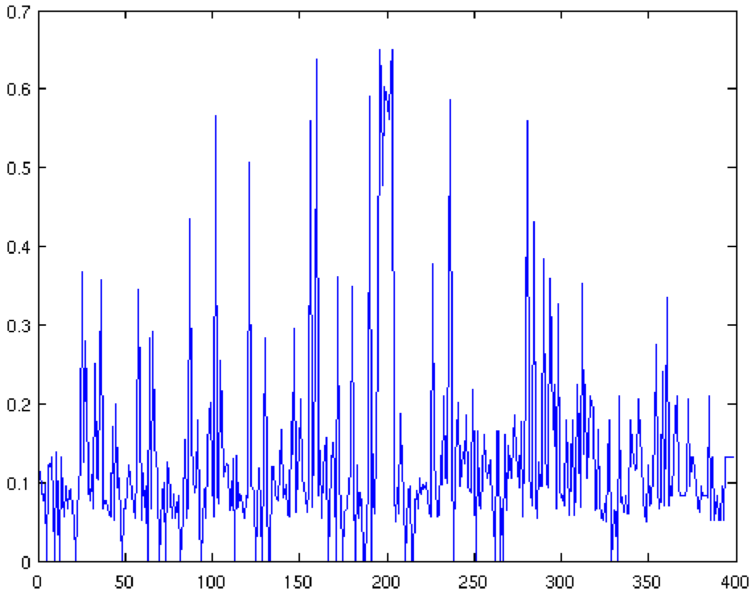


Figure 3.6: Feature vector representation of the hypothetical protein that is formed by inserting a GPCR Kinase motif into a random position of androgen receptor alpha protein. Profile map was generated by using 8 different GPCR Kinase proteins.

3.2.1 Subcellular Localization

The idea of subsequence distribution was first proposed in P2SL ([5]). However, we developed more robust, reliable and efficient method for this idea. In order to be able to show the improvement, we first performed tests on the subcellular localization dataset on which P2SL was trained and tested. Dataset was composed of 4 different classes, namely ER targeted (ER), cytoplasmic (C), mitochondrial (M) and nuclear (N) ([5]). ER targeted and mitochondrial proteins have signal peptides of length 25 and 35 amino acids respectively, at the N-terminal of the proteins. While extracting subsequences for feature map construction we used first 30 amino acids for ER targeted proteins and first 40 amino acids for mitochondrial proteins. Two types of tests were performed. First, in a one-versus-all setting, the areas under the Receiver Operator Characteristic (ROC) curve, which is called the ROC scores, were calculated for each localization and results are given in Table 3.2. ROC score is a measure of discriminative power of a classifier independent of the threshold parameter.

In the second test case, classifiers for each localization were combined using the

Table 3.2: Average ROC scores and standard deviations for subcellular localization predictions.

Localization	**Data Size**	**Mean ROC**	**Std. Dev.**
ER targeted	3115	0.97	0.006
Cytoplasmic	1789	0.95	0.005
Mitochondrial	1148	0.96	0.006
Nuclear	2225	0.96	0.005

Table 3.3: Confusion matrix representing average percentage results of 4-fold prediction tests compared with P2SL results.

Actual		Predicted Label			
		N	C	M	ER
		%	%	%	%
N	*SPMap*	**89.83**	7.5	1.1	1.58
	P2SL	**75.34**	19.94	3.29	1.43
C	*SPMap*	7.14	**89.05**	1.8	2.02
	P2SL	14.66	**79.33**	3.65	2.36
M	*SPMap*	2.09	5.4	**89.29**	3.22
	P2SL	3.31	7.23	**83.80**	5.66
ER	*SPMap*	2.07	2.5	1.41	**94.03**
	P2SL	4.89	6.19	3.29	**85.63**

winner-take-all principle. Each test sample was assigned to the location whose classifier produced the highest SVM score. The confusion matrix obtained by averaging 4-fold cross-validation tests and their comparison with P2SL results are given in Table 3.3 ([5]).

3.2.2 G-protein Coupled Receptor Subfamily Classification

Tests are subsequently carried on G-protein coupled receptor (GPCR) subfamily classification problem that was extensively studied in the literature. Consequently, GPCR subfamily classification constitutes a good benchmark dataset for comparing with other

Table 3.4: Comparison of Accuracy of Various Classifiers at GPCR Level I and II Subfamily Classification.

Classifier	**Level I Accuracy** %	**Level II Accuracy** %
BLAST	83.3	74.5
Decision Tree	77.3	70.8
Fisher-SVM	88.4	86.3
kernNN	64.0	51.0
Naïve Bayes	93.0	92.4
SAM-T2K HMM	69.9	70.0
SPMap	**95.4**	**93.8**

methods. For GPCR subfamily classification, we used the dataset presented in [33] to compare with the results of various classifiers presented in [33] and [12]. Same train and test splits were used for 2-fold cross validation for fairness of comparison. SPMap was tested on level I and level II subfamily classification of GPCR proteins. In level I subfamily classification, there were 1269 sequences from 19 subfamilies within classes A and C in addition to 149 non-GPCR sequences. In level II subfamily classification, there were 1170 GPCR sequences from 70 different level II subfamilies. Some of the sequences in level I subfamily classification have no level II subfamily classification and some of the level II subfamilies only have one protein so they are grouped as other sequences with non-GPCR sequences. Datasets and train and test splits are available at [26].

The comparison of accuracy of various classifiers and SPMap is presented in Table 3.4. Fisher-SVM, BLAST, SAM-T2K HMM, and kernNN methods were presented in [33] and Decision Tree and Naïve Bayes methods were presented in [12].

3.2.3 Enzyme Class Classification

Finally we evaluated the performance of SPMap on enzyme class classification. Enzymes play a central role in many of the biological functions in a cell. They are indispensable for understanding the molecular systems in a cell and are important drug targets. Hence

Table 3.5: Comparison of success rates of various classifiers on 6 major enzyme classes calculated with leave-one-out cross-validation.

Classes		**Lu et al**	**Blast**	**Psi-Blast**	**SVM-Prot**	**SPMap**
	total	Success(%)	Success(%)	Success(%)	Success(%)	Success(%)
Oxidoreductase	436	**93.53**	89.68	91.06	73.62	80.73
Transferase	832	**93.63**	88.46	87.98	82.45	66.23
Hydrolase	741	**94.20**	86.10	86.77	77.33	71.93
Lyase	170	75.29	75.29	70.59	68.82	**94.12**
Isomerase	114	74.56	73.68	73.68	68.42	**96.49**
Ligase	150	89.33	**90.00**	88.67	37.33	88.00

accurate classification is very important in enzyme research.

Dataset for enzyme classification is extracted from Brenda database ([47]). Enzyme Commission of International Congress of Biochemistry developed a numerical classification scheme for enzymes based on the chemical reactions they catalyze. Each enzyme is described by a sequence of 4 numbers (Enzyme Commission (EC) numbers) resulting from a 4 level hierarchy where first number specifies the most general class and the last one specifies the most specific. At the highest level there are 6 major classes of enzymes. Automated prediction methods are successfully applied to enzyme classification according to the first ([39] and second level of EC numbers ([11]). We also performed tests according to the first and second EC numbers. On the first level there are 6 major classes of enzymes. The dataset used for this level is presented in [39]. Each class is filtered so that there are no pair of proteins with more than 25% sequence identity. The success rates for various methods and SPMap for 6 classes with leave-one-out cross-validation is presented in Table 3.5.

We also classified proteins according to their first two EC numbers, resulting in 56 classes. We omitted classes with very few members. Sensitivity and specificity values calculated over 4-fold cross validation are presented in Table 3.6. This classifier for 56 enzyme classes is available as an online service at [50].

Table 3.6: Sensitivity ($TP/(TP + FN)$) and Specificity ($TN/(TN + FP)$) values for 56 Enzyme Class classifiers calculated over 4-fold cross validation

Enzyme Class	**Data Size**	**Sensitivity**	**Specificity**
EC 1.1 Acting on the CH-OH group of donors	8878	95.33	85.05
EC 1.2 Acting on the aldehyde or oxo group of donors	4099	91.63	97.17
EC 1.3 Acting on the CH-CH group of donors	2455	85.75	98.09
EC 1.4 Acting on the CH-NH2 group of donors	1573	88.64	99.74
EC 1.5 Acting on the CH-NH group of donors	1244	81.35	99.72
EC 1.6 Acting on NADH or NADPH	5572	94.54	95.85
EC 1.7 Acting on other nitrogenous compounds as donors	802	83.67	99.93
EC 1.8 Acting on a sulfur group of donors	1699	89.94	99.82
EC 1.9 Acting on a heme group of donors	1620	93.99	98.51
EC 1.10 Acting on diphenols and related substances as donors	813	86.86	99.98
EC 1.11 Acting on a peroxide as acceptor	1267	91.56	99.97
EC 1.12 Acting on hydrogen as donor	243	68.89	99.97
EC 1.13 Acting on single donors / with incorporation of molecular oxygen (oxygenases)	1048	87.66	99.97
EC 1.14 Acting on paired donors, with incorporation / or reduction of molecular oxygen	1909	83.3	98.42
EC 1.15 Acting on superoxide radicals as acceptor	935	93.56	99.99
EC 1.16 Oxidizing metal ions	142	65.71	99.96
EC 1.17 Acting on CH or CH2 groups	1063	90.31	99.92
EC 1.18 Acting on iron-sulfur proteins as donors	745	91.94	99.97
EC 1.20 Acting on phosphorus or arsenic in donors	66	66.67	99.99
EC 1.21 Acting on X-H and Y-H to form an X-Y bond	60	88.89	100
EC 1.97 Other oxidoreductases	169	80.95	99.99
EC 2.1 Transferring one-carbon groups	6061	92.28	90.97
EC 2.2 Transferring aldehyde or ketonic groups	1058	94.32	99.94
EC 2.3 Acyltransferases	6149	92.52	91.55
EC 2.4 Glycosyltransferases	6004	92.65	89.54
EC 2.5 Transferring alkyl or aryl groups, other than methyl groups	5188	93.94	96.73
EC 2.6 Transferring nitrogenous groups	2011	95.22	99.85
EC 2.7 Transferring phosphorus-containing groups	23424	89.78	91.08
EC 2.8 Transferring sulfur-containing groups	982	87.35	99.91
EC 2.9 Transferring selenium-containing groups	72	88.89	100
EC 3.1 Acting on ester bonds	9879	74.79	96.05
EC 3.2 Glycosylases	4789	93.76	91.98
EC 3.3 Acting on ether bonds	363	84.44	99.97
EC 3.4 Acting on peptide bonds (peptidases)	5945	93.4	87.48
EC 3.5 Acting on carbon-nitrogen bonds, other than peptide bonds	5942	90.28	88.25
EC 3.6 Acting on acid anhydrides	7430	96.23	88.22
EC 3.7 Acting on carbon-carbon bonds	66	81.25	100
EC 3.8 Acting on halide bonds	101	49.33	99.98
EC 4.1 Carbon-carbon lyases	7606	93.77	87.95
EC 4.2 Carbon-oxygen lyases	7211	93.23	87.46

Continued on Next Page...

Table 3.6 – Continued

Enzyme Class	Data Size	Sensitivity	Specificity
EC 4.3 Carbon-nitrogen lyases	1264	91.14	99.89
EC 4.4 Carbon-sulfur lyases	626	82.91	99.8
EC 4.6 Phosphorus-oxygen lyases	614	91.28	99.9
EC 4.99 Other lyases	297	90.99	99.98
EC 5.1 Racemases and epimerases	2030	92.18	99.66
EC 5.2 cis-trans-Isomerases	1232	92.86	99.92
EC 5.3 Intramolecular isomerases	2910	90.65	99.18
EC 5.4 Intramolecular transferases (mutases)	2195	88.57	99.37
EC 5.5 Intramolecular lyases	135	71.72	99.98
EC 5.99 Other isomerases	1418	95.57	99.96
EC 6.1 Forming carbon—oxygen bonds	6285	97.05	98.39
EC 6.2 Forming carbon—sulfur bonds	1112	93.17	99.91
EC 6.3 Forming carbon—nitrogen bonds	6784	94.53	95.25
EC 6.4 Forming carbon—carbon bonds	785	94.9	99.87
EC 6.5 Forming phosphoric ester bonds	433	89.2	99.97
EC 6.6 Forming nitrogen—metal bonds	118	90.81	99.97

3.3 Discussion

3.3.1 Computational Complexity

SPMap is composed of two main parts. First part is the subsequence profile map construction. It is only performed once for a new classifier to be trained. Hence, its efficiency does not affect the performance during the classification of new sequences. The most expensive part of the map construction is the clustering of subsequences. Most of the standard clustering algorithms require numerical vectors to work on. More specifically, they require a metric to calculate the distance between the cluster representations and data points and a method to update these cluster representations throughout the course of the algorithm. These methods usually perform $O(nk)$ distance calculations where n is the number of data points and k is the number of clusters. They require the number of clusters k to be given at the start. There are also clustering algorithms that use only pairwise distances between data points. They don't require the number of clusters k as a parameter but they have to perform $O(n^2)$ pairwise distance calculations and that might be very inefficient in terms of time and memory for large n. Note that n in this case is the total number of subsequences extracted from all of the positive training examples, which is roughly the number of amino acids in the positive training examples. However, Algorithm 1 can be implemented in $O(nk)$. The critical step is the calculation of the average distance of subsequence x_i to the cluster u given in Equation 3.7.

$$s_u(x_i) = \frac{\sum_{x_j \in C_u} s(x_i, x_j)}{|C_u|} \tag{3.7}$$

With this definition, Algorithm 1 requires n^2 pairwise subsequence similarity calculations. Combining Equation 3.1 and Equation 3.7 we get

$$\psi(x, C_k) = \frac{\sum_{x_j \in C_k} \sum_{t=1}^{l} M(x(t), x_j(t))}{|C_k|} \tag{3.8}$$

$$\psi(x, C_k) = \sum_{t=1}^{l} \sum_{x_j \in C_k} \frac{M(x(t), x_j(t))}{|C_k|} \tag{3.9}$$

The inner sum in Equation 3.9 only depends on the counts of amino acids at each position of the subsequences in cluster C_k not the subsequences themselves. So $\psi(x, C_k)$ can be re-written as

$$\psi(x, C_k) = \sum_{t=1}^{l} \sum_{i=1}^{20} \frac{\phi_k(t, i)}{|C_k|} M(x(t), i) \tag{3.10}$$

where $\phi_k t, i$ is the count of amino acid i at position t of subsequences in cluster C_k. $\frac{\phi_k(t,i)}{|C_k|}$ is nothing but the frequency of amino acid i at position k of subsequences in cluster C_k. Hence,

$$s_u(x_i) = \sum_{t=1}^{l} \sum_{j=1}^{20} f_u^t(a_j) M(x_i(t), a_j) \tag{3.11}$$

where $x_i(t)$ denotes the amino acid appearing at the t^{th} position of the subsequence x_i and $M(x_i(t), a_j)$ is the entry of similarity matrix for amino acids $x_i(t)$ and a_j. The number $f_u^t(a_j)$ represents the frequency of amino acid a_j at the t^{th} position of subsequences in cluster u. The complexity of the Algorithm 1 becomes $O(nkl)$ where l is the length of the subsequences, k is the number of clusters, and n is the total length of all of the proteins in positive training set. Since l, is an arbitrary but fixed parameter, it can be said that it is $O(nk)$ with respect to the number of the input sequences. k is dependent on the threshold value δ given in Algorithm 1; but it is around 1800 for the default δ value, 8. It is almost constant or varying very slowly with the data size. The second part of the presented method is construction of the feature vectors. Since the probability of each subsequence of the protein against all of the subsequence profiles must be calculated, it again can be implemented in $O(nk)$ time. In this case, n represents the length of the given protein to be mapped and k is the number of subsequence profiles. SPMap is linear in the size of the input data. It is very efficient and scalable to handle large datasets.

3.3.2 Performance Test Results

SPMap has a significant improvement over P2SL for subcellular localization classification. The improvement is both in terms of accuracy and computational efficiency. In order to discretize the subsequence space, P2SL uses self organizing maps (SOM) which are hard to train because of the necessity of large training data and convergence problems. As a result different runs on SOM might result in different feature spaces. P2SL is prone to missing some important subsequences since it does not consider all possible subsequences. Since SOM requires numerical vectors, P2SL encodes amino acids as 20 dimensional vectors which causes a 5 length subsequence to be represented as a 100 dimensional vector further complicating the SOM training. SPMap uses clusters of all possible subsequences for discretization of subsequence space instead of SOM in P2SL. Similarity between subsequences are calculated using an amino acid similarity matrix and standard string similarity calculation methods, avoiding high dimensional encoding of subsequences. One of the advantages of SPMap is that it works well on wide range of different classification tasks with the default parameter values. This makes it easier to use without expertise and optimization. Furthermore, our feature space mapping algorithm have only one parameter, the threshold value δ, which has a well performing default value in general.

We also investigated the performance of SPMap on functional classification tasks other than subcellular localization. In order to assess and compare the capabilities of SPMap, we performed tests on G-protein coupled receptor (GPCR) subfamily level classification. GPCRs are very important targets in drug design but known to be hard to classify, because they have highly diverse family at the sequence level ([41]). It can be seen that SPMap outperformed other classifiers in both level I and level II GPCR subfamily classification. To our knowledge, at the time of writing this paper, Naïve Bayes approach of [12] was the best performing method on the benchmark dataset presented in [33].

The application of SPMap on enzyme class classification demonstrated that our method too generates comparable or better results to those obtained by previous studies. The dataset used for the test on 6 major enzyme classes was filtered so that there are no pair of proteins with more than 25% sequence identity. This makes the classification task more difficult especially for the methods that only use sequence or subsequence similarity. Furthermore, SPMap depends solely on the available training data to generate the subsequence feature map, where the method presented in [39] uses domains that are already available in the databases. Nevertheless, results were interestingly complementary. SPMap achieved very high accuracy when the other methods performed poorly and vice versa. For the second level of enzyme hierarchy SPMap achieved high sensitivity in most of the classes. We used all the available data in 4-fold cross validation. As a result, a few classes with comparably large data sizes were biased towards false positives, hence relatively low specificity. Selecting a representative training subset for large classes might enhance the specificity of the classifier.

3.3.3 Perspectives

Since supervised discriminative methods model the differences between families of positive and negative examples explicitly, they provide better solutions for most of the problems of function classification. Most widely used discriminative method is the support vector machines (SVMs) combined with an appropriate kernel or feature space mapping ([14]). The main issue in classification of proteins according to their primary sequences is to find a kernel or a feature mapping that captures the information hidden in the important discriminative regions of the given sequences. Since, functionally important regions (catalytic sites, binding sites, structural motifs) are conserved over much wider taxonomic distances than the sequences themselves, conserved subsequences among different proteins are strong indicators of functional or structural similarity. Hence, SPMap pursued a new approach based on distribution of subsequences over a map constructed using the actual protein sequences in the positive training set.

The idea of constructing similarity graphs of subsequences and extracting motifs from the clusters of these graphs was already exploited for DNA sequences ([22]). In SPMap, we did not try to identify the motifs explicitly. We just let the classification algorithm learn which subsequence distributions are in fact discriminative. One advantage of SPMap is that it allows further investigation of these constructed profiles to identify motifs of positive training family. As a feature study, constructed profiles can be investigated to see how similar or different they are, compared to the aligned regions resulting from a multiple sequence alignment of that family of proteins.

One further step may be identifying disordered regions and extracting subsequences from these regions. Most of the active sites, catalytic sites, etc. lies along disordered regions ([19, 56]). This would reduce the number of unrelated subsequences hence the noise during the feature map construction.

3.4 Conclusions

We described a discriminative system for functional classification of protein sequences. It uses a subsequence similarity based feature space mapping, SPMap, to convert protein sequences into vector representations. The main idea was to consider the distribution of the subsequences of a given protein over a set of subsequence profiles as its feature representation. SPMap outperformed P2SL tool in subcellular localization and various well known methods in GPCR subfamily classification. In enzyme class classification SPMap produced better or at least comparable results to some of the existing methods.

Our results showed that using subsequence distributions over a quantized space as a feature space for classification of proteins is an effective method in a wide range of different classification problems. Furthermore, the proposed method is computationally efficient and capable of handling large datasets.

It is also important to note that we fixed all parameters for optimized values. This makes SPMap easier to apply for the biologists.

CHAPTER 4

COMBINING CLASSIFIERS ON THE GO

One reason the discriminative methods do not receive as much attention among the biologists compared to the standard sequence alignment methods is the requirement of handling large number of functional classes. There are more than 5000 terms in the molecular function aspect of Gene Ontology (GO). Yet, the use of discriminative classifiers in the literature is confined to selecting the correct function among a small set of functional classes. In order to develop a general annotation system with a classification approach, one should cover enough number of important functional terms. It is not necessary to cover all functional terms since GO describes gene products with fine granularity resulting in thousands of terms. Many terms have none or very few gene products and most of them are not very critical to handle in a large scale annotation of proteins. One should carefully filter and generate relevant classes for the classification system. Secondly, GO allows directed acyclic graphs in its hierarchy, further complicating the selection of training data for each term.

In this chapter, we present a method to prepare training data for the terms defined in GO framework. Then, we investigate the effect of combining different methods for protein function classification. We focus on annotation of proteins with 300 GO molecular function terms. We formulate this problem as a classification problem with 300 classes where proteins can be assigned to more than one class. Although GO defines the relations between the terms, each term is treated as a separate classification problem in a one-versus-all setting. The relations between terms were taken into account while preparing positive and negative training data for each class. We applied 3 different methods to this classification problem. In a one-versus-all setting, usually the size of negative training dataset is much larger than that of the positive training dataset. In order to avoid a bias towards larger negative class, we present a threshold relaxation method that not only shifts the threshold towards a more appropriate classification boundary but also maps the output of the classifier to a probability that states how probable it is that the given sample is a member of the target classes. Finally, we investigated different classifier combination methods and results showed that combination improved the performance for about 93% of the classifiers while yielding similar results to the best performing method for the rest of the classifiers.

4.1 Dataset

One of the well-known and most widely used attempt to standardize protein function terms and to define their relations is GO. GO provides ontology in 3 aspects: *molecular function*, *biological process*, and *cellular location*. In this study, we focus on *molecular function* aspect. GO organizes molecular functions as nodes on a directed acyclic graph (DAG). Each node is a more specific case of its parent node or nodes. A node may have more than one parent. Here, we present a way of establishing positive and negative training data for each class by using evidence codes provided by the GO Annotation (GOA) project and by considering the structure of the GO DAG. While preparing training data, we used Uniprot release 13.0 as the source for protein sequences([6]). Annotations are obtained from October, 2007 version of GOA mapping file and again October 2007 version of GO ontology is used as the bases of our functional terms and their relations in our system.

4.1.1 Positive Training Set

Preparing positive training dataset is relatively easy compared to negatives. First we extracted all proteins that are annotated with the target term or one of its descendants connected with a *is_a* relation by the Gene Ontology Annotation (GOA) project. In order to populate a training dataset without any bias towards computational prediction methods

Table 4.1: List of evidence codes for Gene Ontology annotations.

Abbreviation	Name
IDA	Inferred from Direct Assay
IPI	Inferred from Physical Interaction
IMP	Inferred from Mutant Phenotype
IGI	Inferred from Genetic Interaction
IEP	Inferred from Expression Pattern
ISS	Inferred from Sequence or structural Similarity
IGC	Inferred from Genomic Context
RCA	inferred from Reviewed Computational Analysis
TAS	Traceable Author Statement
NAS	Non-traceable Author Statement
IC	Inferred by Curator
ND	No biological Data available
IEA	Inferred by Electronic Annotation

and to reduce the noise in the training data as much as possible, we filtered out those proteins that are annotated with one of IC, IEA, ISS, IGC, NAS, ND evidence codes. These codes refer to annotations either obtained by electronic means or have ambiguity in their origin ([20]). The rest of the evidence codes IDA, IEP, IGI, IMP, IPI, RCA, and TAS refer to experimental evidences or reviewed computational analysis which we think are more reliable. Table Table 4.1 shows a list of the evidence codes.

4.1.2 Negative Training Set

Theoretically, an annotation for a protein only specifies what function it performs. This is not (generally) an indication of what it doesn't perform. For a protein not having a specific functional label might be merely due to lack of knowledge or experiment. Although this may not be a severe problem in practice, it helps us to understand the difficulties of constructing a negative training dataset for a target term. As a result, each protein that does not have the annotation of the target class or one of its descendants is

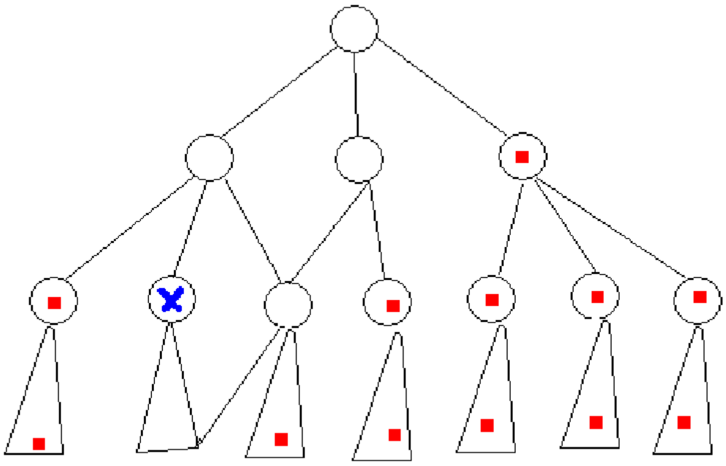


Figure 4.1: Sample graphical representation of possible negative terms for a target term. Blue cross represents the target term and the red squares represent the possible negative terms.

a possible negative training sample. Including all such proteins in the negative training dataset is neither useful nor necessary. First of all, sizes of the positive and negative training sets may become very unbalanced in such a case. For some functional classes, the size of positive training dataset is on the order of tens of proteins, whereas it is about tens of thousands for the negative dataset. Second, computational cost increases with the size of the negative training dataset.

Since we trained our classifiers in one-versus-all setting for 300 GO molecular function terms, our strategy was to select random representative sequences (at most 10) from each term other than the target term. We imposed two constraints on the selected random representative sequences:

1. A sequence shouldn't be annotated with the target term or one of it is descendant terms.

2. If a sequence is annotated with one of the ancestors of the target term, it should also have been annotated with a sibling of the target term.

The first constraint is trivial since we don't want to include protein sequences that are already in the positive training data. Second constraint is imposed in order to avoid including prospective positive training data in to the negative dataset. Ideally, each protein should be annotated with a GO term on a leaf node, in other words, with most specific annotation. If a protein is annotated only up to an internal node, this means either there is lack of evidence for a more specific annotation or an appropriate GO term for that protein has not been added to the ontology yet. Thus, we excluded proteins that are annotated by an ancestor GO term but not with a sibling. Figure 4.1 shows the possible terms to collect negative data for a target term.

4.2 Methods

After preparing positive and negative training data for each of 300 GO molecular function terms, we applied three classification methods representing three approaches:

- BLAST *k*-nearest neighbor (BLAST-kNN) for homology-based approach,
- Subsequence Profile Map (SPMap) for subsequence-based approach,
- Peptide statistics combined with SVMs (PEPSTATS-SVM) for feature-based approach.

4.2.1 BLAST-kNN

In order to classify the target protein, we used *k*-nearest neighbor algorithm ([16]). Similarities between the target protein and proteins in the training data were calculated using NCBI-BLAST tool. We extracted *k*-nearest neighbors having the highest *k* BLAST score. The output of BLAST-*k*NN, O_B for a target protein is calculated as:

$$O_B = \frac{S_p - S_n}{S_p + S_n} \tag{4.1}$$

where S_p is the sum of BLAST scores of proteins in *k*-nearest neighbors that are in the positive training data. Similarly, S_n is the sum of scores of *k*-nearest neighbor proteins that are in the negative training data. Note that the value of O_B is between -1 and +1. The output is 1 if all *k* nearest proteins are the elements of positive training dataset and -1 if all *k* proteins are from negative training dataset. Instead of directly using O_B with a fixed threshold we used the threshold relaxation algorithm given in Section 4.2.4.

4.2.2 SPMap

SPMap maps protein sequences to a fixed-dimensional feature vector where each dimension represents a group of similar fixed-length subsequences. In order to obtain groups of similar subsequences, SPMap first extracts all possible subsequences from the positive training data and clusters similar subsequences. A probabilistic profile or a position specific scoring matrix is then generated for a cluster. The number of clusters determine the dimension of the feature space. Generation of these profiles is called the construction of the feature space map. Once this map is constructed, it is used to represent protein sequences as fixed dimensional vectors. Each dimension of the feature vector is the probability calculated by the best matching subsequence of the protein sequence to the corresponding probabilistic profile. If the sequence to be mapped contains a subsequence similar to a specific group, the value of the corresponding dimension will be high. Note that this representation reflects the information of subsequences that are highly conserved among the positive training data. After the construction of the feature vectors, SVMs are used as to train classifiers. Further information on SPMap is found in [44].

4.2.3 PEPSTATS-SVM

Pepstats tool which is a part of the European Molecular Biology Open Software Suite (EMBOSS) is used to extract peptide statistics of the proteins ([43]). Each protein is represented by a 37 dimensional vector. Peptide features and their dimensions are given in 4.2.3. These features are scaled using the ranges of positive training data and finally fed to an SVM classifier.

4.2.4 Threshold Relaxation

SVM finds a separating decision surface (hyperplane) between the two classes that maximizes the margin, which is the distance of that hyperplane to the nearest samples. For a new sample, output of the SVM is the distance of the hyperplane to the new sample. Sign of the output determines which side of the hyperplane the new sample resides. Hence, the natural threshold for SVM is zero. Optimization algorithm of SVM that finds the hyperplane maximizing the margin is data-driven and may have bias towards the classes

Table 4.2: Features used in PEPSTATS-SVM and their dimensions.

Feature	Dimension
Molecular Weight	1
Number of residues	1
Average residues weight	1
Isoelectric point	1
Charge	1
A280 Molar Extinction Coefficient	1
A280 Extinction Coefficient 1mg/ml	1
Improbability of expression in inclusion bodies	1
Dayhoff Statistics for each aminoacid	20
Percent of tiny residues	1
Percent of small residues	1
Percent of aliphatic residues	1
Percent of aromatic residues	1
Percent of non-polar residues	1
Percent of polar residues	1
Percent of charged residues	1
Percent of basic residues	1
Percent of acidic residues	1
total	37

with more training samples. As a result, using the natural threshold usually results in poor sensitivity if the sizes of the positive and negative training datasets are unbalanced. This is exactly the case in our problem. There are many studies in the literature about threshold relaxation towards smaller class ([58, 3, 48]). In our study, instead of adjusting the threshold value, we present a method that defines probability $P(x)$ of a sample x to be in the positive class.

First, we split the test data into two sets, a *helper set*, to calculate the probability $P(x)$ and a held-out *validation set*, to evaluate the performance of the method. Since, the number of positive test samples is outnumbered by the negative test samples, our method should handle this unbalanced situation. Thus, we calculated a confidence value for the new sample for being positive and negative separately and then we combined these confidences into a single probability. The confidence for a new sample being positive $C_p(x)$ is calculated as the ratio of positive samples in helper set having a classifier output lower than that of the new sample. The confidence for being negative $C_n(x)$ is calculated similarly (Equation 4.2 and Equation 4.3). These ratios are combined to calculate the probability of the new sample to be in positive class (Equation 4.4). A new sample is predicted as positive if $p(x \epsilon Positives) > 0.5$ and as negative, otherwise.

$$C_p(x) = \frac{\sum_{y \epsilon Y_p} I(\phi(x) >= \phi(y))}{|Y_p|} \tag{4.2}$$

$$C_n(x) = \frac{\sum_{y \epsilon Y_n} I(\phi(x) <= \phi(y))}{|Y_n|} \tag{4.3}$$

$$p(x \epsilon Positives) = \frac{C_p}{C_p + C_n} \tag{4.4}$$

Y_p and Y_n are the positive and negative test samples in the helper set, respectively. $\phi(x)$ denotes the output of the classifier for sample x. I operator returns 1 if the condition holds, 0 otherwise. Note that this method implicitly adjusts the threshold. Furthermore, it provides the user a measure to assess how probable it is that the sample is a member of the given class.

4.2.5 Classifier Combination

Observations on many classification problems with different classification methods have shown that although there is usually a best performing method on a specific problem, the samples that are correctly classified or misclassified by different methods may not necessarily overlap ([35]). This observation led to the idea of classifier combination in order to achieve a higher accuracy ([35, 49]). In this study we investigated four classifier combination techniques for three different classification methods each one representing one of the three approaches stated in Section 1:

1. Voting

2. Mean

3. Weighted Mean

4. Addition

Voting, also known as majority voting, simply decides the class of the new sample by counting positive and negative votes from each classifier. Note that votes of the methods have equal weight and the output value of the classifiers are not taken into account.

For the *Mean* combination method, the mean of the probability values calculated by Equation 4.4 is used to decide the class of the new sample. If this mean value is greater than 0.5 sample is labeled as positive.

