Nucleic Acids Research, 2005, Vol. 33, Database issue ***D553–D555***
doi:10.1093/nar/gki056

ArrayExpress—a public repository for microarray gene expression data at the EBI

H. Parkinson*, U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. Garcia Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone and A. Brazma

European Bioinformatics Institute, EMBL-EBI Wellcome Trust Genome Campus, Hinxton CB10 1SD, UK

Received September 22, 2004; Revised and Accepted October 1, 2004

ABSTRACT

ArrayExpress is a public repository for microarray data that supports the MIAME (Minimum Information About a Microarray Experiment) requirements and stores well-annotated raw and normalized data. As of November 2004, ArrayExpress contains data from ~12 000 hybridizations covering 35 species. Data can be submitted online or directly from local databases or LIMS in a standard format, and password-protected access to prepublication data is provided for reviewers and authors. The data can be retrieved by accession number or queried by various parameters such as species, author and array platform. A facility to query experiments by gene and sample properties is provided for a growing subset of curated data that is loaded in to the ArrayExpress data warehouse. Data can be visualized and analysed using Expression Profiler, the integrated data analysis tool. ArrayExpress is available at http://www.ebi.ac.uk/arrayexpress.

INTRODUCTION

ArrayExpress is an international public repository for microarray data established at the European Bioinformatics Institute (EBI) in 2002 (1). ArrayExpress supports standards and recommendations developed by the Microarray Gene Expression Data (MGED) society (www.mged.org), including the Minimum Information About a Microarray Experiment (MIAME) (2) and Microarray Gene Expression Mark up Language (MAGE-ML) (3). Along with Gene Expression Omnibus (4) and CIBEX (5), it is one of the three repositories recommended by the MGED society (6) for storing data related to publications. The ArrayExpress suite of databases and applications comprises: (i) MIAMExpress, a web-based MIAME supportive data-submission tool; (ii) the ArrayExpress repository that provides public and password-protected access to the submitted data; (iii) a query optimized data warehouse containing a curated subset of normalized data; and (iv) Expression Profiler, an integrated online visualization and analysis tool. All the software in the ArrayExpress suite is open source. Here we will focus on describing MIAMExpress, the repository and the data warehouse; Expression Profiler has been reviewed recently (7).

As the number of journals requiring submission to public repositories is growing, the cost of microarray experiments is falling and as data submission tools are improving, the volume of data in ArrayExpress is growing rapidly. During the last 12 months the ArrayExpress content has grown more than 10-fold (Figure 1a), and as of November 2004, the repository contains ~12 000 hybridizations comprising more than 300 studies from 35 species (Figure 1b). The majority of studies concern samples from *Homo sapiens* or *Mus musculus*. Slightly more than 25% of the experiments have been performed using Affymetrix arrays. Although the majority of experiments study gene expression, there is a growing volume of ChIP on Chip and Comparative Genome Hybridization data in ArrayExpress.

SUBMISSION AND CURATION

There are two major submission routes to ArrayExpress: (i) online via the MIAMExpress data submission tool, and (ii) via a MAGE-ML-based pipeline set-up with an external application or database. Currently, more than a half of all submissions have been submitted online. MIAMExpress is primarily aimed at users with no substantial local bioinformatics support and with no access to a local database providing direct deposition. No prior knowledge of the MIAME guidelines is required, as contextual help on the information required and help on the use of MIAMExpress is provided via links from the web interface. Submitters progress through a series of simple web forms to describe their experiment and upload the data files.

*To whom correspondence should be addressed. Tel: +44 1223 494672; Fax: +44 1223 494468; Email: parkinson@ebi.ac.uk

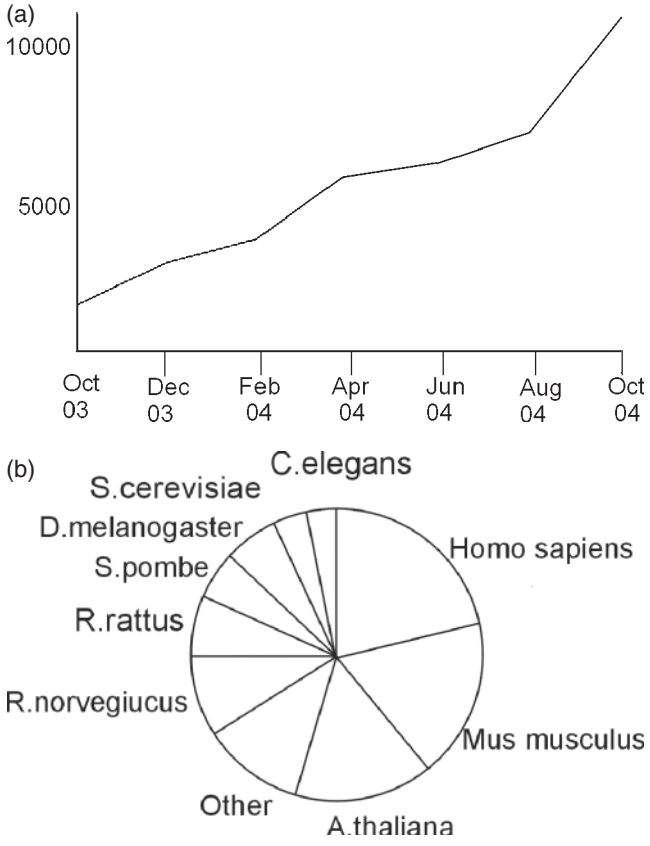

Figure 1. (**a**) The number of hybridizations from October 2003 to September 2004. (**b**) The content of the database is shown broken down by species.

