An Integrated Strategy for Functional Analysis in Large-scale Proteomic Research by Gene Ontology*

LI Dong[1, 2], LI Jian-Qi[1], OUYANG Shu-Guang[1], WU Song-Feng[1],
WANG Jian[1], Xu Xiao-Jie[2], ZHU Yun-Ping[1]**, HE Fu-Chu[1]**
([1]*Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, Beijing* 100850, China;
[2]*College of Chemistry and Molecular Engineering, Peking University, Beijing* 100871, China)

Abstract Data analysis poses a significant challenge to the large-scale proteomics studies. Based on the structured and controlled vocabularies-Gene Ontology (GO), and the GO annotation from related databases, a strategy composed of several programs and local databases is developed to identify the functional distribution and the significantly enriched functional categories of the proteomic expression profile. It would be helpful for understanding the overall functions of these identified proteins and supply the fundamental information for further bioinformatics exploration. This strategy has been successfully used in the Human Fetal Liver (HFL) proteomic research, which is available online at http://www.hupo.org.cn/GOfact/.

Key words bioinformatics, proteomics, expression profile, gene ontology, protein function

Large-scale proteomics study is a breakthrough in experimental biology. It brings up a great amount of proteins that present a significant challenge for the biologists to explore their biological functions. The traditional "literature mining" method is often time-consuming and low efficient, thus great efforts should be made to develop more effective bioinformatics strategies. In recent years, a hierarchical, dynamically controlled vocabulary Gene Ontology (GO) has been constructed to describe known molecular function, subcellular location and biological role of proteins, and gained significant success in biological research, especially in large-scale experimental research [1]. Currently, GO has been widely used to annotate the function of several organisms and become the standard for function annotation. Together with the development of Gene Ontology itself, lots of effective GO applications appeared, such as GoMiner [2], GO-Mapper [3], and GOStat [4]. But they are often used for interpreting microarray results, and none of them has practical application strategy in the proteomics research. During the proteomic research, we have developed a strategy for the functional analysis of the large-scale proteomic results based on the GO hierarchy and the GO annotation supplied by the GO consortium (http://www.geneontology.org/). This strategy is intended to calculate the functional distribution and identify the significantly enriched category for the proteomic data. It has been successfully used in the research of Human Fetal Liver (HFL) proteomic expression profile.

1 Overview of this strategy

This strategy consists of PERL programs and several datasets. The data flow chart is illustrated in Figure 1, and the main steps are implemented as follow.

1.1 Data collection and database construction

GO and GO annotation (GOA) flat format text files, which store the GO hierarchical vocabulary and human protein functional annotation respectively, are downloaded from the GO website and stored in the local relational database. In our proteomic research, IPI (International Protein Index) database is selected for protein validation due to its high quality, high coverage and low redundant degree [5]. Meanwhile, GOA files contain the annotation information of the IPI human proteins. The identified proteins' IPI accession numbers are correlated with GO terms in

*This work was supported by grants from The National Basic Research Program of China(2001CB510209), Hi-Tech Research and Development Program of China (2002BA711A11, 2004BA711A21), Beijing Municipal Science and Technology Project (H030230280590), and The National Natural Science Foundation of China (321003).

**Corresponding author .
HE Fu-Chu.Tel: 86-10-68171208, E-mail: hefc@nic.bmi.ac.cn
ZHU Yun-Ping.Tel: 86-10-66932248, E-mail: zhuyp@hupo.org.cn

Received: April 28, 2005 Accepted: August 31, 2005

GO hierarchy by database retrieve. Therefore, these identified proteins are annotated by GO terms at different levels of GO hierarchy. Besides IPI IDs, this strategy supports more types of IDs, such as Uniprot, RefSeq and Ensembl IDs, and it can convert them into IPI IDs.

