Lecture Notes in Bioinformatics 4774

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Jagath C. Rajapakse Bertil Schmidt
Gwenn Volkert (Eds.)

Pattern Recognition in Bioinformatics

Second IAPR International Workshop, PRIB 2007
Singapore, October 1-2, 2007
Proceedings

Springer

Volume Editors

Jagath C. Rajapakse
Nanyang Technolocial University, Singapore
E-mail: asjagath@ntu.edu.sg

Bertil Schmidt
University of New South Wales Asia, Singapore
E-mail: bertil.schmidt@unswasia.edu.sg

Gwenn Volkert
Kent State University, USA
E-mail: volkert@cs.kent.edu

Library of Congress Control Number: Applied for

CR Subject Classification (1998): H.2.8, I.5, I.4, J.3, I.2, H.3, F.1-2

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-75285-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-75285-1 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12166909 06/3180 5 4 3 2 1 0

Preface

The advancements of computational and informational techniques have enabled in silico testing of many lab-based experiments in life sciences before performing them in in vitro or in vivo. Though computational techniques are not capable of mimicking all wet-lab experiments, bioinformatics will inevitably play a major role in future medical practice. For example, in the pursuit of new drugs it can reduce the costs and complexity involved in expensive wet-lab experiments. It is expected that by 2010, sequencing of individual genomes will be affordable generating an unprecedented increase of life sciences data, in the form of sequences, expressions, networks, images, literature. Pattern recognition techniques lie at the heart of discovery of new insights into biological knowledge, as the presence of particular patterns or structure is often an indication of its function.

The aim of the workshop series Pattern Recognition in Bioinformatics (PRIB) is to bring pattern recognition scientists and life scientists together to promote pattern recognition applications to solve life sciences problems. This volume presents the proceedings of the 2nd IAPR Workshop PRIB 2007 held in Singapore, October 1–2, 2007. It includes 38 technical contributions that were selected by the International Program Committee from 125 submissions. Each of these rigorously reviewed papers was presented orally at the workshop. The proceedings consists of six parts. Part 1: Sequence Analysis; Part 2: Prediction of Protein Structure, Interaction, and Localization; Part 3: Gene Expression Analysis; Part 4: Pathway Analysis; Part 5: Medical Informatics; and Part 6: Bioimaging.

Part 1 of the proceedings contains seven chapters on sequence analysis. Tang et al. propose a new design of BLAST-based gene ontology (GO) term annotator which incorporates data mining techniques and rough sets to deduce biological functions from DNA sequences. A design of ClustalW, using field programmable gate arrays (FPGA) is developed by Aung et al. to perform sequence alignment in real-time applications. Stepanova, Lin, and Lin develop a two-phase artificial neural network, and present its FPGA implementation, for genome-wide detection of response elements in steroid hormone receptors. Greene, Bill, and Moore propose an expert knowledge-guided mutation operator for the detection of genome-wide variations of DNA, using genetic programming. Luthra et al. find a conserved motif PMNYM of the transmembrane TM5 domain involved in dimerization of the A2a receptor, with a PROSITE search. Deng, Deng, and Havukkala find a strong GC and AT skew correlation in the chicken genome, using a novel visualization technique. Pearson et al. compare interval mapping to a hierarchical Bayesian method for quantitative trait loci analysis on *Arabidopsis thaliana*.

Part 2 of the proceedings contains nine chapters on the prediction of protein structure, interaction, and localization. Shi et al. propose multiple support vector machines (SVM) to handle different features and then decision templates to combine predictions so as to detect protein subcellular localization. Hoque, Chetty, and Dooley

propose a generalized schemata theorem incorporating twin removal for genetic algorithms (GA) to predict protein structure. Zhang, Wei, and Ding use a fuzzy SVM to improve the prediction of structural classes of low-homology proteins. Singh and Ramani demonstrate a method to predict right-handed β-helix fold from protein sequences using SVM and report improved performance measures.

