BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 19 2009, pages 2566–2572
doi:10.1093/bioinformatics/btp422

Gene expression

aGEM: an integrative system for analyzing spatial-temporal gene-expression information

Natalia Jiménez-Lozano*,†, Joan Segura†, José Ramón Macías, Juanjo Vega and José María Carazo

GN7 of the National Institute for Bioinformatics (INB) and Biocomputing Unit of the National Centre for Biotechnology (CNB-CSIC), Darwin 3, Campus de Cantoblanco, 28049 Madrid, Spain

Received on October 20, 2008; revised and accepted on July 7, 2009

Advance Access publication July 9, 2009

Associate Editor: David Rocke

ABSTRACT

Motivation: The work presented here describes the 'anatomical Gene-Expression Mapping (aGEM)' Platform, a development conceived to integrate phenotypic information with the spatial and temporal distributions of genes expressed in the mouse. The aGEM Platform has been built by extending the Distributed Annotation System (DAS) protocol, which was originally designed to share genome annotations over the WWW. DAS is a client-server system in which a single client integrates information from multiple distributed servers.

Results: The aGEM Platform provides information to answer three main questions. (i) Which genes are expressed in a given mouse anatomical component? (ii) In which mouse anatomical structures are a given gene or set of genes expressed? And (iii) is there any correlation among these findings? Currently, this Platform includes several well-known mouse resources (EMAGE, GXD and GENSAT), hosting gene-expression data mostly obtained from *in situ* techniques together with a broad set of image-derived annotations.

Availability: The Platform is optimized for Firefox 3.0 and it is accessed through a friendly and intuitive display: http://agem.cnb.csic.es

Contact: natalia@cnb.csic.es

Supplementary information: Supplementary data are available at http://bioweb.cnb.csic.es/VisualOmics/aGEM/home.html and http://bioweb.cnb.csic.es/VisualOmics/index_VO.html and *Bioinformatics* online.

1 INTRODUCTION

Gene expression is the process by which heritable information from a gene is made into a functional gene product, such as protein or RNA. Precise regulation of spatio-temporal gene expression is crucial during the development of an organism. It is essential to know the exact timing and location of gene transcripts when studying the functions of genes involved in developmental processes.

There is a broad range of freely accessible gene-expression databases (GXDs). Each is devoted either to a whole organism [ZFIN for zebrafish, BDGP and FlyBase for Drosophila, MEPD for Medaka, Anissed for Ciona, XDB3 for Xenopus, GXD and EMAGE for mouse (Haudry *et al*., 2008)], or to a specific part of the organism at a specific developmental stage or during the period between two developmental stages. Thus, gene-expression data are broadly dispersed among many research groups distributed around the world. An effort at integration is needed to share the results and analyses among all groups in order to obtain a general and complete landscape of the gene-expression field. During recent years, several initiatives have tried to address the issue of data integration. They can be grouped into two main movements. The first is devoted to constructing a main repository that collects all the information available for a single organism (e.g. A.C. elegans database: ACeDB, http://www.acedb.org/) or for many organisms (Ensembl; Hubbard *et al*., 2007). The basis of this kind of effort is centralization. In contrast, the foundation for the second type of initiative is decentralization, which occurs when research projects and institutions maintain and provide their own data separately (Cuticchia, 2000; Letovsky *et al*., 1998; Shoman *et al*., 1995, Skupski *et al*., 1999) This integration model requires a protocol that will allow the user to retrieve information in a robust manner.

The platform described in this article exploits decentralization by using the Distributed Annotation System (reference DAS). DAS defines a simple protocol that enables clients to retrieve data and annotations from multiple and disperse servers in a homogeneous form. It describes how data should be represented and communicated. Data sources in DAS are implemented as web services using eXtensible Markup Language (XML) for data representation and the hyper-text transfer protocol (HTTP) for data transport. DAS clients send out HTTP requests to the DAS servers in charge of processing the queries and returning an HTTP response that contains an XML representation of the resources.

The platform described in this article is a DAS system that integrates the following resources: EMAGE, GXD, GENSAT and OMIM. These resources will now be explained.

GXD and EMAGE databases were the result of an innovative project started in 1994 in the UK MRC Human Genetics Unit in Edinburgh, together with the Jackson Laboratory in the USA (Ringwald *et al*., 1994). This project (named EMAP: Edinburgh Mouse Atlas Project) was devoted to creating 3D grey-level voxel images from several mouse developmental stages (named

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

Theiler Stages; Theiler, 1989), to which image-mapped gene-expression data from *in situ* experiments were related. Obviously, this correspondence was only possible thanks to the development of a well curated mouse anatomical ontology, which describes the anatomical terms for each of the 26 embryonic Theiler Stages (EMAP ontology; Baldock and Burger, 2005). This ontology is designed to capture the structural changes that occur during embryonic development and consists of a set of 26 hierarchies, one for each developmental stage. Each stage is characterized by the internal and external morphological features that are recognizable in an embryo during that period of development (Burger *et al*., 2004).

