Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W673–W676***
doi:10.1093/nar/gki480

GBA server: EST-based digital gene expression profiling

Xin Wu[1], Michael G. Walker[2,3], Jingchu Luo[1] and Liping Wei[1,2,*]

[1]Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University, Beijing 100871, P. R. China, [2]Biomedical Informatics, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA and [3]Walker Bioscience, Sunnyvale, CA, USA

Received February 14, 2005; Revised and Accepted April 15, 2005

ABSTRACT

Expressed Sequence Tag-based gene expression profiling can be used to discover functionally associated genes on a large scale. Currently available web servers and tools focus on finding differentially expressed genes in different samples or tissues rather than finding co-expressed genes. To fill this gap, we have developed a web server that implements the GBA (Guilt-by-Association) co-expression algorithm, which has been successfully used in finding disease-related genes. We have also annotated UniGene clusters with links to several important databases such as GO, KEGG, OMIM, Gene, IPI and HomoloGene. The GBA server can be accessed and downloaded at http://gba.cbi.pku.edu.cn.

INTRODUCTION

The sequencing and analysis of Expressed Sequence Tags (ESTs) is one of the three most important techniques used to study gene expression, the other two being DNA microarray and SAGE. Vast amounts of EST data are now available, and the volume is growing rapidly. Currently there are over 25 million EST sequences in NCBI's dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/), among which six million were added in 2004 alone. Because the same gene may be represented by many different EST sequences, the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene) (1) was developed to partition nucleotide sequences into a non-redundant set of gene-oriented clusters.

Although it can be difficult to use data from more than one microarray experiment because the platforms are often different, it is straightforward to use EST data across different libraries, tissues and disease stages. Analysis of these EST data, often with the help of the UniGene database, holds tremendous value for the study of gene expression. For instance, Baranova *et al*. (2) developed software named HsAnalyst that identified novel tumor markers and potential targets for anti-tumor therapy using data in dbEST and UniGene. Stanton *et al*. (3) built differential gene expression profiles from dbEST and UniGene to identify tissue-enriched genes in mouse pancreas, mammary gland and heart. Ewing *et al*. (4) built digital gene expression profiles of the rice genome using dbEST and analyzed these profiles using Pearson's correlation coefficient; they found two co-expressed clusters of contigs encoding proteins with seed-related functions. Walker (5) developed the Guilt-by-Association (GBA) algorithm, which uses the Fisher exact test to find genes that have similar expression patterns to that of a query gene across EST libraries; using known disease-related genes as a query (bait), they successfully identified new genes that are associated with schizophrenia, Parkinson's disease and prostate cancer (5,6). Thompson *et al*. (7) applied the GBA algorithm to identify groups of genes involved in common cellular processes, which they named 'functional modules', in pregnancy, breast cancer and ovarian cancer, and validated the results using real-time PCR.

Despite the proven importance of EST-based expression profiling, there is a shortage of web servers that conduct such analysis. The deficiency is especially clear when contrasted with the large number of available web servers for microarray analysis. Furthermore, the EST servers that do exist, including Digital Differential Display (DDD, http://www.ncbi.nlm.nih.gov/UniGene/info_ddd.html), cDNA Digital Gene Expression Displayer (DGED, http://cgap.nci.nih.gov/Tissues/GXS), xProfiler (http://cgap.nci.nih.gov/Tissues/xProfiler) and GEPIS (8), all aim to find differentially expressed genes in different pools of tissues or samples. For instance Scheurle *et al*. (9) used DDD to find up-regulated and down-regulated genes in solid tumors. Currently there is no web server that can find co-expressed genes based on ESTs in cDNA libraries and UniGene. In particular, the GBA algorithm is not publicly available as a web server or standalone software.

Given the value of finding co-expressed genes, we have developed a new web server, named GBA, that has two types

*To whom correspondence should be addressed. Tel: +86 10 6276 4970; Fax: +86 10 6275 2438; Email: weilp@mail.cbi.pku.edu.cn

of functions. First, given a gene sequence, it uses the GBA algorithm to find other genes (clustered by UniGene) that have statistically similar expression pattern across all EST libraries. Thus a user can input a novel sequence and find the most closely co-expressed genes that can offer valuable information on the input's function. Or the user can input a gene known to be involved in a disease and find new genes co-expressed with it that may also be involved in the same disease. Second, we have linked UniGene clusters to a variety of other databases, including Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene), IPI (http://www.ebi.ac.uk/IPI/IPIhelp.html) (10), HomoloGene (http://www.ncbi.nih.gov/HomoloGene), OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) (11), GOA (http://www.ebi.ac.uk/GOA/index.html) (12) and KEGG (http://www.genome.jp/kegg/) (13). The server allows a user to input a UniGene ID and retrieve extensive information about the gene such as functional categories and known pathways. Alternatively, it also allows a user to input an ID from GO, KEGG, etc. and retrieve all UniGene clusters involved in the particular GO function or KEGG pathway.