The combination method *Mean* treats each method equally. But the performance of the methods vary for different functional classes. Thus in the *weighted mean* method, we assigned weights to each method depending on their classification performance on the functional class for which the classifier combination is performed. To assess the performance of the methods we made use of the area under the Receiver Operating Characteristic (ROC) curve, which is called the ROC score. ROC score is a widely used as measure to evaluate the performance of classification methods. ROC score gives an estimation of the discriminative power of the method independent of the threshold value. To calculate the ROC score of each method we used the *helper test set*. Then, we assigned a weight to each method calculated by Equation 4.5.

$$W(m) = \frac{R_m^4}{R_{BLAST-kNN}^4 + R_{SPMap}^4 + R_{Pepstats-svm}^4} \tag{4.5}$$

$W(m)$ denotes weight of method m. R_m is the ROC score for method m. Note that we used the 4^{th} power of ROC scores to assign higher weight to the method with a better ROC score.

In the *Addition* method, output value of the classification methods are summed directly. The probability defined in Section 4.2.4 is then calculated using these added values.

4.3 Results and Discussion

Tests were performed for 300 GO terms in one-versus-all setting. For each GO term, statistics are obtained by averaging results from 5-fold cross-validation. In order to calculate the probability described in Section 4.2.4, we used leave-one-out cross validation in the test set. In other words, we used all available test dataset but one as the *helper set* and one held-out sample as the *validation set*. This is performed for all of the test dataset.

In order to compare the methods and combination strategies, we made use of F_1 statistics. When the sizes of the positive and negative test sets are unbalanced several common statistics such as, sensitivity, specificity, and accuracy may overstate or understate the performance of the classification. F_1 measure is the harmonic mean between precision and sensitivity. It is robust in case of uneven datasets ([30]).

$$Precision = \frac{TP}{TP + FP} \tag{4.6}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.7}$$

$$F_1 = \frac{2 \times Precision \times Sensitivity}{Sensitivity + Precision} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{4.8}$$

TP, FP, TN, and FN denotes true positive, false positive, true negative and false negative, respectively.

Weighted mean method performed best in 279 of 300 classifiers, with an average F_1 score of 0.77. Thus, it is chosen to be the basis combination method for our online tool *GOPred*. *Addition* was the best for 8 classes. *Voting* and *mean* were the best methods for 1 and 3 of the classes, respectively. On the overall, combination improved the performance for 291 of 300 classes. One should note that for the rest of the cases, at least one combination method performed very similar to the best performing single method. Average sensitivity, specificity and F_1 scores over 300 classes is given in 4.3. With respect to F_1 scores, BLAST-*k*NN turned out to be best performing single method for a majority of the functional terms while outperformed by SPMap only at a small fraction of functional terms. SPMap is especially effective for detecting remote homology situations with the use of conserved subsequences. But, as we prepare training and test samples for 300

Table 4.3: F_1 scores, sensitivity and specificity values averaged over 300 GO functional term classifiers.

Method	F_1	Sensitivity	Specificity
SPMap	0.62	89.12	88.92
BLAST-kNN	0.70	92.07	92.53
Pepstats-SVM	0.39	75.47	75.48
Voting	0.71	90.50	92.85
Mean	0.74	91.11	93.74
Weighted Mean	0.77	91.82	94.79
Addition	0.70	92.72	92.49

GO terms, we did not impose any constraints for sequence similarity between test and training examples. As a result, for most test samples, there were very similar sequences in the training data, which is very advantageous for BLAST-*k*nn. Pepstats-SVM was the weakest method in all functional classes. It seems that simple peptide statistics are not sufficient for accurate classification of GO functional terms. Nevertheless, it turned out to be that samples correctly classified by each of the methods are not the same for all methods. This explains the success of the combination methods. As a future work, Pepstats-SVM will be replaced by a more powerful feature-based classification method.

In order to investigate the effect of *threshold relaxation* method presented in Section 4.2.4 we repeated the whole experiment by using natural threshold 0 for all methods. Figure 4.2 shows the comparison of sensitivity and specificity values with and without threshold relaxation averaged over 300 GO terms. Pepstats-SVM turned out to be the most benefiting method which is actually useless without threshold adjustment. BLAST-*k*NN is the less effected method which is not surprising since *k*-nearest neighbors method do not generate a single decision boundary. After threshold relaxation there is a small decrease in specificity but a much larger increase in sensitivity. This conforms with our expectation that there will be a bias towards the class with more training samples. In majority of the 300 GO terms, positive training dataset was highly outnumbered by the negative training dataset. Thus, samples tend to be classified as negative. This explains the

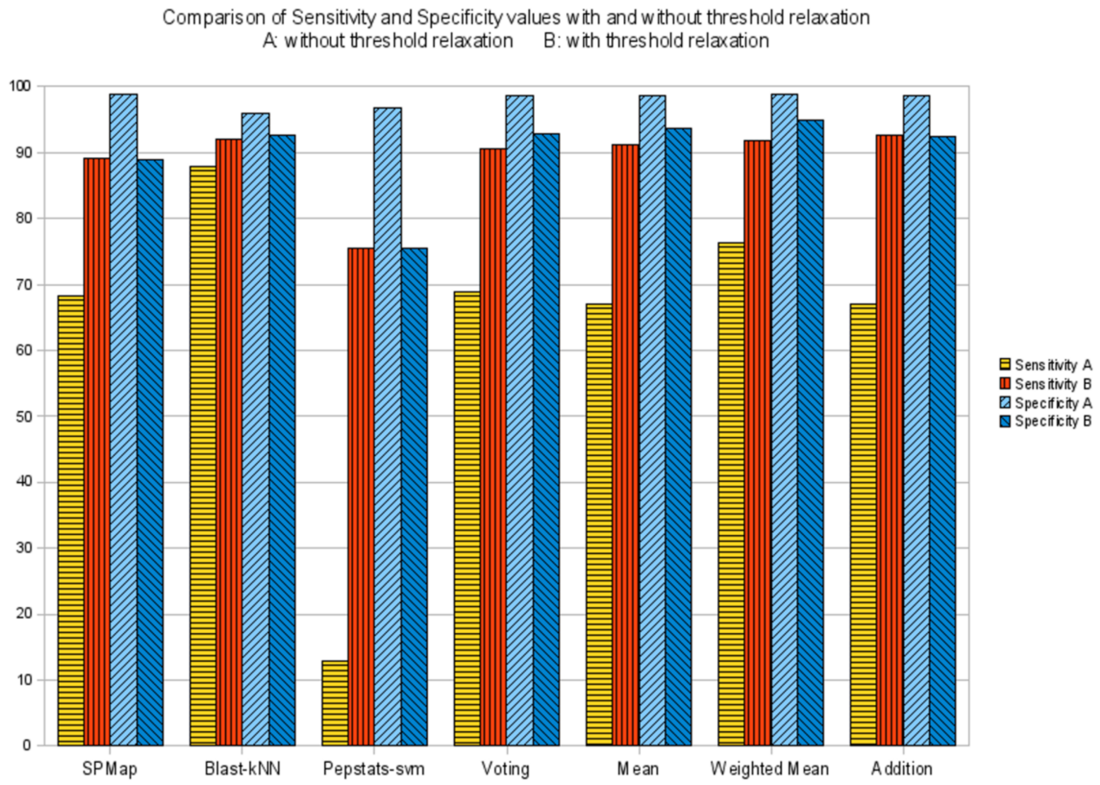


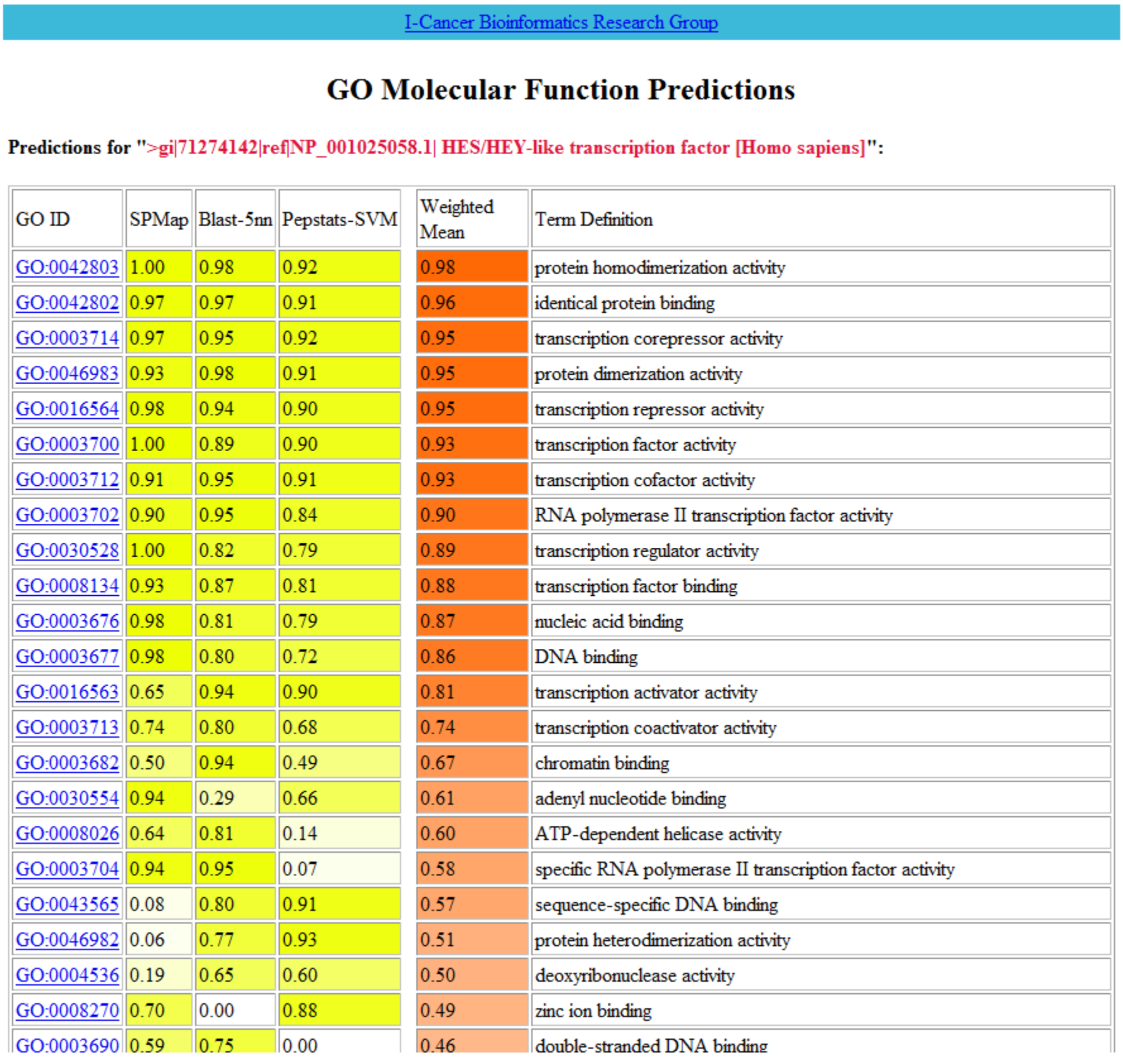
Figure 4.2: Comparison of average sensitivity and specificity values with and without threshold relaxation.

very high specificity and low sensitivity values when threshold relaxation was not used. Automated function prediction tools are generally used to have a rough idea about the protein's possible functions before conducting further in vitro experiments. We believe that failing to detect an important annotation is a more severe problem than assigning a wrong annotation. Thus, increasing sensitivity without a detrimental effect to specificity is a very important achievement. Detailed statistics (Dataset sizes, TP, FP, TN, FN, Sensitivity, Specificity, ROC score, F_1 score) for all of the methods on each GO functional term can be found in Appendix B.

The actual challenge for an automated annotation tool is the annotation of newly identified sequences or genomes. Thus, we applied our method to the prediction of functions of 8 newly reported Homo Sapiens proteins to NCBI in the last year. The combined classifiers were able to predict the reported functions of the proteins in all of the cases. This is a good indication of the effectiveness of the method. 4.3 shows proteins, their reported functions, and annotations of GOPred along with the probabilities calculated by

Table 4.4: GOPred annotations for 8 newly validated human gene entries from NCBI gene database.

Gene Symbol	Reported Function	GOPred annotations:Probability
killin	Nuclear inhibitor of DNA synthesis with high affinity DNA binding [13]	Exonuclease activity: **0.95**
glrx1	glutaredoxin-like, oxidoreductase[21]	oxidoreductase activity: **0.97**
fnip2	AMPK and FLCN interaction[28]	enzyme activator activity: **0.61** enzyme binding: **0.71**
kif18b	microtubule associated motor protein which use ATP[57]	microtubule binding: **0.88** motor activity: **0.83** nucleotide binding: **0.91**
helt	transcription regulator activity[46]	protein homodimerization activity: **0.98** transcription corepressor activity: **0.95**
rgl4	guanin nucleotide dissociation[9]	guanyl-nucleotide exchange factor: **0.79** small GTPase binding: **0.73**
pgap1	GPI inositol-deacylase[51]	lipase activity: **0.89** hydrolase activity acting on ester bonds: **0.89** acyltransferase activity: **0.79**
cobra1	member of negative elongation factor complex during transcription, inhibitor of AP1[40]	ribonucleotide binding: **0.91** enzyme regulator activity: **0.81**



I-Cancer Bioinformatics Research Group

GO Molecular Function Predictions

Predictions for ">gi|71274142|ref|NP_001025058.1| HES/HEY-like transcription factor [Homo sapiens]":

GO ID	SPMap	Blast-5nn	Pepstats-SVM	Weighted Mean	Term Definition
GO:0042803	1.00	0.98	0.92	0.98	protein homodimerization activity
GO:0042802	0.97	0.97	0.91	0.96	identical protein binding
GO:0003714	0.97	0.95	0.92	0.95	transcription corepressor activity
GO:0046983	0.93	0.98	0.91	0.95	protein dimerization activity
GO:0016564	0.98	0.94	0.90	0.95	transcription repressor activity
GO:0003700	1.00	0.89	0.90	0.93	transcription factor activity
GO:0003712	0.91	0.95	0.91	0.93	transcription cofactor activity
GO:0003702	0.90	0.95	0.84	0.90	RNA polymerase II transcription factor activity
GO:0030528	1.00	0.82	0.79	0.89	transcription regulator activity
GO:0008134	0.93	0.87	0.81	0.88	transcription factor binding
GO:0003676	0.98	0.81	0.79	0.87	nucleic acid binding
GO:0003677	0.98	0.80	0.72	0.86	DNA binding
GO:0016563	0.65	0.94	0.90	0.81	transcription activator activity
GO:0003713	0.74	0.80	0.68	0.74	transcription coactivator activity
GO:0003682	0.50	0.94	0.49	0.67	chromatin binding
GO:0030554	0.94	0.29	0.66	0.61	adenyl nucleotide binding
GO:0008026	0.64	0.81	0.14	0.60	ATP-dependent helicase activity
GO:0003704	0.94	0.95	0.07	0.58	specific RNA polymerase II transcription factor activity
GO:0043565	0.08	0.80	0.91	0.57	sequence-specific DNA binding
GO:0046982	0.06	0.77	0.93	0.51	protein heterodimerization activity
GO:0004536	0.19	0.65	0.60	0.50	deoxyribonuclease activity
GO:0008270	0.70	0.00	0.88	0.49	zinc ion binding
GO:0003690	0.59	0.75	0.00	0.46	double-stranded DNA binding

Figure 4.3: GOPred output for *helt* (HES/HEY-like transcription factor) protein [Homo Sapiens].

our method that the protein can be annotated with the corresponding GO term. Figure 4.3 shows the output of our online classification tool for *helt* protein. Furthermore, GOPred is also applied to annotation of 73 newly reported genes from Ovis Aries (Sheep). Results are available on GOPred web site ([25]).

4.4 Conclusions

Automated functional annotation of proteins is an important and difficult problem in computational biology. Most of the function prediction tools, aside from those that uses simple *transfer* approach, defines the annotation problem as a classification problem.

Thus, they require positive and negative training data and the success of the resulting classifier relies on the representative power of this dataset. In this study, we first presented a method to construct accurate positive and negative training data using DAG structure of GO and annotations and evidence codes provided by GOA project.

When using functional classifiers as an annotation system, one has to implement a classifier for each functional class in a one-versus-rest setting because as the number of functions increase it becomes intractable to train one-versus-one classifiers. However, one-versus-rest setting in a classifier renders positive and negative samples highly unbalanced. We present a threshold relaxation method that not only avoids the bias towards the class with more training data but also assigns a probability to the prediction which provides a way of assessing the strength of the annotation.

There is a rich literature on automated function prediction methods each of which have different strengths and weaknesses. We investigated the effects of combining different classifiers for accurate annotation of proteins with functional terms defined in molecular function aspect of GO. Resulting combined classifier clearly outperformed constituent classifiers. Test results also showed that the best combination strategy is the *weighted mean* classifier combination method which assigns different weights to classifiers depending on their discriminative strengths on a specific functional term. Although training more than one classifier requires additional effort and time, it has to be performed only once. Once the classifier is trained, classifying a new sample using the trained classifier imposes only a very small overhead which is totally acceptable if one considers the importance of accurate annotations.

It is also important to note that we do not merely give annotations but also provide a measure for each functional class that states how probable the query protein is a member of that class. This means we also provide less probable functional annotations. This information may help the biologist to build a road map before conducting expensive in vitro experiments.

Finally, proposed classifier combination approach was made publicly available as an online annotation system, named *GOPred*, covering 300 GO terms. Since classifier for each GO term was trained in a one-versus-rest manner, independent of other terms, *GOPred* can be easily extended to cover annotations for more GO terms.

CHAPTER 5

CONCLUSION

Automated functional annotation of proteins is an important and difficult problem in computational biology. Most of the function prediction tools, aside from those that uses simple *transfer approach*, defines the annotation problem as a classification problem. This *classifier approach* is capable of producing more accurate annotations. Methods defined under this approach can be grouped in 3 main categories depending on the features they use:

1. Homology-based methods,
2. Subsequence-based methods,
3. Feature-based methods.

In this thesis, we first described a system for functional classification of protein sequences that combines strengths of both subsequence-based methods and homology-based methods while avoiding the main drawbacks of both. Our method uses a subsequence similarity based feature space mapping, SPMap, to convert protein sequences into vector representations. The main idea was to consider the distribution of the subsequences of a given protein over a set of subsequence profiles as its feature representation. Our results showed that using subsequence distributions over a quantized space as a feature space for classification of proteins is an effective method in wide range of different classification problems. Furthermore, the proposed method is computationally efficient and capable of handling large datasets.

Classifier approach requires positive and negative training data and the success of the resulting classifiers rely on the representative power of this dataset. Thus, we presented a method to construct accurate positive and negative training data using DAG structure of GO and annotations and evidence codes provided by GOA project.

When using functional classifiers as an annotation system, one has to implement a classifier for each functional class in a one-versus-rest setting because as the number of functions increase it becomes intractable to train one-versus-one classifiers. However, one-versus-rest setting in a classifier renders positive and negative samples highly unbalanced. We present a threshold relaxation method that not only avoids the bias towards the class with more training data but also assigns a probability to the prediction which provides a way of assessing the strength of the annotation.

There is a rich literature on automated function prediction methods each of which have different strengths and weaknesses. We investigated the effects of combining different classifiers for accurate annotation of proteins with functional terms defined in molecular function aspect of GO. Resulting combined classifier clearly outperformed constituent classifiers. Test results also showed that the best combination strategy is the *weighted mean* classifier combination method which assigns different weights to classifiers depending on their discriminative strengths on a specific functional term.

It is also important to note that we do not merely give annotations but also provide a measure for each functional class that states how probable the query protein is a member of that class. This means we also provide less probable functional annotations. This information may help the biologist to build a road map before conducting expensive in vitro experiments.

Finally, proposed classifier combination approach was made publicly available as an online annotation system, named *GOPred*, covering 300 GO terms. Since classifier for each GO term was trained in a one-versus-rest manner, independent of other terms, *GOPred* can be easily extended to cover annotations for more GO terms.

The proposed feature extraction method, *SPMap*, may be used on other classification problems that requires the mining of implicit common patterns/motifs. One such area would be identification of transcription factor binding sites on DNA sequences. Of course this would require careful adjustment of clustering module in order to be effective with 4 letter alphabet of DNA.

The results also showed that although SPMap outperformed BLAST in most of the cases when BLAST is used with simple *transfer approach*, the BLAST-*k*NN approach used in Chapter 4 was better than SPMap for classifying most of the GO terms. This may be an indication of the strength of *classifier approach* over simple similarity based *transfer approach*. Such a conclusion may be justified by carefully designing controlled experiments that compares simple BLAST versus BLAST-*k*NN.

REFERENCES

[1] Altschul,S.F., Gish,W., Miller,W., Myers,E.W., Lipman,D.J (1990) A basic local alignment search tool, *Journal of Molecular Biology*, **215**, 403-410.

[2] Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acid Research*, **25**, 3389-3402.

[3] Arampatzis,A. (2002) Unbiased S-D Threshold Optimization, Initial Query Degradation, Decay, and Incrementality, for Adaptive Document Filtering, *Tenth Text Retrieval Conference* (TREC-2001), 596-605.

[4] Ashburner,M., Ball,C., Blake,J., Botstein,D., Butler,H., Cherry,J., Davis,A., Dolinski,K., Dwight,S., Eppig,J. et. al, Gene Ontology: tool for the unification of biology, 2000, Nature Genetics 25, 25-29.

[5] Atalay,V., & Cetin-Atalay,R., Implicit Motif Distribution based Hybrid Computational Kernel for Sequence Classification, 2005, Bioinformatics 21(8), 1429-1436.

[6] Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M., Martin,M.J., Natale,D.A., O'Donovan,C., Redaschi,N., Yeh,L.S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Research*. **33** D154-D159.

[7] Ben-hur,A., Brutlab,D. (2003) Remote homology detection: a motif based approach, *Bioinformatics*, **19**, 26-33.

[8] Blekas,K., Fotiadis,D.I., and Likas,A. (2005) Motif-based protein sequence classification using neural networks. *J Comput Biol.*, **12(1)**, 64-82.

[9] Bodemann,B.O., White,M.A. (2008) Ral GTPases and cancer: linchpin support of the tumorigenic platform, *Nat. Rev. Cancer*, **8(2)**, 133-140.

[10] Burges, C.J.C. (1998) A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, **2**, 121-167.

[11] Cai,C., Han,L., Ji,Z., Chen,X., and Chen,Y. (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence, *Nucleic Acids Research*, **31(13)**, 3692-3697.

[12] Cheng,B.Y.M., Carbonell,J.G., and Klein-Seetharaman,J. (2005) Protein classification based on text document classification techniques. *Proteins*, **58(4)**, 955-970.

[13] Cho,Y.J., Liang,P. (2008) Killin is a p53-regulated nuclear inhibitor of DNA synthesis, *Proc Natl. Acad. Sci. USA*, **105(14)**, 5396-5401.

[14] Cristianini,N., & Shawe-Taylor,J., An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, 2000, Cambridge University Press.

[15] Cortes, C. and Vapnik, V. (1995) Support vector networks, *Machine Learning*, **20**, 273-297.

[16] Cover,T.M. and Hart,P.E. (1967) Nearest neighbor pattern classification. *IEEE Trans. IT*, *13(1)*, 21-27.

[17] Devos,D. and Valencia, A. (2000) Practical limits of function prediction, *PROTEINS: Structure, Function, and Genetics*, **41**, 98-107.

[18] Duda,R.O., Hart,P.E., and Stork,D.G. (2000), *Pattern Classification (2nd Edition)*, Wiley-Interscience.

[19] Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M., & Obradovic,Z., Intrinsic disorder and protein function, 2002, Biochemistry 41, 6573-6582.

[20] Eisner,R., Poulin,B., Szafron,D., Lu,P., and Greiner,R., (2005) Improving Protein Function Prediction using the Hierarchical Structure of the Gene Ontology, *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*.

[21] Fernandes,A.P., Holmgren,A. (2004) Glutaredoxins: glutathione-dependent redox enzymes with functions far beyond a simple thioredoxin backup system, *Antioxid Redox Signal*, **6(1)**, 63-74.

[22] Fratkin,E., Naughton,B.T., Brutlag,D.L., & Batzoglou,S., MotifCut: regulatory motifs finding with maximum density subgraphs, 2006, Bioinformatics 22(14), e150-e157.

[23] Friedberg,I. (2006) Automated protein function prediction - the genomic challenge *Briefings in Bioinformatics*, **7**, 225-242.

[24] Gilks,W.R., Audit,B., de Angelis,D., Tsoka,S., Ouzounis,C.A. (2005) Percolation of annotation errors through hierarchically structured protein sequence databases, *Math Biosci.*, **193**, 223-234.

[25] GOPred web site: http://kinaz.fen.bilkent.edu.tr/gopred/ovisaries.html, accessed-date:02/08/2008.

[26] GPCR subfamily classification benchmark dataset: http://www.soe.ucsc.edu/research/compro/gpcr/subfamily_seqs, accessed-date: 04/05/2007.

[27] Hannenhalli,S.S. and Russell,R.B. (2000) Analysis and prediction of functional sub-types from protein sequence alignments, *J Mol Biol.*, **303(1)**, 61-76.

[28] Hasumi,H., Baba,M., Hong,S.B., Hasumi,Y., Huang,Y., Yao,M., Valera,V.A., Linehan,W.M., Schmidt,L.S. (2008) Identification and characterization of a novel folliculin-interacting protein FNIP2, *Gene*, **415(1-2)**, 60-67.

[29] Henikoff,S., & Henikoff,J.G., Amino acid substitution matrices from protein blocks, 1992, Proc. Natl. Acad. Sci. USA 89, 10915-10919.

[30] Holloway,D.T., Kon,M.A., and DeLisi,C. (2006) Machine Learning Methods for Transcription data Integration, *IBM Journal of Research and Development*, **50(6)**, 631-643.

[31] Jensen,L.J., Gupta,R., Blom,N., Devos,D., Tamames,J., Kesmir,C., Nielsen, H., Staerfeldt,H.H., Rapacki,K., Workman,C., Andersen,C.A.F., Knudsen,S., Krogh,A.,

Valencia,A., and Brunak, S. (2002) Prediction of human protein function from post-translational modifications and localization features, *J Mol Biol.*, **319(5)**, 1257-1265.

[32] Joachims,T., Making large-Scale SVM Learning Practical (Book Chapter), 1999, Advances in Kernel Methods - Support Vector Learning, MIT Press.

[33] Karchin,R., Karplus,K., and Haussler,D. (2002) Classifying G-protein coupled receptors with support vector machines, *Bioinformatics*, **18(1)**, 147-159.

[34] King,R.D., Karwath,A., Clare,A., and Dehaspe,L. (2000) Accurate prediction of protein functional class from sequence in the Mycobacterium tuberculosis and Escherichia coli genomes using data mining, *Yeast*, **17(4)**, 283-293.

[35] Kittler,J.,Hatef,M.,Duin,R.P.W, Matas,J. (1998) On Combining Classifiers, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **20(3)**.

[36] Kunik,V., Solan,Z., Edelman,S., Ruppin,E., and Horn,D. (2005) Motif extraction and protein classification, *In Proc. Computational Systems Bioinformatics (CSB)*, 80-85.

[37] Leslie,C.S., Eskin,E., Cohen,A., Weston,J., Noble,W.S. (2004) Mismatch string kernels for discriminative protein classification, *Bioinformatics*, **20**, 467-476.

[38] Liu,A.H. and Califano,A. (2001) Functional classification of proteins by pattern discovery and top-down clustering of primary sequences, *IBM Systems Journal*, **40(2)**, 379-393.

[39] Lu, L., Qian,Z., Cai, Y., Yixue Li ECS: An automatic enzyme classifier based on functional domain composition, 2007, Computational Biology and Chemistry 31, 226-232.

[40] McChesney,P.A., Aiyar,S.E., Lee,O.J., Zaika,A., Moskaluk,C., Li,R., El-Rifai,W. (2006) Cofactor of BRCA1: a novel transcription factor regulator in upper gastrointestinal adenocarcinomas *Cancer Research*, **66(3)**, 1346-53.

[41] Moriyama,E.N. & Kim,J., Protein family classification with discriminant function analysis, 2006, In Genome Exploitation: Data Mining the Genome, J. P. Gustafson, Ed. Springer.

[42] Pasquier,C., Promponas,V.J., and Hamodrakas,S.J. (2001) PRED-CLASS: cascading neural networks for generalized protein classification and genome-wide applications, *Proteins*, **44(3)**, 361-369.

[43] Rice,P., Longden,I., and Bleasby,A. (2000) The European Molecular Biology Open Software Suite, *Trends in Genetics*, **16(6)**, 276-277.

[44] Sarac,O.S., Yuzugullu,O.G., Cetin-Atalay,R., and Atalay,V. (2008) Subsequence-based feature map for protein function classification, *Computational Biology and Chemistry*, **32**, 122-130.

[45] Sasson,O., Kaplan,N., Linial,M. (2006) Functional annotation prediction: All for one and one for all, *Protein Science*, **15**, 1-16.

[46] Schwanbeck,R., Schroeder,T., Henning,K., Kohlhof,H., Rieber,N., Erfurth,M.L., Just,U. (2008) Notch Signaling in Embryonic and Adult Myelopoiesis, *Cell Tissues Organs*, Epub ahead of print.

[47] Schomburg,I., Chang,A., & Schomburg,D., BRENDA, enzyme data and metabolic information, 2002, Nucleic Acids Research 30(1), 47-49.

[48] Shanahan,J.G., Roma,N., (2003) Boosting support vector machines for text classification through parameter-free threshold relaxation, *Proc. of* 12^th *int. conf. on Information and knowledge management*, LA, USA, 247-254.

[49] Sohn,S.Y., Shin,H.W. (2007) Experimental study for the comparison of classifier combination methods, *Pattern Recognition*, **40(1)**, 33-40.

[50] SPMap enzyme classifier web site: http://gen.ceng.metu.edu.tr/spmap/cgi-bin/enzyme.cgi, accessed-date: 02/07/2008.

[51] Tanaka,S., Maeda,Y., Tashima,Y., Kinoshita,T. (2004) Inositol deacylation of glycosylphosphatidylinositol-anchored proteins is mediated by mammalian PGAP1 and yeast Bst1p, *J. Biol. Chem*, **279(14)**, 14256-63.

[52] Tomii,K., & Kanehisa,M., Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, 1996, Protein Eng. 9, 27-36.

[53] Wang,J.T.L., Ma,Q., Shasha,D., & Wu,C.H., New techniques for extracting features from protein sequences, 2001, IBM Systems Journal 40(2), 426–441.

[54] Wang,X., Schroeder,D., Dobbs,D., and Honavar,V.G. (2003) Automated data-driven discovery of motif-based protein function classifiers, *Inf. Sci*, **155(1-2)**, 1-18.

[55] Wikipedia protein entry: http://en.wikipedia.org/wiki/Proteins, accessed-date: 04/06/2008.

[56] Wright,P.E. & Dyson,H.J., Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm, 1999, J. Mol. Biol. 293, 321-331.

[57] Yildiz,A., Selvin,P.R. (2005) Kinesin: walking, crawling and sliding along?, *Trends Cell Biol.*, **15(2)**, 112-120.

[58] Zhai,C., Jansen,P., Stoica,E., Grot,N., Evans,D.A. (1999) Threshold Calibration in CLARIT Adaptive Filtering, *Seventh Text Retrieval Conference* (TREC-7), 149-156.

APPENDIX A

BLOSUM62 MATRIX

Table A.1: Blossum 62 Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

APPENDIX B

DETAILED STATISTICS 300 GO TERMS

Table B.1: Detailed statistics of 7 classification methods for each of 300 GO terms. True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN), Sensitivity (SENS), Specificity (SPEC), Median Ratio False Positives (MedRFP), ROC score (ROC), F1 statistics (F1), Positive Predictive Value (PPV).