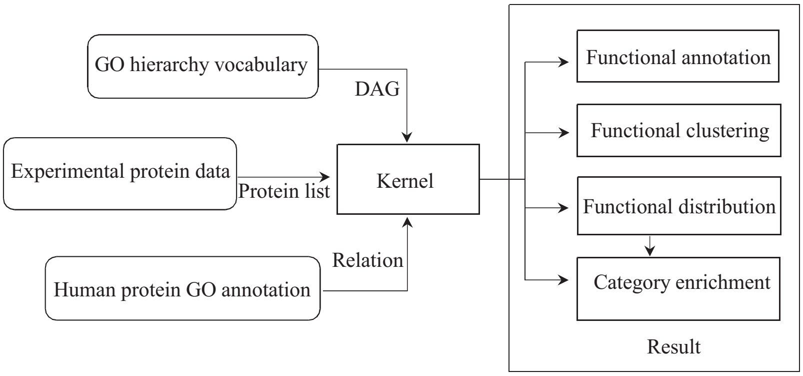

Fig. 1 Flow chart of this strategy

The kernel is a set of PERL scripts, which parses the input data and requirements and then gives out the corresponding results.

1.2 Calculation of functional distribution based on GO DAG structure

To calculate the functional distribution, the number of proteins of each category and its progeny categories should be obtained first. GO is designed to be of Directed Acyclic Graph (DAG) structure, i.e. compared with the classical tree structure, some categories in GO often have more than one parent categories, thus it makes the count more difficult than if the GO were stored in a classical tree structure (Figure 2). Here we introduce the concept of "GO term path", the corresponding path of each protein's GO term is obtained according to the GO hierarchy, then the path is split into several separated sub paths, so a protein can be assigned to its ancestor categories; finally, any duplicated binary relation between the protein and the GO term is removed to avoid the double-count along the traversal.

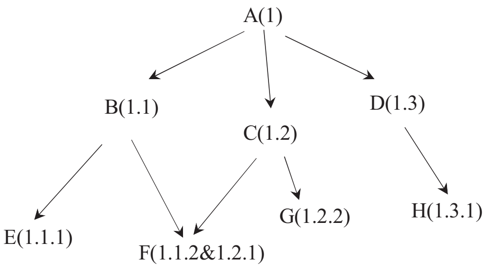

Fig. 2 DAG structure of GO

The capital letters represent the GO terms. The label in the parenthesis besides each GO term represents the path of this GO term, defining the list of GO terms from the top level and the annotated level. Because of the special DAG structure of GO, there are often several paths for one certain GO term, such as GO term "F".

1.3 Identification of significantly enriched/depleted categories by hypergeometry distribution model

For a given expression profile, it is more important to identify the category enrichment/depletion in three ontologies than only the function distribution. The hypergeometry distribution model, which has been used in the Microarray data analysis[6], is introduced in this strategy, and all the IPI human proteins are proposed as the reference dataset. By the program described in **1.2**, we first count the number of proteins having an ontology annotation in the IPI human databases (N_{ipi}) and in the proteomic expression profile (N_{pro}). Then, for each class within the ontology, we count the number of proteins belonging to this class in IPI human database (n_{ipi}) and in the profile (n_{pro}). The standard hypergeometric probability for observing these counts by chance is evaluated as $P(N_{\mathrm{ipi}}, N_{\mathrm{pro}}, n_{\mathrm{ipi}}, n_{\mathrm{pro}})=C(n_{\mathrm{ipi}}, n_{\mathrm{pro}})C(N_{\mathrm{ipi}}-n_{\mathrm{ipi}}, N_{\mathrm{pro}}-n_{\mathrm{pro}})/C(N_{\mathrm{ipi}}, N_{\mathrm{pro}})$, where $C(n, k)$ is the binomial coefficient for n chooses k. The lower and upper tail probabilities are calculated by summing $P(N_{\mathrm{ipi}}, N_{\mathrm{pro}}, n_{\mathrm{ipi}}, n')$ for values of n' from MAX $(0, N_{\mathrm{pro}}+n_{\mathrm{ipi}}-N_{\mathrm{ipi}})$ to n_{pro} and then from n_{pro} to MIN $(N_{\mathrm{pro}}, n_{\mathrm{ipi}})$. The smaller p-value is retained for single-test. It should be pointed out the current GO, GOA and IPI human databases, on which this strategy is heavily dependent, are far from completeness. This may affect the final analysis result. But this limitation is inevitable now, and will vanish gradually as the related databases expand in the near future.

