AN EXTENSION TO GOPRED TO ANNOTATE SWISS-PROT AND TREMBL SEQUENCES FOR ALL GENE ONTOLOGY CATEGORIES AND EC NUMBERS

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

BY

AHMET SÜREYYA RİFAİOĞLU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

FEBRUARY 2015

Approval of the thesis:

AN EXTENSION TO GOPRED TO ANNOTATE SWISS-PROT AND TREMBL SEQUENCES FOR ALL GENE ONTOLOGY CATEGORIES AND EC NUMBERS

submitted by **AHMET SÜREYYA RİFAİOĞLU** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** ________________

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

Assoc. Prof. Dr. Rengül Çetin-Atalay
Co-supervisor, **Informatics Institute, METU** ________________

Examining Committee Members:

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

Prof. Dr. Mehmet Volkan Atalay
Computer Engineering Department, METU ________________

Assoc. Prof. Dr. Tolga Can
Computer Engineering Department, METU ________________

Assist. Prof. Dr. Ömer Sinan Saraç
Computer Engineering Department, İstanbul Technical University ________________

Dr. Tunca Doğan
European Bioinformatics Institute, University of Cambridge ________________

Date: ________________

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

Name, Last Name: AHMET SÜREYYA RİFAİOĞLU

Signature :

ABSTRACT

AN EXTENSION TO GOPRED TO ANNOTATE SWISS-PROT AND TREMBL SEQUENCES FOR ALL GENE ONTOLOGY CATEGORIES AND EC NUMBERS

RİFAİOĞLU, Ahmet Süreyya
M.S., Department of Computer Engineering
Supervisor : Prof. Dr. Mehmet Volkan Atalay
Co-Supervisor : Assoc. Prof. Dr. Rengül Çetin-Atalay

February 2015, 59 pages

Traditional methods cannot keep up with annotating proteins as the number of proteins whose sequences known is increasing exponentially. For this reason, automated protein annotation became an important research area in bioinformatics. In this thesis, GOPred method is extended to annotate Swiss-Prot and TrEMBL sequences for all Gene Ontology (GO) categories and EC Numbers. GOPred consists of SPMap, Blast-*k*NN and Pepstats methods which are subsequence, similarity and feature based methods, respectively. Previous version of GOPred method for functional classification of proteins was used for 300 molecular function Gene Ontology (GO) terms. In this study, improved system is trained for 514 molecular function GO terms, 2909 biological process GO terms and 438 cellular component GO terms. The system is also applied on functional prediction of enzymes for 851 Enzyme Commission (EC) Numbers. Each term is trained as a separate classifier with its own training data. All Swiss-Prot annotations that have experimental evidences are used in data preparation of terms. GOPred gives three scores for each classification method, when a sequence is given as input. Subsequently, obtained scores are combined and a weighted mean score is calculated. Since performances of terms are different, we used a new method to calculate optimal decision threshold for each term and only the predictions whose weighted mean scores are over the determined thresholds are presented. Performance of each term is measured separately and their average is calculated for each GO category and EC. F-Score values are calculated as 0.86, 0.75 and 0.80 for molecular function, biological process and cellular component categories of GO, respectively. F-Score value is 0.96 for EC. To the best of our knowledge, this is the best performance achieved for EC number prediction in the literature. GO term prediction results show that the performance of our system is better for prediction of multi-functional proteins. We also showed that combination of different classification methods enhances the prediction results. Finally, improved system is tested on about 58 million TrEMBL proteins. Predictions that are given by the improved system are compared with the annotations of TrEMBL reference systems which are EMBL, HAMAP, PDB, PIR, PIRNR and RuleBase. Results are consistent with the annotations of TrEMBL reference systems.

Keywords: Protein Function Prediction, Gene Ontology Term, Enzyme Commission Numbers, Decision Threshold

ÖZ

TÜM GEN ONTOLOJİSİ VE EC NUMARALARI İÇİN SWISS-PROT VE TREMBL DİZİLERİNİ ANLAMLANDIRMAK AMACIYLA GOPRED YÖNTEMİNİN GENİŞLETİLMESİ

RİFAİOĞLU, Ahmet Süreyya
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Mehmet Volkan Atalay
Ortak Tez Yöneticisi : Doç. Dr. Rengül Çetin-Atalay

Şubat 2015 , 59 sayfa

Dizisi bilinen proteinlerin sayısı üstel olarak arttığı için geleneksel yöntemler, dizisi bilinen proteinlerin anlamlandırılmasında yetersiz kalmıştır. Bu yüzden, proteinlerin otomatik olarak anlamlandırılması biyoenformatik alanında önemli bir araştırma konusu olmuştur. Bu tezde, tüm Gen Ontolojisi (GO) kategorileri ve EC numaraları için Swiss-Prot ve TrEMBL dizilerini anlamlandırmak amacıyla GOPred yöntemi genişletilmiştir. GOPred yöntemi sırasıyla altdizi, benzerlik ve özellik tabanlı olan SPMap, Blast-*k*NN ve Pepstats yöntemlerinden oluşmaktadır. GOPred yöntemini önceki versiyonunda 300 moleküler işlev GO terimleri için protein işlev sınıflandırılması yapılmıştır. Bu çalışmada, geliştirilen sistem 514 moleküler işlev, 2909 biyolojik süreç ve 438 hücresel bileşen GO terimleri için eğitilmiştir. Sistem ayrıca, 851 Enzim Komisyonu (EC) numarası için enzimlerin işlev tahminine uygulanmıştır. Her terim kendi eğitim verileri kullanılarak ayrı sınıflandırıcılar olarak eğitilmiştir. Terimlerin eğitim verilerinin hazırlanması, Swiss-Prot veritabanındaki deneysel kanıtlara dayanan protein anlamlandırmaları kullanılarak hazırlanmıştır. GOPred'de kullanılan sınıfladırma yöntemleri girdi olarak verilen her diziye üç sonuç vermektedir. Daha sonra elde edilen sonuçlar birleştirilerek tek bir ağırlıklı sonuç hesaplanmaktadır. Terimlerin başarım değerleri farklı olduğu için, eğitilen her terim için en uygun karar eşik değeri, yeni bir yöntem kullanılarak hesaplanmıştır. Sadece ağırlıklı sonuç değerleri belirlenen eşik değerlerinin üzerinde olan tahminler sunulmuştur. Terimlerin başarım değerleri ayrı olarak ölçülmüş ve her GO grubu ve EC için ortalama başarım değerleri hesaplanmıştır. F-ölçütü değerleri moleküler işlev GO terimleri, biyolojik süreç GO terimleri ve hücresel bileşen GO terimleri için sırasıyla 0.86, 0.85 ve 0.80 olarak hesaplanmıştır. EC için F-ölçütü değeri 0.96 olarak hesaplanmıştır. Bildiğimiz kadarıyla, bu sonuç EC numaraları tahmini konusunda kaynaklar içerisinde elde edilmiş en iyi sonuçtur. GO terimleri tahmin sonuçları, sistemin başarımının çok işlevli proteinlerde daha iyi olduğunu gösteriyor. Ayrıca farklı sınıflandırma yöntemlerinin birleştirilmesinin tahmin sonuçlarını iyileştirdiği gösterilmiştir. Son olarak, geliştirilen sistem yaklaşık 58 milyon TrEMBL proteinleri için denenmiştir. Geliştirilen sistemin verdiği tahminler, TrEMBL için protein anlamlandırması yapan yapan EMBL, HAMAP, PDB, PIR, PIRNR ve RuleBase referans sistemleriyle karşılaştırılmıştır. Sonuçlar, TrEMBL referans sistemlerinin sonuçlarıyla örtüşmüştür.

Anahtar Kelimeler: Protein İşlev Tahmini, Gen Ontolojisi Terimleri, Enzim Komisyon Numaraları, Karar Eşik Değeri

To my family

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my supervisor, Dr. Volkan Atalay, for his guidance, positive attitude and patience during my master thesis. I am also grateful to Dr. Rengül Atalay, my co-supervisor, for encouraging and motivating me whilst helping me to develop my background in biological aspects of this study. Their guidance helped me in all the time of this research and writing of this thesis.

I am extremely indebted to Dr. Sinan Saraç, for his valuable advice, constructive criticism and his extensive discussions around this thesis and GOPred. I also want to thank Dr. Maria Martin and Dr. Tunca Doğan for inviting me European Bioinformatics Institute for three months and providing necessary infrastructure and resources to accomplish my research work. It was a wonderful experience for me to be part of UniProt Automatic Annotation Team.

I would like to thank my friends Abdullah Doğan, Aybike Şimşek and Murat Öztürk for all the fun times that we have spent in METU-CENG. I am especially thankful to Murat Öztürk for his comments about my presentation and sharing his Latex knowledge.

I am especially grateful to my parents and sisters for their constant support and love throughout my life. I also want to thank my dearest wife, İnci, for sharing this life with me and standing by me in good and bad times. Her support and understanding made this thesis possible. Thank you for giving me your endless love. My son, Semih, always made me smile when I was overstressed. He taught me how to work effectively in short period of times. I hope you will have a happy and successful life.

TABLE OF CONTENTS

CHAPTERS

APPENDICES

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

GO	Gene Ontology
MF	Molecular Function
BP	Biological Process
CC	Cellular Component
EC	Enzyme Comission
SVM	Support Vector Machines
EMBOSS	The European Molecular Biology Open Software
EBI	European Bioinformatics Institute

CHAPTER 1

INTRODUCTION

Proteins are building blocks that play important role in every process in living cells. Proteins are made of a chain of amino acids which are organic compounds that constitute the basic structures of various biological components. Each amino acid has its own chemical properties. There are 20 different amino acids. Information regarding to amino acids is given in Appendix A.1.

Amino acids form peptide bonds, which are chemical bonds that occur by formation of two amino acids. Polypeptide bonds are constructed by the combination of peptide bonds. Proteins consist of one more polypeptide bonds. Therefore, proteins are polymers whereas amino acids are monomers. Each protein has its own order of amino acids, which is called the "sequence" of that protein. Sequence of a protein can be considered as a string consisting of 20 letters. The order of amino acids in the sequence determines the three dimensional (3-D) structure and functions of corresponding protein. There are various types of proteins such as enzymes, antibodies, storage proteins, transport proteins, hormonal proteins etc. Proteins play critical roles such as DNA replication, identification of the molecules going in/out and forming other macromolecules. Therefore, they are basically involved in all of the functions occur within cells.

There are four levels of proteins structure that are primary, secondary, tertiary and quaternary structure. Primary structure of a protein is a chain of amino acids. Secondary structure of proteins occurs by the formation of hydrogen bonds between amino acids. Tertiary structure is formed when all secondary structure items are folded together to construct 3-D structure of proteins. Quaternary structure occurs when a protein has more than one amino acid chain. Quaternary structure of a protein determines its function. Structure of protein 1TUI is represented in Figure 1.1.

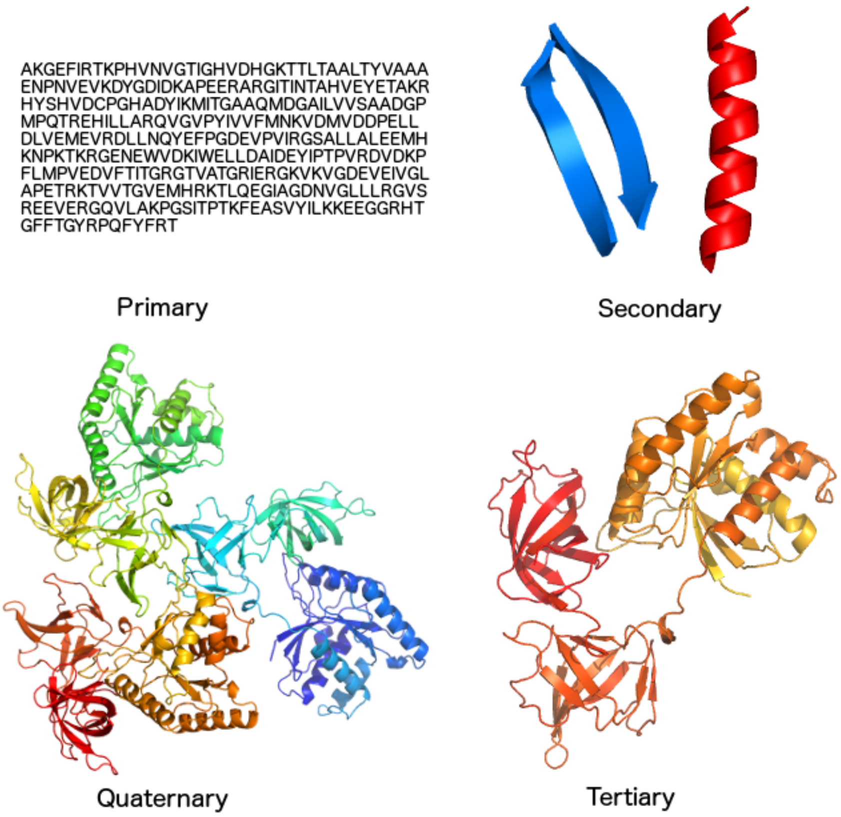

AKGEFIRTKPHVNVGTIGHVDHGKTTLTAALTYVAAA
ENPNVEVKDYGDIDKAPEERARGITINTAHVEYETAKR
HYSHVDCPGHADYIKMITGAAQMDGAILVVSAADGP
MPQTREHILLARQVGVPYIVVFMNKVDMVDDPELL
DLVEMEVRDLLNQYEFPGDEVPVIRGSALLALEEMH
KNPKTKRGENEWVDKIWELLDAIDEYIPTPVRDVDKP
FLMPVEDVFTITGRGTVATGRIERGKVKVGDEVEIVGL
APETRKTVVTGVEMHRKTLQEGIAGDNVGLLLRGVS
REEVERGQVLAKPGSITPTKFEASVYILKKEEGGRHT
GFFTGYRPQFYFRT
Primary
Secondary
Quaternary
Tertiary

Figure 1.1: Protein structure of 1TUI

Enzymes are catalysts for almost all of the biological reactions. They accelerate biological reactions providing a different pathway by lowering activation energy. Each enzyme has a part named active site, which has special formation and functional groups. Substrates are specific types of molecules that are bound to active sites of an enzyme to start enzymatic reactions. Products and enzyme itself are the outputs of enzymatic reactions. The structure of enzymes does not change after reactions. Therefore, it can be used for other enzymatic reactions later on. An illustration of an enzymatic reaction is given in Figure 1.2. Enzymes are also proteins. The function of an enzyme is highly dependent on the order of amino acid sequence of corresponding enzyme.

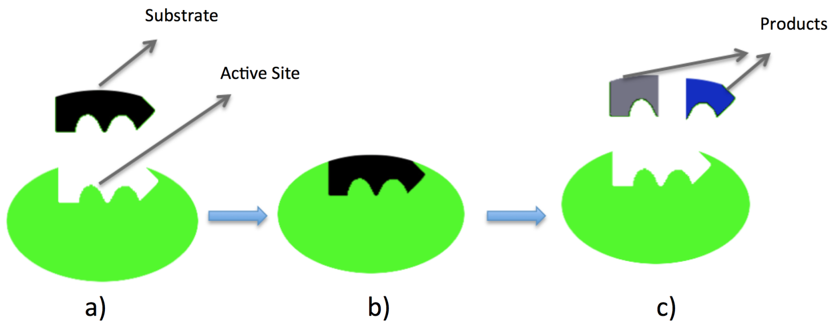


Figure 1.2: An enzyme catalyzes a reaction and two products are created.

