GoPipe: Streamlined Gene Ontology Annotation for Batch Anonymous Sequences With Statistics*

CHEN Zuo-Zhou[1,2], XUE Cheng-Hai[3], ZHU Sheng[1,2], ZHOU Feng-Feng[4], XUEFENG BRUCE LING[5], LIU Guo-Ping[6], CHEN Liang-Biao[2]**

([1]*College of Life Science, Zhejiang University, Hangzhou* 310029, China;
[2]*Institute of Genetics and Developmental Biology, The Chinese Academy of Sciences, Beijing* 100080, China;
[3]*Institute of Automation, The Chinese Academy of Sciences, Beijing* 100080, China;
[4]*Department of Computer Science, National High Performance Computing Center, University of Science and Technology of China, Hefei* 230027, China;
[5]*Tularik Inc., 1120 Veterans Blvd, South San Francisco*, CA 94080, USA;
[6]*School of Electronics, University of Glamorgan, Pontypridd CF37 1DL*, UK)

Abstract Accelerated availability of new sequences, especially ESTs, calls for computational methods to link sequences with Gene Ontology (GO) terms in a batch mode. There is currently no program for such purpose except Goblet, an online tool which uses BLAST to interpret query sequence with proper GO terms, but has a restriction of upload sequence files less than 100 kilobytes in size. GoPipe is a standalone package that integrates BLAST and InterProScan results to obtain Gene Ontology annotation with built-in statistical options. GoPipe takes any number of BLAST and/or InterProScan output files simultaneously and launches jobs sequentially to perform parsing, data integration, redundancy removal, GO distributions calculation and graphic display. A very high annotation specificity of 99.1% was achieved for a test dataset when the program was run in the "intersection" mode, which intersects the BLAST and InterProScan results, outperforming the specificity (81.1%) obtained from the InterProScan only. Statistical tools are also provided to compare GO distributions between different inputs, so that GO distributions of different sets of batch sequences can be compared, and differentially represented GO terms can be easily displayed. High specificity, speed and flexibility make GoPipe an ideal tool for streamlined GO annotation for batch sequences. The package is freely available at http://gopipe.fishgenome.org/ or by contacting the authors.

Key words Gene Ontology, functional genomics, EST, BLAST, InterProScan, GOA

Large-scale sequencing projects call for computational methods to link sequences with functions and other categorical information. This procedure, named electronic annotation, can be achieved through computational analysis with a set of vocabulary that describes the characteristics of genes and/or gene products.

Gene Ontology (GO), as such a set of structured, controlled and dynamic vocabulary, is becoming the *de facto* standard for describing molecular functions, biological processes, and cellular components of genes and gene products [1]. The well defined GO terms not only provide a set of unified vocabulary but also give us hierarchies of the vocabularies, which make it possible to compare the distributions between two sets of sequence annotations at various ontological levels.

A variety of approaches have been designed to associate sequences with GO terms, such as domain mapping, vocabulary mapping, textual mining, expression profiling, and the integration of these methods [2~4]. But for anonymous sequences, such as ESTs of non-model organisms, of which characteristic information is limited, the homology searching tool, BLAST [5] is usually used to search GO annotated databases to link these new sequences with GO terms. Based on this, several web services have been provided (the GOst software tool (http://www.godatabase.org), Goblet [6] and OntoBlast [7]). Alternatively new sequences can be scanned for various signatures (domains, families, repeats, etc.) that have already been mapped to GO terms. The most widely used tool is InterProScan, which combines different protein signature recognition methods native to the InterPro member databases into one resource to look up for corresponding InterPro and GO annotation [8].

To our knowledge, however, there is currently no program for batch Gene Ontology annotation except a new version of Goblet, which only uses BLAST to associate query sequence with GO terms, but has a

*This work was supported by a grant from The National Natural Science Foundation of China (30330080).
**Corresponding author.
Tel: 86-10-62554807, Fax: 86-10-62554807
E-mail: lbchen@genetics.ac.cn
Received: October 9, 2004 Accepted: January 7, 2005

restriction of upload sequence files less than 100 kilobytes in size. Considering that 1000 EST sequences probably in the size of 500 kilobytes, EST projects that produce tens or hundreds of thousands of ESTs usually needs to write programs by their own. In addition, the relationship between sequence similarity and function is not adequately addressed, and the accuracy of GO prediction is low when only BLAST is used.

Batch Gene Ontology annotation is different from online annotation in a manner of one sequence after another. It is much more computer intensive, has no human-computer interactions, requires pre-defined criteria. Further more, batch EST sequences are often produced from biological samples in question, and statistical analyses are usually desirable for in-depth revealing of the biological meanings of the sequences.