Taguchi and Gromiha investigate several amino acid features and find amino acid occurrences improve the recognition of protein fold recognition significantly over the other features. Ou, Shao, and Chen propose an efficient RBF network to identify interface residues of interacting proteins, based on PSSM profiles and biochemical properties. Ahmad presents dynamic outlier exclusion training algorithm for neural networks to enhance sequence-based predictions in residue level protein properties. Gromiha analyzes amino acid sequences of transmembrane β barrel proteins (TMBs) and finds a significantly higher occurrence of Ser, Asn and Gln in TMBs than in globular proteins. Ahmed estimates the evolutionary average hydrophobicity profile from a family of protein sequences.

Part 3 of the proceedings contains nine chapters on gene expression analysis. Yuriy et al. develop an online database for Affymetrix probe mapping and annotation (APMA) for interactive access, search, and visualization of target sequences mapping and annotation. Blanco, Martin-Merino, and Rivas combine different kinds of dissimilarity-based classifiers for the identification of cancerous samples from microarray data and illustrate its efficacy over existing classifiers. Stiglic, Khan, and Kokol propose small ensemble classifiers to visually interpret microarray data for easy comprehension of their functionality. The method is illustrated in a case-study of leukemia samples. Zhou et al. propose ant-MST, an ant-based minimum spanning tree for gene expression data clustering. McGarry, Sarfraz, and McIntyre integrate GO measures to SOM classification of gene expression data to obtain biologically meaningful clusters of genes.

Teng and Chan find order preserving clusters in gene expression data by converting each gene vector into an ordered label sequence. A method is then proposed by finding the frequent orders by iteratively combining the most frequent prefixes and suffixes in a statistical way. Mao and Tang propose correlation-based relevancy and redundancy measures for efficient gene selection and show promising results in six gene expression problems. Mundra and Rajapakse present relevancy and redundancy criteria for gene selection with an SVM-recursive feature elimination (RFE) method which selects gene subsets with better classification accuracy and generalization capability compared to the SVM-RFE method. Oja obtains digital expression profiles of human endogenous retroviruses.

Part 4 of the proceedings contains four chapters on pathway analysis. Ram and Chetty propose a framework for path analysis in gene regulatory networks by first finding the network structure by causal modeling and then enhancing the network by post-processing. Sehgal et al. reconstruct transcriptional gene regulatory network reconstruction through cross-platform fusion of gene networks. Ling et al. reconstruct protein–protein interaction pathways by mining subject-verb-objects intermediates in biological texts. Chaturvedi, Sakharkar, and Rajapakse propose a validation technique for gene regulatory networks with protein–protein interaction data by using a GA.

They demonstrate the potential of the method in an application to cell-cycle regulation.

Part 5 of the proceedings contains four chapters in medical informatics. Kurzynski and Zolnierek introduce and compare rough set- and fuzzy set-based methods for sequential medical diagnostic problems. Perumal, Lim and Sakharkar propose a comparative genomic approach for metabolic pathway analysis for in silico identification of putative drug targets in *Pseudomonas aeruginosa*. You et al. compare four methods of affinity prediction models for HLA-binding peptides and T-cell epitope identification, and find that non-liner models perform better than linear predictors. Rajapakse and Feng propose a method to identify peptides binding to MHC molecules by simultaneously optimizing entropy and evolutionary distance. Further, the binding motifs are determined by the optimal alignment of binding sites.

Part 6 of the proceedings contains five chapters on bioimaging. Dufour et al. develop a automated nuclear morphometric analysis of 3D fluorescence microscopy images by using active meshes. They also propose shape descriptors and evaluate their robustness and independence on fluorescent beads and on two cell lines. Kumar and Rajapakse propose a time-frequency-based method for detection of activation in functional MRI time-series and discuss the advantages over earlier methods. Dehzangi, Zolghadri, and Boostani develop a weighted distance neural network for high-performance classification of two imagery tasks in the cue-based brain computer interface. Zheng and Rajapakse tract the anatomical connectivity of the brain, using sequential sampling and resampling of diffusion tensor MR images. The method does not adopt fractional anisotropy as the stopping criteria and regularizes the fiber-tracking process by assigning high confidence values at low curvature points. Gong et al. develop an automated pipeline for classification of CT brain images of different head trauma, which is useful for building a content-based medical image retrieval system.

