Annotated Confirmation Report

Protein Subcellular Localization Prediction Based on Gene Ontology and SVM

The Hong Kong Polytechnic University

Protein Subcellular Localization Prediction Based on Gene Ontology and SVM

A thesis submitted in partial fulfillment of

Confirmation of Doctor of Philosophy

Department of Electronic and Information Engineering

The Hong Kong Polytechnic University

2011

ABSTRACT

Protein subcellular localization is an essential step to annotate proteins as well as to design drugs. Computational methods are required to replace the laborious and time-consuming experimental processes for fast and reliable prediction in proteomics research. This report proposes two different approaches to predicting the subcellular locations of proteins.

✓ *Clearly states research topic, importance, and key technical terms*

✓ *States research aims*

The first approach uses profile alignment scores and the occurrences of some predefined Gene Ontology (GO) terms as features and uses support vector machines (SVMs) as classifiers. The scores from the profile-alignment SVM and the GO SVM are fused to enhance classification performance. To make the best use of the GO terms, different approaches to constructing GO vectors from the GO terms returned from InterProScan were investigated. The results demonstrate that the performance of GO methods is comparable to profile-alignment methods and outperforms those based on amino-acid compositions. Also, the fusion of these two methods can outperform individual methods.

✓ *Starts with clear topic sentence which links to aims in first paragraph*

✓ *Outlines methods used in research*

The second approach uses the accession number (AN) of a query protein and the accession numbers of homologous proteins returned from PSI-BLAST as the query strings to search against the Gene Ontology Annotation (GOA) database. The occurrences of a set of predefined GO terms are used to construct the GO vectors for classification by SVMs. Again, different approaches to constructing GO vectors were investigated. Experimental results based on a recent benchmark dataset suggest that using the accession numbers of homologous proteins as the query strings can achieve an accuracy of 93.97%, which is significantly higher than all published results based on the same dataset. The accuracy can be further increased to 98.89% if the accession numbers of the query proteins are also used as query strings.

✓ *Starts with clear topic sentence, with language repeated from above, to improve flow of ideas*

✓ *Abstract outlines major findings for each of methods used*

💡 *Add why these findings are important, and what further research is needed.*

✓ *The table of contents has clear headings and sub-headings.*

💡 *Use correct page numbers, and include a list of abbreviations.*

Contents

Special Note:
The page numbers may not exactly match those in the report due to the annotations.

List of Figures

List of Tables

Special Note:
The page numbers may not exactly match those in the report due to the annotations.

Chapter 1

1 Introduction

Protein subcellular localization is one of the most essential and indispensable topics in proteomics research. Recent years have witnessed the incredibly fast development of molecular biology and computer science, which makes it possible to utilize computational methods to determine the subcellular locations of proteins. This chapter introduces the background knowledge about proteins, their subcellular locations as well as subcellular localization prediction. Different conventional methods for subcellular localization prediction are introduced, and finally our proposed methods are outlined.

✓ Clearly states research topic, its importance, and uses evaluative language, e.g. "incredibly fast development"

✓ Uses present simple tense to clearly preview contents of chapter

1.1 Proteins and Subcellular Locations

Proteins, which are essential biological macromolecules for organisms, participate in virtually every process within cells. Proteins are important in many biological processes, including metabolism catalyzing, cell signalling, immune responses, cell adhesion, and digestion. Most of the biological activities performed by proteins occur in cellular compartments, or subcellular locations. In eukaryotic cells, major subcellular locations include mitochondria, chloroplast, cytoplasm, nucleus, extracellular space, endoplasmic reticulum (ER), Golgi apparatus, peroxisome, vacuoles, cytoskeleton, nucleoplasm, lysosome and plasma membrane. Proteins can perform normal functions only if they are located in proper subcellular compartments. Moreover, the subcellular locations of proteins can have significant influence on identifying their functional characteristics, which is one of the fundamental targets in bioinformatics. Therefore, subcellular localization is one of the indispensable steps in proteomics research.

✓ Explains key terms and gives examples of their application

💡 *Lists major locations but if this is not a complete list, is such a long list necessary?*

✓ Final sentence summarises clearly and restates importance using "indispensable"

1.2 Subcellular Localization Prediction

Protein subcellular localization prediction, is to determine the cellular compartment(s) that a protein will be transported to. Traditionally, this problem is solved by purely experimental means through time-consuming and laborious laboratory tests [1]. However, the number of newly found protein sequences has been growing rapidly in the post-genomic era. Therefore, more reliable, efficient and automatic methods are highly required for the prediction of where a protein resides in a cell. The knowledge thus obtained can help biologists to use these newly discovered protein sequences for both basic biological research and drug design [2].

✓ Uses clear headings for sections 1.1 and 1.2 to show division of key concepts

✓ Clearly explains how and why research is needed

1.2.1 Conventional Prediction Methods

Over the years, a number of in-silico methods have been proposed to deal with this problem. Conventional methods can be generally divided into four categories described below.

✓ Provides a strong analysis based on clear comparisons and evaluations for each of methods discussed.

✓ Uses names of authors for most significant studies

✓ Develops background by showing how this research has built on existing knowledge. This is repeated in each sub-section.

(1) Composition-based methods are one of the earliest methods for subcellular localization prediction. This category focuses on the relationship between subcellular locations and the information embedded in the amino acid sequences such as amino-acid compositions (AA) [3],[4], amino-acid pair compositions (PairAA) [3], and gapped amino-acid pair compositions (GapAA) [5] [6]. Nakashima and Nishikawa [3] pioneered the prediction of proteins by using a simple odds-ratio statistics to discriminate between soluble intracellular and extracellular proteins based on AA and PairAA information. In the AA method, each sequence can be represented by a 20-Dimensional AA composition vector for subsequent classification. It was found that a simple odds-ratio statistics based on amino-acid composition and residue-pair frequencies can be used to discriminate between soluble intracellular and extracellular proteins. To further include the sequence-order information in the sequence vectors, PairAA [3] has also been used in the prediction. Later, Park and Kanehisa [5] used GapAA method to obtain much more sequential information. Based on these early approaches, Chou [7] proposed a method called pseudo amino-acid composition (PseAA) using a

sequence-order correlation factor to discover more biochemical properties from protein sequences.

(2) Sorting-signals based methods predict the localization via the recognition of N-terminal sorting signals in amino acid sequences [8]. These cleavable peptides contain the information about where the protein should be transported, either to the secretory pathway (in which case they are called signal peptides) or to mitochondria and chloroplast (in which they are called transit peptides). Nakai and Kanehisa in 1991 [9] proposed the earliest predictor using sorting signals —PSORT, and in 2006 they extended PSORT to WoLF PSORT [10]. PSORT is a knowledge-based program for predicting protein subcellular localization, and WoLF PSORT utilizes the information contained in sorting signals, amino acid composition and functional motifs to convert amino acid sequences into numerical features. Later, methods using signal peptides, mitochondrial targeting peptides and chloroplast transit peptides have also been proposed [11] [12]. Among these predictors, TargetP [13], which uses Hidden Markov Models (HMMs) and neural networks to learn the relationship between subcellular locations and amino acid sequences, is the most popular.

✓ Avoids overusing link words, e.g. moreover and uses reference words e.g. "they" and "these predictors" to improve flow of text

✓ Uses present simple tense to represent facts

✓ Uses present perfect tense with a non-specific time "over the years"

💡 Use present perfect tense with "recently" (also non-specific time), i.e. "has recently been proposed".

(3) Homology-based methods use the fact that homologous sequences are more likely to reside in the same subcellular location. In this group of methods, a query sequence is first used to search through a protein database for homologs [14] [15], and then the subcellular location of this query sequence is determined as the one to which the homologs belong. This kind of methods can achieve a very high accuracy as long as the homologs of the query sequences can be found in protein databases [16]. Over the years, a number of homology-based predictors have been proposed. For example, Proteome Analyst [17] computes the feature vectors for classification by using the presence or absence of some tokens from certain fields of the homologous sequences in the Swiss-Prot database. Kim et al. [18] demonstrates that feature vectors can be created by aligning an unknown protein sequence with every training sequence (with known subcellular locations). Recently, a predictor called PairProSVM was proposed by Mak et al. [19], which applies profile alignment to detect weak similarity between

protein sequences. For each query sequence, a profile can be generated by PSI-BLAST [20]. Then the obtained profile is aligned with the profile of each training sequence to form a score vector, which is classified by SVMs. It was found that profile alignment is more sensitive to detecting the weak similarity between protein families than sequence alignment.

(4) Functional-domain based methods make use of the correlation between the function of a protein and its subcellular location. Euk-OET-PLoc, proposed by Chou et al. [21], demonstrates that this category can achieve a higher performance than any other existing methods. In [22], a sequence is mapped into the GO database so that a feature vector can be formed by determining which GO terms the sequence holds. Moreover, based on deeper biological knowledge, [23] proposes a searching algorithm called GOmining to discover the informative GO terms and classify them into instructive GO terms and essential GO terms to leverage the information in the GO database. The authors also propose using BLAST [24] to retrieve homologs of the datasets to generate GO terms for those newly found proteins without known accession numbers, which made the algorithms more powerful than the previous ones.

✓ Starts with clear topic sentence

💡 Use names here because numbers 23 and 24 make the reference to "the authors" unclear. Use of "propose" and "also propose" increases confusion.

1.2.2 Comparing Conventional Methods

Among all the methods mentioned above, composition-based methods are easy to implement and have obvious biological reasoning; but in most cases these methods perform poorly, which demonstrates that amino acid sequence information is not sufficient for protein subcellular localization. Besides, sorting-signal based methods can determine the subcellular locations of proteins from the sequence segments containing the localization information, leading these methods to be more biologically plausible and robust. However, this type of methods could only deal with proteins that contain signal sequences. For example, the popular TargetP [13], [25] could only detect three locations: chloroplast, mitochondria and secretory pathway (extracellular). Homology-based

✓ Presents good comparison and analysis of previous studies, e.g. "perform poorly" and "plausible and robust"

💡 Better to use fewer link words in short paragraphs. Use more comparison links i.e. "TargetP is less able to detect location" or "Homology-based methods are better able to detect locations."

methods, on the other hand, theoretically can detect as many locations as appeared in the training data and can achieve comparatively high accuracy [26]. But when the training data contains sequences with low sequence similarity or the numbers of samples in different classes are imbalanced, the performance is still very poor. While the functional-domain based methods can often outperform sequence-based methods (as they can leverage the annotations in functional domain databases), they can only be applied to datasets where the sequences possess the required information as so far not all sequences are functionally annotated. Thus, they must be complemented by other types of methods.

✓ Identifies need for further research in the final sentence: this is the research gap

1.3 Our Proposed Methods

The methods mentioned earlier have their own advantages and disadvantages. Here, we propose two different approaches to overcoming the disadvantages.

✓ Explains how methods will be used to study research gap

The first approach fuses functional-domain based methods and homology-based methods. For the former, given a query sequence, we used InterProScan[1] to retrieve its gene ontology (GO) terms and construct a GO vector for SVM classification, and therefore we refer to the predictor as InterProGOSVM. This predictor makes use of the rich information available in various protein signature databases and the function annotations in InterPro [27]; as a result, its performance can be significantly better than those based on amino-acid compositions only. However, InterProGoSVM can only be applied to sequences that have been functionally annotated in InterPro, i.e., it is only applicable for the proteins with valid GO terms. For the homolog-based method, we align the profile of the query sequence against the profiles of a set of training sequences to form a alignment-score vector for SVM classification, and therefore we refer to the predictor as PairProSVM. This predictor can detect weak similarity between protein sequences and their remote homologs. It can be applied to all protein sequences with or without GO terms. Experimental results show that these two methods can provide strongly complementary information to each other.

✓ Explains how this study will overcome limitations of existing methods

The second proposed approach, namely GOASVM, is a functional-domain based method that makes full use of the Gene Ontology Annotation (GOA) database to predict the subcellular locations of proteins. For proteins with known accession numbers (ANs), their ANS are utilized to search the GO terms in the GOA database. While for those proteins without ANs, PSI-BLAST is used to search for the homologs and then their ANS are utilized to retrieve GO terms for further classification. In this case, for those proteins that do not have valid GO information, their homologous proteins, which are functionally annotated, are used for prediction. Thus, it is not necessary to use other methods as a backup. Experimental results show that the performance of GOASVM is significantly better than other existing methods.

The rest of this thesis is organized as follows. In Chapter 2, the procedures of constructing GO vectors from sequences using InterProScan and post-processing the raw GO vectors are detailed. The profile alignment S VM and the fusion with InterProGOSVM are explained. In Chapter 3, the GOASVM method is presented in details. In Chapter 4, we describe the experimental setup, including datasets and the performance metrics. In Chapter 5 and Chapter 6, results and analysis are presented. In Chapter 7, conclusions and the future works are presented.

✓Explains how report is organized

💡Use the passive voice, i.e. “are presented”, rather than personal “we”? Check if your supervisor has a preference.

Chapter 2

2 Functional-Domain vs. Homology-Based Methods

✓ *Clearly organises chapter 2 and 3. Each chapter groups models and explains their theoretical background*

Functional-domain based methods that use Gene Ontology (GO) and homology-based methods that use profile alignment use different information for protein subcellular localization. This chapter describes these two types of methods in detail and investigates how they can be combined to improve the prediction performance.

