UTMGO: A Tool for Searching a Group of Semantically Related Gene Ontology Terms and Application to Annotation of Anonymous Protein Sequence

Razib M. Othman, Safaai Deris, and Rosli M. Illias

***Abstract*—**Gene Ontology terms have been actively used to annotate various protein sets. SWISS-PROT, TrEMBL, and InterPro are protein databases that are annotated according to the Gene Ontology terms. However, direct implementation of the Gene Ontology terms for annotation of anonymous protein sequences is not easy, especially for species not commonly represented in biological databases. UTMGO is developed as a tool that allows the user to quickly and easily search for a group of semantically related Gene Ontology terms. The applicability of the UTMGO is demonstrated by applying it to annotation of anonymous protein sequence. The extended UTMGO uses the Gene Ontology terms together with protein sequences associated with the terms to perform the annotation task. GOPET, GOtcha, GoFigure, and JAFA are used to compare the performance of the extended UTMGO.

***Keywords*—**Anonymous protein sequence, Gene Ontology, Protein sequence annotation, Protein sequence alignment

I. Introduction

THE Gene Ontology (GO) [1] is a project to provide a rich and comprehensive unified vocabulary to describe genes and their functions and products. The vocabulary is formed as a hierarchy of terms in three main categories: molecular function, biological process, and cellular component. Currently the GO comprises more than 20 thousand terms and is updated every 30 minutes, which tally with the growth activities in the bioinformatics field. The GO terms have been widely used for annotating various protein sets such as in DRTF [2], a database of rice transcription factor; SCOPPI [3], a database of protein domain-domain interactions; NOPdb [4], a database of nucleolar proteome; and Organelle DB [5], a database of protein localization and function.

One of the advantages of the GO terms is that it can cope with synonyms and can describe biological function. Furthermore, the GO terms are linked with approximately 8.2 million associations, 1.9 million different gene products, and with the largest set covering around 1.8 million protein sequences from 0.3 million species. Thence, specific protein sets can easily be compared with respect to common functional features [6], [7], protein databases such as MiGenes [8] and PA-GOSUB [9] can be explored through complicated queries, and large-scale protein database can simply be annotated [10], [11] based on the GO terms. However, direct use of the GO terms to annotate anonymous protein sequences is not easy, especially from small sequencing projects or for species not commonly represented in biological databases. Furthermore, for small group of scientists with little computational background or without appropriate facilities it is a tedious task to annotate those protein sequences.

In this paper, we present UTMGO, a tool for searching a group of semantically related GO terms. The basic structure of the UTMGO is extended to show how it could be applicable to annotation of anonymous protein sequence. Generally, the UTMGO consists of two primary components. The first component is applied to cluster the monolithic GO RDF/XML file into a set of more accessible and understandable files. The second component is employed to search for semantically related GO terms together with protein sequences associated with the terms from the fragmented GO RDF/XML files. The extended version of the UTMGO integrates JAligner engine [12] to perform protein sequence alignment. The JAligner engine has been modified to comply with the extended UTMGO. JAligner is a Java implementation of protein pairwise sequence alignment algorithms. This tool applies the dynamic programming algorithm Smith-Waterman. Both of the versions are described in detail in Section III.

This paper is organized as follows. Section II presents existing tools for annotating anonymous protein sequence. Section III gives the structure of the UTMGO and its extended version. Section IV demonstrates the working of UTMGO

Manuscript received June 1, 2006. This material is based upon work supported by the Malaysian Ministry of Science, Technology, and Innovation (MOSTI) in part under Intensification of Research in Priority Areas (IRPA) grants (Project No. 04-02-06-0057-EA001 and 04-02-06-10050-EAR) and in part under Short Term Research (STR) grant (Project No. 75162).

Razib M. Othman is with the Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, MALAYSIA (corresponding author; phone: 607-5532358; fax: 607-5565044; e-mail: razib@fsksm.utm.my).

Safaai Deris is with the School of Graduate Studies, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, MALAYSIA (e-mail: safaai@fsksm.utm.my).

Rosli M. Illias is with the Faculty of Chemical and Natural Resources Enginneering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, MALAYSIA (e-mail: r-rosli@utm.my).