We would like to sincerely thank all authors who spent their time and effort to make important contributions to this book. Many thanks go to the reviewers whose comments have enhanced the quality of the chapters. Our gratitude also goes to the *LNBI editors* and the *managing editor* for their most kind support and help in editing this book.

We would also like to thank all individuals and institutions that contributed to the success of PRIB 2007, especially the authors for submitting the papers and all the sponsors for generously providing financial support for the workshop. We are very grateful to IAPR for the sponsorship and the IAPR Technical Committee (TC-20) on Pattern Recognition for Bioinformatics for their support and advice. Our gratitude goes to the School of Computer Engineering, Nanyang Technological University, Singapore, for supporting the workshop in many ways.

We would like to express our gratitude to all PRIB 2007 International Program Committee members and other invited reviewers for their objective and thorough reviews of the submitted papers. We fully appreciate the PRIB 2007 Organizing Committee for their time and excellent work. We thank Publicity Co-chairs, Feng Lin and Sy Loi Ho, for their hard work in getting the proceedings ready on time. We are grateful to Norhana Ahmad, PRIB 2007 secretary, for coordinating all the logistics of the workshop. Our thanks also go to Ang Linda for maintaining the workshop Web

site, Tan Sing Yau for the technical support, and Jean Tan for his help in graphics design.

Last but not least, we wish to convey our sincere thanks to Springer for providing excellent professional support in preparing this volume.

October 2007

Jagath C. Rajapakse
Raj Acharya
Bertil Schmidt
Gwenn Volkert

Organization

IAPR Technical Committee (TC-20) on Pattern Recognition for Bioinformatics

Raj Acharya (Vice-chair)	Pennsylvania State University, USA
Fransisco Azuaje	University of Ulster, UK
Vladimir Brusic	University of Queensland, Australia
Phoebe Chen	Deakin University, Australia
David Corne	Heriot-Watt University, UK
Elena Marchiori	Vrije University of Amsterdam, The Netherlands
Mariofanna Milanova	University of Arkansas at Little Rock, USA
Gary B. Fogel	Natural Selection, Inc., USA
Saman K. Halgamuge	University of Melbourne, Australia
Visakan Kadirkamanathan	University of Sheffield, UK
Nik Kasabov	Auckland University of Technology, New Zealand
Irwin King	Chinese University of Hong Kong, Hong Kong
Alex V. Kochetov	Russian Academy of Sciences, Russia
Graham Leedham	Nanyang Tech. University, Singapore
Ajit Narayanan	University of Exeter, UK
Marimuthu Palaniswami	University of Melbourne, Australia
Jagath C. Rajapakse (Chair)	Nanyang Tech. University, Singapore
Gwenn Volkert	Kent State University, USA
Roy E. Welsch	Massachusetts Inst. of Technology, USA
Kay C. Wiese	Simon Fraser University, Canada
Limsoon Wong	National University of Singapore, Singapore
Jiahua (Jerry) Wu	Wellcome Trust Sanger Inst., UK
Yanqing Zhang	Georgia State University, USA
Qiang Yang	Hong Kong University of Science and Technology, Hong Kong

PRIB 2007 Organization

General Chair

Jagath C. Rajapakse (Co-chair)	Nanyang Technological University, Singapore

General Co-chair

Raj Acharya	Pennsylvania State University, USA

Program Chairs

Bertil Schmidt	University of New South Wales Asia, Singapore
Gwenn Volkert	Kent State University, USA

Special Session Chairs

Shandar Ahmad	National Institute of Biomedical Innovation, Japan
Madhu Chetty	Monash University, Australia
Elena Marchiori	Vrije University of Amsterdam, The Netherlands