We developed GoPipe, a standalone Perl-based package integrating BLAST and/or InterProScan results to produce a non-redundant list of associations between sequences and GO-Ids. It also provides tools to calculate GO distributions for the query sequences, with respects to all the GO terms or GO slim terms. In addition, differentially distributed GO terms between two sets of sequences are statistically and easily displayed with built-in programs in this package.

1 Materials and methods

1.1 Programs

The data flow of GoPipe is shown in Supplementary material 1. GoPipe takes the resultant files of batch query sequences either produced by running BLAST against the GO annotated, SWISS-PROT and Trembl database [9] (ftp://ftp.ebi.ac.uk/pub/databases/sp_tr_nrdb/fasta) or from the InterProScan signature searching tool (ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan/, http://www.ebi.ac.uk/InterProScan/). The InterProScan is set to return files in XML format and the Blast in the standard text format (pairwise format). Both formats are suitable for GoPipe input, and there is no limit in the number of files of each format.

GoPipe is composed of five sub-programs to conduct batch GO annotation and subsequent statistical analyses. The first sub-program parser.pl extracts necessary information according to the formats of input files. When GoPipe is fed with InterProScan files, the links between query sequences and corresponding GO-Ids are directly extracted; alternatively, when it is fed with BLAST resultant files, parser.pl takes two more parameters from the command line: the E-value cut-off, which specifies the E-value below which the hits could be accepted to obtain GO associations, and the number "*n*", to set the upper limit number of BLAST hits, of which at least one record exist in GO association data files of GOA [10]. The second sub-program clean.pl is used to remove redundancies caused by duplicated sequence names and Go-Ids, or from sequences associated with two or more GO terms on the same path due to the "true path rule" (http://www.geneontology.org/GO.usage.html). The third program distribution.pl calculates the number and the frequency of the sequences annotated to each GO term. The same content for the GO Slim terms are also calculated by the forth sub-program slim.pl. The fifth sub-program then automatically draws graphs to display the statistical results for the GO Slim terms (exemplified in Supplementary material 2).

Two additional programs query.pl and comparecr.pl are also included in the package to facilitate the searching and comparing efforts. The query.pl is used to search the newly annotated data for sequence names that are associated to any GO-Id of interest. The comparecr.pl is designed to reveal differentially distributed GO terms between two sets of query sequences using Chi-square test and Fisher's Exact test. Since we assess all categories of GO, the problem of multiple testing should be concerned, therefore, in addition to the raw P-values, the corrected P-values for multiple testing are also provided using the linear step-up procedure of Benjamini and Hochberg (1995) of false discovery rate (FDR) control[11].

GoPipe produces four text and six graph files. The first is a tab-delimited file consisting non-redundant GO annotations for query sequences. The other three text files contain the number and the proportion of sequences for each GO term, generic GO Slim term, and GOA GO Slim term. The six graph files are bar graphs for statistics of "molecular function", "biological process" and "cellular component" of the GOA GO Slim terms and generic GO Slim terms, respectively. The output files of query.pl and compare.pl are also tab-delimited files containing corresponding information.

In this package, we use several data files extracted from several resources. "SPTR_GO" contains accession numbers of UniProt protein database and their associated GO-Ids, which is derived from the GO association files of Uniprot database of GOA project. "Graph" contains path information, and "Term" contains all the GO terms and two lists of GO Slim terms. Both are derived from the Gene Ontology database (http://www.godatabase.org/dev/database/).

By setting a command-line parameter, users can switch the main program gopipe.pl from the normal mode (union mode) to the intersection mode, in which the sub-program parser.pl is substituted by intersection.pl. Intersection.pl parses the BLAST and the InterProScan results separately to obtain two lists of GO annotations, and then redundancy is removed in a similar way as in the clean.pl. The program then proceeds to compare the two sets of annotations to extract the common GO terms for the same query sequence. For those GO terms that are different for the same sequence in the two sets, if those terms are on the same path, higher-level terms are assigned to the sequence since they are regarded as common terms in the two data sets due to the "true path rule".

Efforts were made to improve the speed of GoPipe and to make it flexible and easy to use. We adopted the binary search technique to speed up the program by thousand folds. GoPipe needs neither configuration nor database access and has no limit on the number of input files. The outputs are simple tab-delimitated flat files. A graphical user interface is also provided for researchers (see Supplentary material 3) to use the program conveniently. GoPipe can be run on most UNIX/Linux systems with Perl 5.0 or higher version installed. We tested it on Linux RadHat with Perl 5.8.0.