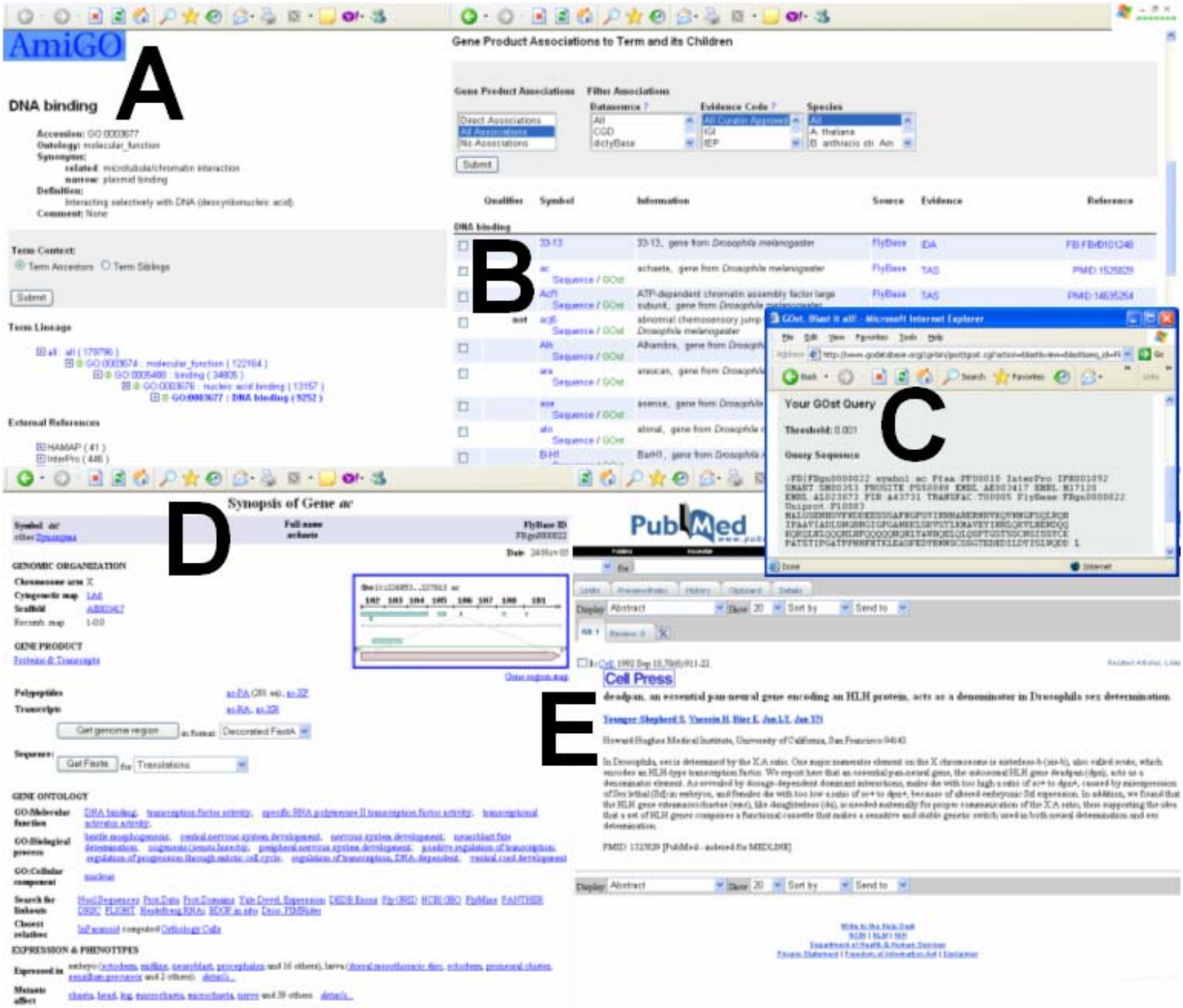

Fig. 1 Example of GO term. (A) The GO term, "DNA binding" (GO:0003677). (B) Gene *ac* (*achaete*), a gene product associated to "DNA binding" based on GOA. (C) Protein sequence of gene *ac* in FASTA format. (D) Synopsis of gene *ac* in FlyBase database. (E) Reference that relate "DNA binding" and gene *ac* in accordance with Traceable Author Statement (TAS) evidence.

(both basic and extended versions) and its comparison with other tools. Section V concludes with a discussion and a brief outline of future improvement directions.

II. Existing Protein Sequence Annotation Tools

Annotation is a process of associating additional information with a particular point in a piece of information. Instead of annotating genomes and protein sets, it has been widely applied in annotating image [13] and web [14]. In the post-genomic era, annotation of protein sequence should be inferred from annotations of the nucleotide sequence, analogies with already understood proteins, plus references to patterns and motifs characteristic of particular protein functions. This can be achieved using the GO terms that have associations with gene products provided by Gene Ontology Annotation (GOA), see example in Fig. 1. In the GOA project which can be accessed at www.ebi.ac.uk/GOA/, electronic mappings and manual curation are used to assign the GO terms with all proteomes that exist in the UniProt Resource (UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, and PIR-PSD), Ensembl, and other biological databases. The GOA statistics for *Homo sapiens* and other species are shown in Fig. 2.

A number of tools have been developed for annotation of anonymous protein sequence based on the GO terms. These

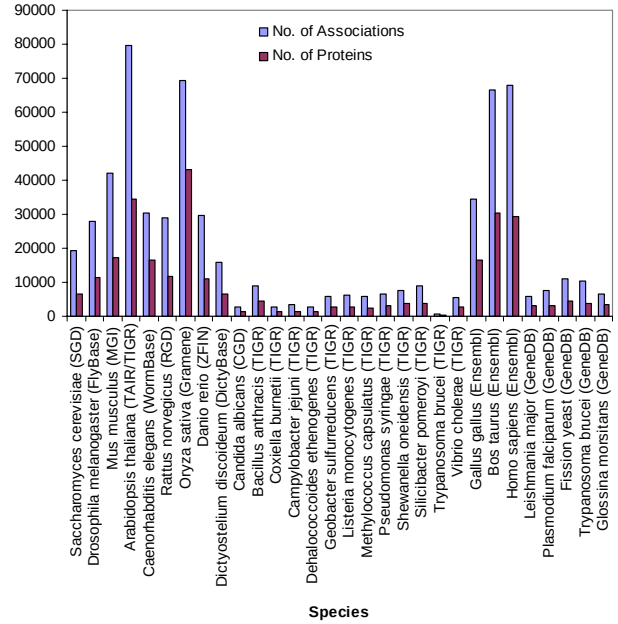

Fig. 2 GOA statistics

tools include:

1) GOPET [15] is an automated annotation tool for assigning the GO terms to cDNA or protein query sequences. It utilizes the GO for annotation terms, BLAST (Basic Local Alignment Search Tool) and GO-mapped protein databases for performing homology searches, and support vector machines for the prediction and the assignment of confidence values.
2) GOtcha [16] is a tool that provides a prediction of a set of GO terms for a given query sequence (DNA or protein). BLAST is used to get the initial score of each GO term and the scores calibrated against term-specific probability (P-score) to give higher accuracy.
3) GOblet [17] is a tool that offers annotation for anonymous cDNA or protein sequences according to the GO terms. It uses the GOA together with a series of protein databases and then employs BLAST to perform annotation by sequence similarity searches.
4) GoFigure [18] is a tool that accepts unknown DNA or protein sequence as an input and then uses BLAST to predict the GO terms by identifying homologous sequences in the GO annotated databases.
5) JAFA [19] is a meta-server that uses several function prediction programs such as GOtcha, GOblet, GoFigure, InterProScan [20], and Phydbac [21]. It accepts a protein sequence and provides predictions based on the GO terms.

In the meantime, an extensive study has been done by Verspoor *et al*. [22] and Xie *et al*. [23] counting on computational linguistic techniques. On the other hand, a group of similar tools to annotate the anonymous protein sequence without relying on the GO terms are also available such as QuasiMotiFinder [24], MyHits [25], Pfaat [26], GeneAtlas [27], and SMART [28]. There are also works using Bayesian method, statistical method, and C4.5 that have been carried out by [29]−[31] respectively.

However, little effort has been done to develop a tool with the following features to overcome the existing weaknesses:

1) Not dependent on BLAST, requires low cost and minimum hardware specifications, and with reasonable amount of execution time.
2) Not dependent on FASTA (Fast Alignment) format, the query protein sequence can be a newly sequenced one and not necessarily represented in existing biological databases.
3) Not dependent on RDBMS (Relational Database Management Systems), does not require user to setup the RDBMS software and to import the data or sources into the RDBMS format.
4) Fully based on the GO data without requiring download of GOA data or sequence sets from various sources.
5) Annotation is not only dependent on molecular function terms but also on biological process and cellular component terms.
6) Sequence alignment is not carried out to all protein sequences but only to sequences with higher outguessed similarity.

