Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Biologies 326 (2003) 1075–1078
COMPTES RENDUS BIOLOGIES

Molecular biology and genetics

ArrayExpress: a public database of gene expression data at EBI

Philippe Rocca-Serra, Alvis Brazma *, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Sergio Contrino, Jaak Vilo, Niran Abeygunawardena, Gaurab Mukherjee, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, Susanna-Assunta Sansone

EMBL Outstation – Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

Received 16 September 2003; accepted 23 September 2003

Presented by François Gros

Abstract

ArrayExpress is a public repository for microarray-based gene expression data, resulting from the implementation of the MAGE object model to ensure accurate data structuring and the MIAME standard, which defines the annotation requirements. ArrayExpress accepts data as MAGE–ML files for direct submissions or data from MIAMExpress, the MIAME compliant web-based annotation and submission tool of EBI. A team of curators supports the submission process, providing assistance in data annotation. Data retrieval is performed through a dedicated web interface. Relevant results may be exported to Expression-Profiler, the EBI based expression analysis tool available online (http://www.ebi.ac.uk/arrayexpress). ***To cite this article: P. Rocca-Serra et al., C. R. Biologies 326 (2003).***

Résumé

ArrayExpress : une base publique de données d'expression génique à l'EBI. ArrayExpress est une base d'archivage publique pour les données d'analyse d'expression génique par microréseaux, résultant de l'implémentation du modèle objet MAGE assurant la structuration correcte des données, et du standard MIAME, qui définit les prérequis en matière d'annotation. ArrayExpress accepte les données directement sous format MAGE–ML ou via MIAMExpress, l'utilitaire de soumission en ligne de l'EBI. Une équipe d'annotateurs accompagne le processus de soumission et fournit une assistance à l'annotation des données. L'interrogation de la base est réalisée via une interface Web dédiée permettant d'exporter les résultats des requêtes vers *Expression Profiler*, l'outil d'analyse en ligne développé par l'EBI (http://www.ebi.ac.uk/arrayexpress). ***Pour citer cet article : P. Rocca-Serra et al., C. R. Biologies 326 (2003).***

Keywords: ArrayExpress; database; Expression Profiler; MAGE–ML; MIAME standard; microarray; MGED ontology

Mots-clés : ArrayExpress ; base de données ; *Expression Profiler* ; MAGE–ML ; Ontologie MGED ; MIAME standard ; microréseaux

* Corresponding author.
E-mail address: brazma@ebi.ac.uk (A. Brazma).

1631-0691/$ – see front matter
doi:10.1016/j.crvi.2003.09.026

1. Introduction

Though microarray techniques have been available for several years and that large amounts of data have been gathered, major breakthroughs are still yet to come. If heterogeneity in technology, platforms and computing options may be blamed for the delay [1], the lack of thought through exchange infrastructure represents the major hurdle. So far, most microarray data have been published on specific web sites. These resources are usually of limited value due to lack of annotation, both in quantity and quality. These limitations, by preventing cross platform analysis and mining, make it almost impossible to fully exploit the data so far accumulated. The genuine complexity and size of data produced by microarray technology has therefore generated a need for setting up guidelines to achieve data exchangeability. To this end, two standards have been devised to solve, first the problem of the structure of the data and second the problem of the amount of information required for microarray experiment annotation. ArrayExpress aims to provide a public repository by implementing these standards and supplying the infrastructure that should favor microarray data exchange and interpretation.

2. Structuring the microarray data

Exchanging data requires common standards to describe, structure and format data in a way that could be implemented irrespective to technical choices. The MAGE–OM object model is a platform-independent data model capable of describing the intrinsic complexity of the microarray-based experiment. The MAGE–ML language, an XML derived language, and its related Data Type Definition has been generated from the MAGE–OM object model [2]. These three elements have now been granted the status of Bioscience standards by the OMG and are gaining broader acceptance among the most prominent industrial and academic players of the field. ArrayExpress and its environment is the first functional implementation of the MAGE–OM object model allowing data submission in MAGE–ML format.

3. The challenge of data annotation

In addition to defining the standards for data structure and modeling, the huge challenge of annotation has to be addressed both in quantity and quality to ensure complete data compatibility and reusability. The MIAME requirements standing for Minimal Information About a Microarray Experiment [3] have been developed to tackle the issue of the amount of information to be supplied. The standard defines for every critical element of a microarray-based experiment, the necessary information to be provided by anyone willing to share the results of his work.

When dealing with quality of annotation, a critical issue is the need for machine processable descriptions. To achieve automated treatment of the information, consistent annotation is a paramount for mining agents to work efficiently; synonyms and free text should therefore be avoided. To this end, an effort has been carried out to develop field specific ontologies, which capture knowledge, and controlled vocabularies to perform efficient microarray experiment annotation. Among those, the Biomaterial Ontology, established by the MGED society, provides a standard way for annotating biological samples from which mRNA are extracted and used in microarray experiments. The ontology itself relates and cross-references to several controlled vocabulary projects and annotation database thus taking advantage of existing effort. The MGED Biomaterial ontology is available at http://www.cbil.upenn.edu/Ontology/MGEDontology.html.