1.2 Evaluation methodology

The sensitivity and specificity of GoPipe were benchmarked by comparing the predicted results to those curated by human. Due to the complex structure of GO, extra efforts are needed to determine a "true" or a "false" prediction. When comparing the predicted set of GO associations with the human curated associations of the same sequence, three possibilities could occur as shown in Figure 1. Here, GoPipe predicted GO associations of a sequence are in the "set A", and the human curated ones are listed in the "set B". The three possible situations are: "same", "same path" or "no relation". The number of GO terms associated with the above three situations in both sets can be counted through using the 7 scalars named "Common", "Higher_A", "Lower_A", "Special_A", "Higher_B", "Lower_B" and "Special_B". The specificity of GoPipe is therefore defined as: Specificity = (Common + Higher_A + Lower_A)/ (Common + Higher_A+ Lower_A + Special_A), and it represents the percentage of "true" GO associations out of all predictions. Here the numbers in the "Common", "Higher_A", and "Lower_A" categories are added up to present the "true" GO associations. The sensitivity of GoPipe is defined as: Sensitivity = (Common + Higher_A + Lower_A)/ (Common + Higher_A+ Lower_A +Special_B), for benchmarking the proportion of true GO assignments in all true associations of the dataset.

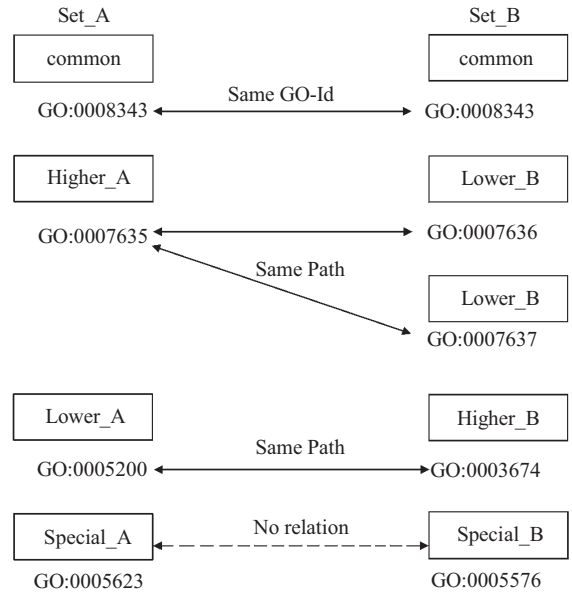

Fig.1 Possible relations between GoPipe predicted GO annotations (Set A) and the human curated annotations (Set B)

GoPipe predicted results could be the "same", "same path" or "no relation" with the standard annotations for the same sequence. 7 scalars ("Common", "Higher_A", "Lower_A", "Special_A", "Higher_B", "Lower_B" and "Special_B") are used for tally the three situations. The GO-Ids under each scalar is an example to denote the possible relationships between set A and Set B.

2 Evaluation of GoPipe

In order to evaluate the performance of GoPipe, a set of 684 human proteins with the GO annotation evidence code of "TAS" (traceable author statement) from the "Human GOA file" (version 16.0) of GOA was complied to use as test sequences. With these sequences, BLAST was first run against a local SWISS-PROT and Trembl non-redundant database (SWISS-PROT Release 42.8, TrEMBL Release 25.8), and then InterProScan were performed online with XML outputs. Using these resultant files, four combinations of inputs were fed to GoPipe: BLASTP output files, InterProScan output files, BLASTP and InterProScan files at union mode or at intersection mode, respectively.

Specificity and sensitivity were measured for each of the above inputs by an evaluation system when all associations of human proteins were removed from the GO association database. GO annotation specificity and sensitivity produced by BLAST using different parameters are shown in Supplementary materials 4~7.

Results derived from BLAST, InterProScan, union mode and intersection mode, are presented in Table 1. GO associations derived from InterProScan files showed high specificity but low sensitivity, while the BLAST yielded relatively low specificity and moderate sensitivity. In the "intersection mode", when GO associations were established by intersecting the Blast and the InterProScan results, specificity arose dramatically to 99.1%. We observed that Blast removed many false positive GO associations but few true positive ones (data not shown). The choice of execution mode is left for researchers to decide according to their own emphasis on specificity or sensitivity. GoPipe therefore possesses high specificity, speed and flexibility, which makes it suitable for streamlined annotation of batch sequences.

Table 1 Performances of GoPipe using different input files

	Specificity	Sensitivity
BLAST (1E-5, *n*=5)	31.6%	39.0%
InterProScan	81.1%	11.8%
Union mode	31.7%	39.1%
Intersection mode	99.1%	11.8%

3 Discussion

As mentioned above, batch annotation for anonymous sequences is different from that for single ones, and our work has following novelties: GO associations predicted by BLAST search and InterProScan can be integrated for better specificity, especially at the intersection mode; redundant parental GO associations for the same sequence are removed according to the "true path rule"; the number of sequences for each GO term is calculated with plotting tools provided; two sets of GO predictions can be compared to assess GO terms being over- or under-represented, which is especially useful to study functional differentiation between two sets of ESTs.