III. The Structure of the Basic and Extended UTMGO

The heart of the UTMGO is its searching and clustering engines, namely SSMGA [32] and SMAGA [33] respectively as shown in Fig. 3. Brief explanations of processing behind the UTMGO are as follows:

1) Public GO data in MySQL and RDF/XML formats are downloaded from the GO website.
2) The single GO RDF/XML file is split into smaller files by SMAGA. The number of files created and the size and the GO terms delegated to each file will be determined automatically during the execution of the clustering process.
3) Corresponding gene products together with protein sequences associated with the GO terms either based on IEA (Inferred from Electronic Annotation) or non-IEA evidence from the GO MySQL database are added into the fragmented GO RDF/XML files.
4) The UTMGO requires the user to enter a GO term and the number of matched GO terms to be returned N_{t1}.
5) Results which are N_{t1} number of GO terms with higher term similarity score to the query GO term will be generated by SSMGA.

The UTMGO is a Linux based tool. Its engines are developed using C/C++ languages and it uses JSP (Java Server Pages) scripts for submitting queries via a web form. The extended version of the UTMGO, as shown in Fig. 4, is specifically designed for annotation of anonymous protein sequence. This extended version comprises of the following steps:

1) Get an anonymous protein sequence, the number of GO terms to be returned N_{t2}, a term similarity threshold, a number of protein sequences associated with each GO

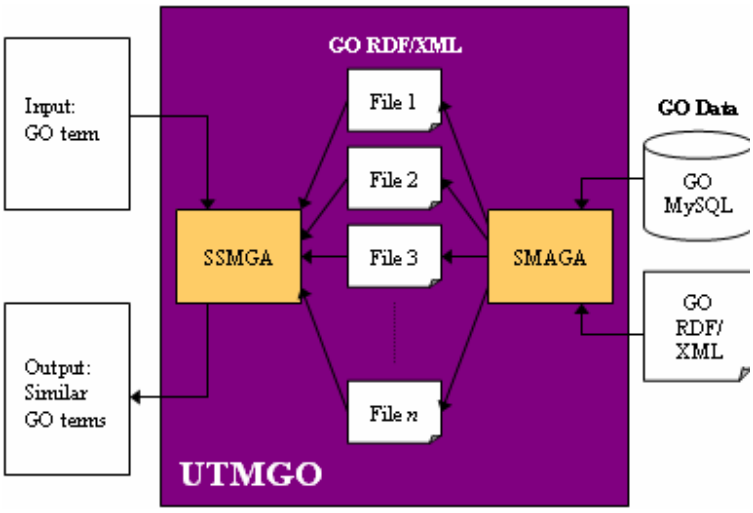

Fig. 3 The structure of the UTMGO for searching a group of semantically related GO terms

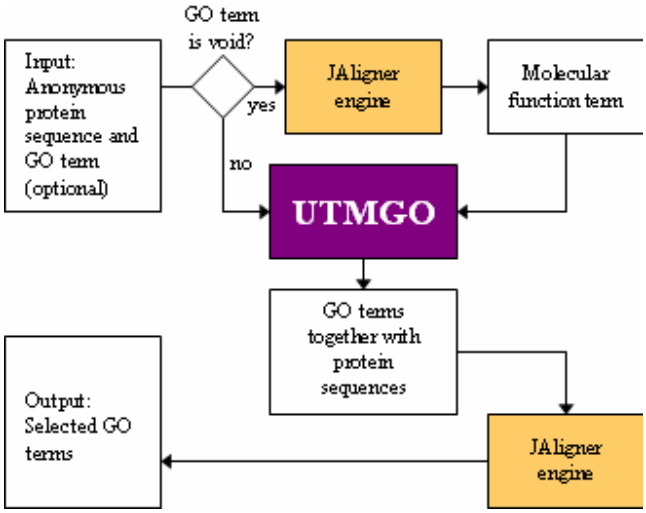

Fig. 4 The structure of the extended UTMGO for annotation of anonymous protein sequence

term to be returned N_s, and optionally a GO term from the user.

2) If the GO term is void, then go to step 3. Otherwise, go to step 6.
3) Get input from the user for appropriate species, matrix type either BLOSUM (Blocks Substitution Matrix) or PAM (Point Accepted Mutations), and open and extend gap penalties to limit the search.
4) Perform sequence similarity search using the JAligner engine for the given anonymous protein sequence from step 1. The search is executed for protein sequences from the fragmented GO RDF/XML files that relate to the molecular function terms. The output will be a protein sequence with the highest sequence alignment score.
5) Select molecular function term with maximum cardinal of association with the protein sequence obtained from step 4 for the next step. If there is more than one term, the user will make the selection.
6) Submit the GO term either from step 1 or step 5 to the UTMGO and then perform term similarity search.
7) Return N_{t2} number of GO terms with term similarity score higher than the term similarity threshold from step 1,