1.1 Problem Definition

Proteins generally have more than one function. Literature is abundant in terms of experiments and studies to determine the functions of proteins. In addition, there are many biocurators who examine the published material to find out functions of proteins. In recent years, protein sequences of many organisms have been extracted. However, the functions of the proteins cannot be determined at the same time the sequences are found. Functions of only a small proportion of the proteins have been determined by experiments. Since there is massive data and it is exponentially growing, it is almost impossible to annotate protein functions by traditional ways such as experiments and literature search.

Recently, protein function prediction is emerged as an important research area whose aim is to determine functions of proteins by computational methods using the distinctive features of the proteins such as motifs, domains, sequence similarities, physical and chemical structures, protein-protein interactions etc. Most of the protein function prediction methods are based on ontologies such as Gene Ontology and Enzyme Commission Numbers.

Protein function prediction is a challenging and important problem for many reasons. First, all functionalities of a protein cannot be determined by a single experiment and a protein may have more than one function. Besides, biocurators may misinterpret the literature information and erroneously annotate protein functions. Some functions of proteins may not be observable, when the experiments are performed. Therefore, protein function prediction methods can be used as a guide to conduct expensive and sensitive experiments.

1.2 Extensions

In this study, previously developed protein function prediction method GOPred [1] is extended to all GO categories with modifications. Method is also applied on classification of Enzyme Commission (EC) Numbers. In addition, hierarchical evaluation of predictions is proposed for EC numbers. Extensions can be summarized as follows:

- A restriction rule, based on number of protein's annotations is added for negative data preparation of GO term prediction in order to improve the performance.
- Number of MF GO terms is extended from 300 to 514 GO terms. Biological Process and Cellular Component aspect of GO terms are added.
- A balancing method is applied to calculate optimal decision thresholds for GO terms instead of giving a single global threshold.
- EC number prediction is performed using the same classification methods with data preparation method proposed in [2].
- Fourth-level EC classification is also included for training.
- A hierarchical evaluation method is proposed to present predictions for EC classification.
- The system is trained with new extensions.

CHAPTER 2

BACKGROUND INFORMATION

2.1 UniProt and UniProtKB

UniProt is the comprehensive protein sequence and annotation database. The main objective of UniProt is to create reliable, comprehensive and qualified protein databases using well-defined curation methods that is followed by experienced biocurator teams. The correctness of each entry is verified manually and literature is searched continuously to check if there are erroneous or conflicting information. In addition, databases are updated periodically to reflect recent changes. UniProt GO annotation (UniProt-GOA) is a program that provides high-quality electronic and manual annotations for proteins in UniProt Knowledgebase (UniProtKB) [3].

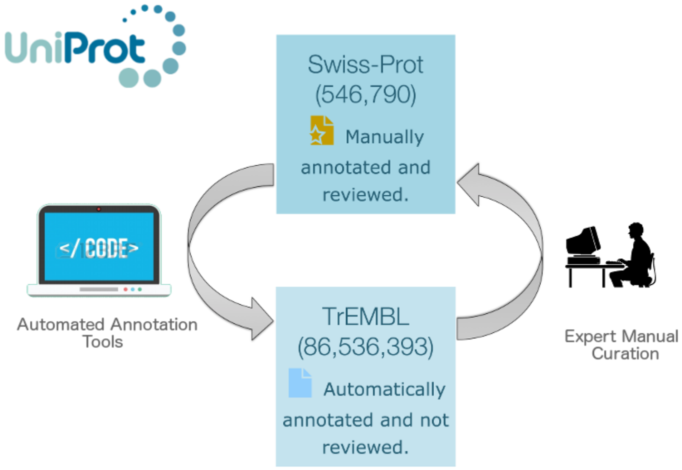

UniProt
Swiss-Prot (546,790)
Manually annotated and reviewed.
</ CODE>
Automated Annotation Tools
TrEMBL (86,536,393)
Automatically annotated and not reviewed.
Expert Manual Curation

Figure 2.1: Overview of UniProtKB

There are two databases under UniProtKB. Proteins in SwissProt database are manually curated and annotated while TrEMBL database consists of proteins that are annotated by automated annotation tools. Therefore, proteins that are in TrEMBL database have not been reviewed by curators. There are currently 546,790 proteins in SwissProt whereas there are 86,536,393 proteins in TrEMBL database (UniProtKB Release 2014/10). TrEMBL proteins become SwissProt proteins if they satisfy the manual annotation rules that are carried out by expert curators. General overview of the UniProtKB is illustrated in Figure 2.1. Since, proteins and annotations in SwissProt database are manually inspected, growth rate of SwissProt database is slow whereas growth rate of TrEMBL database are exponential. The growth rate of SwissProt and TrEMBL databases is given in Figure 2.2.

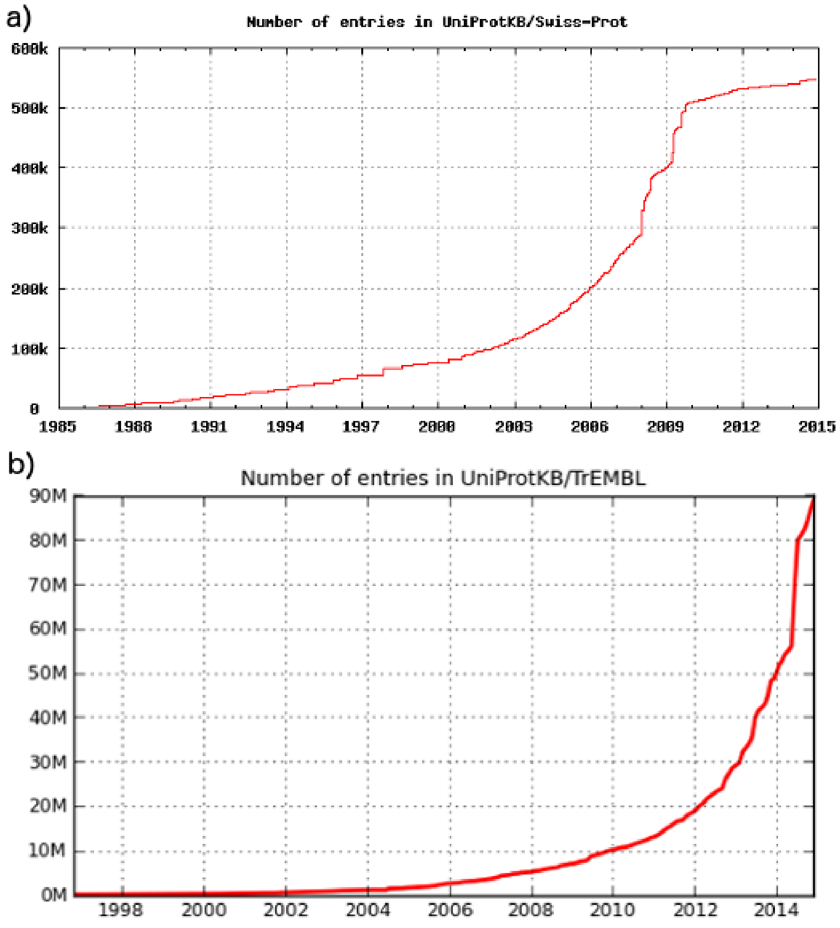


Figure 2.2: Growth rate of a) SwissProt and b) TrEMBL over years ([4],[5].

2.2 Gene Ontology Project and Structure of GO Terms

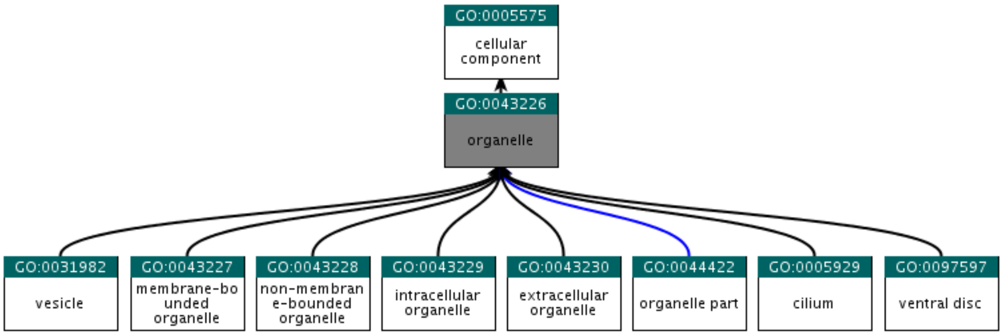
The Gene Ontology (GO) project, created by the GO Consortium, is the most popular and versatile collaborative initiative whose goal is to create a standard, dynamically controlled vocabulary to represent genes and gene product properties by using GO terms [6]. There are three main categories of GO that are biological process (BP), molecular function (MF) and cellular component (CC). Biological process GO terms represent activities that occur by formation of one or more molecular functions. Activities that are performed by gene products in a molecular level are represented by molecular function GO terms. Finally, cellular component GO terms represent parts in cell where the protein mainly carries out its function such as "ribosome".

2.3 Structure of Gene Ontology Terms

Each GO term has a unique identifier in the form of "GO:xxxxxxx" where x represents a digit between 0 and 9. As it is mentioned in previous section, there are three main ontologies in GO. Each category has its own root and they are formed as directed acyclic graphs. Nodes in the graphs represent GO terms. An example part of GO graph is given in Figure 2.3. GO terms are connected to each other with four type of relationships which are *is-a*, *part of*, *has part* and *regulates*.

- If there is an *is-a* relationship between GO_A and GO_B, then GO_B is a more specific term than GO_A. Namely, GO_A is a parent of GO_B. For example, there is an *is-a* relationship between GO:0005515 (protein binding) and GO:0005488 (binding). So, if a protein is annotated by GO:0005515, then it is inherently annotated by GO:0005488.
- If there is a *part of* relationship between GO_A and GO_B, then GO_B has to be a part of GO_A. For example, cytoplasm is part of cell.
- *has part* relationship is specifically used when GO_B always has part of GO_A. Namely, if GO_A exists, then GO_B always exists.
- Finally, *regulates* relationship occurs if a GO term activity directly impacts the

expression or behaviour of another GO term.



QuickGO - http://www.ebi.ac.uk/QuickGO

Figure 2.3: Structure of GO terms ([7])

2.4 Enzyme Commission (EC) Number

Enzyme Commission numbers are recommended by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). EC numbers are used for functional classification of enzymes based on a hierarchy [8]. EC numbers show the biological or chemical reactions of enzymes and they are represented as four numbers/dashes separated by periods. The first three digits show type of the reactions. Last digit of each EC number shows the substrate information [9]. Structure of EC numbers is represented as a tree in Figure 2.4. There is also *is-a* relationship between EC numbers as GO terms. Therefore, an EC number should have functions of its parents.

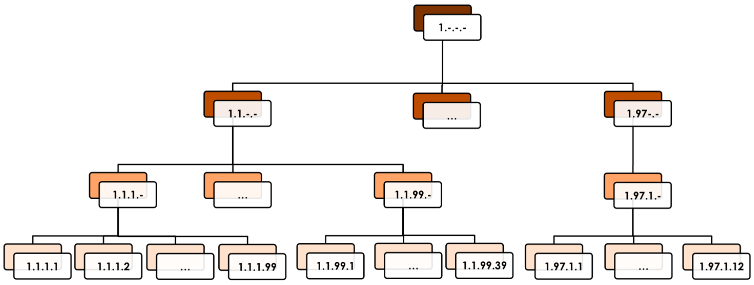


Figure 2.4: Structure of EC numbers

2.5 Domains and Motifs

Domains are discriminative functional or structural blocks that are related to various important activities of proteins such as interactions, functions etc. Proteins may have single or multi-domains and functions of proteins are highly dependent of the domains that they contain. In most of the cases, single domains determine function of a protein. However, multiple domains can collaborate to determine a single function, too [10]. Motifs are conserved regions in proteins, which are short compared to domains. Motifs can be used characterization of proteins that belong to the same groups. Motifs and domains are related because a domain may have common motifs and functions of a protein can be determined by identifying motifs that exist in corresponding domain. Therefore, both domains and motifs can be used to determine functions of proteins. An example of representation of a domain and motif of a protein over 3-D structure of corresponding protein is given in Figure 2.5.

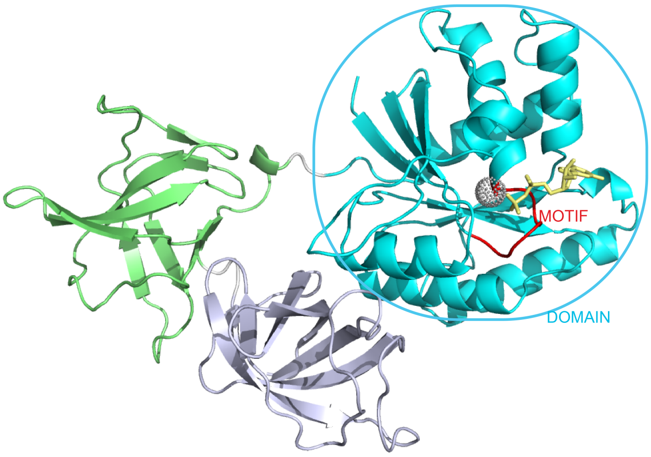


Figure 2.5: Demonstration of domain and motif over 3-D structure of a protein

InterPro is a database where domains, sites and protein family membership information of proteins are stored. Several sources generate protein family and domain information of proteins, called signatures, based on different methods for InterPro database [11]. Most of the protein function prediction methods that give annotations for UniProtKB generate predictions based on InterPro signatures.

CHAPTER 3

RELATED WORK

Several studies are proposed in the literature for functional predictions of enzymatic and non-enzymatic proteins ([12], [13]). There are various data sources that provide many types of biological information which can be used for functional classification of proteins such as sequence, protein structure, physicochemical, protein-protein interactions, gene expression information ([3],[14],[15],[16]). Some of the methods use only one type of biological data whereas others use combination of different types of biological information to predict functions of proteins. Automated protein function prediction methods that are based on protein sequences can be divided into three groups [13] :

- Homology-based approaches
- Subsequence-based approaches
- Feature-based approaches

Homology-based approaches determine function of a protein by aligning target sequence with sequences whose functions are known and transferring functions based on degree of similarity. Subsequence-based approaches use important regions of sequences such as domains and motifs that are highly related to functions of corresponding proteins. Feature-based methods converts sequence of proteins into biological features such as molecular weight, polarity etc. Computational methods that are used for determining functions of proteins can be considered in two groups which are transfer approaches and classification approaches. Transfer approaches determine functions of proteins based on homology and biological relations between proteins.

Automated function prediction techniques that use biological data with data mining and machine learning methods are called classification approaches. In classification approaches, functions of proteins are considered as classes and training data is prepared for each functional class separately. Classification approaches became more popular, since they give more accurate predictions than transfer approaches. There are several methods that use transfer approach and classification approaches. Some methods give predictions by combining the results of other prediction systems. When a protein sequence is queried, the predictions that are agreed by majority or all of the prediction systems are given as output. In recent years, many computational methods are proposed for protein function annotation using GO terms and EC numbers.

