International Journal of Bio-Science and Bio-Technology
Vol.7, No.2 (2015), pp.57-70
http://dx.doi.org/10.14257/ijbsbt.2015.7.2.06

A Review of Software for Predicting Gene Function

Swee Kuan Loh[a], Swee Thing Low[a], Mohd Saberi Mohamad[a,*], Safaai Deris[a], Shahreen Kasim[b], Choon Yee Wen[a], Zuwairie Ibrahim[c], Bambang Susilo[d], Yusuf Hendrawan[d], Agustin Krisna Wardani[d]

[a] *Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia,81310, Skudai, Johor, Malaysia.*
[b] *Faculty of Computer Science and Information Technology, Web Technology Department, Universiti Tun Hussein Onn, 86400, Parit Raja, Batu Pahat, Johor, Malaysia.*
[c] *Faculty of Electrical and Electronics Engineering, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia.*
[d] *Faculty of Agricultural Technology, University of Brawijaya, Jl. Veteran Malang, ZIP 65145, Indonesia.*
helenloh.sk@gmail.com, lowsweething@gmail.com, saberi@utm.my, safaai@utm.my, shahreen@uthm.edu.my, ywchoon2@live.utm.my, zuwairie@ump.edu.my, bmsusilo@gmail.com, yusufhendrawan@yahoo.com, wardani8@yahoo.com
** Corresponding author: Email address: saberi@utm.my, Tel: 60-7-5533153, Fax: 60-7-5565044.*

Abstract

A rich resource of information on functional genomics data can be applied to annotating the thousands of unknown gene functions that can be retrieved from most sequenced. High-throughput sequencing can lead to increased understanding of proteins and genes. We can infer networks of functional couplings from direct and indirect interactions. The development of gene function prediction is one of the major recent advances in the bioinformatics fields. These methods explore genomic context by major recent advances in the bioinformatics fields rather than by sequence alignment. This paper reviews software related to predicting gene function. Most of these programs are freely available online. The advantages and disadvantages of each program are stated clearly in order for the reader to understand them in a simple way. Web links to the software are provided as well.

Keywords: *Gene Function, Artificial Intelligence, Bioinformatics, Cancer, Functional Genomics*

1. Introduction

The increase in the amount of genomic sequence data thanks to the Genome Project has increased the numbers of unknown gene. Hence, it is necessary to develop computational approach to discover the functions of unknown genes using the vast amount of information that is produced by genomic projects, such as gene expression data. Gene expression is the process involving splicing that is the central dogma for the production of RNA and gene products [1]. The increase in functional proteomic and functional genomic methods has increased the need for analysis of gene function. Examples of functions that are focused on in this field are gene translation, transcription, interaction among proteins and others. Gene function is important in bringing information of RNA transcripts, function of DNA at the level of genes and protein products. Discovering gene function serves as a goal that reveals relationships between the

ISSN: 2233-7849 IJBSBT

phenotype and genome of an organism. Hence, a blueprint about how each gene functions in the cellular and molecular process is clearly provided. This leads to the possibility of understanding which genes and how they are involved in a particular mutation that is implicated in a human genetic disease. The work of discovering gene functions is not only important to understanding the biology of a particular organism but also plays an important role in identifying potential therapeutic targets. Gene function is difficult to process in their natural languages. However, there are several research areas that provide assistance in understanding gene function, such as gene regulatory network construction. A gene regulatory network represents the functional interactions among genes in a graphical view and applies functional genomics to predicting novel gene functions [2]. Genetic interaction is very informative for interpreting gene-gene functional linkages, as stated in You *et al*. [3]. It is extensively exploited for the dissection of specific pathway structures and annotating gene functions. A readable compute program should be developed to deal with gene data that can automatically annotate the huge streams of sequence produced by the genome sequencing project, such as in Delcher et al. [4] and Alexandersson et al. [5].

According to Zhou et al. [6], one novel data mining algorithm is Ontology-based pattern identification (OPI). Gene function expression patterns are identified based on the best existing knowledge. Gene ontology is a widely used in the study of gene function prediction. There are three ontologies in gene ontology: molecular function, the cellular component and biological process. A number of programs have been built to search, browse and process Gene Ontology. The predicting gene function is one of the major challenges in the field of biology.

Gene function prediction is essentially the study of a classification problem based on machine learning techniques. Computational prediction methods give assistance in suggesting a restricted set of candidate functions that are verified through experimentation. Moreover, new hypotheses are generated directly. Also, guidance in exploring promising hypotheses is one of the benefits of using computational methods in prediction. The great availability of biological data has led to new directions in predicting function from sequence information and genomic analysis. The prediction of gene function can be accomplished by integrating several data sources. Genes with similar functions may not be clustered as a group in the clustering process. The specific interaction between genes needs to be obtained. Hence, several programs have been developed which contribute to this task. In this paper, software used to predict gene function will be introduced and summarized. The popular gene function prediction programs, like GeneMark and GeneFAS will be introduced. Some web-based programs have their own databases to allow the user to insert a dataset, such as a protein query, to be analysed. Software is written in different languages such as C language and Perl language. The search for gene function is initiated either in protein form or web form. The advantages and disadvantages are stated in order to let the user understand in a simple and clear way. A link to a web page with the software is also provided. Which software is more reliable and performs better at predicting gene function will be explained. The user should concern him/herself about the dataset to be provided and the parameters to be used in the predictions.

