BIOINFORMATICS APPLICATIONS NOTE
Vol. 19 no. 5 2003, pages 653–654
DOI: 10.1093/bioinformatics/btg044

DigiNorthern, digital expression analysis of query genes based on ESTs

Jianxin Wang and Ping Liang*

Department of Cancer Genetics and the Biomathematics & Biostatistics Core
Facility, Roswell Park Cancer Institute, Elm & Carlton Streets, Buffalo, NY 14263, USA

Received on September 3, 2002; revised on November 11, 2002; accepted on November 22, 2002

ABSTRACT

Summary: DigiNorthern (DN) is a web-based tool for virtually displaying expression profiles of query genes based on EST sequences. Two utilities are available: DN1 takes one query gene and quantitatively display its expression levels in tissues/organs that express the gene with comparison between normal and neoplastic status of each tissue; DN2 takes two sequences as query genes and compares their expression profiles side by side.

Availability: DN is currently available for analyses of human and mouse genes and is accessible at http://falcon.roswellpark.org/DN. Options for other species may become available in future depending on user's request and EST data availability.

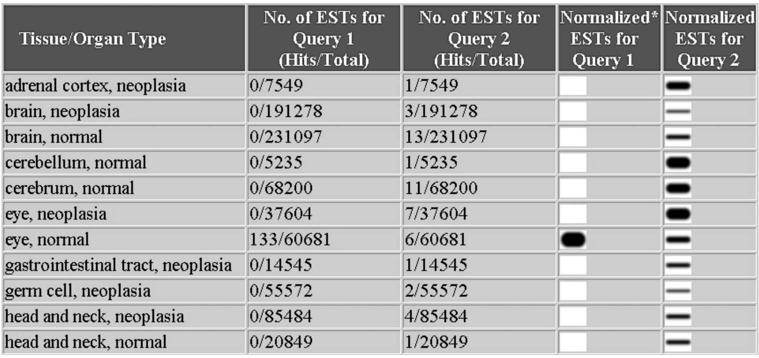
Contact: ping.liang@roswellpark.org

Analysis of gene expression refers to the detection and quantification of a gene transcript in different tissues/cells including those under different developmental, physiological, and pathological conditions. The availability of comprehensive data generated by high-throughput functional genomics approaches, mainly expressed sequence tag (EST) and serial analysis of gene expression (SAGE), provides the feasibility to study gene expression through in silico analysis (Boguski *et al.*, 1993; Lash *et al.*, 2000). We have so far accumulated over 4 million EST sequences for human and 2 million for mouse from diverse organ-, tissue-, and disease-derived cDNA libaries, mostly generated through the Cancer Anatomy Genome Project (CGAP) (http://cgap.nci.nih.gov) (Boon, 2002). Several data mining tools have been developed by NCBI to facilitate the use of these EST and SAGE data (http://www.ncbi.nlm.nih.gov/ncigap) (Scheurle *et al.*, 2000; Strausberg *et al.*, 2001). These tools include Digital Differential Display (DDD), cDNA xProfiler, and cDNA Digital Gene Expression Displayer (DGED) for EST data and SAGEmap tools for SAGE data, allowing users to screen genes that are differentially expressed among selected lists of cDNA libraries/tissues. In addition, digital expression data is available for each UniGene through a pre-computed data set based on SAGE and/or ESTs. We here report the availability of DigiNorthern that digitally displays the expression data for user given query genes based on dynamically collected EST data.

The strategy is to first collect all ESTs representing user's query gene by a sequence similarity search using a locally installed BLAST program (Altschul *et al.*, 1990) against constantly updated EST databases. The BLAST output is then filtered using optimized parameters to make sure that all matched ESTs are true representation of the query gene while keeping all true matches. Then the identifiers (GI) of these matched ESTs are retrieved to identify the cDNA library generating each EST based on its GenBank information. The tissue/organ name and pathological status (normal vs. different cancer stages) for each EST are then obtained by searching a cDNA library lookup table. This lookup table links each cDNA library with indexed standard tissue keywords created by combining information from several sources including the data files describing all CGAP EST libraries from the NCI website (http://cgap.nci.nih.gov/Info/CGAPDownload). The number of matched ESTs for the query gene in each type of tissue is then counted. To quantitatively represent the relative expression level of the query gene in each tissue, the ratio between the number of matched ESTs vs. the total number of ESTs available for the tissue, which is obtained from another lookup table, is calculated. The two lookup-tables used in this process are constantly updated to cover any new cDNA libraries for ESTs deposited into NCBI dbEST. To enhance the visualization of the results, virtual 'northern gel' pictures with densities ranging in 26 scales from level 0, indicating no expression, to level 25, the highest level of expression, are also used to accordingly represent the values of matched EST/total ESTs (Figure 1).