Supplementary materials of this article can be downloaded from our web site (http://gopipe.fishgenome.org/supplement1.htm).

Acknowledgements We would like to thank GOA project and GO Consortium for their data files. Some codes used to calculate chi-square p-values are from Michael Kospach and StatLib.

References

1 Harris M A, Clark J, Ireland A, *et al*. The Gene Ontology (GO) database and informatics resource. Nucl Acids Res, 2004, **32**: D258-D261

2 Pouliot Y, Gao J, Su Q J, *et al*. DIAN: a novel algorithm for genome ontological classification. Genome Res, 2001, **11** (10): 1766~1779

3 Xie H, Wasserman A, Levine Z, *et al*. Large-scale protein annotation through gene ontology. Genome Res, 2002, **12** (5): 785~794

4 Lagreid A, Hvidsten T R, Midelfart H, *et al*. Predicting gene ontology biological process from temporal gene expression patterns. Genome Research, 2003, **13** (5): 965~979

5 Altschul S F, Madden T L, Schaffer A A, *et al*. Gapped BLAST and PSIBLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, **25** (17): 3389~3402

6 Hennig S, Groth D, Lehrach H. Automated Gene Ontology annotation for anonymous sequence data. Nucleic Acids Res, 2003, **31** (13): 3712~3715

7 Zehetner G. OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res, 2003, **31** (13): 3799~3803

8 Mulder N J, Apweiler R, Attwood T K, *et al*. The InterPro Database, 2003 brings increased coverage and new features. Nucl Acids Res, 2003, **31** (1): 315~318

9 Apweiler R, Bairoch A, Wu C H, *et al*. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res, 2004, **32**: D115-D119

10 Camon E, Magrane M, Barrell D, *et al*. The Gene Ontology Annotation (GOA) project: implementation of GO in Swiss-Prot, TrEMBL and InterPro. Genome Research, 2003, **13** (4): 662~672

11 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B, 1995, **57**: 289~300

GoPipe: 批量序列的 Gene Ontology 注释和统计分析 *

陈作舟[1,2] 薛成海[3] 朱 晟[1,2] 周丰丰[4]
XUEFENG BRUCE LING[5] 刘国平[6] 陈良标[2]**

([1]浙江大学生命科学院，杭州 310029；[2]中国科学院遗传与发育生物学研究所，北京 100080；
[3]中国科学院自动化研究所，北京 100080；[4]中国科学技术大学计算机系，国家高性能计算中心，合肥 230027；
[5]*Tularik Inc., 1120 Veterans Blvd, South San Francisco*, CA 94080, USA
[6]*School of Electronics, University of Glamorgan, Pontypridd CF37 1DL*, UK)

摘要 随着后基因组时代的到来，批量的测序，特别是 EST 的测序，逐渐成为普通实验室的日常工作. 这些新的序列往往需要进行批量的 Gene Ontology (GO)的注释及随后的统计分析. 但是目前除了 Goblet 以外，并没有软件适合对未知序列进行批量的 GO 注释，而 GoBlet 因为具有上载量的限制，以及仅仅利用 BLAST 作为预测工具，所以仍有许多不足之处. 开发了一个软件包 GoPipe，通过整合 BLAST 和 InterProScan 的结果来进行序列注释，并提供了进一步作统计比较的工具. 主程序接收任意个 BLAST 和 InterProScan 的结果文件，并依次进行文本分析、数据整合、去除冗余、统计分析和显示等工作. 还提供了统计的工具来比较不同输入对 GO 的分布来挖掘生物学意义. 另外，在交集工作模式下，程序取 InterProScan 和 BLAST 结果的交集， 在测试数据集中，其精确度达到 99.1%，这大大超过了 InterProScan 本身对 GO 预测的精确度，而敏感度只是稍微下降. 较高的精确度、较快的速度和较大的灵活性使它成为对未知序列进行批量 Gene Ontology 注释的理想的工具. 上述软件包可以在网站 (http://gopipe.fishgenome.org/) 免费获得或者与作者联系获取.

关键词 Gene Ontology，功能基因组学，EST，BLAST，InterProScan，GOA
学科分类号 Q811.4

---

*国家自然科学基金资助项目 (30330080) .
**通讯联系人. Tel: 010-62554807; Fax: 010-62554807, E-mail: lbchen@genetics.ac.cn
收稿日期: 2004-10-09, 接受日期: 2005-01-07