

大规模 GO 注释的生物信息学流程

黄子夏,柯才焕,陈 军*

(厦门大学海洋与环境学院,福建 厦门 361005)

摘要: 随着新一代测序技术的不断发展,海量的序列数据将为生物学者挖掘基因信息提供巨大的资源. 信息挖掘的一项重要工作是对序列进行功能注释,其中最重要的功能注释方式是基因本体论(Gene Ontology, GO)的注释. 利用生物信息学方法和软件工具集成了针对 EST 序列的大规模 GO 注释流程(large-scale GO annotation pipeline, LSGAP). 该流程集合了 BLAST、B2g4pipe 以及 Wego 等软件和 Swissprot、Nr 或 Interpro 等常用蛋白数据库. 用户可以将 EST 序列通过此流程最终获得可视化的 GO 分类统计图表,直观地显示基因在不同过程中的参与情况. 为了验证 LSGAP 的准确性,对 2007 年发表的美洲牡蛎(*Crassostrea Virginica*)的 EST 序列进行了 LSGAP 分析,结果表明 GO 分析非常准确有效. 通过与 Blast2go 和 GoBlast 等 GO 注释软件进行比较,LSGAP 流程具有可以本地化运行 BLAST、对硬件要求低和运行时间短等诸多优势,因此 LSGAP 流程是科研人员进行基因功能挖掘的有效工具.

关键词: LSGAP; GO 注释; 基因功能; 生物信息学

中图分类号: Q 7; TP 319

文献标志码: A

文章编号: 0438-0479(2012)01-0139-05

随着新一代测序技术的不断发展,越来越多物种的基因组和转录组获得了高通量测序,给人们带来了海量的核酸序列数据. 如何对海量序列中隐含的生物信息进行高通量的解读是研究者要面对的问题. 序列信息解读的一个核心内容是对基因进行功能注释,具体操作上就涉及到一个概念:基因的本体论^[1](Gene Ontology, GO)注释. 大多数基因在不同生物中的同源基因拥有相同的主要生物学功能,因此,在某些物种里已知的基因功能信息可以用来解释其他物种对应的同源基因. 但是这些已知的功能信息包含在先前积累的浩瀚文献之中,不同的文献会用不同的词汇来描述同一生物学功能,这为功能检索和注释带来诸多不便. 而 GO 项目^[1]就是建立一套特定的词汇集合来描述生物学功能,以此对基因功能注释统一化. 用于描述生物学功能的词汇必然要反映生物学功能的本质,此即本体论(Ontology)的由来.

GO 项目对基因功能进行了一致性描述,开发了可控制的词汇表,且无物种特异性. 目前已经建立了三大独立的本体论词汇表:生物过程(biological process)、细胞组分(cellular component)和分子功能

(molecular function). 这三大本体论词条下面又可以独立出不同的亚层次,以“有向非循环图(directed acyclic graphs)”的方式层层向下,将每一个本体论词条串联起来形成树状结构. 将 GO 词条分配给基因序列的行为即 GO 注释. 目前,GO 注释已经成为高通量测序序列的主要注释方法.

综上所述,一个基因编码的蛋白质可以在多种水平上被赋予本体论定义,GO 注释的原理就是通过计算机程序建立基因产物与用于定义它们的本体论词条之间的联系. 自从 1998 年基因本体论项目^[1](Gene Ontology Project)创立以来,已经有多款 GO 注释软件发表,实现了基因产物的标准化注释. 比如 2004 年发布的 GOstat^[2]和 2006 年发布的 GoBlast^[3],主要以序列网页递交方式进行注释,首先将序列上传于网络服务器,后台运行分析后,再将 GO 注释信息返回用户界面,用户可下载获取结果. 其缺点是一次只能上载少于 100 条待注释序列,大批量基因功能注释的效率很低;又如 2005 年发布的 Blast2go^[4]和 2008 年发布的 Annot8r^[5],主要通过本地化方法进行注释,首先将序列通过与 NCBI 蛋白数据库的比对获得比对报告,接着将庞大的 GO 注释信息导入本地化数据库,最后将比对报告与本地化 GO 数据库建立联系,从而完成序列的 GO 注释. 其缺点是需要下载大量与 GO 注释相关的数据库,内存需求大,对注释序列的数目存在限制,且生成 GO 注释统计图表需要有较强的硬件系统支持,故限制了应用范围,不利于在中小型实验室普