✓ *Gives a clear preview of content of chapter*

2.1 Functional-Domain Based Methods

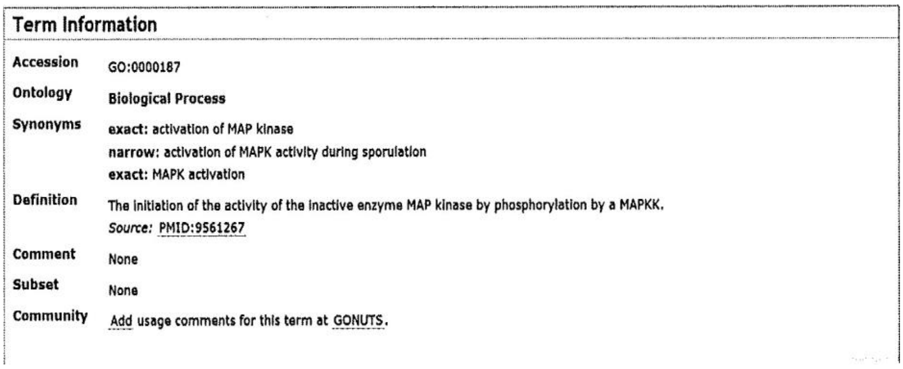
Gene Ontology (GO) [i] [28] is a set of standardized vocabularies that annotate the function of genes and gene products across different species. The term 'ontology' originally refers to a systematic account of existence. In the GO database, the annotations of gene products are organized in three related ontologies: cellular components, biological processes, and molecular functions. A cellular component is a component of a cell. It is a part of some larger objects such as an anatomical structure or a gene product group. A biological process is a sequence of events achieved by one or more ordered assemblies of molecular functions. A molecular function is achieved by activities that can be performed by individual or by assembled complexes of gene products at the molecular level. Fig. 2.1 shows an example of a GO term (GO:0000187) obtained from the GO website. As can be seen, GO:0000187 belongs to the 'Biological Process' ontology, and the specific definitions, synonyms and other related information can be also found from this GO term. This suggests that GO terms correlate with biological information of proteins and thus could be used for protein subcellular localization.

✓ *Gives clear definition of key terms related to methods discussed*

💡 *Do not define key terms with the same words, e.g. "A Cellular component is a component of a cell". Use "the parts a cell is composed of".*

✓ *Explains and justifies choice of method*

Label this as a table and refer to Table 2.1 in the paragraph, i.e. "Table 2.1 shows..."
Label tables with text above them, with the source cited below: text refers to "GO website" as the source so it must be cited.



Term Information

Accession	GO:0000187
Ontology	Biological Process
Synonyms	exact: activation of MAP kinase narrow: activation of MAPK activity during sporulation exact: MAPK activation
Definition	The initiation of the activity of the inactive enzyme MAP kinase by phosphorylation by a MAPKK. *Source:* PMID:9561267
Comment	None
Subset	None
Community	Add usage comments for this term at GONUTS.

Figure 2.1: Information of a GO term (GO:0000187).

Although the 'Cellular Component' ontology is directly related to the subcellular localization, we cannot simply use its GO terms to annotate the subcellular locations of proteins. The reason is that the percentage of proteins that have annotation of cellular components in the GO database is less than the percentage of proteins that have subcellular locations annotations in the Swiss-Prot database [29]. In fact, for those proteins that are annotated as 'Subcellular Location Unknown' in Swiss-Prot, many of them have GO terms also labelled as 'Cellular Component Unknown' in the GO database. On the other hand, proteins with subcellular locations clearly annotated in Swiss-Prot may still be marked as 'Cellular Component Unknown' in the GO database [29]. Because of this limitation, it is

✓ Makes clear comparison of three ontologies, introducing contrast with "on the other hand"

✓ Gives theoretical reasons for processes

Uses "we" to refer to work in this study which is OK, but also uses "we" refer to people in general which is less suitable, e.g. "we cannot simply use". This "we" is people in general.

necessary to make use of the other two ontologies as they are also relevant (although not directly) to the subcellular localization of proteins.

We have investigated several approaches to extracting subcellular localization information from the GO database. This is realized through a GO Processor, which consists of two parts: GO vector construction and GO vector post-processing.

✓ *Ends section with outline of following section*

2.1.1 Construction of GO Vectors

The construction of GO vectors is divided into two steps. First, a collection of distinct GO terms is obtained by presenting all of the sequences in a dataset to InterProScan. [1] For each query sequence, InterProScan returns a file containing the GO terms found by various protein-signature recognition algorithms (we used all available algorithms in this work). Using the first dataset described in Chapter 4, we found 1203 distinct GO terms, from GO:0019904 to GO:0016719. These GO terms form a GO Euclidean space with 1203 dimensions.

✓ *Gives clear first sentence to explain what is in Section 2.1.1. Links two main ideas well, e.g. "First" and "In the second step"*

In the second step, for each sequence in the dataset, we constructed a GO vector by matching its GO terms to all of the 1203 GO terms determined in the first step. We have investigated four approaches to determining the elements of the GO vectors.

1. 1-0 value. In this approach, each of the 1203 GO terms represents one canonical basis of a Euclidean space, and a protein sequence is represented by a point with coordinates equal to either 0 or 1. Specifically, the GO vector of the i-th protein is denoted as:"

✓ *Indents text to highlight technical terms*
✓ *Uses specific terms that appear in later discussion*

$$\mathbf{p}_i = \begin{bmatrix} a_{i,1} \\ \vdots \\ a_{i,j} \\ \vdots \\ a_{i,1203} \end{bmatrix} \quad \text{where } a_{i,j} = \begin{cases} 1 & \text{, GO hit} \\ 0 & \text{, otherwise} \end{cases} \qquad (2.1)$$

where 'GO hit' means that the GO term appears in the file returned from InterProScan using the i-th protein sequence as the input.

2. Term-frequency. This approach is similar to the 1-0 value approach in that a protein is represented by a point in a Euclidean space. However, unlike the 1-0 approach, it uses the number of occurrences of individual GO terms as the coordinates. Specifically, the GO vector of the i-th protein is defined as:

$$\mathbf{p}_i = \begin{bmatrix} b_{i,1} \\ \vdots \\ b_{i,j} \\ \vdots \\ b_{i,1203} \end{bmatrix} \quad \text{where } b_{i,j} = \begin{cases} f_{i,j} & \text{, GO hit} \\ 0 & \text{, otherwise} \end{cases} \qquad (2.2)$$

where fi,j is the number of occurrences of the j-th GO term (term-frequency) in the i-th protein sequence. The rationale is that the term-frequencies may also contain important information for classification and therefore should not be quantized to either 0 or 1. Note that bi,j's are analogous to the term-frequencies commonly used in document retrieval [30].

Fig. 2.2 illustrates how a GO vector is constructed from the obtained GO terms for each protein sequence. Suppose there are 3 sequences in the dataset. First, by using InterProScan, we obtain the GO terms for the 3 sequences as shown in the figure, There are 5 distinct GO terms among the 3 sequences. Then, we form a 5-dim GO Euclidean space (Note that the element order should be the same for all sequences). Next, we count the frequency of occurrences for each GO term in each sequence to form GO vectors PI, P2 and P3.

💡 *Indicate where Figure 2.2 is if it is not on the same page, i.e. "Figure 2.2, on the previous page, illustrates..."*

✓ *Refers to figures using wide range of language, e.g. "shows", "illustrates", "is shown"*

3. Inverse Sequence-Frequency (ISF). In this approach, a protein is represented by a point with coordinates determined by the existence of GO terms and the inverse sequence-frequency (ISP). Specifically, the GO vector Pi of the i-th protein is defined as:

$$\mathbf{p}_i = \begin{bmatrix} c_{i,1} \\ \vdots \\ c_{i,j} \\ \vdots \\ c_{i,1203} \end{bmatrix}, \; c_{i,j} = a_{i,j} \log \left(\frac{N}{|\{k : a_{k,j} \neq 0\}|} \right) \tag{2.3}$$

where N is the number of protein sequences in the dataset. The denominator in Eq. 2.3 is the number of GO vectors (among all GO vectors in the dataset) having a non-zero entry in their j-th element, or equivalently the number of sequences with the j-th GO term as determined by InterProScan.

Term-Frequency (TF): seql AN: 088978 GO:0005737 GO:0005737 seq 2 AN: 912945 GO:0005515 GO:0005515 GO:0005515 GO:0005488 GO:0005515 GO:00055i5 seq3 AN:Q7M359 GO:0005737 GO:0005737 GO:0005737 GO:0008270 GO:0046872

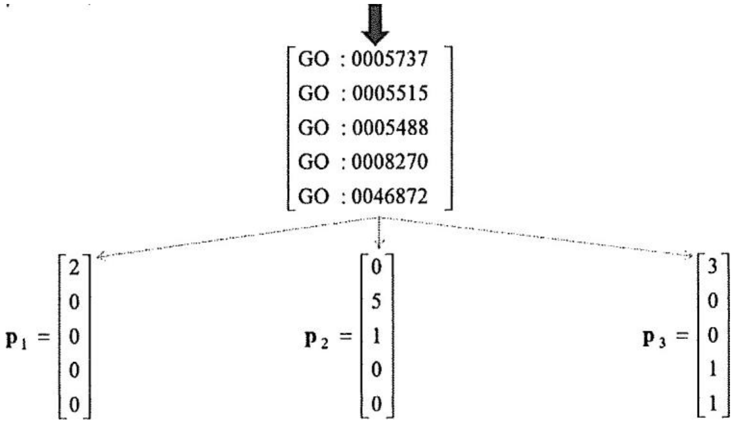


Figure 2.2: An example illustrating the construction of GO vectors based on term-frequencies.

Note that the logarithmic term in Eq. 2.3 is analogous to the inverse document frequency commonly used in document retrieval [30]. The idea is to emphasize (resp. suppress) the GO terms that have a low (resp. high) frequency of occurrences in the protein sequences. The reason is that if a GO term occurs in every sequence, it is not very useful for classification.

Start a sentence with a subject rather than the verb “note”, i.e. “It should be noted that...”

4. Term Frequency—Inverse Sequence Frequency (TF-ISF). This approach combines term-frequency (TF) and inverse sequence frequency (ISF) mentioned above. Specifically, the GO vector Pi of the i-th protein is defined as:

$$\mathbf{p}_i = \begin{bmatrix} d_{i,1} \\ \vdots \\ d_{i,j} \\ \vdots \\ d_{i,1203} \end{bmatrix}, \; d_{i,j} = b_{i,j} \log\left(\frac{N}{|\{k : b_{k,j} \neq 0\}|}\right) \qquad (2.4)$$

where bi,j is defined in Eq. 2.2.

✓ *States following about Section 2.1.2:*

a. Rationale for section

b. Preview of contents

c. Explanation of methods

2.1.2 Post-processing of GO Vectors

Although the raw GO vectors can be directly applied to support vector machines (SVMs) for classification, better performance may be obtained by post-processing the raw vectors before SVM classification. Here we introduce two post-processing methods: (1) vector norm and (2) geometric mean.

1. Vector Norm. Given the i-th GO training vector Pi, the vector is normalized as:

$$\mathbf{x}_i^{(v)} = [x_{i,1}^{(v)}, \ldots, x_{i,1203}^{(v)}]^{\mathsf{T}} \text{ where } x_{i,j}^{(v)} = \frac{p_{i,j}}{||\mathbf{p}_i||} \qquad (2.5)$$

where the superscript (v) stands for vector norm, and Pi,j is the j-th element of Pi. In case llPill = 0, we set all the element of c(.^{v}.) 0. Similarly, given the i-th test vector p/i, the GO test vector is normalized as:

$$\mathbf{x}_i^{(v)\prime} = \left[x_{i,1}^{(v)\prime}, \ldots, x_{i,1203}^{(v)}{}'\right]^{\mathsf{T}} \text{ where } x_{i,j}^{(v)\prime} = \frac{p'_{i,j}}{||\mathbf{p}'_i||} \qquad (2.6)$$

2. Geometric Mean. This method involves pairwise comparison of GO vectors, followed by normalization.

   -Pairwise Comparison: Denote P = [PI, P2, . . as a T x 1203 matrix whose rows are the raw GO vectors of T training sequences. Given the i-th GO training vector Pi, we compute the dot products between Pi and each of the training GO vectors to obtain a T-dim vector:

$$\mathbf{x}_i = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T]^{\mathsf{T}} \mathbf{p}_i = \mathbf{P}\mathbf{p}_i \ , \ i = 1, \ldots, T. \tag{2.7}$$

   During testing, given the i-th test vector pl., we compute

$$\mathbf{x}'_i = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_T]^{\mathsf{T}} \mathbf{p}'_i = \mathbf{P}\mathbf{p}'_i \ , \ i = 1, \ldots, T' \tag{2.8}$$

   where T' is the number of test vectors (sequences).

   -Normalization: The j-th elements of Xi is divided by the geometric mean of the i-th element of Xi and the j-th element of xj, leading to the normalized vectors:

$$\mathbf{x}_i^{(g)} = [x_{i,1}^{(g)}, \ldots, x_{i,T}^{(g)}]^{\mathsf{T}} \text{ where } x_{i,j}^{(g)} = \frac{x_{i,j}}{\sqrt{x_{i,i} x_{j,j}}} \tag{2.9}$$

   where the superscript (g) stands for geometric mean. Note that pairwise comparison guarantees that the elements and cj j exist for i, j.

2.1.3 Multiclass S VM Classification

Support Vector Machines (SVMs) were originally proposed by Vapnik [31] to tackle binary classification problems. An SVM classifier maps a set of input patterns into a high-dimensional space and then finds the optimal separating hyperplane and the margin of separations in that space. The obtained hyperplane is able to classify the patterns into two categories and maximize their distance from the hyperplane. To tackle the multi-class problems, the

✓ Evaluates techniques used, e.g. "successfully"

✓ Describes features of method

✓ Places adverbs, i.e. _____ly words next to the verb, not at the start of the sentence, e.g. "originally proposed", and "typically used"

one-vs-rest approach described below is typically used.