ConFunc uses conserved residues of sequences and generates position specific scoring matrices for GO terms by considering hierarchical structure of GO terms to make protein function predictions [17]. ConFunc groups sequences that have the same GO terms and conserved subsequences are found within the same groups. Then, using GO terms and conserved subsequences position specific scoring matrices are constructed to be able to give predictions. ConFunc is applied on MF GO terms for 7150 sequences. Precision and recall values are obtained as 0.77 and 0.41, respectively.

CombFunc is another tool that gives predictions based on GO terms which uses gene expression, protein sequence and protein-protein interactions by incorporating ConFunc [18]. Protein-protein interaction data is used by applying neighbor counting algorithm including indirect neighbors. Co-expressed genes are extracted from gene expression data and frequencies of GO terms in the co-expressed gene set are calculated. Then, support vector machines are used for classifications. CombFunc is tested on 1686 protein set and algorithm is applied for BP and MF GO terms. Precision and recall values for BP GO terms are 0.74 and 0.41, respectively. For MF GO terms, precision and recall values are calculated as 0.71 and 0.64, respectively.

PFP assigns a score for each predicted GO term after running PSI-BLAST to get similar sequences to the target sequence using e-values. Then, predictions are given using data mining methods by considering GO hierarchy [19].

GOtcha predicts GO terms for a given protein or DNA sequence by calculating term-specific probabilities and gives scored matches as output [20]. Term-specific probabilities are calculated according to frequencies of GO terms taken from result of BLAST search of target protein.

JAFA is a meta-server created for protein function annotation problem that uses five function prediction methods and gives scores according to agreement of predicted functions for queried sequences. It also gives the level of the predicted GO term on the GO hierarchy. As the number of level increases, the given prediction becomes more specific and the scores of predictions are calculated by taking into account the GO hierarchy [21].

Roy et al. proposed a novel approach to protein function prediction, COFACTOR, which takes 3-D structure of the proteins as input and gives predictions based on EC numbers, GO terms and protein-ligand binding sites as output. Predictions are given by applying two phases. The former one is global structural alignment of the queried protein. Global structural alignment is done by applying a heuristic algorithm followed by a variation of Needleman-Wunsch algorithm. The second phase is to identify local functional sites of queried protein by applying another heuristic algorithm to find best functional sites between query and training proteins [22].

Vinayagam et al. proposed GOPET for automated protein function annotation problem based on Gene Ontology terms. Protein/nucleotide sequences are blasted against 16 well-known GO annotation databases and proteins are separated into two classes according to similarity search and their functions. Then, support vector machines are used to make predictions. GOPET gives a confidence value for each prediction [23].

HAMAP (High-quality Automated and Manual Annotation of Proteins) is proposed by Gattiker et al. which is originally created for automated protein function annotation of microbial proteins by generating rules. HAMAP is created as a supporting tool for manual annotations for UniProtKB/SwissProt. The system is integrated into UniProt annotation pipeline. HAMAP procedure includes similarity searches against SwissProt/TrEMBL, InterPro searches, profile searches and the system is extended to eukaryotic proteins [24].

Kretschmann et al. proposed SpearMint system which is an automated rule generation system for protein function annotation. Rules are generated by considering length of the sequences, InterPro signatures and taxonomic information using entries in UniProtKB/SwissProt [25]. SpearMint system is improved and the name of the system is changed as Statistical Automatic Annotation System (SAAS). The system is integrated into UniProt pipeline and continuously updated. SAAS system gives prediction for many properties of proteins including GO terms and EC numbers [26].

The Unified Rule (UniRule) system is another system that is currently used at EBI to create, store, and apply manual rules by accompanying other systems that are used. UniRule system gives annotations for several properties of proteins including functional annotations such as GO terms and EC numbers. Since, the created rules are constructed and tested by expert curators the annotations that are given by this system is highly reliable [26].

EnzML is an automated method that is proposed for multi-label prediction of EC numbers using InterPro signatures. A modified version of *k*-Nearest Neighbor (kNN) using binary relevance algorithm is used as classification algorithm. Training dataset is converted into fixed dimensional dataset using Binary Relevance algorithm where number of dimensions are the number of EC numbers to be predicted [27].

Silveria et al. proposed ENZYMatic Annotation Predictor (ENZYMAP) for characterizing and predicting EC numbers. The aim of the system is to annotate enzymes based on the annotation changes between different releases of UniProtKB/SwissProt EC number annotations. Organism Classification (OC), Reference Position (RP) and Keyword (KW) properties of annotations are used to characterize the changes EC number annotations between releases. Three classification techniques are applied which are Naïve Bayes, *k*-Nearest Neighbor (*k*NN) and C4.5 decision tree algorithm to classify EC numbers and their comparisons are performed [28].

Several approaches and computational methods are proposed for functional classification of proteins. When we investigate automated function prediction methods that use protein sequences, we see that homology-based approaches are the most extensively used methods in protein function prediction, since they are fast and easy-to-implement. There are various methods that use transfer approach based on similarities and guilt-by-association methods. However, when we investigate the results of employed methods, we see that computational methods that use machine learning and data mining techniques give more accurate predictions than transfer approaches. In addition, combination of different approaches that use different types of biological data enhance the prediction results. However, most of the available methods are trained and tested on small datasets due to lack of data and computational resources. Therefore, comprehensive studies that use large datasets are insufficient in the literature.

CHAPTER 4

MATERIALS AND METHODS

4.1 Materials

4.1.1 Data Preparation For GO Terms

In this study, GO term annotations from *UniProtKB/SwissProt Release 2014_10* are used for training. There are 14 columns that represent different properties of annotations in Swiss-Prot database. We used *Gene Product ID*, *GO Identifier*, *GO Term Name*, *Aspect* and *Evidence Code* columns in our study. *Evidence Code* column shows how annotations are done and which methods are followed. There are 21 types of evidence codes and only Inferred from Electronic Annotation (IEA) evidence codes are not based on manual curation. The remaining evidence codes are manual evidence codes. There are four categories of manual evidence codes:

- Experimental evidence codes
- Computational Analysis evidence codes
- Author Statement evidence codes
- Curational Statement evidence codes.

Protein functions that are assigned by experimental evidence codes are based on experimental data and they are considered as highly reliable annotations. Annotations that are done by evidence codes that are in the other categories are less reliable although they are annotated by experienced curators. The explanations experimental evidence codes is given in Table 4.1.

Table 4.1: Experimental evidence codes

Evidence Name	**Evidence Code**
Inferred from Experiment	EXP
Inferred from Direct Assay	IDA
Inferred from Physical Interaction	IPI
Inferred from Mutant Phenotype	IMP
Inferred from Genetic Interaction	IGI
Inferred from Expression Pattern	IEP

In previous version of GOPred, annotations based on EXP, IDA, IPI, IMP, IGI and IEP and TAS (Tracable Author Statement) evidence codes were included in positive training dataset. In this study, we removed TAS evidence code and used annotations that have only experimental evidence codes which are EXP, IDA, IPI, IMP, IGI and IEP. The remaining 15 evidence codes are not used since annotations based on other evidence codes have much more noisy data. All of the evidence codes are available in Appendix B.1.

Positive and negative data preparation is very important in classification problems. In protein function prediction problem, positive data preparation is simple. But, a reasonable method should be applied to prepare negative training data. Data preprocessing and positive data preparation can be summarized as follows:

After we downloaded all protein annotations from Swiss-Prot database, annotations whose evidence codes are different from experimental evidence codes are removed. Subsequently, all annotations are propagated to the parents of annotated GO terms according to the "true path rule", which defines inheritance relationships between GO terms [6]. In other words, if a protein is annotated by a GO term, it is considered as it is also annotated by the parents of corresponding GO term. Subsequently, duplicate annotations are removed and number of GO terms to be trained are determined by counting number of proteins that are associated with GO terms. GO terms that have 50 or more unique protein associations with experimental evidence codes are selected for training. We added a new rule to negative data preparation method that is proposed by Sarac et al. and applied to our problem [1]. A protein annotation should fulfill the following properties to be included in the negative set of a GO term:

- Protein should have at least 5 unique annotations by any evidence codes
- GO term annotations of the protein should not contain target GO term or any of its children.
- If protein is annotated by parent/s of the target GO term, it should be annotated by a sibling of the target GO term.

The first rule is added and it is applied as a precondition to the existing rules. Its aim is to increase the probability of a protein being in negative training dataset of a GO term so that it would never have the function of corresponding GO term. Negative dataset preparation for each GO term is presented in the Figure 4.1. Each node represents a GO term and the nodes marked with “X” shows the target GO term. Nodes that are marked with “A” shows annotations that are done for the candidate protein. In the upper graph, since protein is not annotated by target GO term or any of its descendants, it is included in negative training dataset of target protein. In the other graph, protein is not annotated by target GO term or any of its descendants. However, it is annotated by an ancestor of target term. So, we have to check if it is annotated by a sibling of target term or not. Since it is annotated by the sibling of the target term, it is included in the negative training dataset as well.

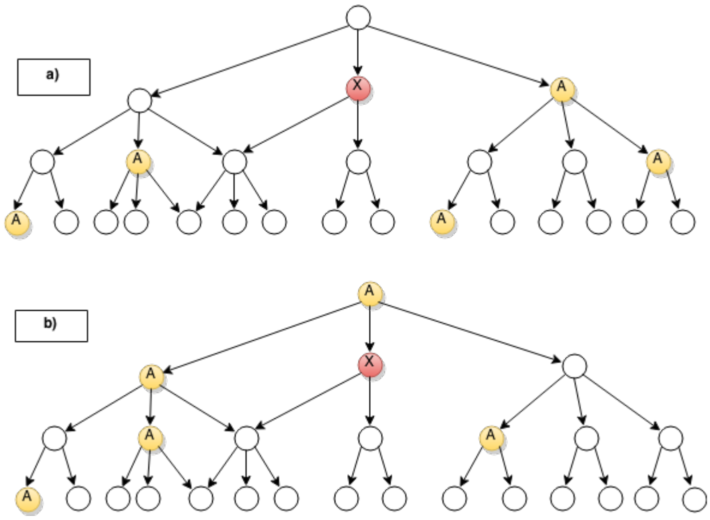


Figure 4.1: In a) protein is not annotated by a parent of X. So, only first and second rules are checked. In b) since protein is annotated by a parent of the target term, annotations should obey all rules. Further information can be found in text.

4.1.2 Data Preparation For EC Numbers

We used EC number annotations from Swiss-Prot database. There are 547,085 proteins available in in *UniProtKB/SwissProt Release 2014_11* and 256,692 of them are classified as enzymes. EC number annotations have no evidence code attribute. So, we used all the enzyme annotations available in Swiss-Prot. Only most specific annotations are given in the database in order to reduce redundancy. Therefore, we propagated annotations to their parents. EC numbers that have more than 50 enzyme associations are determined and selected for training. We prepared training data for a classifier based on its level which is proposed in [2]. Positive and negative dataset for classifiers are prepared according to the rules below:

- Positive training data for EC number X:
  - Proteins that are associated with X and proteins associated with descendants of X
- Negative training data for EC number X:
  - Proteins that are associated with siblings of X and proteins associated with descendants of siblings of X

Positive and negative training data preparation method for EC numbers is illustrated in Figure 4.2. Since, there are four levels in EC hierarchy, positive and negative training data preparation is employed for each level, separately. The aim of this data preparation method is to discriminate the function of an EC number against its siblings without considering other EC numbers. In the figure, dark green EC numbers represent target EC numbers. Proteins that are associated with green EC numbers are included in positive training dataset of target EC number. Proteins that are annotated by red EC numbers are included in negative training dataset of target EC number. Proteins that are associated with grey EC numbers are not considered in training data preparation.

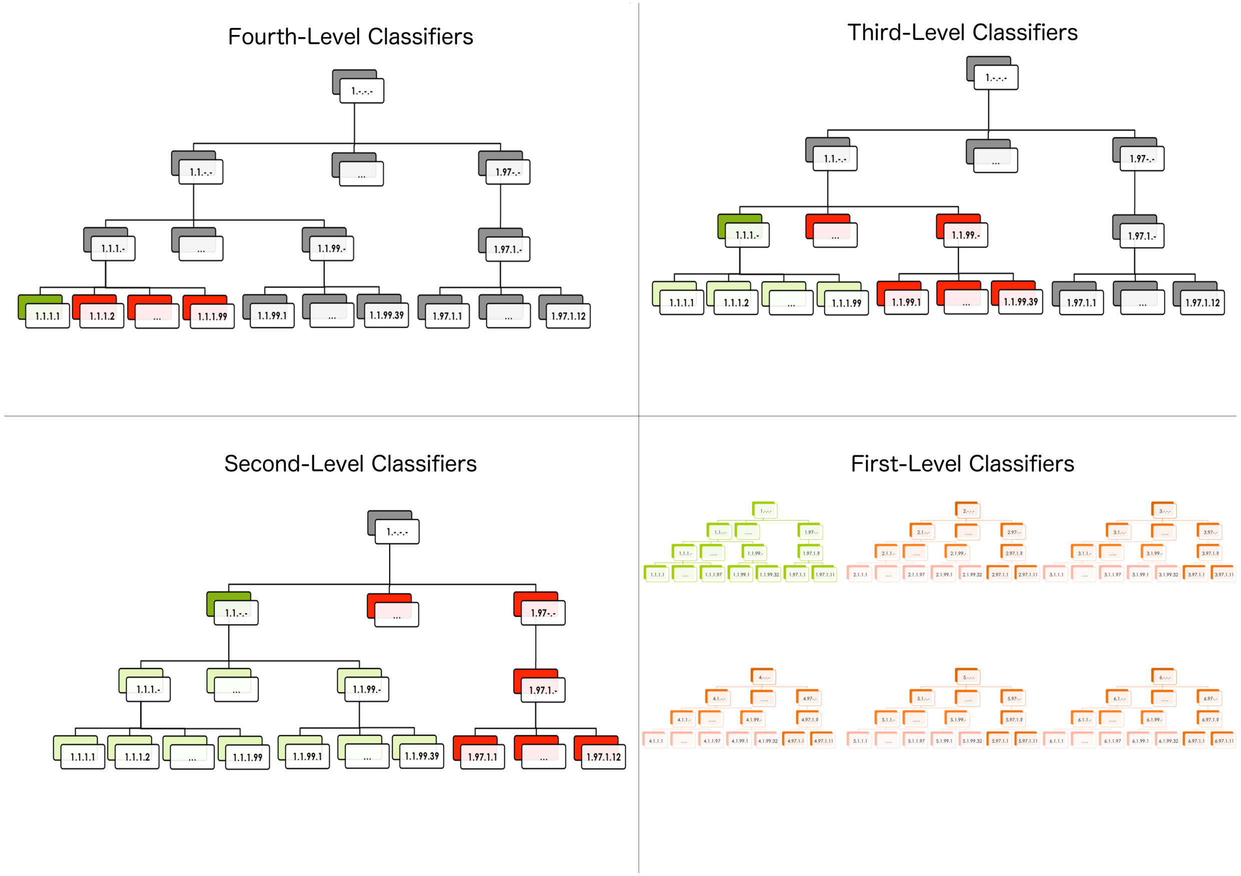


Figure 4.2: Data preparation of EC numbers based on hierarchy

4.2 Methods

After training data is prepared with the new modification, twenty percent of samples are separated as validation dataset. GOPred consists of methods from different approaches which are Blast-*k*NN, SPMap and Pepstats-SVM [1]. In this part, these methods are explained briefly. GOPred is applied on classification of GO terms for all GO aspects. It is also employed functional classification of enzymes using EC numbers. In previous version of GOPred, only probabilities were given for each predicted term. In this study, we employed two methods to determine the predictions to be given instead of probabilities. First, a single global threshold is determined. Subsequently, different decision thresholds are determined, since terms are trained with different training data. The results of two methods are compared. We proposed a hierarchical evaluation method for EC numbers whose aim is to give predictions based data preparation method. Finally, overview of the system is explained.