Several problems have been found in predicting gene function. Very large numbers of functional classes can be found in Gene Ontology (over 1000) and FunCat (over 100). Taxonomies of gene function included Gene Ontology (GO) and Functional Catalogue (FunCat). Gene Ontology is a functional term set with some annotated genes that is widely used in the research area. Gene Ontology is widely used to classify gene function [7]. Gene Ontology is comprised of biological process ontology, cellular component ontology and molecular function ontology. Each gene or protein is associated by Gene Ontology with some functional term. The output in Gene Ontology is constructed based on the direct acyclic graph structure. The predicting gene function for an unknown gene involves assigning some Gene Ontology functional term to it. In the machine learning

community, this classification is called a multi-label classification problem. In contrast, FunCat produces a hierarchical tree-like structured output. FunCat only concentrates on the functional process. Hence, it is more stable than Gene Ontology.

This paper provides basic concepts about gene function and its importance in the field of research. This review paper provides and reviews several programs involved in predicting gene function. A brief description of each program is given in this review paper to allow users to gain better understanding of these programs. Web links that access these programs together with some other features such as the platform and the programming language implemented in the program are also provided in table form. Moreover, the main advantages and disadvantages of each program are listed and discussed in the discussion section.

2. Gene Function Prediction Tools

Many programs have been proposed to infer gene function. Nowadays, gene function prediction software provides inference services based on many features such as sequence and structure information, functional ontologies and heterogeneous data sources. Table 1 shows the sources of the gene function prediction programs that are discussed in this paper.

Table 1. Sources of Gene Function Prediction Software

Gene function prediction software	Author / Year	Web Link	Platform	Implemented Language
SNAPper	Kolesov *et al.*[8]	http://pedant.gst.de/snapper	Web-based	C
Funcassociate	Berrizl *et al.* [9]	http://llama.mshri.on.ca/cgi/func1/funcassociate	Linux, Windows	C, Perl
OntoBlast	Zehetner [10]	http://functionalgenomics.de/ontogate	Web-based	C
GAIN	Karaoz *et al.* [11]	http://genomics10.bu.edu/gain/supplements/cv.html	Window, Linux	C++, Perl
GeneFAS	Joshi *et al.*[12]	http://digbio.missouri.edu/genefas	Window, Linux, MacOsX	Java
GFINDer	Masseroli *et al.*[13]	http://www.medinfopoli.polimi.it/GFINDer	Web-based	Java
GOToolBox	Martin *et al.*[14]	http://gin.univ-mrs.fr/GOToolBox	Web-based	Perl
Blast2GO	Conesa *et al.*[15]	http://www.blast2go.de	Window, Linux, MacOsX	Java
Biopixie	Myers *et al.*[16]	http://neurolex.org/wiki/Category:Resource:bioPIXIE	Linux, Windows	PHP, Perl
GeneMark	Basemer and Borodovsky [17]	http://exon.gatech.edu/	MacOS X,Linux, AIX, Solaris	C
Phydbac "Gene Function Predictor"	Enault *et al.*[18]	http://www.igs.cnrs-mrs.fr/phydbac/indexPS.html	Web-based	Perl
VIRGO	Massjouni *et al.*[19]	http://whipple.cs.vt.edu:8080/virgo	Web-based	Java

GOBlast	Wang and Mo[20]	http://bioq.org/goblast	Linux	Perl
SynFPS	Li *et al.*[21]	http://www.synteny.net	Windows	C#
ChemGenome	Singhal *et al.*[22]	http://www.scfbio-iitd.res.in/chemgenome/download.htm	Linux, Windows, Solaris	C
HCGene	Valentini and Cesa-Bianchi [23]	http://homes.di.unimi.it/~valenti/SW/hcgene/node1.html	Linux, Windows, MacOsX	R
Prosecutor	Blom *et al.*[24]	http://www.prosecutor.nl	Window, Linux, MacOsX	Java
GeneMANIA	Warde-Farley *et al.*[25]	http://www.genemania.org	Web-based	Java
GO-*At*	Bradford *et al.*[26]	http://www.bioinformatics.leeds.ac.uk/goat	Web-based	Java
PANTHER 7	Mi *et al.*[27]	http://www.pantherdb.org	Web-based	Java
PhyloProf	Valentini *et al.*[28]	http://dna.cs.byu.edu/phyloprof/cli.php	Web-based	Java
PlasmoPredict	Tedder *et al.*[29]	http://www.bioinformatics.leeds.ac.uk/~bio5pmrt/PlasmoPredict/PlasmoPredict.html	Web-based	Java
AraNet	Hwang *et al.* [30]	http://www.functionalnet.org/aranet	Web-based	Java
Eukaryotic GeneMark	Borodovsky and Lomsadze [31]	http://topaz.gatech.edu/eukhmm.cgi	Sun Solaris, Unix, Linux, AIX	C
Argot2	Falda *et al.*[32]	http://www.medcomp.medicina.unipd.it/Argot2	Web-based	Java
FunCoup	Alexeyenko *et al.*[33]	http://funcoup.sbc.su.se/	Web-based	Java

2.1. *SNAPper*

SNAPper is software developed by Kolesov *et al.* [8] to predict gene function according to the conservation of gene order. A SNAPper search can be performed by inserting a protein sequence. SNAPper is part of PEDANT, a server analysis of the genome. Relational databases store PEDANT data and can be assessed by the SNAPper server if allowed. Properties of gene products can be corrected by using SQL queries. Parameters like number of gene neighbours and finding orthologs are set by default to decrease the number of false positives. SNAPper can be implemented in C or with a UNIX environment. The service is developed on the PEDANT, which is used as a genome analysis server. Each gene product can be subjected to searches for protein motifs, protein features and homology.