After GO vector construction and post-processing, the vectors Pi, XP), or XP) can be used for training one-vs-rest SVMs. Specifically, for an M-class problem (here M is the number of subcellular locations), M independent SVMs are trained. During testing, given an unknown protein with GO vector p', the output of the rn-th SVM is:

$$s_m^{\mathrm{GO}}(\mathbf{p}') = \sum_{r \in \mathrm{SV}_m^{\mathrm{GO}}} \alpha_{m,r}^{\mathrm{GO}} y_{m,r}^{\mathrm{GO}} K^{\mathrm{GO}}(\mathbf{p}_r, \mathbf{p}') + b_m^{\mathrm{GO}} \qquad (2.10)$$

where $SV^{G}m^{O}$ is the set of support vector indexes corresponding to the m-th SVM, e {—1, +1} are the class labels, are the Lagrange multipliers, and K^{GO} (pr, p') is a kernel function. The form of K^{G0} (pr, p') depends on the post-processing method being used. For example, if vector norm is used for normalization, the kernel becomes:

$$K^{\mathrm{GO}}(\mathbf{p}_r, \mathbf{p}') = \left\langle \mathbf{x}_r^{(v)}, \mathbf{x}^{(v)\prime} \right\rangle \qquad (2.11)$$

The SVM score can be combined with the score of the profile alignment SVM described next.

2.2 Homology-Based Methods

Kernel techniques based on profile alignment have been used successfully in detecting remote homologous proteins [32] and in predicting subcellular locations of eukaryotic proteins [19]. Instead of extracting feature vectors directly from sequences, profile alignment method trains an SVM classifier by using the scores of local profile alignment.

✓ *Gives general background and evaluates work using "successfully"*

This method, namely PairProSVM, extracts the features from protein sequences by aligning the profiles of the sequences with each of the training profiles [19]. A profile is a matrix in which elements in a column (sequence position) specify the frequency of individual amino acids appeared in the corresponding position of some homologous

✓ *Describes features of method*

✓ *Links ideas between paragraphs well using "This method"* sequences. Given a sequence, a profile can be derived by aligning it with a set of similar sequences. The similarity score between a known and an unknown sequence can be computed by aligning the profile of the known sequence with that of the unknown sequence [32]. Since the comparison involves not only two sequences but also their closely related sequences, the score is more sensitive to detecting weak similarity between protein families.

The profile of a sequence can be obtained by presenting the sequence to PSIBLAST [33] that searches against a protein database for homologous sequences. The information pertaining to the aligned sequences is represented by two matrices: position-specific scoring matrix (PSSM) and position-specific frequency matrix (PSFM). Each entry of a P SSM represents the log-likelihood of the residue substitutions at the corresponding position in the query sequence. The PSFM contains the weighted observation frequencies of each position of the aligned sequences.

Fig. 2.3 illustrates the flow of the profile alignment method for subcellular localization. Given a query sequence, we first obtain its profile by presenting it to PSI-BLAST. Then we align it with the profile of each training sequence to form an alignment score vector, which is further used as inputs to an SVM classifier for classification. Mathematically, given the i-th test protein sequence, we align its profile with each of the training

✓ Links ideas using "which"

💡 Use a passive verb to avoid we, i.e. "Then it is aligned…", place adverbs next to the verb, i.e. "are mathematically aligned".

profiles to obtain a profile-alignment test vector.

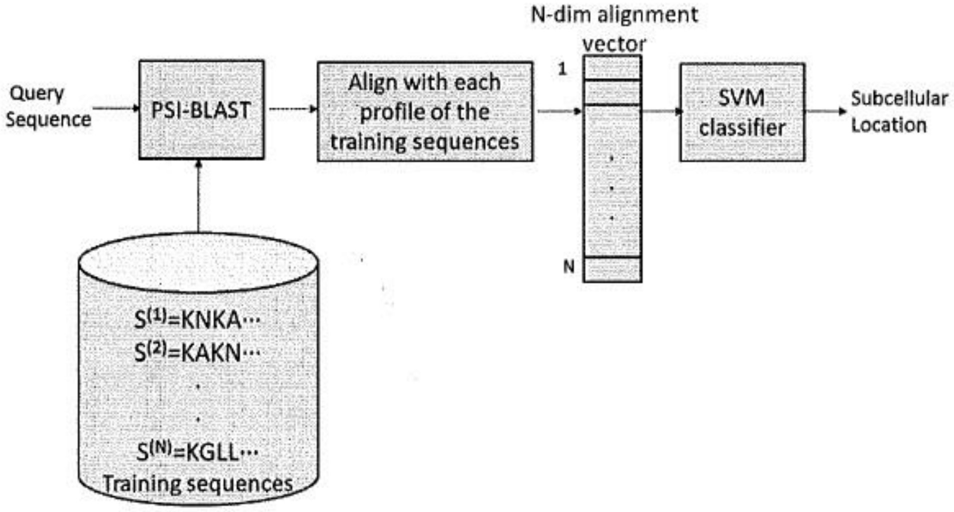


Figure 2.3: Flowchart of profile alignment method.

q/i, whose elements are then normalized by the geometric mean as follows:

$$\mathbf{q}_i^{(g)\prime} = [q_{i,1}^{(g)\prime}, \ldots, q_{i,T}^{(g)\prime}]^\mathsf{T} \text{ where } q_{i,j}^{(g)\prime} = \frac{q'_{i,j}}{\sqrt{q'_{i,i} q'_{j,j}}}. \tag{2.12}$$

Similar to the GO method, a one-versus-rest SVM classifier was used to classify the profile-alignment vectors. Specifically, the score of the rn-th profile alignment SVM is

$$s_m^{\text{PA}}(\mathbf{q}') = \sum_{r \in \text{SV}_m^{\text{PA}}} \alpha_{m,r}^{\text{PA}} y_{m,r}^{\text{PA}} K^{\text{PA}}(\mathbf{q}_r, \mathbf{q}') + b_m^{\text{PA}} \tag{2.13}$$

which is to be fused with the score of the GO SVM.

2.3 Fusion of Functional-Domain and Homology Based Methods

Fig. 2.4 illustrates the fusion of InterProGOSVM and PairProSVM. The GO and profile alignment scores produced by the GO and profile alignment SVMs are normalized by Z-norm:

Explain the main idea in the text as well as in the sub-heading, i.e. "There are also methods that fuse functional-domain and homology-based methods."

Give an outline of the content at the start of the section.

$$\tilde{s}_m^{\text{GO}}(\mathbf{p}') = \frac{s_m^{\text{GO}}(\mathbf{p}') - \mu_m^{\text{GO}}}{\sigma_m^{\text{GO}}} \text{ and } \tilde{s}_m^{\text{PA}}(\mathbf{q}') = \frac{s_m^{\text{PA}}(\mathbf{q}') - \mu_m^{\text{PA}}}{\sigma_m^{\text{PA}}} \tag{2.14}$$

where (pm GO , amGO) and (um$^{\text{PA}}$, ãm$^{\text{PA}}$) are respectively the mean and standard derivation of the GO and profile alignment SVM scores derived from the training sequences. The normalized GO and profile-alignment SVM scores are fused:

$$\tilde{s}_m^{\text{Fuse}}(\mathbf{p}', \mathbf{q}') = w^{\text{GO}}\tilde{s}_m^{\text{GO}}(\mathbf{p}') + w^{\text{PA}}\tilde{s}_m^{\text{PA}}(\mathbf{q}') \tag{2.15}$$

where W$^{\text{G}}$O + WPA — 1. Finally, the predicted class of the test sequence is given by

$$m^* = \arg \max_{m=1}^{M} \tilde{s}_m^{\text{Fuse}}(\mathbf{p}', \mathbf{q}'). \tag{2.16}$$

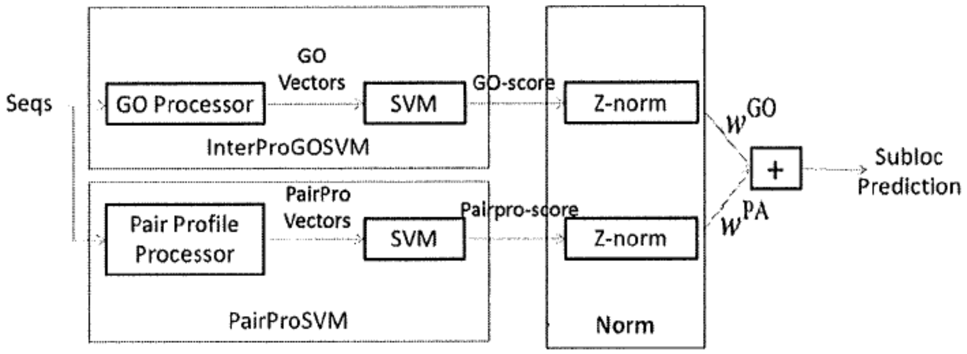


Figure 2.4: Fusion of InterProGOSVM and PairProSVM

Chapter 3

3 GOASVM Method

This chapter proposes a functional-domain method which retrieves the GO terms directly from the Gene Ontology Annotation (GOA) database instead of indirectly from InterProScan. The high performance of this method demonstrates its superiority over the previous InterProGOSVM methods. For those proteins that do not have accession numbers, this chapter proposes using PSI-BLAST to find the homologs and to use their accession numbers (ANs) to retrieve the GO terms from the GOA database.

✓ Explains what is in rest of chapter, using present simple tense "proposes"

3.1 Gene Ontology Annotation Database

As a result of the GO Consortium annotation effort, the Gene Ontology Annotation (GOA) database[1] has become a large and comprehensive resource for proteomics research [34]. The database provides structured annotations to nonredundant proteins using standardized GO vocabularies. Proteins of different species in UniProt Knowledgebase [35], which includes Swiss-Prot [36], TrEMBL [36] and P IR-PSI) [37], have been annotated through a combination of electronic and manual techniques. The large-scale assignment of GO terms to UniProKB entries (or ANs, short for Accession Numbers) has been made possible by successfully converting a proportion of the existing knowledge held within the UniProKB database into GO terms [34]. GOA also includes a series of specific bi-directional cross-references to other databases. For example, the majority of UniProtKB entries contain cross-references to an InterPro identification number and vice versa. InterPro is a key database maintained by European Bioinformatics Institute (EBI) [38]. The GO assignments are released monthly, in accordance with a format standardized by the GO Consortium, in a 'gene association file'. As a result, using different releases of the same database may bring

✓ Provides general background, using present perfect tense, e.g. "has become"

✓ Explains how database works using key terms

✓ Refers to previous research using evaluative language, e.g. comprehensive, successfully

✓ Gives an example

✓ Explains example

different or even significantly distinct results. Typically, the newer the version of GOA database is, the better the results will be.

The systematic integration of GO annotations and the UniProtKB can be exploited for subcellular localization. Specifically, given the accession number of a protein, a set of GO terms can be retrieved from the GOA database file. [ii] In UniProKB, each protein has a unique accession number (AN), and in the GOA database, each AN may associate with zero, one or more distinct GO terms. Conversely, one GO term may associate with zero, one, or many different ANS. This means that the mappings between ANS and GO terms are many-to-many.

✓ Uses variety of verbs to refer to figures, e.g. "shows", "suggest"

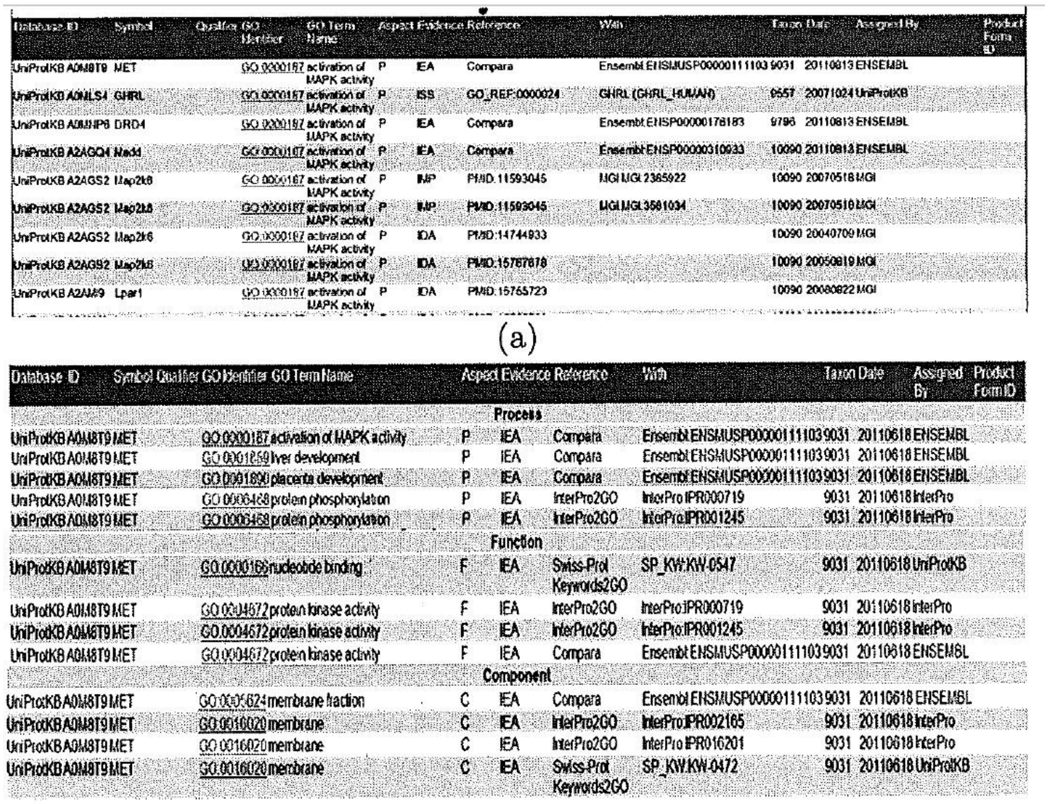
Fig. 3.1(a) shows the query result (under Protein Annotation) of the GOA webserver using the GO terms GO:0000187 as the searching key, and Fig. 3.1(b) displays the query results using the accession number AOM8T9 as the searching key. As can be seen, the same GO term—GO:0000187—can be associated with UniProtKB ID or ANs AOM8T9, AOMLS4, AOMNP6, etc. The same UniProtKB ID AOM8T9 can be associated with GO:0000187, GO:0001889, GO:0001890, etc. These two examples suggest that the mappings between ANS and GO terms are many-to-many, which enables us to make full use of them for classification of proteins. These figures also suggest that GO annotations have different degree of reliability or 'evidence'. The evidences are based on the information sources from which the annotations are produced. The sources include IEA (Inferred from Electronic Annotation), ISS (Inferred from Structural and Sequence Similarity), IMP (Inferred from Mutant Phenotype), IDA (Inferred from Direct Assay), etc.