4. Submission routes to ArrayExpress

Based on the experience gained from the sequencing projects [4], adequate submission procedures have been devised depending on submitter's needs. MAGE–ML pipelines have been tailored for institutions involved in high-throughput projects (e.g., The Sanger Center, TIGR, Affymetrix) or microarray computing projects such as BASE [5]. For smaller scale projects or with limited bioinformatics support, MIAME express, a MIAME compliant web-based tool for submission and annotation, is available. MIAMExpress can be used as a submission tool when all experiments are completed or alternately on a daily basis,

as an electronic lab-book. The tool provides a simple and robust tool for submitting experiments, protocols and arrays while ensuring appropriate formatting and annotation. The complexity of the MAGE–ML format conversion is taken care of by the tool so that researchers using MIAMExpress are at one click ahead of submission. MIAMExpress is implemented using perl-cgi scripts and stores the data temporarily in a mySQL database. This transient storage has two purposes: (1) store pending submissions and (2) enable quality control of annotation and structure by the microarray curation team. Throughout the submission process, submitters are assisted and guided by the curation team available at arraysubs@ebi.ac.uk. Last, MIAME express can also be set up as a standalone tool and is available as open-source from http://www.sourceforge.net/.

5. Accessing and mining the data

ArrayExpress data can be viewed through a dedicated query form (http://www.ebi.ac.uk/arrayexpress). All submission types can be queried on accession numbers. Type specific (Experiment, Array and Protocol) query fields allow case insensitive searches on e.g. authors, experimental factor, experiment type and species. Results are displayed as short summaries containing a series of links to the different objects. From there, numerical data corresponding to the gene expression levels are made available as tab-separated file. These can then be directed to ExpressionProfiler (http://www.ebi.ac.uk/microarray/ExpressionProfiler/ep.html), the EBI online analysis tool for further analysis and visualization [6]. Finally, MAGE–ML documents can be downloaded as a compressed file directly from the result interface. Note that sequence or gene identifier based queries are not yet supported and further work is needed to implement those. The task is complex and requires integration of a broad variety of resources from within the EBI and other institutions and will require the development of a specific datawarehouse for ArrayExpress. The MAGE–ML formatted content of Array Express database is available on request from arraysubs@ebi.ac.uk.

6. ArrayExpress future

Even though ArrayExpress is now fully functional, allowing submission, query and export to analysis tool, it is still a tool under development and does not yet take advantage of the full power of the MAGE object model. Hence, work is still ongoing to enhance the query capabilities, especially those related to gene and reporter that should enable cross platform and reporter reliability assessment. Integration of query capabilities based on ontology annotations is also scheduled as part as query functionalities. To achieve microarray data exchange, interconnection with other microarray databases such as GEOmnibus at the NCBI [7] and with the CIBEX project at the DDBJ has to be implemented. This requires devising a MAGE–ML export function. A variant application of that export function could be used to transfer MAGE–ML files to private databases in order to perform local assessments. In addition to software related efforts, we are actively working with different centers and consortia to generate high quality MIAME compliant data, examples of these include the International Genomics Consortium (IGC) [8] who intend to profile thousands of tumor samples and deposit the data in ArrayExpress and ILSI Toxicogenomics projects.

Acknowledgements

The ArrayExpress project is funded by EMBL, a grant from the European Commission (TEMBLOR), and a Toxicogenomics database grant from ILSI. Initial funding was provided by Incyte and we particularly thank Lee Grower.

References

[1] A. Brazma, A. Robinson, G. Cameron, M. Ashburner, One-stop shop for microarray data, Nature 403 (2000) 699–700.

[2] P.T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W.L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B.J. Aronow, A. Robinson, D. Bassett, C.J. Stoeckert, A. Brazma, Design and implementation of microarray gene expression markup language (MAGE–ML), Genome Biol. 3 (2002) 46.1–46.9.

[3] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C.A. Ball, H.C. Causton, T. Gaasterland, P. Glenisson, F.C.P. Holstege, I.F. Kim, V. Markowitz, J.C. Matese, H. Parkinson, A. Robinson, S. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, M. Vingron, Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, Nature Genet. 29 (2001) 365–371.

[4] G. Stoesser, W. Baker, A.E. van den Broek, E. Camon, P. Hingamp, P. Sterk, M.A. Tuli, The EMBL nucleotide sequence database, Nucleic Acids Res. 28 (2000) 19–23.

[5] L.H. Saal, C. Troein, J. Vallon-Christersson, S. Gruvberger, A. Borg, C. Peterson, BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data, Genome Biol. 3 (2002) 3.1–3.6.

[6] J. Vilo, M. Kapushesky, P. Kemmeren, U. Sarkans, A. Brazma, Expression profiler, in: G. Parmigiani, E.S. Garrett, R. Irizarry, S.L. Zeger (Eds.), The Analysis of Gene Expression Data: Methods and Software, Springer-Verlag, in press.

[7] R. Edgar, M. Domrachev, A. Lash, Gene Expression Omnibus: NCBI gene expression and hybridisation array data repository, Nucleic Acids Res. 30 (2002) 207–210.

[8] J. Knight, Cancer comes under scrutiny in fresh genomics initiative, Nature 4 (2001) 855.