1.4 High level representation of the protein's function

In the functional analysis of a proteomic expression profile, particularly reporting the analytical result, it is more useful to have a high-level view than the full hierarchy view of each of the three ontologies. These subsets of GO are known as "GO slim"[7]. For the description of a dataset and comparison of several datasets, GO consortium has made some general and specific GO slims available for specific analysis. The related program in this strategy can cluster the proteins and present the statistical information by the GO slim terms. Of course, these selected terms can also flexibly meet the users' interest.

2 Application of this strategy in proteomic expression profile research

This strategy takes the list of protein's IPI IDs as input, and gives out the corresponding results. Figure 3 shows an example of the analysis results in one proteomic expression profile research.

The identified proteins are automatically annotated by this strategy and organized as the fundamental information. They can be presented in table form for further analysis (Figure 3a). The proteins sharing a common functional category within

(a)

Protein		Functional category	Ontology
IPI00000010	GO:0000074	Regulation of cell cycle	P
	GO:0003924	GTPase activity	F
	GO:0005525	GTP binding	F
	GO:0007264	Small GTPase mediated signal transduction	P
	GO:0008151	Cell growth and/or maintenance	P
IPI00000110	GO:0003677	DNA binding	F
	GO:0005634	Nucleus	C
	GO:0006355	Regulation of transcription\,DNA-dependent	P
	GO:0008270	Zinc ion binding	F
IPI00000138	GO:0000139	Golgi membrane	C
	GO:0003827	Alpha-1\,3-mannosylglycoprotein 2-beta-N-acetylglucos	F
	GO:0005975	Carbohydrate metabolism	P
	GO:0006023	Aminoglycan biosynthesis	P
	GO:0006487	N-linked glycosylation	P
	GO:0016021	Integral to membrane	C
	GO:0016757	Transferase activity\, transfering glycosyl groups	F
IPI00000335	GO:0005739	Mitochondrion	C

(b)

Functional category	Count	Proteins
GO:0000050	5	IPI00003389 IPI00011062 IPI00220267 IPI00291560 IPI00295363
GO:0000059	3	IPI00001639 IPI00185146 IPI00329200
GO:0000060	2	IPI00001639 IPI00007307 IPI00005791 IPI00291939 IPI00306400
GO:0000074	8	IPI00000010 IPI00004390 IPI00007307 IPI00008380 IPI00008810 IPI00013393 IPI00013890 IPI00017334
GO:0000075	1	IPI00291939
GO:0000122	7	IPI00006282 IPI00013216 IPI00013394 IPI00027151 IPI00028828 IPI00030404 IPI00298887
GO:0000139	6	IPI00000138 IPI00004671 IPI00016720 IPI00026530 IPI00031583 IPI00220219

(c)

Ontology	Functional category	n_{pro}	PER	p-Value	DIR	n_{IPI}	N_{IPI}	N_{pro}
Process	GO:0008152:metabolism	1 181	70.93%	2.427E−36	++	10 253	18 094	1 665
Process	GO:0006350:metabolism	275	16.52%	2.703E−01	+	2 889	18 094	1 665
Component	GO:0005634:nucleus	535	35.01%	2.996E−06	++	4 717	1 5787	1 528
Component	GO:0016020:membrane	337	22.05%	4.095E−39	−−	5 817	1 5787	1 528
Function	GO:0004871:signal transducer activity	145	7.62%	1.599E−37	−−	3 565	2 0606	1 902
Function	GO:0005488:binding	1 164	61.20%	1.660E−04	++	11 809	2 0606	1 902
Function	GO:0003676:nucleic acid binding	529	27.81%	4.252E−05	++	4 959	2 0606	1 902
Function	GO:0003824:catalytic activity	765	40.22%	4.412E−13	++	6 750	2 0606	1 902
Function	GO:0003754:chaperone activity	54	2.84%	2.111E−13	++	199	2 0606	1 902
Function	GO:0030234:enzyme regulator activity	66	3.47%	5.049E−01	+	712	2 0606	1 902
Function	GO:0003774:motor activity	40	2.10%	8.284E−05	++	232	2 0606	1 902

Fig. 3 Example of functional analysis results of proteomic data

(a) Functional annotation for the identified proteins. (b) Functional cluster of identified proteins. (c) The functional distribution and category enrichment by the reduced GO terms. Categories are "ONTOLOGY" (biological process, molecular function or cellular component), PER (percentage of proteins belonging to this category), "PVAL" (p-value from hypergeometric distribution prior to correcting for single testing) and "DIR" (note for the enrichment/depletion of the function classes, "++" for significantly enriched, "+" for enriched, "−" for depleted and "−−" for significantly depleted).

the hierarchy could be grouped into several clusters by corresponding programs. Users can specify the GO terms at any level within the hierarchy they are interested in. (Figure 3b).