4.2.1 Blast-kNN

k-Nearest Neighbor algorithm [29] is used with Blast [30] as the first method. Similarity search is done among training dataset of each functional term and *k* Blast scores are taken. The score for an input protein is calculated by Equation 4.1.

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

S_p is sum of *k*-nearest positive Blast scores and S_n is sum of *k*-nearest negative Blast scores of target functional term. This equation gives a score between -1 and 1. Negative scores mean that target protein is more similar to proteins that are in negative set. If the calculated O_B score is positive, target protein is more similar to proteins in positive training dataset.

4.2.2 SPMap

Sarac et al. proposed a subsequence-based method for functional prediction of proteins [31]. SPMap consists of three main modules:

- Subsequence Extraction Module: Fixed-length subsequences of protein sequences in positive training set are extracted. Extraction is done by selecting fixed-length subsequence starting from the first amino acid and shifting one by one.
- Clustering Module: Extracted subsequences are clustered based on their similarities. Similarities of subsequences are calculated using BLOSUM62 matrix. If similarity of a subsequence is over a certain threshold, it is included in the most similar cluster. Otherwise, a new cluster is created.
- Probabilistic Profile Construction: Probabilistic profiles are constructed for each obtained cluster based on the amino acid positions of subsequences within corresponding cluster. Number of probabilistic profiles are the same as number of clusters.

After construction of probabilistic profiles, proteins are represented as fixed-dimensional feature vectors using probabilistic profiles where number of dimensions is equal to number of probabilistic profiles. Finally, proteins that are converted into feature vectors are given as input to SVM classifier. SVM-light software is used as SVM classifier [32].

4.2.3 Pepstats-SVM

Pepstats is a tool available in The European Molecular Biology Open Software (EMBOSS) that calculates statistics for proteins such as molecular weight, number of residues, charge etc [33]. Proteins are represented as 37-dimensional vectors using Pepstats and obtained dataset is given as input to SVM classifier.

4.2.4 Calculation of Combined Prediction Scores

Prediction results of Blast-*k*NN, SPMap and Pepstats-SVM are converted to probabilities using threshold relaxation method [31]. Threshold relaxation method is also used to overcome imbalanced positive and negative training dataset problem. After prediction scores are converted into probabilities, the results of three methods are combined and a single score called Weighted Mean (WMean) is calculated. WMean assigns a weight for each method and it represents a probability value. WMean favors best performed method for the input sequence.

4.2.5 Defining A Global Decision Threshold

When a protein sequence is given as input, we assign a probability value (WMean) for each trained term and generate a prediction file as output for the query protein. The output file contains four columns where columns are GO ID, WMean, SPMap, Blast-*k*NN and Pepstats-SVM scores, respectively. Each row represents a term that are sorted in descending order by WMean value.

In this part of the study, the aim is to find a general threshold value for predictions according to WMean. As it is mentioned in section 4.1.1 and section 4.1.2, hierarchical structures of GO and EC are considered while preparing training datasets. Therefore, we determined optimal decision thresholds based on GO and EC hierarchy. Our aim is to determine the terms that are associated with corresponding protein by considering predictions over the specified threshold. The pseudocode of the algorithm that is used to determine optimal threshold is given in Algorithm 1 which can be summarized as follows:

First, we run extended system for protein sequences in the validation set and get the prediction scores. Swiss-Prot annotations are also determined for corresponding proteins. Optimal threshold and F-Score value of the optimal threshold are assigned to 0 at the beginning. Subsequently, true positive, false positive, false negative predictions are determined for each threshold value as follows: Predicted terms whose scores are greater than the specified threshold value are marked as true positives for each protein, if the protein is annotated by corresponding term or one of its parents. Otherwise, it is marked as false positive. If prediction score is less than the threshold and protein is annotated by predicted term in Swiss-Prot database, prediction is marked as false negative. Optimal threshold and corresponding F-Score values are updated, if calculated F-Score for selected threshold is greater. Finally, best F-score and decision threshold is found and it is used as general optimal threshold.

Algorithm 1 Pseudocode of Global Decision Threshold Algorithm

```
Require: $protein\_predictions$: is a map where keys are protein ids and values are
    list of lists. Each sub-list holds a predicted term and corresponding prediction
    score, $protein\_annotations$: is a map where keys are protein ids and values are
    list of terms and their parents that are associated with corresponding protein
  $optimal\_threshold \leftarrow 0.00$
  $fscore \leftarrow 0.00$        // $fscore$ holds calculated fscore value for optimal threshold
  $threshold \leftarrow 1.00$
  while $threshold \geq 0.00$ do
    for all $P \in validation\_dataset$ do
      for all $Prediction \in protein\_prediction[P].keys()$ do
        $term$ holds predicted term and $SCORE$ holds prediction score
        if $SCORE \geq threshold$ then
          if $term \in protein\_annotations[P]$ then
            $TP \leftarrow TP + 1$
          else
            $FP \leftarrow FP + 1$
          end if
        else if $term \in protein\_annotations[P]$ then
          $FN \leftarrow FN + 1$
        end if
      end for
    end for
    $temp\_fscore \leftarrow F - Scoreforthreshold$
    if $temp\_fscore > threshold$ then
      $fscore \leftarrow temp\_fscore$
      $optimal\_threshold \leftarrow threshold$
    end if
    $threshold \leftarrow threshold - 0.01$
  end while
  return ($optimal\_threshold$,$fscore$)
```

4.2.6 Determining Optimal Thresholds For Terms

Determining optimal threshold values for probabilistic classifiers is an important and difficult problem, especially when training data is unbalanced and classes of some examples are undefined. Datasets can be unbalanced when number of samples belonging a class significantly more than the samples of other classes. In some cases, data is unlabeled and classes for corresponding data are created using heuristics. Therefore, labeled samples by heuristics may not be relevant to assigned classes. In addition, evaluation metrics vary among different problems. There are various evaluation metrics such as accuracy, sensitivity (recall), specificity, precision. Accuracy shows correctly classified samples over all of the samples. Sensitivity represents correctly classified samples in positive dataset over all samples in positive dataset whereas specificity shows correctly classified samples in negative dataset over all samples negative dataset. Finally, precision shows correctly classified samples in positive dataset over all samples that are classified as positive. In some cases, accuracy value is calculated for different thresholds and threshold with highest accuracy value is chosen as optimal threshold. In some other problems, the goal is to achieve high sensitivity while in others the goal is high specificity. In most of the biological applications, majority of the data is negative such as protein function prediction problem. In addition, even proteins in negative training datasets created using logical methods, we are not sure whether they are negative. Therefore, performance measure should be done considering these issues.

In protein function prediction problem based on GO terms and EC numbers, obtaining positive data is trivial. However, even negative training data is chosen artificially by considering hierarchical structure of terms, it is not guaranteed that every single protein in negative set is actually negative. Therefore, a good balancing method should be applied in order to avoid bias. Namely, system should be penalized more for false negative predictions where it should be penalized less for false positive predictions. Chen et al. proposed a method for decision threshold adjustments in classification problems and conducted experiments on different datasets for different classification methods [34]. The method is also used to determine optimal decision threshold for Receiver Operating Characteristic (ROC) analysis [35]. ROC curves are created by plotting false positive rate against true positive rates for different thresholds. It is extensively used evaluation of performances of classifiers and finding optimal decision thresholds. In the study, the following equation is derived to determine optimal threshold for probabilistic classifiers where optimal threshold is set to the threshold with minimum error:

$$P_{ERR}(\tau) = 1 - (SN(\tau) \times \pi_1 + SP(\tau) \times \pi_0) \tag{4.2}$$

In the equation, τ, SN, SP, π_1, π_0 stands for threshold, sensitivity, specificity, prior probabilities of positive and negative samples, respectively. P_{ERR} represents error value regarding performance of classifier for selected threshold τ. In our study, we modified the equation and add a new coefficient Θ as a balancing factor between false negative and false positive predictions. The modified equation can be seen below:

$$P_{ERR}(\tau) = 1 - (SN(\tau) \times \pi_1 \times \Theta + SP(\tau) \times \pi_0) \tag{4.3}$$

Our aim is to minimize the obtained error value using Equation 4.3. The added coefficient is used to relax threshold to increase the number of true positive predictions. When we relax the threshold, number of false positive predictions increases as well as number of true positive predictions. However, small number of increase in false positive predictions is negligible, when we consider ambiguity of proteins' class in negative datasets. The value of Θ is determined by selecting highest F-Score with minimum error. F-Score is a statistical measure which is a harmonic mean of precision and recall values.

In this section, the aim is to find a threshold value for each term individually. Each term is considered as a separate classifier and therefore, performance measure is done separately. The procedure that is followed to find true positive, false positive, true negative and false negative predictions for each term is given in Algorithm 2. The pseudocode of the algorithm can be explained as follows :

First, proteins in validation set are separated as positive and negative sets for each term. Positive and negative sets are prepared according to data preparation methods presented in section 4.1.1 and section 4.1.2. However, when we separate sets, negative validation sets become too large than positive validation sets. To eliminate bias, number of proteins in negative validation sets is set to be same as the number of proteins in positive validation sets by randomly selecting equal number of proteins. Each protein that is in positive validation set of a term is marked as true positive, if its prediction score is above the threshold. Otherwise, it is marked as false negative. Subsequently, each protein that is in negative validation set of a term is marked as false positive, if its prediction score is above the threshold. Otherwise, it is marked as true negative. After true positive, false positive, true negative and false negative predictions are found, F-Score and error values are calculated for selected threshold. Finally, threshold with minimum error and maximum F-Score value is selected as optimal threshold. This algorithm is run five times and average of the values are considered.

4.2.7 Hierarchical Evaluation of Predictions

In this section, a hierarchical evaluation method is proposed to give predictions for EC numbers. Proposed method is used to determine predictions by applying an algorithm consistent with the data preparation method. The pseudocode of the algorithm is given in Algorithm 3. The algorithm is run for all predictions of input sequences separately. The steps that is followed to apply hierarchical evaluation of EC number can be summarized as follows:

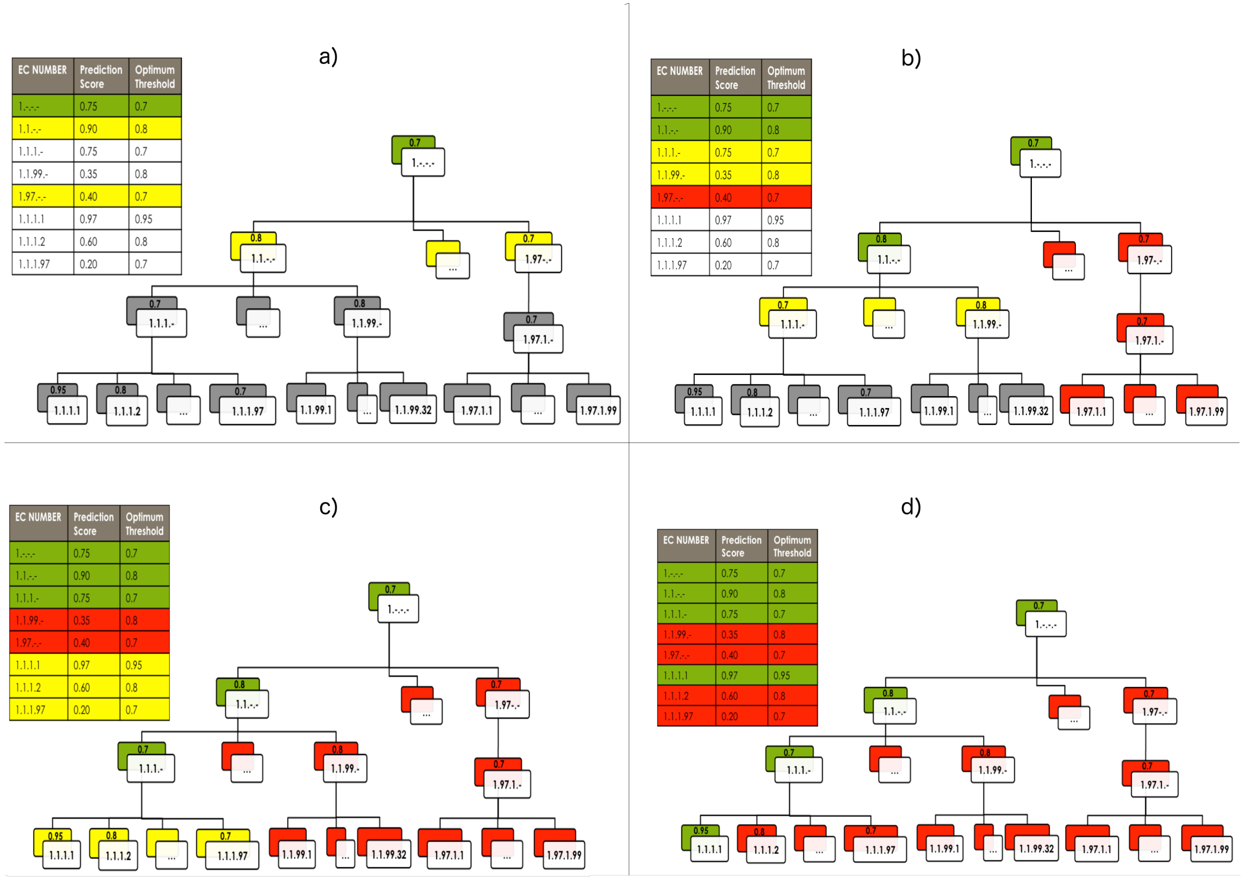
After training, optimal decision thresholds are calculated for each EC number according to the method explained in Section 4.2.6. First, predicted EC numbers and parents of corresponding EC numbers are extracted for each protein. Scores of each predicted EC number and its parents are compared with optimal thresholds of corresponding EC numbers starting from first-level parents up to the level of predicted EC number. If prediction scores of predicted EC and all of its parents are greater than the optimal thresholds of corresponding EC number, predicted term is selected as output for corresponding protein. The hierarchical evaluation method that is used is illustrated in Figure 4.3.