GXD first full production version was released in 1999 (Ringwald *et al*., 1999). GXD is a database component of the Mouse Genome Informatics (MGI) database resource http://www.informatics.jax.org, the community model organism database for the laboratory mouse. GXD currently stores gene-expression data from *in situ* and other techniques such as northern blotting, real-time polymerase chain reaction, etc., and refers not only to the mouse embryo (Theiler Stages 1–26) but also to the adult mouse (postnatal mouse corresponding to Theiler Stage 28). The GXD uses the EMAP ontology together with a compatible anatomical dictionary for the adult mouse (MA ontology; Hayamizu *et al*., 2005).

EMAGE database was launched in 2003 (Baldock *et al*., 2003) in the framework of EMAP project to provide spatially mapped gene-expression data referred only to mouse embryo (Venkataraman *et al*., 2008).

More recently, Gong and co-workers have created the Gene-expression Nervous System Atlas (GENSAT; Gong *et al*., 2003), a resource of gene-expression data relating to the developing and adult mouse nervous system at the cellular level, with data supplied by the National Institute of Neurological Disorders and Stroke. The aim of the GENSAT project is to map the expression of genes in the mouse central nervous system (CNS), using both *in situ* hybridization and transgenic mouse techniques. It is a collection of pictorial gene-expression maps of the mouse brain and spinal cord. GENSAT catalogs images of histological sections of the developing (embryonic days 10.5 and 15.5) and adult mouse brain, in which tags such as Enhanced Green Fluorescence Protein have been used to visualize the relative degrees of *in situ* expression of a wide array of genes (Wheeler *et al*., 2008). Unlike EMAGE and GXD, annotations in GENSAT are not referred to a well-established ontology but to a vocabulary that includes anatomical terms together with cellular and sub-cellular types.

Of course, this new set of spatial gene-expression data must be related to the wealth of existing biological information already organized into a wide variety of databases. As an example, in this work we have used OMIM, the 'Online Mendelian Inheritance in Man' database (OMIM; Hamosh *et al*., 2002) to link genes involved in disease to (mouse) gene-expression data from the EMAGE, GXD and GENSAT databases. OMIM catalogs all known diseases with a genetic component, and links them whenever possible to the relevant genes in the human genome. It also provides references for further research and tools for genomic analysis of the catalog gene.

In this article we present the 'anatomical Gene-Expression Mapping (aGEM)' Platform, the aim of which is to facilitate access to information related to the anatomical pattern of gene expression in the mouse. It can therefore complement many functional genomics studies. The aGEM Platform extends the Distributed Annotation System (DAS) protocol, which was conceived to share genome annotations.

The main biologically relevant questions that can be answered by the 'aGEM Platform' are as follows. (i) Which genes are expressed in a given anatomical component? (ii) In which anatomical structures is a given gene (or set of genes) expressed? And (iii) is there any correlation among these findings?

Before studying the 'aGEM Platform' in depth, it is necessary to explain what the term 'annotation' means in the context of this work. The literal definition of annotation is: 'extra information inserted at a particular point in a document or other piece of information' (http://en.wikipedia.org/wiki/Annotation). In this context, the central piece of information in the aGEM Platform is the set of spatial gene-expression data, relating information about gene expression to the list of anatomical components in which a given gene may be expressed. Experts then interpret the results that provide annotations, e.g. additional information such as expression pattern, expression level, image description or details about the experiment. Thus, pieces of objective information directly derived from experiments are called 'data', and subjective 'expert' interpretations are called 'annotations'.

Owing to the diversity of gene-expression resources integrated in the aGEM Platform, it is necessary to design a coherent way of referring to the genes and to the anatomical components in which a gene is expressed. Gene nomenclature is an active work area in bioinformatics, but despite efforts in this direction (Tamames and Valencia, 2006), genes and their products are occasionally named inconsistently, resulting in a confusing set of synonyms and homonyms. Much effort is focused on disambiguating names, mainly led by text-mining tools, which need uniform gene and protein nomenclature rules in order to extract these symbols from the scientific literature. Because the aGEM Platform is primarily designed to store spatial and temporal gene-expression information for the mouse, the Mouse Genome Informatics Identifiers (MGI) established by the Mouse Genome Informatics Resource have been taken as the standard.