💡 Give a few examples rather than use "etc.", and use "and a number of other sources" rather than a long list.

3.2 Retrieval of GO Terms

Because some proteins, such as those newly discovered proteins, may not have a GO term in the GOA database, it is essential to develop strategies to handle these special cases. Here, we introduce three different approaches—using ANS only, using sequences only, and using both ANS and sequences—to generating valid GO terms.

💡 Link ideas by changing the word than repeat the word "proteins", i.e. "Because some proteins, such as newly discovered ones,"



Database ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
UniProtKB A0M8T9	MET		GO:0000187	activation of MAPK activity	P	IEA	Compara	Ensembl:ENSMUSP00000111103	9031	20110813	ENSEMBL	
UniProtKB A0MLS4	GHRL		GO:0000187	activation of MAPK activity	P	ISS	GO_REF:0000024	GHRL (GHRL_HUMAN)	9557	20071024	UniProtKB	
UniProtKB A0MNP8	DRD4		GO:0000187	activation of MAPK activity	P	IEA	Compara	Ensembl:ENSP00000176183	9796	20110813	ENSEMBL	
UniProtKB A2AGQ4	Nadd		GO:0000187	activation of MAPK activity	P	IEA	Compara	Ensembl:ENSP00000310933	10090	20110813	ENSEMBL	
UniProtKB A2AGS2	Map2k6		GO:0000187	activation of MAPK activity	P	IMP	PMID:11593045	MGI:MGI:2385922	10090	20070518	MGI	
UniProtKB A2AGS2	Map2k6		GO:0000187	activation of MAPK activity	P	IMP	PMID:11593045	MGI:MGI:3581034	10090	20070518	MGI	
UniProtKB A2AGS2	Map2k6		GO:0000187	activation of MAPK activity	P	IDA	PMID:14744933		10090	20040709	MGI	
UniProtKB A2AGS2	Map2k6		GO:0000187	activation of MAPK activity	P	IDA	PMID:15767678		10090	20050819	MGI	
UniProtKB A2AM49	Lpar1		GO:0000187	activation of MAPK activity	P	IDA	PMID:15765723		10090	20080822	MGI	

(a)

Database ID	Symbol	Qualifier	GO Identifier	GO Term Name	Aspect	Evidence	Reference	With	Taxon	Date	Assigned By	Product Form ID
					Process							
UniProtKB A0M8T9	MET		GO:0000187	activation of MAPK activity	P	IEA	Compara	Ensembl:ENSMUSP00000111103	9031	20110618	ENSEMBL	
UniProtKB A0M8T9	MET		GO:0001889	liver development	P	IEA	Compara	Ensembl:ENSMUSP00000111103	9031	20110618	ENSEMBL	
UniProtKB A0M8T9	MET		GO:0001890	placenta development	P	IEA	Compara	Ensembl:ENSMUSP00000111103	9031	20110618	ENSEMBL	
UniProtKB A0M8T9	MET		GO:0006468	protein phosphorylation	P	IEA	InterPro2GO	InterPro:IPR000719	9031	20110618	InterPro	
UniProtKB A0M8T9	MET		GO:0006468	protein phosphorylation	P	IEA	InterPro2GO	InterPro:IPR001245	9031	20110618	InterPro	
					Function							
UniProtKB A0M8T9	MET		GO:0000166	nucleotide binding	F	IEA	Swiss-Prot Keywords2GO	SP_KW:KW-0547	9031	20110618	UniProtKB	
UniProtKB A0M8T9	MET		GO:0004672	protein kinase activity	F	IEA	InterPro2GO	InterPro:IPR000719	9031	20110618	InterPro	
UniProtKB A0M8T9	MET		GO:0004672	protein kinase activity	F	IEA	InterPro2GO	InterPro:IPR001245	9031	20110618	InterPro	
UniProtKB A0M8T9	MET		GO:0004672	protein kinase activity	F	IEA	Compara	Ensembl:ENSMUSP00000111103	9031	20110618	ENSEMBL	
					Component							
UniProtKB A0M8T9	MET		GO:0005624	membrane fraction	C	IEA	Compara	Ensembl:ENSMUSP00000111103	9031	20110618	ENSEMBL	
UniProtKB A0M8T9	MET		GO:0016020	membrane	C	IEA	InterPro2GO	InterPro:IPR002165	9031	20110618	InterPro	
UniProtKB A0M8T9	MET		GO:0016020	membrane	C	IEA	InterPro2GO	InterPro:IPR016201	9031	20110618	InterPro	
UniProtKB A0M8T9	MET		GO:0016020	membrane	C	IEA	Swiss-Prot Keywords2GO	SP_KW:KW-0472	9031	20110618	UniProtKB	

(b)

Figure 3.1: The query results of GOA webserver (http://www.ebi.ac.uk/GOA) using (a) a GO term (GO:0000187) and (b) an accession number (AOM8T9) as the searching key.

3.2.1 GO Terms Retrieval Using ANS Only

For proteins with known ANs, we can directly retrieve the GO terms from the many-to-many mapping between ANS and GO terms in the GOA database. This approach is similar to the one described in Chapter 2 in that both aim to retrieve GO terms from databases. However, there are also important differences. In particular, the InterProScan in Chapter 2 uses various algorithms to search for relevant GO terms from different protein-signature databases, while the approach described here uses the AN of a protein as the

✓ Provides general background to introduce section

✓ Refers to previous section

✓ Uses "However" to introduce the main topic of this section

✓ Uses "In particular" to introduce details

✓ Evaluates approach

searching key to search against the GOA database to retrieve the GO terms. The latter approach is more direct because the GOA database is more comprehensive than the databases used by InterProScan and it focuses on GO related information. Therefore, the information extracted from GOA is undoubtedly richer than that extracted from InterProScan.

✓ Makes evaluative comment on data, e.g. "Undoubtedly richer"
✓ Refers to work in previous chapter

While proteins with known ANS have already been labelled, they may not be functionally annotated. In fact, there are proteins that have ANS but their ANS do not associate with any GO terms in the GOA database.

Fig. 3.2 illustrates the flow of GOASVM using only accession number (ANs). After retrieving the GO terms, similar to Chapter 2, we construct the GO vectors using 1-0 value, TF, ISF and TF-ISF. Then, the obtained GO vectors are postprocessed by Vector Norm or Geometric Mean. The raw GO vectors (without post-processing) or normalized vectors are then directly recognized by SVM classifers. The only difference between GOASVM and InterProGOSVM mentioned in Chapter 2 is that the GO terms are retrieved by searching the GOA database using the accession numbers (ANs) as the keys, whereas for the latter, GO terms are retrieved via InterProScan. As the GOA database contains biological annotation of proteins while InterProScan relies on computational algorithms, the performance achieved by using the GO terms extracted from the GOA database is expected to be better than that by using the GO terms extracted by InterProScan.

✓ Predicts possible outcome of approach

3.2.2 GO Terms Retrieval Using Sequences Only

Note that the GOA database does not contain any amino-acid sequences. As a result, it is impossible to use amino acid sequences as searching keys. However, newly discovered proteins may only have amino acid sequences and do not have an accession number. Apparently, the GO terms retrieval technique mentioned in Section 3.2.1 could not be used. Fortunately, we can use PSI-BLAST [20] to find the

💡*Put the main point first to strengthen the topic sentence, i.e. "It is impossible to use …because the GOA database does not… ".*

✓ Clear structure to section
✓ Explains problem
✓ Gives possible solution

homologs of these proteins and use their ANS as the searching keys to retrieve the GO terms from the GOA database.

PSI-BLAST can find remote homologs for the unknown proteins. We can adjust the parameters of PSI-BLAST (e.g. varying the value of the option E- value) to control remoteness of the homologs with respect to the unknown proteins and the number of remote homologs.

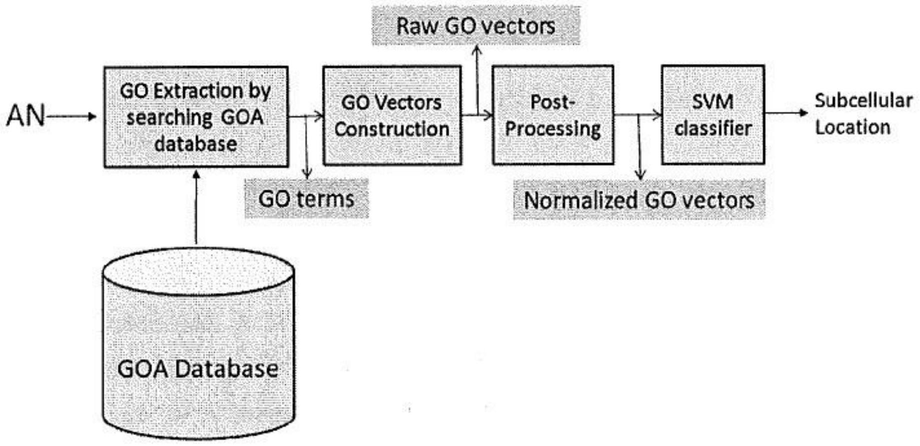


Figure 3.2: Flowchart of GOASVM method using only accession numbers (ANs)

Although a large number of homologs can provide more information, some of the information could be redundant or even irrelevant (noise). This raises another question: how many homologs should we take for each protein sequence? Here, we take the top homolog because it is the most relevant and more homologs are likely to bring us more irrelevant (or noise) information than useful information.

✓ *Compares solution with alternative*

✓ *Predicts possible results*

In general, suppose PSI-BLAST finds n homologs for a sequence. We use the obtained n ANS as searching keys to retrieve n sets of GO terms for the sequence, which results in n GO vectors.

✓ *Use "suppose" in complete sentence, i.e. "Suppose PSI BLAST finds..., then we use the obtained..."*

Fig. 3.3 illustrates the flow of GOASVM using only sequences as input. There are n ANS and n GO vectors for each sequence, which result in n scores for each protein sequence. Then, n scores are linearly combined to obtain a weighted score for classification. Mathematically, if we choose n ANS for each sequence, the number of training vectors for the SVMs will be n times bigger. Denote (p'i,j) as the score of the m-th SVM for the i-th test protein by using the j-th AN. Then we fuse these n SVM scores as:

$$\tilde{s}_m^{\mathrm{GO}}(\mathbf{p}'_i) = \sum_{j=1}^{n} w_j s_m^{\mathrm{GO}}(\mathbf{p}'_{i,j}) \tag{3.1}$$

wherej = 1, . . . ,n, and 237:1 wj : 1. For convenience, we set wj = 1/n.

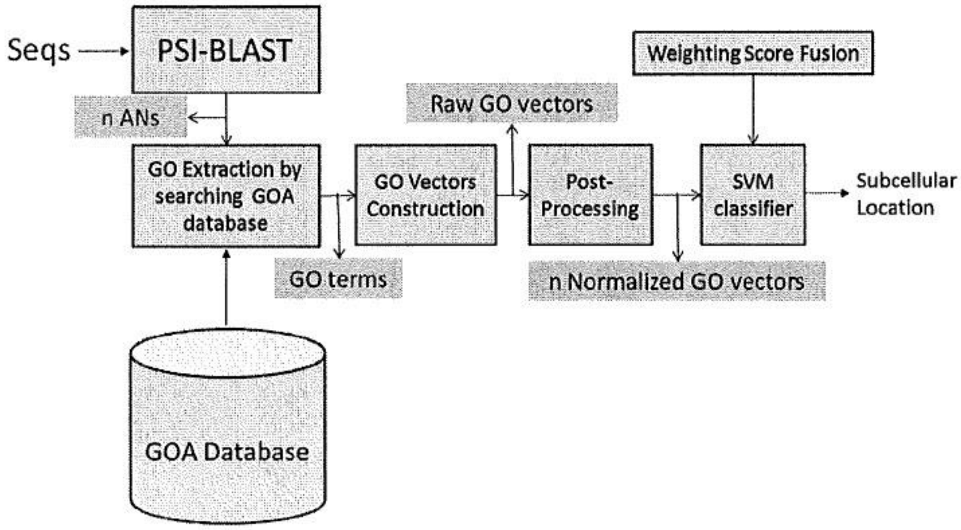


Figure 3.3: Flowchart of GOASVM method using only sequences

(Better weighting factors will be found in further research). Finally, the predicted class of the test sequence is given by

$$m^* = \arg\max_{m=1}^{M} \tilde{s}_m^{\mathrm{GO}}(\mathbf{p}'_i), \tag{3.2}$$

m=l where M is the number of subcellular locations.

3.2.3 GO Terms Retrieval Using both ANS and Sequences

Actually the two retrieval methods described in Section 3.2.1 and Section 3.2.2 are complementary and enough for prediction of subcellular localization. But the proteins that are suitable for the first method (using ANs) are also applicable to the second method (using sequences). It is logical to ask: For the proteins that are applicable to both cases, can the combination of these two methods further improve classification performance? This brings us the third method—using both ANS and sequences.

Do not start a sentence with "actually". Place the adverb before the verb i.e. "actually are".
✓ Suggests alternative method as research question after reviewing both existing methods

In this method, for each protein sequence, we retrieve the GO terms separately using only ANS and using only sequences from the same GOA database. In general, suppose we find n ANS for each sequence. That means each sequence will generate (n + 1) set of GO terms—one is from the true AN and the other n from the ANS of n homologs. In this case, we should combine them together to determine the total distinct GO terms, which will be used to determine the number of the vector dimension.