Publicity Chairs

Saman K. Halgamuge	University of Melbourne, Australia
Roberto Tagliaferri	Università Di Salerno, Italy
Wei Wang	Fudan University, China
Yanqing Zhang	Georgia State University, USA

Publication Chairs

Sy-Loi Ho	Nanyang Technological University, Singapore
Feng Lin	Nanyang Technological University, Singapore

Local Chair

Graham Leedham	University of New South Wales Asia, Singapore

Local Organization Committee

Byron Koon Kau Choi	Nanyang Technological University, Singapore
Yulan He	Nanyang Technological University, Singapore
Hwee Kuan Lee	Bioinformatics Institute, Singapore
Jinming Li	Nanyang Technological University, Singapore

Secretariat

Norhana Binte Ahmad	Nanyang Technological University, Singapore

System Administration

Linda Ang Ah Giat	Nanyang Technological University, Singapore

Program Committee

Tatsuya Akutsu	Kyoto University, Japan
Guillaume Bourque	Genome Institute of Singapore, Singapore
Timo Rolf Bretschneider	Nanyang Technological University, Singapore
Zehra Cataltepe	Istanbul Technical University, Turkey
Phoebe Chen	Deakin University, Australia
Francis Y.L. Chin	University of Hong Kong, Hong Kong
Peter Clote	Boston College, USA
David Corne	Heriot-Watt University, UK
Carlos Cotta	University of Malaga, Spain
Antoine Danchin	Institut Pasteur, France
Joaquín Dopazo	Centro de Investigación Príncipe Felipe, Spain
James G. Evans	Massachusetts Institute of Technology, USA
Alexandru Floares	Oncological Institute Cluj-Napoca, Romania
Mikhail S. Gelfand	Institute for Information Transmission Problems, Russia
Ilkka Havukkala	Auckland University of Technology, New Zealand
Jaap Heringa	Vrije Universiteit, The Netherlands
Lisa Holm	University of Helsinki, Finland
Ming-Jing Hwang	Academia Sinica, Taiwan
Visakan Kadirkamanathan	University of Sheffield, UK
Nikola Kasabov	Auckland University of Technology, New Zealand
Irwin King	The Chinese University of Hong Kong, Hong Kong

Alex V. Kochetov	Russian Academy of Sciences, Russia
Vladimir A. Kuznetsov	Genome Institute of Singapore, Singapore
Chee Keong Kwoh	Nanyang Technological University, Singapore
Wing-Ning Li	University of Arkansas, USA
Alan Wee-Chung Liew	Chinese University of Hong Kong, Hong Kong
Frederique Lisacek	Swiss Institute of Bioinformatics, Switzerland
Hiroshi Matsuno	Yamaguchi University, Japan
Martin Middendorf	Universität Leipzig, Germany
Mariofanna Milanova	University of Arkansas at Little Rock, USA
Aleksandar Milosavljevi	Baylor College of Medicine, USA
Satoru Miyano	University of Tokyo, Japan
Jason H. Moore	Dartmouth Medical School, USA
Parvin Mousavi	Queen's University, Canada
See-Kiong Ng	Institute for Infocomm Research, Singapore
Yanay Ofran	Columbia University, USA
Christos Ouzounis	European Bioinformatics Institute, UK
Zoran Obradovic	Temple University, USA
Nikhil R. Pal	Indian Statistical Institute, India
Laxmi Parida	IBM T.J. Watson Research Center, USA
Mihail Popescu	University of Missouri, USA
Predrag Radivojac	Indiana University, USA
Nikolaus Rajewsky	Max Delbruck Center for Molecular Medicine, Germany
Jem Rowland	University of Wales Aberystwyth, UK
Meena Kishore Sakharkar	Nanyang Technological University, Singapore
Akinori Sarai	Kyushu Institute of Technology, Japan
Alexander Schliep	Max Planck Institute for Molecular Genetics, Germany
Christian Schoenbach	Nanyang Technological University, Singapore
N.Srinivasan	Indian Institute of Science, India
P. N. Suganthan	Nanyang Technological University, Singapore
Wing Kin Sung	National University of Singapore, Singapore
Anna Tramontano	University of Rome "La Sapienza", Italy
Michael Wagner	Cincinnati Children's Hospital Research Foundation, USA
Haiying Wang	University of Ulster at Jordanstown, UK
Lusheng Wang	City University of Hong Kong, Hong Kong
Michael Q. Zhang	Cold Spring Harbor Laboratory, USA