Consistency in the spatial localisation of gene expression has been achieved by using standardised anatomical ontologies for each organism. The 26 EMAP embryo ontologies describing mouse development together with the adult mouse ontology constitute the foundations to which gene expression has to be referred. In the current version of aGEM, the vocabulary from the GENSAT database has been semi-automatically mapped to these reference ontologies. Therefore, the procedure of including new annotation resources in aGEM will require a careful study, which will include ontology mapping and the resolution of gene nomenclature discrepancies.

2 SYSTEM AND METHODS

The Distributed Annotation System (DAS) is a lightweight system for integrating biological data and annotations hosted in multiple and disperse resources. DAS became an open standard implemented as a client-server system in which a single client integrates information from multiple servers. The logic of DAS is based on the rule 'dumb server, clever client'. This means that the server's behavior is as simple as possible, responding to a limited set of commands and returning data in a well-defined, uniform format. On the other hand, clients will be in charge of retrieving, analyzing and displaying the data to the user in the correct way.

DAS was originally designed to share annotations of genomes (Dowell *et al*., 2001), but its usefulness has led to extension to other fields such as protein sequences (Jones *et al*., 2005), protein structures (Prlic *et al*., 2005) and even 3D volumes obtained from electron microscopy with fitted structures (Macías *et al*., 2007).

The basic DAS architecture consists of a 'reference' service and one or more 'annotation' services. The reference service hosts the entities on which to base annotations as a hierarchy of 'entry points' (in this specific case: the anatomical terms or the genes), while the annotation services host all the features for the corresponding reference entities.

All DAS requests take the form of a URL with the following format:
site_prefix/das/data_source/command[?param1=value1[&paramn=valuen]]

That is, a site-specific prefix (http://biocomp.cnb.csic.es:9090, for the services described here), followed by a standardized path that includes the word 'das' and a query string containing the data source name and a command. Some commands may require parameters in the form param=value, which are then placed after a question mark. When more than one parameter is needed, an ampersand is used as separator.

The response from the server to the client consists of a standard HTTP header with DAS status information, followed by an XML document. An extension of the DAS protocol is presented in this article exemplified by a DASTERM.xml document (term retrieval). This document is the response from the reference server to the query for a specific anatomical term or gene.

Typical DAS sessions start with the client application querying the registry for a list of active servers. Then one or more entry points are requested from the corresponding reference servers. All the annotation servers storing information about the requested entry points are then queried. Finally, the client collects and processes the data and displays them.

Proserver, a Perl-based, lightweight server (Finn *et al*., 2007), has been used to implement the reference and annotation servers to be described in this article. Proserver is widely used within the DAS community owing to its simplicity, easy configuration and versatility. A detailed description of each server plus the Visual Genomic client composing the aGEM Platform is presented in the following section.

3 ALGORITHM

3.1 Reference server

In the extension of the DAS protocol described in this article, two reference services are hosted in the reference server. For the first, the entities of reference are the set of terms from the EMAP and MA ontologies and the GENSAT vocabulary describing the anatomical structures of the mouse embryo, postnatal mouse and mouse central nervous system, respectively. The GENSAT vocabulary is more specific because it reaches cellular and sub-cellular levels. EMAP and MA ontologies are the result of a joint effort between the MRC Human Genetics Unit in Edinburgh and the Jackson Laboratory in USA. Therefore, these two ontologies share the same hierarchical structure. However, the GENSAT anatomical structure list is not considered as an ontology but rather as a vocabulary. The problem of storing EMAP and MA ontologies together with GENSAT vocabulary arises when two different terms (synonyms) are employed to refer to the same anatomical entity. This problem has been solved by manually aligning the anatomical terms from both repositories. The second reference service has as entry points the whole set of genes from the EMAGE, GXD, GENSAT and OMIM databases. An example query to these reference services is:

http://biocomp.cnb.csic.es:9090/das/EMAP_Reference/term
?query=EMAP800
http://biocomp.cnb.csic.es:9090/das/MGI_Reference/term?query
=MGI98330

Here, 'term' is the new command, extending the DAS protocol, to retrieve the anatomical term data by EMAP ID, MA ID or GENSAT ID or gene data by Mouse Genome Informatics ID (MGI ID; Ringwald *et al*., 1999) or UniProt ID (the UniProt Consortium, 2009). The 'term' command requires the 'query' parameter, followed by the identifier of the entry to be retrieved.