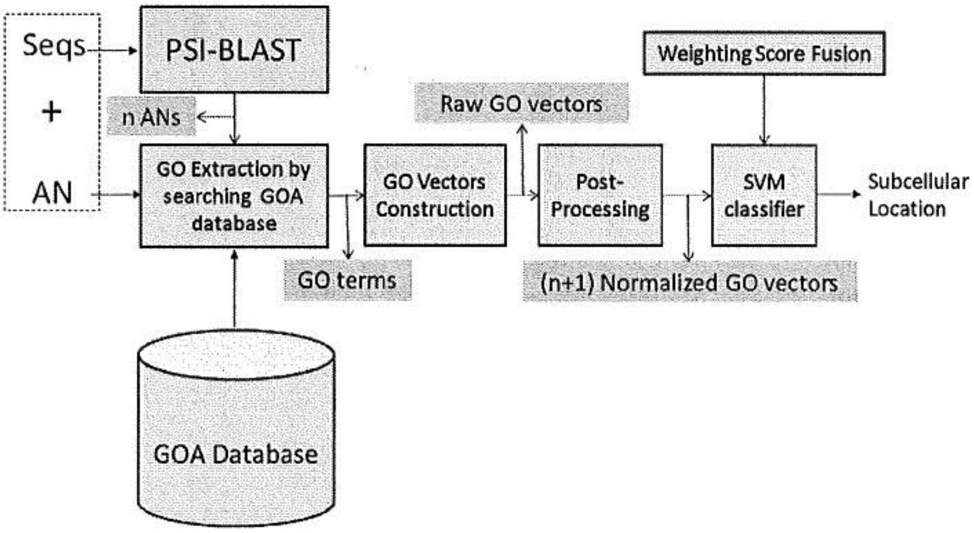


Figure 3.4: Flowchart of GOASVM method using both accession numbers (ANs) and sequences

Fig. 3.4 illustrates the flow of GOASVM using both ANS and sequences. the scoring system in SVM classifiers is almost the same as the case using only sequences. Similarly, suppose is the score of the m-th SVM for the i-th test protein by using the j-th AN. Then the fusion score of the ^{t}m-th SVM for the i-th sequence is:

$$\tilde{s}^{\mathrm{GO}}_{m}(\mathbf{p}'_{i}) = \sum_{j=1}^{n+1} w_{j} s^{\mathrm{GO}}_{m}(\mathbf{p}'_{i,j}) \qquad (3.3)$$

where j — 1, (n + 1), and X--'0+1 'Wj 1. Here, as we only select one AN for each sequence, the weighting factors for the true AN and the homologous AN is 0.5 : 0.5. So wj = 0.5 for j - — 1,2. Uses ‘actually’ to start a sentence.

Chapter 4

4 Experimental Setup

This chapter is divided into two parts: dataset construction and performance metrics. The former describes the details of how the datasets are constructed, while the latter specifies the performance evaluation measures.

4.1 Dataset Construction

For the fusion of InterProGOSVM and PairProSVM, the performance was evaluated on Huang and LPs dataset [39], which was created by selecting all eukaryotic proteins with annotated subcellular locations from Swiss-Prot 41.0. The dataset comprises 3572 proteins with 11 classes. The breakdown of the dataset is shown in Table 4.1. Specifically, there are 622 cytoplasm, 1188 nuclear, 424 mitochondria, 915 extracellular, 26 Golgi apparatus, 225 chloroplast, 45 endoplasmic reticulum, 7 cytoskeleton, 29 vacuole, 47 peroxisome, and 44 lysosome. The sequence similarity is cut off at 50%.

✓*Clear explanation of how data set is constructed*

💡*Start the topic sentence with the subject, not start with for: "For the fusion," and again later "For GOASVM, construction," Use "The performance was evaluated for the fusion of…"*

Among the 3572 protein sequences, only 3120 sequences have valid GO vectors (with at least one non-zero element in the GO vectors). For the remaining 452 sequences, InterProScan cannot find any GO terms. Therefore, we only used sequences with valid GO vectors in our experiments and reduced the dataset size to 3120 protein sequences.

For GOASVM, the performance was evaluated on Chou's dataset [21]. It consists of 4150 protein sequences, of which there are 2423 training protein sequences and 1727 testing protein sequences distributed into 16 subcellular compartments (classes).

💡*Do not start a sentence with "for". The subject is 'The performance of GOASVM' and the sentence should be " The performance of GOASVM was evaluated on ..."*

Label	Subcellular Location	No. of Sequence
1	Cytoplasm	622
2	Nuclear	1188
3	Mitochondria	424
4	Extracellular	915
5	Golgi apparatus	24
6	Chloroplast	225
7	Endoplasmic reticulum	45
8	Cytoskeleton	7
9	Vacuole	29
10	Peroxisome	47
11	Lysosome	44
Total		3572

Table 4.1: Breakdown of the dataset used in the fusion of InterProGOSVM methods and PairProSVM. This dataset is extracted from Swiss-Prot 41.0 and the sequence similarity is cut off to 50%.

💡*Place table headings above a table.*

The protein sequences in this dataset were collected from the SwissProt 48.2 according to their experimentally annotated subcellular locations. To obtain high-quality, well-defined working datasets, the data were screened strictly according to some criteria described below [21]:

- ✓ *Good to specify criteria for screening data*
- ✓ *Uses past simple tense and passive voice to describe methodology*

1. Only protein sequences annotated with 'eukaryotic' were included, since the current study only focused on eukaryotic proteins;

2. Sequences annotated with ambiguous or uncertain terms, such as 'probably', 'maybe', 'probable', 'potential', or 'by similarity', were excluded;

💡 *Do not use vague words such as "some". Better to use "the criteria" or "a number of criteria including those described below."*

3. Those protein sequences labelled with two or more subcellular locations were excluded because of the lack of uniqueness;

4. Sequences annotated with 'fragments' were excluded and also, sequences with less than 50 amino acid residues were removed since these proteins might just be fragments;

5. To avoid any homology bias, the sequence similarity in the same subcellular location among the obtained dataset was cut off at 25% operated by a culling program [40] to winnnow the redundant sequences;

6. Subcellular locations (subsets) containing less than 20 protein sequences were left out because of lacking statistical significance.

After strictly following the criteria mentioned above, only 4150 protein sequences were found, of which there are 25 cell wall, 21 centriole, 258 chloroplast, 97 cyanelle, 718 cytoplasm, 25 cytoskeleton, 113 endoplasmic reticulum, 806 extracellular, 85 Golgi apparatus, 46 lysosome, 228 mitochondrion, 1169 nucleus, 64 peroxisome, 413 plasma membrane, 38 plastid, and 44 vacuole. Then, this dataset was further divided into training dataset (2423 sequences) and testing dataset (1727 sequences). And the specific numbers of proteins within each compartment of the training and testing datasets are shown in Table 4.2. As can be seen, both the training and testing datasets are quite imbalanced. The number of proteins in different subcellular locations vary significantly (from 4 to 695). Further, the datasets are

💡 *Use lack as a verb or a noun, i.e. "because of a lack of statistical ..."*
Or "because it lacks statistical..."

✓ *Clear explanation of procedures adopted to identify protein sequences.*

💡 *Avoid using "and" to start a sentence.*

💡 *Avoid using "we" and use passive, i.e. Gene Ontology Annotation was used as..."*

💡 *Cite Chou's dataset.*

both in low sequence similarity and in 16 subcellular locations. Thus, the properties of the training and testing dataset are imbalanced, multi-class distributed and in low sequence similarity, which make conventional methods difficult to classify.

In the experiments, we used the Gene Ontology Annotation (GOA) database (released on 08-March-2011) as the retrieval database. When using the ANS of the proteins in Chou's dataset as the searching keys to search against this database, 5450 distinct GO terms were found. When using the ANS of homology proteins found by PSI-BLAST as the searching key, the number of distinct GO terms is 5430. When using both ANS in the dataset and the ANS found by PSI-BLAST, 5465 distinct GO terms were found.

Use the past simple tense when describing your own methodology. The present tense (i.e. "may have" refers to the standard way the procedure is always done). Write "The last part might have..."

4.2 Performance Metrics

Five-fold cross validation was used for performance evaluation. This ensures that every sequence in the dataset will be tested. In the five-fold cross validation, the whole dataset was randomly divided into 5 disjoint parts with equal size [41]. The last part may have 1-4 more examples than the former 4 parts in order for each example to be evaluated on the model. Then one part of the dataset was used as the test set and the remained parts are jointly used as the training set. This procedure is repeated for five times, and each time a different part was chosen as the test set.

✓ *Explains five-fold cross validation and justifies the approach*

Label	Subcellular Location	Training Dataset	Testing Dataset
1	Cell Wall	20	5
2	Centriole	17	4
3	Chloroplast	207	51
4	Cyanelle	78	19
5	Cytoplasm	384	334
6	Cytoskeleton	20	5
7	Endoplasmic reticulum	91	22
8	Extracellular	402	404
9	Golgi apparatus	68	17
10	Lysosome	37	9
11	Mitochondrion	183	45 695 12
12	Nucleus	474	90
13	Peroxisome	52	7
14	Plasma membrane	323	8
15	Plastid	31	
16	Vacuole	36	
Total		2423	1727

Table 4.2: Breakdown of the dataset used in the GOASVM method. This dataset is extracted from Swiss-Prot 48.2 and the sequence similarity is cut off to 25%.

Explain the content of the table in a paragraph. "The table shows that..." and place the heading above the table.

For GOASVM, both five-fold cross validation and independent tests were performed. For five-fold cross validation, the training dataset was used. For the independent tests, the whole training dataset was used for training the SVMs and the independent test set was used for performance evaluation.

We used several performance measures, including the overall accuracy (ACC) , overall Mathew's correlation coefficient (OMCC) [19] and weighted average Mathew's correlation (WAMCC) [19]. The latter two measures OMCC and WAMCC are based on Mathew's correlation coefficient (MCC) [42]. MCC can overcome the shortcoming of accuracy on imbalanced data and have the advantage of avoiding the performance to be dominated by the majority classes. For example, a classifier which predicts all samples as positive cannot be regarded as a

✓ Clear description of performance measures adopted.

✓ Rationale for choice of MCC as performance measure provided

Do not use "for example". This is a comparison and it would be better to use "In contrast".

good classifier unless it can also predict negative samples accurately. In this case, the accuracy and MCC of the positive class are 100% and 0%, respectively. Therefore, MCC is a better measure for imbalanced classification.

Denote M e R^{Cxc} as the confusion matrix of the prediction results, where C is the number of subcellular locations. Then Mij(l i, j < C) represents the number of proteins that actually belong to class i but are predicted as class j.

Then, we further denote:

$$p_c = M_{c,c}, \tag{4.1}$$

$$q_c = \sum_{i=1,i\neq c}^{C} \sum_{j=1,j\neq c}^{C} M_{i,j}, \tag{4.2}$$

$$r_c = \sum_{i=1,i\neq c}^{C} M_{i,c}, \tag{4.3}$$

$$s_c = \sum_{j=1,j\neq c}^{C} M_{c,j}, \tag{4.4}$$

where c(l < c < C) is the index of a particular subcellular location. For class c, pc is the number of true positives, qc is the number of true negatives, Tc is the number of false positives, and sc is the number of false negatives. Based on these notations, the ACC, MCCc for class c, OMCC and WAMCC are defined respectively as:

$$\mathrm{ACC} = \frac{\sum_{c=1}^{C} M_{c,c}}{\sum_{i=1}^{C}\sum_{j=1}^{C} M_{i,j}}, \tag{4.5}$$

$$\mathrm{MCC}_c = \frac{p_c q_c - r_c s_c}{\sqrt{(p_c + s_c)(p_c + r_c)(q_c + s_c)(q_c + r_c)}}, \tag{4.6}$$

$$\mathrm{OMCC} = \frac{\hat{p}\hat{q} - \hat{r}\hat{s}}{\sqrt{(\hat{p} + \hat{r})(\hat{p} + \hat{s})(\hat{q} + \hat{r})(\hat{q} + \hat{s})}}, \tag{4.7}$$

$$\mathrm{WAMCC} = \sum_{c=1}^{C} \frac{p + s}{N} \mathrm{MCC}_c, \tag{4.8}$$

where $N = \sum_{c=1}^{C} p_c + s_c, \hat{p} = \sum_{c=1}^{C} p_c, \hat{q} = \sum_{c=1}^{C} q_c, \hat{r} = \sum_{c=1}^{C} r_c, \hat{s} = \sum_{c=1}^{C} s_c$.

Chapter 5

✓*Gives clear description of findings and explains clearly data presented as charts and tables*

5 Results and Analysis

This chapter details the performance and analysis of the two proposed methods mentioned in the Chapter 2 and Chapter 3, including the fusion of functional domain based method (InterProGOSVM) and homology-based methods (PairProSVM), as well as the functional-domain based method (GOASVM).

Do not use “mentioned” and “including”. Mentioned refers to something that is not the main point. Write “described in Chapter 2”. Including means not all. Here it is not needed.

5.1 Performance of FYIsion of InterProGOSVM and PairProSVM

5.1.1 Performance of PairProSVM

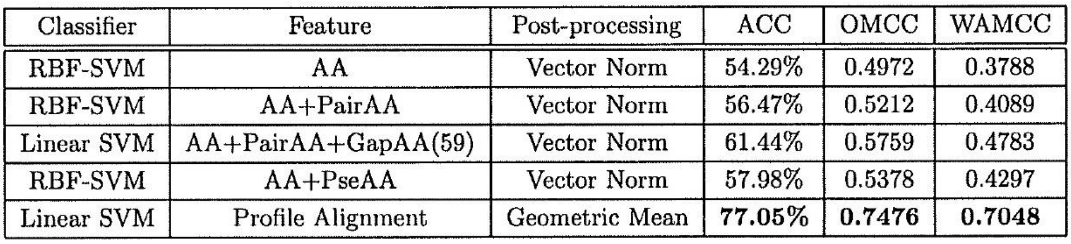
Table 5.1 shows the performance of different SVMs using various features extracted from the protein sequences. The features include amino acid composition (AA) [3], amino-acid pair composition (PairAA) [3], AA composition with the maximum gap length equal to 59 (the minimum length of all of the 3120 sequences is 61) [5], pseudo AA composition [7], and profile alignment scores. The penalty factor for training the SVMs was set to 1 for both linear SVM and RBF-SVM. For RBF-SVMs the kernel parameter was set to 1. As AA and PairAA produce low-dimensional feature vectors, the performance achieved by RBF-SVM is better than that of the linear SVM. So, we just present the performance of RBF-SVM.