Reviewers

Konagaya Akihiko	RIKEN, Genomic Sciences Centre, Japan
Mundra Piyushkumar Arjunlal	Nanyang Technological University, Singapore
Wendy Ashlock	University of Guelph, Canada
Sansanee Auephanwiriyakul	Chiangmai University, Thailand
Jung-Hsien Chiang	National Cheng Kung University, Taiwan
Kai-Bo Duan	Center for Drug Discovery, Singapore
Julien Epps	University of New South Wales Asia, Singapore
Margaret J. Eppstein	University of Vermont, Canada
Bruno Gaeta	University of New South Wales, Australia
Shinn-Ying Ho	National Chiao Tung University, Taiwan
Masoud Jamei	Simcyp Limited, UK
Vert Jean-Philippe	Ecole des Mines de Paris, France
Vinny Just	Ohio University, USA
Marta Kasprzak	Poznan University of Technology, Poland
Kyung Joong Kim	Yonsei University, Korea
Prasanna Ratnakar Kolatkar	Genomic Institute of Singapore, Singapore
Lukasz Kurgan	University of Alberta, Canada
Weiguo Liu	Nanyang Technological University, Singapore
Pasi Luukka	Lappeenranta University of Technology, Finland
Jianmin Ma	Nanyang Technological University, Singapore
Nawar Malhis	University of British Columbia, Canada
Bernard Moret	Ecole Polytechnique Federale de Lausanne, France
Ngoc Minh Nguyen	Nanyang Technological University, Singapore
Merja Oja	University of Helsinki, Finland
Menaka Rajapakse	Institute of Infocomm Research, Singapore
Carmelina Ruggiero	University of Genoa, Italy
Muhammad Shoaib B. Sehgal	Monash University, Australia
Scott Smith	Boise State University, USA
Yuchun Tang	Georgia State University, USA
Thanos Vasilakos	University of Western Macedonia,Greece
Chandra Verma	Bioinformatics Institute, Singapore
Tiffani Williams	Texas A&M Engineering, USA
Gwan-Su Yi	Information and Communications University, Korea
Rui Xu	University of Missouri-Rolla, USA
Runxuan Zhang	Institut Pasteur, France
Shuigeng Zhou	Fudan University, China

Table of Contents

Part I: Sequence Analysis

Part II: Prediction of Protein Structure, Interaction and Localization

Part IV: Pathway Analysis

Part V: Medical Informatics

Part VI: Bioimaging

Automated Methods of Predicting the Function of Biological Sequences Using GO and Rough Set

Xu-Ning Tang[1], Zhi-Chao Lian[2], Zhi-Li Pei[2,3], and Yan-Chun Liang[2,*]

[1] College of Software, Jilin University, Changchun 130012, China
[2] College of Computer Science and Technology, Jilin University, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Changchun 130012, China
[3] College of Mathematics and Computer Science, Inner Mongolia University for Nationalities, Tongliao 028043, China
ycliang@jlu.edu.cn

Abstract. With the extraordinarily increase in genomic sequence data, there is a need to develop an effective and accurate method to deduce the biological functions of novel sequences with high accuracy. As the use of experiments to validate the function of biological sequence is too expensive and hardly to be applied to large-scale data, the use of computer for prediction of gene function has become an economical and effective substitute. This paper proposes a new design of BLAST-based GO term annotator which incorporates data mining techniques and utilizes rough set theory. Moreover, this method is an evolution against the traditional methods which only base on BLAST or characters of GO Terms. Finally, experimental results prove the validity of the proposed rough set-based method.

