BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no. 12 2003, pages 1594–1595
DOI: 10.1093/bioinformatics/btg198

Digital extractor: analysis of digital differential display output

Stephen F. Madden[1], Barry O'Donovan[2], Simon J. Furney[1], Hugh R. Brady[1], Guenole Silvestre[2] and Peter P. Doran[1,*]

[1]Human Genomics and Bioinformatics Research Unit, Department of Medicine and Therapeutics, University College Dublin, Mater Misericordiae Hospital, The Dublin Molecular Medicine Centre, 41 Eccles St, Dublin 7, Ireland and [2]Department of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland

Received on June 24, 2002; accepted on March 3, 2003

ABSTRACT

Summary: Digital Extractor is a program for the high-throughput processing of data sets derived from digital differential display-based comparisons of EST libraries. These comparisons can be utilized to identify discrete subsets of genes whose expression is restricted to distinct tissue types. The program facilitates these investigations by permitting parallel annotation of genes identified as being differentially expressed.

Availability: The executable program, suitable for use on all UNIX-based platforms is freely available to non-profit users

Contact: pdoran.genome@mater.ie

Digital Differential Display (DDD) is an Internet based resource for the identification of genes whose expression is altered between different tissue types (http://www.ncbi.nlm.nih.gov/UniGene). This resource exploits the large number of publicly available cDNA libraries corresponding to different tissues, cancers, etc. This online system permits: (a) selection of cDNA libraries to be compared (e.g. cancer versus normal tissue); (b) comparison of the constituent sequences; and (c) output of a list of differentially expressed sequences. This resource offers exciting new avenues for exploration in the search for novel genes in health and disease; indeed it has recently been applied to the identification of cancer-associated transcripts (Scheurle *et al.*, 2000). Whilst DDD represents an important tool for the biomedical research community, its main limitation rests in the cumbersome nature of the subsequent data analysis.

Here we present a new program Digital Extractor, for the processing of data obtained from DDD-based investigations of differential gene expression. This program refines the DDD approach by permitting rapid, automated annotation of output gene lists.

*To whom correspondence should be addressed.

Extractor is written in PERL and can be implemented on all UNIX platforms with PERL version 5.0 or greater. The application can be executed using either a Java application or a command line interface. Digital Extractor integrates and utilizes a number of tools including: (a) CAP3 (Huang and Madan, 1999), for assembly of EST clusters into contigs; (b) RepeatMasker (Smit, 1996), for masking of repetitive elements within the assembled EST contigs; and (c) BLAST (Altschul *et al.*, 1990), for homology searching.

The results file produced from a DDD experiment is an HTML page detailing all of the ESTs, whose expression differs between the conditions of interest with a link provided to each UNIGENE cluster. Digital Extractor uses the DDD output HTML page as input. The page is loaded into the application and scanned to extract the accession numbers of the UNIGENE clusters, representing differentially expressed genes. These accession numbers are then used to complete automated extraction of the UNIGENE clusters from the locally-stored databases. The user can specify application parameters such as the database to be searched, and the *e*-values for the BLAST searches.

Each UNIGENE cluster extracted from the database contains all of the cDNA sequences that correspond to an individual gene. The number of sequences in each cluster can range from a few dozen to many thousand. To date no attempt has been made to produce contigs from the representative sequences in these clusters due to a number of reasons including the presence in the UNIGENE cluster of all the splicing variants of the gene of interest and the inclusion of 5′ and 3′ reads from the same gene. However, if the information produced from the DDD is to be of use in the identification of differentially expressed genes, in particular the annotation of unknown transcripts, it is necessary to produce contiguous sequences, for database searching, by cluster assembly. This step is crucial to reduce the number of BLAST runs per experiment thus

improving throughput. Having obtained the complete UNIGENE clusters representing each hit in the DDD experiment, Digital Extractor integrates and utilizes the CAP3 EST assembly program to produce contiguous sequences. The result of this procedure is the production of a sequence for BLAST analysis. This step obviates the need for BLAST searches of all the individual sequences within each UNIGENE cluster, thus dramatically increasing the efficiency of the application.

To facilitate the rapid, parallel identification of the gene corresponding to each assembled contig, by means of the BLAST algorithm, it is necessary to optimize the sequence inputs. To achieve this, the assembled contigs are masked for repeat elements and low complexity DNA using the Repeat Masker Application. This step substantially improves the performance of the BLAST-based sequence identification. Having assembled and masked each of the contiguous sequences corresponding to genes whose expression is altered between the conditions of interest, BLAST is used to search for homologous sequences in the non-redundant nucleotide database. The output of Digital Extractor is a results page with the identity of all annotated sequences, with links to the NCBI database for further information.

To determine the speed and accuracy of Digital Extractor, a test analysis was performed on a data set produced from a DDD-based comparison of cDNA libraries from human kidney and human pancreas. The output from this comparison was a web page containing 171 links to UNIGENE clusters. The web page was downloaded and submitted to the program via the Java GUI. The option to extract only un-annotated clusters was chosen, which focused the analysis on a subset of 15 UNIGENE clusters. It was also decided to compare the clusters to the nr database with an *e*-value of 0.001 (choosing a smaller database, and a lower *e*-value would obviously decrease run time of this aspect of the program). It takes approximately 6 minutes, to extract, assemble, mask, and blast each cluster, on a 1 GHz processor with 216 MB of memory. The major limiting factor being the assembly process. A detailed analysis of the data set can be viewed on our web-site, http://www.hgbru.org/worked_example.htm.

In summary Digital Extractor represents an efficient, user-friendly platform for the rapid annotation of data derived from DDD-based experiments to analyze differential gene expression.

ACKNOWLEDGEMENTS

We acknowledge the National Centre for Biotechnology Information for provision and curation of sequence databases, used herein. These studies were supported by the Irish Government's Programme for Research in Third Level Institutions, the European Union Fifth Framework Programme and the Punchestown Kidney Research Fund.

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.

Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.

Scheurle,D., DeYoung,M.P., Binninger,D.M., Page,H., Jahanzeb,M. and Naraynan,R. (2000) Cancer gene discovery using digital differential display. *Cancer Res.*, **60**, 4037–4043.

Smit,A.F.A. (1996) The origin of interspersed repeats in the human genome. *Curr. Op. Gen. Dev.*, **6**, 743–748.