BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 11 2000
Pages 1038–1039

DRAGON: Database Referencing of Array Genes Online

Christopher M.L.S. Bouton [1, 2] and Jonathan Pevsner [1, 2,*]

[1]Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA and [2]Department of Neurology, The Kennedy Krieger Institute, Baltimore, MD 21205, USA

Received on April 27, 2000; revised on August 11, 2000; accepted on August 15,2000

Abstract

Summary: *'Database Referencing of Array Genes ONline' or 'DRAGON' is a web-accessible database that aids in the analysis of differential gene expression data as a biological annotation tool. Users of DRAGON can submit data sets containing large lists of genes and then choose particular characteristics that DRAGON supplies for all genes on the list rapidly and simultaneously.*

Availability: *The DRAGON database is available for queries on the DRAGON web site www.kennedykrieger.org/pevsnerlab/dragon.htm.*

Contact: *pevsner@kennedykrieger.org or cbouton@jhmi.edu*

Introduction

Recent technologies such as cDNA microarrays and serial analysis of gene expression (SAGE) make possible the analysis of the mRNA expression of thousands of genes simultaneously. Results of microarray experiments often consist of lists of expression values for hundreds to thousands of genes that are differentially regulated between two samples. When confronted with these lists one issue is how to define the biological characteristics of individual genes. A second issue is how to identify biological relationships between the genes on the list. What genes perform similar functions? What genes participate in the same cellular pathways? Answers to these questions can provide important clues to the types of biological processes that may be associated with the experimental system being investigated. However, determining the answers to such questions is usually an arduous process that is left to the investigator's knowledge of the system in question, literature searches and the tedious process of researching the biological characteristics of individual genes (and their encoded proteins) in public databases via the World Wide Web (WWW).

We have developed a WWW-accessible database called 'Database Referencing of Array Genes ONline' or DRAGON. DRAGON takes advantage of the fact that a wide range of information pertaining to the biological characteristics of large numbers of genes is available in multiple public databases. DRAGON extracts information that is directly relevant to the analysis of differential gene expression data from these databases and makes the information available in a single location. Users of DRAGON can use the 'Annotate' page on the DRAGON web site to submit a large list of genes that includes a Genbank accession number for each gene as well as expression values for each gene. They can then request that the list be annotated with specific types of information that might be interesting in light of the experimental system being studied. Instead of having to research the characteristics of genes one at a time, DRAGON annotates the entire list simultaneously.

Some types of information provided by DRAGON are specific to individual genes and their encoded proteins on a gene list. For example, the SWISS-PROT database provides descriptions of the functions of individual proteins. Other types of information can be used to categorize genes into groups based upon shared biological characteristics of the genes or their encoded proteins. For example, accession numbers from the Pfam database (http://www.sanger.ac.uk/Software/Pfam/) classify hundreds to thousands of proteins into families that share homologous domains. Additionally, the SWISS-PROT database (http://www.expasy.ch/sprot/) defines known functions of proteins using a controlled vocabulary of keywords (ftp://expasy.proteome.org.au/databases/swiss-prot/release/keywlist.txt).

The information provided by DRAGON can be integrated with other expression data analysis in order to gain a better understanding of how biological characteristics are related to gene expression patterns. For example, microarray data sets are typically analyzed with a variety of clustering techniques (e.g. K-means or hierarchical clustering). With these methods expression data is used to identify groups of genes that are co-regulated over time or across samples. One inference that is often made

*To whom correspondence should be addressed.

about co-regulated groups of genes is that they may be functionally related. Previous studies have identified functional relationships such as shared promoter elements, transcription factors, chromosomal loci or cellular functions of encoded proteins between co-regulated genes (Eisen *et al.*, 1998; Heyer *et al.*, 1999; Spellman *et al.*, 1998). By first using DRAGON to annotate a data set and then clustering the data by expression values, investigators can rapidly determine whether subsets of co-regulated genes in their experimental system are related by specific biological characteristics. For example, if a group of co-regulated genes was also found to share a specific Pfam accession number, that could indicate that their encoded proteins shared common functional properties.

DRAGON is distinct from some other types of microarray-related web sites. Databases such as ArrayExpress (http://www.ebi.ac.uk/arrayexpress/), the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) and ExpressDB (http://arep.med.harvard.edu/ExpressDB/) are repositories for microarray data, and ArrayDB (http://genome.nhgri.nih.gov/arraydb/) is an example of a microarray data management system. DRAGON does not store microarray data, instead DRAGON is a WWW-accessible tool for the annotation of microarray data with a biological information.

System, methods and implementation

The files contained in the DRAGON database can be downloaded (or ordered on CD-ROM) from the DRAGON web site as tab-delimited text files. A Dell PowerEdge 6300, dual Xeon 550 MHz processor-based server running Red Hat Linux 6.2 serves the DRAGON web site with Apache (http://www.apache.org). MySQL (http://www.mysql.com) is used as the relational database management system. DRAGON is composed of multiple tables derived from flat-files (indicated in parentheses following the names of each database) provided by the UniGene (ftp://ftp.ncbi.nlm.nih.gov/repository/UniGene/Hs.data.Z, Rn.data.Z, Mm.data.Z and Dr.data.Z), SWISS-PROT (ftp://expasy.proteome.org.au/databases/swiss-prot/release/sprot39.dat) and Pfam (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/PfamA.full.gz) databases. Further additions to DRAGON will include information derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.ad.jp/kegg/), Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim/), Transfac (http://www.cbi.pku.edu.cn/TRANSFAC/), Interpro (http://www.ebi.ac.uk/interpro/), the Biological and Biochemical Image Database (BBID; http://bbid.grc.nia.nih.gov/) and multiple yeast databases.

In order to create the DRAGON database tables, a cron process is used to automatically run a set of Perl scripts (http://www.perl.com) which use the Net::FTP module to check for updated versions of each database flat-file at the respective web sites on a daily basis. If a database file has been updated, the Perl script downloads the new file, extracts specific information from it, and saves the information into a set of output files. Continual updating of DRAGON minimizes errors in the cross-referencing of the databases that can occur due to inconsistencies such as retired accession numbers. The output files are imported into a MySQL database using the DBI module. The MySQL database is queried via common gateway interface (.cgi) scripts also written in Perl. As the user selects the types of information they wish to annotate their data set with, an SQL statement is constructed which selects and joins the appropriate tables in the database. When the query is submitted, the constructed SQL statement is passed to the MySQL database, processed and returned in one of three possible output formats.

Conclusion

The DRAGON database associates biologically relevant information derived from numerous public databases with data generated by large-scale gene expression experiments. An eventual goal in the development of DRAGON is the comprehensive definition of biological data regarding each gene on a list through the interconnection of as many public databases as possible. Future work is also focused on developing novel methods of integrating the information provided by DRAGON with standard types of microarray data analysis.

Acknowledgements

We thank Mr George W. Henry for his significant contributions to the design and production of the DRAGON web site. We thank Mr Larry Frelin for his help with the preparation of this manuscript. Additionally, we thank Dr Kirby Smith for his suggestions and support. Supported by NIEHS grant PO1-ES08131, MRDDRC grant HD 24061 and a gift from Merck Research Laboratories to J.P.

References

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, **95**, 14 863–14 868.

Heyer,L.J., Kruglyak,S. and Yooseph,S. (1999) Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.*, **9**, 1106–1115.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.