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
GO:0003676	spmap	5354	370	2276	870	86.02	86.02	0.03	0.93	0.8962	0.94
	blast	5653	240	2406	571	90.83	90.93	0.05	0.97	0.9331	0.96
	peps	4489	737	1896	1709	72.43	72.01	0.15	0.79	0.7859	0.86
	voting	5509	324	2322	715	88.51	87.76	0.02	0.97	0.9138	0.94
	mean	5571	302	2344	653	89.51	88.59	0.01	0.96	0.9211	0.95
	wmean	5646	250	2396	578	90.71	90.55	0.01	0.97	0.9317	0.96
	add	5572	275	2371	652	89.52	89.61	0.01	0.95	0.9232	0.95
GO:0016787	spmap	4079	387	1856	855	82.67	82.75	0.02	0.91	0.8679	0.91
	blast	4482	205	2038	452	90.84	90.86	0.07	0.97	0.9317	0.96
	peps	3103	832	1396	1824	62.98	62.66	0.29	0.67	0.7003	0.79
	voting	4273	282	1961	661	86.60	87.43	0.02	0.97	0.9006	0.94
	mean	4326	247	1996	608	87.68	88.99	0.01	0.94	0.9101	0.95
	wmean	4458	224	2019	476	90.35	90.01	0.01	0.96	0.9272	0.95
	add	4425	230	2013	509	89.68	89.75	0.01	0.96	0.9229	0.95
GO:0016740	spmap	3955	398	2089	756	83.95	84.00	0.01	0.92	0.8727	0.91
	blast	4348	192	2295	363	92.29	92.28	0.06	0.98	0.9400	0.96
	peps	3002	902	1567	1697	63.89	63.47	0.23	0.69	0.6979	0.77
	voting	4132	271	2216	579	87.71	89.10	0.01	0.97	0.9067	0.94
	mean	4154	236	2251	557	88.18	90.51	0.00	0.96	0.9129	0.95
	wmean	4305	193	2294	406	91.38	92.24	0.00	0.97	0.9350	0.96
	add	4297	218	2269	414	91.21	91.23	0.00	0.97	0.9315	0.95
GO:0003677	spmap	3715	359	2394	556	86.98	86.96	0.01	0.94	0.8904	0.91
	blast	3819	289	2464	452	89.42	89.50	0.05	0.96	0.9116	0.93
	peps	3063	778	1961	1190	72.02	71.60	0.14	0.79	0.7569	0.80
	voting	3789	353	2400	482	88.71	87.18	0.03	0.97	0.9007	0.91
	mean	3812	327	2426	459	89.25	88.12	0.01	0.95	0.9065	0.92
	wmean	3842	279	2474	429	89.96	89.87	0.01	0.96	0.9156	0.93
	add	3782	315	2438	489	88.55	88.56	0.01	0.95	0.9039	0.92
GO:0004871	spmap	3624	314	2169	526	87.33	87.35	0.01	0.94	0.8961	0.92
	blast	3752	237	2246	398	90.41	90.46	0.06	0.97	0.9220	0.94
	peps	3028	671	1801	1113	73.12	72.86	0.11	0.79	0.7724	0.82
	voting	3667	250	2233	483	88.36	89.93	0.01	0.97	0.9091	0.94
	mean	3709	226	2257	441	89.37	90.90	0.00	0.96	0.9175	0.94
	wmean	3743	196	2287	407	90.19	92.11	0.00	0.96	0.9255	0.95

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	3717	258	2225	433	89.57	89.61	0.01	0.95	0.9150	0.94
GO:0060089	spmap	3628	312	2168	522	87.42	87.42	0.01	0.94	0.8969	0.92
	blast	3746	240	2240	404	90.27	90.32	0.05	0.97	0.9208	0.94
	peps	3015	678	1794	1127	72.79	72.57	0.11	0.78	0.7696	0.82
	voting	3667	234	2246	483	88.36	90.56	0.01	0.98	0.9109	0.94
	mean	3712	215	2265	438	89.45	91.33	0.01	0.96	0.9192	0.95
	wmean	3747	177	2303	403	90.29	92.86	0.01	0.97	0.9282	0.95
	add	3731	248	2231	419	89.90	90.00	0.01	0.95	0.9179	0.94
GO:0030528	spmap	3361	324	2358	462	87.92	87.92	0.02	0.94	0.8953	0.91
	blast	3437	271	2411	386	89.90	89.90	0.04	0.97	0.9128	0.93
	peps	2857	678	1988	961	74.83	74.57	0.09	0.82	0.7771	0.81
	voting	3398	294	2388	425	88.88	89.04	0.02	0.98	0.9043	0.92
	mean	3426	264	2418	397	89.62	90.16	0.01	0.96	0.9120	0.93
	wmean	3463	244	2438	360	90.58	90.90	0.01	0.97	0.9198	0.93
	add	3414	286	2396	409	89.30	89.34	0.01	0.95	0.9076	0.92
GO:0004872	spmap	3172	239	2330	328	90.63	90.70	0.00	0.96	0.9180	0.93
	blast	3276	164	2405	224	93.60	93.62	0.05	0.98	0.9441	0.95
	peps	2685	599	1960	805	76.93	76.59	0.10	0.83	0.7927	0.82
	voting	3210	199	2370	290	91.71	92.25	0.01	0.99	0.9292	0.94
	mean	3233	183	2386	267	92.37	92.88	0.00	0.98	0.9349	0.95
	wmean	3263	153	2416	237	93.23	94.04	0.00	0.98	0.9436	0.96
	add	3210	213	2356	290	91.71	91.71	0.01	0.97	0.9273	0.94
GO:0005215	spmap	2987	260	2022	385	88.58	88.61	0.00	0.95	0.9026	0.92
	blast	3096	186	2096	276	91.81	91.85	0.05	0.97	0.9306	0.94
	peps	2381	671	1602	985	70.74	70.48	0.13	0.77	0.7420	0.78
	voting	3002	221	2061	370	89.03	90.32	0.01	0.98	0.9104	0.93
	mean	3042	185	2097	330	90.21	91.89	0.00	0.96	0.9220	0.94
	wmean	3096	146	2136	276	91.81	93.60	0.00	0.97	0.9362	0.95
	add	3040	224	2058	332	90.15	90.18	0.01	0.96	0.9162	0.93
GO:0016772	spmap	2574	325	2324	361	87.70	87.73	0.00	0.95	0.8824	0.89
	blast	2755	160	2489	180	93.87	93.96	0.03	0.98	0.9419	0.95
	peps	1920	916	1713	1009	65.55	65.16	0.22	0.70	0.6661	0.68
	voting	2636	190	2459	299	89.81	92.83	0.01	0.98	0.9151	0.93
	mean	2648	163	2486	287	90.22	93.85	0.00	0.97	0.9217	0.94
	wmean	2702	137	2512	233	92.06	94.83	0.00	0.98	0.9359	0.95
	add	2673	235	2414	262	91.07	91.13	0.00	0.97	0.9149	0.92
GO:0022892	spmap	2445	255	2111	293	89.30	89.22	0.00	0.96	0.8992	0.91
	blast	2517	190	2176	221	91.93	91.97	0.03	0.98	0.9245	0.93
	peps	1970	663	1692	763	72.08	71.85	0.13	0.78	0.7343	0.75
	voting	2461	196	2170	277	89.88	91.72	0.01	0.98	0.9123	0.93
	mean	2487	177	2189	251	90.83	92.52	0.00	0.97	0.9208	0.93
	wmean	2533	138	2228	205	92.51	94.17	0.00	0.98	0.9366	0.95
	add	2487	218	2148	251	90.83	90.79	0.00	0.97	0.9138	0.92
GO:0004888	spmap	2599	194	2408	210	92.52	92.54	0.00	0.98	0.9279	0.93
	blast	2658	140	2462	151	94.62	94.62	0.03	0.99	0.9481	0.95
	peps	2257	511	2077	544	80.58	80.26	0.07	0.87	0.8106	0.82
	voting	2628	144	2458	181	93.56	94.47	0.01	0.99	0.9418	0.95
	mean	2639	140	2462	170	93.95	94.62	0.00	0.98	0.9445	0.95
	wmean	2653	123	2479	156	94.45	95.27	0.00	0.99	0.9500	0.96
	add	2614	179	2423	195	93.06	93.12	0.00	0.98	0.9332	0.94
GO:0016301	spmap	2314	279	2410	270	89.55	89.62	0.00	0.96	0.8940	0.89
	blast	2451	136	2553	133	94.85	94.94	0.03	0.98	0.9480	0.95
	peps	1708	911	1760	872	66.20	65.89	0.23	0.71	0.6570	0.65
	voting	2355	165	2524	229	91.14	93.86	0.01	0.98	0.9228	0.93
	mean	2370	138	2551	214	91.72	94.87	0.00	0.97	0.9309	0.94
	wmean	2413	115	2574	171	93.38	95.72	0.00	0.98	0.9441	0.95
	add	2384	207	2482	200	92.26	92.30	0.00	0.97	0.9214	0.92
GO:0022857	spmap	2324	218	2072	243	90.53	90.48	0.00	0.97	0.9098	0.91
	blast	2396	149	2141	171	93.34	93.49	0.04	0.98	0.9374	0.94
	peps	1905	590	1695	655	74.41	74.18	0.11	0.81	0.7537	0.76
	voting	2343	186	2104	224	91.27	91.88	0.01	0.98	0.9195	0.93

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	2363	164	2126	204	92.05	92.84	0.00	0.97	0.9278	0.94
	wmean	2388	126	2164	179	93.03	94.50	0.00	0.98	0.9400	0.95
	add	2348	195	2095	219	91.47	91.48	0.00	0.97	0.9190	0.92
GO:0016773	spmap	2154	255	2439	227	90.47	90.53	0.00	0.97	0.8994	0.89
	blast	2244	152	2542	137	94.25	94.36	0.03	0.98	0.9395	0.94
	peps	1541	950	1730	838	64.78	64.55	0.23	0.70	0.6329	0.62
	voting	2180	154	2540	201	91.56	94.28	0.01	0.98	0.9247	0.93
	mean	2197	132	2562	184	92.27	95.10	0.00	0.97	0.9329	0.94
	wmean	2235	113	2581	146	93.87	95.81	0.00	0.98	0.9452	0.95
	add	2216	187	2507	165	93.07	93.06	0.00	0.97	0.9264	0.92
GO:0022891	spmap	2084	222	2175	214	90.69	90.74	0.00	0.97	0.9053	0.90
	blast	2137	167	2230	161	92.99	93.03	0.03	0.98	0.9287	0.93
	peps	1670	637	1748	622	72.86	73.29	0.10	0.80	0.7262	0.72
	voting	2099	171	2226	199	91.34	92.87	0.01	0.99	0.9190	0.92
	mean	2123	142	2255	175	92.38	94.08	0.00	0.98	0.9305	0.94
	wmean	2144	118	2279	154	93.30	95.08	0.00	0.98	0.9404	0.95
	add	2120	184	2213	178	92.25	92.32	0.00	0.97	0.9213	0.92
GO:0004672	spmap	1895	233	2506	177	91.46	91.49	0.00	0.97	0.9024	0.89
	blast	1959	149	2590	113	94.55	94.56	0.03	0.99	0.9373	0.93
	peps	1369	916	1813	700	66.17	66.43	0.23	0.72	0.6288	0.60
	voting	1907	164	2575	165	92.04	94.01	0.00	0.99	0.9206	0.92
	mean	1913	128	2611	159	92.33	95.33	0.00	0.98	0.9302	0.94
	wmean	1958	101	2638	114	94.50	96.31	0.00	0.98	0.9480	0.95
	add	1938	176	2563	134	93.53	93.57	0.00	0.98	0.9259	0.92
GO:0016491	spmap	1752	340	2349	254	87.34	87.36	0.00	0.95	0.8551	0.84
	blast	1876	174	2515	130	93.52	93.53	0.02	0.98	0.9250	0.92
	peps	1359	847	1829	641	67.95	68.35	0.18	0.73	0.6462	0.62
	voting	1810	264	2425	196	90.23	90.18	0.01	0.98	0.8873	0.87
	mean	1835	217	2472	171	91.48	91.93	0.00	0.97	0.9044	0.89
	wmean	1871	163	2526	135	93.27	93.94	0.00	0.98	0.9262	0.92
	add	1841	221	2468	165	91.77	91.78	0.00	0.97	0.9051	0.89
GO:0003700	spmap	1777	282	2494	200	89.88	89.84	0.00	0.96	0.8806	0.86
	blast	1823	215	2561	154	92.21	92.26	0.03	0.98	0.9081	0.89
	peps	1460	709	2057	512	74.04	74.37	0.10	0.81	0.7051	0.67
	voting	1801	261	2515	176	91.10	90.60	0.02	0.98	0.8918	0.87
	mean	1809	237	2539	168	91.50	91.46	0.01	0.97	0.8993	0.88
	wmean	1814	199	2577	163	91.76	92.83	0.00	0.97	0.9093	0.90
	add	1794	257	2519	183	90.74	90.74	0.01	0.96	0.8908	0.87
GO:0015075	spmap	1680	219	2226	165	91.06	91.04	0.00	0.97	0.8974	0.88
	blast	1718	168	2277	127	93.12	93.13	0.02	0.98	0.9209	0.91
	peps	1312	686	1749	526	71.38	71.83	0.14	0.78	0.6840	0.66
	voting	1691	177	2268	154	91.65	92.76	0.01	0.99	0.9109	0.91
	mean	1712	160	2285	133	92.79	93.46	0.00	0.98	0.9212	0.91
	wmean	1726	127	2318	119	93.55	94.81	0.00	0.98	0.9335	0.93
	add	1696	199	2246	149	91.92	91.86	0.00	0.97	0.9070	0.89
GO:0016788	spmap	1379	399	2188	251	84.60	84.58	0.00	0.92	0.8093	0.78
	blast	1510	191	2396	120	92.64	92.62	0.02	0.98	0.9066	0.89
	peps	1007	973	1601	621	61.86	62.20	0.28	0.67	0.5582	0.51
	voting	1437	290	2297	193	88.16	88.79	0.01	0.97	0.8561	0.83
	mean	1454	238	2349	176	89.20	90.80	0.00	0.96	0.8754	0.86
	wmean	1488	154	2433	142	91.29	94.05	0.00	0.97	0.9095	0.91
	add	1507	193	2394	123	92.45	92.54	0.00	0.97	0.9051	0.89
GO:0030234	spmap	1222	477	2184	265	82.18	82.07	0.01	0.90	0.7671	0.72
	blast	1327	290	2371	161	89.18	89.10	0.03	0.95	0.8548	0.82
	peps	927	985	1666	557	62.47	62.84	0.29	0.68	0.5459	0.48
	voting	1289	417	2244	199	86.63	84.33	0.08	0.96	0.8071	0.76
	mean	1303	364	2297	185	87.57	86.32	0.01	0.94	0.8260	0.78
	wmean	1319	273	2388	169	88.64	89.74	0.01	0.95	0.8565	0.83
	add	1325	292	2369	163	89.05	89.03	0.01	0.96	0.8535	0.82
	spmap	1284	401	2302	222	85.26	85.16	0.00	0.93	0.8048	0.76
	blast	1360	262	2441	146	90.31	90.31	0.02	0.96	0.8696	0.84

GO:0043167

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	954	972	1720	547	63.56	63.89	0.21	0.69	0.5568	0.50
	voting	1313	342	2361	193	87.18	87.35	0.02	0.97	0.8307	0.79
	mean	1333	295	2408	173	88.51	89.09	0.01	0.95	0.8507	0.82
	wmean	1348	233	2470	158	89.51	91.38	0.00	0.96	0.8733	0.85
	add	1358	268	2435	148	90.17	90.09	0.01	0.96	0.8672	0.84
GO:0005102	spmap	1281	458	2286	256	83.34	83.31	0.01	0.91	0.7821	0.74
	blast	1329	370	2374	208	86.47	86.52	0.01	0.94	0.8214	0.78
	peps	1008	925	1810	522	65.88	66.18	0.19	0.72	0.5822	0.52
	voting	1299	406	2338	238	84.52	85.20	0.03	0.95	0.8014	0.76
	mean	1311	372	2372	226	85.30	86.44	0.01	0.93	0.8143	0.78
	wmean	1330	315	2429	207	86.53	88.52	0.00	0.94	0.8360	0.81
	add	1321	386	2358	215	86.00	85.93	0.01	0.93	0.8147	0.77
GO:0000166	spmap	1303	371	2306	210	86.12	86.14	0.00	0.94	0.8177	0.78
	blast	1379	235	2442	134	91.14	91.22	0.03	0.97	0.8820	0.85
	peps	998	892	1772	512	66.09	66.52	0.20	0.71	0.5871	0.53
	voting	1333	250	2427	180	88.10	90.66	0.01	0.97	0.8611	0.84
	mean	1342	220	2457	171	88.70	91.78	0.00	0.96	0.8728	0.86
	wmean	1362	172	2505	151	90.02	93.57	0.00	0.97	0.8940	0.89
	add	1370	253	2424	143	90.55	90.55	0.00	0.96	0.8737	0.84
GO:0008324	spmap	1363	216	2377	123	91.72	91.67	0.00	0.97	0.8894	0.86
	blast	1396	162	2431	90	93.94	93.75	0.02	0.98	0.9172	0.90
	peps	1034	769	1812	448	69.77	70.21	0.16	0.76	0.6295	0.57
	voting	1366	176	2417	120	91.92	93.21	0.01	0.99	0.9022	0.89
	mean	1375	156	2437	111	92.53	93.98	0.00	0.98	0.9115	0.90
	wmean	1395	115	2478	91	93.88	95.56	0.00	0.99	0.9312	0.92
	add	1380	185	2408	106	92.87	92.87	0.00	0.98	0.9046	0.88
GO:0004930	spmap	1430	134	2559	75	95.02	95.02	0.00	0.99	0.9319	0.91
	blast	1462	76	2617	43	97.14	97.18	0.01	1.00	0.9609	0.95
	peps	1334	289	2389	168	88.81	89.21	0.01	0.94	0.8538	0.82
	voting	1435	83	2610	70	95.35	96.92	0.00	1.00	0.9494	0.95
	mean	1439	68	2625	66	95.61	97.47	0.00	0.99	0.9555	0.95
	wmean	1443	64	2629	62	95.88	97.62	0.00	0.99	0.9582	0.96
	add	1438	122	2571	67	95.55	95.47	0.00	0.99	0.9383	0.92
GO:0043169	spmap	1196	362	2371	183	86.73	86.75	0.00	0.93	0.8144	0.77
	blast	1255	247	2486	124	91.01	90.96	0.02	0.96	0.8712	0.84
	peps	887	959	1761	490	64.42	64.74	0.21	0.70	0.5504	0.48
	voting	1222	322	2411	157	88.61	88.22	0.02	0.97	0.8361	0.79
	mean	1232	279	2454	147	89.34	89.79	0.00	0.96	0.8526	0.82
	wmean	1251	221	2512	128	90.72	91.91	0.00	0.97	0.8776	0.85
	add	1254	249	2484	125	90.94	90.89	0.00	0.96	0.8702	0.83
GO:0016817	spmap	1206	292	2345	151	88.87	88.93	0.00	0.95	0.8448	0.81
	blast	1249	211	2426	108	92.04	92.00	0.02	0.98	0.8868	0.86
	peps	907	858	1764	449	66.89	67.28	0.17	0.73	0.5812	0.51
	voting	1223	228	2409	134	90.13	91.35	0.01	0.98	0.8711	0.84
	mean	1230	205	2432	127	90.64	92.23	0.01	0.96	0.8811	0.86
	wmean	1245	165	2472	112	91.75	93.74	0.01	0.97	0.8999	0.88
	add	1259	191	2446	98	92.78	92.76	0.00	0.97	0.8970	0.87
GO:0016818	spmap	1187	318	2310	164	87.86	87.90	0.00	0.95	0.8312	0.79
	blast	1246	202	2426	105	92.23	92.31	0.01	0.98	0.8903	0.86
	peps	914	840	1776	437	67.65	67.89	0.18	0.73	0.5887	0.52
	voting	1211	235	2393	140	89.64	91.06	0.02	0.97	0.8659	0.84
	mean	1228	211	2417	123	90.90	91.97	0.00	0.96	0.8803	0.85
	wmean	1239	158	2470	112	91.71	93.99	0.00	0.97	0.9017	0.89
	add	1253	191	2437	98	92.75	92.73	0.00	0.98	0.8966	0.87
GO:0016462	spmap	1196	283	2383	141	89.45	89.38	0.00	0.95	0.8494	0.81
	blast	1236	200	2466	101	92.45	92.50	0.02	0.98	0.8915	0.86
	peps	904	846	1805	432	67.66	68.09	0.18	0.74	0.5859	0.52
	voting	1202	232	2434	135	89.90	91.30	0.01	0.98	0.8676	0.84
	mean	1214	195	2471	123	90.80	92.69	0.00	0.96	0.8842	0.86
	wmean	1228	149	2517	109	91.85	94.41	0.00	0.97	0.9049	0.89

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	1247	182	2484	90	93.27	93.17	0.00	0.98	0.9017	0.87
GO:0008233	spmap	1204	312	2328	161	88.21	88.18	0.00	0.95	0.8358	0.79
	blast	1288	148	2492	77	94.36	94.39	0.02	0.99	0.9197	0.90
	peps	923	838	1787	442	67.62	68.08	0.15	0.74	0.5905	0.52
	voting	1217	193	2447	148	89.16	92.69	0.01	0.98	0.8771	0.86
	mean	1239	149	2491	126	90.77	94.36	0.00	0.97	0.9001	0.89
	wmean	1274	104	2536	91	93.33	96.06	0.00	0.98	0.9289	0.92
	add	1262	199	2441	103	92.45	92.46	0.00	0.97	0.8931	0.86
GO:0017076	spmap	1133	385	2301	188	85.77	85.67	0.00	0.93	0.7982	0.75
	blast	1206	233	2453	115	91.29	91.33	0.02	0.97	0.8739	0.84
	peps	873	893	1777	445	66.24	66.55	0.21	0.71	0.5661	0.49
	voting	1158	275	2411	163	87.66	89.76	0.01	0.97	0.8410	0.81
	mean	1160	239	2447	161	87.81	91.10	0.01	0.96	0.8529	0.83
	wmean	1181	186	2500	140	89.40	93.08	0.00	0.97	0.8787	0.86
	add	1205	238	2448	116	91.22	91.14	0.00	0.97	0.8719	0.84
GO:0017111	spmap	1139	247	2433	117	90.68	90.78	0.00	0.96	0.8622	0.82
	blast	1164	198	2483	92	92.68	92.61	0.02	0.98	0.8892	0.85
	peps	863	827	1842	393	68.71	69.01	0.17	0.75	0.5859	0.51
	voting	1144	206	2475	112	91.08	92.32	0.02	0.98	0.8780	0.85
	mean	1153	185	2496	103	91.80	93.10	0.01	0.97	0.8890	0.86
	wmean	1167	149	2532	89	92.91	94.44	0.00	0.97	0.9075	0.89
	add	1176	170	2511	80	93.63	93.66	0.00	0.98	0.9039	0.87
GO:0005198	spmap	1142	350	2409	166	87.31	87.31	0.00	0.93	0.8157	0.77
	blast	1168	295	2464	140	89.30	89.31	0.01	0.96	0.8430	0.80
	peps	904	805	1949	385	70.13	70.77	0.14	0.75	0.6031	0.53
	voting	1172	328	2431	136	89.60	88.11	0.03	0.97	0.8348	0.78
	mean	1175	290	2469	133	89.83	89.49	0.01	0.96	0.8475	0.80
	wmean	1178	253	2506	130	90.06	90.83	0.01	0.96	0.8602	0.82
	add	1175	280	2479	133	89.83	89.85	0.01	0.96	0.8505	0.81
GO:0032553	spmap	1115	352	2317	168	86.91	86.81	0.00	0.94	0.8109	0.76
	blast	1169	238	2431	114	91.11	91.08	0.02	0.97	0.8691	0.83
	peps	837	902	1750	442	65.44	65.99	0.21	0.71	0.5547	0.48
	voting	1121	256	2413	162	87.37	90.41	0.01	0.97	0.8429	0.81
	mean	1135	213	2456	148	88.46	92.02	0.01	0.96	0.8628	0.84
	wmean	1165	172	2497	118	90.80	93.56	0.00	0.97	0.8893	0.87
	add	1182	209	2460	101	92.13	92.17	0.00	0.97	0.8841	0.85
GO:0032555	spmap	1108	371	2341	175	86.36	86.32	0.00	0.94	0.8023	0.75
	blast	1178	221	2490	105	91.82	91.85	0.02	0.97	0.8784	0.84
	peps	838	921	1776	443	65.42	65.85	0.22	0.71	0.5513	0.48
	voting	1135	291	2421	148	88.46	89.27	0.02	0.98	0.8379	0.80
	mean	1148	265	2447	135	89.48	90.23	0.00	0.96	0.8516	0.81
	wmean	1155	183	2529	128	90.02	93.25	0.00	0.97	0.8813	0.86
	add	1185	213	2499	98	92.36	92.15	0.00	0.97	0.8840	0.85
GO:0003723	spmap	1054	351	2413	152	87.40	87.30	0.00	0.95	0.8074	0.75
	blast	1065	322	2443	141	88.31	88.35	0.01	0.96	0.8214	0.77
	peps	887	713	2044	315	73.79	74.14	0.09	0.82	0.6331	0.55
	voting	1080	321	2444	126	89.55	88.39	0.02	0.97	0.8285	0.77
	mean	1093	278	2487	113	90.63	89.95	0.01	0.96	0.8483	0.80
	wmean	1106	251	2514	100	91.71	90.92	0.00	0.97	0.8631	0.82
	add	1092	259	2505	114	90.55	90.63	0.00	0.96	0.8541	0.81
GO:0042802	spmap	750	823	1944	317	70.29	70.26	0.14	0.78	0.5682	0.48
	blast	788	723	2044	279	73.85	73.87	0.09	0.83	0.6113	0.52
	peps	645	1076	1681	419	60.62	60.97	0.30	0.65	0.4632	0.37
	voting	777	768	1999	290	72.82	72.24	0.28	0.88	0.5949	0.50
	mean	792	744	2023	275	74.23	73.11	0.09	0.82	0.6085	0.52
	wmean	793	702	2065	274	74.32	74.63	0.08	0.83	0.6190	0.53
	add	793	710	2057	274	74.32	74.34	0.09	0.83	0.6171	0.53
GO:0001584	spmap	1068	97	2593	39	96.48	96.39	0.00	0.99	0.9401	0.92
	blast	1080	65	2625	27	97.56	97.58	0.00	1.00	0.9591	0.94
	peps	1011	222	2456	95	91.41	91.71	0.01	0.96	0.8645	0.82
	voting	1068	59	2631	39	96.48	97.81	0.00	1.00	0.9561	0.95

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	1073	43	2647	34	96.93	98.40	0.00	0.99	0.9654	0.96
	wmean	1073	40	2650	34	96.93	98.51	0.00	0.99	0.9667	0.96
	add	1069	94	2596	38	96.57	96.51	0.00	0.99	0.9419	0.92
GO:0022803	spmap	984	206	2416	84	92.13	92.14	0.00	0.97	0.8716	0.83
	blast	1015	130	2492	53	95.04	95.04	0.01	0.99	0.9173	0.89
	peps	745	782	1831	322	69.82	70.07	0.15	0.75	0.5744	0.49
	voting	996	140	2482	72	93.26	94.66	0.00	0.99	0.9038	0.88
	mean	1006	110	2512	62	94.19	95.80	0.00	0.98	0.9212	0.90
	wmean	1010	81	2541	58	94.57	96.91	0.00	0.99	0.9356	0.93
	add	1024	108	2514	44	95.88	95.88	0.00	0.99	0.9309	0.90
GO:0015267	spmap	987	197	2405	81	92.42	92.43	0.00	0.98	0.8766	0.83
	blast	1012	138	2464	56	94.76	94.70	0.01	0.99	0.9125	0.88
	peps	748	768	1825	319	70.10	70.38	0.15	0.75	0.5792	0.49
	voting	1001	133	2469	67	93.73	94.89	0.00	0.99	0.9092	0.88
	mean	1003	114	2488	65	93.91	95.62	0.00	0.98	0.9181	0.90
	wmean	1015	86	2516	53	95.04	96.69	0.00	0.99	0.9359	0.92
	add	1019	120	2482	49	95.41	95.39	0.00	0.99	0.9234	0.89
GO:0030554	spmap	883	400	2319	153	85.23	85.29	0.00	0.93	0.7615	0.69
	blast	943	245	2474	93	91.02	90.99	0.02	0.97	0.8480	0.79
	peps	686	903	1803	349	66.28	66.63	0.20	0.70	0.5229	0.43
	voting	920	300	2419	116	88.80	88.97	0.02	0.97	0.8156	0.75
	mean	925	253	2466	111	89.29	90.70	0.01	0.96	0.8356	0.79
	wmean	937	187	2532	99	90.44	93.12	0.00	0.97	0.8676	0.83
	add	952	223	2496	84	91.89	91.80	0.00	0.97	0.8611	0.81
GO:0022838	spmap	954	202	2421	78	92.44	92.30	0.00	0.98	0.8720	0.83
	blast	988	115	2508	44	95.74	95.62	0.02	0.99	0.9255	0.90
	peps	725	770	1845	306	70.32	70.55	0.15	0.75	0.5740	0.48
	voting	966	130	2493	66	93.60	95.04	0.00	0.99	0.9079	0.88
	mean	972	110	2513	60	94.19	95.81	0.00	0.98	0.9196	0.90
	wmean	985	85	2538	47	95.45	96.76	0.00	0.99	0.9372	0.92
	add	991	103	2520	41	96.03	96.07	0.00	0.99	0.9323	0.91
GO:0032559	spmap	872	377	2341	140	86.17	86.13	0.00	0.93	0.7713	0.70
	blast	931	217	2501	81	92.00	92.02	0.02	0.97	0.8620	0.81
	peps	681	867	1830	330	67.36	67.85	0.20	0.71	0.5322	0.44
	voting	892	271	2447	120	88.14	90.03	0.01	0.97	0.8202	0.77
	mean	903	244	2474	109	89.23	91.02	0.00	0.96	0.8365	0.79
	wmean	915	188	2530	97	90.42	93.08	0.00	0.97	0.8652	0.83
	add	933	215	2503	79	92.19	92.09	0.00	0.97	0.8639	0.81
GO:0008092	spmap	807	375	2372	128	86.31	86.35	0.00	0.93	0.7624	0.68
	blast	832	302	2445	103	88.98	89.01	0.03	0.95	0.8043	0.73
	peps	664	773	1961	268	71.24	71.73	0.12	0.78	0.5606	0.46
	voting	833	331	2416	102	89.09	87.95	0.02	0.97	0.7937	0.72
	mean	834	301	2446	101	89.20	89.04	0.01	0.95	0.8058	0.73
	wmean	837	261	2486	98	89.52	90.50	0.01	0.96	0.8234	0.76
	add	840	278	2468	95	89.84	89.88	0.00	0.96	0.8183	0.75
GO:0046873	spmap	905	215	2462	79	91.97	91.97	0.00	0.98	0.8603	0.81
	blast	938	126	2551	46	95.33	95.29	0.01	0.99	0.9160	0.88
	peps	700	756	1911	282	71.28	71.65	0.14	0.76	0.5742	0.48
	voting	922	147	2530	62	93.70	94.51	0.00	0.99	0.8982	0.86
	mean	925	118	2559	59	94.00	95.59	0.00	0.98	0.9127	0.89
	wmean	929	81	2596	55	94.41	96.97	0.00	0.99	0.9318	0.92
	add	941	115	2562	43	95.63	95.70	0.00	0.99	0.9225	0.89
GO:0005524	spmap	869	358	2370	131	86.90	86.88	0.00	0.94	0.7804	0.71
	blast	924	210	2518	76	92.40	92.30	0.02	0.97	0.8660	0.81
	peps	669	881	1828	329	67.03	67.48	0.20	0.71	0.5251	0.43
	voting	898	265	2463	102	89.80	90.29	0.01	0.97	0.8303	0.77
	mean	906	226	2502	94	90.60	91.72	0.00	0.96	0.8499	0.80
	wmean	910	167	2561	90	91.00	93.88	0.00	0.97	0.8763	0.84
	add	929	195	2533	71	92.90	92.85	0.00	0.98	0.8748	0.83
GO:0005216	spmap	923	194	2438	71	92.86	92.63	0.00	0.98	0.8745	0.83
	blast	947	129	2503	47	95.27	95.10	0.02	0.99	0.9150	0.88

