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 (IMAGE, 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.cnbc.csic.es

Contact: natalia@cnb.csic.es


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 (GXD). 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
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 expands 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 anatomical coordinates of its expression. 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 (Márias 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.cnbc.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 status 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 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., 2004), 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:


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

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 Thiel stage 14, EMAP3056 which corresponds to ‘heart’ in Thiel stage 18 and EMAP7570 that is ‘hindbrain’ in Thiel stage 23:


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:


of aGEM, it is convenient to introduce some notation. From numerical values for each database (Table 1). These different schemas by converting the different text scores into expression strength in different ways. In this work we have unified considering that EMAGE, GXD and GENSAT evaluate gene-combined into an ‘expression strength’ value, not an easy task strength information from the different databases is stored and GENSAT has its own vocabulary, so we have mapped it to the by EMAGE and GXD, this approach is straightforward, since databases integrated in the Platform use the same gene identifiers expressed (gene-expression strength).

The query by gene identifier (gene ID) is answered by the aGEM Platform explicitly considers the different levels of 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 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.

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 \( [g, r, t, v] \), will be referred to as a score item. The set of score items is represented as: \( \{ [g, r, t, v] | g \in G, r \in R, t \in \{1, \ldots, 27\}, v \in \{0, 7\} \} \). 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, the gene-expression level between 0 = not expressed to 7 = highly expressed.

The aGEM Platform can perform three types of statistical analysis. The first type of analysis fixes the anatomical structure, \( s_t \in S \) and \( \{ [g, s_t, r, v] | g \in G, r \in R, t \in \{1, \ldots, 27\}, v \in \{0, 7\} \} \), 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 correlation matrix in which genes with similar expression patterns (strongly co-expressed) are represented by red 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 \( t_v \in \{1, \ldots, 27\} \) is fixed, generating a view showing the expression strength (represented as a color code

<table>
<thead>
<tr>
<th>Table 1. Expression strength scores in the different databases and their corresponding values in aGEM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GXD</strong></td>
</tr>
<tr>
<td>Absent</td>
</tr>
<tr>
<td>0 1 2 3 4 5 6 7</td>
</tr>
<tr>
<td><strong>EMAGE</strong></td>
</tr>
<tr>
<td>Not detected</td>
</tr>
<tr>
<td>0 1.4 2.8 4.2 5.6 7</td>
</tr>
<tr>
<td><strong>GENSAT</strong></td>
</tr>
<tr>
<td>Undetectable</td>
</tr>
<tr>
<td>0 2.3 5.75</td>
</tr>
</tbody>
</table>
The white color on the vertical bar represents the lack of available data; the matrix is displayed following a color code gradient (below the matrix).

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.

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.
ApoE in hippocampus or even expressed in other substructures in AD in the hippocampus. Seventeen genes were identified with our Atp8a1 gene

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>activin A receptor type II-like 1</td>
<td>Acvr1l</td>
<td>MGI:1338946</td>
</tr>
<tr>
<td>Adiponectin receptor 2</td>
<td>Adipor2</td>
<td>MGI:93830</td>
</tr>
<tr>
<td>alpha 1C subunit of the voltage-dependent calcium channel type L</td>
<td>Cacna1c</td>
<td>MGI:103013</td>
</tr>
<tr>
<td>Estrogen-related receptor gamma</td>
<td>Esrrg</td>
<td>MGI:1347056</td>
</tr>
<tr>
<td>G protein-coupled receptor 6</td>
<td>Gpr6</td>
<td>MGI:2155249</td>
</tr>
<tr>
<td>kainate 5 ionotropic receptor</td>
<td>Grik5</td>
<td>MGI:95818</td>
</tr>
<tr>
<td>metabotropic receptor</td>
<td>Gmra</td>
<td>MGI:1351340</td>
</tr>
<tr>
<td>glycoprotein synthase kinase 3 beta</td>
<td>Gsk3b</td>
<td>MGI:1861437</td>
</tr>
<tr>
<td>myotubularin related protein 2</td>
<td>Mtn2</td>
<td>MGI:1924366</td>
</tr>
<tr>
<td>protein phosphatase 1F</td>
<td>Ppm1</td>
<td>MGI:1918464</td>
</tr>
<tr>
<td>semaphorin 6C</td>
<td>Sema6c</td>
<td>MGI:1338032</td>
</tr>
<tr>
<td>cationic amino acid transporter</td>
<td>Slc9a6</td>
<td>MGI:2443511</td>
</tr>
<tr>
<td>member 6 of the solute carrier family 9</td>
<td>Slc9a6</td>
<td>MGI:1355323</td>
</tr>
</tbody>
</table>

ApoE gene

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

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. The other three genes listed

<table>
<thead>
<tr>
<th>Name</th>
<th>Symbol</th>
<th>Gene ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>activin A receptor type II-like 1</td>
<td>Acvr1l</td>
<td>MGI:1338946</td>
</tr>
<tr>
<td>Adiponectin receptor 2</td>
<td>Adipor2</td>
<td>MGI:93830</td>
</tr>
<tr>
<td>alpha 1C subunit of the voltage-dependent calcium channel type L</td>
<td>Cacna1c</td>
<td>MGI:103013</td>
</tr>
<tr>
<td>Estrogen-related receptor gamma</td>
<td>Esrrg</td>
<td>MGI:1347056</td>
</tr>
<tr>
<td>G protein-coupled receptor 6</td>
<td>Gpr6</td>
<td>MGI:2155249</td>
</tr>
<tr>
<td>kainate 5 ionotropic receptor</td>
<td>Grik5</td>
<td>MGI:95818</td>
</tr>
<tr>
<td>metabotropic receptor</td>
<td>Gmra</td>
<td>MGI:1351340</td>
</tr>
<tr>
<td>glycoprotein synthase kinase 3 beta</td>
<td>Gsk3b</td>
<td>MGI:1861437</td>
</tr>
<tr>
<td>myotubularin related protein 2</td>
<td>Mtn2</td>
<td>MGI:1924366</td>
</tr>
<tr>
<td>protein phosphatase 1F</td>
<td>Ppm1</td>
<td>MGI:1918464</td>
</tr>
<tr>
<td>semaphorin 6C</td>
<td>Sema6c</td>
<td>MGI:1338032</td>
</tr>
<tr>
<td>cationic amino acid transporter</td>
<td>Slc9a6</td>
<td>MGI:2443511</td>
</tr>
<tr>
<td>member 6 of the solute carrier family 9</td>
<td>Slc9a6</td>
<td>MGI:1355323</td>
</tr>
</tbody>
</table>

ApoE gene

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

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 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 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.

### 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.

### REFERENCES

N. Jiménez-Lozano et al.