Functional distribution and categories enrichment are demonstrated in Figure 3c. The categories with large number proteins suggest that the proteins in these categories are very active in this profile and expressed in high or at least middle abundance. Compared with all human proteins deposited in IPI, the enriched categories mean the proteins in these categories are overpresented in this profile and reflect the biological specific categories of these data. To gain the comprehensive understanding of the protein functions， these two aspects should be considered together. Sometimes the two aspects agree with each other: the large category is enriched; but sometimes they don’t: proteins in one category have a relatively large proportion in the expression profile, but compared with the full dataset, they are significantly depleted. This contrasting phenomenon may suggest the proteins related to this category have high tissue/organ specificity.

3 Discussion

One of the major challenges in high-throughput proteomic experiments is to elucidate the biological function under the high volume data. Here we report an integrated GO application strategy in proteomic expression profile analysis. The automatic functional annotation and clustering provide the fundamental functional information, the statistical analysis presents the further understanding of the expression profile’s function, and the induction of “GO-slim” facilitate the users to interpret and report the analysis result. Although this strategy is developed during the analysis of human liver proteomic expression profiles, it can also be used in other organisms, such as yeast, worm and fly. As we know, this is the most comprehensive and practical GO application strategy in proteomic analysis.

References

1 Ashburner M, Ball C, Blake J, *et al*. Gene ontology: tool for the unification of biology. Nat Genet, 2000, **25** (1): 25~29

2 Zeeberg B R, Feng W, Wang G, *et al*. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol, 2003, **4** (4): R28

3 Smid M, Dorssers L C. GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. Bioinformatics, 2004, **20** (16): 2618~2625

4 Beissbarth T, Speed T P. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics, 2004, **20** (9): 1464~1465

5 Kersey P J, Duarte J, Williams A, *et al*. The international protein index: an integrated database for proteomics experiments. Proteomics, 2004, **4** (7): 1985~1988

6 Draghici S, Khatri P, Bhavsar P, *et al*. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. Nucl Acids Res, 2003, **31** (13): 3775~3781

7 Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research, 2004, **32** (Database Issue): D258~D261

高通量蛋白质组学研究中一种基于 GO 的蛋白质功能分析策略 *

李 栋[1,2] 荔建琦[1)] 欧阳曙光[1)] 吴松锋[1)] 王 建[1)] 徐筱杰[2)] 朱云平[1)**] 贺福初[1)**]

([1)]北京放射医学研究所，北京 100850；[2)]北京大学化学与分子工程学院，北京 100871)

摘要 挖掘高通量实验数据蕴含的生物学意义是蛋白质组学研究面临的一大挑战. 基于等级化结构化的词汇表 GO (Gene Ontology) 和相关数据库中的蛋白质功能注释，发展了一种对蛋白质组学研究中得到的表达谱 (Expression profile) 进行功能分析的策略. 在对蛋白质表达谱进行功能注释的基础上给出蛋白质表达谱中蛋白质功能的分布，同时给出感兴趣功能类别的统计信息. 这有助于对表达谱蛋白质功能的整体理解和深入的生物信息学分析. 该策略已经成功应用胎肝蛋白表达谱研究中，用户可以通过访问网址 http://www.hupo.org.cn/GOfact/ 使用或者下载我们的程序.

关键词 生物信息学，蛋白质组学，表达谱， GO，蛋白质功能

学科分类号 Q61

*国家重点基础研究发展计划(973)(2001CB510209), 国家高技术研究发展计划(863)(2002-BA711A11, 2004BA711A21), 北京市重大科技专项(H030230280590)和国家自然科学基金资助项目(321003).

** 通讯联系人. 贺福初. Tel: 010-68171208, E-mail: hefc@nic.bmi.ac.cn; 朱云平. Tel: 010-66932248, E-mail: zhuyp@hupo.org.cn

收稿日期：2005-04-28, 接受日期：2005-08-31