LU TP 03-24
December 5, 2003

ACID: a database for microarray clone information

Markus Ringnér[1], Srinivas Veerla[1,2], Samuel Andersson[1,2],
Johan Staaf[3] and Jari Häkkinen[1,2,*]

[1]Complex Systems Division and [2]Lund Swegene Bioinformatics Facility,
Department of Theoretical Physics, Lund University, SE-223 62 Lund, Sweden
and
[3]Department of Oncology, Lund University Hospital, SE-221 00 Lund, Sweden

submitted to *Bioinformatics*

*To whom correspondance should be addressed.

ABSTRACT

Summary: ACID is an online database for information about microarray cDNA clones. For each clone, the database contents include assigned UniGene cluster(s), location in the full-length transcript, assigned Gene Ontology terms, and position in the genome assembly.

Availability: http://bioinfo.thep.lu.se/acid.html

Contact: jari@thep.lu.se

INTRODUCTION

For microarray data analysis it is of great value to be able to extract up-to-date information about all the clones present on the arrays in a streamlined fashion. The information may be of interest for several purposes, but perhaps primarily to facilitate further data analysis. For example, to investigate where in the genome clones of interest are located, or whether these clones are significantly associated with specific Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000). Information about clones can also be useful for quality control, for example, to address where in the full-length transcript a specific clone is located. To address these issues, we have developed the Array Clone Information Database (ACID) to provide a searchable resource for information about cDNA clones. The information is accessible through a set of applications with web interfaces. Other resources that provide information about microarray clones include Resourcerer (Tsai *et al.*, 2001) and DIG (https://dig.cgt.duke.edu/index.php).

DATABASE

The database contains all *Homo sapiens* and *Mus musculus* cDNA clones present in UniGene for which there is at least one EST annotated as a 3' or 5' read of the clone (http://www.ncbi.nlm.nih.gov/UniGene). In particular the IMAGE (http://image.llnl.gov/) and RIKEN (http://genome.rtc.riken.go.jp/) cDNA clone sets are included, together with GenBank (Benson *et al.*, 2002) accession numbers for all their 3' and 5' reads. For human (using UniGene Hs build 163), the database contains 3.1 million cDNA clones and 3.4 million EST sequences based on 127,835 UniGene clusters. 16,173 of the clusters are represented by at least one RefSeq (Pruitt and Maglott, 2001) mRNA sequence. For mouse (using Mm build 130), the corresponding numbers are 2.7 million clones, 3.2 million ESTs, and 93,645 clusters (15,161 with a RefSeq sequence). The database is implemented in MySQL and applications are written in Perl and C++. Details of how the contents in ACID are assembled and generated can be found on the ACID web site.

APPLICATIONS

Currently, there are two main applications available.

Clone information application. In this application, ACID takes a list of clone identifiers, GenBank accession numbers, gene symbols, or UniGene clusters and associates them with the corresponding UniGene cluster information. For each item in the input list, minimally the UniGene cluster, the gene name, the gene symbol, the chromosome, the cytoband location, the LocusLink identifier, and the RefSeq mRNA accession number are provided. Additional information includes associated GO terms and the OMIM identifier (http://www.ncbi.nlm.nih.gov/omim/). Moreover, for UniGene clusters that contain a RefSeq full-length transcript sequence, information about the position of the RefSeq sequence in the genome assemblies in the genome browser database at the University of California Santa Cruz (Karolchik *et al.*, 2003) is provided. This position information includes chromosome, strand, and start and end bases.

Clone location application. For each cDNA clone in UniGene, we have used bl2seq (Tatusova and Madden, 1999) to align its 3' and 5' sequence reads with the longest RefSeq sequence in the corresponding UniGene cluster. In this application, ACID takes a list of clone identifiers and the aligned position of each clone in its corresponding full length transcript is provided. Alternatively, this application takes a list of RefSeq mRNA accession numbers, gene symbols or UniGene clusters and lists all clones matched in UniGene to the corresponding clusters, together with the location of the clones in the full-length transcripts. To make the results easily interpretable, a graphical display of the locations of the clones can be generated (Fig. 1). In the graphical display, the full-length transcript is annotated with coding region and exon boundaries, together with their genomic positions.