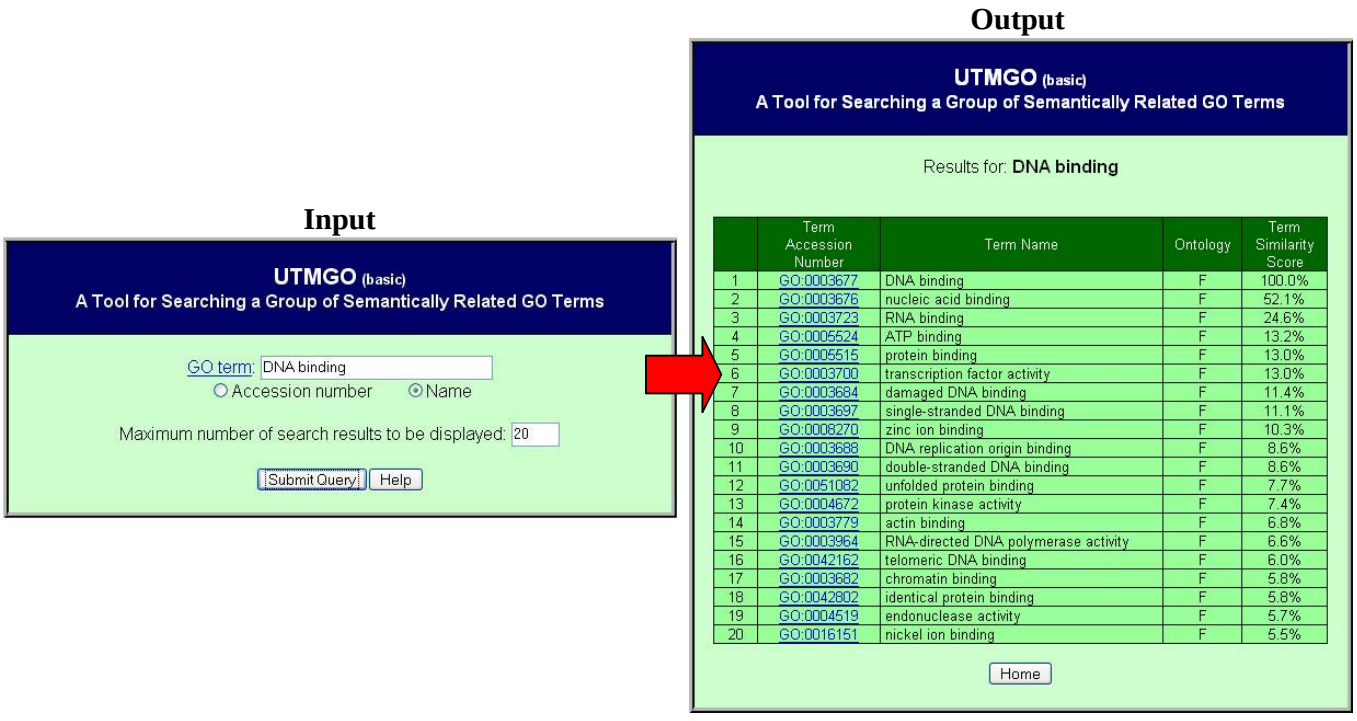

Fig. 5 Screenshot of the UTMGO

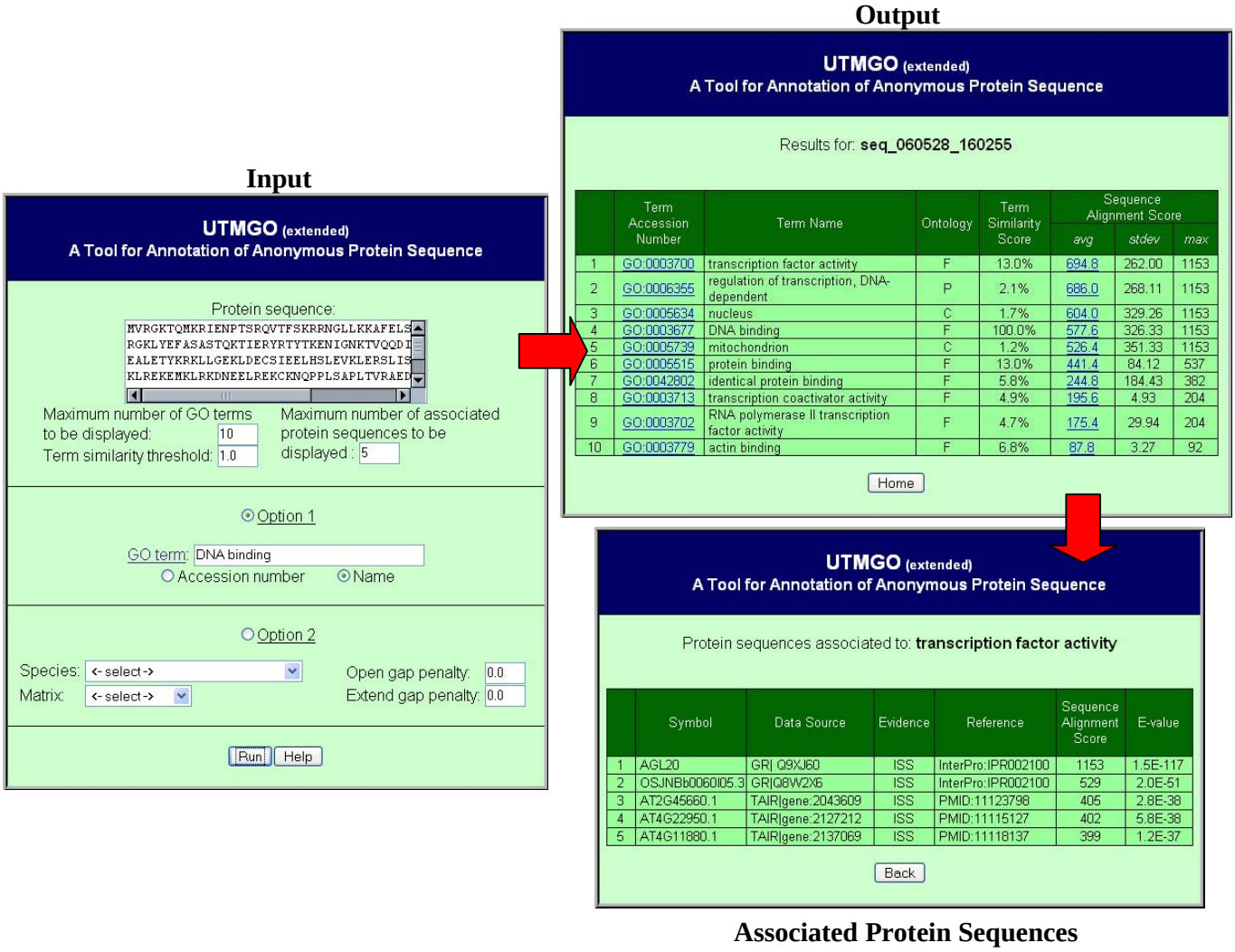

Fig. 6 Screenshot of the extended UTMGO with Option 1

thus, together with protein sequences associated with them.

8) Compute sequence alignment score between the query sequence and all sequences for each GO term gained from previous step using the JAligner engine. Display only N_s number of protein sequences with higher sequence alignment score for each GO term.

The UTMGO and its extended version are available upon request to the corresponding author. The user will be supplied with a set of latest GO RDF/XML files which have been fragmented by the SMAGA, bundled with binary code of the SSMGA and the JAligner engine. Note that, the latest fragmented GO RDF/XML files can be requested depending on the GO monthly releases. Moreover, the user is also advised to have a low-cost PC cluster using MPICH libraries [34] to accelerate the execution speed. Otherwise, the UTMGO and its extended version can be run on a single PC with the following minimum requirements: Fedora Core 2 or Red Hat Linux 8.0 with Pentium 4 processor 2.8 GHz and 512 MB RAM, Linux web browser such as Firefox, Opera, or Konqueror, and Apache Tomcat and JRE (Java Runtime Environment) need to be installed in order to run the JSP page and the JAligner engine respectively.

IV. Testing and Comparison

GO data released in May 2006 is used to test the UTMGO and its extended version. They are executed using a low-cost PC cluster that consists of 25 Pentium IV 2.8 GHz processors with 512 MB RAM and 100 Mbps NIC. The operating system used is Fedora Core 5. Screenshot of the UTMGO is shown in Fig. 5. The example shows the query results for "DNA binding" (GO:0003677), which is a molecular function term interacting selectively with DNA. The results depict 20 GO terms that have higher term similarity score to the query GO term. The first column presents the GO term accession number, followed by short description of the GO term, its aspect (either cellular component (C), molecular function (F), or biological process (P)), and the term similarity score. Clicking on the respective GO term accession number shows all information related to the term. The information is displayed via AmiGO (www.godatabase.org).