2.2. *Funcassociate*

Funcassociate is a web-based program built by Berrizl et al. [9] to assist users in applying Gene Ontology (GO) attributes to characterize large datasets of genes. Its distinguishing features include a Monte Carlo simulation and the ability to handle ranked input lists. The source code was developed in C and Perl language. Currently, Funcassociate can support ten organisms (*Shewanellaoneidensis, Vibrio cholerae,*

Saccharomyces cerevisiae, Caenorhaebditiselegans, Arabidopsis thaliana, Schizosaccharomycespombe, Drosophila melanogaster, Rattusnorvegicus, Homo sapiens and *Musmusculus*). An adjusted p-value is estimated from 1000 simulated queries of the null hypothesis result. The ability to restrict the genes in the universe used in the analysis with other features is not found in other similar programs. It is written in C language, while Perl is used in the CGI script that deals with requests over the web.

2.3. *OntoBlast*

Zehetner [10] introduced OntoBlast, which is an inner function of Ontologies TO GenomeMatrix software for functional prediction purposes. OntoBlast utilizes functional ontology and sequence similarities to predict gene function. The objective is to overcome the challenge from gene functional inference by integrating gene sequence data. Gene Ontology is used by OntoBlast to support the functional annotation process. Search frame and list frame are two major aspects of OntoBlast. For the search frame, a result from the BLAST server, which is based on sequence similarity, is generated. Meanwhile, a list of functional ontology terms with different weighting numbers is provided in the list frame. Then the BLAST result and ontology terms are associated together for further prediction of gene function. A smaller weighting number indicates strong evidence for the similarity of their sequence data. Two genes that share the functional annotation may indicate they are sharing a similar function.

2.4. *GAIN*

Karaoz *et al.* [11] developed GAIN, the Gene Annotation using Integrated Networks program to predict gene function using a computational approach. It operates in a functional linkage network. In the graph, nodes represent genes and edges represent how they are connected. In GAIN, each gene is predicted independently and multiple predictions can be made for the same genes. A single organism of FLN is constructed by integrating functional genomic information. Several algorithms are included in GAIN to systematically propagate annotations. Most of them are semi-supervised learning machine languages. Positive and unknown examples are easy to generate in GAIN but it is hard to generate negative samples. Positive samples are gene-function pairs, to which unknown samples with some of the same categorization are considered.

2.5. *GeneFAS*

GeneFAS was introduced by Joshi *et al*. [12] and stands for Gene Function Annotation System. It combines various types of high throughput biological data sources such as microarray data, interaction among proteins, information about protein complexes and functional annotation for gene function prediction based on a probabilistic method. GeneFAS assesses functional relationships in the Gene Ontology hierarchy and correlations between proteins from microarray data. GeneFAS carries out a two-step mechanism for functional predictions. The observed frequency related to the unknown functions of genes based on the high throughput data information is the *a-priori* probability. The estimations of *a-priori* probabilities are applied to the investigations of functional similarities and interactions between genes for each data set with high throughput. Subsequently, prediction of function for an uncharacterized gene is carried out based on the estimated *a-priori* probabilities.

2.6. *GFINDer*

Masseroli *et al.* [13] developed Genome Function Integrated Discover (GFINDer), which is a web server for functional prediction. Annotations of functional categories are extracted by GFINDer from multiple data sources. The main feature of this software is providing an evaluation of the annotation relevance for estimating the functional bias that exists in a microarray gene set. The statistical significance values based on the gene's membership in a particular functional category are calculated. This statistical analysis on extracted functional categories is then applied to classify the gene's function. The three layer architecture of GFINDer includes a data layer, processing layer and user layer. Various types of data sources and annotations are handled by the data layer. The processing layer is the main layer responsible for managing client requests. However, the user layer provides a graphical user interface for users to access the processing layer.

2.7. *GOToolBox*

Martin *et al.* [14] developed a collection of software and methods for gene function prediction called GOToolBox. PERL is the programming language used to write GOToolBox programs. Overrepresented and underrepresented gene terms in a dataset can be identified according to the Gene Ontology. The genomic frequency of Gene Ontology functional terms embedded with genes is computer for the creation of dataset in gene function prediction. GO-Family and GO-Proxy are two of the major functions in this toolbox. Those genes that are functionally related are clustered together by the GO-Proxy program using calculations of gene distance based on Gene Ontology annotation terms followed by a clustering algorithm. However, GO-Family is responsible for searching the genes that share the same Gene Ontology annotation terms with the gene in the user query based on measurements of functional similarity.