MIAMExpress is an open source software that can be customized for use by a single laboratory, or for particular application domains. Examples of customization hosted at the EBI include the toxicology (8) and plant-specific MIAMExpress versions. MIAMExpress is now installed in 35 locations worldwide. Source code and installation information can be found at http://sourceforge.net/projects/miamexpress/.

The ArrayExpress curation team processes each submission before it is loaded into the repository. Submissions are checked for MIAME compliance, accuracy and completeness of biological information provided, as well as for data consistency (e.g. it is checked if submitted data files match the specified array designs). During the curation process, the curation team may contact the submitter if inconsistencies in the data are found. Once the data are successfully loaded into ArrayExpress repository, the experiment is issued an accession number and a password is provided to the submitter if requested. The data in the repository are owned by the submitter, released on the date specified or upon publication in a journal and no changes are made without the submitter's consent. Where array designs are commercially available these are pre-loaded into ArrayExpress by the curators in response to user request. Information on custom-made arrays is submitted as a tab-delimited file containing position information and annotation information.

The ArrayExpress curators work with external databases when setting up a direct data-submission pipeline to ensure that data are MIAME compliant and well formatted. Once a pipeline is established, the submissions are curated at the source database and monitored by ArrayExpress curators. MAGE-ML-based pipelines have been established from 15 external databases, manufacturers or tools, including the Stanford Microarray Database (SMD) (9), MIDAS at TIGR (10), from externally installed MIAMExpress systems at Cambridge University and the European Molecular Biology Laboratory at Heidelberg, as well as the array manufacturers, Affymetrix and Agilent.

Further data curation is performed when populating the ArrayExpress warehouse from the repository. The curators select data based on their MIAME compliance, presence of normalized data and the quality of the biological annotation. Array designs are additionally annotated to the current version of the sequence databases at the EBI and up-to-date gene annotation, such as InterPro (11) Gene Ontology (GO) terms (12) and gene names are added, while the original array annotation supporting the publication is maintained in the repository.

DATA ACCESS AND QUERY

The highest level of organization in the ArrayExpress repository is the Experiment, which consists of one or more hybridizations, usually linked to a publication. The ArrayExpress query interface provides the ability to query for Experiments, Protocols and Array designs by their various attributes, such as species, authors or array platforms. Once an experiment has been selected the users can examine the description of the samples and protocols by navigating through the experiment, or they can download the data for analysis locally. The data can also be analysed and visualized online using Expression Profiler. Password-protected access to pre-publication data is provided for submitters and reviewers.

The ArrayExpress data warehouse [which is based on the BioMart technology (11)] supports queries on gene attributes, such as gene names, gene function (GO annotations) or information on which family a gene belongs to or the motifs and domains it contains (InterPro terms), and on sample properties. The user can retrieve and visualize the gene expression values for multiple experiments. For example, querying the gene name 'jun' and sample property 'leukemia' retrieves all the experiments that contain data for a gene annotated with this name and that have been studied in experiments described using the term 'leukemia'. A list of genes that match the query is returned. These can be visualized using line plots and data can be selected for further analysis. Links are provided back to the repository where users can access the full annotation and supporting raw data. Experimental data and corresponding array designs selected by the curators on the basis of MIAME compliance, annotation quality and comparability are loaded periodically into the warehouse. A schematic diagram of the software architecture is shown in Figure 2.

FUTURE

The online submission tool MIAMExpress is being extended to allow a spreadsheet based data batch uploading to facilitate