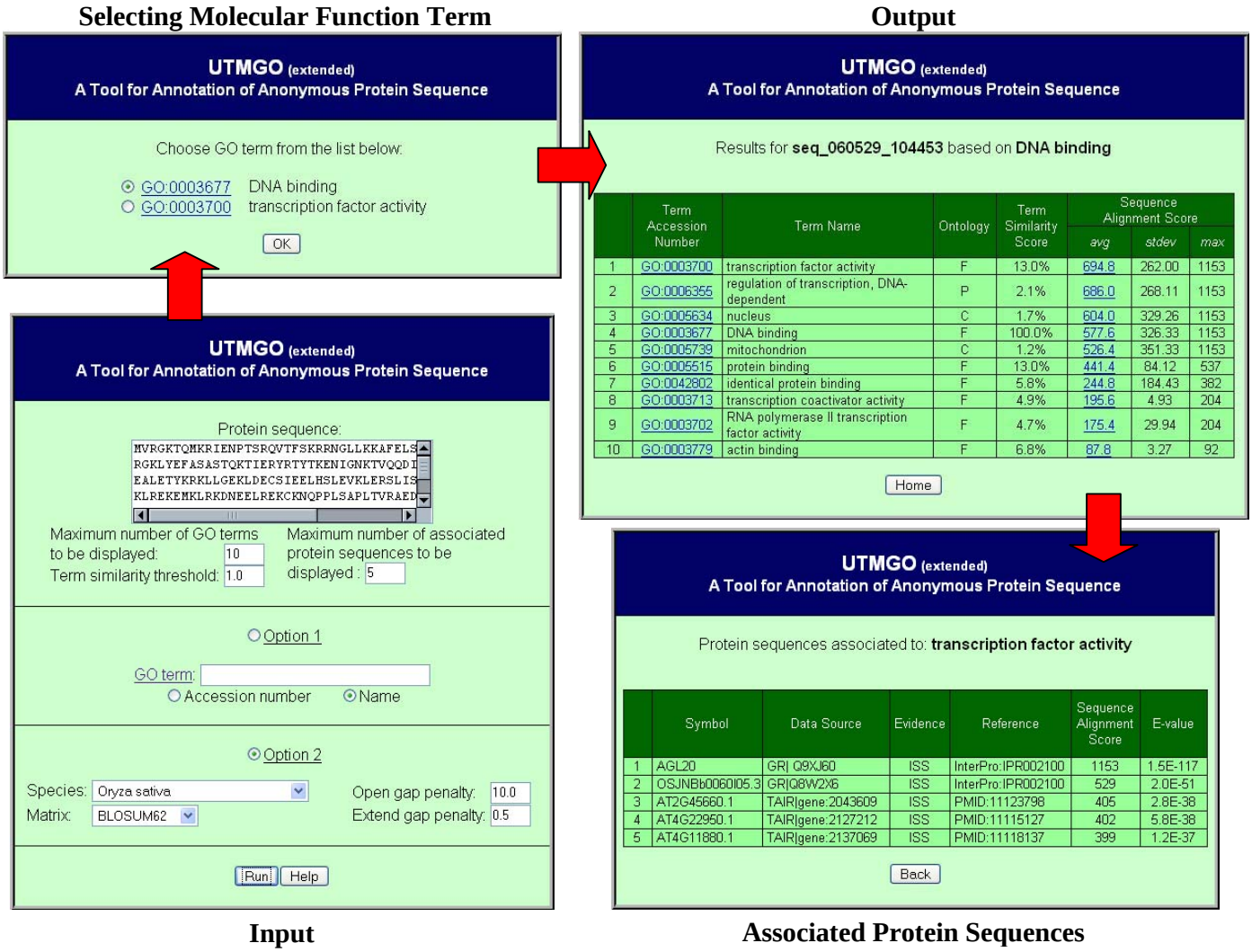

Fig. 7 Screenshot of the extended UTMGO with Option 2

The input protein sequence to demonstrate the extended UTMGO is as follows:

MVRGKTQMKRIENPTSRQVTFSKRRNGLLKKAFE
LSVLCDAEVALIVFSPRGKLYEFASASTQKTIERYR
TYTKENIGNKTVQQDIEQVKADADGLAKKLEALE
TYKRKLLGEKLDECSIEELHSLEVKLERSLISIRGR
KTKLLEEQVAKLREKEMKLRKDNEELREKCKNQP
PLSAPLTVRAEDENPDRNINTTNDNMDVETELFIG
LPGRSRSSGGAAEDSQAMPHS

The protein sequence belongs to *AGL20* (*MADS box-like protein*), an *Oryza sativa* species that is obtained from Gramene database [35]. The operation of extended UTMGO is divided into two cases: with (Option 1) or without (Option 2) a GO term entered by the user. The first case is depicted in Fig. 6. Same as with the UTMGO, the output consists of the GO term accession number and its short description, aspect, and term similarity score. Value-added information is arithmetic mean (*avg*), standard deviation (*stdev*), and the largest value (*max*) of the sequence alignment score of N_s number of protein sequences that are attached to the GO terms. The user can explore details of the GO term by clicking its accession number. In addition, the N_s number of protein sequences associated with the GO term can be viewed by clicking the arithmetic mean. The protein sequences are displayed with its symbol, data source, evidence, and reference; thence, with sequence alignment score and E-value between them and the query sequence. The second case is depicted in Fig. 7. The GO term proposed by the JAligner engine or selected by the user that is given to the UTMGO is shown at the top of the results table.

To show the capability of the UTMGO, its output is compared with other GO browsers such as AmiGO, GenNav (mor.nlm.nih.gov/perl/gennav.pl), TAIR Keyword Browser (www.arabidopsis.org/), and QuickGO (www.ebi.ac.uk/ego/). Table I tabulates the GO terms returned by those browsers for search of "DNA binding" (GO:0003677). The term similarity score is computed using semantic similarity measure proposed by Razib *et al.* [32]. The output is arranged based on first 10 search results generated by each browser. The results show that the GO terms with higher term similarity score such as "nucleic acid binding" (GO:0003676), "RNA binding" (GO:0003723), and "ATP binding" (GO:0005524) are