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	716	725	1897	277	72.10	72.35	0.13	0.77	0.5883	0.50
	voting	937	125	2507	57	94.27	95.25	0.00	0.99	0.9115	0.88
	mean	941	105	2527	53	94.67	96.01	0.00	0.98	0.9225	0.90
	wmean	950	84	2548	44	95.57	96.81	0.00	0.99	0.9369	0.92
	add	954	104	2527	40	95.98	96.05	0.00	0.99	0.9298	0.90
GO:0004674	spmap	809	223	2537	71	91.93	91.92	0.00	0.97	0.8462	0.78
	blast	826	169	2591	54	93.86	93.88	0.01	0.98	0.8811	0.83
	peps	632	768	1982	248	71.82	72.07	0.13	0.79	0.5544	0.45
	voting	811	150	2610	69	92.16	94.57	0.01	0.98	0.8810	0.84
	mean	815	118	2642	65	92.61	95.72	0.00	0.98	0.8991	0.87
	wmean	827	98	2662	53	93.98	96.45	0.00	0.98	0.9163	0.89
	add	832	147	2613	48	94.55	94.67	0.00	0.98	0.8951	0.85
GO:0042578	spmap	787	343	2382	114	87.35	87.41	0.00	0.95	0.7750	0.70
	blast	848	164	2561	53	94.12	93.98	0.01	0.98	0.8866	0.84
	peps	627	810	1902	273	69.67	70.13	0.17	0.75	0.5366	0.44
	voting	818	195	2530	83	90.79	92.84	0.01	0.98	0.8548	0.81
	mean	830	160	2565	71	92.12	94.13	0.00	0.97	0.8778	0.84
	wmean	842	114	2611	59	93.45	95.82	0.00	0.98	0.9068	0.88
	add	853	146	2579	48	94.67	94.64	0.00	0.98	0.8979	0.85
GO:0004175	spmap	804	314	2407	104	88.55	88.46	0.00	0.95	0.7937	0.72
	blast	858	155	2566	50	94.49	94.30	0.01	0.98	0.8933	0.85
	peps	624	843	1869	284	68.72	68.92	0.13	0.75	0.5255	0.43
	voting	823	197	2524	85	90.64	92.76	0.01	0.98	0.8537	0.81
	mean	829	170	2551	79	91.30	93.75	0.00	0.97	0.8694	0.83
	wmean	852	127	2594	56	93.83	95.33	0.00	0.98	0.9030	0.87
	add	862	136	2585	46	94.93	95.00	0.00	0.98	0.9045	0.86
GO:0008134	spmap	739	493	2262	161	82.11	82.11	0.02	0.90	0.6932	0.60
	blast	746	477	2278	154	82.89	82.69	0.01	0.92	0.7028	0.61
	peps	624	827	1916	274	69.49	69.85	0.14	0.77	0.5313	0.43
	voting	759	470	2285	141	84.33	82.94	0.05	0.94	0.7130	0.62
	mean	759	433	2322	141	84.33	84.28	0.03	0.92	0.7256	0.64
	wmean	773	402	2353	127	85.89	85.41	0.02	0.93	0.7451	0.66
	add	774	392	2363	126	86.00	85.77	0.02	0.93	0.7493	0.66
GO:0016887	spmap	726	284	2424	84	89.63	89.51	0.00	0.95	0.7978	0.72
	blast	749	206	2502	61	92.47	92.39	0.01	0.98	0.8487	0.78
	peps	601	695	2003	209	74.20	74.24	0.10	0.79	0.5708	0.46
	voting	728	202	2506	82	89.88	92.54	0.02	0.98	0.8368	0.78
	mean	728	164	2544	82	89.88	93.94	0.00	0.96	0.8555	0.82
	wmean	737	142	2566	73	90.99	94.76	0.00	0.97	0.8727	0.84
	add	756	179	2529	54	93.33	93.39	0.00	0.97	0.8665	0.81
GO:0016563	spmap	717	411	2366	124	85.26	85.20	0.01	0.92	0.7283	0.64
	blast	723	391	2386	118	85.97	85.92	0.01	0.94	0.7396	0.65
	peps	628	684	2079	212	74.76	75.24	0.09	0.82	0.5836	0.48
	voting	741	373	2404	100	88.11	86.57	0.05	0.96	0.7581	0.67
	mean	739	341	2436	102	87.87	87.72	0.02	0.94	0.7694	0.68
	wmean	740	327	2450	101	87.99	88.22	0.02	0.95	0.7757	0.69
	add	735	352	2425	106	87.40	87.32	0.02	0.94	0.7624	0.68
GO:0016874	spmap	670	370	2363	104	86.56	86.46	0.00	0.94	0.7387	0.64
	blast	723	181	2552	51	93.41	93.38	0.01	0.98	0.8617	0.80
	peps	501	954	1772	272	64.81	65.00	0.24	0.69	0.4497	0.34
	voting	694	254	2479	80	89.66	90.71	0.01	0.98	0.8060	0.73
	mean	695	198	2535	79	89.79	92.76	0.00	0.97	0.8338	0.78
	wmean	711	146	2587	63	91.86	94.66	0.00	0.98	0.8719	0.83
	add	726	173	2560	48	93.80	93.67	0.00	0.98	0.8679	0.81
GO:0022804	spmap	720	199	2408	60	92.31	92.37	0.00	0.97	0.8476	0.78
	blast	739	135	2472	41	94.74	94.82	0.01	0.99	0.8936	0.85
	peps	586	637	1964	192	75.32	75.51	0.04	0.84	0.5857	0.48
	voting	727	144	2463	53	93.21	94.48	0.01	0.99	0.8807	0.83
	mean	732	108	2499	48	93.85	95.86	0.00	0.98	0.9037	0.87
	wmean	735	86	2521	45	94.23	96.70	0.00	0.99	0.9182	0.90

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	740	134	2473	40	94.87	94.86	0.00	0.98	0.8948	0.85
GO:0005261	spmap	760	184	2556	55	93.25	93.28	0.00	0.98	0.8641	0.81
	blast	777	126	2614	38	95.34	95.40	0.01	0.99	0.9045	0.86
	peps	596	725	2006	217	73.31	73.45	0.13	0.78	0.5586	0.45
	voting	768	139	2601	47	94.23	94.93	0.01	0.99	0.8920	0.85
	mean	772	122	2618	43	94.72	95.55	0.00	0.98	0.9035	0.86
	wmean	780	82	2658	35	95.71	97.01	0.00	0.99	0.9302	0.90
	add	784	104	2636	31	96.20	96.20	0.00	0.99	0.9207	0.88
GO:0046914	spmap	647	340	2428	91	87.67	87.72	0.00	0.95	0.7501	0.66
	blast	680	217	2551	58	92.14	92.16	0.01	0.98	0.8318	0.76
	peps	488	922	1839	248	66.30	66.61	0.14	0.73	0.4548	0.35
	voting	665	262	2506	73	90.11	90.53	0.01	0.98	0.7988	0.72
	mean	672	221	2547	66	91.06	92.02	0.00	0.97	0.8240	0.75
	wmean	680	159	2609	58	92.14	94.26	0.00	0.98	0.8624	0.81
	add	689	184	2584	49	93.36	93.35	0.00	0.98	0.8554	0.79
GO:0046983	spmap	515	621	2159	148	77.68	77.66	0.02	0.85	0.5725	0.45
	blast	555	453	2327	108	83.71	83.71	0.03	0.92	0.6643	0.55
	peps	407	1059	1712	254	61.57	61.78	0.25	0.66	0.3827	0.28
	voting	529	508	2272	134	79.79	81.73	0.13	0.93	0.6224	0.51
	mean	527	458	2322	136	79.49	83.53	0.02	0.89	0.6396	0.54
	wmean	535	396	2384	128	80.69	85.76	0.01	0.91	0.6713	0.57
	add	553	463	2317	110	83.41	83.35	0.01	0.91	0.6587	0.54
GO:0016564	spmap	664	338	2445	91	87.95	87.85	0.01	0.94	0.7558	0.66
	blast	668	323	2460	87	88.48	88.39	0.01	0.96	0.7652	0.67
	peps	554	726	2048	199	73.57	73.83	0.08	0.82	0.5450	0.43
	voting	682	308	2475	73	90.33	88.93	0.03	0.97	0.7817	0.69
	mean	679	274	2509	76	89.93	90.15	0.01	0.96	0.7951	0.71
	wmean	681	242	2541	74	90.20	91.30	0.01	0.96	0.8117	0.74
	add	684	260	2523	71	90.60	90.66	0.01	0.96	0.8052	0.72
GO:0022836	spmap	704	217	2456	63	91.79	91.88	0.00	0.97	0.8341	0.76
	blast	732	121	2552	35	95.44	95.47	0.01	0.99	0.9037	0.86
	peps	561	711	1956	205	73.24	73.34	0.13	0.78	0.5505	0.44
	voting	717	150	2523	50	93.48	94.39	0.00	0.99	0.8776	0.83
	mean	722	117	2556	45	94.13	95.62	0.00	0.98	0.8991	0.86
	wmean	727	84	2589	40	94.78	96.86	0.00	0.99	0.9214	0.90
	add	735	111	2562	32	95.83	95.85	0.00	0.99	0.9113	0.87
GO:0042623	spmap	610	304	2428	75	89.05	88.87	0.00	0.95	0.7630	0.67
	blast	631	217	2515	54	92.12	92.06	0.01	0.98	0.8232	0.74
	peps	498	734	1991	185	72.91	73.06	0.10	0.78	0.5201	0.40
	voting	616	198	2534	69	89.93	92.75	0.01	0.98	0.8219	0.76
	mean	624	169	2563	61	91.09	93.81	0.00	0.97	0.8444	0.79
	wmean	628	133	2599	57	91.68	95.13	0.00	0.97	0.8686	0.83
	add	635	203	2529	50	92.70	92.57	0.00	0.98	0.8339	0.76
GO:0016791	spmap	636	266	2485	67	90.47	90.33	0.00	0.96	0.7925	0.71
	blast	670	131	2620	33	95.31	95.24	0.01	0.99	0.8910	0.84
	peps	506	757	1980	197	71.98	72.34	0.14	0.78	0.5148	0.40
	voting	651	146	2605	52	92.60	94.69	0.00	0.99	0.8680	0.82
	mean	650	117	2634	53	92.46	95.75	0.00	0.98	0.8844	0.85
	wmean	662	86	2665	41	94.17	96.87	0.00	0.99	0.9125	0.89
	add	671	123	2628	32	95.45	95.53	0.00	0.99	0.8965	0.85
GO:0004713	spmap	665	189	2552	49	93.14	93.10	0.00	0.98	0.8482	0.78
	blast	685	110	2632	29	95.94	95.99	0.01	0.99	0.9079	0.86
	peps	546	627	2106	165	76.79	77.06	0.04	0.84	0.5796	0.47
	voting	673	105	2637	41	94.26	96.17	0.00	0.99	0.9021	0.87
	mean	672	88	2654	42	94.12	96.79	0.00	0.98	0.9118	0.88
	wmean	678	69	2673	36	94.96	97.48	0.00	0.98	0.9281	0.91
	add	682	124	2618	32	95.52	95.48	0.00	0.98	0.8974	0.85
GO:0030695	spmap	588	306	2418	73	88.96	88.77	0.00	0.95	0.7563	0.66
	blast	620	168	2556	41	93.80	93.83	0.01	0.98	0.8558	0.79
	peps	481	729	1981	179	72.88	73.10	0.11	0.80	0.5144	0.40
	voting	604	225	2499	57	91.38	91.74	0.00	0.99	0.8107	0.73

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	609	191	2533	52	92.13	92.99	0.00	0.98	0.8337	0.76
	wmean	619	131	2593	42	93.65	95.19	0.00	0.98	0.8774	0.83
	add	627	142	2582	34	94.86	94.79	0.00	0.98	0.8769	0.82
GO:0005509	spmap	565	318	2449	71	88.84	88.51	0.00	0.96	0.7439	0.64
	blast	590	203	2564	46	92.77	92.66	0.01	0.97	0.8258	0.74
	peps	433	862	1889	202	68.19	68.67	0.17	0.74	0.4487	0.33
	voting	585	240	2527	51	91.98	91.33	0.01	0.98	0.8008	0.71
	mean	583	196	2571	53	91.67	92.92	0.00	0.97	0.8240	0.75
	wmean	587	146	2621	49	92.30	94.72	0.00	0.98	0.8576	0.80
	add	597	171	2596	39	93.87	93.82	0.00	0.98	0.8504	0.78
GO:0003712	spmap	554	482	2285	115	82.81	82.58	0.02	0.91	0.6499	0.53
	blast	560	456	2311	109	83.71	83.52	0.02	0.93	0.6647	0.55
	peps	478	786	1974	190	71.56	71.52	0.14	0.78	0.4948	0.38
	voting	569	440	2327	100	85.05	84.10	0.05	0.95	0.6782	0.56
	mean	569	402	2365	100	85.05	85.47	0.03	0.93	0.6939	0.59
	wmean	575	379	2388	94	85.95	86.30	0.02	0.94	0.7086	0.60
	add	576	388	2379	93	86.10	85.98	0.01	0.94	0.7055	0.60
GO:0042277	spmap	582	258	2493	61	90.51	90.62	0.00	0.95	0.7849	0.69
	blast	606	156	2595	37	94.25	94.33	0.01	0.98	0.8626	0.80
	peps	479	690	2047	164	74.49	74.79	0.05	0.82	0.5287	0.41
	voting	591	180	2571	52	91.91	93.46	0.01	0.98	0.8359	0.77
	mean	590	149	2602	53	91.76	94.58	0.00	0.97	0.8538	0.80
	wmean	595	119	2632	48	92.53	95.67	0.00	0.97	0.8769	0.83
	add	593	218	2533	50	92.22	92.08	0.00	0.97	0.8157	0.73
GO:0016757	spmap	522	271	2489	56	90.31	90.18	0.00	0.96	0.7615	0.66
	blast	550	141	2619	28	95.16	94.89	0.01	0.99	0.8668	0.80
	peps	447	616	2135	130	77.47	77.61	0.05	0.86	0.5451	0.42
	voting	532	160	2600	46	92.04	94.20	0.01	0.99	0.8378	0.77
	mean	533	137	2623	45	92.21	95.04	0.00	0.98	0.8542	0.80
	wmean	537	103	2657	41	92.91	96.27	0.00	0.98	0.8818	0.84
	add	548	148	2612	30	94.81	94.64	0.00	0.98	0.8603	0.79
GO:0003702	spmap	490	430	2344	89	84.63	84.50	0.01	0.92	0.6538	0.53
	blast	497	389	2385	82	85.84	85.98	0.01	0.94	0.6785	0.56
	peps	452	595	2167	126	78.20	78.46	0.08	0.85	0.5563	0.43
	voting	503	363	2411	76	86.87	86.91	0.04	0.96	0.6962	0.58
	mean	510	343	2431	69	88.08	87.64	0.02	0.94	0.7123	0.60
	wmean	509	330	2444	70	87.91	88.10	0.02	0.94	0.7179	0.61
	add	512	318	2456	67	88.43	88.54	0.01	0.95	0.7268	0.62
GO:0019899	spmap	427	627	2104	128	76.94	77.04	0.03	0.84	0.5308	0.41
	blast	440	566	2165	115	79.28	79.27	0.03	0.89	0.5637	0.44
	peps	348	1004	1723	205	62.93	63.18	0.24	0.67	0.3654	0.26
	voting	436	548	2183	119	78.56	79.93	0.20	0.92	0.5666	0.44
	mean	435	524	2207	120	78.38	80.81	0.04	0.88	0.5746	0.45
	wmean	451	477	2254	104	81.26	82.53	0.01	0.90	0.6082	0.49
	add	450	516	2215	105	81.08	81.11	0.01	0.90	0.5917	0.47
GO:0003779	spmap	448	356	2420	65	87.33	87.18	0.00	0.94	0.6803	0.56
	blast	466	254	2522	47	90.84	90.85	0.02	0.97	0.7559	0.65
	peps	356	838	1930	155	69.67	69.73	0.13	0.76	0.4176	0.30
	voting	456	276	2500	57	88.89	90.06	0.02	0.97	0.7325	0.62
	mean	461	249	2527	52	89.86	91.03	0.01	0.96	0.7539	0.65
	wmean	467	204	2572	46	91.03	92.65	0.00	0.97	0.7889	0.70
	add	473	223	2553	40	92.20	91.97	0.00	0.96	0.7825	0.68
GO:0019199	spmap	525	125	2619	25	95.45	95.44	0.00	0.99	0.8750	0.81
	blast	542	49	2695	8	98.55	98.21	0.00	1.00	0.9500	0.92
	peps	458	437	2298	90	83.58	84.02	0.02	0.90	0.6348	0.51
	voting	533	61	2683	17	96.91	97.78	0.00	1.00	0.9318	0.90
	mean	534	49	2695	16	97.09	98.21	0.00	1.00	0.9426	0.92
	wmean	537	36	2708	13	97.64	98.69	0.00	1.00	0.9564	0.94
	add	537	70	2674	13	97.64	97.45	0.00	1.00	0.9283	0.88
GO:0008047	spmap	412	529	2220	97	80.94	80.76	0.00	0.89	0.5683	0.44
	blast	450	329	2420	59	88.41	88.03	0.01	0.96	0.6988	0.58

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	345	864	1875	162	68.05	68.46	0.18	0.74	0.4021	0.29
	voting	432	381	2368	77	84.87	86.14	0.02	0.96	0.6536	0.53
	mean	438	324	2425	71	86.05	88.21	0.01	0.93	0.6892	0.57
	wmean	445	260	2489	64	87.43	90.54	0.00	0.94	0.7331	0.63
	add	452	312	2437	57	88.80	88.65	0.00	0.94	0.7101	0.59
GO:0016879	spmap	408	386	2390	66	86.08	86.10	0.00	0.94	0.6435	0.51
	blast	446	173	2603	28	94.09	93.77	0.00	0.98	0.8161	0.72
	peps	325	853	1913	147	68.86	69.16	0.15	0.75	0.3939	0.28
	voting	418	232	2544	56	88.19	91.64	0.01	0.98	0.7438	0.64
	mean	431	194	2582	43	90.93	93.01	0.00	0.98	0.7843	0.69
	wmean	441	143	2633	33	93.04	94.85	0.00	0.98	0.8336	0.76
	add	447	162	2614	27	94.30	94.16	0.00	0.99	0.8255	0.73
GO:0008270	spmap	443	346	2448	62	87.72	87.62	0.00	0.94	0.6847	0.56
	blast	462	236	2558	43	91.49	91.55	0.00	0.98	0.7681	0.66
	peps	368	739	2043	135	73.16	73.44	0.06	0.81	0.4571	0.33
	voting	451	244	2550	54	89.31	91.27	0.01	0.98	0.7517	0.65
	mean	455	213	2581	50	90.10	92.38	0.00	0.96	0.7758	0.68
	wmean	464	171	2623	41	91.88	93.88	0.00	0.97	0.8140	0.73
	add	470	191	2603	35	93.07	93.16	0.00	0.97	0.8062	0.71
GO:0003735	spmap	473	200	2588	37	92.75	92.83	0.00	0.97	0.7997	0.70
	blast	463	221	2567	47	90.78	92.07	0.00	0.98	0.7755	0.68
	peps	451	305	2473	58	88.61	89.02	0.01	0.94	0.7130	0.60
	voting	485	138	2650	25	95.10	95.05	0.01	1.00	0.8561	0.78
	mean	485	115	2673	25	95.10	95.88	0.00	0.99	0.8739	0.81
	wmean	485	108	2680	25	95.10	96.13	0.00	0.99	0.8794	0.82
	add	488	118	2670	22	95.69	95.77	0.00	0.99	0.8746	0.81
GO:0016746	spmap	383	422	2279	71	84.36	84.38	0.00	0.91	0.6084	0.48
	blast	420	205	2496	34	92.51	92.41	0.01	0.98	0.7785	0.67
	peps	312	829	1862	141	68.87	69.19	0.14	0.76	0.3915	0.27
	voting	396	241	2460	58	87.22	91.08	0.01	0.98	0.7259	0.62
	mean	398	212	2489	56	87.67	92.15	0.00	0.96	0.7481	0.65
	wmean	408	164	2537	46	89.87	93.93	0.00	0.96	0.7953	0.71
	add	418	214	2487	36	92.07	92.08	0.00	0.97	0.7698	0.66
GO:0004714	spmap	484	119	2664	21	95.84	95.72	0.00	0.99	0.8736	0.80
	blast	494	58	2725	11	97.82	97.92	0.01	1.00	0.9347	0.89
	peps	416	466	2305	87	82.70	83.18	0.01	0.90	0.6007	0.47
	voting	487	71	2712	18	96.44	97.45	0.00	1.00	0.9163	0.87
	mean	491	54	2729	14	97.23	98.06	0.00	1.00	0.9352	0.90
	wmean	493	45	2738	12	97.62	98.38	0.00	1.00	0.9453	0.92
	add	495	65	2718	10	98.02	97.66	0.00	1.00	0.9296	0.88
GO:0016829	spmap	383	325	2398	52	88.05	88.06	0.00	0.94	0.6702	0.54
	blast	406	182	2541	29	93.33	93.32	0.01	0.98	0.7937	0.69
	peps	326	660	2053	107	75.29	75.67	0.11	0.82	0.4595	0.33
	voting	392	186	2537	43	90.11	93.17	0.01	0.98	0.7739	0.68
	mean	395	158	2565	40	90.80	94.20	0.00	0.97	0.7996	0.71
	wmean	400	131	2592	35	91.95	95.19	0.00	0.98	0.8282	0.75
	add	406	180	2543	29	93.33	93.39	0.00	0.97	0.7953	0.69
GO:0005125	spmap	408	321	2448	53	88.50	88.41	0.00	0.94	0.6857	0.56
	blast	421	237	2532	40	91.32	91.44	0.01	0.98	0.7525	0.64
	peps	349	652	2107	110	76.03	76.37	0.08	0.84	0.4781	0.35
	voting	414	228	2541	47	89.80	91.77	0.01	0.98	0.7507	0.64
	mean	413	203	2566	48	89.59	92.67	0.00	0.97	0.7669	0.67
	wmean	418	166	2603	43	90.67	94.01	0.00	0.97	0.8000	0.72
	add	422	231	2538	39	91.54	91.66	0.00	0.97	0.7576	0.65
GO:0022890	spmap	374	342	2400	53	87.59	87.53	0.00	0.94	0.6544	0.52
	blast	385	273	2469	42	90.16	90.04	0.00	0.98	0.7097	0.59
	peps	307	754	1980	119	72.07	72.42	0.09	0.77	0.4129	0.29
	voting	377	263	2479	50	88.29	90.41	0.02	0.97	0.7067	0.59
	mean	379	228	2514	48	88.76	91.68	0.00	0.96	0.7331	0.62
	wmean	388	183	2559	39	90.87	93.33	0.00	0.97	0.7776	0.68

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	391	233	2509	36	91.57	91.50	0.00	0.97	0.7441	0.63
GO:0004857	spmap	330	459	2287	67	83.12	83.28	0.00	0.90	0.5565	0.42
	blast	339	398	2348	58	85.39	85.51	0.00	0.96	0.5979	0.46
	peps	259	939	1799	137	65.40	65.70	0.17	0.69	0.3250	0.22
	voting	329	375	2371	68	82.87	86.34	0.05	0.96	0.5976	0.47
	mean	334	342	2404	63	84.13	87.55	0.00	0.92	0.6226	0.49
	wmean	336	271	2475	61	84.63	90.13	0.00	0.93	0.6693	0.55
	add	340	392	2354	57	85.64	85.72	0.00	0.93	0.6023	0.46
GO:0004721	spmap	398	231	2541	36	91.71	91.67	0.00	0.97	0.7488	0.63
	blast	414	134	2638	20	95.39	95.17	0.00	0.99	0.8432	0.76
	peps	335	622	2141	99	77.19	77.49	0.09	0.83	0.4817	0.35
	voting	403	126	2646	31	92.86	95.45	0.00	0.99	0.8370	0.76
	mean	407	101	2671	27	93.78	96.36	0.00	0.99	0.8641	0.80
	wmean	408	79	2693	26	94.01	97.15	0.00	0.99	0.8860	0.84
	add	413	131	2641	21	95.16	95.27	0.00	0.99	0.8446	0.76
GO:0003924	spmap	411	133	2650	20	95.36	95.22	0.00	0.98	0.8431	0.76
	blast	417	98	2685	14	96.75	96.48	0.01	0.99	0.8816	0.81
	peps	335	610	2163	96	77.73	78.00	0.05	0.85	0.4869	0.35
	voting	412	83	2700	19	95.59	97.02	0.01	0.99	0.8898	0.83
	mean	411	56	2727	20	95.36	97.99	0.00	0.99	0.9154	0.88
	wmean	414	45	2738	17	96.06	98.38	0.00	0.99	0.9303	0.90
	add	418	94	2689	13	96.98	96.62	0.00	0.98	0.8865	0.82
GO:0016758	spmap	359	283	2472	39	90.20	89.73	0.00	0.96	0.6904	0.56
	blast	376	162	2593	22	94.47	94.12	0.01	0.98	0.8034	0.70
	peps	328	474	2274	69	82.62	82.75	0.03	0.90	0.5471	0.41
	voting	370	145	2610	28	92.96	94.74	0.00	0.98	0.8105	0.72
	mean	370	125	2630	28	92.96	95.46	0.00	0.98	0.8287	0.75
	wmean	372	105	2650	26	93.47	96.19	0.00	0.98	0.8503	0.78
	add	374	163	2592	24	93.97	94.08	0.00	0.97	0.8000	0.70
GO:0009055	spmap	357	451	2321	69	83.80	83.73	0.00	0.90	0.5786	0.44
	blast	377	320	2452	49	88.50	88.46	0.01	0.96	0.6714	0.54
	peps	283	913	1854	141	66.75	67.00	0.15	0.72	0.3494	0.24
	voting	361	321	2451	65	84.74	88.42	0.02	0.97	0.6516	0.53
	mean	364	305	2467	62	85.45	89.00	0.01	0.94	0.6648	0.54
	wmean	368	252	2520	58	86.38	90.91	0.00	0.95	0.7036	0.59
	add	377	317	2455	49	88.50	88.56	0.00	0.94	0.6732	0.54
GO:0003682	spmap	389	341	2460	53	88.01	87.83	0.00	0.94	0.6638	0.53
	blast	396	291	2510	46	89.59	89.61	0.00	0.96	0.7015	0.58
	peps	319	747	2046	119	72.83	73.25	0.11	0.77	0.4242	0.30
	voting	394	276	2525	48	89.14	90.15	0.02	0.97	0.7086	0.59
	mean	398	262	2539	44	90.05	90.65	0.01	0.95	0.7223	0.60
	wmean	394	226	2575	48	89.14	91.93	0.00	0.96	0.7420	0.64
	add	401	262	2539	41	90.72	90.65	0.00	0.96	0.7258	0.60
GO:0042803	spmap	308	584	2188	81	79.18	78.93	0.01	0.86	0.4809	0.35
	blast	315	532	2240	74	80.98	80.81	0.02	0.91	0.5097	0.37
	peps	239	1041	1724	147	61.92	62.35	0.25	0.66	0.2869	0.19
	voting	307	511	2261	82	78.92	81.57	0.12	0.92	0.5087	0.38
	mean	306	503	2269	83	78.66	81.85	0.02	0.87	0.5109	0.38
	wmean	304	449	2323	85	78.15	83.80	0.01	0.89	0.5324	0.40
	add	316	527	2245	73	81.23	80.99	0.01	0.88	0.5130	0.37
GO:0030246	spmap	334	332	2435	45	88.13	88.00	0.00	0.93	0.6392	0.50
	blast	346	243	2524	33	91.29	91.22	0.00	0.98	0.7149	0.59
	peps	277	728	2032	100	73.47	73.62	0.06	0.81	0.4009	0.28
	voting	341	239	2528	38	89.97	91.36	0.02	0.97	0.7112	0.59
	mean	339	219	2548	40	89.45	92.09	0.00	0.95	0.7236	0.61
	wmean	340	181	2586	39	89.71	93.46	0.00	0.96	0.7556	0.65
	add	344	255	2512	35	90.77	90.78	0.00	0.96	0.7035	0.57
GO:0016881	spmap	336	361	2401	49	87.27	86.93	0.00	0.94	0.6211	0.48
	blast	364	161	2601	21	94.55	94.17	0.01	0.99	0.8000	0.69
	peps	282	732	2018	103	73.25	73.38	0.12	0.79	0.4031	0.28
	voting	351	211	2551	34	91.17	92.36	0.01	0.99	0.7413	0.62

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	355	175	2587	30	92.21	93.66	0.00	0.98	0.7760	0.67
	wmean	363	127	2635	22	94.29	95.40	0.00	0.98	0.8297	0.74
	add	365	148	2614	20	94.81	94.64	0.00	0.98	0.8129	0.71
GO:0001653	spmap	415	88	2667	14	96.74	96.81	0.00	0.99	0.8906	0.83
	blast	421	50	2705	8	98.14	98.19	0.00	1.00	0.9356	0.89
	peps	387	262	2482	42	90.21	90.45	0.02	0.95	0.7180	0.60
	voting	417	48	2707	12	97.20	98.26	0.00	1.00	0.9329	0.90
	mean	417	43	2712	12	97.20	98.44	0.00	0.99	0.9381	0.91
	wmean	417	41	2714	12	97.20	98.51	0.00	0.99	0.9402	0.91
	add	412	110	2645	17	96.04	96.01	0.00	0.99	0.8665	0.79
GO:0022832	spmap	395	168	2579	25	94.05	93.88	0.00	0.98	0.8037	0.70
	blast	397	154	2593	23	94.52	94.39	0.01	0.99	0.8177	0.72
	peps	308	724	2014	111	73.51	73.56	0.09	0.80	0.4245	0.30
	voting	392	120	2627	28	93.33	95.63	0.00	0.99	0.8412	0.77
	mean	392	94	2653	28	93.33	96.58	0.00	0.99	0.8653	0.81
	wmean	397	74	2673	23	94.52	97.31	0.00	0.99	0.8911	0.84
	add	400	132	2615	20	95.24	95.19	0.00	0.99	0.8403	0.75
GO:0005244	spmap	390	191	2555	29	93.08	93.04	0.00	0.98	0.7800	0.67
	blast	397	142	2603	22	94.75	94.83	0.01	0.99	0.8288	0.74
	peps	308	721	2016	110	73.68	73.66	0.11	0.80	0.4257	0.30
	voting	391	140	2606	28	93.32	94.90	0.01	0.99	0.8232	0.74
	mean	396	117	2629	23	94.51	95.74	0.00	0.99	0.8498	0.77
	wmean	399	85	2661	20	95.23	96.90	0.00	0.99	0.8837	0.82
	add	400	126	2619	19	95.47	95.41	0.00	0.99	0.8466	0.76
GO:0008528	spmap	402	81	2669	12	97.10	97.05	0.00	0.99	0.8963	0.83
	blast	406	57	2693	8	98.07	97.93	0.00	1.00	0.9259	0.88
	peps	379	225	2513	35	91.55	91.78	0.02	0.95	0.7446	0.63
	voting	404	48	2702	10	97.58	98.25	0.00	1.00	0.9330	0.89
	mean	401	44	2706	13	96.86	98.40	0.00	0.99	0.9336	0.90
	wmean	401	40	2710	13	96.86	98.55	0.00	0.99	0.9380	0.91
	add	400	105	2645	14	96.62	96.18	0.00	0.99	0.8705	0.79
GO:0016747	spmap	316	459	2277	63	83.38	83.22	0.00	0.90	0.5477	0.41
	blast	348	225	2511	31	91.82	91.78	0.00	0.98	0.7311	0.61
	peps	263	810	1918	114	69.76	70.31	0.13	0.77	0.3628	0.25
	voting	332	255	2481	47	87.60	90.68	0.01	0.97	0.6874	0.57
	mean	334	227	2509	45	88.13	91.70	0.00	0.95	0.7106	0.60
	wmean	340	166	2570	39	89.71	93.93	0.00	0.96	0.7684	0.67
	add	346	242	2494	33	91.29	91.15	0.00	0.96	0.7156	0.59
GO:0008415	spmap	306	503	2207	70	81.38	81.44	0.00	0.88	0.5165	0.38
	blast	346	225	2485	30	92.02	91.70	0.01	0.98	0.7307	0.61
	peps	261	812	1890	114	69.60	69.95	0.11	0.77	0.3605	0.24
	voting	327	283	2427	49	86.97	89.56	0.01	0.97	0.6633	0.54
	mean	325	219	2491	51	86.44	91.92	0.00	0.95	0.7065	0.60
	wmean	339	173	2537	37	90.16	93.62	0.00	0.96	0.7635	0.66
	add	345	229	2481	31	91.76	91.55	0.00	0.96	0.7263	0.60
GO:0004386	spmap	367	167	2590	24	93.86	93.94	0.00	0.98	0.7935	0.69
	blast	373	125	2632	18	95.40	95.47	0.01	0.99	0.8391	0.75
	peps	309	564	2177	81	79.23	79.42	0.06	0.85	0.4893	0.35
	voting	369	108	2649	22	94.37	96.08	0.00	0.99	0.8502	0.77
	mean	369	88	2669	22	94.37	96.81	0.00	0.99	0.8703	0.81
	wmean	371	58	2699	20	94.88	97.90	0.00	0.99	0.9049	0.86
	add	376	107	2650	15	96.16	96.12	0.00	0.99	0.8604	0.78
GO:0008234	spmap	363	215	2567	30	92.37	92.27	0.00	0.96	0.7477	0.63
	blast	376	130	2652	17	95.67	95.33	0.00	0.99	0.8365	0.74
	peps	289	731	2044	104	73.54	73.66	0.09	0.81	0.4091	0.28
	voting	367	136	2646	26	93.38	95.11	0.00	0.98	0.8192	0.73
	mean	367	108	2674	26	93.38	96.12	0.00	0.98	0.8456	0.77
	wmean	372	74	2708	21	94.66	97.34	0.00	0.98	0.8868	0.83
	add	375	126	2656	18	95.42	95.47	0.00	0.98	0.8389	0.75
GO:0008237	spmap	358	198	2556	26	93.23	92.81	0.00	0.97	0.7617	0.64
	blast	371	93	2661	13	96.61	96.62	0.00	0.99	0.8750	0.80