Keywords: GO BLAST Rough Set Theory.

1 Introduction

Along with the development of modern sequencing technology, the number of gene sequence is increasing everyday. A report coming from GenBank, a major repository of genomic data, shows an exponential increase in sequence data, during the last decade. As a result, biologists have to waste amount of time in finding out some useful information within specific domain. Even worse, different biological database might use different nomenclatures, which like some dialects, making information search, especially for computer-based information search, unavailable. So, how to store and take advantage of the information has become many biologists' common concern.

1.1 Gene Ontology

The emergence of Gene Ontology (GO) project has been used to solve the nomenclature problem. Gene Ontology project provides a set of unified, standard and hierarchical terms to note the functional characters of gene products [1]. People can use

* Corresponding author.

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 1–10, 2007.

nomenclature provided by GO project to annotate the biological functions of biological sequences.

Each item in GO database is composed with three key parts: gene product ID, GO terms and evidence code. Among them, gene product ID uniquely identifies the sequence of a gene product. Moreover, as sequence data alone is of limited use to biologists, GO project annotates the functions of gene products from three points of view. They are biological process, cellular component and molecular function. At last, evidence code indicates how annotation to a particular term is supported.

Essentially, each of these three types of terms can be separated into more detailed sub-categories, so that those terms construct a DAG (directed acyclic hierarchical graph), shown in Figure 1. Generally speaking, GO is a unified biological tool which can annotate gene product's function with a set of dynamic controlled vocabulary and it can keep on upgrading with the development of biology.

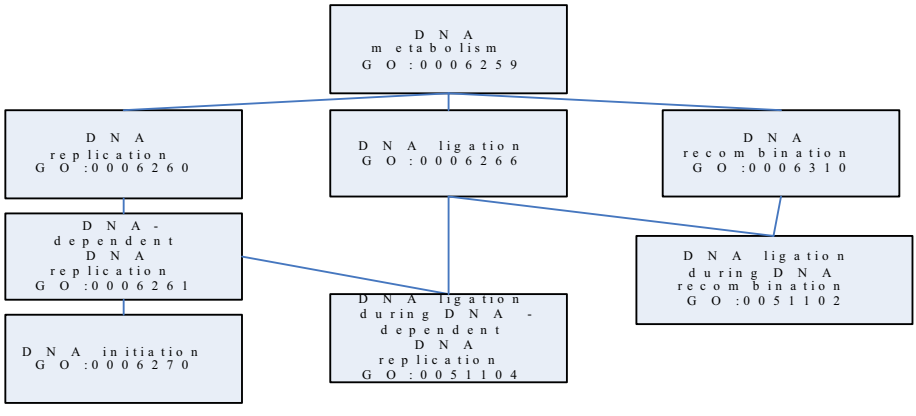

Fig. 1. Directed acyclic hierarchical graph of GO term

1.2 Basic Theory About Rough Set

Rough set has been introduced as a mathematical tool for dealing with fuzzy and uncertain knowledge in artificial intelligence application.

For convenience, we will introduce some basic concepts of rough set at first [2].

Definition 1. Given a knowledge system K= (U, R), for each subset $X \subseteq U$ and an equivalence relation $R \in ind(K)$, define two subsets:

Lower approximation: $\underline{R}X = \bigcup\{Y \in U / R \mid Y \subseteq X\}$

Upper approximation: $\overline{R}X = \bigcup\{Y \in U / R \mid Y \cap X \neq \varnothing\}$

Any subset defined by its lower and upper approximation is called a rough set.

Definition 2. Positive region: Let P and Q be equivalence relations within U, $pos_P(Q)$ is called the P-positive region of Q, such that $pos_p(Q) = \bigcup_{X \in U/Q} \underline{P}X$.