收稿日期:2011-05-04

基金项目:国家重点基础研究发展计划(973)项目(2010CB126403);

国家自然科学基金项目(40976093);福建省青年科技人才

创新项目(2008F3098)

* 通信作者:chenjun@xmu.edu.cn

及. 受益于第二代高通量测序技术的商业化, 目前许多小型实验室也能获得宏量的基因组或转录组数据. 但小型实验室往往会受限于硬件和网络环境, 无法展开高效的 GO 注释, 因此开发适合小型实验室使用的高效 GO 注释流程就很有必要. 本文针对这个需求, 首次开发和集成了基于 PC/UNIX 系统、完全免费的软件和公共数据库的大规模 GO 注释流程 (large-scale GO annotation pipeline, LSGAP), 通过串联的方式将几种软件有机结合, 可以在当前主流 PC 机或笔记本电脑上, 以较短的运行时间完成高通量的 GO 注释. LSGAP 的优点是: 序列高通量、注释速度快、对硬件要求低且注释准确, 为中小型实验室提供了有效可行的 GO 注释方法, 大大降低了成本.

1 系统与方法

1.1 开发环境

硬件环境: 联想 LENOVO F41M 型号双 CPU 笔记本电脑, 2 GB 内存, 120 GB SATA 硬盘, 2 M 电信宽带.

操作系统: 为 Windows XP, 以 VMware Player 3.0 搭载 Ubuntu10.04 虚拟 Linux 系统.

1.2 下载软件及数据库

1) BLAST2. 2. 18 序列比对软件, 从 <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LAT-EST/> 获取;

2) B2g4pipe2. 3. 4^[4] GO 注释软件, 为 LSGAP 的核心程序, 从 http://www.blast2go.org/_media/b2gpipe234.tar.gz 获取;

3) 浏览器程序, 为 Browser/Server 架构;

4) Java 6. 0, 从 <http://www.java.com/download> 获取;

5) Swissprot 蛋白数据库和 Nr 非冗余蛋白数据库, 从 <ftp://ftp.ncbi.nlm.nih.gov/blast/db/> 获取;

6) Annot8r1. 1. 1 GO 注释软件, 从 <http://www.nematodes.org/bioinformatics/annot8r/> 获取.

1.3 LSGAP 的大规模 GO 注释流程

LSGAP 流程 (图 1) 整合了多个基因识别和 GO 注释软件, 如 BLAST^[6]、B2g4pipe 和 Wego^[7] 等, 以及多个生物序列数据库, 如 Swissprot、非冗余蛋白数据库 Nr 等. 首先利用 BLAST 软件将 FASTA 格式^[8] 的 EST 序列进行相似性比对, 生成 xml 格式的 BLAST 报告文件; B2g4pipe 分析 BLAST 报告文件, 通过网络

服务器进行 GO 注释, 生成 GO 注释文件; 接着利用 PERL^[9] 脚本语言将 GO 注释文件转化成 Wego 程序可识别的格式; 最后利用 Wego 网页程序生成 GO 注释的统计图表.



图 1 LSGAP 流程图

Fig. 1 LSGAP flow diagram

2 数据演示

2.1 测试数据

本系统采用美洲牡蛎 (*Crassostrea virginica*) EST 序列^[10] 作为测试数据, 总共 5 542 条序列, 总碱基数为 2 622 701 bp.

2.2 测试过程

2.2.1 BLAST 序列相似性比对

在 Ubuntu10.04 系统中, 进入工作目录, 此目录下包含美洲牡蛎 EST 序列文件 EST.fasta 以及格式化的 Swissprot 蛋白数据库, 通过命令行运行 BLAST 程序, 即:

```
$ blastall -p blastx -i EST.fasta -d Swissprot -o blast_result.xml -e 1e-5 -v 2 -m 7
```

该命令的参数解析如下: -p: 使用 blastx 程序; -i: 输入的 EST 序列文件名; -d: 所使用的数据库名称; -o: 存放 BLAST 报告的文件名称; -e: E 值; -v: 仅显示分值最高的两条比对序列; -m 7: 输出的 BLAST 报告为 xml 格式.