DESCRIPTION OF FUNCTIONS

The GBA server supports four functions: GBA Engine and Gene Matcher, which implement the GBA algorithm, and UniGene Annotation and Annotation Linker, which link UniGene clusters to several important molecular databases.

GBA Engine

We have downloaded and parsed all cDNA libraries from EMBL (ftp://ftp.ebi.ac.uk/pub/databases/embl/release/) (14) and CGAP (http://cgap.nci.nih.gov/Info/CGAPDownload/) (15) and stored the data in a relational database. Because libraries with too few entries are under-sequenced and do not adequately reflect the true expression levels of genes, GBA Engine allows the user to specify the minimum number of cDNA sequences in a library in order for it to be included in the analysis. Table 1 shows how using a different minimum number of cDNA sequences affects the number of libraries included in the analysis. Based on previous experience, we recommend using a cutoff of 500 or 1000.

A gene is considered present (expressed) in a library if at least one cDNA sequence corresponding to the gene is found in the library. The presence/absence of a gene in all libraries forms a vector that represents its expression profile. For a pair of genes, *A* and *B*, GBA Engine converts their expression profiles into a 2×2 contingency table, showing the numbers of libraries where both *A* and *B* are present, where *A* is present but *B* is absent, where *B* is present but *A* is absent, and where both *A* and *B* are absent. GBA Engine then applies the Fisher exact test to test the null hypothesis that there is no association between *A* and *B* and determines a *P*-value for statistical significance. Because multiple statistical tests are performed, GBA Engine can optionally apply a Bonferroni correction to the *P*-value to reduce the false positive rate. Given a gene (represented by its UniGene ID), GBA Engine returns a list of other UniGene clusters that have similar expression pattern across cDNA libraries, ranked by their *P*-value.

Table 1. Number of human and mouse libraries in the CGAP database using different minimum numbers of cDNA sequences

No. of cDNAs in library	No. of human libraries	No. of mouse libraries
>5000	276	253
>2000	488	401
>1000	713	516
>500	1117	584
>100	3774	651
Total	8570	965

Gene Matcher

If a user has an anonymous gene sequence, the first requirement is to find its corresponding UniGene cluster before running GBA Engine or UniGene Annotation. Gene Matcher provides an interface that runs BLAST (16) to query the gene sequence against all UniGene sequences to identify the right UniGene cluster.

UniGene Annotation and Annotation Linker

To help users better understand the functions of UniGene clusters, we have integrated UniGene clusters with several molecular databases including Gene, IPI, HomoloGene, OMIM, GO and KEGG, which we downloaded, parsed and stored in a relational database. Given a UniGene ID, UniGene Annotation returns detailed information on the gene locus, orthologs, disease association, GO categories and pathways. Alternatively, given an ID from the other databases, Annotation Linker returns all UniGene clusters that are associated with the locus, ortholog, GO category and pathway, respectively.