Definition 3. Let $DT = < U, C \cup D, V, f >$ be a Decision table, where C and D stand for conditional and decision attributes subsets, $C \cap D = \varnothing$, U is a non-empty, finite set called universe, V is called the value set, f stand for information function.

Definition 4. Let $\varnothing \subseteq X \subseteq C$, $\varnothing \subseteq Y \subseteq D$, $U/Y \neq \{U\}$, given $x \in X$, define significance of x with X (comparing with Y): $sig^Y_{X-\{x\}}(x) = (|S_X(Y)| - |S_{X-\{x\}}(Y)|)/|U|$.

2 Relative Work and Background

Although the emergence of GO project has been used to solve the problem of unification of nomenclature successfully, there is another remarkable problem about how to apply these nomenclatures on large-scale data effectively.

At present, a number of automated BLAST-based GO term prediction applications have been published. BLAST is the most widely used sequence alignment tool [3, 4]. It permits the user to find similar sequence according to high degrees of local similarity. Normally, it is very likely that similar sequences might be homological; therefore, the similar sequences may have the same or similar functions. For these reasons BLAST has been employed to assign GO terms to a novel sequence. Nowadays, there are several methods with the idea of predicting the function of gene product using BLAST and GO, such as TOP BLAST, GOtach, GOFigue, Goblet and some others [5-10]. These approaches can be roughly divided into several main kinds: graph-based, discriminant function-based and term distance concordance-based and so on. Among them the TOP BLAST is the most commonly used approach. However, TOP BLAST is not so accurate and convincing. As a result, this paper recommends a new design of BLAST-based GO term annotator which incorporates data mining techniques and utilizes rough set theory. Under the strict criterion, the new approach provides higher quality and more accurate functional prediction for a novel sequences than TOP BLAST can.

3 Rough Set-Based Method

3.1 Data Collection

The Gene Ontology data were downloaded and divided into three parts: training set, test set and BLAST-able database. This data consist of protein sequence data and their GO term associations. UniPort annotations, proteins and their GO term associations are

submitted by UniPort, is referred to as BLAST-able database. This data, consisting of 107,632 proteins, have high quality annotation. Non UniPort annotations, consisting of 3,537 proteins and their GO term associations are submitted by other sources, are referred to as training set and test set. In order to examine our method's validity, we employ cross-validation method. Each time we randomly select 1,200 proteins as test set and the other 2,337 proteins as training set.

Evidence code indicates how annotation to a particular term is supported. Some are supported by experiments, some are supported by literature and some are supported by computation method. According to different evidence codes, for training set and test set respectively we constructed 2 different experimental sets: one experimental set, called 7-evidence set, includes GO terms supporting by evidence codes such as: TAS, IDA, IC, IMP, IGI, IPI and IEP. Another experimental set, called NoIEA set, includes GO terms supporting by all evidence codes except IEA. For the reason that all GO terms within 7-evidence set are supported by evidence code which have high reliability, meanwhile the GO terms within NoIEA set just preclude those supported by evidence code of IEA, there is no doubt that GO terms in 7-evidenc are more reliable and accurate than those in NoIEA.

3.2 Accuracy Metrics

As we employ the strict evaluation method, precision and recall rate are defined as:

$$\text{Precision: } P = \frac{c}{p}$$

Where c is the number of correct predicted term assignments and p is the total number of predicted assignments.

$$\text{Recall rate: } R = \frac{c}{t}$$

Where c is the number of correct predicted term assignments and t is the total number of correct term.

$$\text{Harmonic Mean: } H = \frac{2}{1/P + 1/R}$$

Only if the predicted term is the right term which the source sequence indeed has, we count it as a correct prediction. Otherwise, prediction hit on either its parent term or its children term is considered as a false prediction.

3.3 Preparation

Before deducing rules from decision table, there are some preparation works to do.

3.3.1 Basic Concept

(1) Source sequence: we define those protein sequences which need prediction of function in training set as source sequence.