OUTLOOK

ACID will be updated as new builds of UniGene or assemblies of the genomes become available. Support for additional organisms present in UniGene may be provided in the future. We anticipate that ACID will expand to contain further information as well as additional applications, and these additions will be available as they are developed. In particular, we are planning to extend the contents of the database with more functional annotations than GO terms. This plan includes mapping the clones to the KEGG pathway database (Kanehisa *et al.*, 2002). In addition, to provide a useful analysis environment, we are planning to link ACID with BASE (Saal *et al.*, 2002), the open-source database for array data maintained by our group. The list of potential additions can be made very long. Nevertheless, ACID has already proven to be very useful in our own research, and we expect it to be of great value for anyone interested in microarray data analysis.

ACKNOWLEDGMENTS

We thank the CBI404 students at Lund University for prototyping. This work was in part supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium.

REFERENCES

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A.and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.* **30**, 17-20.

Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42-46.

Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J., Weber,R.J., Haussler,D. and Kent, W.J. (2003) The UCSC Genome Browser Database *Nucl Acids Res.* **31**, 51-54.

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29**, 137-140.

Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,Å. and Peterson,C. (2002) Bioarray software environment: a platform for comprehensive management and analysis of microarray data. *Genome Biol.* **3**, software0003.1-0003.6.

Tatusova,T.A. and Madden,T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.* **174**, 247-50.

The Gene Ontology Consortium. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet.* **25**, 25-29.

Tsai,J., Sultana,R., Lee,Y., Pertea,G., Karamycheva,S., Antonescu.V., Cho,J., Parvizi,B., Cheung,F. and Quackenbush,J. (2001) RESOURCERER: a database for annotating and linking microarray resources within and across species. *Genome Biol.* **2**, software0002.1-0002.4.

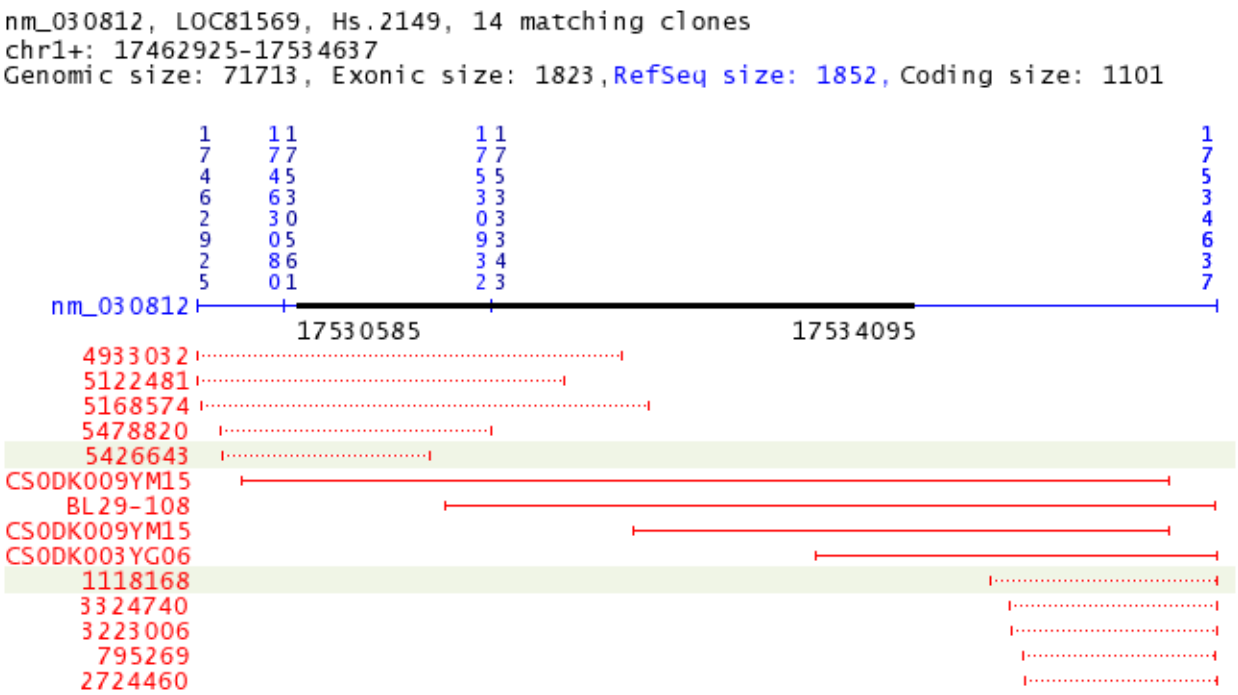

Figure 1: The graphical display available in the clone location application. UniGene cluster Hs.2149 is shown as an example. A detailed description of the display is available as part of the database information at the ACID website.