Article ID: 1671-8224(2007)03-0151-04

To cite this article: WANG Cheng-gang , MO Zhi-hong. Integrating Gene Ontology and Blast to predict gene functions [J]. J Chongqing Univ: Eng Ed (ISSN 1671-8224), 2007, 6(3): 151-154.

Integrating Gene Ontology and Blast to predict gene functions

WANG Cheng-gang [1,a]***, MO Zhi-hong*** [2]

[1] *College of Bioengineering, Chongqing University, Chongqing 400030, P.R. China*

[2] *College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400030, P.R. China*

Received 11 November 2006; revised 17 January 2007

Abstract: A GoBlast system was built to predict gene function by integrating Blast search and Gene Ontology (GO) annotations together. The operation system was based on Debian Linux 3.1, with Apache as the web server and Mysql database as the data storage system. FASTA files with GO annotations were taken as the sequence source for blast alignment, which were formatted by wu-formatdb program. The GoBlast system includes three Bioperl modules in Perl: a data input module, a data process module and a data output module. A GoBlast query starts with an amino acid or nucleotide sequence. It ends with an output in an html page, presenting high scoring gene products which are of a high homology to the queried sequence and listing associated GO terms beside respective gene poducts. A simple click on a GO term leads to the detailed explanation of the specific gene function. This avails gene function prediction by Blast. GoBlast can be a very useful tool for functional genome research and is available for free at http://bioq.org/goblast.

Keywords: gene ontology; Blast; gene function prediction

CLC number: Q7 Document code: A

1 Introduction

Recent efforts in high-throughput sequencing have achieved a rapid increase of sequences available in public databases. To use these sequences to efficiently predict gene genic functions has become a focus. To enhance the reliability of gene annotation and the speed of functional genomic research, we developed GoBlast system, which is a software package that integrates Blast search and gene ontology (GO) annotation to predict gene functions. It combines, in a single analysis step, blast search results directly with annotations from GO.

The Gene Ontology [1] project is a collaborative effort to address the need for consistent descriptions of gene products in different databases. Biologists currently waste a lot of time and effort in searching for all of the available information about each small area of research. This is hampered further by the wide variations in terminology that may be commonly used sometime; the discrepancies inhibit effective searches by computers as well as people. For example, if you are searching for new targets for antibiotics, you may want to find all the gene products that are involved in bacterial protein synthesis, and that have significantly different sequences or structures from those in humans. But if one database describes these molecules as being involved in "translation", whereas another uses the phrase "protein synthesis", it will be difficult for you—and even harder for a computer—to find functionally equivalent terms. The GO collaborators [2] are developing three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. The three organizing principles of GO are molecular function, biological process and cellular component. A gene product has

[a] WANG Cheng-gang (王成刚): Male; Born 1978; PhD candidate; Research interest: genetic engineering; E-mail: wangcg@stu.cqu.edu.cn, cqwcg@163.com; Mobile phone: +86-13883079895.

one or more molecular functions and is used in one or more biological processes; it may be associated with one or more cellular components. Now, Gene Ontology is a standard tool to annotate gene function.

Blast [3] is a basic tool to predict novel gene function, but scientists can not find gene GO annotations based on a raw result of blast. A user first does blast search, then uses the blast search result to query the gene GO annotation on the website of Gene Ontology. This is time-taking, especially when the number of blast results is very large. To enhance the speed, we developed a GoBlast system. It combines Blast and GO search in one action, and gives both blast results and GO annotations in GoBlast results.

2 System construction

2.1 Software and hardware

We used open source software, Debian Linux 3.1 (http://www.debian.org) [4], as the operation system. The hardware of the computer comprised AMD 2800+CPU, 1G DDR RAM, and 160G SATA hard disk. The system architecture was B/S (Browser/Server). This is the most common architecture of bioinformatics system. A client computer can run the GoBlast software over the network only when it has a web browser. We selected Apache server (http://www.apache.org) [5] as the web server and Mysql database [6] as the data storage system, and encoded GoBlast in Perl in a PC containing a 2800+AMD athlon and linux operation system.

2.2 Data sets

We downloaded FASTA [7] files with GO annotations to use as the sequence source for blast alignment. They were from http://archive.godatabase.org, which had 85 662 protein sequences. The FASTA files were formatted by *wu-formatdb* program (http://blast.wustl.edu) to satisfy the format of *wublast* program [8]. We downloaded the most recent GO database (2005-Sep-04) from http://archive.godatabase.org, which had 219 599 GO terms, and loaded its data into our local Mysql database as the source for integrating GO terms with blast search result.