IMPLEMENTATION

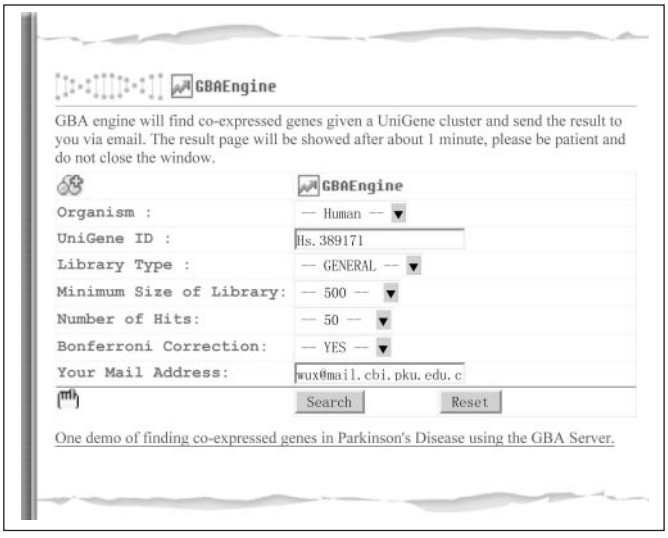
The GBA server is a three-tier application developed in Java. In the Database Tier, we developed the GBA tool Java package to parse the molecular databases integrated in the GBA server and used the POSTGRES database system to store the data. In the Logic Tier, we developed Java programs for data processing and statistical tests. In the Web Tier, we used Apache Tomcat (http://jakarta.apache.org/tomcat/index.html) for the web server, Struts (http://struts.apache.org/) as the MVC (Model–View–Control) framework and Java Mail (http://java.sun.com/products/javamail/community/links/index.html) to handle emailing of the results.

We update our database whenever the NCBI UniGene database is updated. We track the external databases such as GO and OMIM regularly and update the links in our database when these databases are updated. Complete updates of our database are scheduled monthly. In addition, users can download and install the GBA server locally and update it according to their own needs. (See GBA Administrator Guide on the website for help.)

EXAMPLE APPLICATION

To demonstrate the use of the GBA server, we used it to identify genes that may be associated with Parkinson's disease, one of the most common neurodegenerative disorder

(a)



(b)

GBA: Digital EST Expression Profiling
WebSite: http://gba.cbi.pku.edu.cn

INPUT

Organism: human
UniGene ID: Hs.389171 PTEN induced putative kinase 1
Library Type: general
Minimum ESTs Size: 500
Number of Hits: 50
Bonferroni Correction: true

OUTPUT

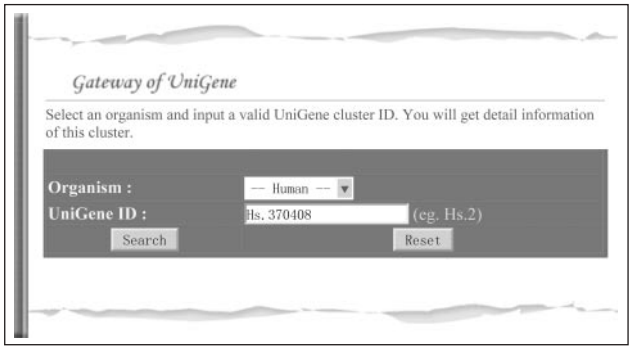
PP (Number of libs that both target gene and output gene are present)
PA (Number of libs that target gene is present but output gene is absent)
AP (Number of libs that target gene is absent but output gene is present)
AA (Number of libs that both target gene and output gene are absent)

Index	Similar UniGenes Found	PP	PA	AP	AA	P-value	Description
1	Hs.517949	129	60	188	736	1.88 E-27	Microtubule-associated protein 4
2	Hs.474982	131	58	196	728	2.16 E-27	Aconitase 2, mitochondrial
3	Hs.370408	108	81	123	801	4.59 E-26	Catechol-O-methyltransferase
4	Hs.124126	111	78	136	788	2.99 E-25	Actin related protein 2/3 complex, subunit 1A, 41kDa
5	Hs.513145	111	78	137	787	4.93 E-25	Mesenchymal stem cell protein DSC92
6	Hs.374503	126	63	192	732	1.48 E-24	Mortality factor 4 like 1
7	Hs.255935	122	67	179	745	3.42 E-24	B-cell translocation gene 1, anti-proliferative
8	Hs.485262	121	68	178	746	1.08 E-23	Mitochondrial carrier homolog 1 (C. elegans)
9	Hs.183706	118	71	168	756	1.53 E-23	Adducin 1 (alpha)
10	Hs.474833	99	90	108	816	2.51 E-23	Casein kinase 1, epsilon
11	Hs.520943	120	69	178	746	5.19 E-23	Williams-Beuren syndrome chromosome region 1
12	Hs.473761	130	59	216	708	5.19 E-23	Reticulon 3
13	Hs.501309	126	63	200	724	6.16 E-23	Cold inducible RNA binding protein
14	Hs.42806	106	83	132	792	1.09 E-22	Calcium binding protein Cab45 precursor
15	Hs.301419	97	92	106	818	1.92 E-22	Telomeric repeat binding factor 2, interacting protein
16	Hs.518827	138	51	253	671	1.98 E-22	Cyclin I
17	Hs.529892	152	37	321	603	2.01 E-22	Sequestosome 1
18	Hs.525232	102	87	121	803	2.19 E-22	Low density lipoprotein receptor-related protein 10
19	Hs.246413	125	64	200	724	2.21 E-22	RNA binding motif protein 12
20	Hs.237536	81	108	65	859	2.70 E-22	Hypothetical protein MGC20781