2.8. *Blast2GO*

Conesa *et al.* [15] proposed a desktop application program written in the Java language for functional annotation, Blast2GO. It aims to provide high throughput and automated sequence annotation for unannotated sequences based on similarity searches and allows functions to be associated with annotated genes. Blast2GO finds homologs using BLAST software. Gene Ontology functional terms that are retrieved according to the Blast hit sequences are assigning to a query sequence based on an annotation rule. An annotation score which by possibility of abstraction and the highest similarity hit of Gene Ontology term is calculated for each Gene Ontology candidate term for annotation assignment. Gossip is used to introduce the gene functionality in Blast2GO. The main feature of this software is its ability to visualize the analysis of functional genes in graphic form.

2.9. *Biopixie*

Myers *et al.* [16] developed bioPIXIE, which can predict the components in a novel network based on different types of input data but does not simply recapitulate known biology. A general probabilistic system is developed for query-based discovery of pathway-specific networks through integration of diverse genome-wide data. bioPIXIE is a biological network predictor for *S. cerevisiae* which is a comprehensive and public system for visualization, integration and analysis. It is publicly available over the World Wide Web. It is convenient for researchers to use. The advantage of this software is it can incorporate different types of genomic data. The result shows that bioPIXIE dramatically increases the number of network components recovered with any of the individual types of evidence allowed. Researchers are enabled to study the interactions and novel pathways of components.

2.10. *GeneMark*

Basemer and Borodovsky [17] developed the GeneMark web software, which provides an interface to their program designed to predict genes in eukaryotic, prokaryotic and viral genomic sequences. The server can analyse approximately 200 prokaryotic and more than 10 eukaryotic using pre-computed gene models and species-specific versions of the software. Genes in prokaryotic sequences from novel genomes can be determined using models derived on the spot upon sequence submission. This can either be done using the full-fledged self-training program GeneMarkS or a simple heuristic approach. A database of re-annotations of more than 1000 viral genomes by the GeneMarkS program is also available from the website. In order to provide the latest versions of the gene models and software, the GeneMark website is frequently updated.

2.11. *Phydbac "Gene Function Predictor"*

The Phydbac "Gene Function Predictor" is gene annotation software introduced by Enault *et al.* [18]. This gene function predictor links to the Phydbac web server, which is written in Perl language. It consists of two major operating modes which are prediction and database gathering. For prediction, a consensus profile for the query sequence is generated and profiles that have the greatest similarity are retrieved. In addition, the query sequence is compared with all organisms to identify conserved neighbours on chromosomes. For database gathering, results that are related to the processed organisms are gathered. The results of database gathering are then retrieved using gene names or annotation keywords. The level of prediction confidence is displayed discriminately in terms of colours and keywords. The gene ontology functional annotation assignment is carried out after the association partner is identified.

2.12. *VIRGO*

Massjouni *et al.* [19] presented a function prediction web server called VIRtual Gene Ontology implemented in Java language, due to the massive existence of unknown gene functions. At first, microarray data are associated with molecular interaction data in order to generate a network of functional associations. Gene ontology annotations are embedded in each node in the network graph. Finally, the annotation labels are propagated across the interaction network to make predictions about gene function. This software employs Gene Annotation using Integrated Networks (GAIN), implemented in C++ language, as the prediction engine for functional genes. Leave-one-out cross validation is performed to evaluate the performance of the GAIN algorithm. Another significant feature of this service is confidence estimates are provided to indicate the quality or priority of prediction for each unknown gene. A propagation diagram is also provided through this prediction service. However, *H. sapiens* and *S. cerevisiae* are the two organisms that VIRGO currently supports, since very large datasets are available.

2.13. *GOBlast*

Wang and Mo [20] presented a software system that combines Gene Ontology annotations and Blast search for gene functional inference called GOBlast. There are three Bioperl modules that together constitute GOBlast functionality. For the data input module, a nucleotide or amino acid sequence in FASTA format with Gene Ontology annotation terms is used as input. For the data processing module, the input sequence with gene ontology annotation is used to implement a Blast search for sequence alignment. The Gene Ontology terms within the Blast results are selected for the retrieval of Gene Ontology information in order to submit to the next module. For the data output module, the genes scoring the highest on homology with the query gene and associated annotation

terms are shown. The associated annotation terms provide a description of the specific function of the gene.

2.14. *SynFPS*

Li *et al.* [21] developed SynFPS, the Synteny-based Function Prediction System, in C# programming language. According to resemblances in gene distribution, a set of weakly related genomes is grouped together using a K-means clustering algorithm. Genes with similar functions are identified using regular expressions rather than sequence similarity. Distances between genes are taken into account in identifying genomes that are closely related. Subsequently, negative and positive training datasets are extracted from each group for supervised learning on gene function by a support vector machine. The negative training dataset is the neighbours to positive data, whereas the positive training dataset consists of gene groups that were identified by the system previously when matching the genes that were added manually with regular expression. This program only aims for functional predictions in the whole genome context based on the synteny of genes or the order of gene conservation and distances between genes.