2.2.2 B2g4pipe 实现 GO 注释

B2g4pipe 是一款基于 Blast2go 的命令行界面 (command line interface) GO 注释软件, 是 LSGAP 流程的核心. 与 Blast2go (GUI) 相比, 它最大的特点就是注释速度快, 对硬件要求低, 且注释准确. 确保网络畅通, 进入工作目录, 此目录下包含 B2g4pipe 软件以及 BLAST 报告文件, 通过命令行运行程序, 即:

```
$ java -jar blast2go.jar -in blast_result.xml -out  
GO.annot -a
```

该命令的参数解析如下: -in: 输入的 xml 格式的 BLAST 报告; -out: 输出文件的名称; -a: 程序产生 GO 注释文件. 有关 B2g4pipe 其他的一些重要参数列于表 1. 有关 B2g4pipe 软件的具体安装需求和运用方法详见软件压缩包的 Readme 文件.

通过以上程序的运行, 5 542 条美洲牡蛎 EST 序列中有 1 477 条序列有 GO 注释词条, 而美洲牡蛎 EST 序列分析文章^[10] (Jonas 等) 中有 1 104 条序列存在 GO 注释. 相比之下本流程获得的注释更加丰富. GO 注释文件的部分内容列于表 2.

2.2.3 Wego 程序生成可视化统计图表

由 B2g4pipe 程序生成的 GO 注释结果 GO.annot 文件, 通过 PERL 脚本语言编程, 转化成 Wego 网页程序可以识别的格式, 将注释文件上载于 Wego 网页^[7], 运行结果如图 2.

为了评估 LSGAP 流程的准确性, 针对美洲牡蛎 EST 分析文章^[10] 的部分二级或三级词条进行百分比分析, 结果比较列于表 3. 由表 3 可知, 二者 GO 注释各词条下序列所占百分比基本一致, 故证明了 LSGAP 流程的准确性.

3 几种不同的 GO 注释软件效率比较

本文将 LSGAP 流程同几款不同的 GO 注释软件

表 1 B2g4pipe 的一些重要参数

Tab. 1 The important parameters of B2g4pipe

参数	描述
-v	即时打印程序运行信息报告
-d	产生 GO 注释项目文件(.dat)
-prop	应用程序的参数设置文件
-Xms	运行程序时所用的最小内存(MB)
-Xmx	运行程序时所用的最大内存(MB)

表 2 GO 注释文件部分结果

Tab. 2 The partial results of GO annotation file

序列 GI 号	GO 号	描述
gi 152812318	GO:0016740	snf1-相关激酶
gi 152812336	GO:0031902	escrt-i 复合亚基 vps28
gi 152812357	GO:0006950	热休克 70 ku 同源蛋白
gi 152812618	GO:0051903	乙醇脱氢酶-3
gi 152815191	GO:0044464	谷氨酸缩肽酶-2
gi 152815210	GO:0008152	过氧化物酶体肌氨酸氧化酶
gi 152815266	GO:0043066	线粒体超氧化物歧化酶
gi 152817646	GO:0005516	肌球蛋白重横纹肌
gi 152818063	GO:0000398	剪切因子 3a 亚基 1
gi 152819004	GO:0043234	微管蛋白 beta-2c 链