Algorithm 2 Pseudocode of Determining Optimal Thresholds For Terms

```
Require: $positive\_proteins$: is a map where keys are terms, values are proteins
  that are associated with corresponding terms, $negative\_proteins$: is a map where
  keys are terms and values are proteins that are negative according to negative data
  preparation rules,$term\_list$: list of trained terms
  for all $term \in term\_list$ do
    $optimal\_threshold \leftarrow 0.00$
    $threshold \leftarrow 1.00$
    while $threshold \geq 0.00$ do
      for all $pos\_P \in positive\_proteins[term]$ do
        $SCORE$ holds prediction score of $term$ for $pos\_P$
        if $SCORE \geq threshold$ then
          $TP \leftarrow TP + 1$
        else
          $FN \leftarrow FN + 1$
        end if
      end for
      for all $neg\_P \in negative\_proteins[term]$ do
        $SCORE$ holds prediction score of $term$ for $neg\_P$
        if $SCORE \geq threshold$ then
          $FP \leftarrow FP + 1$
        else
          $TN \leftarrow TN + 1$
        end if
      end for
      Calculate error according to the equation 4.3
      if $error < err$ then
        $err \leftarrow error$
      end if
      $threshold \leftarrow threshold - 0.01$
    end while
  end for
```

Algorithm 3 Pseudocode of Hiearchical Evaluation of EC Numbers

```
Require: $list\_of\_ec$: is a list where items are trained EC numbers, $optimal\_thresholds$: is a map where keys are trained EC numbers and values are optimal thresholds of corresponding EC numbers, $prediction\_scores$: is a map where keys are trained EC numbers and values are prediction scores of corresponding EC numbers, $parents\_ec$: is a map where keys are trained EC numbers and values are parents of corresponding EC number
  $real\_predictions$:= { }    // $real\_predictions$ holds final predictions given for protein
  $is\_all\_predicted \leftarrow True$  // a flag that indicates whether prediction scores of all of the parents of $ec$ are over optimal thresholds
  for all $ec \in list\_of\_ec$ do
    $score \leftarrow prediction\_scores[ec]$    // $score$ holds prediction score of $ec$
    if $score$ > $optimal\_thresholds[ec]$ then
      for all $parent \in parents\_ec[ec]$ do
        if $prediction\_scores[parent]$ < $optimal\_thresholds[parent]$ then
          $is\_all\_predicted \leftarrow False$
        end if
      end for
    end if
    if $is\_all\_predicted = True$ then
      $real\_predictions.add(ec)$
    end if
  end for
  return $real\_predictions$
```



a)

EC NUMBER	Prediction Score	Optimum Threshold
1.-.-.-	0.75	0.7
1.1.-.-	0.90	0.8
1.1.1.-	0.75	0.7
1.1.99.-	0.35	0.8
1.97.-.-	0.40	0.7
1.1.1.1	0.97	0.95
1.1.1.2	0.60	0.8
1.1.1.97	0.20	0.7

b)

EC NUMBER	Prediction Score	Optimum Threshold
1.-.-.-	0.75	0.7
1.1.-.-	0.90	0.8
1.1.1.-	0.75	0.7
1.1.99.-	0.35	0.8
1.97.-.-	0.40	0.7
1.1.1.1	0.97	0.95
1.1.1.2	0.60	0.8
1.1.1.97	0.20	0.7

c)

EC NUMBER	Prediction Score	Optimum Threshold
1.-.-.-	0.75	0.7
1.1.-.-	0.90	0.8
1.1.1.-	0.75	0.7
1.1.99.-	0.35	0.8
1.97.-.-	0.40	0.7
1.1.1.1	0.97	0.95
1.1.1.2	0.60	0.8
1.1.1.97	0.20	0.7

d)

EC NUMBER	Prediction Score	Optimum Threshold
1.-.-.-	0.75	0.7
1.1.-.-	0.90	0.8
1.1.1.-	0.75	0.7
1.1.99.-	0.35	0.8
1.97.-.-	0.40	0.7
1.1.1.1	0.97	0.95
1.1.1.2	0.60	0.8
1.1.1.97	0.20	0.7

Figure 4.3: Hierarchical evaluation of EC numbers

4.2.8 General Overview of The System

General overview of the system is given in Figure 4.4. Structure of the system can be summarized as follows:

First, training and validation datasets are created using Swiss-Prot database. Datasets are created for each term separately according to methods explained in Section 4.1.1 and 4.1.2. After training datasets are created, each term is trained and prediction models are generated for each term. Subsequently, validation data is given as input to created models and predictions are obtained for validation data. Finally, system is evaluated and optimal thresholds are determined for each term based on prediction results. When a sequence is given as input, predictions whose scores are over predefined thresholds are given as output for GO terms. If the input sequence is queried for EC numbers, hierarchical evaluation method described in previous section is applied to give predictions.

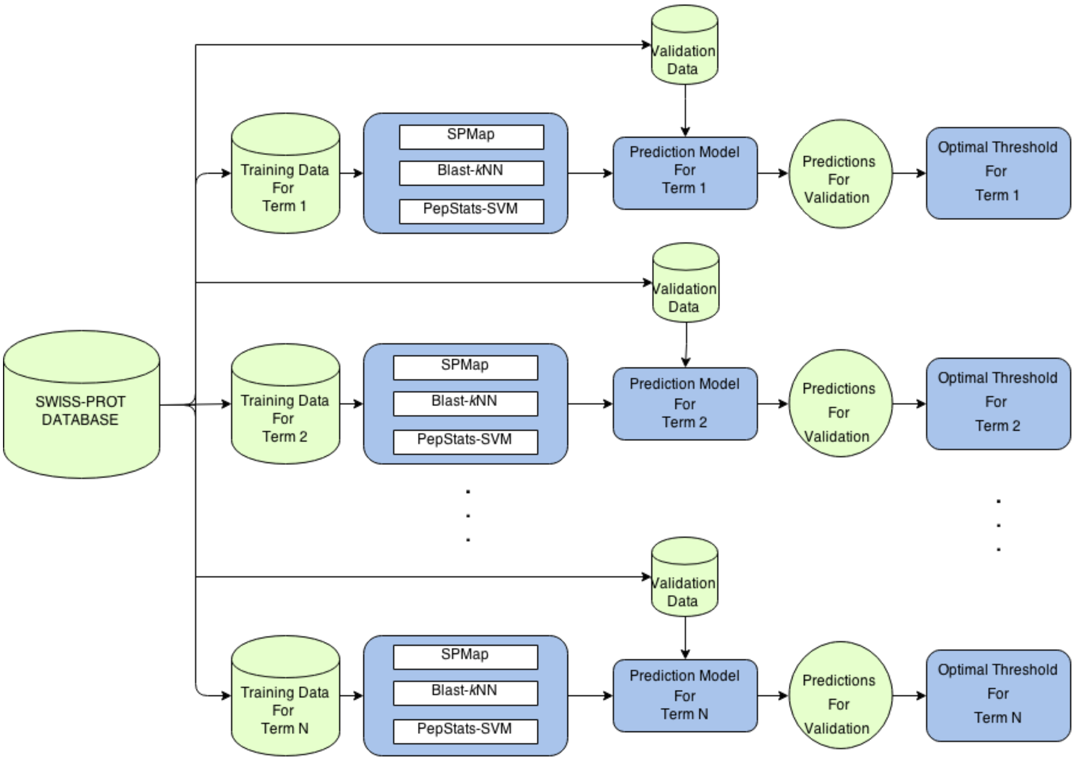

Validation Data
SWISS-PROT DATABASE
Training Data For Term 1
SPMap
Blast-kNN
PepStats-SVM
Prediction Model For Term 1
Predictions For Validation
Optimal Threshold For Term 1
Validation Data
Training Data For Term 2
SPMap
Blast-kNN
PepStats-SVM
Prediction Model For Term 2
Predictions For Validation
Optimal Threshold For Term 2
Validation Data
Training Data For Term N
SPMap
Blast-kNN
PepStats-SVM
Prediction Model For Term N
Predictions For Validation
Optimal Threshold For Term N

Figure 4.4: General overview of the system

4.2.9 Training and Testing

Our system is trained and tested under European Bioinformatics Institute (EBI) cluster system. EBI is a part of European Molecular Biology Laboratory that is a worldly-known research institute. EBI provides various types of biological and molecular databases, including UniProt [36]. EBI cluster is a computer farm that have hundreds of computational nodes. There are several terabytes of RAM and storage. EBI cluster system has the following properties:

- 760 nodes consisting of 21,000 hyper-threaded CPU cores.
- 1 node with 1 TeraByte of RAM and 64 CPUs.
- 6 nodes with 2 TeraByte of RAM and 64/128 CPUs.
- 550 nodes with 128 Gigabyte of RAM.

We parallelized GOPred system for training and testing on EBI cluster. Each term is trained on a different core and the system is distributed over 500 cores for training and testing.

CHAPTER 5

RESULTS

In this study, we used ontology terms as our classifiers for GO and EC. We trained 514 molecular function GO terms, 2909 biological process GO terms and 438 cellular component GO terms. In EC category, 851 EC numbers are also trained using hierarchical data preparation and evaluation method. Optimal decision thresholds are determined using the Equation 4.3 which shows an error value according to performance of classifiers based on different thresholds. Statistical significance of results are shown using F-Score which is combination of precision and recall. In the formulas, TP, FP, TN and FN represent true positive, false positive, true negative and false negative, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

$$Specificity = \frac{TN}{FP + TN} \tag{5.3}$$

$$F - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5.4}$$

5.1 General Optimal Thresholds and Terms as Classifiers

General optimal thresholds are calculated for MF, BP and CC GO terms according to the algorithm that is given in Section 4.2.5. F-Score values are calculated as 0.80, 0.69 and 0.72 for for MF, BP and CC categories of GO, when general optimal thresholds are used. F-Score value for EC numbers is calculated as 0.85 for the determined general optimal threshold. Since, terms are trained separately with their own training data, we determined optimal thresholds for each term according to the algorithm given in Section 4.2.6. ROC curves and determined cut-off points for GO:0019899, GO:0017076, GO:0015276 and GO:0043167 are given in Figure 5.1. Cut-off points that are calculated for GO terms are marked with red crosses over the curves. F-Score values for determined cut-off points are 0.68, 0.92, 0.89 and 0,79 for GO:0019899, GO:0017076, GO:0015276 and GO:0043167, respectively. When we examine the positions of decision thresholds, we see that they are put on plausible places.

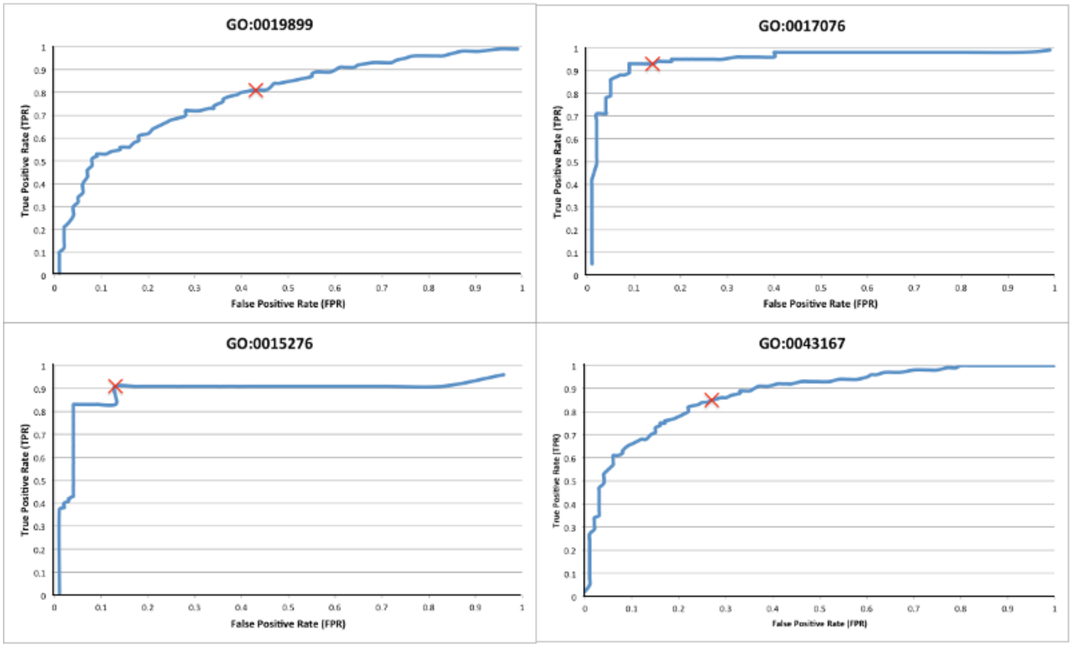


Figure 5.1: ROC curves for GO:0019899, GO:0017076, GO:0015276 and GO:0043167. Red crosses over blue lines show the decision threshold points for corresponding classifiers

5.2 GO Term Prediction Results

Positive and negative data preparation is applied for three different aspects of GO terms separately. Number of GO terms to be trained is determined as 514, 2909 and 438 for MF, BP and CC aspects of GO terms, respectively. Total number of proteins used in positive sets of each GO category and numbers of annotations that are used for corresponding categories are summarized in Table 5.1.

Table 5.1: Summary of training data statistics

GO Aspect	# of GO Terms	# of Proteins	# of Annotations
MF	514	32192	197,665
BP	2909	41994	1,357,805
CC	438	39455	399,952

Twenty percent of positive and negative training data is not included training set and used as validation set. After training, we gave validation set as input to our system and calculated F-Score values for each GO term. F-Score values are determined by changing WMean value and calculating F-Score for corresponding WMean value. Subsequently, maximum F-Score value is selected for GO terms. GO term vs. F-Score plots for MF, BP and CC GO terms can be seen in Figure 5.2, Figure 5.3 and Figure 5.4, respectively. GO terms are ordered in descending order according to F-Score values and they are separated into groups based on their F-Score values. GO terms whose F-Score values are between 1.0 and 0.9 are coloured as red. GO terms whose F-Score values between 0.9 and 0.8 are coloured as blue and so on. F-Score intervals and corresponding colours is given in corresponding figures. Average F-Score values are calculated as 0.86, 0.75 and 0.80 for molecular function, biological process and cellular component aspects of GO respectively. Results show that our method can predict molecular function GO terms with a higher performance. In addition, performance results for biological process and cellular component GO terms are satisfactory.

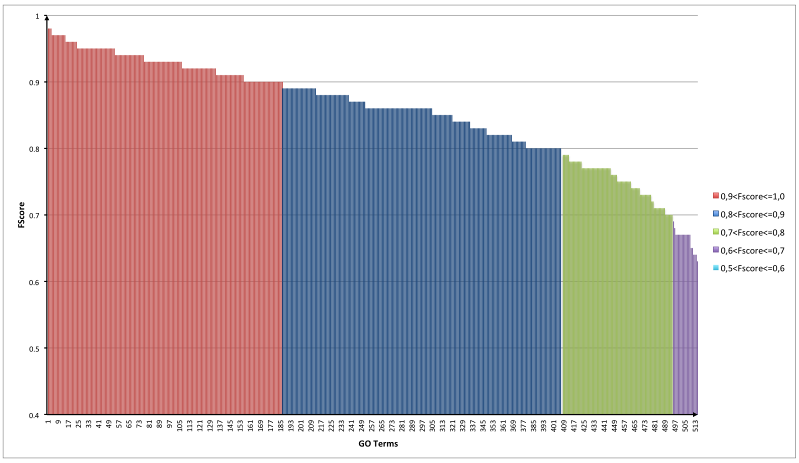


Figure 5.2: Plot of MF GO terms versus their F-score values.