2.15. *ChemGenome*

ChemGenome is ab-initio gene prediction software developed by Singhal *et al*. [22]. Physico-chemical properties are used to construct a three dimensional vector to predict genes in Prokaryotic Genomes by obtaining the stacking energies and base pairing for each codon from the report. There are 64 codons in the CG2 model, each of which is assigned to a biology interaction such as stacking, hydrogen bonding and propensity interaction based on molecular dynamic simulation. The parameters of the software are calculated from MD stimulation. ChemGenome shows better results in differentiating genes from non-genes at an equivalent level. It brings out the possibility of useful and unique ab initio. Its results show that differentiating genes from non-genes gain similar better than the knowledge model trained. ChemGenome can be run in a Linux environment. The input requires a FASTA format genome sequence. The threshold value for a small genome should be set at a lower value and vice versa. Several methods are used in ChemGenome which include *swissprot* space, protein space and DNA space. It will predict prokaryotic coding regions if the user enters all or part of a genome.

2.16. *HCGene*

Valentini and Cesa-Bianchi [23] introduced Hierarchical Classification of Genes (HCGene) for gene function prediction. It is an R package software library for supervised gene function classification. Subtrees related to biological problems, subgraph extraction, gene labelling and multiple gene products are allowed in HCGene. It also supports collaboration with semi-supervised and unsupervised methods. HCGene can analyse ontology taxonomies, which are Functional Catalogue and Gene Ontology for labelling functional classes. It provides three major capabilities, namely, generation of multi-labels, graphing and data processing. HCGene associates functional class labels that are derived from ontology taxonomies and different types of biological data, such as gene expression data, to the queried genes for functional inference. Several methods can be included in the extraction of Gene Ontology subgraphs and Functional Catalogue subgraphs.

2.17. *Prosecutor*

Prosecutor is a Java based gene function inference engine that is mainly used for prokaryotes and was proposed by Blom *et al*. [24]. It is parameter-free software. Prosecutor treats each functional category as individual; hence, a trusted set of protein interactions with function information is not necessary. It fully utilizes various annotation

data, gene expression data and extra biological knowledge in a genomic context to predict gene function. Prosecutor predicts the functions of genes based on an iterative guilt-by-association principle without a fixed cutoff. Genes with unknown functions are linked to genes with annotation terms at a high association rate using a sensitive algorithm. The significance of all relationships between functional annotation and genes is measured. As a result, the predictions of gene function visualize as a force-directed network layout graph using a Prefuse toolkit.

2.18. *GeneMANIA*

The GeneMANIA prediction server is a gene function prediction web interface introduced by Warde-Farley *et al.* [25]. GeneMANIA generates hypothetical gene function, gene list analysis and prioritization of genes. According to available genomic and proteomic information, genes that have similar function as the query gene are appended into the gene list and visualized as a functional linkage network. In addition, the value of a prediction can be represented by assigning it a weight. By default, the weighting method that is assigned based on the query genes is employed for weight assignment. This assigned weight of the query genes indicates the predictive value of each dataset. All non-query genes are scored using label propagation and finally, for ranking purposes. GeneMANIA currently supports six different types of organisms.

2.19. *Go-At*

Bradford *et al.* [26] presented Gene Ontology prediction in *Arabidopsis thaliana* (GO-*At*), which is a two-step web-based application since *Arabidopsis thaliana* is still poorly characterized. It utilizes the confluence of five different data sources, namely, the gene neighbourhood, interaction information, phylogenetic profile, sequence information and co-expression data in order to make functional inferences. In the first step, the probability of a functional interaction with the gene queried by the user is applied to automatically construct a ranked gene list in descending order of the probability of an interaction with the query gene and an associated score of prediction for each function. The assignment of the prediction score is based on the observed frequency of a function in the gene list that most probably represents the query gene function. In the second step, the predicted gene function with the best score is chosen.

2.20. *PANTHER 7*

Mi *et al*. [27] proposed PANTHER 7, which denotes Protein Analysis Through Evolutionary Relationships, for functional inference. PANTHER 7 predicts gene functions according to relationships across evolution and genes with known functions determined from experiments. Complete genome sequences enable analysis of evolution and the understanding of phylogenetic relationships [34]. PANTHER 7 generates phylogenetic trees using multiple protein alignments and associates the trees with annotation terms to explain the evolution of gene function. The phylogenetic trees are estimated using multiple alignments of proteins by the MAFFT alignment program. Events of gene duplication and speciation can also be represented by enhanced phylogenetic trees. Gene duplication events and gene trees are predicted using the GIGA algorithm. Determination of orthologs is also included in the gene function predictions.

2.21. *PhyloProf*

According to Valentini *et al*. [28], PhyloProf is a four-stage process which includes a fuzzy genome profile constructed for each participating genome, after which the fuzzy profile is discretized, initial profile data is de-noised and the intra-inter genome distance is calculated. A distance diagram is produced and divided into four areas: bottom right

quadrant, top left quadrant, main diagonal and bottom left corner. The most interesting genes are presented based on their intra-inter distances. The interface was built using a combination of CSS, XHTML and PHP server scripting. It performs well for small tasks like querying the database of nucleotides. PhyloProf places an emphasis on the user experience and an aesthetic user interface. The results page shows each query gene. The query gene indicates which genomes had a gene homolog. It is very simple to use and yet a very powerful program for gene regulatory networks.