3.2 Annotation servers

The annotation servers host all the features known about the entry points produced and controlled by a specific annotation provider, which may be a research laboratory or an established database. The system described in this article includes a separate annotation service for each data source. Four annotation services have been integrated, corresponding to the EMAGE, GXD, GENSAT and OMIM databases. The first and second services store annotations that refer to anatomical terms from the EMAP and MA ontologies. The third service provides gene-expression information from the mouse central nervous system, which reaches cellular and sub-cellular resolution and is referred to the GENSAT vocabulary. The fourth service complements data retrieved by the three aforementioned gene-expression annotation servers providing information about the pathological process in which the gene/s under study is/are involved. As the OMIM database stores only human gene data, an intermediate step has been considered in order to transform human gene identifiers into their mouse homologs using BioConductor (Gentleman *et al*., 2004). The decision to include OMIM has been driven by the need to link gene-expression data with human disease information. Owing to the clearly distinct functions of the EMAGE, GXD and GENSAT servers with respect to the OMIM server, a simple classification has been established that divides the servers into gene-expression servers and general biological information servers. The annotation servers can be accessed via an http request in the same way as the reference servers:

Prefix/das/Annotation_Server/features?segment=entry_point
_indentifier

where 'features' is the DAS command for retrieving annotations at the chosen entry point, which is passed through the 'segment' parameter. The entry points can be anatomical terms, for example EMAP800, which corresponds to the 'future midbrain' present during Theiler stage 14, EMAP3056 which corresponds to 'heart' in Theiler stage 18 and EMAP7570 that is 'hindbrain' in Theiler stage 23:

http://biocomp.cnb.csic.es:9090/das/EMAGE_Features/features
?segment=EMAP800
http://biocomp.cnb.csic.es:9090/das/GXD_Features/features
?segment=EMAP3056
http://biocomp.cnb.csic.es:9090/das/GENSAT_Features/features
?segment=EMAP7570

or genes, as for CD44 antigen with MGI identifier 88338 nonagouti gene with MGI identifier 87853, the RAR-related orphan receptor beta gene (MGI:1343464) or PAX6:

http://biocomp.cnb.csic.es:9090/das/EMAGE_Features/features
?segment=MGI88338
http://biocomp.cnb.csic.es:9090/das/GXD_Features/features
?segment=MGI87853

http://biocomp.cnb.csic.es:9090/das/GENSAT_Features/features
?segment=MGI1343464
http://biocomp.cnb.csic.es:9090/das/OMIM_Features/features
?segment=PAX6

Note that a separate request has to be issued for each annotation server.

4 IMPLEMENTATION

This section will be devoted to explaining the functionalities of the aGEM client, which, like all DAS clients, is in charge of gathering up all the information (in the XML document) from the different services and overlaying it, producing a single integrated view.

aGEM client v2.0 not only gives an integrated view of the four databases mentioned above, but also allows the experimentalist to retrieve relevant statistical information relating gene expression to anatomical structure (space) and developmental stage (time). aGEM is accessible in this URL: http://agem.cnb.csic.es and can answer the following questions. (i) Which genes are expressed in a given anatomical component? (ii) In which anatomical structures is a given gene (or set of genes) expressed? And (iii) is there any correlation among these findings? To answer the first question, the client launches a query to each of the three gene-expression annotation servers, as shown in the previous section.

The query by gene identifier (gene ID) is answered by the gene-expression annotation servers and by the general biological information server OMIM. In both cases the server response is an XML document (DASTERM.xml) containing all expression assays that comply with the requisites imposed by the client. The client then converts the XML document into a HTML document by applying a convenient formatting style sheet for Mozilla Firefox version 3.0.

From the data structure point of view, the portal codifies each entry in the gene-expression databases by a 4D '*tuple*' containing the gene identifier, the anatomical structure (spatial component), the developmental stage (represented as a discrete time component corresponding to the Theiler Stage) and a value that quantifies the gene-expression level between 0 = not expressed to 7 = highly expressed (gene-expression strength).

Obtaining the first element of the tuple is quite easy since all databases integrated in the Platform use the same gene identifiers (those given by MGI; Ringwald *et al.*, 1999). However, the situation is different for spatial and temporal data. Our Platform stores this information using the EMAP and MA ontologies. For data provided by EMAGE and GXD, this approach is straightforward, since they use those ontologies to identify structural terms. Nevertheless, GENSAT has its own vocabulary, so we have mapped it to the appropriate EMAP and MA ontology term.

The aGEM Platform explicitly considers the different levels of gene-expression strength in its statistical analysis. Gene-expression strength information from the different databases is stored and combined into an 'expression strength' value, not an easy task considering that EMAGE, GXD and GENSAT evaluate gene-expression strength in different ways. In this work we have unified these different schemas by converting the different text scores into numerical values for each database (Table 1).