✓ *Uses good clear topic sentence. Then describes details*
✓ *Refers to table*

Do not use “just” and “so” they are informal, better to use “only” i.e. “Therefore, only the performance of is presented”

✓ *Donot break a sentence in half to insert a table.*

Table 5.1 shows that amino-acid composition and its variant are not good features for subcellular localization. AA method only explores the amino acid composition information, so it performs the worst. PairAA, GapAA and the extended PseAA extract the sequence-order information, so their combinations achieve a slightly better prediction performance. Among the amino acid based methods, the highest accuracy is only 61.44%. On the other hand, the homology-based method that exploits the homologous sequences in protein databases (via PSI-BLAST) achieves a significant better performance. This suggests that the information pertaining to the amino acid sequences is limited. On the contrary, homology-based method PairProSVM can extract much more valuable information about protein subcellular localization than amino acid based methods. The higher OMCC and WAMCC also suggest that PairProSVM can handle imbalanced problems better.



Classifier	Feature	Post-processing	ACC	OMCC	WAMCC
RBF-SVM	AA	Vector Norm	54.29%	0.4972	0.3788
RBF-SVM	AA+PairAA	Vector Norm	56.47%	0.5212	0.4089
Linear SVM	AA+PairAA+GapAA(59)	Vector Norm	61.44%	0.5759	0.4783
RBF-SVM	AA+PseAA	Vector Norm	57.98%	0.5378	0.4297
Linear SVM	Profile Alignment	Geometric Mean	**77.05%**	**0.7476**	**0.7048**

Table 5.1: Performance obtained by using amino acid composition (AA) [3], amino-acid pair composition (PairAA) [3], AA composition with gap (length = 59) (GapAA) [5], pseudo AA composition (PseAA) [7], and profile alignment scores as feature vectors and different SVMs as classifiers. The last row corresponds to the PairProSVM proposed in [19].

✓ Clearly links ideas and builds the paragraph:
a. details key findings
b. compares key findings
c. explains key findings
d. evaluates key findings
using the following structures:
"so", "on the other hand"
"This suggests"
"also"

5.1.2 Performance of Different GO Vector Construction Methods and Normalization Methods

✓ *Uses "can" and "may" to show it is not certain*

Use more formal expressions to describe numbers than "quite a few" and "a lot of". Better alternatives are "a large number of" or a significant number of".

Table 5.2 shows the performance of 12 InterProGOSVM methods. For ease of reference, we label these methods as GO-I, GO-2,. . . ,GO-12. Linear SVMs were used in all cases and the penalty factor was also set to 1. When using vector norm or geometric mean to post-process the GO vectors, the inverse sequence frequency can produce more discriminated GO vectors, as evident in the higher accuracy, OMCC and WAMCC corresponding to GO-6 and GO-10. As there may be quite a few redundant GO terms existing in a lot of protein sequences,

30

Method ID	GO Vectors Construction	Post-processing	ACC	OMCC	WAMCC
GO-I	1-0 value	None	72.21%	0.6943	0.6467
GO-2	ISF	None	71.89%	0.6908	0.6438
GO-3		None	71.99%	0.6919	0.6451
GO-4	TF-ISF	None	71.15%	0.6827	0.6325
GO-5	1-0 value	Vector Norm	71.25%	0.6837	0.6335
GO-6	ISF	Vector Norm	72.02%	0.6922	0.6427
GO-7		Vector Norm	70.96%	0.6806	0.6293
GO-8	TF-ISF	Vector Norm	71.73%	0.6890	0.6386
GO-9	1-0 value	Geometric Mean	70.51%	0.6756	0.6344
GO-IO	ISP	Geometric Mean	72.08%	0.6929	0.6468
GO-II		Geometric Mean	70.64%	0.6771	0.6290
GO-12	TF-ISF	Geometric Mean	71.03%	0.6813	0.6391

Table 5.2: Performance of InterProGOSVM methods using different approaches to constructing the raw GO vectors and different post-processing approaches to normalizing the raw GO vectors. 'None' in Post-processing means that the raw GO vectors Pi are used as input to the SVMs. ISF: inverse sequence-frequency; TF: term-frequency; TF-ISF: term-frequency inverse sequence frequency.

Have a new paragraph before "Except" to show you are now introducing the findings.

✓*Interprets data ("Tables 5.1 and 5.2 suggest that") instead of only referring to it*

using ISF can remove or weaken their impact on final prediction of subcellular locations. Except for ISF, using the raw GO vectors as the SVM input achieves the best performance, as evident in the higher accuracy, OMCC and WAMCC corresponding to GO-I, GO-3, and GOA. This suggests that post-processing could remove some of the subcellular localization information pertaining to the raw GO vectors. GO-I achieves the best performance, suggesting that postprocessing is not necessary. Table 5.2 and Table 5.1 suggest that InterProGOSVM outperforms the amino-acid-composition methods and InterProGOSVM is also comparable, although a bit inferior, to PairProSVM.

5.1.3 Performance of Fusion Predictor

Table 5.3 shows the performance of fusing the InterProGOSVM and PairProSVM. The performance was obtained by optimizing the fusion weights w^{co} (based on the test dataset). The results show that the combination of PairProSVM and GO-IO (ISF with geometric mean) achieves the highest accuracy—79.04%, which

Method I	Optimal w		OMCC	WAMCC
GO-I	0.4490	78.91%	0.7680	0.7322
GO-2	0.2643	78.56%	0.7641	0.7260
GO-3	0.3970	78.75%	0.7662	0.7291
GO-4	0.3693	78.72%	0.7659	0.7285
GO-5	0.3711	78.78%	0.7666	0.7293
GO-6	0.3428	78.78%	0.7666	0.7294
	0.4263	78.81%	0.7670	o. 7289

Place the table heading above it.

GO-8	0.2947	78.40%	0.7624	0.7234
GO -9	0.4186	78.97%	o. 7687	0.7318
GO-IO	0.4515	79.04%	0.7694	0.7335
GO-II	0.3993	78.37%	o. 7620	o. 7222
GO-12	0.3670	78.62%	o. 7648	0.7263

Table 5.3: Performance of the fusion of InterProGOSVM and PairProSVM .

is significant better than PairProSVM (77.05%) and the InterProGOSVM method (72.21%) alone. The results also suggest that fusion of PairProSVM and any configuration of InterProGOSVM can outperform the individual methods. This is mainly because the information obtained from homology search and from functional domain databases has different perspectives and is therefore complementary to each other.

Use the adverb "significantly" before comparisons, i.e. significantly better, significantly stronger".

✓ *Uses "mainly" and "seems to" to show it is not certain*

Surprisingly, fusing the best performing InterProGOSVM and profile-alignment method does not give the best performance. And for different fusion methods, the best performance is achieved at different optimal $w^{c\ o}$. Since the performance of PairProSVM seems to be a bit better than that of InterProGOSVM, it is reasonable to give less weight to InterProGOSVM and more to PairProSVM.

Do not start a sentence with "And" better to use "In addition".

Do not use "bit". It is spoken English. Better to use "slightly".

5.1.4 Correlation between the Weighting Factor and Fusion Performance

As mentioned above, the w^{co} will significantly influence the final performance of each fusion method. It is necessary to discover how the parameter impacts

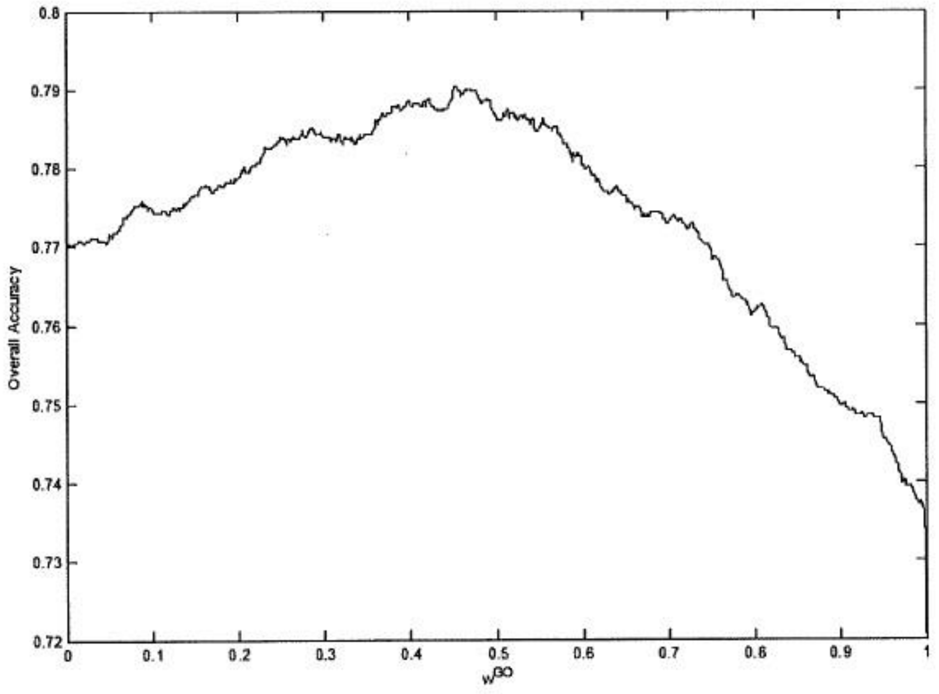


Figure 5.1: Performance of fusing of GO-IO and PairProSVIM using different fusion weight w^{c0}.

the accuracy of fusion methods. Here, we chose the fusion method with the best performance—GO-10. Fig. 5.1 shows the performance of fusing GO-IO and PairProSVM by varying w^{cO} from 0 to 1. As can be seen, the performance changes steadily with the change of w^{co} . It suggests that w ^{co} would not impact the final performance of the fusion method abruptly and the improvement of the fusion method over PairProSVM exists for a wide range of w ^{co} Further, to show that the improvement of the fusion methods over each individual method is statistically significant, we also calculated the p-value between them. The pvalue between the accuracy of the fusion system (GO-IO and PairProSVM) and the PairProSVM system is 0.0055, which suggests that the

💡 *Do not "would" when describing actual results. Use "will" to describe real results.*

✓ *Explains p-value by giving reasons and states result* performance of the fusion predictor is significantly better than that of the PairProSVM predictor.

GO Vector Construction Method	OMCC	WAMCC	ACC
1-0 value	0.9208	O. 9144	92.57%
	0.9401	0.9367	94.39%
ISF	0.9023	0.8965	90.84%
TF-ISF	0.9221	0.9181	92.70%

Table 5.4: Performance of different GO vectors construction methods without post-processing based on 5-fold cross validation on Chou's training dataset. Refer to Eqs. 4.5, Eqs. 4.7 and Eqs. 4.8 for the definition of Acc, OMCC and WAMCC. The higher these three evaluation measures, the better the performance.

5.2 Performance of GOAS VM Method

5.2.1 Performance of Different GO Vector Construction Methods and Normalization Methods

Table 5.4 shows the performance of the four GO vectors construction methods without post-processing based on five-fold cross validation using Chou's training dataset. Linear SVMs were used in all cases, and the penalty factor was set to 1. The results show that term-frequency (TF) performs almost 2% better than other three methods, which demonstrates that the frequencies of occurrences of GO terms could also provide information for subcellular locations. The results are also biologically relevant because proteins of the same subcellular localization are expected to have a similar number of occurrences of the same GO term. In this regard, the 1-0 value approach is inferior because it quantizes the number of occurrences of a GO term to 1. The results also suggest that inverse

✓ Links ideas well
a. explains the results
b. compares results
c. explains why they are important
d. highlights possible problem
e. suggests possible solution

sequence frequency (ISP) is detrimental to classification performance, despite its proven effectiveness in the field of document retrieval. We conjecture that this is because ISF could only take effect when at least most of the sequences have some identical GO terms but actually the occurrences of GO terms in different proteins are not so frequent. We have found that even for the most frequently appearing GO term, less than 1/4 (around 600) protein sequences have this GO term.

Donot use "actually" at the start of a sentence. Better to use "In fact" or place it next to the verb, i.e. "are not actually so frequent".

GO Vector Construction Method	Post-Processing	OMCC	WAMCC	ACC
1-0 value	None	0.9208	O. 9144	92.57%
	Vector Norm	0.9247	0.9163	92.94%
	Geometric Mean	0.9313	0.9260	93.56%
	None	0.9401	0.9367	94.39%
	Vector Norm	0.9322	0.9253	93.64%
	Geometric Mean	0.9300	0.9250	93.44%
ISP	None	0.9023	0.8965	90.84%
	Vector Norm	0.9181	0.9090	92.32%
	Geometric Mean	0.9234	0.9168	92.82%
TF-ISF	None	0.9221	0.9181	92.70%
	Vector Norm	0.9335	0.9271	93.77%
	Geometric Mean	0.9331	0.9283	93.73%

Table 5.5: Performance of different post-processing methods in GOASVM.

Table 5.5 shows the performance of applying three post-processing methods to the GO vectors constructed by four different methods. The results demonstrate that applying the post-processing can improve performance except for the TFconstructed GO vectors. Moreover, the improvement achieved by vector norm is not significantly different from that achieved by geometric mean. And among all the combinations, TF without any post-

Do not start the sentence with "And among." Place it at the end of the sentence, i.e. "TF without...performance among all the combinations."

processing achieves the best (94.39%) performance, which suggests that post-processing of GO terms of TF may remove some important information, thus leading to deteriorated accuracies.