在相同硬件和网络环境下比较运行时间, 发现 LSGAP 流程在运行相同的数据量所耗费的时间最短, 效

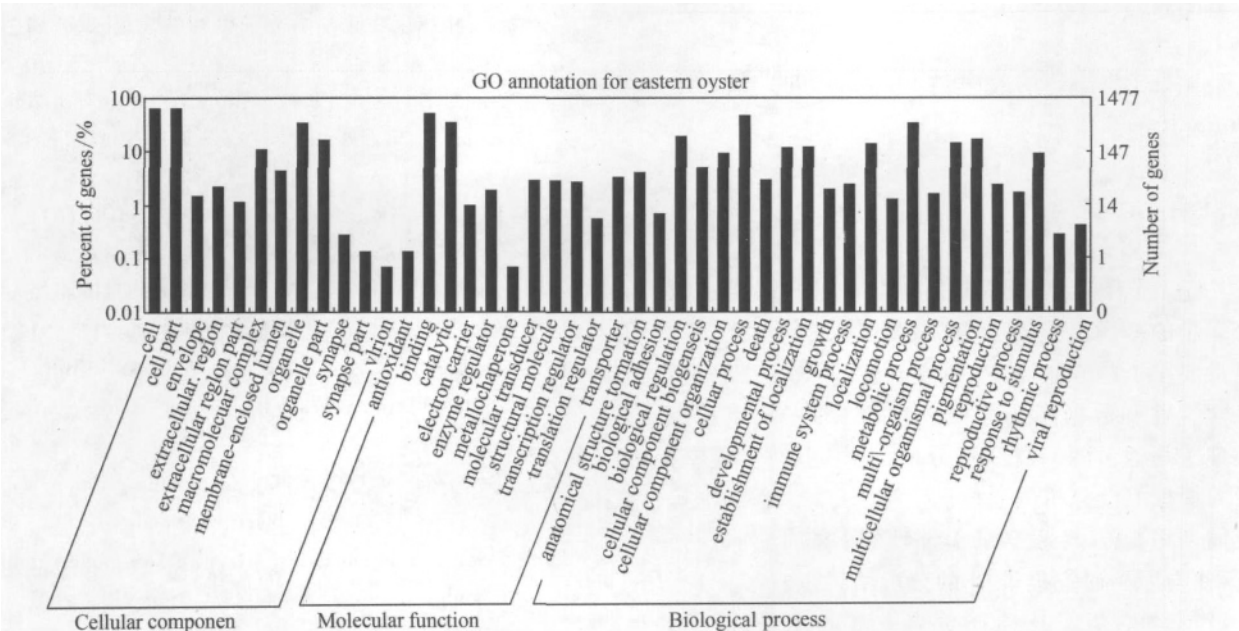


图 2 美洲牡蛎 EST GO 注释柱状图

Fig. 2 GO annotation histogram for eastern oyster EST of *Crassostrea Virginica*

表 3 2 种不同方法 GO 注释序列百分比部分结果比较

Tab. 3 Comparisons between partial results of two different GO annotation software

细胞组分 GO 词条	Jonas 等 ^[10]	LSGAP
注释基因数目	1104	1477
细胞部分	96%	95%
膜结合的细胞器	45%	41%
非膜结合的细胞器	15%	15%
膜泡	2%	3%
细胞器膜	2%	2%
细胞器内腔	6%	7%
胞外基质	2%	1%

率最高. 实验用 LSGAP 比较了 Blast2go v3(不下载 GO 数据库,通过网络注释)以及 Annot8r1.1.1 软件,分别采用 50,500,1 000 条平均长度在 300 bp 左右的序列进行测试,测试结果列于表 4(表中的时间均扣除了序列 BLAST 使用的时间,仅为注释时间).

表 4 几种不同 GO 软件运行效率的比较

Tab. 4 Effectiveness comparison among different GO annotation software min

序列数目	LSGAP	Blast2go	Annot8r
50	4	18	10
500	32	160	107
1000	65	291	220

可见 LSGAP 的效率是 Blast2go 的 4 倍以上,是 Annot8r 的 3 倍以上.

4 结 论

GO 注释是对基因产物功能研究的重要手段,随着海量的基因组或转录组数据的涌现,对 GO 注释的通量要求也越来越高. 本研究建立了将免费的生物信息学软件和数据库整合在一起的 LSGAP 流程,最终实现了高通量的 GO 注释分析,为中小型实验室提供了大规模 GO 分析的手段. 为了测试本流程的准确性,我们利用 2007 年发表的美洲牡蛎 EST 序列进行了测试,并与原始研究报告^[10]的分析结果进行了比较. LSGAP 流程获得了比原作更为丰富的注释:原作获得 1 104 条 EST 序列的 GO 注释,而我们得到了 1 477 条 EST 序列的 GO 注释,比原作高了 34%. 这一方面说

明了数据库较几年前有了一定的更新,另一方面也说明了本流程的有效. 同时将 LSGAP 与其他 GO 注释软件进行了运行速度的对比,发现在相同硬件和网络环境下,LSGAP 流程注释的速度和效率比 Blast2go 和 Annot8r 有成倍的提升,对大批量(大于 10 000 条序列)GO 注释的实现是重要的推进. 因此 LSGAP 是科研人员进行基因功能开发的实用工具.