2.22. *PlasmoPredict*

Tedder *et al.* [29] presented a web-based gene function prediction program called PlasmoPredict, since the gene functions of *Plasmodium falciparum* are for the most part still unknown. This software utilizes various types of data in a genomic context, such as high throughput data and structure or sequence data for functional inference. The gene ID of *Plasmodium falciparum* acts as the input. The output is a gene list that is functionally related to the query gene. PlasmoPredict is implemented based on the guilt by association principle. Moreover, the transfer of gene ontology annotation is also provided in this software. The prediction results are visualized as a network graph by the Medusa program. Each network edge is assigned a confidence value that indicates how reliably the two genes share a similar function. The confidence value can be measured by Bayesian classifier and functional annotations.

2.23. *AraNet*

Hwang *et al.* [30] developed AraNet, a network-based functional prediction software for the plant *Arabidopsis thaliana* operating through a web interface. AraNet provides the prioritisation of genes for the functional screening based on the candidate genes. Two search options are provided by the AraNet website which can either search for new candidates from the pathway or predict a gene function according to its neighbours in the network. AraNet produces a functional linkage network with probabilistic criteria via integrating heterogeneous data and modified Bayesian network. Each edge in the functional linkage network is a log-likelihood score to indicate the probability of the positive linkage representation. AraNet collects novel information regarding gene function from the network's neighbours for prediction.

2.24. *Eukaryotic GeneMark*

Borodovsky and Lomsadze [31] developed Eukaryotic GeneMark, one of several programs used to find eukaryotic genes. Two protocols are used to predict gene function in GeneMark program. The first protocol is called GeneMark.hmm. The second protocol is called GeneMark-ES, which performs self-training on the anonymous input sequence and an estimation of species specific parameters of the HMM model is generated. The programs GeneMark-ES and GeneMark.hmm-E are used to find genes in eukaryotic genomes. The GeneMark program may provide valuable information about analysis, especially in eukaryotic genomes. The latest version of the software can be run under the operating system of the LINUX environment. In comparing GeneMark-E and GeneMark.hmm-E, GeneMark-ES requires more resources.

2.25. *Argot2*

Argot2 is a program introduced by Falda *et al.* [32] to support gene functional inference at a large scale. It stands for Annotation Retrieval of Gene Ontology Terms. It provides a function prediction service through a web interface. The input of Argot2 is the list of sequences, which is in FASTA format. A weighting value is assigned to each annotation term based on its BLAST and HMMER e-values. Furthermore, the most

accurate gene ontology terms are selected by clustering methods and finally associated with those input sequences. Semantic similarities are taken into account in the processing of weighted gene ontology terms. The reference databases currently supported by Argot2 are Pfam and UniProt. The database server can be accessed by three different methods, i.e. interactive analysis, batch analysis and consensus analysis.

2.26. *FunCoup*

FunCoup is a database developed by Alexeyenko *et al*. [33] that visualizes and maintains global protein or gene networks of functional coupling. Bayesian integration is used to construct the diverse high-throughput data. Functional coupling information is transferred between species. In version 2 of FunCoup, the new input datasets have been improved in term of network quality and coverage. There is a dramatic increase in the number of high-confidence network links. In the latest version, the human network has been increased eight-fold. FunCoup performs comparative interactomics such as aligning sub-networks between different species using orthologs. There are major increases in data of higher quality and in comprehensive data. This brings benefits such as an increase in total evidence and more accurate predictions. For example, using the FunCoup network, a gene's functional relatedness to Alzheimer's disease can be analysed or determined by the enrichment of common interactors.

3. Discussion

Many programs have been designed for gene function prediction based on different criteria. However, each program has its own strengths and weaknesses. Some programs only support gene function prediction for one organism while others can support more than a single species. For example, AraNet only supports one organism, *Arabidopsis thaliana*. Then, GeneFAS only supports *Sacchromycescerevisiae* whereas the Phydbac "Gene Function Predictor" only supports *Escherichia coli*. In addition, there are only a very few programs that focus on plants. This may be due to the scattered resources and literature in the plant area [35]. Furthermore, some programs predict gene function based on a functional linkage network or phylogenetic trees. However, most programs use Gene Ontology annotation and integrate heterogeneous data sources to facilitate gene function prediction. Table 2 shows the main advantages and limitations of gene function prediction software.

Table 2. Main advantages and limitations of gene function prediction software

Gene function prediction software	Main advantages	Main limitations
Funcassociate	The CGI script of FuncAssociate's and is written in Perl while the software is written in C using standard modules	Funcassociate can support up to a maximum of ten organisms
OntoBlast	Allows comparison of all genes from nine organisms simultaneously	Leads to false positive ontology terms due to insignificant sequence similarities
GAIN	Operating in functional linkage network so that multiple predictions can be made for the same genes	Negative samples are hard to generate
GeneFAS	Predictions can search by either partial or complete matching yeast gene names	Limited to *S. cerevisiae*
GFINDer	Provides evaluation of statistical significance of functional categories	The analysis and annotation are only as accurate as the