TABLE I
COMPARISON OF SEARCH RESULTS BETWEEN UTMGO WITH OTHER GO BROWSERS

UTMGO		AmiGO		GenNav		TAIR Keyword Browser		QuickGO	
Term Accession Number	Term Similarity Score	Term Accession Number	Term Similarity Score	Term Accession Number	Term Similarity Score	Term Accession Number	Term Similarity Score	Term Accession Number	Term Similarity Score
GO:0003677	100.0%	GO:0003680	5.4%	GO:0003680	5.4%	GO:0003680	5.4%	GO:0003677	100.0%
GO:0003676	52.1%	GO:0050692	1.9%	GO:0003681	4.2%	GO:0003677	100.0%	GO:0006260	3.4%
GO:0003723	24.6%	GO:0003677	100.0%	GO:0019237	4.6%	GO:0003681	4.2%	GO:0051880	2.5%
GO:0005524	13.2%	GO:0051880	2.5%	GO:0031490	3.6%	GO:0003684	11.4%	GO:0003899	5.3%
GO:0005515	13.0%	GO:0003681	4.2%	GO:0003684	11.4%	GO:0003690	8.6%	GO:0003887	5.0%
GO:0003700	13.0%	GO:0019237	4.6%	GO:0050692	1.9%	GO:0003691	3.7%	GO:0050692	1.9%
GO:0003684	11.4%	GO:0031490	3.6%	GO:0003677	100.0%	GO:0003692	4.6%	GO:0003908	3.3%
GO:0003697	11.1%	GO:0003684	11.4%	GO:0003690	8.6%	GO:0003695	3.1%	GO:0003964	6.6%
GO:0008270	10.3%	GO:0003690	8.6%	GO:0003691	3.7%	GO:0000182	4.9%	GO:0008534	3.0%
GO:0003688	8.6%	GO:0003691	3.7%	GO:0051880	2.5%	GO:0003696	5.2%	GO:0003886	3.0%

TABLE II
COMPARISON OF PERFORMANCE BETWEEN UTMGO WITH OTHER GO BROWSERS

	UTMGO	AmiGO	GenNav	TAIR Keyword Browser	QuickGO
Precision	83.80%	72.84%	71.62%	72.13%	63.54%
Recall	70.35%	67.78%	66.18%	66.96%	74.39%
Running time	0.13s	0.09s	0.06s	0.08s	0.22s

ordered at the top of the list by the UTMGO. The overall performance is shown in Table II for the same input set of 398 molecular function terms, 551 biological process terms, and 93 cellular component terms. Hence, the UTMGO showed a better precision (83.80%) and the QuickGO provided a better recall (74.39%), whereas the GenNav gives the best running time (0.06 seconds).

The comparison between the extended UTMGO and other protein sequence annotation tools are shown in Table III and Table IV. The GOblet is not in the comparison list due to its service being temporarily unavailable. The results are obtained with *AGL20* as the input protein sequence. The arithmetic mean and the largest value of the sequence alignment score for the protein sequences associated with the GO terms are used to analyze those tools. This is due to the fact that quality of the GO terms relies on the similarity between the query protein sequence and the protein sequences associated with them. Thus, higher is better. As depicted in Table 3, almost all GO terms with higher average of sequence alignment score are listed as top 10 GO terms that are returned by the extended UTMGO. However, GO terms such as "flower development" (GO:0009908) and "cytoplasm" (GO:0005737) are out of the extended UTMGO radar since their term similarity score is 0.9% and 0.6% respectively. These values are lower than the term similarity threshold (1.0%) set for this testing session. Furthermore, as shown in Table 4, all GO terms that have been linked to protein sequence with the highest sequence alignment score (1,153) are returned by the extended UTMGO. Note that, even though GO terms such as "positive regulation of transcription from RNA polymerase II promoter" (GO: 0045944), "DNA bending activity" (GO: 0008301), and "peptidase activity" (GO: 0008233) have maximum sequence alignment score higher than "actin binding" (GO: 0003779) but they are not listed in the top 10 GO terms that are returned by the extended UTMGO. The reason is their average sequence alignment score is lower than the value for "actin binding" (GO: 0003779). To evaluate the performance of the extended UTMGO as compared to other protein sequence annotation tools, a set of protein sequences were chosen from Gramene (www.gramene.org), a database of *Oryza sativa*; Esembl (www.ensembl.org), a database of *Homo sapiens*; SGD (www.yeastgenome.org), a database of *Saccharomyces cerevisiae*; and TAIR (www.arabidopsis.org), a database of *Arabidopsis thaliana*. These protein sequences were selected randomly with 50 protein sequences from each database. The results, as depicted in Table V, show that the extended UTMGO provides a better precision (91.04%) and the GOtcha offered a better recall (89.74%). The best running time is 127 seconds that is taken by the GoFigure.

V. CONCLUSION

The UTMGO has been presented as an alternative way to search the GO terms. The search is done by finding a group of semantically similar GO terms that relate to the user request. This semantic similarity search is not based on word matching but according to degree of relationships between the GO terms. A gene product that associates with one or more GO terms is used to calculate the amount of information the GO terms share in common. Hence, it gives the degree of relationships. The search results have indicated that the UTMGO is capable to find functionally related GO terms as compared to other existing GO browsers which are based on keyword queries. The applicability of the UTMGO has been shown by its extended version. The extended UTMGO has the ability to annotate anonymous protein sequence. The protein sequences associated with the GO terms that are returned by the extended UTMGO have higher sequence alignment score

TABLE III
COMPARISON OF ARITHMETIC MEAN OF SEQUENCE ALIGNMENT SCORE BETWEEN EXTENDED UTMGO WITH OTHER PROTEIN SEQUENCE ANNOTATION TOOLS

Extended UTMGO		GOPET		GOtcha		GoFigure		JAFA	
Term Accession Number	*avg* of Sequence Alignment Score	Term Accession Number	*avg* of Sequence Alignment Score	Term Accession Number	*avg* of Sequence Alignment Score	Term Accession Number	*avg* of Sequence Alignment Score	Term Accession Number	*avg* of Sequence Alignment Score
GO:0003700	694.8	GO:0006355	686.0	GO:0003677	577.6	GO:0003700	694.8	GO:0045944	86.4
GO:0006355	686.0	GO:0003677	577.6	GO:0030528	0.0	GO:0003677	577.6	GO:0006657	0.0
GO:0005634	604.0	GO:0003700	694.8	GO:0003700	694.8	GO:0007275	0.0	GO:0004402	0.0
GO:0003677	577.6	GO:0006139	0.0	GO:0006139	0.0	GO:0009908	113.0	GO:0008362	0.0
GO:0005739	526.4	GO:0006350	0.0	GO:0006350	0.0	GO:0006350	0.0	GO:0007144	0.0
GO:0005515	441.4	GO:0045944	86.4	GO:0006355	686.0	GO:0006355	686.0	GO:0007129	36.2
GO:0042802	244.8	GO:0006357	85.2	GO:0005622	38.6	GO:0005634	604.0	GO:0007020	19.0
GO:0003713	195.6	GO:0003936	0.0	GO:0008233	29.6	-	-	GO:0007004	0.0
GO:0003702	175.4	GO:0008301	57.4	GO:0005215	17.0	-	-	GO:0007015	20.2
GO:0003779	87.8	-	-	GO:0005737	93.8	-	-	GO:0006430	16.6