Figure 1. Input and output of GBA Engine using *PINK1* as an example. (**a**) Example input of GBA Engine. (**b**) Example output of GBA Engine showing genes having similar expression patterns to *PINK1*.

diseases, given genes already known to be associated with Parkinson's disease, such as PTEN induced putative kinase 1 (*PINK1*). It has been suggested that a mutant form of *PINK1* damages neurons by stress-induced apoptosis and mitochondrial dysfunction (17). Using *PINK1* (UniGene ID Hs.389171) as bait, we searched the cDNA libraries with >500 sequences for the top 50 genes that have similar expression patterns, using GBA Engine with a Bonferroni correction (Figure 1a). Part of the result is shown in Figure 1b. In particular, we found that hit #3 (UniGene ID Hs.370408), in turn, has a large number of the same co-expressed genes (measured

(a)



(b)

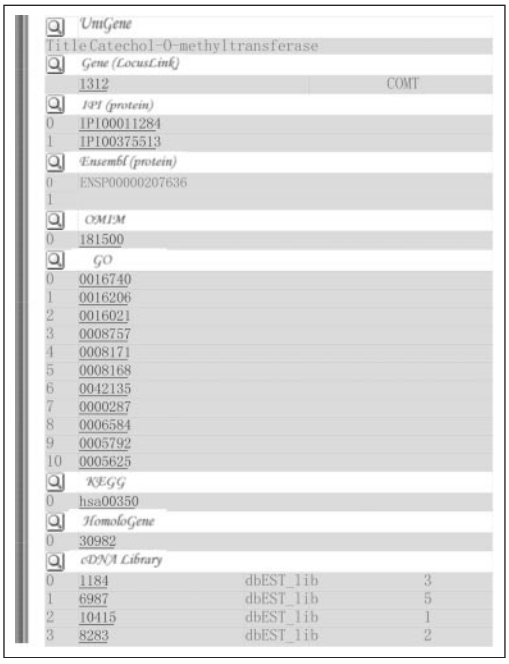
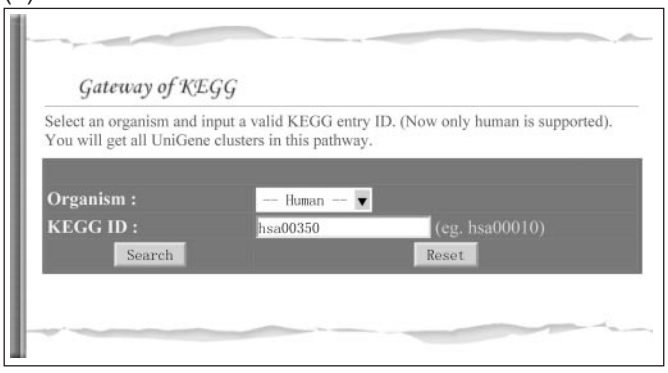


Figure 2. Input and output of UniGene Annotation using Hs.370408 as an example. (**a**) Example input of UniGene Annotation. (**b**) Example output of UniGene Annotation showing detailed information on the gene.

by GBA) as several genes known to be associated with Parkinson's disease, such as alpha-synuclein (*SNCA*) (data not shown) (18). We looked for more information on this gene using UniGene Annotation (Figure 2). The linked entry in the Gene (LocusLink) database shows that the gene is *Catechol-O-methyltransferase* (*COMT*), which encodes a protein that 'catalyzes the transfer of a methyl group from S-adenosylmethionine to catecholamines, including the neurotransmitters dopamine, epinephrine, and norepinephrine' (19). The linked entry in the OMIM database shows that '*COMT* is important in the metabolism of catechol drugs used in the treatment of hypertension, asthma, and Parkinson disease.' Figure 2 shows that *COMT* is involved in KEGG pathway hsa00350 (Tyrosine Metabolism). Figure 3 shows the input and output using the KEGG Gateway of Annotation Linker to find all UniGene clusters involved in the pathway. This example demonstrates the power of using GBA Engine to find additional genes that are involved in a disease process and can serve as potential drug targets. It also demonstrates the value of using UniGene Annotation and Annotation Linker for further in-depth analysis.