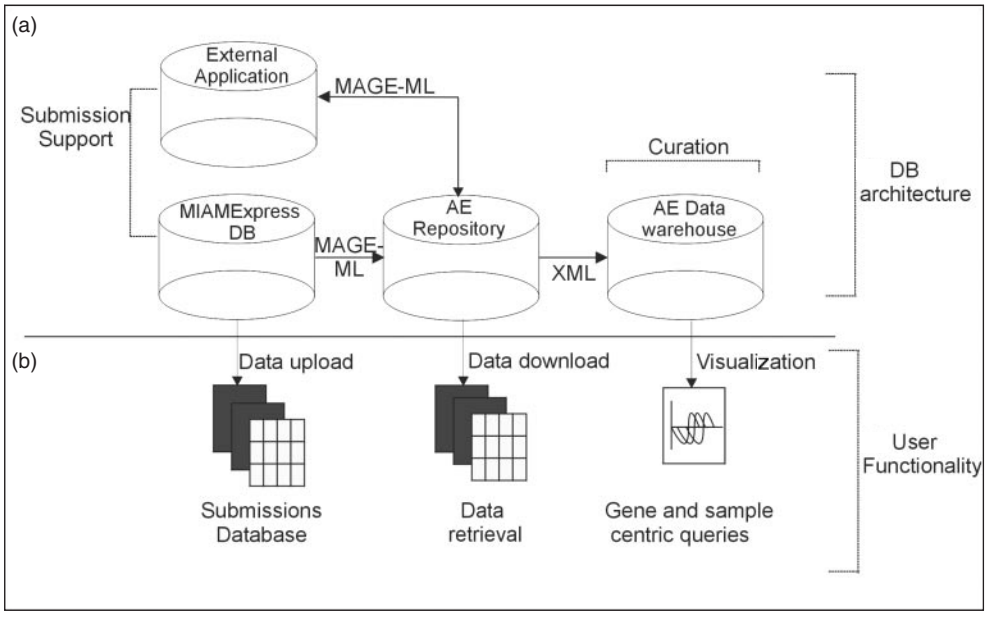

Figure 2. (**a**) The ArrayExpress architecture and database side activities are shown. (**b**). The functionality experienced by the user is shown.

large-scale experiment submissions. A graph-based visualization tool is being added to MIAMExpress and ArrayExpress. The ArrayExpress repository and data warehouse interfaces will be unified. The gene-based query facility in the warehouse will be used as the basis for integrating ArrayExpress into all EBI services more closely, for instance expression data will be accessible from UniProt and Ensembl databases via a Distributed Annotation System (DAS) (http://www.biodas.org) server. As the volume of submissions continues to grow, we expect that the curation phase at the point of submission to the repository will be fully automated and curation efforts will focus on adding value to submitted data made available through the data warehouse.

ACKNOWLEDGEMENTS

We would like to acknowledge the work of Patrick Kemmeren, Catherine Leroy, Pierre Marguerite, Bhuwan Tiwari, Jaak Vilo, Cath Brooksbank, Peteri Jokkinen and the EBI systems group. The EMBL, the European Commission (TEMBLOR, CAGE), International Life Sciences Institute (ILSI) and the National Institutes of Health (NIH) support ArrayExpress Development.

REFERENCES

1. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al*. (2003) Array Express: a public for microarray gene expression data at the EBI. *Nucleic Acids Res*., **31**, 68–71.
2. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al*. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genet*., **29**, 365–371.
3. Spellman,P.T., Miller,M., Stewart,J., Troup,C., Sarkans,U., Chervitz,S., Bernhart,D., Sherlock,G., Ball,C., Lepage,M. *et al*. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*., **3**, RESEARCH0046.
4. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*., **30**, 207–210.
5. Ikeo,K., Ishi-i,J., Tamura,T., Gojobori,T. and Tateno,Y. (2003) CIBEX: center for information biology gene expression database. *C. R. Biol*., **326**, 1079–1082.
6. Ball,C.A., Brazma,A., Causton,H., Chervitz,S., Edgar,R., Hingamp,P., Matese,J.C., Parkinson,H.E., Quackenbush,J., Ringwald,M., Sansone,S., Sherlock,G., Spellman,P., Stoeckert,C., Tateno,Y., Taylor,R., White,J. and Winegarden,N. (2004) Submission of microarray data to public repositories. *PLoS Biol*., **2**, 1276–1277.
7. Kapushesky,M., Kemmeren,P., Culhane,A.C., Durinck,S., Ihmels,J., Korner,C., Kull,M., Torrente,A., Sarkans,U., Vilo,J. *et al*. (2004) Expression Profiler: next generation—an online platform for analysis of microarray data. *Nucleic Acids Res*., **32**, W465–W470.
8. Mattes,W.B., Pettit,S.D., Sansone,S.A., Bushel,P.R. and Waters,M.D. (2004) Database development in toxico genomics: issues and efforts. *Environ. Health Perspect*., **112**, 495–505.
9. Sherlock,G., Hernandez-Boussard,T., Kasarskis,A., Binkley,G., Matese,J.C., Dwight,S.S., Kaloper,M., Weng,S., Jin,H., Ball,C.A. *et al*. (2001) The Stanford Microarray. *Nucleic Acids Res*., **29**, 152–155.
10. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M., Sturn,A., Snuffin,M., Rezantsev,A., Popov,D., Ryltsov,A., Kostukovich,E., Borisovsky,I., Liu,Z., Vinsavich,A., Trush,V. and Quackenbush,J. (2003) TM4: a free, open-soure system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.
11. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*., **31**, 315–318.
12. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*., **32**, D258–D261.
13. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: ageneric system for fast and flexible access to biological data. *Genome Res*., **14**, 160–169.