To begin our discussion on the data analysis capabilities of aGEM, it is convenient to introduce some notation. From now on, a 4-tuple, represented as (g,s,t,v), will be referred to as a score item. The set of score items is represented as: $\{(g,s,t,v)/g\in G, s\in S, t\in\{1,\ldots,27\}, v\in[0,7]\}$, where G is a set of gene IDs, S is a set of anatomical structures, the set $\{1,\ldots,27\}$represents the 27 different Theiler stages and the interval $[0,7]$ is the possible range of gene-expression level values measured in a assay. Since there may be many experiments for the same gene, anatomical structure and Theiler stage, when the client queries the aGEM server, the score items from all the selected-expression assays (those that comply with the requisites imposed by the user in the query) are grouped into subsets with equal g,s and t.The average gene-expression strength values are then calculated for each subset. The result is the average strength score for each subset$(g,s,t,\bar{v})$.

Table 1. Expression strength scores in the different databases and their corresponding values in aGEM

GXD							
Absent	Ambiguous	Trace	Weak	Present	Moderate	Strong	Very strong
0	1	2	3	4	5	6	7

EMAGE					
Not detected	Possible	Detected	Weak	Moderate	Strong
0	1.4	2.8	4.2	5.6	7

GENSAT		
Undetectable	Weak signal	Moderate to strong signal
0	2.3	5.75

The aGEM Platform can perform three types of statistical analysis.

The first type of analysis fixes the anatomical structure, $s_0\in S$ and $\{(g,s_0,t,\bar{v})/g\in G, t\in\{1,\ldots,27\}, \bar{v}\in[0,7]\}$, generating three views. The first of these corresponds to a matrix in which the rows are the different genes assayed in the selected structure, while the columns correspond to each of the 27 Theiler stages. The matrix is displayed following a color code gradient that depends on the gene-expression strength value (Fig. 1A). In this way a very intuitive representation is obtained of 'when' different genes start to be expressed in a given anatomical component during development. The second view addresses the issue of gene co-expression in a given anatomical component. To this end, the Pearson's correlation coefficient is calculated among average score items of the selected anatomical structure $s_0\in S$. The result is a color gradient correlation matrix in which genes with similar expression patterns (strongly co-expressed) are represented by red positions and genes with opposed expression patterns (negatively correlated genes) are represented by light blue positions (Fig. 1B). Filtering by P-value can now be applied, generating a focused third view (Fig. 1C), allowing us to analyze those significantly co-expressed genes further.

In the second type of analysis performed by the aGEM Platform, a specific developmental stage $t_0\in\{1,\ldots,27\}$ is fixed, generating a view showing the expression strength (represented as a color code

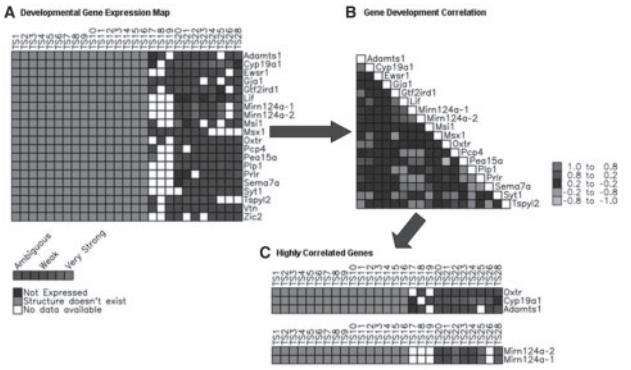

Fig. 1. (**A**) Matrix showing the gene-expression level for a set of genes during the development for a fixed anatomical structure (in this case, brain). The matrix is displayed following a color code gradient (below the matrix) that depends on the gene-expression strength value. There are two bars below the matrix. The horizontal colored bar denotes the expression level color code that ranges from red (strong expression) to blue (weak expression). The white color on the vertical bar represents the lack of available data; the grey color denotes the inexistence of the anatomical structure under study in a particular developmental stage and the black color means no expression. (**B**) Gene correlation matrix. The information displayed in the gene matrix shown in A can be summarized in the gene correlation matrix shown here. The color code for matrix elements is shown on right hand side of the matrix. The result is a color gradient correlation matrix in which genes with similar expression patterns (strongly co-expressed) are represented by red positions and the cyan positions correspond to genes with opposed expression patterns (negative correlated genes). (**C**) From the gene correlation matrix in B, the system can infer several comparative matrices showing the expression pattern during the 27 developmental stages of the correlated genes. Observe the similar expression pattern of genes Mirn124a-1 and Mirn124a-2.