TABLE IV
COMPARISON OF THE LARGEST VALUE OF SEQUENCE ALIGNMENT SCORE BETWEEN EXTENDED UTMGO WITH OTHER PROTEIN SEQUENCE ANNOTATION TOOLS

Extended UTMGO		GOPET		GOtcha		GoFigure		JAFA	
Term Accession Number	*max* of Sequence Alignment Score	Term Accession Number	*max* of Sequence Alignment Score	Term Accession Number	*max* of Sequence Alignment Score	Term Accession Number	*max* of Sequence Alignment Score	Term Accession Number	*max* of Sequence Alignment Score
GO:0003700	1,153	GO:0006355	1,153	GO:0003677	1,153	GO:0003700	1,153	GO:0045944	153
GO:0006355	1,153	GO:0003677	1,153	GO:0030528	0	GO:0003677	1,153	GO:0006657	0
GO:0005634	1,153	GO:0003700	1,153	GO:0003700	1,153	GO:0007275	0	GO:0004402	0
GO:0003677	1,153	GO:0006139	0	GO:0006139	0	GO:0009908	565	GO:0008362	0
GO:0005739	1,153	GO:0006350	0	GO:0006350	0	GO:0006350	0	GO:0007144	0
GO:0005515	537	GO:0045944	153	GO:0006355	1,153	GO:0006355	1,153	GO:0007129	92
GO:0042802	382	GO:0006357	151	GO:0005622	101	GO:0005634	1,153	GO:0007020	95
GO:0003713	204	GO:0003936	0	GO:0008233	148	-	-	GO:0007004	0
GO:0003702	204	GO:0008301	153	GO:0005215	85	-	-	GO:0007015	101
GO:0003779	92	-	-	GO:0005737	134	-	-	GO:0006430	83

TABLE V
COMPARISON OF PERFORMANCE BETWEEN EXTENDED UTMGO WITH OTHER PROTEIN SEQUENCE ANNOTATION TOOLS

	Extended UTMGO	GOPET	GOtcha	GoFigure	JAFA
Precision	91.04%	82.63%	85.09%	84.77%	68.31%
Recall	88.83%	80.39%	89.74%	80.62%	79.15%
Running time	293s	270s	518s	127s	302s

to the query protein sequence. The SSMGA and the SMAGA components in the UTMGO play an important role in accelerating the search. Moreover, the extended UTMGO is not dependent on BLAST and RDBMS, can be used for not-yet-annotated protein sequences, fully based on the GO data, and relate to all categories in the GO.

Future development direction for the UTMGO is applying it to predict protein function and protein-protein interactions. In the case of the extended UTMGO, enhancement includes the ability to support more than one protein sequence per query and to accept DNA sequence as an input.

REFERENCES

[1] 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. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25-29, May 2000.

[2] G. Gao, Y. Zhong, A. Guo, Q. Zhu, W. Tang, W. Zheng, X. Gu, L. Wei, and J. Luo, "DRTF: a database of rice transcription factors," *Bioinformatics*, vol. 22, no. 10, pp. 1286-1287, May 2006.

[3] C. Winter, A. Henschel, W.K. Kim, and M. Schroeder, "SCOPPI: a structural classification of protein-protein interfaces," *Nucleic Acids Res.*, vol. 34, no. 1, pp. D310-D314, Jan. 2006.

[4] A.K. Leung, L. Trinkle-Mulcahy, Y.W. Lam, J.S. Andersen, M. Mann, and A.I. Lamond, "NOPdb: Nucleolar Proteome Database," *Nucleic Acids Res.*, vol. 34, no. 1, pp. D218-D220, Jan. 2006.

[5] N. Wiwatwattana and A. Kumar, "Organelle DB: a cross-species database of protein localization and function," *Nucleic Acids Res.*, vol. 33, no. 1, pp. D598-D604, Jan. 2005.

[6] X. Wu, L. Zhu, J. Guo, D.Y. Zhang, and K. Lin, "Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations," *Nucleic Acids Res.*, vol. 34, no. 7, pp. 2137-2150, Apr. 2006.

[7] R.M. Salasznyk, A.M. Westcott, R.F. Klees, D.F. Ward, Z. Xiang, S. Vandenberg, K. Bennett, and G.E. Plopper, "Comparing the protein expression profiles of human mesenchymal stem cells and human osteoblasts using gene ontologies," *Stem Cells Dev.*, vol. 14, no. 4, pp. 354-366, Aug. 2005.

[8] S. Basu, E. Bremer, C. Zhou, and D.F. Bogenhagen, "MiGenes: a searchable interspecies database of mitochondrial proteins curated using Gene Ontology annotation," *Bioinformatics*, vol. 22, no. 4, pp. 485-492,