2.3 Building GoBlast

We designed the GoBlast system using Bioperl [9] modules in Perl. The system was made up of three modules: a data input module, a data process module and a data output module. We uploaded GoBlast online at http://bioq.org/goblast, with all source data and help documents available.

2.3.1 Data input module

A user starts a GoBlast query with an amino acid or nucleotide sequence. GoBlast selects an appropriate algorithm for either case. A FASTA sequence can be either pasted into the query window or loaded from a local hard-disk using the browsing function (Fig. 1). The sequence in the FASTA file is used as the query sequence. The threshold for query has to be selected.

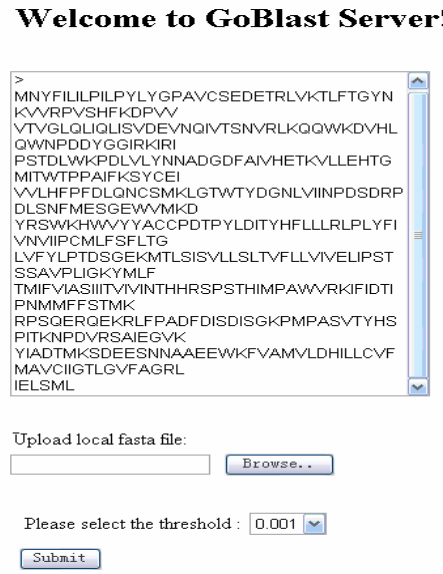

Fig. 1 Data input page

2.3.2 Data process module

The data process module calls Blast to align the input sequence according to the sequence database, and creates a temporary file with the blast results. Then, it selects GO terms from the temporary file, and

searches detailed GO information from the Mysql database using the GO terms. The blast and GO results are submitted to the data output module.

2.2.3 Data output module

The GoBlast results are output in an html page (Fig. 2a). The top of the page are the threshold and the query sequence. Below is a table showing a list of gene products that have a high homology to the query sequence. The first column of the table shows the symbols of the gene products of high homology. The next column shows the GO data sources where the gene-product information is from, if available. The third and fourth columns show the terms associated to the corresponding gene product in the GO and the evidence codes used to link them. Clicking on a gene symbol leads a user to the page of details of a gene product. Clicking on a term name brings a user the page about the term. Finally, the user sees the raw results of the Blast query (Fig. 2b). A user can compare the GoBlast results with the raw BLAST results to understand what GoBlast brings us and how GoBlast increases the search time.

a

Your GoBlast Query

Threshold: 0.001

Query Sequence

```
>               MNYFILILPILPYLYGPAVCSEDETRLVKTLFTGYNKVVRPVSHFKDPVV
VTVGLQLIQLISVDEVNQIVTSNVRLKQQWKDVHLQWNPDDYGGIRKIRI PSTDLWKPDLVLYNNADGDFAIVHETKVLLEHTGMITWTPPAIFKSYCE
VVLHFPFDLQNCSMKLGTWTYDGNLVIINPDSDRPDLSNFMESGEWVMKD YRSWKHWVYYACCPDTPYLDITYHFLLLRLPLYFIVNVIIPCMLFSFLT
LVFYLPTDSGEKMTLSISVLLSLTVFLLVIVELIPSTSSAVPLIGKYMLF TMIFVIASIIITVIVINTHHRSPSTHIMPAWVRKIFIDTIPNMMFFSTM
RPSQERQEKRLFPADFDISDISGKPMPASVTYHSPITKNPDVRSAIEGVK YIADTMKSDEESNNAAEEWKFVAMVLDHILLCVFMAVCIIGTLGVFAGR
IELSML
```

High Scoring Gene Products

Gene Symbol	Score (P)	Datasource	Associated Terms	Aspect
☐ ACHA_BRARE	1.9e-250	UniProt	cation channel activity	molecular_function
			motor axon guidance	biological_process
			neuromuscular synaptic transmission	biological_process
			regulation of axon extension	biological_process
☐ chrna1	1.9e-250	ZFIN	cation channel activity	molecular_function
			motor axon guidance	biological_process
			neuromuscular synaptic transmission	biological_process
			regulation of axon extension	biological_process