		employed annotation databases
GOToolBox	Provides GO-Diet software that allows a certain ontology depth to be selected so as to restrict GO terms and reduce the number of GO terms in forming a slim GO hierarchy	It does not support a directed acyclic graphical output option
Blast2GO	Easy to distribute and low maintenance	Limited to Gene Ontology annotations
Biopixie	Multiple types of genomic data are integrated for 3 diverse KEGG pathways, and robust enhanced integration of different types of genomic data	Pre-processed data cannot be classified
GeneMark	Pre-computed gene models can be processed and models derived on sequence submission identified prokaryotic gene sequences from novel genomes	Higher quality data are needed
Phydbac "Gene Fucntion Predictor"	Useful for incomplete sequences due to blast mode	Limited to *Escherichia coli*
VIRGO	Assigns confidence estimates to allow biologists to prioritize gene function predictions for further analysis	Only supports *H. sapiens* and *S. cerevisiae* currently
GOBlast	Gene Ontology and Blast search are combined into a single step to save time	The input sequences need to satisfy the format for the *wublast* program
SynFPS	Determines genes that are functionally similar using regular expressions	Only extendable to other organisms that have properties similar to phage genomes
ChemGenome	Only simple parameters are needed for predictions and the stability of hydrogen bonding between the DNA bases. Prokaryotic and eukaryotic genome gene finders are enhanced	There is no different data is integrated
HCGene	Capable of integrating heterogeneous data sources and processing graphs and multi-labels. Annotations for gene products and genes are given. Both hierarchies function at different degrees of resolution and multi classes involved are provided	Does not have a hierarchical classification algorithm since it is more prone to facilitate gene classification algorithms
Prosecutor	Parameter free and has additional prokaryote-specific information layers	Only focus on prokaryotes and occurrence of false positives is proportional to the number of performed tests
SNAPper	SNAP function prediction is accepted for querying a protein sequence. Representation of hyperlinked graphical is rendered	Definitive of function is not included in the result.
GeneMANIA	Provides gene prioritization and weight assignment	Only supports six different organisms currently
GO-*At*	Predicts unknown gene functions by determining ten putative golgins/Golgi-associated proteins	The prediction performance for biological process and the cellular component still need improvement
PANTHER 7	A phylogenetic tree represents all the	Only model domains that are

	evolutionary events in the gene family and experimental results keep changing	conserved across a single subfamily
PhyloProf	A distance diagram is produced and divided into four areas	Has not been well developed
PlasmoPredict	Includes a wide range of functional genomics data	Only focuses on parasitology
AraNet	Provides gene set enrichment analysis of the valid query genes	It is not suitable for some predictions of particular function and prediction power is influenced by genes that are not well connected
Eukaryotic GeneMark	Gene prediction to find exact exon/intron and gene boundaries has improved using GeneMark.hmm-E and the absence of a training data problem has been solved using the GeneMark-ES algorithm	A longer time cost for predictions
Argot2	Highly scalable and able to annotate from small gene sets to whole genomes	User needs to provide the BLAST and HMMER search results
FunCoup	Great increase in higher quality data; facilities for subnetwork conservation analysis are provided; possibility of performing comparative interactomics	The need to set parameter values such as network distance and identify the gene set

4. Conclusion

In the previous sections, numerous approaches for the computational prediction of gene function from various types of biomedical data were discussed. The analysis of biological data involves handling a number of challenges, many of which have only been partly addressed. Some of the most prominent challenges are widely varying sizes of functional classes and most classes being very small, hierarchical arrangement of functional labels as in Gene Ontology, and incompleteness and various types and extents of noise in biological data. Each type of biological data usually has a strong correspondence with a certain type of function that can be best predicted using data sets of that type. The Gene Ontology is increasingly being established as the most appropriate functional classification scheme for protein function prediction research because of its several desirable properties, and the forward looking attitude of its curators who are keeping it up-to-date with latest research. In particular, several GO-friendly approaches have recently been proposed, which incorporate the hierarchical structure of GO in the prediction technique so as to exploit the parent-child relationships between various functional classes. GO contains separate ontologies for three different types of protein function, namely cellular component, molecular function and biological process, thus making it easier to identify the most appropriate functional hierarchy to be used for making predictions from biological data of a certain type. Even though many advances have been made in the field of gene function prediction, there is still a lack of understanding of the most appropriate prediction technique for any particular category of genes. Thus, there is a great need for the creation of benchmark datasets and the adoption of a consistent evaluation methodology, standardization will help in the identification of the most appropriate function predictions strategy in a certain context, and the current weaknesses and needs of the field. Last but not the least, we firmly believe that an efficient scientific workflow can be established, in which, first, hypotheses are generated by executing the appropriate function prediction algorithm on the available biological data, then, these hypotheses are validated experimentally, thus leading to confident predictions of a gene's function.

Acknowledgements

We would like to thank the Universiti Teknologi Malaysia for supporting this research through a GUP research grant (Grant number: Q.J130000.2507.05H50) and a Matching grant (Grant number: Q.J130000.3007.00M27).

Conflict of Interest

There is no conflict of interest in this manuscript.