DN offers two utilities: DN1 for a single query sequence and DN2 for two genes side-by-side comparison. User inputs include DNA sequence(s) for the query gene(s), choice of genome, and the optional changes of parameters for BLAST search, while the output is a table in html

*To whom correspondence should be addressed.



Tissue/Organ Type	No. of ESTs for Query 1 (Hits/Total)	No. of ESTs for Query 2 (Hits/Total)	Normalized* ESTs for Query 1	Normalized ESTs for Query 2
adrenal cortex, neoplasia	0/7549	1/7549	□	▬
brain, neoplasia	0/191278	3/191278	□	▬
brain, normal	0/231097	13/231097	□	▬
cerebellum, normal	0/5235	1/5235	□	▬
cerebrum, normal	0/68200	11/68200	□	▬
eye, neoplasia	0/37604	7/37604	□	▬
eye, normal	133/60681	6/60681	●	▬
gastrointestinal tract, neoplasia	0/14545	1/14545	□	▬
germ cell, neoplasia	0/55572	2/55572	□	▬
head and neck, neoplasia	0/85484	4/85484	□	▬
head and neck, normal	0/20849	1/20849	□	▬

Fig. 1. A screen snapshot of DN2 result page showing the expression profiles of two query genes, human alpha-A-crystallin (query 1) and human alpha-tubulin (query2).

format with graphic visualization as shown in Figure 1. A link to the actual BLAST output is provided to help users to examine the actual sequence alignments to determine if it is necessary to adjust the BLAST search parameters.

The accuracy of DN was examined using genes with known expression profiles. As shown in Figure 1, DN accurately detected the highly specific expression of human alpha-A-crystallin gene in eye and the universal expression of alpha-tubulin gene. DN2 can be particularly useful to study the different expression patterns between paralogous genes. Compared to Virtual Northern data from the NCI web site, which is based on pre-computed UniGene clusters, that may not be updated timely, DN always uses the most recent EST data and takes any query gene, including newly predicted genes that have not been collected into UniGene database.

In conclusion, DN is a useful tool complementary to the tools available at NCI CGAP for digital analysis of expression profile for given genes allowing quick and inexpensive identification of unusual or distinct gene expression patterns possessed by certain genes, especially for those related to cancers. It can provide very useful preliminary results for guiding the design of further experimental analysis. Due to the fact that EST data is still incomplete in many aspects, results generated by DN may not be very accurate in some cases, especially for genes expressed in very low levels or when the available cDNA libraries are small. Therefore, it is strongly recommended that, whenever possible, users should verify the results by experimental methods.

Acknowledgements

This work is partially supported by shared resources of the Roswell Park Cancer Center support grant P30 CA16056 from NCI and by Roswell Park Alliance Foundation to PL.

REFERENCES

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.

Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST-database for ‘expressed sequence tags’. *Nat. Genet.*, **4**, 332–333.

Boon,K., Osorio,E.C., Greenhut,S.F., Schaefer,C.F., Shoemaker,J., Polyak,K., Morin,P.J., Buetow,K.H., Strausberg,R.L., De Souza,S.J. and Riggins,G.J. (2002) An anatomy of normal and malignant gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 11287–11292.

Scheurle,D., DeYoung,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.

Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.

Strausberg,R.L,, Greenhut,S.F., Grouse,L.H., Schaefer,C.F. and Buetow,K.H. (2001) In silico analysis of cancer through the Cancer Genome Anatomy Project. *Trends Cell Biol.*, **11**, S66–71.