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	292	647	2095	91	76.24	76.40	0.06	0.84	0.4418	0.31
	voting	358	105	2649	26	93.23	96.19	0.00	0.99	0.8453	0.77
	mean	362	80	2674	22	94.27	97.10	0.00	0.99	0.8765	0.82
	wmean	366	58	2696	18	95.31	97.89	0.00	0.99	0.9059	0.86
	add	373	88	2666	11	97.14	96.80	0.00	0.99	0.8828	0.81
GO:0015291	spmap	366	148	2538	21	94.57	94.49	0.00	0.98	0.8124	0.71
	blast	372	104	2582	15	96.12	96.13	0.01	0.99	0.8621	0.78
	peps	318	471	2207	68	82.38	82.41	0.02	0.90	0.5413	0.40
	voting	370	106	2580	17	95.61	96.05	0.00	1.00	0.8575	0.78
	mean	374	85	2601	13	96.64	96.84	0.00	0.99	0.8842	0.81
	wmean	373	79	2607	14	96.38	97.06	0.00	0.99	0.8892	0.83
	add	372	103	2583	15	96.12	96.17	0.00	0.99	0.8631	0.78
GO:0022843	spmap	363	168	2608	23	94.04	93.95	0.00	0.98	0.7917	0.68
	blast	369	128	2648	17	95.60	95.39	0.01	0.99	0.8358	0.74
	peps	288	690	2075	97	74.81	75.05	0.11	0.81	0.4226	0.29
	voting	367	107	2669	19	95.08	96.15	0.01	0.99	0.8535	0.77
	mean	368	91	2685	18	95.34	96.72	0.00	0.99	0.8710	0.80
	wmean	369	76	2700	17	95.60	97.26	0.00	0.99	0.8881	0.83
	add	370	112	2664	16	95.85	95.97	0.00	0.99	0.8525	0.77
GO:0046982	spmap	234	712	2067	80	74.52	74.38	0.03	0.82	0.3714	0.25
	blast	259	493	2286	55	82.48	82.26	0.01	0.94	0.4859	0.34
	peps	191	1087	1686	123	60.83	60.80	0.24	0.64	0.2399	0.15
	voting	241	539	2240	73	76.75	80.60	0.14	0.92	0.4406	0.31
	mean	242	520	2259	72	77.07	81.29	0.02	0.87	0.4498	0.32
	wmean	252	438	2341	62	80.25	84.24	0.01	0.89	0.5020	0.37
	add	255	521	2258	59	81.21	81.25	0.01	0.89	0.4679	0.33
GO:0004518	spmap	281	508	2224	63	81.69	81.41	0.00	0.90	0.4960	0.36
	blast	313	249	2483	31	90.99	90.89	0.01	0.98	0.6909	0.56
	peps	243	794	1932	101	70.64	70.87	0.13	0.77	0.3519	0.23
	voting	299	291	2441	45	86.92	89.35	0.01	0.97	0.6403	0.51
	mean	301	276	2456	43	87.50	89.90	0.00	0.95	0.6536	0.52
	wmean	309	206	2526	35	89.83	92.46	0.00	0.96	0.7194	0.60
	add	314	253	2479	30	91.28	90.74	0.00	0.96	0.6894	0.55
GO:0008639	spmap	298	339	2435	42	87.65	87.78	0.00	0.94	0.6100	0.47
	blast	317	184	2590	23	93.24	93.37	0.01	0.98	0.7539	0.63
	peps	252	718	2048	88	74.12	74.04	0.10	0.79	0.3847	0.26
	voting	304	207	2567	36	89.41	92.54	0.01	0.98	0.7145	0.59
	mean	307	192	2582	33	90.29	93.08	0.00	0.97	0.7318	0.62
	wmean	309	140	2634	31	90.88	94.95	0.00	0.98	0.7833	0.69
	add	316	199	2575	24	92.94	92.83	0.00	0.98	0.7392	0.61
GO:0019787	spmap	304	289	2490	35	89.68	89.60	0.00	0.95	0.6524	0.51
	blast	319	169	2610	20	94.10	93.92	0.01	0.99	0.7715	0.65
	peps	237	837	1938	102	69.91	69.84	0.14	0.76	0.3355	0.22
	voting	310	212	2567	29	91.45	92.37	0.01	0.98	0.7201	0.59
	mean	311	189	2590	28	91.74	93.20	0.00	0.97	0.7414	0.62
	wmean	316	143	2636	23	93.22	94.85	0.00	0.98	0.7920	0.69
	add	321	145	2634	18	94.69	94.78	0.00	0.98	0.7975	0.69
GO:0016614	spmap	320	148	2616	18	94.67	94.65	0.00	0.98	0.7940	0.68
	blast	328	82	2681	10	97.04	97.03	0.00	1.00	0.8770	0.80
	peps	279	485	2276	59	82.54	82.43	0.05	0.90	0.5064	0.37
	voting	323	77	2687	15	95.56	97.21	0.00	1.00	0.8753	0.81
	mean	324	54	2710	14	95.86	98.05	0.00	1.00	0.9050	0.86
	wmean	325	50	2714	13	96.15	98.19	0.00	1.00	0.9116	0.87
	add	329	76	2687	9	97.34	97.25	0.00	0.99	0.8856	0.81
GO:0016741	spmap	284	401	2361	46	86.06	85.48	0.00	0.92	0.5596	0.41
	blast	301	239	2524	29	91.21	91.35	0.00	0.98	0.6920	0.56
	peps	222	889	1866	107	67.48	67.73	0.19	0.72	0.3083	0.20
	voting	294	274	2489	36	89.09	90.08	0.01	0.97	0.6548	0.52
	mean	294	244	2519	36	89.09	91.17	0.00	0.95	0.6774	0.55
	wmean	295	177	2586	35	89.39	93.59	0.00	0.96	0.7357	0.62

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	300	260	2503	30	90.91	90.59	0.00	0.96	0.6742	0.54
GO:0016798	spmap	272	268	2516	28	90.67	90.37	0.00	0.95	0.6476	0.50
	blast	280	183	2601	20	93.33	93.43	0.00	0.99	0.7339	0.60
	peps	228	664	2115	72	76.00	76.11	0.03	0.84	0.3826	0.26
	voting	277	156	2628	23	92.33	94.40	0.01	0.99	0.7558	0.64
	mean	281	130	2654	19	93.67	95.33	0.00	0.97	0.7904	0.68
	wmean	283	96	2688	17	94.33	96.55	0.00	0.98	0.8336	0.75
	add	284	148	2636	16	94.67	94.68	0.00	0.97	0.7760	0.66
GO:0003713	spmap	285	560	2233	71	80.06	79.95	0.02	0.88	0.4746	0.34
	blast	287	550	2243	69	80.62	80.31	0.02	0.91	0.4811	0.34
	peps	267	691	2093	89	75.00	75.18	0.11	0.81	0.4064	0.28
	voting	300	441	2352	56	84.27	84.21	0.04	0.95	0.5469	0.40
	mean	299	444	2349	57	83.99	84.10	0.02	0.91	0.5441	0.40
	wmean	303	422	2371	53	85.11	84.89	0.02	0.92	0.5606	0.42
	add	300	439	2354	56	84.27	84.28	0.01	0.91	0.5479	0.41
GO:0008289	spmap	275	480	2271	58	82.58	82.55	0.00	0.89	0.5055	0.36
	blast	291	354	2397	42	87.39	87.13	0.01	0.96	0.5951	0.45
	peps	206	1045	1703	127	61.86	61.97	0.28	0.64	0.2601	0.16
	voting	285	359	2392	48	85.59	86.95	0.04	0.95	0.5834	0.44
	mean	288	343	2408	45	86.49	87.53	0.00	0.92	0.5975	0.46
	wmean	290	281	2470	43	87.09	89.79	0.00	0.93	0.6416	0.51
	add	291	351	2400	42	87.39	87.24	0.00	0.93	0.5969	0.45
GO:0005085	spmap	292	279	2461	33	89.85	89.82	0.00	0.96	0.6518	0.51
	blast	303	191	2549	22	93.23	93.03	0.01	0.98	0.7399	0.61
	peps	247	652	2076	78	76.00	76.10	0.08	0.82	0.4036	0.27
	voting	296	178	2562	29	91.08	93.50	0.01	0.98	0.7409	0.62
	mean	297	147	2593	28	91.38	94.64	0.00	0.97	0.7724	0.67
	wmean	303	109	2631	22	93.23	96.02	0.00	0.98	0.8223	0.74
	add	306	162	2578	19	94.15	94.09	0.00	0.98	0.7718	0.65
GO:0016779	spmap	276	369	2446	41	87.07	86.89	0.00	0.93	0.5738	0.43
	blast	295	197	2618	22	93.06	93.00	0.00	0.99	0.7293	0.60
	peps	224	820	1988	93	70.66	70.80	0.17	0.75	0.3292	0.21
	voting	285	226	2589	32	89.91	91.97	0.01	0.98	0.6884	0.56
	mean	287	201	2614	30	90.54	92.86	0.00	0.96	0.7130	0.59
	wmean	292	147	2668	25	92.11	94.78	0.00	0.97	0.7725	0.67
	add	292	225	2590	25	92.11	92.01	0.00	0.97	0.7002	0.56
GO:0005083	spmap	292	331	2418	40	87.95	87.96	0.00	0.94	0.6115	0.47
	blast	314	147	2602	18	94.58	94.65	0.01	0.98	0.7919	0.68
	peps	251	675	2066	81	75.60	75.37	0.12	0.81	0.3990	0.27
	voting	304	206	2543	28	91.57	92.51	0.01	0.98	0.7221	0.60
	mean	305	164	2585	27	91.87	94.03	0.00	0.97	0.7615	0.65
	wmean	311	136	2613	21	93.67	95.05	0.00	0.98	0.7985	0.70
	add	314	152	2597	18	94.58	94.47	0.00	0.98	0.7870	0.67
GO:0008168	spmap	281	362	2386	42	87.00	86.83	0.00	0.93	0.5818	0.44
	blast	296	232	2517	27	91.64	91.56	0.01	0.98	0.6957	0.56
	peps	211	933	1801	112	65.33	65.87	0.20	0.71	0.2877	0.18
	voting	282	247	2502	41	87.31	91.01	0.01	0.98	0.6620	0.53
	mean	286	215	2534	37	88.54	92.18	0.00	0.96	0.6942	0.57
	wmean	291	147	2602	32	90.09	94.65	0.00	0.97	0.7648	0.66
	add	300	199	2550	23	92.88	92.76	0.00	0.97	0.7299	0.60
GO:0004842	spmap	282	340	2441	39	87.85	87.77	0.00	0.94	0.5981	0.45
	blast	299	194	2587	22	93.15	93.02	0.00	0.99	0.7346	0.61
	peps	237	736	2043	84	73.83	73.52	0.11	0.79	0.3663	0.24
	voting	290	221	2560	31	90.34	92.05	0.01	0.98	0.6971	0.57
	mean	292	181	2600	29	90.97	93.49	0.00	0.97	0.7355	0.62
	wmean	297	137	2644	24	92.52	95.07	0.00	0.98	0.7868	0.68
	add	301	174	2607	20	93.77	93.74	0.00	0.98	0.7563	0.63
GO:0016616	spmap	296	149	2648	15	95.18	94.67	0.00	0.99	0.7831	0.67
	blast	301	87	2710	10	96.78	96.89	0.00	1.00	0.8612	0.78
	peps	257	483	2307	53	82.90	82.69	0.05	0.89	0.4895	0.35
	voting	297	76	2721	14	95.50	97.28	0.00	1.00	0.8684	0.80

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	298	65	2732	13	95.82	97.68	0.00	0.99	0.8843	0.82
	wmean	302	52	2745	9	97.11	98.14	0.00	0.99	0.9083	0.85
	add	302	78	2719	9	97.11	97.21	0.00	0.99	0.8741	0.79
GO:0004091	spmap	274	317	2448	35	88.67	88.54	0.00	0.95	0.6089	0.46
	blast	289	182	2583	20	93.53	93.42	0.00	0.98	0.7410	0.61
	peps	225	740	2013	84	72.82	73.12	0.12	0.79	0.3532	0.23
	voting	280	176	2590	29	90.61	93.64	0.00	0.98	0.7320	0.61
	mean	281	149	2617	28	90.94	94.61	0.00	0.97	0.7605	0.65
	wmean	283	121	2645	26	91.59	95.63	0.00	0.97	0.7938	0.70
	add	286	204	2561	23	92.56	92.62	0.00	0.97	0.7159	0.58
GO:0043492	spmap	283	254	2477	29	90.71	90.70	0.00	0.96	0.6667	0.53
	blast	295	160	2571	17	94.55	94.14	0.01	0.99	0.7692	0.65
	peps	227	733	1989	85	72.76	73.07	0.08	0.81	0.3569	0.24
	voting	287	161	2570	25	91.99	94.10	0.00	0.98	0.7553	0.64
	mean	288	117	2614	24	92.31	95.72	0.00	0.97	0.8033	0.71
	wmean	291	96	2635	21	93.27	96.48	0.00	0.98	0.8326	0.75
	add	295	155	2576	17	94.55	94.32	0.00	0.98	0.7743	0.66
GO:0043565	spmap	251	427	2361	45	84.80	84.68	0.00	0.93	0.5154	0.37
	blast	245	462	2326	51	82.77	83.43	0.01	0.95	0.4885	0.35
	peps	220	707	2071	75	74.58	74.55	0.09	0.82	0.3601	0.24
	voting	255	355	2433	41	86.15	87.27	0.03	0.96	0.5629	0.42
	mean	256	315	2473	40	86.49	88.70	0.01	0.94	0.5905	0.45
	wmean	255	288	2500	41	86.15	89.67	0.00	0.95	0.6079	0.47
	add	257	370	2418	39	86.82	86.73	0.00	0.94	0.5569	0.41
GO:0017171	spmap	294	197	2559	23	92.74	92.85	0.00	0.97	0.7277	0.60
	blast	303	123	2633	13	95.89	95.54	0.00	0.99	0.8167	0.71
	peps	251	566	2182	66	79.18	79.40	0.04	0.88	0.4427	0.31
	voting	297	107	2649	20	93.69	96.12	0.01	0.99	0.8239	0.74
	mean	297	83	2673	20	93.69	96.99	0.00	0.99	0.8522	0.78
	wmean	299	68	2688	18	94.32	97.53	0.00	0.99	0.8743	0.81
	add	304	130	2626	13	95.90	95.28	0.00	0.99	0.8096	0.70
GO:0015399	spmap	273	273	2434	30	90.10	89.92	0.00	0.95	0.6431	0.50
	blast	288	151	2556	15	95.05	94.42	0.00	0.98	0.7763	0.66
	peps	221	719	1980	81	73.18	73.36	0.07	0.82	0.3559	0.24
	voting	275	163	2544	28	90.76	93.98	0.00	0.99	0.7422	0.63
	mean	275	124	2583	28	90.76	95.42	0.00	0.98	0.7835	0.69
	wmean	279	96	2611	24	92.08	96.45	0.00	0.98	0.8230	0.74
	add	288	135	2572	15	95.05	95.01	0.00	0.98	0.7934	0.68
GO:0015405	spmap	272	287	2463	31	89.77	89.56	0.00	0.95	0.6311	0.49
	blast	283	180	2570	20	93.40	93.45	0.00	0.98	0.7389	0.61
	peps	220	746	1992	83	72.61	72.75	0.09	0.80	0.3467	0.23
	voting	275	163	2587	28	90.76	94.07	0.01	0.98	0.7422	0.63
	mean	276	119	2631	27	91.09	95.67	0.00	0.97	0.7908	0.70
	wmean	278	86	2664	25	91.75	96.87	0.00	0.98	0.8336	0.76
	add	283	188	2562	20	93.40	93.16	0.00	0.97	0.7313	0.60
GO:0016820	spmap	278	223	2503	25	91.75	91.82	0.00	0.97	0.6915	0.55
	blast	289	130	2597	14	95.38	95.23	0.00	0.99	0.8006	0.69
	peps	224	706	2010	79	73.93	74.01	0.09	0.80	0.3633	0.24
	voting	282	121	2606	21	93.07	95.56	0.00	0.99	0.7989	0.70
	mean	282	100	2627	21	93.07	96.33	0.00	0.98	0.8234	0.74
	wmean	286	66	2661	17	94.39	97.58	0.00	0.98	0.8733	0.81
	add	289	124	2603	14	95.38	95.45	0.00	0.98	0.8073	0.70
GO:0016810	spmap	256	357	2380	37	87.37	86.96	0.00	0.94	0.5651	0.42
	blast	272	206	2532	21	92.83	92.48	0.01	0.98	0.7056	0.57
	peps	216	732	2000	77	73.72	73.21	0.08	0.81	0.3481	0.23
	voting	264	196	2543	29	90.10	92.84	0.01	0.99	0.7012	0.57
	mean	264	154	2585	29	90.10	94.38	0.00	0.97	0.7426	0.63
	wmean	268	112	2627	25	91.47	95.91	0.00	0.97	0.7964	0.71
	add	273	186	2552	20	93.17	93.21	0.00	0.97	0.7261	0.59
GO:0008236	spmap	290	151	2612	16	94.77	94.53	0.00	0.98	0.7764	0.66
	blast	297	95	2668	9	97.06	96.56	0.01	1.00	0.8510	0.76

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	244	562	2195	62	79.74	79.62	0.05	0.88	0.4388	0.30
	voting	293	64	2699	13	95.75	97.68	0.00	1.00	0.8839	0.82
	mean	291	57	2706	15	95.10	97.94	0.00	0.99	0.8899	0.84
	wmean	295	46	2717	11	96.41	98.34	0.00	0.99	0.9119	0.87
	add	295	107	2656	11	96.41	96.13	0.00	0.99	0.8333	0.73
GO:0016853	spmap	256	346	2467	35	87.97	87.70	0.00	0.94	0.5733	0.43
	blast	274	178	2635	17	94.16	93.67	0.00	0.99	0.7376	0.61
	peps	205	826	1977	86	70.45	70.53	0.14	0.75	0.3101	0.20
	voting	267	224	2589	24	91.75	92.04	0.01	0.99	0.6829	0.54
	mean	267	170	2643	24	91.75	93.96	0.00	0.98	0.7335	0.61
	wmean	269	115	2698	22	92.44	95.91	0.00	0.98	0.7970	0.70
	add	271	191	2622	20	93.13	93.21	0.00	0.98	0.7198	0.59
GO:0042626	spmap	266	270	2501	27	90.78	90.26	0.00	0.96	0.6417	0.50
	blast	279	144	2627	14	95.22	94.80	0.00	0.99	0.7793	0.66
	peps	215	726	2039	77	73.63	73.74	0.07	0.81	0.3487	0.23
	voting	268	135	2636	25	91.47	95.13	0.01	0.99	0.7701	0.67
	mean	274	106	2665	19	93.52	96.17	0.00	0.98	0.8143	0.72
	wmean	276	69	2702	17	94.20	97.51	0.00	0.98	0.8652	0.80
	add	282	116	2655	11	96.25	95.81	0.00	0.99	0.8162	0.71
GO:0019001	spmap	272	263	2504	29	90.37	90.50	0.00	0.96	0.6507	0.51
	blast	283	163	2604	18	94.02	94.11	0.00	0.99	0.7577	0.63
	peps	228	669	2088	73	75.75	75.73	0.06	0.81	0.3806	0.25
	voting	272	135	2632	29	90.37	95.12	0.00	0.99	0.7684	0.67
	mean	278	107	2660	23	92.36	96.13	0.00	0.97	0.8105	0.72
	wmean	279	85	2682	22	92.69	96.93	0.00	0.98	0.8391	0.77
	add	285	148	2619	16	94.68	94.65	0.00	0.98	0.7766	0.66
GO:0042165	spmap	303	118	2631	14	95.58	95.71	0.00	0.99	0.8211	0.72
	blast	310	70	2679	7	97.79	97.45	0.01	0.99	0.8895	0.82
	peps	271	397	2348	45	85.76	85.54	0.03	0.93	0.5508	0.41
	voting	306	72	2677	11	96.53	97.38	0.00	1.00	0.8806	0.81
	mean	308	59	2690	9	97.16	97.85	0.00	0.99	0.9006	0.84
	wmean	308	51	2698	9	97.16	98.14	0.00	0.99	0.9112	0.86
	add	307	86	2663	10	96.85	96.87	0.00	0.99	0.8648	0.78
GO:0043566	spmap	232	510	2248	53	81.40	81.51	0.00	0.90	0.4518	0.31
	blast	246	382	2377	39	86.32	86.15	0.01	0.95	0.5389	0.39
	peps	204	735	2013	77	72.60	73.25	0.13	0.78	0.3344	0.22
	voting	238	362	2398	47	83.51	86.88	0.02	0.95	0.5379	0.40
	mean	242	322	2438	43	84.91	88.33	0.01	0.93	0.5701	0.43
	wmean	236	292	2468	49	82.81	89.42	0.00	0.94	0.5806	0.45
	add	247	374	2386	38	86.67	86.45	0.00	0.94	0.5453	0.40
GO:0022834	spmap	287	175	2587	19	93.79	93.66	0.00	0.98	0.7474	0.62
	blast	294	106	2656	12	96.08	96.16	0.00	0.99	0.8329	0.73
	peps	231	671	2083	75	75.49	75.64	0.06	0.82	0.3825	0.26
	voting	292	104	2658	14	95.42	96.23	0.00	0.99	0.8319	0.74
	mean	292	89	2673	14	95.42	96.78	0.00	0.98	0.8501	0.77
	wmean	293	71	2691	13	95.75	97.43	0.00	0.99	0.8746	0.80
	add	295	107	2655	11	96.41	96.13	0.00	0.98	0.8333	0.73
GO:0015276	spmap	288	159	2592	18	94.12	94.22	0.00	0.99	0.7649	0.64
	blast	295	102	2649	11	96.41	96.29	0.00	0.99	0.8393	0.74
	peps	240	588	2155	66	78.43	78.56	0.05	0.84	0.4233	0.29
	voting	293	75	2676	13	95.75	97.27	0.00	1.00	0.8694	0.80
	mean	293	65	2686	13	95.75	97.64	0.00	0.99	0.8825	0.82
	wmean	295	46	2705	11	96.41	98.33	0.00	0.99	0.9119	0.87
	add	294	106	2645	12	96.08	96.15	0.00	0.99	0.8329	0.73
GO:0032561	spmap	260	269	2525	27	90.59	90.37	0.00	0.95	0.6373	0.49
	blast	272	158	2636	15	94.77	94.35	0.00	0.99	0.7587	0.63
	peps	216	689	2096	71	75.26	75.26	0.08	0.79	0.3624	0.24
	voting	261	132	2662	26	90.94	95.28	0.01	0.99	0.7676	0.66
	mean	262	102	2692	25	91.29	96.35	0.00	0.97	0.8049	0.72
	wmean	267	76	2718	20	93.03	97.28	0.00	0.98	0.8476	0.78

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	274	137	2657	13	95.47	95.10	0.00	0.98	0.7851	0.67
GO:0015631	spmap	244	352	2426	34	87.77	87.33	0.00	0.93	0.5584	0.41
	blast	246	318	2461	32	88.49	88.56	0.01	0.96	0.5843	0.44
	peps	204	716	2047	73	73.65	74.09	0.08	0.82	0.3409	0.22
	voting	245	291	2488	33	88.13	89.53	0.02	0.98	0.6020	0.46
	mean	244	272	2507	34	87.77	90.21	0.01	0.95	0.6146	0.47
	wmean	244	247	2532	34	87.77	91.11	0.00	0.96	0.6346	0.50
	add	249	301	2478	29	89.57	89.17	0.00	0.96	0.6014	0.45
GO:0005342	spmap	274	173	2582	17	94.16	93.72	0.00	0.99	0.7425	0.61
	blast	276	153	2602	15	94.85	94.45	0.01	0.99	0.7667	0.64
	peps	256	332	2414	35	87.97	87.91	0.02	0.94	0.5825	0.44
	voting	275	123	2632	16	94.50	95.54	0.01	0.99	0.7983	0.69
	mean	278	94	2661	13	95.53	96.59	0.00	0.99	0.8386	0.75
	wmean	278	89	2666	13	95.53	96.77	0.00	0.99	0.8450	0.76
	add	279	122	2633	12	95.88	95.57	0.00	0.99	0.8064	0.70
GO:0005057	spmap	254	375	2414	38	86.99	86.55	0.00	0.93	0.5516	0.40
	blast	254	361	2428	38	86.99	87.06	0.01	0.96	0.5601	0.41
	peps	219	682	2100	72	75.26	75.49	0.09	0.82	0.3674	0.24
	voting	255	248	2541	37	87.33	91.11	0.01	0.97	0.6415	0.51
	mean	254	210	2579	38	86.99	92.47	0.00	0.95	0.6720	0.55
	wmean	255	197	2592	37	87.33	92.94	0.00	0.95	0.6855	0.56
	add	260	314	2475	32	89.04	88.74	0.00	0.94	0.6005	0.45
GO:0004197	spmap	278	212	2559	23	92.36	92.35	0.00	0.97	0.7029	0.57
	blast	288	133	2638	13	95.68	95.20	0.00	0.99	0.7978	0.68
	peps	222	724	2039	79	73.75	73.80	0.09	0.79	0.3561	0.23
	voting	280	118	2653	21	93.02	95.74	0.01	0.99	0.8011	0.70
	mean	280	89	2682	21	93.02	96.79	0.00	0.98	0.8358	0.76
	wmean	282	68	2703	19	93.69	97.55	0.00	0.98	0.8664	0.81
	add	287	135	2636	14	95.35	95.13	0.00	0.98	0.7939	0.68
GO:0005525	spmap	255	234	2529	24	91.40	91.53	0.00	0.96	0.6641	0.52
	blast	267	126	2637	12	95.70	95.44	0.00	1.00	0.7946	0.68
	peps	216	618	2136	63	77.42	77.56	0.07	0.82	0.3881	0.26
	voting	261	99	2664	18	93.55	96.42	0.00	0.99	0.8169	0.72
	mean	262	83	2680	17	93.91	97.00	0.00	0.98	0.8397	0.76
	wmean	266	64	2699	13	95.34	97.68	0.00	0.98	0.8736	0.81
	add	268	117	2646	11	96.06	95.77	0.00	0.98	0.8072	0.70
GO:0046943	spmap	266	151	2621	14	95.00	94.55	0.00	0.98	0.7633	0.64
	blast	268	118	2654	12	95.71	95.74	0.00	0.99	0.8048	0.69
	peps	243	361	2397	37	86.79	86.91	0.02	0.94	0.5498	0.40
	voting	268	88	2684	12	95.71	96.83	0.01	1.00	0.8428	0.75
	mean	269	80	2692	11	96.07	97.11	0.00	0.99	0.8553	0.77
	wmean	268	76	2696	12	95.71	97.26	0.00	0.99	0.8590	0.78
	add	268	117	2655	12	95.71	95.78	0.00	0.98	0.8060	0.70
GO:0008565	spmap	239	404	2371	41	85.36	85.44	0.00	0.92	0.5179	0.37
	blast	261	189	2586	19	93.21	93.19	0.01	0.98	0.7151	0.58
	peps	221	578	2189	59	78.93	79.11	0.06	0.87	0.4096	0.28
	voting	251	201	2574	29	89.64	92.76	0.01	0.98	0.6858	0.56
	mean	251	172	2603	29	89.64	93.80	0.00	0.96	0.7141	0.59
	wmean	255	146	2629	25	91.07	94.74	0.00	0.96	0.7489	0.64
	add	259	213	2562	21	92.50	92.32	0.00	0.97	0.6888	0.55
GO:0030594	spmap	285	112	2623	11	96.28	95.90	0.00	0.99	0.8225	0.72
	blast	289	69	2666	6	97.97	97.48	0.00	1.00	0.8851	0.81
	peps	256	372	2355	40	86.49	86.36	0.03	0.93	0.5541	0.41
	voting	289	65	2670	7	97.64	97.62	0.01	1.00	0.8892	0.82
	mean	288	56	2679	8	97.30	97.95	0.00	0.99	0.9000	0.84
	wmean	289	52	2683	7	97.64	98.10	0.00	0.99	0.9074	0.85
	add	288	82	2653	8	97.30	97.00	0.00	0.99	0.8649	0.78
GO:0004553	spmap	225	171	2634	14	94.14	93.90	0.00	0.98	0.7087	0.57
	blast	225	170	2634	14	94.14	93.94	0.01	0.99	0.7098	0.57
	peps	194	519	2280	45	81.17	81.46	0.01	0.87	0.4076	0.27
	voting	226	102	2703	13	94.56	96.36	0.00	1.00	0.7972	0.69

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	227	86	2719	12	94.98	96.93	0.00	0.99	0.8225	0.73
	wmean	229	66	2739	10	95.82	97.65	0.00	0.99	0.8577	0.78
	add	229	114	2690	10	95.82	95.93	0.00	0.99	0.7869	0.67
GO:0005096	spmap	249	296	2496	30	89.25	89.40	0.00	0.94	0.6044	0.46
	blast	259	206	2586	20	92.83	92.62	0.00	0.98	0.6962	0.56
	peps	219	595	2188	60	78.49	78.62	0.07	0.85	0.4007	0.27
	voting	257	200	2592	22	92.11	92.84	0.01	0.98	0.6984	0.56
	mean	256	180	2612	23	91.76	93.55	0.00	0.97	0.7161	0.59
	wmean	256	155	2637	23	91.76	94.45	0.00	0.97	0.7420	0.62
	add	258	208	2584	21	92.47	92.55	0.00	0.97	0.6926	0.55
GO:0005006	spmap	286	35	2749	3	98.96	98.74	0.00	1.00	0.9377	0.89
	blast	288	14	2770	1	99.65	99.50	0.00	1.00	0.9746	0.95
	peps	269	169	2596	19	93.40	93.89	0.00	0.98	0.7410	0.61
	voting	286	5	2779	3	98.96	99.82	0.00	1.00	0.9862	0.98
	mean	287	3	2781	2	99.31	99.89	0.00	1.00	0.9914	0.99
	wmean	288	2	2782	1	99.65	99.93	0.00	1.00	0.9948	0.99
	add	289	6	2778	0	100.00	99.78	0.00	1.00	0.9897	0.98
GO:0008509	spmap	256	186	2557	18	93.43	93.22	0.00	0.97	0.7151	0.58
	blast	260	140	2603	14	94.89	94.90	0.00	0.99	0.7715	0.65
	peps	219	555	2180	55	79.93	79.71	0.04	0.87	0.4179	0.28
	voting	256	112	2631	18	93.43	95.92	0.01	0.99	0.7975	0.70
	mean	257	100	2643	17	93.80	96.35	0.00	0.98	0.8146	0.72
	wmean	257	82	2661	17	93.80	97.01	0.00	0.98	0.8385	0.76
	add	259	148	2595	15	94.53	94.60	0.00	0.98	0.7606	0.64
GO:0005267	spmap	257	210	2566	20	92.78	92.44	0.00	0.97	0.6909	0.55
	blast	265	120	2656	12	95.67	95.68	0.01	0.99	0.8006	0.69
	peps	216	606	2162	60	78.26	78.11	0.06	0.84	0.3934	0.26
	voting	261	91	2685	16	94.22	96.72	0.00	0.99	0.8299	0.74
	mean	260	66	2710	17	93.86	97.62	0.00	0.98	0.8624	0.80
	wmean	264	53	2723	13	95.31	98.09	0.00	0.98	0.8889	0.83
	add	267	102	2674	10	96.39	96.33	0.00	0.98	0.8266	0.72
GO:0003714	spmap	218	529	2233	52	80.74	80.85	0.01	0.89	0.4287	0.29
	blast	223	455	2307	47	82.59	83.53	0.01	0.94	0.4705	0.33
	peps	191	811	1939	79	70.74	70.51	0.13	0.75	0.3003	0.19
	voting	226	414	2348	44	83.70	85.01	0.06	0.95	0.4967	0.35
	mean	228	401	2361	42	84.44	85.48	0.02	0.92	0.5072	0.36
	wmean	228	366	2396	42	84.44	86.75	0.02	0.93	0.5278	0.38
	add	229	419	2342	41	84.81	84.82	0.01	0.92	0.4989	0.35
GO:0005200	spmap	239	264	2510	25	90.53	90.48	0.00	0.96	0.6232	0.48
	blast	241	244	2530	23	91.29	91.20	0.01	0.98	0.6435	0.50
	peps	188	787	1977	75	71.48	71.53	0.12	0.78	0.3037	0.19
	voting	238	227	2547	26	90.15	91.82	0.01	0.98	0.6529	0.51
	mean	241	194	2580	23	91.29	93.01	0.00	0.97	0.6896	0.55
	wmean	242	166	2608	22	91.67	94.02	0.00	0.98	0.7202	0.59
	add	245	200	2574	19	92.80	92.79	0.00	0.97	0.6911	0.55
GO:0008026	spmap	243	171	2625	16	93.82	93.88	0.00	0.97	0.7221	0.59
	blast	247	133	2663	12	95.37	95.24	0.01	0.99	0.7731	0.65
	peps	200	630	2157	59	77.22	77.40	0.06	0.84	0.3673	0.24
	voting	245	112	2684	14	94.59	95.99	0.01	0.99	0.7955	0.69
	mean	246	100	2696	13	94.98	96.42	0.00	0.98	0.8132	0.71
	wmean	246	79	2717	13	94.98	97.17	0.00	0.98	0.8425	0.76
	add	248	124	2672	11	95.75	95.57	0.00	0.99	0.7861	0.67
GO:0005262	spmap	247	149	2610	14	94.64	94.60	0.00	0.98	0.7519	0.62
	blast	249	125	2634	12	95.40	95.47	0.01	0.99	0.7843	0.67
	peps	187	765	1983	73	71.92	72.16	0.10	0.79	0.3086	0.20
	voting	248	123	2636	13	95.02	95.54	0.00	0.99	0.7848	0.67
	mean	251	110	2649	10	96.17	96.01	0.00	0.98	0.8071	0.70
	wmean	250	74	2685	11	95.79	97.32	0.00	0.99	0.8547	0.77
	add	250	113	2646	11	95.79	95.90	0.00	0.99	0.8013	0.69
GO:0004252	spmap	241	127	2621	11	95.63	95.38	0.00	0.98	0.7774	0.65
	blast	244	88	2660	8	96.83	96.80	0.01	0.99	0.8356	0.73