Dec. 2005.
[9] P. Lu, D. Szafron, R. Greiner, D.S. Wishart, A. Fyshe, B. Pearcy, B. Poulin, R. Eisner, D. Ngo, and N. Lamb, "PA-GOSUB: a searchable database of model organism protein sequences with their predicted Gene Ontology molecular function and subcellular localization," *Nucleic Acids Res.*, vol. 33, no. 1, pp. D147-D153, Jan. 2005.
[10] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler, "The Gene Ontology Annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Res.*, vol. 32, no. 1, pp. D262-D266, Jan. 2004.
[11] E. Camon, M. Magrane, D. Barrell, D. Binns, W. Fleischmann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, A. Cox, and R. Apweiler, "The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro," *Genome Res.*, vol. 13, no. 4, pp. 662-672, Apr. 2003.
[12] A. Moustafa (2005, Apr.). JAligner: open source Java implementation of Smith-Waterman. Available: http://jaligner.sourceforge.net.
[13] K.-S. Goh, E.Y. Chang, and B. Li, "Using one-class and two-class SVMs for multiclass image annotation," *IEEE Trans. Knowledge & Data Engineering*, vol. 17, no. 10, pp. 1333-1346, Oct. 2005.
[14] M. Fernandes, M. Alho, J.A. Martins, J.S. Pinto, and P. Almeida, "Web annotation system based on web services," *Int. J. Web Services Practices*, vol. 1, no. 1-2, pp. 101-108, Aug. 2005.
[15] A. Vinayagam, C. del Val, F. Schubert, R. Eils, K. Glatting, S. Suhai, and R. König, "GOPET: a tool for automated predictions of Gene Ontology terms," *BMC Bioinformatics*, vol. 7, no. 1, rec. 161, Mar. 2006.
[16] D.M. 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*, vol. 5, no. 1, rec. 178, Nov. 2004.
[17] D. Groth, H. Lehrach, and S. Hennig, "GOblet: a platform for Gene Ontology annotation of anonymous sequence data," *Nucleic Acids Res.*, vol. 32, no. 1, pp. W313-W317, Jul. 2004.
[18] S. Khan, G. Situ, K. Decker, and C.J. Schmidt, "GoFigure: automated Gene Ontology annotation," *Bioinformatics*, vol. 19, no. 18, pp. 2484-2485, Dec. 2003.
[19] I. Friedberg, T. Harder, and A. Godzik, "JAFA: a protein function annotation meta server," *Nucleic Acids Res.*, to be published.
[20] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez, "InterProScan: protein domains identifier," *Nucleic Acids Res.*, vol. 33, no. 1, pp.W116-W120, Jul. 2005.
[21] F. Enault, K. Suhre, O. Poirot, C. Abergel, and J.M. Claverie, "Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3720-3722, Jul. 2003.
[22] K. Verspoor, J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L.M. Rocha, and T. Simas, "Protein annotation as term categorization in the Gene Ontology using word proximity networks," *BMC Bioinformatics*, vol. 6, suppl. 1, rec. S20, May 2005.
[23] H. Xie, A. Wasserman, Z. Levine, A. Novik, V. Grebinskiy, A. Shoshan, and L. Mintz, "Large-scale protein annotation through Gene Ontology," *Genome Res.*, vol. 12, no. 5, pp. 785-794, May 2002.
[24] R. Gutman, C. Berezin, R. Wollman, Y. Rosenberg, and N. Ben-Tal, "QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns," *Nucleic Acids Res.*, vol. 33, no. 1, pp. W255-261, Jul. 2005.
[25] M. Pagni, V. Ioannidis, L. Cerutti, M. Zahn-Zabal, C.V. Jongeneel, and L. Falquet, "MyHits: a new interactive resource for protein annotation and domain identification," *Nucleic Acids Res.*, vol. 32, no. 1, pp. W332-W335, Jul. 2004.
[26] J.M. Johnson, K. Mason, C. Moallemi, H. Xi, S. Somaroo, and E.S. Huang, "Protein family annotation in a multiple alignment viewer," *Bioinformatics*, vol. 19, no. 4, pp. 544-545, Mar. 2003.
[27] D.H. Kitson, A. Badretdinov, Z.Y. Zhu, M. Velikanov, D.J. Edwards, K. Olszewski, S. Szalma, and L. Yan, "Functional annotation of proteomic sequences based on consensus of sequence and structural analysis," *Brief Bioinform.*, vol. 3, no. 1, pp. 32-44, Mar. 2002.
[28] I. Letunic, L. Goodstadt, N.J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R.R. Copley, C.P. Ponting, and P. Bork, "Recent improvements to the SMART domain-based sequence annotation resource," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 242-244, Jan. 2002.
[29] E.D. Levy, C.A. Ouzounis, W.R. Gilks, and B. Audit, "Probabilistic annotation of protein sequences based on functional classifications," *BMC Bioinformatics*, vol. 6, no. 1, rec. 302, Dec. 2005.
[30] W.G. Krebs and P.E. Bourne, "Statistically rigorous automated protein annotation," *Bioinformatics*, vol. 20, no. 7, pp. 1066-1073, May 2004.
[31] 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*, vol. 17, no. 10, pp. 920-926, Oct. 2001.
[32] R.M. Othman, S. Deris, R.M. Illias, H.T. Alashwal, R. Hassan, and F. Mohamed, "Incorporating semantic similarity measure in genetic algorithm: an approach for searching the Gene Ontology terms," *Int. J. Computational Intelligence*, vol. 3, no. 3, pp. 257-266, May 2006.
[33] R.M. Othman, S. Deris, R.M. Illias, Z. Zakaria, and S.M. Mohamad, "Automatic clustering of Gene Ontology by genetic algorithm," *Int. J. Information Technology*, vol. 3, no. 1, pp. 37-46, Apr. 2006.
[34] W. Gropp, E. Lusk, N. Doss, and A. Skjellum, "A high-performance, portable implementation of the MPI message-passing interface standard," *Parallel Computing*, vol. 22, no. 6, pp. 789-828, Sep. 1996.
[35] P. Jaiswal, J. Ni, I. Yap, D. Ware, W. Spooner, K. Youens-Clark, L. Ren, C. Liang, B. Hurwitz, W. Zhao, K. Ratnapu, B. Faga, P. Canaran, M. Fogleman, C. Hebberd, S. Avraham, S. Schmidt, T. Casstevens, E.S. Buckler, L. Stein, and S. Mccouch, "Gramene: a genomics and genetics resource for rice," *Rice Genetics Newsletter*, vol. 22, no. 1, pp. 9-16, Jan. 2006.

Razib M. Othman is a doctoral candidate at the Faculty of Computer Science and Information Systems, the Universiti Teknologi Malaysia. He received the BSc and MSc degrees in Computer Science both from the Universiti Teknologi Malaysia, in 1999 and 2003 respectively. Currently, he is working for his PhD in Computational Biology. He also has interests in artificial intelligence, software agent, parallel computing, and web semantics. In March 2005, he was awarded the Young Researcher award by the Malaysian Association of Research Scientists (MARS). Two of his inventions, software products named *2D Engineering Drawing Extractor* and *2D Design Structure Recognizer*, have won 5 awards at the 21st Invention and New Product Exposition held in Pittsburgh, USA including the Best Invention of the Pacific Rim, and a gold medal award at the 34th International Exhibition of Inventions of New Techniques and Products held in Geneva, Switzerland.

Safaai Deris is a Professor of Artificial Intelligence and Software Engineering at the Faculty of Computer Science and Information Systems, Deputy Dean at the School of Graduate Studies, and Director of Laboratory of Artificial Intelligence and Bioinformatics at the Universiti Teknologi Malaysia. He received the MEng degree in Industrial Engineering, and the DEng degree in Computer and System Sciences, both from the Osaka Prefecture University, Japan, in 1989 and 1997 respectively. His recent academic interests include the application and development of intelligent techniques in planning, scheduling, and bioinformatics.

Rosli M. Illias is an Associate Professor at the Faculty of Chemical and Natural Resources Engineering at the Universiti Teknologi Malaysia. He received the PhD degree in Molecular Biology from the Edinburgh University, UK in 1997, and the BSc degree in Microbiology from the Universiti Kebangsaan Malaysia in 1992. His research interests are in the areas of microbial technology, molecular enzymology, and molecular genetics.