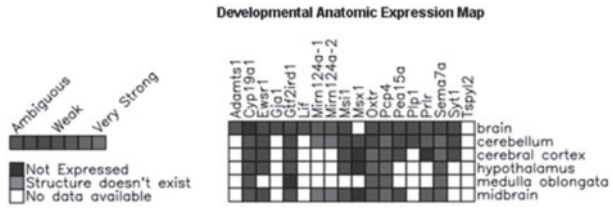

Fig. 2. Matrix showing the expression strength for a set of genes related to brain (columns), for a fixed Theiler stage (TS28). A structure filtered statistical analysis has been carried out to select only six structures. The bars on the left show the color code employed.

gradient) for a specified set of genes (columns) in the different structures (rows) $\{(g,s,t_0,\overline{v})/g\in G, s\in S, \overline{v}\in[0,7]\}$ (Fig. 2).

Finally, the third type of analysis focuses on a specified gene, $g_0\in G$, generating three views. The first of these focuses on the set of anatomical structures in which this gene has been assayed: $\{(g_0,s,t,\overline{v})/s\in S, t\in\{1,...,27\}, \overline{v}\in[0,7]\}$. Rows and columns in this expression matrix correspond to anatomical structures and Theiler Stages (discrete times), respectively (Fig. 3A). The second view computes the Pearson's correlation coefficient among the average score items of the selected gene g_0 among the different anatomical structures (Fig. 3B). These expression matrices are very useful for following the expression pattern of a single gene during embryonic development. Finally, a focused view can be obtained through filtering by *P*-value (Fig. 3C).

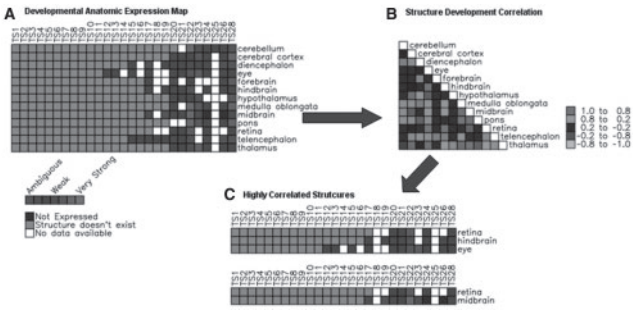

Fig. 3. (**A**) Matrix showing the expression strength during the development for a fixed gene (Pax6). Pax6 is a transcriptional factor involved in the development of central nervous system and retina. A structure filtered statistical analysis has been carried out to, selecting only those structures related to CNS and eye. Observe the similar expression pattern for retina, hindbrain and eye during the developmental stages comprised between TS20 and TS22 (**B**). The information displayed in the structure filtered matrix shown in A can be summarized in the structure correlation matrix shown here. The color code ranges from red to blue. Red color indicates the maximum expression pattern correlation along the development for a given gene in two different structures. Here, maximum correlations can be found for example between eye and hindbrain, between hindbrain and retina, and between retina and eye. (**C**) From the structure correlation matrix in B, the system can infer several comparative matrices showing the expression pattern during the 27 developmental stages of the correlated anatomical components. These comparative matrices highlight the similar expression pattern of Pax6 in central nervous system and eye by filtering by *P*-value.

5 DISCUSSION

aGEM v2.0 is a powerful Platform that integrates several important mouse gene-expression resources. Statistical analyses in aGEM are computed on the gene-expression strength average among all structure–gene pairs fulfilling the user requisites.

The system also offers the possibility to compute the average expression value using recursion through the anatomic ontology by considering gene-expression strength from substructures.

The query by structure also allows the possibility to display the information for the corresponding substructures up to the third level of depth.

The usefulness of the aGEM Platform has been proven with two test cases. The first one corresponds to the third type of analysis explained in the section above, demonstrating, for example, the similar expression pattern of Pax6 in the Central Nervous System and the eye (Fig. 3). This finding is highly consistent with existing literature that describes the important role of Pax6 in the development of both CNS and eye. In humans, Pax6 mutations are associated with aniridia, a congenital abnormality in which there is only a rudimentary iris (Glaser *et al.*, 1994).

The second test case is related to late-onset alzheimer's disease (AD). Neuropathological lesions characteristic of AD consist of amyloid plaques and neurofibrillary tangles in the brain. Plaques are dense, mostly insoluble deposits of amyloid-beta protein and cellular material outside and around neurons, which grow into insoluble twisted fibers within the nerve cells, called tangles.