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	209	473	2266	43	82.94	82.73	0.03	0.88	0.4475	0.31
	voting	242	61	2687	10	96.03	97.78	0.00	1.00	0.8721	0.80
	mean	240	50	2698	12	95.24	98.18	0.00	0.99	0.8856	0.83
	wmean	244	39	2709	8	96.83	98.58	0.00	0.99	0.9121	0.86
	add	246	85	2663	6	97.62	96.91	0.00	0.99	0.8439	0.74
GO:0019904	spmap	193	689	2097	64	75.10	75.27	0.02	0.82	0.3389	0.22
	blast	206	566	2220	51	80.16	79.68	0.02	0.91	0.4004	0.27
	peps	158	1067	1712	99	61.48	61.60	0.24	0.66	0.2132	0.13
	voting	197	536	2250	60	76.65	80.76	0.11	0.91	0.3980	0.27
	mean	201	505	2281	56	78.21	81.87	0.03	0.86	0.4174	0.28
	wmean	204	460	2326	53	79.38	83.49	0.01	0.88	0.4430	0.31
	add	207	540	2246	50	80.54	80.62	0.01	0.87	0.4124	0.28
GO:0015077	spmap	222	334	2433	31	87.75	87.93	0.00	0.92	0.5488	0.40
	blast	225	303	2464	28	88.93	89.05	0.00	0.98	0.5762	0.43
	peps	183	735	2030	68	72.91	73.42	0.07	0.78	0.3131	0.20
	voting	220	256	2512	33	86.96	90.75	0.02	0.98	0.6036	0.46
	mean	224	232	2536	29	88.54	91.62	0.00	0.95	0.6319	0.49
	wmean	222	209	2559	31	87.75	92.45	0.00	0.96	0.6491	0.52
	add	228	278	2488	25	90.12	89.95	0.00	0.96	0.6008	0.45
GO:0004497	spmap	213	132	2651	10	95.52	95.26	0.00	0.98	0.7500	0.62
	blast	218	81	2702	5	97.76	97.09	0.00	0.99	0.8352	0.73
	peps	183	464	2310	38	82.81	83.27	0.03	0.90	0.4217	0.28
	voting	213	56	2727	10	95.52	97.99	0.00	1.00	0.8659	0.79
	mean	213	42	2741	10	95.52	98.49	0.00	1.00	0.8912	0.84
	wmean	214	35	2748	9	95.96	98.74	0.00	1.00	0.9068	0.86
	add	219	63	2720	4	98.21	97.74	0.00	0.99	0.8673	0.78
GO:0008017	spmap	205	350	2437	29	87.61	87.44	0.00	0.94	0.5196	0.37
	blast	204	369	2418	30	87.18	86.76	0.01	0.96	0.5056	0.36
	peps	172	737	2040	62	73.50	73.46	0.08	0.82	0.3010	0.19
	voting	209	304	2483	25	89.32	89.09	0.02	0.97	0.5596	0.41
	mean	209	277	2510	25	89.32	90.06	0.00	0.96	0.5806	0.43
	wmean	207	245	2542	27	88.46	91.21	0.00	0.96	0.6035	0.46
	add	207	319	2468	27	88.46	88.55	0.00	0.96	0.5447	0.39
GO:0019207	spmap	200	489	2279	42	82.64	82.33	0.00	0.90	0.4296	0.29
	blast	210	362	2406	32	86.78	86.92	0.00	0.97	0.5160	0.37
	peps	175	750	2012	65	72.92	72.85	0.12	0.77	0.3004	0.19
	voting	202	342	2426	40	83.47	87.64	0.02	0.96	0.5140	0.37
	mean	207	301	2467	35	85.54	89.13	0.01	0.94	0.5520	0.41
	wmean	204	267	2501	38	84.30	90.35	0.00	0.94	0.5722	0.43
	add	211	360	2408	31	87.19	86.99	0.00	0.93	0.5191	0.37
GO:0019900	spmap	192	575	2198	50	79.34	79.26	0.02	0.86	0.3806	0.25
	blast	203	454	2319	39	83.88	83.63	0.01	0.93	0.4516	0.31
	peps	152	1005	1765	88	63.33	63.72	0.24	0.67	0.2176	0.13
	voting	203	467	2306	39	83.88	83.16	0.12	0.94	0.4452	0.30
	mean	200	428	2345	42	82.64	84.57	0.02	0.90	0.4598	0.32
	wmean	199	365	2408	43	82.23	86.84	0.01	0.92	0.4938	0.35
	add	203	464	2309	39	83.88	83.27	0.01	0.91	0.4466	0.30
GO:0008194	spmap	195	241	2551	17	91.98	91.37	0.00	0.96	0.6019	0.45
	blast	203	135	2657	9	95.75	95.16	0.00	0.99	0.7382	0.60
	peps	181	424	2362	31	85.38	84.78	0.02	0.90	0.4431	0.30
	voting	198	106	2686	14	93.40	96.20	0.00	0.99	0.7674	0.65
	mean	197	82	2710	15	92.92	97.06	0.00	0.98	0.8024	0.71
	wmean	198	69	2723	14	93.40	97.53	0.00	0.98	0.8267	0.74
	add	203	138	2654	9	95.75	95.06	0.00	0.98	0.7342	0.60
GO:0016705	spmap	185	343	2439	26	87.68	87.67	0.00	0.94	0.5007	0.35
	blast	197	184	2598	14	93.36	93.39	0.00	0.98	0.6655	0.52
	peps	163	614	2162	47	77.62	77.88	0.06	0.84	0.3303	0.21
	voting	191	168	2614	20	90.52	93.96	0.01	0.98	0.6702	0.53
	mean	192	126	2656	19	91.00	95.47	0.00	0.97	0.7259	0.60
	wmean	192	100	2682	19	91.00	96.41	0.00	0.98	0.7634	0.66

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	196	201	2581	15	92.89	92.77	0.00	0.98	0.6447	0.49
GO:0045182	spmap	205	354	2417	30	87.23	87.22	0.00	0.93	0.5164	0.37
	blast	213	263	2508	22	90.64	90.51	0.00	0.98	0.5992	0.45
	peps	154	919	1845	79	66.09	66.75	0.17	0.74	0.2358	0.14
	voting	207	265	2506	28	88.09	90.44	0.01	0.97	0.5856	0.44
	mean	209	235	2536	26	88.94	91.52	0.00	0.96	0.6156	0.47
	wmean	212	169	2602	23	90.21	93.90	0.00	0.96	0.6883	0.56
	add	215	249	2522	20	91.49	91.01	0.00	0.97	0.6152	0.46
GO:0015078	spmap	212	324	2457	28	88.33	88.35	0.00	0.95	0.5464	0.40
	blast	217	210	2571	23	90.42	92.45	0.00	0.99	0.6507	0.51
	peps	178	723	2051	62	74.17	73.94	0.06	0.80	0.3120	0.20
	voting	218	217	2564	22	90.83	92.20	0.01	0.98	0.6459	0.50
	mean	217	190	2591	23	90.42	93.17	0.00	0.96	0.6708	0.53
	wmean	218	160	2621	22	90.83	94.25	0.00	0.97	0.7055	0.58
	add	222	215	2566	18	92.50	92.27	0.00	0.97	0.6558	0.51
GO:0030414	spmap	174	250	2525	17	91.10	90.99	0.00	0.96	0.5659	0.41
	blast	181	147	2628	10	94.76	94.70	0.00	0.99	0.6975	0.55
	peps	131	850	1919	59	68.95	69.30	0.07	0.74	0.2237	0.13
	voting	177	153	2622	14	92.67	94.49	0.00	0.98	0.6795	0.54
	mean	177	117	2658	14	92.67	95.78	0.00	0.98	0.7299	0.60
	wmean	181	80	2695	10	94.76	97.12	0.00	0.98	0.8009	0.69
	add	183	121	2654	8	95.81	95.64	0.00	0.99	0.7394	0.60
GO:0003729	spmap	204	305	2485	24	89.47	89.07	0.00	0.96	0.5536	0.40
	blast	198	311	2479	30	86.84	88.85	0.01	0.98	0.5373	0.39
	peps	181	583	2201	46	79.74	79.06	0.04	0.86	0.3653	0.24
	voting	204	211	2579	24	89.47	92.44	0.01	0.98	0.6345	0.49
	mean	210	199	2591	18	92.11	92.87	0.00	0.97	0.6593	0.51
	wmean	209	173	2617	19	91.67	93.80	0.00	0.97	0.6852	0.55
	add	210	231	2559	18	92.11	91.72	0.00	0.97	0.6278	0.48
GO:0005543	spmap	184	449	2312	36	83.64	83.74	0.00	0.92	0.4314	0.29
	blast	198	284	2477	22	90.00	89.71	0.01	0.97	0.5641	0.41
	peps	139	1008	1744	81	63.18	63.37	0.23	0.67	0.2034	0.12
	voting	192	291	2470	28	87.27	89.46	0.04	0.96	0.5462	0.40
	mean	191	260	2501	29	86.82	90.58	0.01	0.94	0.5693	0.42
	wmean	195	213	2548	25	88.64	92.29	0.00	0.95	0.6210	0.48
	add	196	315	2446	24	89.09	88.59	0.00	0.95	0.5363	0.38
GO:0019955	spmap	222	194	2579	16	93.28	93.00	0.00	0.98	0.6789	0.53
	blast	227	136	2637	11	95.38	95.10	0.00	0.99	0.7554	0.63
	peps	202	435	2330	36	84.87	84.27	0.05	0.88	0.4617	0.32
	voting	224	110	2663	14	94.12	96.03	0.00	0.99	0.7832	0.67
	mean	225	89	2684	13	94.54	96.79	0.00	0.99	0.8152	0.72
	wmean	225	79	2694	13	94.54	97.15	0.00	0.99	0.8303	0.74
	add	228	119	2654	10	95.80	95.71	0.00	0.99	0.7795	0.66
GO:0016298	spmap	185	372	2392	29	86.45	86.54	0.00	0.93	0.4799	0.33
	blast	195	252	2512	19	91.12	90.88	0.00	0.98	0.5900	0.44
	peps	159	710	2045	55	74.30	74.23	0.08	0.81	0.2936	0.18
	voting	189	209	2555	25	88.32	92.44	0.01	0.98	0.6176	0.47
	mean	190	174	2590	24	88.79	93.70	0.00	0.96	0.6574	0.52
	wmean	190	130	2634	24	88.79	95.30	0.00	0.96	0.7116	0.59
	add	195	243	2521	19	91.12	91.21	0.00	0.96	0.5982	0.45
GO:0019887	spmap	183	460	2321	36	83.56	83.46	0.00	0.91	0.4246	0.28
	blast	186	378	2403	33	84.93	86.41	0.00	0.97	0.4751	0.33
	peps	156	780	1991	62	71.56	71.85	0.10	0.76	0.2704	0.17
	voting	183	291	2490	36	83.56	89.54	0.04	0.96	0.5281	0.39
	mean	185	273	2508	34	84.47	90.18	0.00	0.94	0.5465	0.40
	wmean	185	223	2558	34	84.47	91.98	0.00	0.94	0.5901	0.45
	add	188	394	2386	31	85.84	85.83	0.00	0.93	0.4694	0.32
GO:0008083	spmap	194	344	2434	27	87.78	87.62	0.00	0.93	0.5112	0.36
	blast	200	241	2537	21	90.50	91.32	0.00	0.98	0.6042	0.45
	peps	157	756	2018	60	72.35	72.75	0.09	0.78	0.2779	0.17
	voting	195	233	2545	26	88.24	91.61	0.01	0.98	0.6009	0.46

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	197	205	2573	24	89.14	92.62	0.00	0.95	0.6324	0.49
	wmean	201	177	2601	20	90.95	93.63	0.00	0.96	0.6711	0.53
	add	205	199	2579	16	92.76	92.84	0.00	0.96	0.6560	0.51
GO:0008757	spmap	172	404	2350	30	85.15	85.33	0.00	0.92	0.4422	0.30
	blast	184	251	2503	18	91.09	90.89	0.01	0.98	0.5777	0.42
	peps	145	760	1980	57	71.78	72.26	0.12	0.78	0.2620	0.16
	voting	178	211	2543	24	88.12	92.34	0.01	0.97	0.6024	0.46
	mean	181	189	2565	21	89.60	93.14	0.00	0.95	0.6329	0.49
	wmean	183	164	2590	19	90.59	94.05	0.00	0.96	0.6667	0.53
	add	185	229	2525	17	91.58	91.68	0.00	0.95	0.6006	0.45
GO:0008081	spmap	189	288	2469	22	89.57	89.55	0.00	0.95	0.5494	0.40
	blast	195	221	2536	16	92.42	91.98	0.00	0.98	0.6220	0.47
	peps	153	740	2009	57	72.86	73.08	0.12	0.77	0.2774	0.17
	voting	189	173	2585	22	89.57	93.73	0.00	0.98	0.6597	0.52
	mean	189	134	2624	22	89.57	95.14	0.00	0.96	0.7079	0.59
	wmean	193	112	2646	18	91.47	95.94	0.00	0.97	0.7481	0.63
	add	196	201	2556	15	92.89	92.71	0.00	0.97	0.6447	0.49
GO:0003774	spmap	209	173	2619	14	93.72	93.80	0.00	0.98	0.6909	0.55
	blast	218	75	2718	5	97.76	97.31	0.01	0.99	0.8450	0.74
	peps	171	652	2136	52	76.68	76.61	0.06	0.84	0.3270	0.21
	voting	214	101	2692	9	95.96	96.38	0.00	1.00	0.7955	0.68
	mean	213	84	2709	10	95.52	96.99	0.00	0.99	0.8192	0.72
	wmean	213	60	2733	10	95.52	97.85	0.00	0.99	0.8589	0.78
	add	219	67	2726	4	98.21	97.60	0.00	0.99	0.8605	0.77
GO:0005529	spmap	167	297	2485	20	89.30	89.32	0.00	0.94	0.5131	0.36
	blast	173	205	2577	14	92.51	92.63	0.00	0.99	0.6124	0.46
	peps	137	744	2028	50	73.26	73.16	0.01	0.82	0.2566	0.16
	voting	168	202	2580	19	89.84	92.74	0.01	0.97	0.6032	0.45
	mean	167	166	2616	20	89.30	94.03	0.00	0.96	0.6423	0.50
	wmean	171	135	2647	16	91.44	95.15	0.00	0.96	0.6937	0.56
	add	172	219	2563	15	91.98	92.13	0.00	0.96	0.5952	0.44
GO:0042923	spmap	217	85	2674	6	97.31	96.92	0.00	1.00	0.8267	0.72
	blast	220	52	2707	3	98.65	98.12	0.00	1.00	0.8889	0.81
	peps	200	278	2471	23	89.69	89.89	0.02	0.95	0.5706	0.42
	voting	218	46	2713	5	97.76	98.33	0.00	1.00	0.8953	0.83
	mean	221	39	2720	2	99.10	98.59	0.00	1.00	0.9151	0.85
	wmean	221	37	2722	2	99.10	98.66	0.00	1.00	0.9189	0.86
	add	219	71	2688	4	98.21	97.43	0.00	0.99	0.8538	0.76
GO:0008188	spmap	217	72	2712	5	97.75	97.41	0.00	0.99	0.8493	0.75
	blast	220	44	2740	2	99.10	98.42	0.00	1.00	0.9053	0.83
	peps	201	265	2514	21	90.54	90.46	0.02	0.95	0.5843	0.43
	voting	219	42	2742	3	98.65	98.49	0.00	1.00	0.9068	0.84
	mean	219	43	2741	3	98.65	98.46	0.00	1.00	0.9050	0.84
	wmean	219	41	2743	3	98.65	98.53	0.00	1.00	0.9087	0.84
	add	218	71	2713	4	98.20	97.45	0.00	1.00	0.8532	0.75
GO:0004866	spmap	151	342	2403	21	87.79	87.54	0.00	0.93	0.4541	0.31
	blast	162	166	2580	10	94.19	93.95	0.00	0.98	0.6480	0.49
	peps	120	806	1934	51	70.18	70.58	0.09	0.75	0.2188	0.13
	voting	156	191	2555	16	90.70	93.04	0.03	0.98	0.6012	0.45
	mean	156	162	2584	16	90.70	94.10	0.00	0.97	0.6367	0.49
	wmean	157	120	2626	15	91.28	95.63	0.00	0.97	0.6993	0.57
	add	162	186	2560	10	94.19	93.23	0.00	0.97	0.6231	0.47
GO:0004519	spmap	162	547	2229	38	81.00	80.30	0.00	0.89	0.3564	0.23
	blast	182	246	2530	18	91.00	91.14	0.01	0.98	0.5796	0.43
	peps	145	752	2016	55	72.50	72.83	0.08	0.80	0.2644	0.16
	voting	177	274	2502	23	88.50	90.13	0.04	0.97	0.5438	0.39
	mean	177	236	2540	23	88.50	91.50	0.00	0.95	0.5775	0.43
	wmean	180	191	2585	20	90.00	93.12	0.00	0.96	0.6305	0.49
	add	185	217	2559	15	92.50	92.18	0.00	0.96	0.6146	0.46
GO:0001871	spmap	173	356	2371	26	86.93	86.95	0.00	0.93	0.4753	0.33
	blast	180	270	2457	19	90.45	90.10	0.01	0.98	0.5547	0.40

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	145	715	2006	52	73.60	73.72	0.07	0.81	0.2744	0.17
	voting	174	239	2488	25	87.44	91.24	0.02	0.97	0.5686	0.42
	mean	178	201	2526	21	89.45	92.63	0.00	0.96	0.6159	0.47
	wmean	179	176	2551	20	89.95	93.55	0.00	0.96	0.6462	0.50
	add	184	237	2490	15	92.46	91.31	0.00	0.96	0.5935	0.44
GO:0019901	spmap	165	552	2234	40	80.49	80.19	0.02	0.86	0.3579	0.23
	blast	165	553	2233	40	80.49	80.15	0.01	0.92	0.3575	0.23
	peps	125	1079	1704	79	61.27	61.23	0.26	0.65	0.1776	0.10
	voting	164	479	2307	41	80.00	82.81	0.14	0.93	0.3868	0.26
	mean	166	445	2341	39	80.98	84.03	0.02	0.89	0.4069	0.27
	wmean	166	404	2382	39	80.98	85.50	0.01	0.90	0.4284	0.29
	add	168	506	2280	37	81.95	81.84	0.01	0.89	0.3823	0.25
GO:0008135	spmap	182	340	2409	25	87.92	87.63	0.00	0.93	0.4993	0.35
	blast	197	157	2592	10	95.17	94.29	0.00	0.99	0.7023	0.56
	peps	149	779	1965	58	71.98	71.61	0.16	0.78	0.2626	0.16
	voting	192	212	2537	15	92.75	92.29	0.03	0.98	0.6285	0.48
	mean	189	172	2577	18	91.30	93.74	0.00	0.97	0.6655	0.52
	wmean	194	122	2627	13	93.72	95.56	0.00	0.97	0.7419	0.61
	add	195	161	2588	12	94.20	94.14	0.00	0.97	0.6927	0.55
GO:0005179	spmap	177	268	2508	19	90.31	90.35	0.00	0.95	0.5523	0.40
	blast	166	210	2567	30	84.69	92.44	0.00	0.98	0.5804	0.44
	peps	152	628	2142	44	77.55	77.33	0.06	0.82	0.3115	0.19
	voting	176	194	2583	20	89.80	93.01	0.02	0.98	0.6219	0.48
	mean	176	193	2584	20	89.80	93.05	0.00	0.97	0.6230	0.48
	wmean	175	179	2598	21	89.29	93.55	0.00	0.97	0.6364	0.49
	add	179	263	2512	17	91.33	90.52	0.00	0.96	0.5611	0.40
GO:0008238	spmap	187	254	2536	19	90.78	90.90	0.00	0.96	0.5781	0.42
	blast	196	132	2658	10	95.15	95.27	0.01	0.99	0.7341	0.60
	peps	171	492	2294	35	83.01	82.34	0.02	0.91	0.3936	0.26
	voting	191	99	2691	15	92.72	96.45	0.00	0.99	0.7702	0.66
	mean	192	65	2725	14	93.20	97.67	0.00	0.99	0.8294	0.75
	wmean	191	58	2732	15	92.72	97.92	0.00	0.99	0.8396	0.77
	add	196	133	2657	10	95.15	95.23	0.00	0.98	0.7327	0.60
GO:0015171	spmap	190	134	2645	9	95.48	95.18	0.00	0.99	0.7266	0.59
	blast	193	88	2691	6	96.98	96.83	0.00	0.99	0.8042	0.69
	peps	175	350	2418	24	87.94	87.36	0.02	0.94	0.4834	0.33
	voting	190	65	2714	9	95.48	97.66	0.00	1.00	0.8370	0.75
	mean	191	60	2719	8	95.98	97.84	0.00	0.99	0.8489	0.76
	wmean	191	56	2723	8	95.98	97.98	0.00	0.99	0.8565	0.77
	add	192	112	2667	7	96.48	95.97	0.00	0.99	0.7634	0.63
GO:0042625	spmap	176	244	2539	17	91.19	91.23	0.00	0.96	0.5742	0.42
	blast	181	183	2600	12	93.78	93.42	0.00	0.99	0.6499	0.50
	peps	148	644	2134	45	76.68	76.82	0.05	0.84	0.3005	0.19
	voting	180	124	2659	13	93.26	95.54	0.00	0.99	0.7243	0.59
	mean	180	93	2690	13	93.26	96.66	0.00	0.98	0.7725	0.66
	wmean	180	74	2709	13	93.26	97.34	0.00	0.98	0.8054	0.71
	add	180	186	2597	13	93.26	93.32	0.00	0.97	0.6440	0.49
GO:0008227	spmap	206	54	2748	4	98.10	98.07	0.00	0.99	0.8766	0.79
	blast	207	40	2762	3	98.57	98.57	0.00	1.00	0.9059	0.84
	peps	186	332	2465	24	88.57	88.13	0.02	0.94	0.5110	0.36
	voting	206	25	2777	4	98.10	99.11	0.00	1.00	0.9342	0.89
	mean	206	20	2782	4	98.10	99.29	0.00	1.00	0.9450	0.91
	wmean	207	14	2788	3	98.57	99.50	0.00	0.99	0.9606	0.94
	add	206	51	2751	4	98.10	98.18	0.00	0.99	0.8822	0.80
GO:0016765	spmap	163	296	2490	19	89.56	89.38	0.00	0.95	0.5086	0.36
	blast	171	165	2621	11	93.96	94.08	0.00	0.99	0.6602	0.51
	peps	143	619	2158	39	78.57	77.71	0.05	0.86	0.3030	0.19
	voting	165	138	2648	17	90.66	95.05	0.00	0.99	0.6804	0.54
	mean	169	114	2672	13	92.86	95.91	0.00	0.98	0.7269	0.60
	wmean	169	87	2699	13	92.86	96.88	0.00	0.98	0.7717	0.66

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	172	170	2616	10	94.51	93.90	0.00	0.98	0.6565	0.50
GO:0004879	spmap	180	303	2463	21	89.55	89.05	0.00	0.95	0.5263	0.37
	blast	180	254	2513	21	89.55	90.82	0.01	0.98	0.5669	0.41
	peps	162	526	2226	39	80.60	80.89	0.01	0.87	0.3645	0.24
	voting	178	148	2619	23	88.56	94.65	0.00	0.98	0.6755	0.55
	mean	179	126	2641	22	89.05	95.45	0.00	0.95	0.7075	0.59
	wmean	178	127	2640	23	88.56	95.41	0.00	0.95	0.7036	0.58
	add	182	259	2507	19	90.55	90.64	0.00	0.95	0.5670	0.41
GO:0016209	spmap	174	261	2499	18	90.62	90.54	0.00	0.96	0.5550	0.40
	blast	178	203	2557	14	92.71	92.64	0.00	0.99	0.6213	0.47
	peps	147	624	2126	44	76.96	77.31	0.07	0.83	0.3056	0.19
	voting	174	141	2619	18	90.62	94.89	0.00	0.98	0.6864	0.55
	mean	175	120	2640	17	91.15	95.65	0.00	0.97	0.7187	0.59
	wmean	177	99	2661	15	92.19	96.41	0.00	0.97	0.7564	0.64
	add	178	201	2559	14	92.71	92.72	0.00	0.97	0.6235	0.47
GO:0016407	spmap	161	521	2222	37	81.31	81.01	0.00	0.88	0.3659	0.24
	blast	171	338	2404	27	86.36	87.67	0.01	0.97	0.4837	0.34
	peps	149	680	2054	49	75.25	75.13	0.09	0.81	0.2902	0.18
	voting	166	287	2456	32	83.84	89.54	0.01	0.97	0.5100	0.37
	mean	169	280	2463	29	85.35	89.79	0.00	0.94	0.5224	0.38
	wmean	172	238	2505	26	86.87	91.32	0.00	0.94	0.5658	0.42
	add	175	323	2420	23	88.38	88.22	0.00	0.93	0.5029	0.35
GO:0004222	spmap	182	177	2628	12	93.81	93.69	0.00	0.98	0.6582	0.51
	blast	188	94	2710	6	96.91	96.65	0.00	1.00	0.7899	0.67
	peps	147	673	2122	47	75.77	75.92	0.04	0.84	0.2899	0.18
	voting	182	92	2713	12	93.81	96.72	0.00	0.99	0.7778	0.66
	mean	183	71	2734	11	94.33	97.47	0.00	0.99	0.8170	0.72
	wmean	185	49	2756	9	95.36	98.25	0.00	1.00	0.8645	0.79
	add	189	88	2716	5	97.42	96.86	0.00	1.00	0.8025	0.68
GO:0004725	spmap	188	150	2621	10	94.95	94.59	0.00	0.97	0.7015	0.56
	blast	192	89	2682	6	96.97	96.79	0.01	0.99	0.8017	0.68
	peps	156	606	2157	42	78.79	78.07	0.05	0.84	0.3250	0.20
	voting	189	73	2698	9	95.45	97.37	0.00	0.99	0.8217	0.72
	mean	189	58	2713	9	95.45	97.91	0.00	0.98	0.8494	0.77
	wmean	189	46	2725	9	95.45	98.34	0.00	0.99	0.8730	0.80
	add	193	85	2686	5	97.47	96.93	0.00	0.99	0.8109	0.69
GO:0030247	spmap	160	337	2431	22	87.91	87.83	0.00	0.93	0.4713	0.32
	blast	162	311	2457	20	89.01	88.76	0.00	0.97	0.4947	0.34
	peps	137	626	2132	42	76.54	77.30	0.04	0.85	0.2909	0.18
	voting	162	225	2543	20	89.01	91.87	0.02	0.97	0.5694	0.42
	mean	161	196	2572	21	88.46	92.92	0.00	0.95	0.5974	0.45
	wmean	162	193	2575	20	89.01	93.03	0.00	0.95	0.6034	0.46
	add	164	273	2495	18	90.11	90.14	0.00	0.95	0.5299	0.38
GO:0015293	spmap	187	108	2608	6	96.89	96.02	0.00	1.00	0.7664	0.63
	blast	190	54	2662	3	98.45	98.01	0.00	1.00	0.8696	0.78
	peps	153	537	2171	39	79.69	80.17	0.01	0.90	0.3469	0.22
	voting	188	78	2638	5	97.41	97.13	0.00	1.00	0.8192	0.71
	mean	189	54	2662	4	97.93	98.01	0.00	1.00	0.8670	0.78
	wmean	190	42	2674	3	98.45	98.45	0.00	1.00	0.8941	0.82
	add	191	42	2674	2	98.96	98.45	0.00	1.00	0.8967	0.82
GO:0032403	spmap	151	562	2217	38	79.89	79.78	0.00	0.88	0.3348	0.21
	blast	157	470	2309	32	83.07	83.09	0.01	0.95	0.3848	0.25
	peps	127	901	1870	62	67.20	67.48	0.22	0.71	0.2087	0.12
	voting	154	416	2363	35	81.48	85.03	0.02	0.94	0.4058	0.27
	mean	153	385	2394	36	80.95	86.15	0.01	0.90	0.4209	0.28
	wmean	154	332	2447	35	81.48	88.05	0.01	0.91	0.4563	0.32
	add	157	470	2309	32	83.07	83.09	0.00	0.90	0.3848	0.25
GO:0005230	spmap	180	112	2675	7	96.26	95.98	0.00	0.99	0.7516	0.62
	blast	182	88	2699	5	97.33	96.84	0.00	1.00	0.7965	0.67
	peps	156	449	2328	31	83.42	83.83	0.02	0.90	0.3939	0.26
	voting	180	48	2739	7	96.26	98.28	0.00	0.99	0.8675	0.79

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	181	38	2749	6	96.79	98.64	0.00	0.99	0.8916	0.83
	wmean	181	32	2755	6	96.79	98.85	0.00	0.99	0.9050	0.85
	add	182	77	2710	5	97.33	97.24	0.00	0.99	0.8161	0.70
GO:0016811	spmap	143	330	2446	18	88.82	88.11	0.00	0.94	0.4511	0.30
	blast	145	255	2521	16	90.06	90.81	0.00	0.98	0.5169	0.36
	peps	123	645	2124	38	76.40	76.71	0.03	0.85	0.2648	0.16
	voting	144	185	2591	17	89.44	93.34	0.00	0.98	0.5878	0.44
	mean	146	155	2621	15	90.68	94.42	0.00	0.97	0.6320	0.49
	wmean	145	133	2643	16	90.06	95.21	0.00	0.98	0.6606	0.52
	add	147	240	2536	14	91.30	91.35	0.00	0.97	0.5365	0.38
GO:0019838	spmap	168	239	2547	16	91.30	91.42	0.00	0.96	0.5685	0.41
	blast	175	148	2638	9	95.11	94.69	0.01	0.99	0.6903	0.54
	peps	149	521	2257	35	80.98	81.25	0.05	0.86	0.3489	0.22
	voting	173	126	2660	11	94.02	95.48	0.01	0.99	0.7164	0.58
	mean	173	115	2671	11	94.02	95.87	0.00	0.98	0.7331	0.60
	wmean	173	86	2700	11	94.02	96.91	0.00	0.99	0.7810	0.67
	add	174	153	2633	10	94.57	94.51	0.00	0.98	0.6810	0.53
GO:0016410	spmap	152	434	2311	28	84.44	84.19	0.00	0.92	0.3969	0.26
	blast	163	272	2473	17	90.56	90.09	0.01	0.98	0.5301	0.37
	peps	133	736	2003	47	73.89	73.13	0.12	0.80	0.2536	0.15
	voting	153	261	2484	27	85.00	90.49	0.01	0.97	0.5152	0.37
	mean	155	219	2526	25	86.11	92.02	0.00	0.95	0.5596	0.41
	wmean	157	193	2552	23	87.22	92.97	0.00	0.96	0.5925	0.45
	add	163	279	2466	17	90.56	89.84	0.00	0.95	0.5241	0.37
GO:0004540	spmap	135	570	2215	34	79.88	79.53	0.00	0.87	0.3089	0.19
	blast	155	238	2547	14	91.72	91.45	0.01	0.98	0.5516	0.39
	peps	129	657	2118	40	76.33	76.32	0.06	0.85	0.2702	0.16
	voting	146	256	2529	23	86.39	90.81	0.01	0.97	0.5114	0.36
	mean	145	199	2586	24	85.80	92.85	0.00	0.96	0.5653	0.42
	wmean	155	159	2626	14	91.72	94.29	0.00	0.97	0.6418	0.49
	add	157	226	2559	12	92.90	91.89	0.00	0.96	0.5688	0.41
GO:0015082	spmap	135	278	2508	14	90.60	90.02	0.00	0.96	0.4804	0.33
	blast	138	206	2580	11	92.62	92.61	0.00	0.99	0.5598	0.40
	peps	118	573	2201	31	79.19	79.34	0.02	0.88	0.2810	0.17
	voting	135	166	2620	14	90.60	94.04	0.01	0.99	0.6000	0.45
	mean	134	136	2650	15	89.93	95.12	0.00	0.98	0.6396	0.50
	wmean	136	111	2675	13	91.28	96.02	0.00	0.98	0.6869	0.55
	add	138	209	2577	11	92.62	92.50	0.00	0.98	0.5565	0.40
GO:0001664	spmap	154	277	2501	17	90.06	90.03	0.00	0.95	0.5116	0.36
	blast	149	238	2540	22	87.13	91.43	0.00	0.98	0.5341	0.39
	peps	125	740	2029	45	73.53	73.28	0.04	0.82	0.2415	0.14
	voting	151	226	2552	20	88.30	91.86	0.01	0.97	0.5511	0.40
	mean	152	196	2582	19	88.89	92.94	0.00	0.95	0.5857	0.44
	wmean	151	169	2609	20	88.30	93.92	0.00	0.95	0.6151	0.47
	add	153	292	2486	18	89.47	89.49	0.00	0.95	0.4968	0.34
GO:0004620	spmap	133	383	2404	21	86.36	86.26	0.00	0.92	0.3970	0.26
	blast	138	281	2506	16	89.61	89.92	0.00	0.98	0.4817	0.33
	peps	114	708	2067	40	74.03	74.49	0.08	0.80	0.2336	0.14
	voting	132	228	2559	22	85.71	91.82	0.01	0.96	0.5136	0.37
	mean	134	198	2589	20	87.01	92.90	0.00	0.94	0.5514	0.40
	wmean	134	161	2626	20	87.01	94.22	0.00	0.95	0.5969	0.45
	add	139	287	2500	15	90.26	89.70	0.00	0.95	0.4793	0.33
GO:0016502	spmap	168	67	2687	3	98.25	97.57	0.00	1.00	0.8276	0.71
	blast	168	82	2672	3	98.25	97.02	0.00	1.00	0.7981	0.67
	peps	157	239	2509	14	91.81	91.30	0.01	0.96	0.5538	0.40
	voting	169	48	2706	2	98.83	98.26	0.01	1.00	0.8711	0.78
	mean	168	41	2713	3	98.25	98.51	0.00	1.00	0.8842	0.80
	wmean	168	39	2715	3	98.25	98.58	0.00	1.00	0.8889	0.81
	add	168	84	2670	3	98.25	96.95	0.00	1.00	0.7943	0.67
GO:0003690	spmap	131	498	2285	28	82.39	82.11	0.00	0.90	0.3325	0.21
	blast	133	408	2374	26	83.65	85.33	0.00	0.96	0.3800	0.25