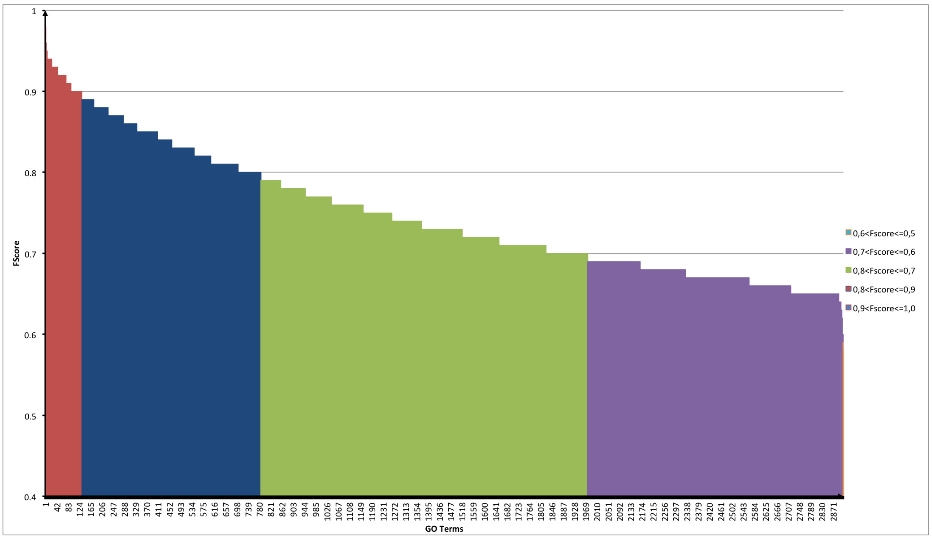


Figure 5.3: Plot of BP GO terms versus their F-score values.

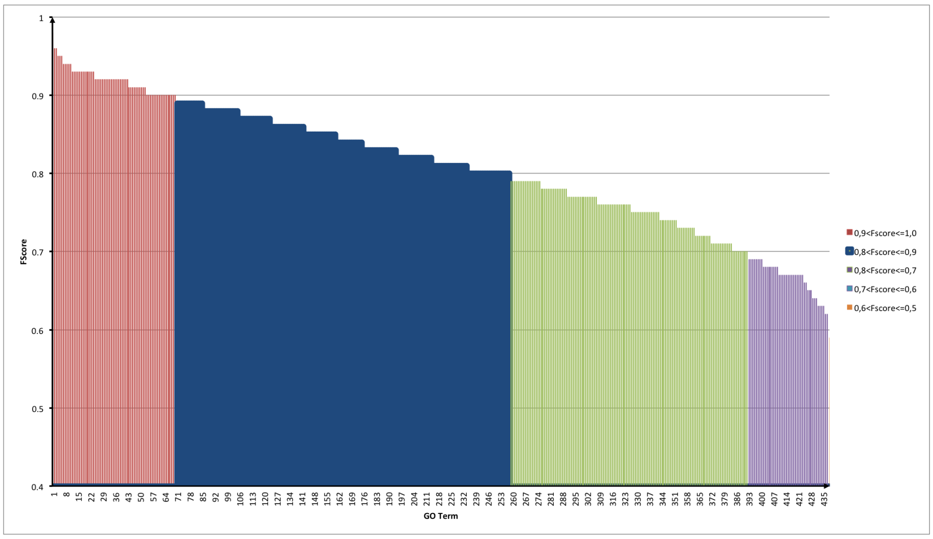


Figure 5.4: Plot of CC GO terms versus their F-score values.

5.3 EC Number Prediction Results

851 EC numbers are trained using our method. There were 901 EC numbers that have 50 or more protein associations. However, there were not enough negative training data for 50 of them. So, they are excluded from trained EC numbers. Number of EC numbers that are trained in each level is given in Table 5.2. We prepared 6 training dataset for first-level classifiers, 50 training dataset for second-level classifiers, 114 training dataset for third level classifiers and 681 training dataset for fourth level classifiers.

Table 5.2: Number of EC numbers trained in each level

Level	**Number of Classifiers**
First	6
Second	50
Third	114
Fourth	681

After training, F-Score values are calculated and average F-Score value is calculated as 0.96 which is higher than the average F-Score values of different aspects of GO terms. Therefore, we can conclude that enzymes carry very distinctive signals on their sequences about their functions and out method identifies these signals successfully

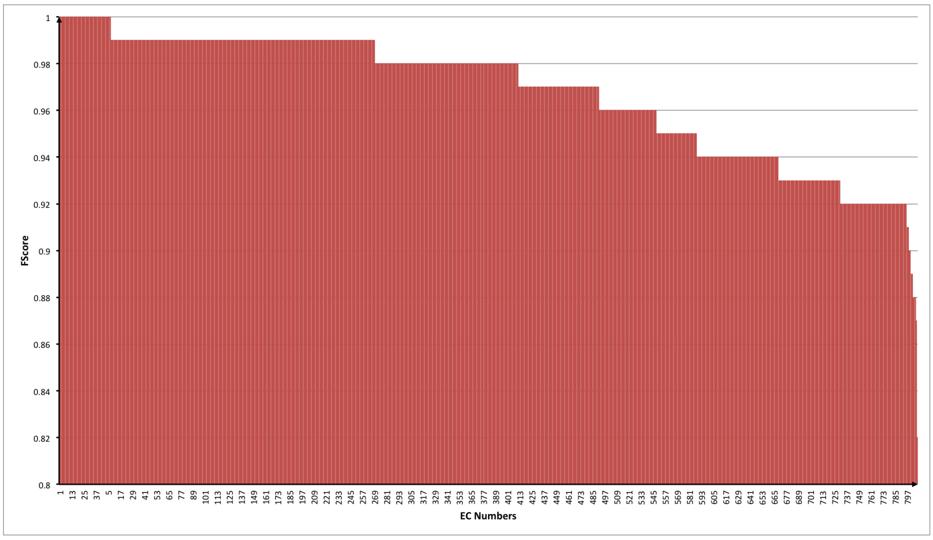


Figure 5.5: Plot of MF GO terms versus their F-score values.

5.4 Testing GOPred On TrEMBL Proteins

System is tested on about 56 million TrEMBL proteins for MF GO terms. After the predictions are obtained for input TrEMBL proteins, optimal thresholds that are calculated in training phase are used to determine the predictions to be given. Predictions whose scores are over the calculated optimal decision thresholds are determined and presented. Our system gave predictions for 17,657,998 proteins within about 56 million proteins. 76,836,472 predictions are given in total for MF GO terms.

In addition, EC number prediction is performed for about 1.7 million enzymes that are selected from TrEMBL database. After the predictions are obtained, hierarchical evaluation method is applied to present predictions for EC numbers. Our system gave predictions for 1,461,145 enzymes. 1,644,656 predictions are given for predicted proteins.

Subsequently, *Comparator* tool is used to compare our predictions with the predictions of TrEMBL reference systems. *Comparator* compares predictions that are given by target system with the annotations that are done by TrEMBL reference systems and it gives statistical information about the predictions that are given by the target system by considering hierarchical structure of GO terms and EC numbers.

5.4.1 Comparison of Predictions

In this section, our aim is to compare the predictions given by our system with the annotations in TrEMBL database. We used *Comparator* system which is a software tool developed by members of UniProt [37]. *Comparator* tool compares predictions that are given by different prediction systems with the predictions given by the automated and semi-automated TrEMBL reference systems. Hierarchical structure of functional terms is also considered, when predictions are evaluated. Therefore, it gives similarity information for predictions. *Comparator* gives two types of output. The former one is protein-based output which consists of three types of information :

- *Number of Predicted Entries* shows number of proteins that are predicted by a prediction system.

- *Entries with NEW predictions* shows number of proteins that does not have any annotations in TrEMBL whereas prediction method gives at least a prediction for corresponding proteins.

- *Others* represents proteins that have at least one annotation in TrEMBL database. The result of comparisons based on prediction proteins can be seen Figure 5.6.

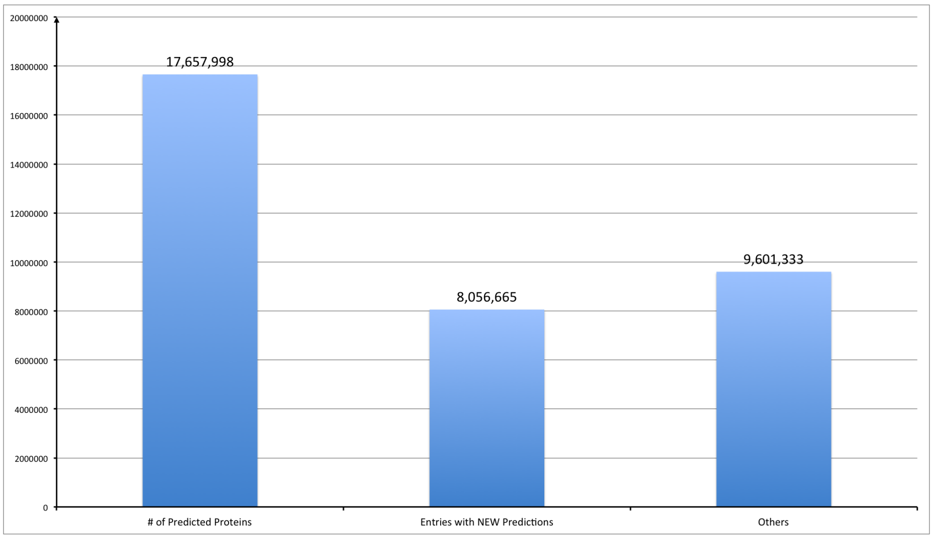


Figure 5.6: Comparator results based on predicted proteins

The second output of *Comparator* shows prediction-based results. It evaluates predictions based on hierarchy and gives five type of information according to the similarities of predictions that the new system gives. The explanation of the five output is as follows:

- *Number of Predictions Compared* represents number of predictions given for input sequences.

- *Predictions with New* shows the number of predictions that are given for proteins that do not have any annotation in TrEMBL database

- *Predictions with IDENTITY* represents the predictions where the same annotations are available in TrEMBL.

- *Predictions with SIMILARITY* shows predictions that are given by our system where the same protein is annotated by descendant or ancestor of predicted GO term.

- *Predictions with MISMATCH* shows predictions that are neither identical nor similar to available predictions. The result of comparisons based on predictions can be seen 5.7

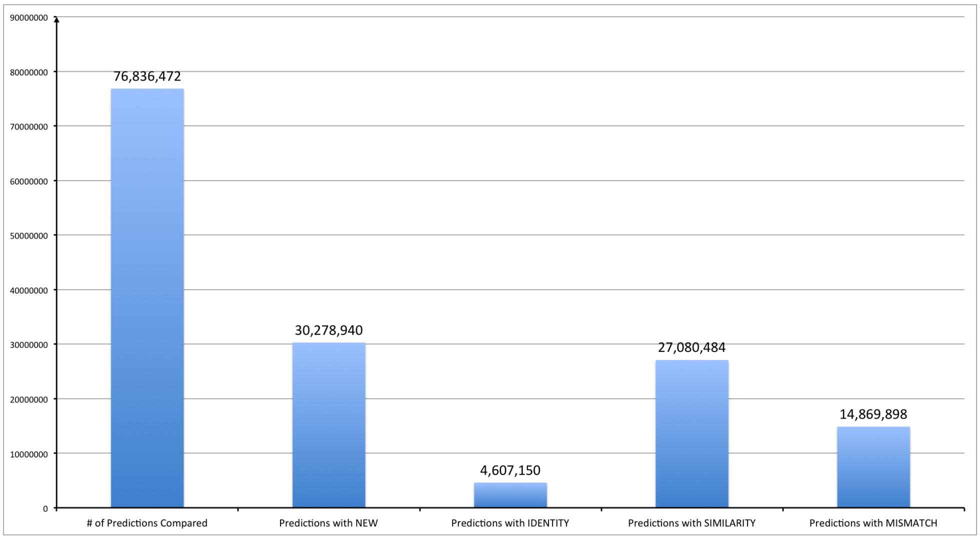


Figure 5.7: Comparator results based on predictions

Number of predictions given by each GO term is calculated to be able to see the distribution. GO terms are separated into four groups according to number of predictions that they gave. Subsequently, predictions that are given by each group is calculated and percentages of predictions that are given by each group is calculated. Percentages of predictions given by each group is given in Figure 5.8. When we look at the results, we see that 33% of the predictions are coming from the GO terms that gave predictions between 1,000,000 and 10,000,000. 52% percent of the predictions are coming from the GO terms that gave predictions between 100,000 and 1,000,000. 14% of the predictions are coming from the GO terms that gave predictions between 10,000 and 100,000. Finally, %1 of predictions is coming from GO terms that gave predictions between 1,000 and 10,000.

Table 5.3: Summary of training data statistics (K = 1,000 M = 1,000,000)

Group Name	Interval	# of GO Terms
Group 1	1 M <# of Predictions<=10 M	16
Group 2	100 K <# of Predictions<= 1 M	135
Group 3	10 K <# of Predictions<=100 K	259
Group 4	1 K <# of Predictions<=10 K	104

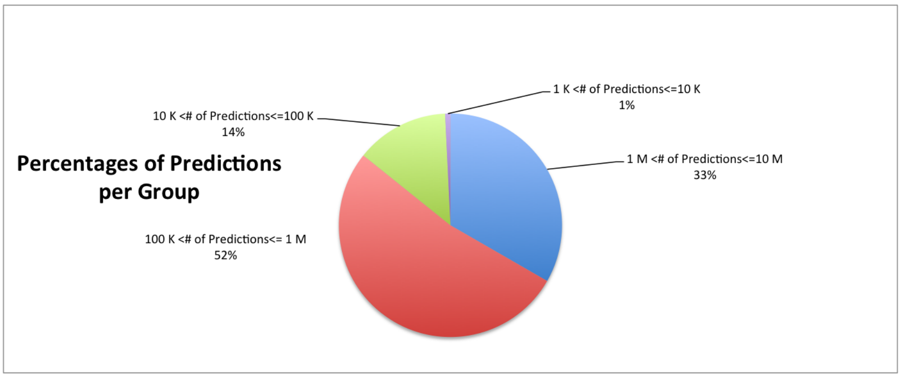


Figure 5.8: Percentages of predictions given by each group

As it is seen in Table 5.3, 33% percent of the predictions are given by only 16 GO terms. When these GO terms are investigated, we see that GO terms that gave predictions between 1,000,000 and 10,000,000 are too general GO terms. For example, GO:0003824 (Catalytic Activity) GO term gives the most predictions among all of the GO terms. When we plot GO:0003824 on GO hierarchy, we see that it is direct descendant of Molecular Function (GO:0003674) GO term which is the most general MF GO term. Some of the examples of GO terms in Group 1 and their prediction numbers can be seen in Figure 5.9.