b

Raw Blast Data

```
BLASTP 2.0MP-WashU [10-May-2005] [linux26-i786-ILP32F64 2005-05-10T21:12:31]

Copyright (C) 1996-2005 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

Reference:  Gish, W. (1996-2005) http://blast.wustl.edu



Database:  go_20050904-seqdblite.fasta
           85,662 sequences; 41,190,799 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                    Smallest
                                                                      Sum
                                                             High  Probabilit
Sequences producing High-scoring Segment Pairs:              Score  P(N)

UNIPROT|Q98880 SPTR:Q98880 symbol:ACHA_BRARE [GO:0005261 ...  2406  1.9e-250
ZFIN|ZDB-GENE-980526-137 SPTR:Q98880 symbol:chrna1 "choli...  2406  1.9e-250
UNIPROT|P25108 SPTR:P25108 symbol:ACHA_RAT [GO:0004889 "n...  1888  1.5e-195
UNIPROT|P04756 SPTR:P04756 symbol:ACHA_MOUSE [GO:0005615 ...  1887  1.9e-195
MGI|MGI:87885 SPTR:P04756 symbol:Chrna1 "cholinergic rece...  1887  1.9e-195
UNIPROT|P02708 SPTR:P02708 symbol:ACHA_HUMAN [GO:0004889 ...  1686  3.0e-192
UNIPROT|O70174 SPTR:O70174 symbol:ACHA4_MOUSE [GO:0004889...  1005  2.6e-117
MGI|MGI:87888 SPTR:O70174 symbol:Chrna4 "cholinergic rece...  1005  2.6e-117
UNIPROT|P43681 SPTR:P43681 symbol:ACHA4_HUMAN [GO:0004889...  1004  3.3e-117
UNIPROT|P32297 SPTR:P32297 symbol:ACHA3_HUMAN [GO:0004889...   996  6.8e-117
UNIPROT|P09483 SPTR:P09483 symbol:ACHA4_RAT [GO:0004889 "...   996  2.3e-116
UNIPROT|Q8R4G9 SPTR:Q8R4G9 symbol:ACHA3_MOUSE [GO:0016021...   994  3.7e-116
MGI|MGI:87887 SPTR:Q8R4G9 symbol:Chrna3 "cholinergic rece...   994  3.7e-116
UNIPROT|P04757 SPTR:P04757 symbol:ACHA3_RAT [GO:0005892 "...   991  7.7e-116
UNIPROT|Q15822 SPTR:Q15822 symbol:ACHA2_HUMAN [GO:0004889...   958  6.9e-113
UNIPROT|Q9R0W9 SPTR:Q9R0W9 symbol:ACHA6_MOUSE [GO:0005615      963  1.4e-112
```

Fig. 2 Results of searching with the amino acid of acetylcholine receptor protein and with a threshold of 0.001: (a) the top of output page; and (b) the bottom of output page: raw blast results

3 Discussion

We tested the GoBlast system with the sequence of acetylcholine receptor protein, setting the Blast threshold at 0.001. The search provided ACHA BRARE [10] the first high-score gene product, with its associated GO terms beside (Fig. 2a). To know the meaning of a GO term, e.g. "cation channel activity", simply clicking on it directly led to the URL (uniform resource locator) of the web page about it. However, on the raw blast result page (Fig. 2b), no links to GO terms were available. GoBlast provides more information than regular Blast. It is easier and more efficient to use for gene function prediction.

References

[1] Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource [J]. Nucleic Acids Research, 2004, 32 (Database issue): 258-261.

[2] Stanford University. Gene ontology: tool for the unification of biology [EB/OL]. [cited 2005-12-04]. http://www.geneontology.org.

[3] Altschul S, Madden TL, Schaffer AA. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs [J]. Nucleic Acids Research, 1997, 17: 3389-3402.

[4] The Debian Project. Debian: the Debian GNU/Linux operating system [EB/OL]. [cited 2006-04-09]. http://www.debian.org.

[5] The Apache Software Foundation. Apache: the Apache HTTP Server [EB/OL]. [cited 2005-11-09]. http://www.apache.org.

[6] MySQL Inc. MYSQL: The open source SQL Database Server [EB/OL]. [cited 2005-04-13]. http://www.mysql.com.

[7] National Center for Biotechnology Information. FASTA format description [ED/OL]. [cited 2005-12-04]. http://www.ncbi.nlm.nih.gov/blast/fasta.shtml.

[8] Washington University. WU-BLAST: Washington University BLAST Server [ED/OL]. [cited 2005-03-22]. http://blast.wustl.edu.

[9] Bioperl Workgroup. Bioperl project [ED/OL]. [cited 2006-03-21]. http://www.bioperl.org.

[10] National Biomedical Research Foundation, the Georgetown University Medical Center. ACHA BRARE: Acetylcholine receptor protein subunit alpha precursor [EB/OL]. [cited 2006-07-04]. http://www.ebi.uniprot.org/entry/ACHA_BRARE.

Edited by LUO Min