Table 2. List of genes expressed in the hippocampus with the same expression pattern as Apolipoprotein E (ApoE)

Name	Symbol	Gene ID
activin A receptor, type II-like 1	*Acvrl1*	*MGI:1338946*
Adiponectin receptor 2	**Adipor2**	**MGI:93830**
Atp8a1 gene	**Atp8a1**	**MGI:1330848**
beta 1,3-galactosyltransferase	*B3galt5*	*MGI:2136878*
brain-specific angiogenesis inhibitor 2	**Bai2**	**MGI:2451244**
alpha 1C subunit of the voltage-dependent calcium channel type L	*Cacna1c*	*MGI:103013*
Estrogen-related receptor gamma	**Esrrg**	**MGI:1347056**
G protein-coupled receptor 6	**Gpr6**	**MGI:2155249**
kainate 5 ionotropic receptor	**Grik5**	**MGI:95818**
metabotropic 3 receptor	**Grm3**	**MGI:1351340**
glycogen synthase kinase 3 beta	**Gsk3b**	**MGI:1861437**
myotubularin related protein 2	*Mtmr2*	*MGI:1924366*
protein phosphatase 1F	Ppm1f	MGI:1918464
semaphorin 6C	Sema6c	MGI:1338032
iron-regulated transporter	*Slc40a1*	*MGI:1315204*
cationic amino acid transporter	Slc7a8	MGI:1355323
member 6 of the solute carrier family 9	*Slc9a6*	*MGI:2443511*

Genes reported in the literature as related to AD are highlighted in bold, those related with other diseases and syndromes are in italic, and genes no related neither with AD and nor with other disease are not highlighted.

Researchers have identified an increased risk of developing late-onset AD related to the apolipoprotein E (APOE) gene (Goedert, 1994; Ma *et al*., 1994). APOE protein interacts strongly with amyloid-beta protein, favoring the formation of tangles.

AD usually affects the hippocampus first and most severely and then other parts of the cortex. As an example application of the aGEM Platform to this case, we can ask for retrieval of all genes in the dataset that best match the expression pattern of APOE, which could be considered a marker for the neurodegenerative process of AD in the hippocampus. Seventeen genes were identified with our homologous pattern search. These results are summarized in Table 2 where eight genes whose relationship with AD has been found in the literature (until January 2009) are highlighted in bold. In the other hand, it is necessary to clarify that there could exist some genes related with the disease but not expressed in the same way than ApoE in hippocampus or even expressed in other substructures in the cerebral cortex and therefore not retrieved in this query to aGEM

The nine other genes identified have not previously been connected to AD. However, six of them have been related to other diseases and syndromes (highlighted in italic, Table 2). Whether this new information has any impact on this specific field of AD research remains to be explored; the topic is outside the scope of the present contribution. No information related to disease has been found for the other three genes listed

aGEM doesn't intent to be a gold standard to retrieve genes potentially related with diseases. The examples shown above only try to prove the utility of aGEM and to promote its use in the gene-expression community. The Platform presented here is an integrative tool because data in aGEM has been processed from source databases. Thus, we propose aGEM as a perfect complement to experimental results.

Summarizing, the aGEM Platform v2.0 is a powerful tool in the gene-expression field. It facilitates access to information related to the anatomical pattern of gene expression in the mouse, so it can complement many functional genomics studies. The platform allows gene-expression data to be integrated with spatial-temporal anatomical data by an intuitive and user-friendly display. The statistical analysis provided by aGEM is very useful but it must be interpreted carefully. Gene-expression strength information in the source databases is submitted by the authors or extracted from the literature by database curators. In the first case, the gene-expression annotation is dependent on the personal perception of the author during the submission process; in the second case it relies on the interpretation of the literature by the curators. In this sense, the aGEM Platform cannot go beyond the limits imposed by the source databases, so the statistical analysis relies on their efficient working practices.

The simplicity of the technology used to build the platform (DAS system) allows it to be extended easily. The next versions of the Platform will expand the range of spatial gene-expression databases being integrated, such as GenePaint (Visel *et al*., 2004), the Mouse Tumor Biology Database (Begley *et al*., 2007), and the 'Electronic Atlas of the Developing Human Brain' (a joint project between the Institute of Human Genetics in Newcastle and the MRC Human Genetics Unit in Edinburgh). The visual interface will also be improved, offering the user the possibility of interacting with 3D models of the organism under study and directly selecting an anatomical structure simply by clicking on it. Interaction with medical image data, specifically in the context of small animal scanners, will also be considered, providing a link between the gene-expression information in animal models and the anatomical information obtained by Positron Emission Tomography (PET) and Computer Tomography (CT). In the meantime, the recently-proposed mapping between human and mouse ontologies (Kruger *et al*., 2007) could allow gene-expression data to be extrapolated from mouse to human.

ACKNOWLEDGEMENTS

Prof. R. Baldock and EMAP team are acknowledged for constant support of the present work. We sincerely thank to the staff of the EMAGE, GXD, GENSAT and OMIM databases for free database access and maintenance and Prof. S Lindsay from Newcastle University for manuscript supervision and for contributing to this article with interesting ideas.