(a)



(b)

Figure 3. Input and output of Annotation Linker using pathway hsa00350 as an example. (**a**) Example input of the KEGG Gateway of Annotation Linker. (**b**) Example output of the KEGG Gateway of Annotation Linker showing all UniGene clusters involved in the pathway.

DISCUSSION

The GBA server is the first to make the GBA algorithm publicly available. It also integrates UniGene clusters with a variety of molecular databases for the first time and provides a public, web-based user interface. With the rapid accumulation of available EST data, and because data in new cDNA libraries from new experiments can be easily used together with existing ones, the effectiveness of the GBA server will continue to increase. Currently the GBA server supports three species that have the most available EST data: human, mouse, and rat. We will add more species in the future that have sufficient EST data available.

ACKNOWLEDGEMENTS

Funding for this work, including payment for the Open Access publication charges for this article, was provided by the China National High-tech 863 Program (2004AA231020, 2004BA711A21).

Conflict of interest statement. None declared.

REFERENCES

1. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
2. Baranova,A.V., Lobashev,A.V., Ivanov,D.V., Krukovskaya,L.L., Yankovsky,N.K. and Kozlov,A.P. (2001) *In silico* screening for tumour-specific expressed sequences in human genome. *FEBS Lett.*, **508**, 143–148.
3. Stanton,J.A., Macgregor,A.B. and Green,D.P. (2003) Identifying tissue-enriched gene expression in mouse tissue using the NIH UniGene database. *Appl. Bioinformatics*, **2** (Suppl. 3), S65–S73.
4. Ewing,R.M., Ben,K.A., Poirot,O., Lopez,F., Audic,S. and Claverie,J.M. (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.*, **9**, 950–959.
5. Walker,M.G. (1999) Pharmaceutical target discovery using Guilt-by-Association: schizophrenia and Parkinson's disease genes. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1999**, 282–286.
6. Walker,M.G., Volkmuth,W., Sprinzak,E., Hodgson,D. and Klingler,T. (1999) Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Res.*, **9**, 1198–1203.
7. Thompson,H.G., Harris,J.W., Wold,B.J., Quake,S.R. and Brody,J.P. (2002) Identification and confirmation of a module of coexpressed genes. *Genome Res.*, **12**, 1517–1522.
8. Zhang,Y., Eberhard,D.A., Frantz,G.D., Dowd,P., Wu,T.D., Zhou,Y., Watanabe,C., Luoh,S.M., Polakis,P., Hillan,K.J., Wood,W.I. and Zhang,Z. (2004) GEPIS—quantitative gene expression profiling in normal and cancer tissue. *Bioinformatics*, **20**, 2390–2398.
9. Scheurle,D., DeYong,M.P., Binninger,D.M., Page,H., Jahanzeb,M. and Narayanan,R. (2000) Cancer gene discovery using digital differential display. *Cancer Res.*, **60**, 4037–4043.
10. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
11. McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD. (2000) Online Mendelian Inheritance in Man, OMIM (TM), http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.
12. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
13. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
14. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M. and Cochrane,G. (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **33**, D29–D33.
15. Lal,A., Lash,A.E., Altschul,S.F., Velculescu,V., Zhang,L., McLendon,R.E., Marra,M.A., Prange,C., Morin,P.J. and Polyak,K. (1999) A public database for gene expression in human cancers. *Cancer Res.*, **59**, 5403–5407.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
17. Hatano,Y., Li,Y., Sato,K., Asakawa,S., Yamamura,Y., Tomiyama,H., Yoshino,H., Asahina,M., Kobayashi,S., Hassin-Baer,S. *et al.* (2004) Novel PINK1 mutations in early-onset parkinsonism. *Ann. Neurol.*, **56**, 424–427.
18. Huang,Y., Cheung,L., Rowe,D. and Halliday,G. (2004) Genetic contributions to Parkinson's disease. *Brain Res. Rev.*, **46**, 44–70.
19. Tai,C.H. and Wu,R.M. (2002) Catechol-*O*-methyltransferase and Parkinson's disease. *Acta Med. Okayama.*, **56**, 1–6.