References

[1] M. Bae, Y. Kim, J. Lee, Y. Jung and H. Kim, International Journal of Genomics, **(2013)**.
[2] C. Li, M. Liakata and R. Schucmann, Briefings in Bioinformatics,vol. 15, no. 5, **(2013)**.
[3] Z. H. You, Z. Yin, K. S. Han, D. S. Huang and X. B. Zhou, Bioinformatics, vol. 11, no. 343, **(2010)**.
[4] A. L. Delcher, D. Harmon, S. Kasif, O. White and S. L. Saizberg, Nucl. Acids Res., vol. 27, no. 23, **(1999)**.
[5] M. Alexandersson, S. Cawley and L. Pachter, Genome Research, vol. 13, no. 3, **(2003)**.
[6] Y. Y. Zhou, J. A. Young, A. Santrosyan, K. S. Chen, S. F. Yan and E. A. Winzeler, Bioinformatics, vol. 21, no. 7, **(2005)**.
[7] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. I. Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubinand and G. Sherlock, Nature Genetics, vol. 25, no. 1, **(2000)**.
[8] G. Kolesov, H. Mewes and D. W. Frishman, Bioinformatics, vol. 18, no. 7, **(2002)**.
[9] G. F. Berrizl, O. D. Kingl, B. Bryant, C. Sander and F. P. Roth, Bioinformatics, vol. 19, no. 18, **(2003)**.
[10] G. Zehetner, Nucl. Acids Res., vol. 31, no. 13, **(2003)**.
[11] U. Karaoz, T. M. Murali, S. Letovsky, Y. Zheng, C. Ding, C. R. Cantor and S. Kasif, Proceedings of the National Academy of Sciences of the United States of America, **(2004)**.
[12] T. Joshi, Y. Chen, J. M. Becker, N. Alexandrov and D. Xu, OMICS, vol. 8, no. 4, **(2004)**.
[13] M. Masseroli, D. Martucci and F. Pinciroli, Nucl. Acids Res., vol. 32, no. 2, **(2004)**.
[14] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry and B. Jacq, Genome Biol., vol. 5, no. 101, **(2004)**.
[15] A. Conesa, S. Götz, J. M. G. Gómez, J. Terol, M. Talón and M. Robles, Bioinformatics, vol. 21, no. 18, **(2005)**.
[16] C. L. Myers, D. Robson, A. Wible, M. A. Hibbs, C. Chiriac, C. L. Theesfeld, K. Dolinski and G. Troyanskaya, Genome Biology, vol. 6, no. 13, **(2005)**.
[17] J. Basemer and M. Borodovsky, Nucl. Acids Res., vol. 33, **(2005)**.
[18] F. Enault, K. Suhre and J. Claverie, BMC Bioinformatics, vol. 6, **(2005)**.
[19] N. Massjouni, C. G. Rivera and T. M. Murali, Nucl. Acids Res., vol. 34, no. 2, **(2006)**.
[20] C. Wang and Z. Mo, Journal of Chongqing University, vol. 6, no. 3, **(2007)**.
[21] J. Li, S. Halgamuge, C. Kells and S. L. Tang, BMC Bioinformatics, vol. 8, no. 6, **(2007)**.
[22] P. Singhal, B. Jayaram, S. B. Dixit and D. L. Beveridg, "Biophysical Journal, vol. 94, no. 11, **(2008)**.
[23] G.Valentini and N.C. Bianchi, Bioinformatics, vol. 24, no. 5, **(2008)**.
[24] E. J. Blom, R. Breitling, K. J. Hofstede, J. B. T. M. Roerdink, S. A. F. T. V. Hijum and O. P. Kuipers, BMC Genomics, vol. 9, **(2008)**.
[25] D.W. Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader and Q. Morris, Nucl. Acids Res., vol. 38, **(2010)**.
[26] J. R. Bradford, C. J. Needham, P. Tedder, M. A. Care, A. J. Bulpitt and D. R.Westhead, The Plant Journal, vol. 61, no. 4, **(2010)**.
[27] H. Mi, Q. Dong, A. Muruganujan, P. Gaudet, S. Lewis and P. D. Thomas, Nucl. Acids Res., vol. 38, **(2010)**.
[28] N.C. Bianchi and G. Valentini, Machine Learning in Systems Biology, vol. 8, **(2010)**.
[29] P. M.Tedder, J. R. Bradford, G. A. McConkey, A. J. Bulpitt and D. R.Westhead, Trends in Parasitology, vol. 26, no. 3, **(2010)**.
[30] S. Hwang, S. Y. Rhee, E. M. Marcotte and I. Lee, Nature Protocol, vol. 6, no. 9, **(2011)**.
[31] M. Borodovsky and A.Lomsadze, Curr. Protoc.in Bioinformatics, **(2011)**.
[32] M. Falda, S. Toppo, A. Pescarolo, E. Lavezzo, B. Di Camillo, A. Facchinetti, E. Cilia, R. Velasco and P. Fontana, BMC Bioinformatics, vol. 13, no. 14, **(2012)**.
[33] A. Alexeyenko, T. Schmitt, T. Tharnberg, D. Guala, O. Frings and L. L.Sonnhammer, Nucl. Acids Res., vol. 40, **(2012)**.
[34] H. C. Rawal, N. K. Singh and T. R. Sharma, International Journal of Genomics, **(2013)**.
[35] V. Sharma and I. N. Sarkar,Briefings in Bioinformatics, vol. 14, no. 2, **(2012)**.