值得一提的是,目前所有的生物信息学注释软件均不能保证完全正确,如 LSGAP 流程依赖的是现有数据库中的序列注释信息,而这些注释信息并不能保证完全准确. 但是,我们可以通过其他互补的方法提高注释的准确度,如将序列平行地同 Swissprot、Nr 或者 CDD(conserved domain database)等数据库进行比对,若均有搜索到相似序列,那么这些序列的比对分值也势必较高,注释结果也越准确;或者可以采用 PSI-BLAST(position-specific iterated BLAST)^[11],此程序灵敏度高于 BLASTP 程序,对于发现远源相似蛋白或者某个蛋白家族的新成员非常有效.

参考文献:

- [1] Harris M A, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resources [J]. Nucleic Acid Res, 2004, 32(Database Issue): 258-261.
- [2] Beissbarth T, Speed T P. Gostat: find statistically over-represented Gene Ontologies within a group of genes [J]. Bioinformatics, 2004, 20(9): 1464-1465.
- [3] 王成刚, 莫志宏. 整合 Blast 搜索与 GO 注释软件 GoBlast [J]. 中国生物化学与分子生物学学报, 2006(12): 1003-1006.
- [4] Ana C, Stefan G. Blast2go: a universal tool for annotation, visualization and analysis in functional genomics research [J]. Bioinformatics, 2005, 21: 3674-3676.
- [5] Ralf S, Mark L. Annot8r: GO, EC and KEGG annotation of EST databases [J]. BMC Bioinformatics, 2008, 9: 180.
- [6] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool [J]. J Mol Biol, 1990, 215: 403-410.
- [7] Ye J, Fang L, Zheng H, et al. WEGO: a web tool for plotting GO annotations [J]. Nucleic Acids Res, 2006, 34: 293-297.
- [8] Bethesda M D. FASTA format description [DB/OL]. [2005-03-22]. <http://blast.wustl.edu/>.
- [9] Jason E S, David B, Kris B, et al. The bioperl toolkit: perl modules for life sciences [J]. Genome Res, 2001, 12(10): 1611-1618.
- [10] Jonas Q, Shaolin W, Ping L, et al. Generation and analysis of ESTs from the eastern oyster, *Crassostrea virginica*

- Gmelin and identification of microsatellite and SNP markers [J]. BMC Genomics, 2007, 8:157.
- [11] Stephen F, Thomas L, Madden A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. Nucleic Acid Res, 1997, 25(17): 3389-3402.

Bioinformatics Procedure of Large-scale GO Annotation

HUANG Zi-xia, KE Cai-huan, CHEN Jun*

(College of Oceanography and Environmental Science, Xiamen University, Xiamen 361005, China)

Abstract: With the fast development of next-generation sequencing technologies, a large number of biological data will provide tremendous sequence resources to biologists in gene exploitation. An important task on data mining is to annotate genes with functions, and the most important method is Gene Ontology (GO) annotation. This research formed the procedure of large-scale GO annotation pipeline for EST sequences, utilizing bioinformatics methodologies and software tools. This procedure encompasses different software like BLAST, B2g4pipe and Wego, together with Swissprot, Interpro or Nr protein databases. Users can put EST sequences with FASTA format through this system and ultimately gain visualized GO distribution statistics diagrams, which demonstrate the situations of the genes involved in different processes. In order to test and verify the preciseness of LSGAP, the EST sequences of eastern oyster published in 2007 were gone through this pipeline, and the results demonstrated that LSGAP procedure was quite accurate and efficient. Compared with other GO annotation software such as Blast2go (Graphical User Interface) and GoBlast, LSGAP procedure has many advantages: running BLAST software locally, without downloading many GO relative databases and consuming less time. All of the results demonstrated that LSGAP is an efficient tool for researchers to do data mining.

Key words: LSGAP; GO annotation; gene function; bioinformatics