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	112	809	1962	47	70.44	70.80	0.15	0.75	0.2074	0.12
	voting	132	356	2428	27	83.02	87.21	0.02	0.96	0.4080	0.27
	mean	133	325	2459	26	83.65	88.33	0.00	0.94	0.4311	0.29
	wmean	132	280	2504	27	83.02	89.94	0.00	0.94	0.4623	0.32
	add	137	383	2400	22	86.16	86.24	0.00	0.92	0.4035	0.26
GO:0015144	spmap	146	142	2616	7	95.42	94.85	0.00	0.99	0.6621	0.51
	blast	145	150	2608	8	94.77	94.56	0.00	0.99	0.6473	0.49
	peps	119	589	2156	33	78.29	78.54	0.02	0.86	0.2767	0.17
	voting	148	100	2658	5	96.73	96.37	0.00	0.99	0.7382	0.60
	mean	148	75	2683	5	96.73	97.28	0.00	0.99	0.7872	0.66
	wmean	146	63	2695	7	95.42	97.72	0.00	0.99	0.8066	0.70
	add	146	141	2617	7	95.42	94.89	0.00	0.98	0.6636	0.51
GO:0005249	spmap	161	139	2637	8	95.27	94.99	0.00	0.98	0.6866	0.54
	blast	163	113	2662	6	96.45	95.93	0.00	1.00	0.7326	0.59
	peps	141	452	2310	28	83.43	83.64	0.03	0.89	0.3701	0.24
	voting	162	48	2728	7	95.86	98.27	0.00	0.99	0.8549	0.77
	mean	163	38	2738	6	96.45	98.63	0.00	0.99	0.8811	0.81
	wmean	163	33	2743	6	96.45	98.81	0.00	0.99	0.8932	0.83
	add	163	103	2672	6	96.45	96.29	0.00	0.99	0.7494	0.61
GO:0045028	spmap	169	57	2683	2	98.83	97.92	0.00	1.00	0.8514	0.75
	blast	169	63	2678	2	98.83	97.70	0.00	1.00	0.8387	0.73
	peps	155	250	2481	16	90.64	90.85	0.01	0.96	0.5382	0.38
	voting	170	43	2698	1	99.42	98.43	0.00	1.00	0.8854	0.80
	mean	170	38	2703	1	99.42	98.61	0.00	1.00	0.8971	0.82
	wmean	170	34	2707	1	99.42	98.76	0.00	1.00	0.9067	0.83
	add	169	41	2699	2	98.83	98.50	0.00	1.00	0.8871	0.80
GO:0001608	spmap	166	92	2675	5	97.08	96.68	0.00	1.00	0.7739	0.64
	blast	168	66	2701	3	98.25	97.61	0.00	1.00	0.8296	0.72
	peps	157	235	2522	14	91.81	91.48	0.01	0.96	0.5577	0.40
	voting	168	50	2717	3	98.25	98.19	0.00	1.00	0.8638	0.77
	mean	168	38	2729	3	98.25	98.63	0.00	1.00	0.8912	0.82
	wmean	168	38	2729	3	98.25	98.63	0.00	1.00	0.8912	0.82
	add	168	62	2705	3	98.25	97.76	0.00	1.00	0.8379	0.73
GO:0001614	spmap	168	51	2700	3	98.25	98.15	0.00	1.00	0.8615	0.77
	blast	171	40	2711	0	100.00	98.55	0.00	1.00	0.8953	0.81
	peps	158	202	2542	13	92.40	92.64	0.01	0.97	0.5951	0.44
	voting	170	30	2721	1	99.42	98.91	0.01	1.00	0.9164	0.85
	mean	170	31	2720	1	99.42	98.87	0.00	1.00	0.9140	0.85
	wmean	170	30	2721	1	99.42	98.91	0.00	1.00	0.9164	0.85
	add	169	43	2708	2	98.83	98.44	0.00	1.00	0.8825	0.80
GO:0016835	spmap	129	466	2304	26	83.23	83.18	0.00	0.90	0.3440	0.22
	blast	140	245	2526	15	90.32	91.16	0.00	0.99	0.5185	0.36
	peps	124	568	2197	31	80.00	79.46	0.07	0.86	0.2928	0.18
	voting	137	197	2574	18	88.39	92.89	0.00	0.97	0.5603	0.41
	mean	137	171	2600	18	88.39	93.83	0.00	0.96	0.5918	0.44
	wmean	139	150	2621	16	89.68	94.59	0.00	0.96	0.6261	0.48
	add	141	257	2513	14	90.97	90.72	0.00	0.95	0.5099	0.35
GO:0005245	spmap	155	112	2682	6	96.27	95.99	0.00	0.99	0.7243	0.58
	blast	157	87	2707	4	97.52	96.89	0.01	1.00	0.7753	0.64
	peps	117	756	2025	44	72.67	72.82	0.08	0.79	0.2263	0.13
	voting	155	79	2715	6	96.27	97.17	0.00	1.00	0.7848	0.66
	mean	155	64	2730	6	96.27	97.71	0.00	0.99	0.8158	0.71
	wmean	155	49	2745	6	96.27	98.25	0.00	1.00	0.8493	0.76
	add	157	88	2706	4	97.52	96.85	0.00	0.99	0.7734	0.64
GO:0015294	spmap	153	83	2670	3	98.08	96.99	0.00	1.00	0.7806	0.65
	blast	152	92	2661	4	97.44	96.66	0.00	0.99	0.7600	0.62
	peps	136	322	2425	19	87.74	88.28	0.01	0.95	0.4437	0.30
	voting	153	48	2705	3	98.08	98.26	0.00	1.00	0.8571	0.76
	mean	153	30	2723	3	98.08	98.91	0.00	1.00	0.9027	0.84
	wmean	153	29	2724	3	98.08	98.95	0.00	1.00	0.9053	0.84

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	153	86	2667	3	98.08	96.88	0.00	1.00	0.7747	0.64
GO:0015662	spmap	138	208	2594	11	92.62	92.58	0.00	0.96	0.5576	0.40
	blast	139	188	2614	10	93.29	93.29	0.00	0.99	0.5840	0.43
	peps	120	575	2222	29	80.54	79.44	0.02	0.87	0.2844	0.17
	voting	137	92	2710	12	91.95	96.72	0.00	0.99	0.7249	0.60
	mean	137	68	2734	12	91.95	97.57	0.00	0.98	0.7740	0.67
	wmean	137	64	2738	12	91.95	97.72	0.00	0.98	0.7829	0.68
	add	140	168	2634	9	93.96	94.00	0.00	0.97	0.6127	0.45
GO:0008066	spmap	149	182	2595	10	93.71	93.45	0.00	0.98	0.6082	0.45
	blast	156	69	2708	3	98.11	97.52	0.01	1.00	0.8125	0.69
	peps	127	530	2238	31	80.38	80.85	0.03	0.87	0.3117	0.19
	voting	152	61	2716	7	95.60	97.80	0.00	1.00	0.8172	0.71
	mean	153	40	2737	6	96.23	98.56	0.00	0.99	0.8693	0.79
	wmean	154	36	2741	5	96.86	98.70	0.00	0.99	0.8825	0.81
	add	156	63	2714	3	98.11	97.73	0.00	0.99	0.8254	0.71
GO:0005088	spmap	131	206	2563	8	94.24	92.56	0.00	0.97	0.5504	0.39
	blast	132	157	2612	7	94.96	94.33	0.00	0.99	0.6168	0.46
	peps	110	587	2171	29	79.14	78.72	0.04	0.85	0.2632	0.16
	voting	131	123	2646	8	94.24	95.56	0.01	0.99	0.6667	0.52
	mean	131	96	2673	8	94.24	96.53	0.00	0.99	0.7158	0.58
	wmean	131	89	2680	8	94.24	96.79	0.00	0.99	0.7298	0.60
	add	133	134	2635	6	95.68	95.16	0.00	0.98	0.6552	0.50
GO:0004702	spmap	145	86	2682	4	97.32	96.89	0.00	0.99	0.7632	0.63
	blast	148	48	2720	1	99.33	98.27	0.00	1.00	0.8580	0.76
	peps	123	493	2271	26	82.55	82.16	0.01	0.89	0.3216	0.20
	voting	146	41	2727	3	97.99	98.52	0.00	1.00	0.8690	0.78
	mean	147	31	2737	2	98.66	98.88	0.00	1.00	0.8991	0.83
	wmean	146	27	2741	3	97.99	99.02	0.00	1.00	0.9068	0.84
	add	148	43	2725	1	99.33	98.45	0.00	1.00	0.8706	0.77
GO:0016875	spmap	148	206	2577	12	92.50	92.60	0.00	0.99	0.5759	0.42
	blast	157	75	2708	3	98.12	97.31	0.00	1.00	0.8010	0.68
	peps	125	613	2157	35	78.12	77.87	0.07	0.83	0.2784	0.17
	voting	153	96	2687	7	95.62	96.55	0.00	1.00	0.7482	0.61
	mean	154	57	2726	6	96.25	97.95	0.00	0.99	0.8302	0.73
	wmean	156	37	2746	4	97.50	98.67	0.00	0.99	0.8839	0.81
	add	157	72	2711	3	98.12	97.41	0.00	0.99	0.8072	0.69
GO:0016876	spmap	153	145	2618	7	95.62	94.75	0.00	0.98	0.6681	0.51
	blast	155	91	2672	5	96.88	96.71	0.00	0.99	0.7635	0.63
	peps	126	601	2152	34	78.75	78.17	0.05	0.85	0.2841	0.17
	voting	153	65	2698	7	95.62	97.65	0.00	0.99	0.8095	0.70
	mean	154	45	2718	6	96.25	98.37	0.00	0.99	0.8579	0.77
	wmean	155	33	2730	5	96.88	98.81	0.00	0.99	0.8908	0.82
	add	155	90	2673	5	96.88	96.74	0.00	0.99	0.7654	0.63
GO:0004812	spmap	153	143	2614	7	95.62	94.81	0.00	0.99	0.6711	0.52
	blast	157	72	2685	3	98.12	97.39	0.00	1.00	0.8072	0.69
	peps	124	620	2132	36	77.50	77.47	0.06	0.84	0.2743	0.17
	voting	156	73	2684	4	97.50	97.35	0.00	1.00	0.8021	0.68
	mean	156	44	2713	4	97.50	98.40	0.00	0.99	0.8667	0.78
	wmean	156	32	2725	4	97.50	98.84	0.00	1.00	0.8966	0.83
	add	157	77	2680	3	98.12	97.21	0.00	1.00	0.7970	0.67
GO:0008080	spmap	132	428	2317	23	85.16	84.41	0.00	0.92	0.3692	0.24
	blast	138	311	2434	17	89.03	88.67	0.01	0.98	0.4570	0.31
	peps	113	758	1981	42	72.90	72.33	0.12	0.78	0.2203	0.13
	voting	135	295	2450	20	87.10	89.25	0.01	0.96	0.4615	0.31
	mean	136	250	2495	19	87.74	90.89	0.00	0.94	0.5028	0.35
	wmean	135	206	2539	20	87.10	92.50	0.00	0.95	0.5444	0.40
	add	140	271	2474	15	90.32	90.13	0.00	0.95	0.4947	0.34
GO:0051082	spmap	128	530	2246	30	81.01	80.91	0.00	0.90	0.3137	0.19
	blast	132	299	2477	26	83.54	89.23	0.00	0.98	0.4482	0.31
	peps	123	613	2153	35	77.85	77.84	0.04	0.85	0.2752	0.17
	voting	130	276	2500	28	82.28	90.06	0.01	0.96	0.4610	0.32

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	134	264	2512	24	84.81	90.49	0.00	0.93	0.4820	0.34
	wmean	135	233	2543	23	85.44	91.61	0.00	0.93	0.5133	0.37
	add	137	366	2410	21	86.71	86.82	0.00	0.92	0.4145	0.27
GO:0005231	spmap	148	84	2714	4	97.37	97.00	0.00	0.99	0.7708	0.64
	blast	148	76	2722	4	97.37	97.28	0.00	1.00	0.7872	0.66
	peps	124	507	2280	28	81.58	81.81	0.02	0.87	0.3167	0.20
	voting	147	45	2753	5	96.71	98.39	0.00	0.99	0.8547	0.77
	mean	147	39	2759	5	96.71	98.61	0.00	0.99	0.8698	0.79
	wmean	147	32	2766	5	96.71	98.86	0.00	0.99	0.8882	0.82
	add	148	74	2724	4	97.37	97.36	0.00	0.99	0.7914	0.67
GO:0030674	spmap	118	557	2219	30	79.73	79.94	0.01	0.86	0.2868	0.17
	blast	122	405	2370	26	82.43	85.41	0.01	0.94	0.3615	0.23
	peps	104	787	1986	42	71.23	71.62	0.12	0.75	0.2006	0.12
	voting	123	373	2403	25	83.11	86.56	0.05	0.94	0.3820	0.25
	mean	125	390	2386	23	84.46	85.95	0.01	0.91	0.3771	0.24
	wmean	124	344	2432	24	83.78	87.61	0.01	0.91	0.4026	0.26
	add	126	422	2353	22	85.14	84.79	0.01	0.91	0.3621	0.23
GO:0016684	spmap	131	259	2514	14	90.34	90.66	0.00	0.95	0.4897	0.34
	blast	137	154	2618	8	94.48	94.44	0.00	0.99	0.6284	0.47
	peps	110	668	2093	35	75.86	75.81	0.01	0.84	0.2384	0.14
	voting	134	122	2651	11	92.41	95.60	0.01	0.98	0.6683	0.52
	mean	134	86	2687	11	92.41	96.90	0.00	0.96	0.7342	0.61
	wmean	136	66	2707	9	93.79	97.62	0.00	0.97	0.7839	0.67
	add	137	154	2618	8	94.48	94.44	0.00	0.97	0.6284	0.47
GO:0004601	spmap	129	303	2472	16	88.97	89.08	0.00	0.95	0.4471	0.30
	blast	136	166	2609	9	93.79	94.02	0.00	0.98	0.6085	0.45
	peps	114	602	2166	30	79.17	78.25	0.02	0.85	0.2651	0.16
	voting	134	128	2647	11	92.41	95.39	0.00	0.98	0.6585	0.51
	mean	134	98	2677	11	92.41	96.47	0.00	0.97	0.7109	0.58
	wmean	134	79	2696	11	92.41	97.15	0.00	0.97	0.7486	0.63
	add	136	188	2587	9	93.79	93.23	0.00	0.97	0.5800	0.42
GO:0005275	spmap	135	210	2576	10	93.10	92.46	0.00	0.97	0.5510	0.39
	blast	142	88	2698	3	97.93	96.84	0.00	0.99	0.7573	0.62
	peps	127	360	2420	18	87.59	87.05	0.02	0.94	0.4019	0.26
	voting	137	114	2672	8	94.48	95.91	0.00	1.00	0.6919	0.55
	mean	139	87	2699	6	95.86	96.88	0.00	0.99	0.7493	0.62
	wmean	139	79	2707	6	95.86	97.16	0.00	0.99	0.7658	0.64
	add	140	117	2669	5	96.55	95.80	0.00	0.99	0.6965	0.54
GO:0003697	spmap	109	505	2275	24	81.95	81.83	0.00	0.90	0.2918	0.18
	blast	115	300	2480	18	86.47	89.21	0.01	0.96	0.4197	0.28
	peps	99	669	2103	32	75.57	75.87	0.08	0.81	0.2202	0.13
	voting	112	293	2487	21	84.21	89.46	0.01	0.96	0.4164	0.28
	mean	114	270	2510	19	85.71	90.29	0.00	0.93	0.4410	0.30
	wmean	117	234	2546	16	87.97	91.58	0.00	0.94	0.4835	0.33
	add	118	323	2456	15	88.72	88.38	0.00	0.94	0.4111	0.27
GO:0046906	spmap	125	372	2409	18	87.41	86.62	0.00	0.93	0.3906	0.25
	blast	131	232	2549	12	91.61	91.66	0.01	0.98	0.5178	0.36
	peps	99	856	1916	44	69.23	69.12	0.14	0.73	0.1803	0.10
	voting	127	233	2548	16	88.81	91.62	0.01	0.98	0.5050	0.35
	mean	126	190	2591	17	88.11	93.17	0.00	0.95	0.5490	0.40
	wmean	130	155	2626	13	90.91	94.43	0.00	0.96	0.6075	0.46
	add	132	216	2565	11	92.31	92.23	0.00	0.96	0.5377	0.38
GO:0020037	spmap	125	345	2415	18	87.41	87.50	0.00	0.93	0.4078	0.27
	blast	131	229	2531	12	91.61	91.70	0.01	0.98	0.5209	0.36
	peps	100	833	1917	43	69.93	69.71	0.10	0.74	0.1859	0.11
	voting	126	204	2556	17	88.11	92.61	0.01	0.97	0.5328	0.38
	mean	125	172	2588	18	87.41	93.77	0.00	0.96	0.5682	0.42
	wmean	128	144	2616	15	89.51	94.78	0.00	0.96	0.6169	0.47
	add	132	219	2541	11	92.31	92.07	0.00	0.96	0.5344	0.38
GO:0051015	spmap	100	295	2471	12	89.29	89.33	0.00	0.96	0.3945	0.25
	blast	100	286	2481	12	89.29	89.66	0.00	0.97	0.4016	0.26

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	88	532	2218	22	80.00	80.65	0.06	0.86	0.2411	0.14
	voting	100	194	2573	12	89.29	92.99	0.01	0.98	0.4926	0.34
	mean	101	157	2610	11	90.18	94.33	0.00	0.97	0.5459	0.39
	wmean	103	148	2619	9	91.96	94.65	0.00	0.97	0.5675	0.41
	add	104	204	2562	8	92.86	92.62	0.00	0.97	0.4952	0.34
GO:0005516	spmap	89	365	2427	14	86.41	86.93	0.00	0.93	0.3196	0.20
	blast	89	360	2434	14	86.41	87.12	0.01	0.96	0.3225	0.20
	peps	72	823	1959	30	70.59	70.42	0.18	0.73	0.1444	0.08
	voting	91	299	2495	12	88.35	89.30	0.04	0.98	0.3692	0.23
	mean	93	258	2536	10	90.29	90.77	0.00	0.96	0.4097	0.26
	wmean	92	223	2571	11	89.32	92.02	0.00	0.96	0.4402	0.29
	add	91	314	2478	12	88.35	88.75	0.00	0.95	0.3583	0.22
GO:0004983	spmap	145	68	2710	3	97.97	97.55	0.00	1.00	0.8033	0.68
	blast	148	27	2751	0	100.00	99.03	0.01	1.00	0.9164	0.85
	peps	135	266	2501	13	91.22	90.39	0.02	0.94	0.4918	0.34
	voting	146	31	2747	2	98.65	98.88	0.00	1.00	0.8985	0.82
	mean	146	27	2751	2	98.65	99.03	0.00	1.00	0.9097	0.84
	wmean	146	22	2756	2	98.65	99.21	0.00	1.00	0.9241	0.87
	add	148	33	2745	0	100.00	98.81	0.00	1.00	0.8997	0.82
GO:0051119	spmap	126	129	2643	6	95.45	95.35	0.00	0.98	0.6512	0.49
	blast	124	166	2606	8	93.94	94.01	0.00	0.99	0.5877	0.43
	peps	101	632	2131	30	77.10	77.13	0.01	0.86	0.2338	0.14
	voting	126	98	2674	6	95.45	96.46	0.00	0.99	0.7079	0.56
	mean	127	83	2689	5	96.21	97.01	0.00	0.98	0.7427	0.60
	wmean	127	57	2715	5	96.21	97.94	0.00	0.98	0.8038	0.69
	add	126	139	2633	6	95.45	94.99	0.00	0.97	0.6348	0.48
GO:0016251	spmap	108	647	2110	33	76.60	76.53	0.03	0.86	0.2411	0.14
	blast	104	385	2372	37	73.76	86.04	0.01	0.95	0.3302	0.21
	peps	111	586	2164	30	78.72	78.69	0.04	0.86	0.2649	0.16
	voting	115	371	2386	26	81.56	86.54	0.02	0.94	0.3668	0.24
	mean	121	395	2362	20	85.82	85.67	0.01	0.92	0.3683	0.23
	wmean	121	376	2381	20	85.82	86.36	0.01	0.92	0.3793	0.24
	add	122	406	2351	19	86.52	85.27	0.00	0.92	0.3647	0.23
GO:0003678	spmap	128	236	2538	10	92.75	91.49	0.00	0.97	0.5100	0.35
	blast	126	245	2529	12	91.30	91.17	0.00	0.98	0.4951	0.34
	peps	110	572	2196	27	80.29	79.34	0.04	0.88	0.2686	0.16
	voting	125	154	2620	13	90.58	94.45	0.01	0.99	0.5995	0.45
	mean	126	128	2646	12	91.30	95.39	0.00	0.97	0.6429	0.50
	wmean	126	112	2662	12	91.30	95.96	0.00	0.98	0.6702	0.53
	add	129	200	2574	9	93.48	92.79	0.00	0.98	0.5525	0.39
GO:0004896	spmap	139	148	2644	7	95.21	94.70	0.00	0.99	0.6420	0.48
	blast	143	56	2736	3	97.95	97.99	0.00	1.00	0.8290	0.72
	peps	125	405	2377	21	85.62	85.44	0.02	0.92	0.3698	0.24
	voting	143	52	2740	3	97.95	98.14	0.00	1.00	0.8387	0.73
	mean	143	44	2748	3	97.95	98.42	0.00	1.00	0.8589	0.76
	wmean	143	38	2754	3	97.95	98.64	0.00	1.00	0.8746	0.79
	add	143	57	2735	3	97.95	97.96	0.00	1.00	0.8266	0.71
GO:0016836	spmap	113	419	2358	20	84.96	84.91	0.00	0.93	0.3398	0.21
	blast	122	199	2578	11	91.73	92.83	0.00	0.99	0.5374	0.38
	peps	110	483	2288	23	82.71	82.57	0.04	0.88	0.3030	0.19
	voting	121	134	2643	12	90.98	95.17	0.01	0.98	0.6237	0.47
	mean	123	125	2652	10	92.48	95.50	0.00	0.98	0.6457	0.50
	wmean	122	110	2667	11	91.73	96.04	0.00	0.98	0.6685	0.53
	add	123	208	2569	10	92.48	92.51	0.00	0.97	0.5302	0.37
GO:0005507	spmap	110	392	2406	18	85.94	85.99	0.00	0.93	0.3492	0.22
	blast	114	190	2608	14	89.06	93.21	0.00	0.99	0.5278	0.38
	peps	90	820	1969	37	70.87	70.60	0.09	0.78	0.1736	0.10
	voting	110	225	2573	18	85.94	91.96	0.01	0.95	0.4752	0.33
	mean	112	197	2601	16	87.50	92.96	0.00	0.93	0.5126	0.36
	wmean	113	140	2658	15	88.28	95.00	0.00	0.94	0.5932	0.45

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	116	269	2529	12	90.62	90.39	0.00	0.95	0.4522	0.30
GO:0016903	spmap	126	102	2670	5	96.18	96.32	0.00	0.98	0.7019	0.55
	blast	128	69	2705	3	97.71	97.51	0.00	1.00	0.7805	0.65
	peps	112	421	2348	19	85.50	84.80	0.01	0.93	0.3373	0.21
	voting	126	43	2731	5	96.18	98.45	0.00	1.00	0.8400	0.75
	mean	127	23	2751	4	96.95	99.17	0.00	1.00	0.9039	0.85
	wmean	127	22	2752	4	96.95	99.21	0.00	1.00	0.9071	0.85
	add	129	63	2711	2	98.47	97.73	0.00	0.99	0.7988	0.67
GO:0015179	spmap	126	154	2596	7	94.74	94.40	0.00	0.98	0.6102	0.45
	blast	127	123	2627	6	95.49	95.53	0.00	0.99	0.6632	0.51
	peps	116	384	2359	17	87.22	86.00	0.01	0.94	0.3665	0.23
	voting	127	76	2674	6	95.49	97.24	0.00	0.99	0.7560	0.63
	mean	127	61	2689	6	95.49	97.78	0.00	0.99	0.7913	0.68
	wmean	127	53	2697	6	95.49	98.07	0.00	0.99	0.8115	0.71
	add	128	115	2635	5	96.24	95.82	0.00	0.99	0.6809	0.53
GO:0051020	spmap	99	517	2257	22	81.82	81.36	0.00	0.88	0.2687	0.16
	blast	103	414	2361	18	85.12	85.08	0.01	0.96	0.3229	0.20
	peps	86	806	1955	35	71.07	70.81	0.11	0.76	0.1698	0.10
	voting	100	364	2411	21	82.64	86.88	0.02	0.95	0.3419	0.22
	mean	104	334	2441	17	85.95	87.96	0.00	0.93	0.3721	0.24
	wmean	104	300	2475	17	85.95	89.19	0.00	0.94	0.3962	0.26
	add	106	369	2406	15	87.60	86.70	0.00	0.93	0.3557	0.22
GO:0008022	spmap	87	783	2013	34	71.90	72.00	0.04	0.79	0.1756	0.10
	blast	91	622	2174	30	75.21	77.75	0.03	0.92	0.2182	0.13
	peps	80	940	1849	41	66.12	66.30	0.19	0.72	0.1402	0.08
	voting	90	598	2198	31	74.38	78.61	0.15	0.90	0.2225	0.13
	mean	90	578	2218	31	74.38	79.33	0.03	0.85	0.2281	0.13
	wmean	92	518	2278	29	76.03	81.47	0.02	0.85	0.2517	0.15
	add	93	639	2157	28	76.86	77.15	0.02	0.84	0.2181	0.13
GO:0004722	spmap	108	331	2446	15	87.80	88.08	0.00	0.94	0.3843	0.25
	blast	114	194	2583	9	92.68	93.01	0.00	0.99	0.5290	0.37
	peps	98	568	2199	25	79.67	79.47	0.03	0.87	0.2484	0.15
	voting	111	127	2650	12	90.24	95.43	0.00	0.98	0.6150	0.47
	mean	112	103	2674	11	91.06	96.29	0.00	0.96	0.6627	0.52
	wmean	112	86	2691	11	91.06	96.90	0.00	0.96	0.6978	0.57
	add	114	198	2579	9	92.68	92.87	0.00	0.96	0.5241	0.37
GO:0016790	spmap	110	328	2408	15	88.00	88.01	0.00	0.93	0.3908	0.25
	blast	118	148	2588	7	94.40	94.59	0.00	0.99	0.6036	0.44
	peps	93	710	2015	32	74.40	73.94	0.08	0.81	0.2004	0.12
	voting	113	130	2606	12	90.40	95.25	0.00	0.98	0.6141	0.47
	mean	113	89	2647	12	90.40	96.75	0.00	0.97	0.6911	0.56
	wmean	117	65	2671	8	93.60	97.62	0.00	0.97	0.7622	0.64
	add	118	152	2584	7	94.40	94.44	0.00	0.97	0.5975	0.44
GO:0016830	spmap	107	353	2413	16	86.99	87.24	0.00	0.92	0.3671	0.23
	blast	113	176	2590	10	91.87	93.64	0.00	0.98	0.5485	0.39
	peps	98	507	2247	23	80.99	81.59	0.07	0.86	0.2700	0.16
	voting	111	118	2648	12	90.24	95.73	0.00	0.99	0.6307	0.48
	mean	113	100	2666	10	91.87	96.38	0.00	0.97	0.6726	0.53
	wmean	114	88	2678	9	92.68	96.82	0.00	0.97	0.7015	0.56
	add	114	214	2552	9	92.68	92.26	0.00	0.96	0.5055	0.35
GO:0004527	spmap	100	445	2333	19	84.03	83.98	0.00	0.89	0.3012	0.18
	blast	104	296	2482	15	87.39	89.34	0.00	0.98	0.4008	0.26
	peps	92	647	2119	27	77.31	76.61	0.08	0.83	0.2145	0.12
	voting	102	211	2567	17	85.71	92.40	0.01	0.98	0.4722	0.33
	mean	105	178	2600	14	88.24	93.59	0.00	0.96	0.5224	0.37
	wmean	105	167	2611	14	88.24	93.99	0.00	0.97	0.5371	0.39
	add	106	316	2461	13	89.08	88.62	0.00	0.96	0.3919	0.25
GO:0019783	spmap	120	193	2556	9	93.02	92.98	0.00	0.96	0.5430	0.38
	blast	125	102	2647	4	96.90	96.29	0.00	0.99	0.7022	0.55
	peps	104	539	2202	25	80.62	80.34	0.03	0.90	0.2694	0.16
	voting	121	89	2660	8	93.80	96.76	0.00	0.99	0.7139	0.58

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	123	57	2692	6	95.35	97.93	0.00	0.99	0.7961	0.68
	wmean	124	51	2698	5	96.12	98.14	0.00	0.99	0.8158	0.71
	add	124	104	2645	5	96.12	96.22	0.00	0.99	0.6947	0.54
GO:0015297	spmap	123	153	2605	6	95.35	94.45	0.00	0.98	0.6074	0.45
	blast	122	141	2617	7	94.57	94.89	0.00	0.99	0.6224	0.46
	peps	109	423	2330	20	84.50	84.63	0.01	0.92	0.3298	0.20
	voting	124	69	2690	5	96.12	97.50	0.00	0.99	0.7702	0.64
	mean	124	62	2697	5	96.12	97.75	0.00	0.99	0.7873	0.67
	wmean	124	53	2706	5	96.12	98.08	0.00	0.99	0.8105	0.70
	add	124	118	2640	5	96.12	95.72	0.00	0.99	0.6685	0.51
GO:0005539	spmap	108	358	2423	16	87.10	87.13	0.00	0.92	0.3661	0.23
	blast	108	328	2454	16	87.10	88.21	0.01	0.97	0.3857	0.25
	peps	94	631	2144	28	77.05	77.26	0.08	0.83	0.2220	0.13
	voting	106	230	2552	18	85.48	91.73	0.01	0.98	0.4609	0.32
	mean	109	191	2591	15	87.90	93.13	0.00	0.95	0.5142	0.36
	wmean	109	185	2597	15	87.90	93.35	0.00	0.95	0.5215	0.37
	add	111	307	2473	13	89.52	88.96	0.00	0.94	0.4096	0.27
GO:0000287	spmap	102	390	2360	17	85.71	85.82	0.00	0.91	0.3339	0.21
	blast	104	348	2402	15	87.39	87.35	0.00	0.97	0.3643	0.23
	peps	88	731	2013	31	73.95	73.36	0.09	0.82	0.1876	0.11
	voting	101	237	2513	18	84.87	91.38	0.02	0.96	0.4420	0.30
	mean	101	208	2542	18	84.87	92.44	0.00	0.93	0.4720	0.33
	wmean	101	197	2553	18	84.87	92.84	0.00	0.94	0.4844	0.34
	add	102	389	2361	17	85.71	85.85	0.00	0.93	0.3344	0.21
GO:0031267	spmap	94	483	2293	19	83.19	82.60	0.01	0.87	0.2725	0.16
	blast	96	405	2371	17	84.96	85.41	0.01	0.96	0.3127	0.19
	peps	82	776	1990	31	72.57	71.95	0.10	0.78	0.1689	0.10
	voting	95	340	2436	18	84.07	87.75	0.02	0.96	0.3467	0.22
	mean	94	327	2449	19	83.19	88.22	0.00	0.93	0.3521	0.22
	wmean	100	272	2504	13	88.50	90.20	0.00	0.94	0.4124	0.27
	add	98	401	2375	15	86.73	85.55	0.00	0.92	0.3203	0.20
GO:0003704	spmap	108	446	2347	20	84.38	84.03	0.01	0.92	0.3167	0.19
	blast	111	324	2469	17	86.72	88.40	0.00	0.97	0.3943	0.26
	peps	104	515	2266	24	81.25	81.48	0.05	0.86	0.2784	0.17
	voting	113	257	2536	15	88.28	90.80	0.02	0.97	0.4538	0.31
	mean	113	245	2548	15	88.28	91.23	0.01	0.96	0.4650	0.32
	wmean	115	224	2569	13	89.84	91.98	0.00	0.96	0.4925	0.34
	add	117	267	2526	11	91.41	90.44	0.00	0.95	0.4570	0.30
GO:0016651	spmap	94	486	2291	19	83.19	82.50	0.01	0.88	0.2713	0.16
	blast	95	221	2556	18	84.07	92.04	0.00	0.98	0.4429	0.30
	peps	88	626	2143	25	77.88	77.39	0.06	0.83	0.2128	0.12
	voting	96	209	2568	17	84.96	92.47	0.03	0.97	0.4593	0.31
	mean	98	231	2546	15	86.73	91.68	0.00	0.94	0.4434	0.30
	wmean	96	203	2574	17	84.96	92.69	0.00	0.93	0.4660	0.32
	add	98	383	2394	15	86.73	86.21	0.00	0.92	0.3300	0.20
GO:0035091	spmap	93	469	2291	19	83.04	83.01	0.00	0.90	0.2760	0.17
	blast	97	325	2435	15	86.61	88.22	0.01	0.98	0.3633	0.23
	peps	81	766	1987	31	72.32	72.18	0.10	0.79	0.1689	0.10
	voting	95	271	2489	17	84.82	90.18	0.03	0.97	0.3975	0.26
	mean	98	253	2507	14	87.50	90.83	0.00	0.95	0.4233	0.28
	wmean	98	208	2552	14	87.50	92.46	0.00	0.96	0.4689	0.32
	add	101	288	2472	11	90.18	89.57	0.00	0.94	0.4032	0.26
GO:0008170	spmap	84	579	2232	21	80.00	79.40	0.00	0.90	0.2188	0.13
	blast	93	299	2513	12	88.57	89.37	0.01	0.98	0.3742	0.24
	peps	70	965	1842	35	66.67	65.62	0.21	0.68	0.1228	0.07
	voting	84	368	2444	21	80.00	86.91	0.01	0.95	0.3016	0.19
	mean	89	305	2507	16	84.76	89.15	0.00	0.93	0.3567	0.23
	wmean	91	219	2593	14	86.67	92.21	0.00	0.95	0.4386	0.29
	add	95	284	2528	10	90.48	89.90	0.00	0.95	0.3926	0.25
GO:0046915	spmap	97	237	2553	8	92.38	91.51	0.00	0.97	0.4419	0.29
	blast	98	156	2634	7	93.33	94.41	0.01	0.99	0.5460	0.39