5.2.2 Comparing with Methods Based on Other Features

Table 5.6 shows the performance of different features and different SVM classifiers. The penalty factor for training the SVMs was set to 1 for both linear SVMs and RBF-SVMs. For RBF-SVMs, the kernel parameter was set to 1. For the first four methods, Vector Norm was adopted for better classification performances. GapAA [43] takes the maximum gap length 48 (the minimum length of all the sequences is 50). As AA, PairAA and PseAA produce low-dimensional feature vectors, the performance achieved by RBF-SVMs is better than that achieved by linear SVMs. So we just present the performance of RBF-SVMs. As can be seen, amino-acid composition and its variant are not good features for subcellular localization.

✓ *Refers to table when describing results*

Classifier	Feature	Post-Processing	OMCC	WAMCC	ACC
RBF SVM RBF SVM		Vector Norm Vector Norm	0.3846 0.4119	0.3124 0.3342	42.30% 44.86
Linear SVM	AA G al)AA (48)	Vector Norm	0.4524	0.3797	% 48.66%
RBF SVM		Vector Norm	0.4185	0.3467	45.48%
Linear SVM	Profile vectors	Geometric Mean	0.5149	0.4656	54.52%
Linear SVM	GO vectors (GOASVM)	None	0.9401	0.9367	94.39%

Table 5.6: Performance of different features and different SVM classifiers. Features include amino acid composition (AA) [3], amino-acid pair composition (PairAA) [3], AA composition with gap (length = 48) (GapAA) [5], pseudo AA composition (PseAA) [7], and profile alignment scores [19].

The highest accuracy is only 48.66%. Moreover, although homology-based method can achieve better accuracy (54.52%) than amino-acid composition based methods, the performance is still very poor, probably because of the low sequence similarity in this dataset. On the other hand, our method can achieve a significantly better performance (94.39%), almost 40% (absolute) better than homology-based method. This suggests that functional-domain based method can provide significantly richer information pertaining to protein subcellular localization than the other methods. The high OMCC and WAMCC also suggest that GOASVM is capable of handling imbalanced classification problems.

Use plural form when more than one method is used, i.e. "homology-based methods" or use "the" if there was only one method, i.e. "the homology-based method"

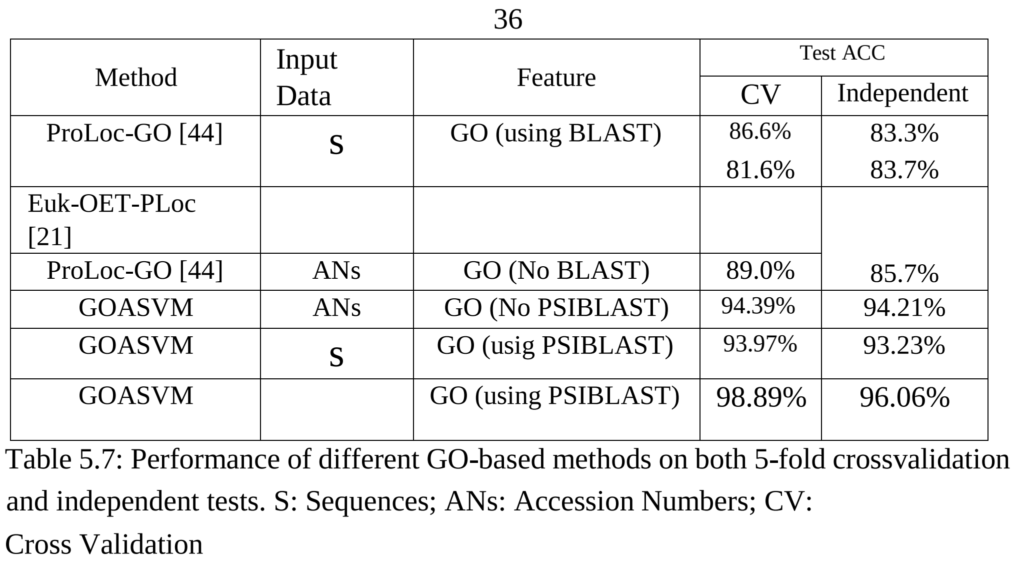
✓ *Use "Their performance" to clarify that "the poor performance" refers to homology-based methods.*

✓ *Avoid overusing linking expressions. "Moreover" can be omitted.*

5.2.3 Comparing with State-of-the-art GO Methods

Table-5.7 compares the performance of three state-of-the-art GO-based methods and the proposed GOASVM method based on 5-fold cross validation on the training dataset and using the whole training set for training and the independent test set for testing. Although the methods mentioned here all belong to GO-based methods, there still exist a lot of differences among them. As Euk-OET-PLoc [21] could not produce valid GO vectors for all proteins, it uses PseAA as a back-up method. Specifically, for those proteins that do not possess a valid GO term, EukOET-PLoc uses PseAA derived from the amino acid sequences of these proteins

✓ Clear statement of purpose and description of action taken



36

Method	Input Data	Feature	Test ACC	
			CV	Independent
ProLoc-GO [44]	S	GO (using BLAST)	86.6% 81.6%	83.3% 83.7%
Euk-OET-PLoc [21]				85.7%
ProLoc-GO [44]	ANs	GO (No BLAST)	89.0%	
GOASVM	ANs	GO (No PSIBLAST)	94.39%	94.21%
GOASVM	S	GO (usig PSIBLAST)	93.97%	93.23%
GOASVM		GO (using PSIBLAST)	98.89%	96.06%

Table 5.7: Performance of different GO-based methods on both 5-fold crossvalidation and independent tests. S: Sequences; ANs: Accession Numbers; CV: Cross Validation

for classification. ProLoc-GO [44] uses either the ANS of proteins as searching keys or uses the ANS of homologous proteins returned from BLAST as searching keys. Our proposed GOASVM can use ANS only, sequences only, or both ANS and sequences as inputs. Given a sequence, PSI-BLAST is used to find the remote homologs and the AN of the highest ranked homolog is used as the searching key. Unlike Euk-OET-PLoc and ProLoc-GO, GOASVM uses PSI-BLAST to find the top-ranked homology even if the protein has an AN. This strategy in fact leads to the best performance (the last 2nd row in Table 5.7).

Table 5.7 also shows that using ANS as input performs slightly better than using sequences as input. The result is biologically plausible because the homologous proteins are not identical to the query protein. As a result, their ANS only represent the close relatives of the query protein. If both ANS and amino acid sequences are available, we should make the best use of them. Our proposed GOASVM achieves this by producing multiple GO vectors based on the original ANS and the ANS of the

✓ *Refers to data in table (the last row in Table 5.7) to support conclusions drawn about results*

✓ *Uses clear topic sentence stating findings*

✓ *Explains findings*

✓ *Gives conclusion*

✓ *Explains consequences of findings*

✓ *Discusses consequences*

homologous proteins returned from PSI-BLAST. The result (last row of Table 5.7) suggests that this strategy can further increase the prediction accuracy to 98.89%, which represents a 4.92% improvement as compared to using the sequences only.

5.2.4 Performance of GOAS VM Using Old GOA Database

The newer the version of GOA database, the more annotation information the database contains. As a result, better performance is expected. So, to avoid taking advantages of using updated version of the GOA database, we performed experiments using the old version of the database and compared with other methods. Table 5.8 shows the performance of GOASVM using an earlier version of the GOA database. For comparison, we used the same version as that used by Euk-OET-PLoc [21] [1] . As ProLoc-GO [44] used a more recent version (released on 2007), we did not compare with it. We also used PseAA as our back-up method for those sequences that cannot generate valid GO terms. As can be seen from the table, our method significantly outperforms Euk-OET-PLoc, almost 7% (absolute) based on cross validation and by 6% (absolute) based on independent tests. From another perspective, it echoes our opinion that using newer versions of the GOA database can achieve better

Method	Input Data	Feature	Test ACC	
			CV	Independent
Euk-OET-PLoc [21]		GO (GOA2005)+PseAA	81.6%	83.7%
GOASVM		GO (GOA2005)+PseAA	88.16%	89.11%
GOASVM	ANs	GO (GOA2011)	94.39%	94.21%

Table 5.8: Performance of GOASVM using old version of GOA database on both 5-fold cross-validation and independent tests.

Put the main point first then the reason, i.e. "We did not compare it as ProLoc-Go…"

✓Links analysis of results when possible to main objective of research (determination of the sub-cellular location of proteins).

Do not use suggest when the findings are clear. Use stronger language, i.e. "strongly suggest" "show" "demonstrates" performance than using older versions (94.39% vs. 88.16%). This suggests that the annotation information is very important to the determination of subcellular locations for proteins.

[I] Actually, we used the version released on 21-0ct-2005, while [21] used the version released on 21-Nov-2005. So strictly speaking, our version was even a bit older

Chapter 6

The discussion section needs to try and explain more about what the findings mean. There could be a discussion of
a. how the results will benefit research. For example, it will help drug companies.
b what problems or difficulties that may have effected results or scope of research
c How the research findings align with results from other research on the topic.
Generally, discussions are ordered in the following way
i. specific findings of the study
ii. how the findings fit with studies/methods discussed in the literature review (6.3))
iii. how the findings relate to the general background to the topic

6 Discussion

6.1 Weighting Factor in the Fusion Method

In the fusion method, Eq. 2.15 suggests that the weighting factors w^{GO} and W^{PA} can influence the fused scores, which in turn affect the performance of fusing the InterProGOSVM methods and PairProSVM. To determine the best weighting factor, we have swept w^{c0} from 0 to 1 at interval of 0.0001.

First provide a brief overview of what the discussion section focuses on.

✓ *Avoid vague terms such as "some". Use "a number of".*

There are some methods that can obtain the optimal fusion weighting factors. Pigeon et al. [45] proposed applying Logistic Regression (LR) to fuse the scores obtained by multiple-feature based systems. The fusion weights can be trained to optimize an objective function based on the training scores. The determination of optimal weighting factors is beyond our focus. More information can be found in [45].

✓ *Refers to other studies to compare possible result*

6.2 Weighting Factor in the GOAS VM Method

In the GOASVM method, the weighting factors (wj in Eq. 3.1) should depend on the remoteness of the homologs. The more remote the homolog is from the test protein, the smaller the weighting factor should be.

As for the case of using both ANS and sequences, since the true ANS will undoubtedly generate more reliable and informative GO terms for the test proteins, larger weights should be given to the SVM scores obtained by ANs-based method and smaller weights for those obtained by homologous sequences. Besides, since we can adjust the E-value and other parameters in the PSI-BLAST, the number of homologs generated by PSI-BLAST can also be controlled.

Put the main point first, i.e. "Larger weights should…in the case of …since…"

✓ *Avoid using two linking expressions together (Besides, since), i.e. "Besides, the number of homologs generated by PSI-BLAST can also be controlled since we can adjust the E-value and other parameters in the PSI-BLAST.*

6.3 Relationship Between the Fusion Method and the GOAS VM Method

We have learned from Chapter 2 and Chapter 3 that GO vectors construction, post-processing, and classification of the InterProGOSVM and the GOASVM methods are the same. The only difference between these two kinds of methods is the way of retrieving the GO terms. Specifically, the InterProGOSVM methods uses InterProScan to search the GO terms, whereas GOASVM uses the ANS of sequences as the keys to search against the GOA database.

Both methods are based on the same notion—taking the advantages of functionaldomain based methods and overcoming their disadvantages. One advantage of functional-domain based methods is that their performance can be significantly better than other

✓ Refers to work in previous chapters, develops clear structure and contextualises present discussion

✓ Provides clear explanation for similarities and differences between two methods

✓ Signals content of discussions (how two proposed methods address disadvantage of functional domain methods)

methods. But the disadvantage is that they are not applicable to all the protein sequences because some sequences may not have any annotated GO terms. The two proposed methods attempt to surmount the disadvantage from different perspectives.

The fusion of InterProGOSVM and PairProSVM attempts to use homologybased method to complement the functional-domain based method. This idea is of course one of the easiest ways we can come up with. And the results suggest that this idea is feasible and successful as we initially expected.

💡 *Avoid multi-word verbs such as "come up with" and "turn out". Better is "propose" and "resulted in a ..."*

The GOASVM method, on the other hand, uses PSI-BLAST to determine the homologous proteins and then retrieve the GO terms using the ANS of the homologs. This idea turns out to be much more efficient and achieves significantly better performance than the fusion method of InterProGOSVM methods and PairProSVM. Moreover, we are also able to determine how many homologs should be used for the best classification. This is also a complementary method, but it 'replaces' the protein sequences rather than the algorithm. Therefore, instead of using homology-based methods as a backup or as a complement, GOASVM uses the homologs of the query sequence as a backup and complement.

✓ *Clearly summarises section with strong summary sentence, "Therefore"*

Chapter 7

💡 *The conclusion chapter can be short and limited to summarizing the main findings and a discussion of future work. The details of the findings have already been described and discussed in chapter 5 and 6.*

✓ *Introduce the focus of this chapter before starting the discussion.*

7 Conclusions and Future Works

7.1 Conclusions

This report investigates two approaches to exploiting gene ontology (GO) for subcellular localization prediction.

In the first approach, namely InterProGOSVM, gene ontology (GO) vectors are produced by presenting protein sequences to InterProScan and considering the GO terms as the axes of a high-dimensional Euclidean space and the existence or number of occurrences of GO terms as coordinates. The GO vectors are further post-processed by normalizing with their vector norm or by the geometric mean of the pairwise dot products. The post-processed vectors are then classified by linear SVMs, and the SVM scores are further combined with those of profile-alignment SVMs to boost prediction accuracy. Results show that homology-based methods that exploit sequence and profile similarities and functional-domain based methods that exploit the GO annotations consider the subcellular localization problem from different perspectives, thus providing significant complementary information for enhancing classification performance. This paper also demonstrates that these two types of methods are far more advantageous than the amino-acid composition based methods.