Funding: National Institute for Bioinformatics (http://www.inab.org), a platform of 'Genoma España'; Industrial and Technological Development Centre (CDTI) under the CENIT Programme (CDTEAM Project); collaborative research project with Integromics SL; EU EMBRACE network
of excellence (LHSG-CT-2004-512092).

Conflict of Interest: none declared.

REFERENCES

Baldock,R.A. *et al*. (2003) EMAP and EMAGE: a framework for understanding spatially organized data. *Neuroinformatics*, **1**, 309–325.

Baldock,R. and Burger,A. (2005) Anatomical ontologies: names and places in biology. *Genome Biol*., **6**, 108.

Begley,D.A. *et al*. (2007) Mouse Tumor Biology Database (MTB): status update and future directions. *Nucleic Acids Res*., **35**, D638–D642.

Burger,A. *et al.* (2004) Formalization of mouse embryo anatomy. *Bioinformatics*, **20**, 259–267.

Cuticchia,A.J. (2000) Future vision of the GDB human genome database. *Hum.* Mutant., **15**, 62–67.

Dowell,R.D. *et al*. (2001) The distributed annotation system. BMC Bioinformatics 2001, **2**, 7.

Finn,R.D. *et al*. (2007) ProServer: a simple, extensible Perl DAS server. *Bioinformatics*, **23**, 1568–1570. Available at: http://www.sanger.ac.uk/Software/analysis/proserver/

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*., **5**, R80.

Glaser,T. *et al*. (1994) PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects. *Nat. Genet*., **7**, 463–471.

Goedert,M. *et al*. (1994) Alzheimer's disease. Risky apolipoprotein in brain. *Nature*, **373**, 45–46.

Gong,S. *et al.* (2003) A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature*, **425**, 917–925.

Hamosh,A. *et al*. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.

Hayamizu,T.F. *et al.* (2005) The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biol*., **6**, R29.

Haudry,Y. *et al.* (2008) 4DXpress: a database for cross-species expression pattern comparisons. *Nucleic Acids Res*., **36**, D847.

Hubbard,T.J. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.

Jones,P. *et al.* (2005) a perfect pair for protein feature visualization. *Bioinformatics*, **21**, 3198–3199. Available at http:www.ebi.ac.uk/dasty/ (last accessed date July 31, 2009).

Kruger,A. *et al.* (2007) Simplified ontologies allowing comparison of developmental mammalian gene expression. *Genome Biol*., **8**, R229.

Letovsky,S.I. *et al.* (1998) GDB: the human genome database. *Nucleic Acids Res*., **26**, 94–99.

Ma,J. *et al.* (1994) Amyloid-associated proteins alpha 1-antichymotrypsin and apolipoprotein E promote assembly of Alzheimer beta-protein into filaments. *Nature*, **372**, 92–94.

Macias,J.R. *et al.* (2007) Integrating electron microscopy information into existing Distributed Annotation Systems. *J. Struct. Biol*., **158**, 205–213. Available at http://biocomp.cnb.csic.es/das (last accessed date July 31, 2009).

Prlic,A. *et al.* (2005) Adding some SPICE to DAS. *Bioinformatics*, **2**, 40–41. Available at http://www.spice-3d.org (last accessed date July 31, 2009).

Ringwald,M. *et al.* (1994) A database for mouse development. *Science*, **265**, 2033–2034.

Ringwald,M. *et al.* (1999) GXD: a gene expression database for the laboratory mouse. The Gene Expression Database Group. *Nucleic Acids Res*., **27**, 106–112.

Shoman,L.M. *et al.* (1995) The Worm Community System, release 2.0 (WCSr2). *Methods Cell Biol*., **48**, 607–625.

Skupski,M.P. *et al.* (1999) The Genome Sequence DataBase: towards an integrated functional genomics resource. *Nucleic Acids Res*., **27**, 35–38

ten Berge,D. *et al.* (1998) Mouse Alx3: an aristaless-like homeobox gene expressed during embryogenesis in ectomesenchyme and lateral plate mesoderm. *Dev. Biol*., **199**, 11–25

Tamames,J. and Valencia,A. (2006) The success (or not) of HUGO nomenclature. *Genome Biol*., **7**, 402.

Theiler,K. (1989) *The House Mouse: Atlas of Embryonic Development.* Springer Verlag, Berlin.

The UniProt Consortium. (2009) The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*., **37**: D169–D174.

Venkataraman,S. *et al.* (2008) EMAGE: Edinburgh Mouse Atlas of Gene Expression: 2008 update. *Nucleic Acids Res*., **36**:D860–D865.

Visel,A. *et al.* (2004) GenePaint.org: an atlas of gene expression patterns in the mouse embryo. Nucleic Acid Res., **32**, D552–D556.

Wheeler,D.L. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **36**, D13–D21.