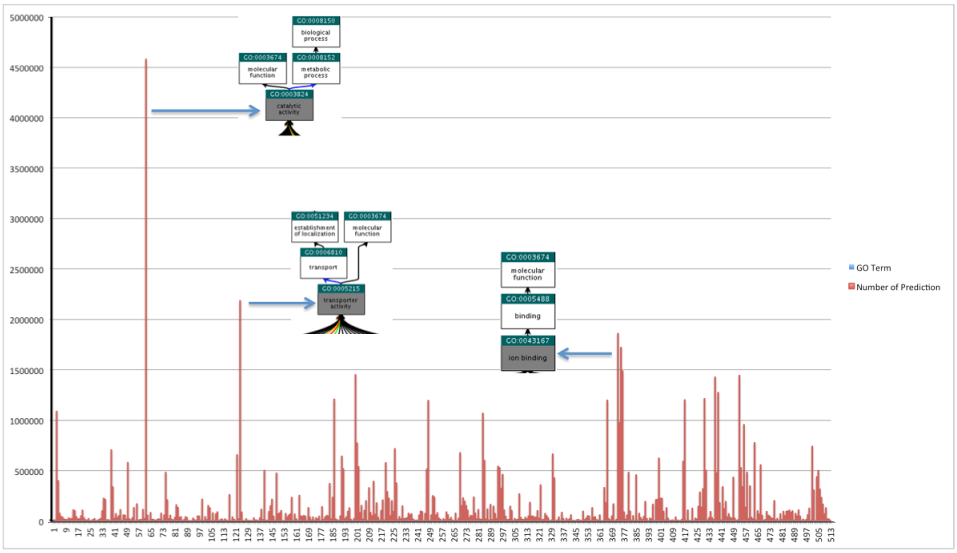


Figure 5.9: Examples of GO terms that gave more predictions

About 1.7 million enzymes from TrEMBL database are selected to be run on *Comparator*. Nearly half of the enzymes have EC number annotations in TrEMBL and the other half does not have any annotations in TrEMBL. So, any predictions that will be done for enzymes that are in the second half will be considered as a new prediction. *Comparator* results for EC number predictions can be seen in Figure 5.10 ve Figure 5.11. Predictions are given according to the Algorithm 3.

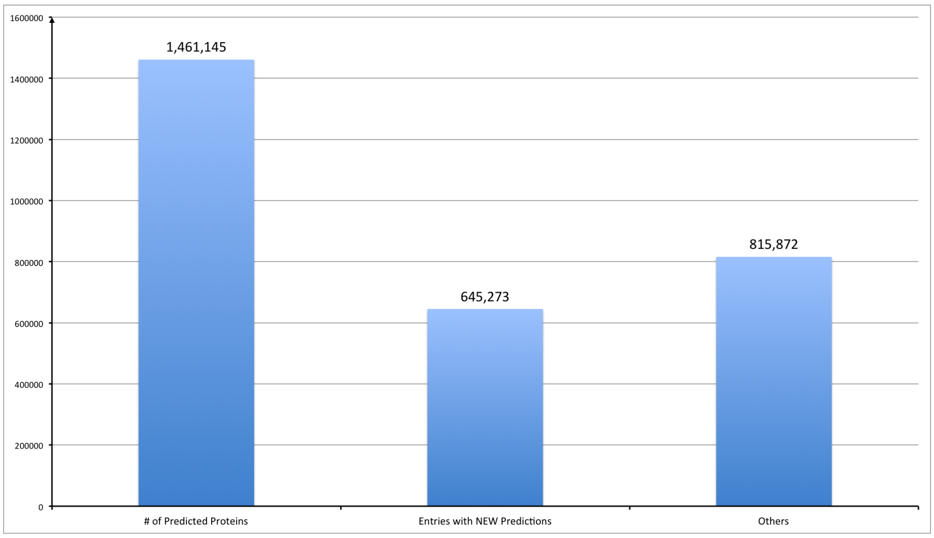


Figure 5.10: EC number comparator results based on predicted proteins

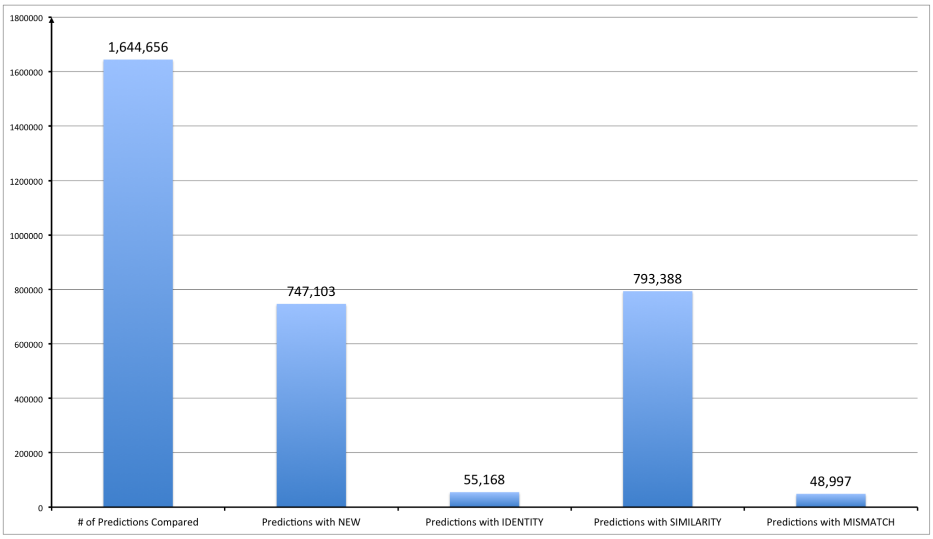


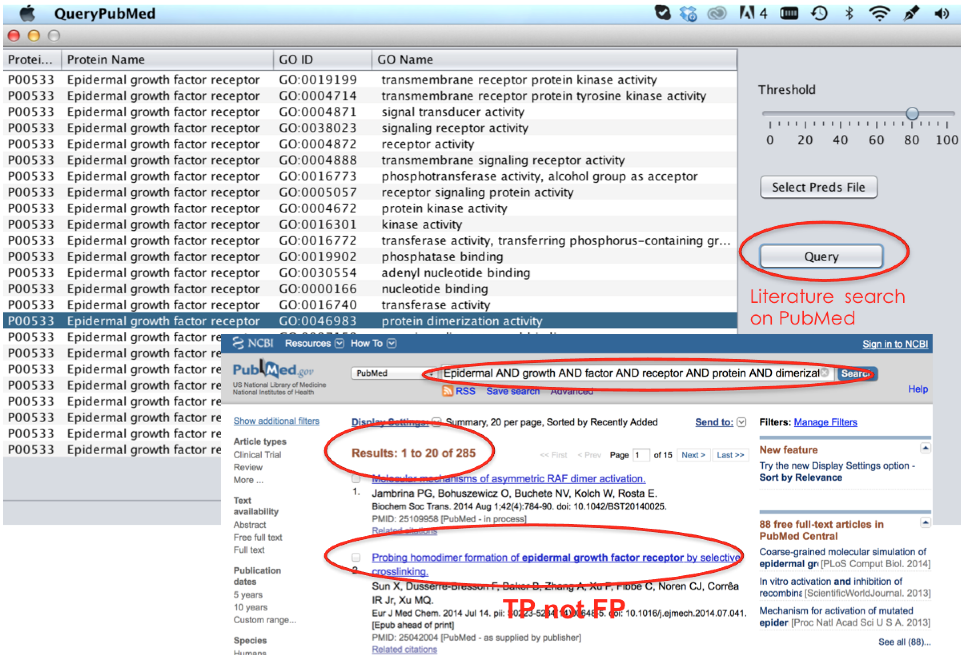
Figure 5.11: EC number comparator results based on predictions

Comparator results of GO terms show that 45% of the predictions are given for the proteins that do not have nay annotation in TrEMBL database. Remaining 55% of the predictions are given for proteins that have at least one annotation in TrEMBL. When the Prediction-based results are examined, 39% of the predictions are given for proteins that do not have any annotation in TrEMBL database. 41% of the predictions are either new or similar to existing annotations. Finally, 20% of the annotations are different from available annotations. EC number prediction results show that 56% of the predictions are given for the enzymes that have EC number annotations in TrEMBL. Prediction-based results show that 45% of the predictions are given for the proteins that do not have any EC number annotation in TrEMBL. 52% of the annotation are similar or identical to available annotations in TrEMBL. When we examine *Comparator* results, it is seen that predictions that are given by our system is consistent with the available predictions. Besides, we can give many new predictions.

5.5 Do False Positives Are Really False Positives?

GO terms are assigned to proteins mainly based on experiments, prediction tools and manual annotations. There are many proteins that have not been annotated by existing GO terms even they have that function. In addition, there may exist annotations based on experiments, but they have not been processed by biocurators. Therefore, predictions that are marked as false positive according to method described in Section 4.2.5 may be actually true positive.

A software tool is prepared to determine if false positive predictions are really false positive or not. The prepared tool searches PubMed to find out if there are existing studies about predictions, which are marked as false positive. When prediction file of a protein is selected, false positive predictions are listed according to the specified threshold. Subsequently, by selecting false positive annotations from the list, protein-GO term pair is searched on PubMed so that publications for selected predictions can be inspected. A set of results are investigated by molecular biologists and some example cases are extracted. For example, in Figure 5.12, it is shown that protein P00533 is associated with GO:0046983 in [38].



Protei...	Protein Name	GO ID	GO Name
P00533	Epidermal growth factor receptor	GO:0019199	transmembrane receptor protein kinase activity
P00533	Epidermal growth factor receptor	GO:0004714	transmembrane receptor protein tyrosine kinase activity
P00533	Epidermal growth factor receptor	GO:0004871	signal transducer activity
P00533	Epidermal growth factor receptor	GO:0038023	signaling receptor activity
P00533	Epidermal growth factor receptor	GO:0004872	receptor activity
P00533	Epidermal growth factor receptor	GO:0004888	transmembrane signaling receptor activity
P00533	Epidermal growth factor receptor	GO:0016773	phosphotransferase activity, alcohol group as acceptor
P00533	Epidermal growth factor receptor	GO:0005057	receptor signaling protein activity
P00533	Epidermal growth factor receptor	GO:0004672	protein kinase activity
P00533	Epidermal growth factor receptor	GO:0016301	kinase activity
P00533	Epidermal growth factor receptor	GO:0016772	transferase activity, transferring phosphorus-containing gr...
P00533	Epidermal growth factor receptor	GO:0019902	phosphatase binding
P00533	Epidermal growth factor receptor	GO:0030554	adenyl nucleotide binding
P00533	Epidermal growth factor receptor	GO:0000166	nucleotide binding
P00533	Epidermal growth factor receptor	GO:0016740	transferase activity
P00533	Epidermal growth factor receptor	GO:0046983	protein dimerization activity
P00533	Epidermal growth factor re		
P00533	Epidermal growth factor re		
P00533	Epidermal growth factor re		
P00533	Epidermal growth factor re		
P00533	Epidermal growth factor re		
P00533	Epidermal growth factor re		
P00533	Epidermal growth factor re		
P00533	Epidermal growth factor re		

Epidermal AND growth AND factor AND receptor AND protein AND dimerizat

Show additional filters

Article types
Clinical Trial
Review
More ...

Text availability
Abstract
Free full text
Full text

Publication dates
5 years
10 years
Custom range...

Species
Humans

Display Settings: Summary, 20 per page, Sorted by Recently Added

Send to:

Results: 1 to 20 of 285

<< First < Prev Page 1 of 15 Next > Last >>

1. Molecular mechanisms of asymmetric RAF dimer activation.
Jambrina PG, Bohuszewicz O, Buchete NV, Kolch W, Rosta E.
Biochem Soc Trans. 2014 Aug 1;42(4):784-90. doi: 10.1042/BST20140025.
PMID: 25109958 [PubMed - in process]
Related citations

2. Probing homodimer formation of epidermal growth factor receptor by selective crosslinking.
Sun X, Dusserre-Bresson F, Baker B, Zhang A, Xu P, Fibbe C, Noren CJ, Corrêa IR Jr, Xu MQ.
Eur J Med Chem. 2014 Jul 14. pii: S0223-5234(14)00648-5. doi: 10.1016/j.ejmech.2014.07.041. [Epub ahead of print]
PMID: 25042004 [PubMed - as supplied by publisher]
Related citations

Filters: Manage Filters

New feature
Try the new Display Settings option - Sort by Relevance

88 free full-text articles in PubMed Central
Coarse-grained molecular simulation of epidermal gr [PLoS Comput Biol. 2014]
In vitro activation and inhibition of recombin [ScientificWorldJournal. 2013]
Mechanism for activation of mutated epider [Proc Natl Acad Sci U S A. 2013]
See all (88)...

Figure 5.12: An example output of PubMedQuery tool

5.6 Individual vs. Combined Classifiers

In this part of the study, we compared performances of individual methods against combination of them. First, best F-Scores are calculated for Blast-kNN, SPMap and PepStats for each functional term. Performances of combination of classifiers based on WMean are also calculated. Subsequently, F-Score values of trained terms are sorted in descending order for each method and the results are plotted. Plots are scaled between 0.80 and 1.00 according to see the performance differences clearly. As it can be seen in Figure 5.13, Figure 5.14, Figure 5.15 and Figure 5.16, the performances of combinations of methods are always higher than the performances of individual methods.In addition performances of Blast-*k*nn and SPMap methods are better than performance of PepStats-SVM.

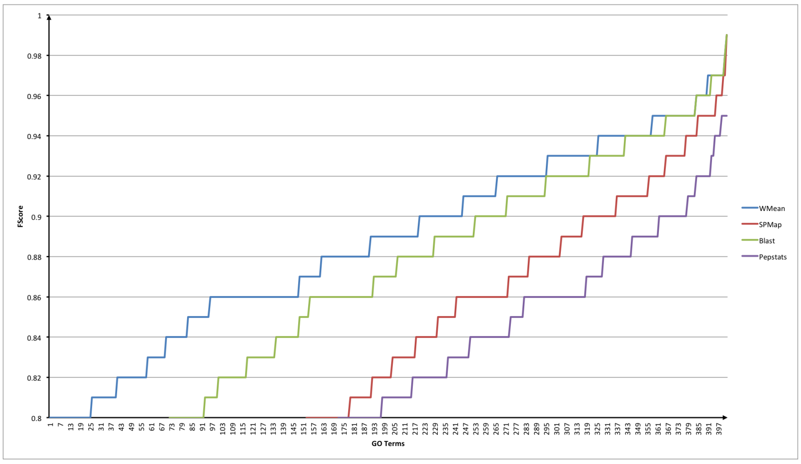


Figure 5.13: Performances of individual and combined classifiers for MF GO terms. x-axis shows MF GO terms, y-axis shows F-Score values

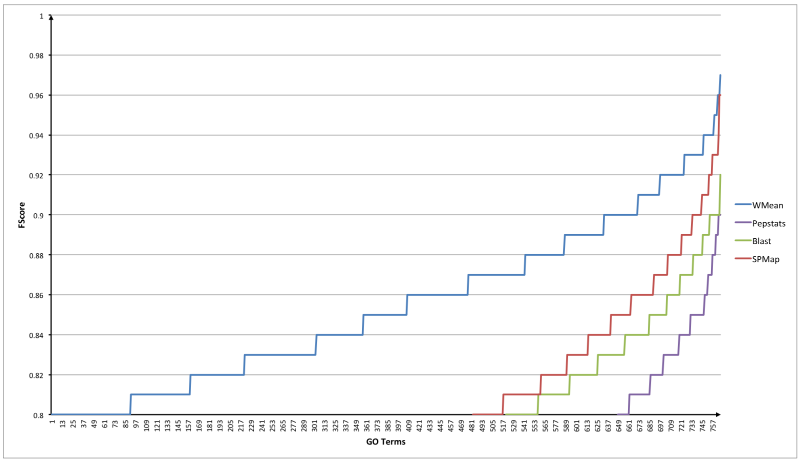


Figure 5.14: Performances of individual and combined methods for BP GO terms. x-axis shows BP GO terms, y-axis shows F-Score values