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	88	446	2340	17	83.81	83.99	0.02	0.91	0.2754	0.16
	voting	97	133	2657	8	92.38	95.23	0.00	0.99	0.5791	0.42
	mean	97	116	2674	8	92.38	95.84	0.00	0.97	0.6101	0.46
	wmean	97	106	2684	8	92.38	96.20	0.00	0.97	0.6299	0.48
	add	99	158	2632	6	94.29	94.34	0.00	0.97	0.5470	0.39
GO:0008094	spmap	90	381	2402	14	86.54	86.31	0.01	0.92	0.3130	0.19
	blast	90	366	2417	14	86.54	86.85	0.00	0.97	0.3214	0.20
	peps	87	450	2322	17	83.65	83.77	0.02	0.91	0.2715	0.16
	voting	92	209	2574	12	88.46	92.49	0.01	0.98	0.4543	0.31
	mean	91	183	2600	13	87.50	93.42	0.00	0.97	0.4815	0.33
	wmean	92	168	2615	12	88.46	93.96	0.00	0.97	0.5055	0.35
	add	94	274	2509	10	90.38	90.15	0.00	0.96	0.3983	0.26
GO:0017016	spmap	88	535	2247	21	80.73	80.77	0.01	0.88	0.2404	0.14
	blast	93	351	2433	16	85.32	87.39	0.00	0.97	0.3363	0.21
	peps	81	705	2070	28	74.31	74.59	0.13	0.80	0.1810	0.10
	voting	91	317	2467	18	83.49	88.61	0.02	0.96	0.3520	0.22
	mean	93	277	2507	16	85.32	90.05	0.00	0.93	0.3883	0.25
	wmean	94	265	2519	15	86.24	90.48	0.00	0.94	0.4017	0.26
	add	95	350	2433	14	87.16	87.42	0.00	0.92	0.3430	0.21
GO:0004428	spmap	107	175	2610	7	93.86	93.72	0.00	0.97	0.5404	0.38
	blast	111	94	2691	3	97.37	96.62	0.01	0.99	0.6959	0.54
	peps	86	690	2088	28	75.44	75.16	0.06	0.84	0.1933	0.11
	voting	109	81	2704	5	95.61	97.09	0.00	0.99	0.7171	0.57
	mean	110	49	2736	4	96.49	98.24	0.00	0.99	0.8059	0.69
	wmean	110	33	2752	4	96.49	98.82	0.00	0.98	0.8560	0.77
	add	110	105	2680	4	96.49	96.23	0.00	0.99	0.6687	0.51
GO:0008276	spmap	90	357	2416	13	87.38	87.13	0.00	0.93	0.3273	0.20
	blast	94	185	2588	9	91.26	93.33	0.01	0.99	0.4921	0.34
	peps	77	707	2053	26	74.76	74.38	0.03	0.83	0.1736	0.10
	voting	90	165	2608	13	87.38	94.05	0.00	0.98	0.5028	0.35
	mean	90	138	2635	13	87.38	95.02	0.00	0.95	0.5438	0.39
	wmean	94	112	2661	9	91.26	95.96	0.00	0.96	0.6084	0.46
	add	96	222	2551	7	93.20	91.99	0.00	0.95	0.4561	0.30
GO:0016782	spmap	117	116	2688	5	95.90	95.86	0.00	0.98	0.6592	0.50
	blast	119	84	2720	3	97.54	97.00	0.01	1.00	0.7323	0.59
	peps	108	340	2459	14	88.52	87.85	0.01	0.96	0.3789	0.24
	voting	117	30	2774	5	95.90	98.93	0.00	0.99	0.8699	0.80
	mean	117	16	2788	5	95.90	99.43	0.00	0.99	0.9176	0.88
	wmean	117	15	2789	5	95.90	99.47	0.00	0.99	0.9213	0.89
	add	119	84	2720	3	97.54	97.00	0.00	0.99	0.7323	0.59
GO:0019213	spmap	101	221	2556	8	92.66	92.04	0.00	0.97	0.4687	0.31
	blast	100	229	2548	9	91.74	91.75	0.00	0.99	0.4566	0.30
	peps	91	459	2308	18	83.49	83.41	0.02	0.93	0.2762	0.17
	voting	104	106	2671	5	95.41	96.18	0.00	1.00	0.6520	0.50
	mean	103	80	2697	6	94.50	97.12	0.00	0.99	0.7055	0.56
	wmean	104	73	2704	5	95.41	97.37	0.00	0.99	0.7273	0.59
	add	103	162	2615	6	94.50	94.17	0.00	0.99	0.5508	0.39
GO:0016627	spmap	97	322	2461	13	88.18	88.43	0.00	0.96	0.3667	0.23
	blast	103	153	2630	7	93.64	94.50	0.00	0.99	0.5628	0.40
	peps	85	650	2129	25	77.27	76.61	0.07	0.83	0.2012	0.12
	voting	101	154	2629	9	91.82	94.47	0.00	0.99	0.5534	0.40
	mean	101	130	2653	9	91.82	95.33	0.00	0.98	0.5924	0.44
	wmean	102	99	2684	8	92.73	96.44	0.00	0.98	0.6559	0.51
	add	104	150	2633	6	94.55	94.61	0.00	0.98	0.5714	0.41
GO:0004263	spmap	119	51	2725	2	98.35	98.16	0.00	1.00	0.8179	0.70
	blast	121	31	2745	0	100.00	98.88	0.01	1.00	0.8864	0.80
	peps	108	321	2447	13	89.26	88.40	0.02	0.94	0.3927	0.25
	voting	119	26	2750	2	98.35	99.06	0.00	1.00	0.8947	0.82
	mean	119	22	2754	2	98.35	99.21	0.00	1.00	0.9084	0.84
	wmean	121	18	2758	0	100.00	99.35	0.00	1.00	0.9308	0.87

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	121	18	2758	0	100.00	99.35	0.00	1.00	0.9308	0.87
GO:0003743	spmap	109	262	2510	10	91.60	90.55	0.00	0.96	0.4449	0.29
	blast	113	149	2623	6	94.96	94.62	0.00	0.99	0.5932	0.43
	peps	96	541	2225	23	80.67	80.44	0.05	0.89	0.2540	0.15
	voting	113	137	2635	6	94.96	95.06	0.00	1.00	0.6125	0.45
	mean	112	111	2661	7	94.12	96.00	0.00	0.99	0.6550	0.50
	wmean	112	94	2678	7	94.12	96.61	0.00	0.99	0.6892	0.54
	add	113	157	2615	6	94.96	94.34	0.00	0.99	0.5810	0.42
GO:0004843	spmap	109	233	2548	10	91.60	91.62	0.00	0.95	0.4729	0.32
	blast	114	137	2644	5	95.80	95.07	0.00	0.99	0.6162	0.45
	peps	96	521	2251	23	80.67	81.20	0.03	0.91	0.2609	0.16
	voting	111	121	2660	8	93.28	95.65	0.00	0.99	0.6325	0.48
	mean	111	78	2703	8	93.28	97.20	0.00	0.98	0.7208	0.59
	wmean	112	71	2710	7	94.12	97.45	0.00	0.99	0.7417	0.61
	add	114	117	2664	5	95.80	95.79	0.00	0.98	0.6514	0.49
GO:0005099	spmap	98	296	2457	11	89.91	89.25	0.00	0.94	0.3897	0.25
	blast	103	193	2560	6	94.50	92.99	0.00	0.99	0.5086	0.35
	peps	85	589	2152	24	77.98	78.51	0.06	0.87	0.2171	0.13
	voting	102	174	2579	7	93.58	93.68	0.01	0.98	0.5299	0.37
	mean	102	136	2617	7	93.58	95.06	0.00	0.97	0.5879	0.43
	wmean	102	117	2636	7	93.58	95.75	0.00	0.97	0.6220	0.47
	add	103	187	2566	6	94.50	93.21	0.00	0.98	0.5163	0.36
GO:0043176	spmap	100	357	2410	12	89.29	87.10	0.00	0.92	0.3515	0.22
	blast	99	230	2537	13	88.39	91.69	0.01	0.98	0.4490	0.30
	peps	82	734	2027	30	73.21	73.42	0.06	0.79	0.1767	0.10
	voting	101	163	2604	11	90.18	94.11	0.02	0.96	0.5372	0.38
	mean	102	151	2616	10	91.07	94.54	0.00	0.93	0.5589	0.40
	wmean	100	139	2628	12	89.29	94.98	0.00	0.93	0.5698	0.42
	add	99	320	2447	13	88.39	88.44	0.00	0.92	0.3729	0.24
GO:0003899	spmap	83	379	2430	12	87.37	86.51	0.00	0.93	0.2980	0.18
	blast	88	173	2636	7	92.63	93.84	0.00	0.99	0.4944	0.34
	peps	72	684	2115	23	75.79	75.56	0.04	0.82	0.1692	0.10
	voting	82	184	2625	13	86.32	93.45	0.00	0.97	0.4543	0.31
	mean	85	142	2667	10	89.47	94.94	0.00	0.96	0.5280	0.37
	wmean	87	104	2705	8	91.58	96.30	0.00	0.97	0.6084	0.46
	add	88	227	2582	7	92.63	91.92	0.00	0.96	0.4293	0.28
GO:0015457	spmap	58	873	1918	25	69.88	68.72	0.09	0.77	0.1144	0.06
	blast	63	314	2477	20	75.90	88.75	0.01	0.96	0.2739	0.17
	peps	50	1108	1677	33	60.24	60.22	0.23	0.62	0.0806	0.04
	voting	57	504	2287	26	68.67	81.94	0.14	0.89	0.1770	0.10
	mean	62	553	2238	21	74.70	80.19	0.03	0.83	0.1777	0.10
	wmean	62	338	2453	21	74.70	87.89	0.01	0.84	0.2567	0.15
	add	65	662	2129	18	78.31	76.28	0.02	0.83	0.1605	0.09
GO:0003823	spmap	101	271	2500	11	90.18	90.22	0.00	0.95	0.4174	0.27
	blast	102	131	2640	10	91.07	95.27	0.01	0.99	0.5913	0.44
	peps	99	335	2425	13	88.39	87.86	0.01	0.93	0.3626	0.23
	voting	101	77	2694	11	90.18	97.22	0.00	0.99	0.6966	0.57
	mean	101	62	2709	11	90.18	97.76	0.00	0.97	0.7345	0.62
	wmean	101	58	2713	11	90.18	97.91	0.00	0.97	0.7454	0.64
	add	103	225	2546	9	91.96	91.88	0.00	0.96	0.4682	0.31
GO:0004867	spmap	79	177	2612	4	95.18	93.65	0.00	0.98	0.4661	0.31
	blast	80	105	2684	3	96.39	96.24	0.00	1.00	0.5970	0.43
	peps	63	730	2055	20	75.90	73.79	0.01	0.82	0.1438	0.08
	voting	80	79	2710	3	96.39	97.17	0.00	0.99	0.6612	0.50
	mean	80	40	2749	3	96.39	98.57	0.00	0.99	0.7882	0.67
	wmean	80	32	2757	3	96.39	98.85	0.00	0.99	0.8205	0.71
	add	81	108	2681	2	97.59	96.13	0.00	0.99	0.5956	0.43
GO:0019200	spmap	107	78	2707	2	98.17	97.20	0.00	0.99	0.7279	0.58
	blast	109	16	2769	0	100.00	99.43	0.00	1.00	0.9316	0.87
	peps	91	488	2284	18	83.49	82.40	0.02	0.91	0.2645	0.16
	voting	108	10	2775	1	99.08	99.64	0.00	1.00	0.9515	0.92

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	108	9	2776	1	99.08	99.68	0.00	1.00	0.9558	0.92
	wmean	108	8	2777	1	99.08	99.71	0.00	1.00	0.9600	0.93
	add	109	14	2771	0	100.00	99.50	0.00	1.00	0.9397	0.89
GO:0004112	spmap	96	181	2611	6	94.12	93.52	0.00	0.97	0.5066	0.35
	blast	97	113	2679	5	95.10	95.95	0.00	1.00	0.6218	0.46
	peps	81	542	2244	20	80.20	80.55	0.04	0.87	0.2238	0.13
	voting	95	63	2729	7	93.14	97.74	0.00	0.99	0.7308	0.60
	mean	96	53	2739	6	94.12	98.10	0.00	0.98	0.7649	0.64
	wmean	96	43	2749	6	94.12	98.46	0.00	0.98	0.7967	0.69
	add	97	152	2640	5	95.10	94.56	0.00	0.98	0.5527	0.39
GO:0016247	spmap	59	756	2016	22	72.84	72.73	0.04	0.82	0.1317	0.07
	blast	61	343	2429	20	75.31	87.63	0.00	0.96	0.2515	0.15
	peps	47	1154	1612	34	58.02	58.28	0.22	0.62	0.0733	0.04
	voting	57	475	2297	24	70.37	82.86	0.15	0.91	0.1860	0.11
	mean	60	510	2262	21	74.07	81.60	0.04	0.86	0.1843	0.11
	wmean	61	362	2410	20	75.31	86.94	0.01	0.87	0.2421	0.14
	add	64	568	2204	17	79.01	79.51	0.00	0.86	0.1795	0.10
GO:0019208	spmap	99	304	2475	10	90.83	89.06	0.00	0.96	0.3867	0.25
	blast	94	208	2571	15	86.24	92.52	0.00	0.98	0.4574	0.31
	peps	77	832	1943	32	70.64	70.02	0.09	0.77	0.1513	0.08
	voting	97	209	2570	12	88.99	92.48	0.01	0.98	0.4675	0.32
	mean	98	195	2584	11	89.91	92.98	0.00	0.96	0.4876	0.33
	wmean	97	160	2619	12	88.99	94.24	0.00	0.97	0.5301	0.38
	add	98	283	2495	11	89.91	89.81	0.00	0.97	0.4000	0.26
GO:0015103	spmap	102	181	2605	7	93.58	93.50	0.00	0.97	0.5204	0.36
	blast	102	142	2644	7	93.58	94.90	0.00	0.98	0.5779	0.42
	peps	96	342	2438	12	88.89	87.70	0.01	0.94	0.3516	0.22
	voting	103	84	2702	6	94.50	96.98	0.00	0.99	0.6959	0.55
	mean	103	81	2705	6	94.50	97.09	0.00	0.98	0.7031	0.56
	wmean	103	70	2716	6	94.50	97.49	0.00	0.98	0.7305	0.60
	add	104	134	2652	5	95.41	95.19	0.00	0.98	0.5994	0.44
GO:0004437	spmap	97	98	2693	3	97.00	96.49	0.00	0.98	0.6576	0.50
	blast	100	30	2761	0	100.00	98.93	0.00	1.00	0.8696	0.77
	peps	82	520	2265	18	82.00	81.33	0.02	0.89	0.2336	0.14
	voting	97	22	2769	3	97.00	99.21	0.00	1.00	0.8858	0.82
	mean	98	15	2776	2	98.00	99.46	0.00	1.00	0.9202	0.87
	wmean	99	11	2780	1	99.00	99.61	0.00	1.00	0.9429	0.90
	add	99	29	2762	1	99.00	98.96	0.00	1.00	0.8684	0.77
GO:0016620	spmap	102	64	2690	2	98.08	97.68	0.00	1.00	0.7556	0.61
	blast	104	21	2733	0	100.00	99.24	0.00	1.00	0.9083	0.83
	peps	87	447	2298	17	83.65	83.72	0.00	0.91	0.2727	0.16
	voting	102	14	2740	2	98.08	99.49	0.00	1.00	0.9273	0.88
	mean	102	9	2745	2	98.08	99.67	0.00	1.00	0.9488	0.92
	wmean	103	6	2748	1	99.04	99.78	0.00	1.00	0.9671	0.94
	add	104	12	2742	0	100.00	99.56	0.00	1.00	0.9455	0.90
GO:0042562	spmap	83	379	2404	13	86.46	86.38	0.00	0.92	0.2975	0.18
	blast	82	300	2483	14	85.42	89.22	0.00	0.97	0.3431	0.21
	peps	65	895	1874	31	67.71	67.68	0.13	0.72	0.1231	0.07
	voting	83	248	2535	13	86.46	91.09	0.04	0.97	0.3888	0.25
	mean	84	243	2540	12	87.50	91.27	0.01	0.95	0.3972	0.26
	wmean	84	169	2614	12	87.50	93.93	0.00	0.95	0.4814	0.33
	add	83	381	2402	13	86.46	86.31	0.00	0.93	0.2964	0.18
GO:0031202	spmap	91	447	2333	17	84.26	83.92	0.00	0.92	0.2817	0.17
	blast	89	316	2465	19	82.41	88.64	0.01	0.97	0.3470	0.22
	peps	90	487	2287	18	83.33	82.44	0.04	0.90	0.2628	0.16
	voting	90	250	2531	18	83.33	91.01	0.01	0.97	0.4018	0.26
	mean	93	254	2527	15	86.11	90.87	0.00	0.95	0.4088	0.27
	wmean	92	237	2544	16	85.19	91.48	0.00	0.96	0.4211	0.28
	add	97	291	2490	11	89.81	89.54	0.00	0.95	0.3911	0.25
GO:0015370	spmap	103	76	2678	2	98.10	97.24	0.00	0.99	0.7254	0.58
	blast	104	41	2713	1	99.05	98.51	0.00	1.00	0.8320	0.72

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	peps	90	386	2363	15	85.71	85.96	0.00	0.95	0.3098	0.19
	voting	104	40	2714	1	99.05	98.55	0.00	1.00	0.8353	0.72
	mean	104	28	2726	1	99.05	98.98	0.00	1.00	0.8776	0.79
	wmean	104	26	2728	1	99.05	99.06	0.00	1.00	0.8851	0.80
	add	104	32	2722	1	99.05	98.84	0.00	1.00	0.8631	0.76
GO:0005089	spmap	90	111	2669	2	97.83	96.01	0.00	1.00	0.6143	0.45
	blast	89	97	2683	3	96.74	96.51	0.00	0.99	0.6403	0.48
	peps	77	496	2278	15	83.70	82.12	0.03	0.90	0.2316	0.13
	voting	89	66	2714	3	96.74	97.63	0.00	1.00	0.7206	0.57
	mean	89	50	2730	3	96.74	98.20	0.00	1.00	0.7706	0.64
	wmean	89	42	2738	3	96.74	98.49	0.00	1.00	0.7982	0.68
	add	89	97	2683	3	96.74	96.51	0.00	0.99	0.6403	0.48
GO:0005178	spmap	92	333	2437	12	88.46	87.98	0.00	0.94	0.3478	0.22
	blast	94	274	2496	9	91.26	90.11	0.01	0.97	0.3992	0.26
	peps	76	752	2007	28	73.08	72.74	0.11	0.79	0.1631	0.09
	voting	95	233	2537	9	91.35	91.59	0.02	0.97	0.4398	0.29
	mean	95	202	2568	9	91.35	92.71	0.00	0.96	0.4738	0.32
	wmean	96	171	2599	8	92.31	93.83	0.00	0.96	0.5175	0.36
	add	96	234	2536	8	92.31	91.55	0.00	0.96	0.4424	0.29
GO:0005253	spmap	87	362	2409	12	87.88	86.94	0.00	0.93	0.3175	0.19
	blast	86	187	2583	13	86.87	93.25	0.00	0.99	0.4624	0.32
	peps	82	510	2255	17	82.83	81.56	0.04	0.87	0.2373	0.14
	voting	85	125	2646	14	85.86	95.49	0.01	0.97	0.5502	0.40
	mean	86	122	2649	13	86.87	95.60	0.00	0.95	0.5603	0.41
	wmean	87	127	2644	12	87.88	95.42	0.00	0.95	0.5559	0.41
	add	90	284	2486	9	90.91	89.75	0.00	0.93	0.3805	0.24
GO:0004536	spmap	67	708	2067	23	74.44	74.49	0.03	0.82	0.1549	0.09
	blast	72	261	2514	18	80.00	90.59	0.00	0.98	0.3404	0.22
	peps	69	629	2141	21	76.67	77.29	0.06	0.85	0.1751	0.10
	voting	73	285	2490	17	81.11	89.73	0.03	0.95	0.3259	0.20
	mean	79	301	2474	11	87.78	89.15	0.00	0.93	0.3362	0.21
	wmean	74	260	2515	16	82.22	90.63	0.00	0.93	0.3491	0.22
	add	77	441	2334	13	85.56	84.11	0.00	0.91	0.2533	0.15
GO:0004295	spmap	106	34	2775	0	100.00	98.79	0.00	1.00	0.8618	0.76
	blast	106	34	2775	0	100.00	98.79	0.00	1.00	0.8618	0.76
	peps	90	424	2376	16	84.91	84.86	0.02	0.90	0.2903	0.18
	voting	106	18	2791	0	100.00	99.36	0.00	1.00	0.9217	0.85
	mean	106	15	2794	0	100.00	99.47	0.00	1.00	0.9339	0.88
	wmean	106	12	2797	0	100.00	99.57	0.00	1.00	0.9464	0.90
	add	106	10	2799	0	100.00	99.64	0.00	1.00	0.9550	0.91
GO:0005201	spmap	100	137	2609	4	96.15	95.01	0.00	0.98	0.5865	0.42
	blast	100	147	2599	4	96.15	94.65	0.00	0.99	0.5698	0.40
	peps	78	654	2082	25	75.73	76.10	0.02	0.82	0.1868	0.11
	voting	101	128	2618	3	97.12	95.34	0.01	0.99	0.6066	0.44
	mean	100	103	2643	4	96.15	96.25	0.00	0.99	0.6515	0.49
	wmean	101	81	2665	3	97.12	97.05	0.00	0.99	0.7063	0.55
	add	101	108	2638	3	97.12	96.07	0.00	0.99	0.6454	0.48
GO:0033558	spmap	85	275	2499	9	90.43	90.09	0.00	0.93	0.3744	0.24
	blast	84	172	2602	10	89.36	93.80	0.01	0.99	0.4800	0.33
	peps	78	471	2295	16	82.98	82.97	0.03	0.92	0.2426	0.14
	voting	87	96	2678	7	92.55	96.54	0.00	0.99	0.6282	0.48
	mean	87	92	2682	7	92.55	96.68	0.00	0.99	0.6374	0.49
	wmean	88	85	2689	6	93.62	96.94	0.00	0.99	0.6592	0.51
	add	88	177	2597	6	93.62	93.62	0.00	0.98	0.4903	0.33
GO:0004468	spmap	87	458	2272	17	83.65	83.22	0.00	0.92	0.2681	0.16
	blast	90	280	2450	14	86.54	89.74	0.01	0.97	0.3797	0.24
	peps	81	599	2126	23	77.88	78.02	0.04	0.86	0.2066	0.12
	voting	89	241	2489	15	85.58	91.17	0.01	0.97	0.4101	0.27
	mean	91	219	2511	13	87.50	91.98	0.00	0.95	0.4396	0.29
	wmean	91	202	2528	13	87.50	92.60	0.00	0.95	0.4584	0.31

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	add	93	306	2424	11	89.42	88.79	0.00	0.95	0.3698	0.23
GO:0004402	spmap	89	413	2335	15	85.58	84.97	0.00	0.92	0.2937	0.18
	blast	89	345	2403	15	85.58	87.45	0.00	0.98	0.3309	0.21
	peps	83	569	2172	21	79.81	79.24	0.06	0.84	0.2196	0.13
	voting	89	251	2497	15	85.58	90.87	0.01	0.96	0.4009	0.26
	mean	91	235	2513	13	87.50	91.45	0.00	0.94	0.4233	0.28
	wmean	91	206	2542	13	87.50	92.50	0.00	0.94	0.4539	0.31
	add	92	320	2428	12	88.46	88.36	0.00	0.94	0.3566	0.22
GO:0004407	spmap	87	206	2588	6	93.55	92.63	0.00	0.96	0.4508	0.30
	blast	82	224	2571	11	88.17	91.99	0.00	0.98	0.4110	0.27
	peps	79	415	2368	14	84.95	85.09	0.02	0.93	0.2692	0.16
	voting	87	99	2696	6	93.55	96.46	0.00	0.99	0.6237	0.47
	mean	87	76	2719	6	93.55	97.28	0.00	0.99	0.6797	0.53
	wmean	87	74	2721	6	93.55	97.35	0.00	0.99	0.6850	0.54
	add	88	197	2598	5	94.62	92.95	0.00	0.98	0.4656	0.31
GO:0008235	spmap	98	118	2669	4	96.08	95.77	0.00	0.98	0.6164	0.45
	blast	100	66	2721	2	98.04	97.63	0.00	1.00	0.7463	0.60
	peps	87	419	2358	15	85.29	84.91	0.01	0.94	0.2862	0.17
	voting	98	28	2759	4	96.08	99.00	0.00	1.00	0.8596	0.78
	mean	99	12	2775	3	97.06	99.57	0.00	1.00	0.9296	0.89
	wmean	99	9	2778	3	97.06	99.68	0.00	1.00	0.9429	0.92
	add	101	47	2740	1	99.02	98.31	0.00	1.00	0.8080	0.68
GO:0005254	spmap	84	332	2453	11	88.42	88.08	0.00	0.94	0.3288	0.20
	blast	84	183	2602	11	88.42	93.43	0.00	0.99	0.4641	0.31
	peps	76	579	2194	19	80.00	79.12	0.02	0.85	0.2027	0.12
	voting	83	129	2656	12	87.37	95.37	0.01	0.96	0.5407	0.39
	mean	83	123	2662	12	87.37	95.58	0.00	0.94	0.5515	0.40
	wmean	83	115	2670	12	87.37	95.87	0.00	0.94	0.5666	0.42
	add	83	354	2431	12	87.37	87.29	0.00	0.93	0.3120	0.19
GO:0060090	spmap	89	409	2368	14	86.41	85.27	0.01	0.93	0.2962	0.18
	blast	90	270	2507	13	87.38	90.28	0.01	0.98	0.3888	0.25
	peps	76	737	2031	27	73.79	73.37	0.13	0.75	0.1659	0.09
	voting	89	239	2539	14	86.41	91.40	0.01	0.97	0.4130	0.27
	mean	90	198	2580	13	87.38	92.87	0.00	0.96	0.4604	0.31
	wmean	91	178	2600	12	88.35	93.59	0.00	0.96	0.4892	0.34
	add	93	290	2486	10	90.29	89.55	0.00	0.95	0.3827	0.24
GO:0008201	spmap	82	465	2326	15	84.54	83.34	0.00	0.89	0.2547	0.15
	blast	84	285	2506	13	86.60	89.79	0.00	0.98	0.3605	0.23
	peps	74	640	2146	21	77.89	77.03	0.08	0.85	0.1829	0.10
	voting	85	239	2552	12	87.63	91.44	0.02	0.97	0.4038	0.26
	mean	87	218	2573	10	89.69	92.19	0.00	0.94	0.4328	0.29
	wmean	86	197	2594	11	88.66	92.94	0.00	0.94	0.4526	0.30
	add	86	313	2478	11	88.66	88.79	0.00	0.93	0.3468	0.22
GO:0009975	spmap	83	115	2678	3	96.51	95.88	0.00	0.98	0.5845	0.42
	blast	84	70	2723	2	97.67	97.49	0.00	1.00	0.7000	0.55
	peps	73	429	2353	13	84.88	84.58	0.01	0.89	0.2483	0.15
	voting	82	40	2753	4	95.35	98.57	0.00	0.99	0.7885	0.67
	mean	82	25	2768	4	95.35	99.10	0.00	0.99	0.8497	0.77
	wmean	83	24	2769	3	96.51	99.14	0.00	0.99	0.8601	0.78
	add	84	100	2693	2	97.67	96.42	0.00	0.99	0.6222	0.46
GO:0004521	spmap	78	539	2232	19	80.41	80.55	0.00	0.87	0.2185	0.13
	blast	85	182	2589	12	87.63	93.43	0.00	0.99	0.4670	0.32
	peps	73	711	2050	24	75.26	74.25	0.06	0.83	0.1657	0.09
	voting	79	257	2515	18	81.44	90.73	0.01	0.97	0.3649	0.24
	mean	81	215	2557	16	83.51	92.24	0.00	0.96	0.4122	0.27
	wmean	86	177	2595	11	88.66	93.61	0.00	0.96	0.4778	0.33
	add	89	272	2498	8	91.75	90.18	0.00	0.96	0.3886	0.25
GO:0016831	spmap	85	264	2517	9	90.43	90.51	0.00	0.96	0.3837	0.24
	blast	88	144	2637	6	93.62	94.82	0.00	0.98	0.5399	0.38
	peps	75	532	2240	17	81.52	80.81	0.04	0.87	0.2146	0.12
	voting	85	83	2699	9	90.43	97.02	0.00	0.99	0.6489	0.51

Table B.1 – Continued

GO Term	Method	TP	FP	TN	FN	SENS	SPEC	MedRFP	ROC	F1	PPV
	mean	87	67	2715	7	92.55	97.59	0.00	0.98	0.7016	0.56
	wmean	88	58	2724	6	93.62	97.92	0.00	0.98	0.7333	0.60
	add	89	147	2634	5	94.68	94.71	0.00	0.97	0.5394	0.38

VITA

PERSONAL INFORMATION

Surname, Name: Saraç, Ömer Sinan
Nationality: Turkish (TC)
Date and Place of Birth: 8 May 1977, Ankara
Marital Status: married, with 1 son
Phone: +90 312 210 55 41
Fax: +90 312 210 55 44
email: sarac@ceng.metu.edu.tr

EDUCATION

Degree	**Institution**	**Year**
Ph.D. in Computer Eng.	Middle East Technical University	2008
M.S. in Computer Eng.	Middle East Technical University	2002
B.S. in Computer Eng.	Middle East Technical University	1999
High School	Üsküdar Fen Lisesi	1995

WORK EXPERIENCE

Year	**Place**	**Title**
2007-2008	Sun Microsystems	Campus Ambassador
2007-2008	Bilkent Univ., Dept. of Molecular Biol. and Gen.	System Admin
1999-2008	METU Dept. of Computer Eng.	Research Assistant
1999-2006	METU Dept. of Computer Eng.	Teaching Assistant

PUBLICATIONS

1. Sarac,O.S., Yuzugullu,O.G., Cetin-Atalay,R., and Atalay,V. Subsequence-based feature map for protein function classification, *Computational Biology and Chemistry*, **32**, 122-130, 2008.

2. Sarac,O.S., Cetin-Atalay, R., and Atalay, V., GOPred: Combining classifiers on the GO, in preparation for *Proteins, Structure, Function and Bioinformatics*, 2008.

3. Temizer, S., Sarac, O.S, Isler, V., Intelligent parallel volume rendering using view coherence, *Proc. of the Int. Symposium on Computer and Information Science (ISCIS'99)*, Kusadasi, Izmir, Turkey, 1999.

4. Sarac, O.S., Gursoy-Yuzugullu, O., Cetin-Atalay, R., and Atalay, V., (2007) Protein Function Annotation by Subsequence based Feature Map, *Automated Function Prediction (AFP) and Biosapiens Special Interest Group (SIG) meeting at ISMB/ECCB 2007*, Vienna, Australia, July 19-20, 2007.

5. Sarac, O.S., Cetin-Atalay, R., Atalay, V., Protein Classification based on subsequences, *Int. Conf. on Health Informatics and Bioinformatics (HIBIT'08)*, 2008.

6. Bezek, P., Sarac, O.S, Atalay, V., Cetin-Atalay, R., Spectral clustering based subsequence fature map for protein classification, 11th *Annual Int. Conf. on Reasearch in Comp. Molecular Biol.*, San Fransicso, 2007.

7. Sarac, O.S., Atalay, V., Cetin-Atalay, R., HMM-based subsequence feature map for protein classification and remote homology detection, 14th *Annual Int. Conf. on Intelligent Systems for Molecular Biology (ISMB'06)*, Forteleza, Brasil, August, 2006.