✓ *Clearly summarises main research in one sentence.*

✓ *Summarises objective of report to contextualize conclusions drawn*

✓ *Reminds reader of details of each approach and results obtained before stating conclusions about findings of study*

💡 *Avoid unclear and vague language such as "different perspectives" "providing …information". Key finding is last sentence so place it at the start of the paragraph.*

In the second approach, namely GOASVM, the accession numbers (ANs) of query proteins are used as keys to search against the Gene Ontology Annotation (GOA) database to find the associated GO terms of the proteins. For proteins without an AN, PSI-BLAST is used to find their homologs and the ANS of these homologs are used as the searching keys. Then, GO vectors are constructed similar to InterProGOSVM methods. Results on a recent dataset demonstrate that GOASVM outperforms the state-of-the-art GO-based methods and amino-acid composition based methods. It was also found that the frequency of occurrences of GO terms (term-frequency) provides useful information for classification, leading to around 2% relative improvement in prediction accuracy. Finally, this study demonstrates that even if the AN of a query protein is known, it is still beneficial to use the ANS of its homologs together with the known AN to construct multiple GO vectors of the query protein.

✓Uses past simple tense to describe details of study, i.e. "was found" uses present simple tense to describe things that are general facts. "It is beneficial to"

The future work section is often part of the discussion chapter.

7.2 Future Works

💡 Use "Future Work". "Future Works may mean future publications or works of art.

✓ Use a clear subtitle "Limitations" for section 1 and 2. This could also appear in

About these two proposed methods, some further research has also been mentioned in Chapter 6, such as how to determine the optimal weighting factor, etc. But more attention should be paid to the works stated as following:

💡 Avoid using "etc" as it is vague, and avoid starting sentences with "but", i.e. "Future research should focus on the following aspects."

1. Feature Selection. According to the experiments and results, we know that the numbers of distinct GO terms can be as many as several thousands. For example, GOASVM can provide 5450 distinct GO terms. This means that each GO vector is of 5450-dimension. It may become extremely large when

the size of the dataset increases. In this case, the computational costs would become prohibitively expensive. But actually, not all GO terms have equally significantly influence on classification. Quite a few GO terms have little contribution to the final prediction performance. So, reducing the redundant GO terms may in turn improve the overall accuracy of the prediction. Therefore, it is also highly required to select the relevant GO terms and disregard the redundant ones.

2. Multi-label Problem. The two proposed methods in this report can only deal with single-label subcellular localization problems. This means the two methods are based on the assumption that a protein is only located in one subcellular location. But in fact, there exist multiplex proteins that can exist in or move among two or more subcellular locations. For example, according to the work of Huh et al [46], the protein with AN 'YBR156C' can be located either in 'microtubule' or 'nucleus'. Proteins with multiple locations are particularly interesting because they may have some very special biological functions intriguing to investigators in both basic research and drug discovery [47]. Therefore, we would like to apply our proposed methods, especially GOASVM method into solving multi-label problems in the future.

✓ Provides clear explanation and rationale for each of aspects selected for further research.

💡 Also have the last sentences of section 1 and 2 at the start of the paragraph as they are the main points. It is easier of the reader to follow.

3. Protein-Protein Interaction. Recent studies [48] have investigated approaches for predicting subcellular localization by utilizing large-scale protein-protein interaction (PPI) networks and have shown that PPI networks can provide accurate localization predictions even without relying on common protein characteristics such as amino acid composition, protein motifs or physio-chemical properties. To interact physically, two proteins must localize

✓ Gives reference to other studies and uses present perfect tense to do so e.g. {have investigated” “have shown”

💡 Do not use “on the other hand” in this situation. It means contrast, here it is better to use “In addition”.

to the same or adjacent cellular compartments, suggesting that interaction may serve as an indicator for subcellular localization. On another hand, the recent availability of large PPI networks in yeast, worm and human [49] [50] makes it possible to utilize PPI for protein subcellular localization. Recent years have also witnessed the exponentially growing PPI measurements. Given these developments, PPI networks have become a basic feature available for many proteins. It is therefore of significant interest to find out whether, and to what extent, PPI can be used in the prediction of subcellular locations. Thus, PPI is also one of our priorities in further research.

Use "References" as the heading for this section, as this is generally applicable to most research papers.

A reference list only includes the sources cited in-text in the report.

A bibliography includes sources cited in-text and all other sources read to inform your research, even though some of these sources you have used may not have been cited in-text.

Bibliography

[1] T. Kleffmann, D. Russenberger, A. von Zychlinski, W. Christopher, K. Sjolander, W. Gruissem, and S. Baginsky, ('The arabidopsis thaliana chloroplast proteome reveals pathway abundance and novel protein functions," Current Biology, vol. 14, no. 5, pp. 354-362, 2004. 2

[2] G. Lubec, L Afjehi-Sadat, J. W. Yang, and J. P. John, "Searching for hypothetical proteins: theory and practice based upon original data and literature," Prog. Neurobiol, vol. 77, pp. 90-127, 2005. 2

[3] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," J. Mol. Biol., pp. 54—61, 1994, 238. 2, 29, 30, 36

[4] K.C. Chou and YD. Cai, "Predicting protein localizaiton in budding yeast," Bioinformatics, pp. 944—950, 2005, 21. 2

[5] K.J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines suing compositions of amiño acid and amino acid paris," Bioinformatics, pp. 1656-1663, 2003, 19. 2, 3, 29, 30, 36

[6] K. Y. Lee, D. W. Kim, D. K. Na, K. H. Lee, and D. H. Lee, "PLPD: reliable protein localization prediction from imbalanced and overlapped datasets," Nucleic Acids Research, vol. 34, no. 17, pp. 4655—4666, 2006. 2

[7] K.C. Chou, "Prediction of protein cellular attributes using pseudo-amino-acidcomposition," Proteins, pp. 246—255, 2001, 43. 3, 29, 30, 36

[8] K. Nakai, "Protein sorting signals and prediction of subcellular localization," Advances in Protein Chemistry, vol. 54, no. 1, pp. 277—344, 2000. 3

[9] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in gram-negative bacteria," Proteins: Structure, Function, and Genetics, vol. 11, no. 2, pp. 95-110, 1991. 3

[10] P. Horton, K. J. Park, T. Obayashi, and K. Nakai, "Protein subcellular localization prediction with WOLF PSORT," in Proc. 4th Annual Asia Pacific Bioinformatics Conference (APBC06), 2006, pp. 39-48. 3

[11] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "A neural network method for identification of prokaryotic and eukaryotic signal perptides and prediction of their cleavage sites," Int. J. Neural Sys., vol. 8, pp. 581—599, 1997. 3

[12] H. Nielsen, S. Brunak, and G. von Heijne, "Machine learning approaches for the prediction of signal peptides and other protein sorting signals," Protein Eng., vol. 12, no. 1, pp. 3—9, 1999. 3

[13] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," J. Mol. Biol., vol. 300, no. 4, pp. 1005-1016, 2000. 3, 4

[14] R. Mott, J. Schultz, P. Bork, and C.P. Ponting, "Predicting protein cellular localization using a domain projection method," Genome research, vol. 12, no. 8, pp. 1168—1174, 2002. 3

[15] M.S. Scott, D. Y. Thomas, and M. T. Hallett, "Predicting subcellular localization via protein motif co-occurrence," Genome research, vol. 14, no. 10a, pp. 1957—1966, 2004. 3

[161 R. Nair and B. Rost, "Sequence conserved for subcellular localization," Protein Science, vol. 11, pp. 2836-2847, 2002. 3

[17] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner, "Predicting subcellular localization of proteins using machine-learned classifiers," Bioinformatics, vol. 20, no. 4, pp. 547—556, 2004. 3

[18] J. K. Kim, G. P. S. Raghava, S. Y. Bang, and S. Choi, "Prediction of subcellular localization of proteins using pairwise sequence alignment and support vector machine," Pattern Recog. Lett., vol. 27, no. 9, pp. 996—1001, 2006. 3

[19] M.W. Mak, J. Guo, and S.Y. Kung, "PairProSVM: Protein subcellular localization based on local pairwise profile alignment and SVM," IEEE/ACM Trans. on Computational Biology and Bioinformatics, vol. 5, no. 3, pp. 416—422, 2008. 3, 14, 27, 30, 36

[20] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," Nucleic Acids Res., vol. 25, pp. 3389—3402, 1997. 4, 20

[21] K. C. Chou and H. B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers," J. of Proteome Research, vol. 5, pp. 1888-1897, 2006. 4, 24, 25, 36, 37, 38

[22] K.C Chou and H.B Shen, "Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization," Biochemical and Biophysical Research Communications, pp. 150-157, 2006, 347. 4

[23] W.L. Huang, C.W. Tung, S.W. Ho, S.F. Hwang, and S.Y. Ho, "ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization," BMC Bioinformatics, 2008. 4

[24] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," J. Mol. Biol., vol. 215, pp. 403—410, 1990. 4

[25] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP, and related tools," Nature Protocols, vol. 2, no. 4, pp. 953—971,

[26] W. Wang, M. W. Mak, and S. Y. Kung, "Speeding up subcellular localization by extracting informative regions of protein sequences for profile alignment," in Proc. Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'IO), 2010, pp. 147—154. 4

[27] N. J. Mulder and R. Apweiler, '(The InterPro database and tools for protein domain analysis," Current Protocols in Bioinformatics, vol. 2, no. 7, pp. 1—18, 2008. 5

[28] The Gene Ontology Consortium, "Gene ontology tool for the unification of biology," Nat. Genet., vol. 25, pp. 25-29, 2000. 7

[29] K.C. Chou and H.B Shen, "Euk-mPloc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," Journal of Proteome Research, pp. 1728-1734, 2007, 6. 8

[30] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF term weights as making relevance decisions," ACM Transactions on Information Systems, vol. 26, no. 3, pp. 1-37, 2008. 10, 11

[31] V. N. Vapnik, "The nature of statistical learning theory," in Springer Verlag, 2000. 13

[32] H. Rangwala and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition," Bioinformatics, vol. 21, no. 23, pp. 4239—4247, 2005. 14

[33] S.F. Altschul, T.L. Madden, A.A. Schafer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database serarch programs," Nucleic Acids Res., vol. 25, pp. 3389—3402, 1997. 14

[34] E. Camon, M. Magrane, D. Barrel, D. Binns, W. Fleischnann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, and A. Cox, "The Gene ontology Annotation (GOA) project:implementation of GO in SWISS-PROT, 'IYEMBL and InterPro," Genome Res., vol. 13, pp. 662-672, 2003. 17

[35] D. Butler, "NIH pledges cash for global protein database," Nature, vol. 419, no. 101, 2002. 17

[36] B. Boeckmann, A. Bairoch, R. Apweiler, M. •C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISSPROT protein knowledgebase and its supplement TrEMBL in 2003," Nucleic Acids Res., vol. 31, pp. 365-370, 2003. 17

[37] C. H. Wu, H. Huang, L. S. Yeh, and W. C. Barker, "Protein family classification and functional annotation," Comput. Biol. Chem., vol. 27, pp. 37—47, 2003. 17

[38] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, and P. Bork, "The InterPro Database, 2003 brings increased coverage and new features," Nucleic Acids Res., vol. 31, pp. 315—318, 2003. 18

[39] Y. Huang and Y. D. Li, "Prediction of protein subcellular locations using fuzzy K-NN method," Bioinformatics, vol. 20, no. 1, pp. 21—28, 2004. 24

[40] G. L. Wang, Jr. Dunbrack, and R. L. PISCES, "A protein sequence culling server," Bioinformatics, vol. 19, pp. 1589—1591, 2003. 26

[41] S. Y. Mei, W. Fei, and S. G. Zhou, "Gene ontology based transfer learning for protein subcellular localization," BMC Bioinformatics, 2011. 26

[42] B. W. Matthews, "Comparison of predicted and observed secondary structure of t4 phage lysozyme," Biochem. Biophys. Acta, vol. 405, pp. 442—451, 1975. 27

[43] K. J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs," Bioinformatics, vol. 19, no. 13, pp. 1656-1663, 2003. 35

[44] W. L. Huang, C. W. Tung, S. W. Ho, S. F. Hwang, and S. Y. Ho, "ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization," BMC Bioinformatics, 2008. 37, 38

[45] S. Pigeon, P Druyts, and P. Verlinde, "Applying logistic regression to the fusion of the NIST'99 1-speaker submissions," Digital Signal Processing, 2000. 39

[46] W. K. Huh, J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea, "Global analysis of protein localization in budding yeast," Nature, vol. 425, pp. 686-691, 2003. 42

[47] K. C. Chou and H. B. Shen, "Euk-PLoc: A fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," Journal of Proteome Research, vol. 6, pp. 1728-1734, 2007. 43

[48] K. Y. Lee, H. Y. Chuang, A. Beyer, M. K. sung, W. K. Huh, B. Lee, and T. Ideker, "Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species," Nucleic Acids Research, vol. 36, no. 20, pp. e136, 2008, 43

[49] A. R. Mendelsohn and R. Brent, "Protein interaction method—toward an edngame," Science, vol. 284, pp. 1948-1950, 1999. 43

[50] P. Uetz, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, and P. Pochart, "A comprehensive analysis of proteinn-protein interaction in Saccharomyces cerevisiae," Nature, vol. 403, pp. 623—627, 2000. 43

Appendix A

AUTHOR'S PUBLICATION

1. S. B. Wan, M. W. Mak, and S. Y. Kung, "Protein Subcellular Localization Prediction Based on Profile Alignment and Gene Ontology" , 2011 IEEE International Workshop on Machine Learning for Signal Processing (MLSP'2011), Beijing, China, sept. 2011

2. S. B. wan, C. Yao, Y. X. Hu, G. M. Zhang , "A Method of Continuous Data Flow Embedded within Speech Signals", The 2nd International Conference on Signal Acquisition and Processing (ICSAP'2010), Bangalore,
India, Feb. 2010