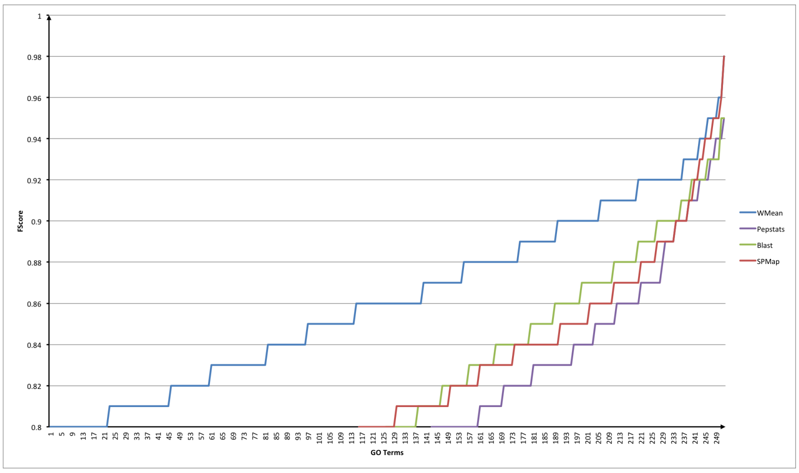


Figure 5.15: Performances of individual and combined methods for CC GO terms. x-axis shows CC GO terms, y-axis shows F-Score values

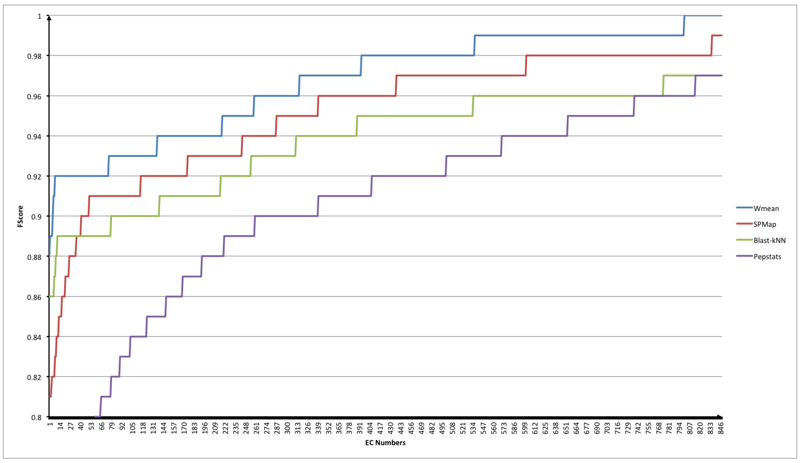


Figure 5.16: Performances of individual and combined methods for EC numbers. x-axis shows EC numbers, y-axis shows F-Score values

5.7 Performance vs. Functionality

In this part of the study, we separated proteins into groups based on their number of annotations. For example, proteins which are in group "Exactly 1" have only 1 annotation is Swiss-Prot database. Proteins which are in group "Exactly 2" have only 2 annotations is Swiss-Prot database and so on. Subsequently, performance of the system is measured for each group separately. The performance results is given in Figure 5.17. x-axis shows threshold value and y-axis shows F-Score for corresponding threshold. When we examine the results, we see that the performance of our system is better for multi-functional proteins. The maximum performance is achieved when the number of annotations that each protein have is 8.

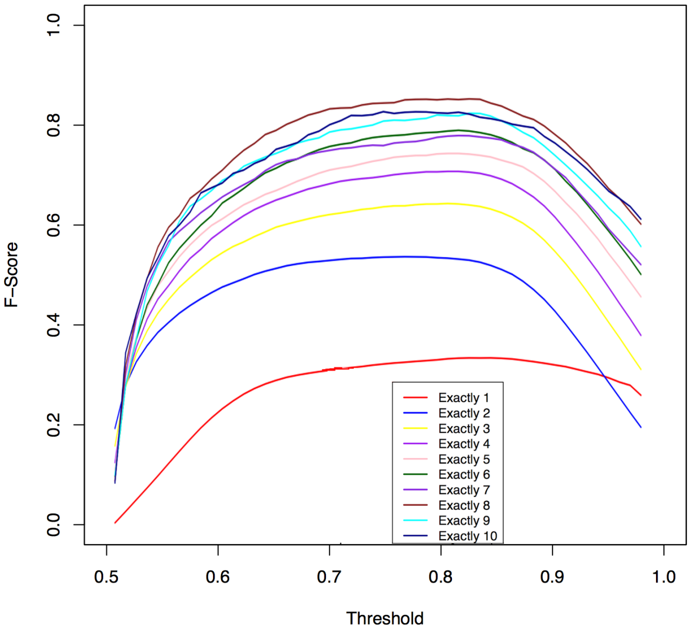


Figure 5.17: Performance curves for different number of annotations.

CHAPTER 6

CONCLUSION AND DISCUSSION

In this study, term prediction is performed using Enzyme Commission and all aspects of Gene Ontology based on GOPred method. Previously proposed GOPred method is improved by adding a safety rule for negative data preparation. We extended number of molecular function Gene Ontology terms from 300 to 514. Biological process and cellular component aspects of Gene Ontology are also added. 2909 biological process and 438 cellular component Gene Ontology terms are trained in addition to 514 molecular function Gene Ontology terms. All system is trained with the new extensions. Results show that average F-Score value of molecular function Gene Ontology terms increased from 0.79 to 0.86, when the new rule is added for negative data preparation.

General optimal threshold is determined by changing threshold (WMean) and selecting threshold with the best F-Score value. General optimal thresholds are defined without considering terms as separate classifiers. The F-Score values are calculated as 0.80, 0.69 and 0.72 for molecular function, biological process and cellular component Gene Ontology terms, when general optimal decision thresholds are used. However, instead of using a general optimal decision threshold, it is more reasonable to determine separate optimal thresholds for each trained term, since performances of classifiers are different than each other. Therefore, we used a method to determine optimal decision thresholds for each trained term. Proposed method is used to decrease number of false negative predictions. F-Score values for Gene Ontology terms are calculated as 0.86, 0,75 and 0.80 for molecular function, biological process and cellular component Gene Ontology terms, respectively. We also showed that performance of our system is better for multi-functional proteins.

Same classification methods is also applied on Enzyme Commission number prediction with hierarchical data preparation and evaluation methods. We trained 851 Enzyme Commission numbers including fourth level Enzyme Commission numbers. Average F-Score value for Enzyme Commission numbers are calculated as 0.96 which is higher than all aspects of Gene Ontology terms. This result shows that function of enzymes can be determined more accurately with our method.To the best of our knowledge this is the best result achieved in the literature.

Trained system is tested on about 56 million proteins TrEMBL proteins for molecular function Gene Ontology terms. Predictions given by our system is compared with the predictions that are given by TrEMBL reference systems. Results show that our system gives 39% new predictions, 5% identical predictions, 35% similar predictions and 21% mismatch predictions. As it is seen, most of the predictions that are given by our system is consistent with the TrEMBL predictions. We investigated the Gene Ontology terms that give significantly more predictions than others and we see that they are too general Gene Ontology terms on the hierarchy. We also tested our system on 1.7 million enzymes from TrEMBL database. Nearly half of the predictions are identical or similar and the other half of the predictions is new predictions.Only a small number of predictions are not consistent with the available predictions.

As a future work, taxonomic restriction can be added to the existing methods. Proteins are classified according to their taxonomies in Swiss-Prot database. In addition, some Gene Ontology terms have taxonomic information and training data can be prepared by considering taxonomies of these GO terms. Predictions that are given for proteins from other taxonomies can be removed to increase the prediction quality for Gene Ontology terms that have taxonomic information. Gene Ontology term and Enzyme Commission number predictions will be investigated by experienced curators from European Bioinformatics Institute (EBI). GOPred system is planned to be integrated into EBI pipeline according to the results of manual inspection.

REFERENCES

[1] Ö.S. Sarac, V. Atalay, and R. Cetin-Atalay. Gopred: Go molecular function prediction by combined classifiers. *PLoS ONE*, 5(8):1–11, 2010.

[2] A. Yaman. Prediction of enzyme classes in a hierarchical approach by using spmap. Master's thesis, Middle East Technical University, 2009.

[3] E.C. Dimmer, R. Huntley, Y. Alam-Faruque, and et al. The uniprot-go annotation database in 2011. *Nucleic acids research*, 40:D565–D570, 2012.

[4] Swiss Institute of Bioinformatics. Uniprotkb/swiss-prot protein knowledgebase release 2015/01 statistics. http://web.expasy.org/docs/relnotes/relstat.html, last visited on January 2015.

[5] EMBL-EBI. Uniprotkb/trembl protein database release 2015/01 statistics. http://www.ebi.ac.uk/uniprot/TrEMBLstats, last visited on January 2015.

[6] The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome research*, 11(8):1425–1433, 2001.

[7] QuickGO. http://www.ebi.ac.uk/QuickGO/, last visited on December 2014.

[8] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Enzyme nomenclature. recommendations 1992. *European Journal of Biochemistry*, 232:1–6, 1995.

[9] K. Tipton and S. Boyce. History of the enzyme nomenclature system. *Bioinformatics*, 16(1):34–40, 2000.

[10] S. Hunter, P. Jones, A. Mitchell, and et al. Interpro in 2011: new developments in the family and domain prediction database. *Nucleic acids research*, 40:D306–D312, 2012.

[11] A. Mitchell, H. Chang, Daugherty L., and et al. The interpro protein families database: the classification resource after 15 years. *Nucleic acids research*, pages 1–9, 2014.

[12] A.K. Tiwari and R. Srivastava. A survey of computational intelligence techniques in protein function prediction. *International journal of proteomics*, 2014:1–22, 2014.

[13] Gaurav P, Vipin Kumar, and Michael Steinbach. Computational approaches for protein function prediction: A survey. https://www.dtc.umn.edu/publications/reports/2007_04.pdf, last visited on January 2015.

[14] Henrick K. Berman, H. and Nakamura H. Announcing the worldwide protein data bank. *Nature Structural and Molecular Biology*, 10(12):980, 2003.

[15] Franceschini A. Kuhn M. Szklarczyk, D. and et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *International journal of proteomics*, 39:D561–D568, 2011.

[16] Burdett T. Fiorelli B. Petryszak, R. and et al. Expression atlas update–a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *International journal of proteomics*, 42:D926–D932, 2014.

[17] M.N. Wass and M.J.E Sternberg. Confunc–functional annotation in the twilight zone. *Bioinformatics*, 24(6):798–806, 2008.

[18] M.N. Wass, G. Barton, and Michael J.E. Sternberg. Combfunc: predicting protein function using heterogeneous data sources. *Nucleic acids research*, 40:798–806, 2012.

[19] T. Hawkins, M. Chitale, S. Luban, and D. Kihara. Pfp: Automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins*, 74(3):566–582, 2009.

[20] D.M.A. Martin, M. Berriman, and G.J. Barton. Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC bioinformatics*, 5:178, 2004.

[21] I. Friedberg, T. Harder, and A. Godzik. Jafa: a protein function annotation meta-server. *Nucleic acids research*, 34:W379–W381, 2006.

[22] Yang J. Roy, A. and Y. Zhang. Cofactor: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research*, 40:W471–W477, 2012.

[23] A. Vinayagam, C. del Val, , F. Schubert, and et al. Gopet: a tool for automated predictions of gene ontology terms. *BMC bioinformatics*, 7:161, 2006.

[24] I. Pedruzzi, C. Rivoire, and A.H. Auchincloss. Hamap in 2013, new developments in the protein family classification and annotation system. *Nucleic acids research*, 41:D584–D589, 2013.

[25] E. Kretschmann, W. Fleischmann, and R. Apweiler. Automatic rule generation for protein annotation with the c4.5 data mining algorithm applied on swiss-prot. *Bioinformatics*, 17(10):D584–D589, 2001.

[26] UniProt. Automatic annotation program. http://www.uniprot.org/program/automatic_annotation, last visited on January 2015.

[27] L. De Ferrari, S. Aitken, J. van Hemert, and I. Goryanin. Enzml: multi-label prediction of enzyme classes using interpro signatures. *BMC bioinformatics*, 13:61, 2012.

[28] S.D.A. Silveira, R. C. de Melo-Minardi, and C.H. da Silveira. Enzymap: exploiting protein annotation for modeling and predicting ec number changes in uniprot/swiss-prot. *PloS one*, 9, 2014.

[29] Cover T.M. Stevens, K.N. and P.E. Hart. Nearest neighbor pattern classification. *IEEE Trans. IT*, 1(13):21–27, 1967.

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

[31] Ö.S. Saraç, Ö. Gürsoy-Yüzügüllü, R. Cetin-Atalay, and V. Atalay. Subsequence-based feature map for protein function classification. *Computational biology and chemistry*, 32(2):122–130, 2008.

[32] T. Joachims. *Making large-Scale SVM Learning Practical (Book Chapter).* MIT Press, 1999.

[33] Longden I. Rice, P. and A. Bleasby. Emboss: the european molecular biology open software suite. *Trends in genetics : TIG*, 16(16):276–277, 2000.

[34] J.J. Chen, C. Tsai, H. Moon, and et al. Decision threshold adjustment in class prediction. *SAR and QSAR in environmental research*, 17(3):337–352, 2006.

[35] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(7):861–874, 2006.

[36] European Bioinformatics Institute. European bioinformatics institute. http://www.ebi.ac.uk/, last visited on January 2015.

[37] UniProt Automatic Annotation Team. Comparator. *R. Saedi, private communication, December 2014.*

[38] X. Sun, F. Dusserre-Bresson, B. Baker, and et al. Probing homodimer formation of epidermal growth factor receptor by selective crosslinking. *European journal of medicinal chemistry*, 88:34–41, 2014.

APPENDIX A

AMINO ACID TABLE

Table A.1: Amino acid table

Amino Acid Name	3-Letter Abbreviation	1-Letter Abbreviation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Asparagine	Asp	D
Aspartic acid	Cys	C
Cysteine	Glu	E
Glutamic acid	Gln	Q
Glutamine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

APPENDIX B

GO EVIDENCE CODES

Table B.1: Experimental evidence codes

Evidence Name	**Evidence Code**
Inferred from Experiment	EXP
Inferred from Direct Assay	IDA
Inferred from Physical Interaction	IPI
Inferred from Mutant Phenotype	IMP
Inferred from Genetic Interaction	IGI
Inferred from Expression Pattern	IEP
Inferred from Sequence or structural Similarity	ISS
Inferred from Sequence Orthology	ISO
Inferred from Sequence Alignment	ISA
Inferred from Sequence Model	ISM
Inferred from Genomic Context	IGC
Inferred from Biological aspect of Ancestor	IBA
Inferred from Biological aspect of Descendant	IBD
Inferred from Key Residues	IKR
Inferred from Rapid Divergence	IRD
Inferred from Reviewed Computational Analysis	RCA
Traceable Author Statement	TAS
Non-traceable Author Statement	NAS
Inferred by Curator	IC
No biological Data available	ND