COMPARING AND ANALYSING GENE EXPRESSION PATTERNS ACROSS ANIMAL SPECIES USING 4D*XPRESS*

YANNICK HAUDRY
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

CHUANG KEE ONG
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

LAURENCE ETTWILLER
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

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

IVICA LETUNIC
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

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

PAUL-DANIEL WEEBER
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

XI WANG
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

JULIEN GAGNEUR
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

CHARLES GIRARDOT
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

DETLEV ARENDT
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

PEER BORK
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

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

EILEEN FURLONG
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

JOACHIM WITTBRODT
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

THORSTEN HENRICH[†]
European Molecular Biology Laboratory EMBL, Meyerhofstrasse 1
69117 Heidelberg, Germany

High-resolution spatial information on gene expression over time can be acquired through whole mount in-situ hybridisation experiments in animal model species such as fish, fly or mouse. Expression patterns of many genes have been studied and data has been integrated into dedicated model organism databases like ZFIN for zebrafish, MEPD for medaka, BDGP for drosophila or MGI for mouse. Nevertheless, a central repository that allows users to query and compare gene expression patterns across different species has not yet been established. For this purpose we have integrated gene expression data for zebrafish, medaka, drosophila and mouse into a central public repository named 4D*Xpress* (http://ani.embl.de/4DXpress). 4D*Xpress* allows to query anatomy ontology based expression annotations across species and quickly jump from one gene to the orthologs in other species based on ensembl-compara relationships. We have set up a linked resource for microarray data at ArrayExpress. In addition we have mapped developmental stages between the species to be able to compare corresponding developmental time phases. We have used clustering algorithms to classify genes based on their expression pattern annotations. To illustrate the use of 4D*Xpress* we systematically analysed the relationships between conserved regulatory inputs and spatio-temporal gene expression derived from 4D*Xpress* and found significant correlation between expression patterns of genes predicted to have similar regulatory elements in their promoters.

[†] corresponding author

Introduction

Embryonic development is a process in which cells signal to each other and thereby acquire different identities which is necessary to establish the basic body plan of the organism. This process results in amazingly complex gene expression patterns that can be visualised by whole mount in situ hybridisation experiments. Whoever has seen expression patterns of typical developmental regulators like FGF, HOX, or PAX genes will understand that the spatial regulation of such genes can not be analysed by microarray experiments.

To know the exact time and location of gene transcripts is crucial when studying the functions of genes involved in development as well as for trying to decipher the code of cis-regulatory modules. It is important to store images, which lets biologists see and judge expression together with an organized annotation. Ontology based annotations let users query the data and make data accessible to computational analysis.

Expression localisation data has been gathered in the model species databases but a central resource, which allows users to compare gene expressions in different species, has not yet been established until recently. We have set up such a platform for cross species expression pattern comparisons (1), comprising annotations on 16505 genes. This is the largest collection available to date. Our vision is that in a few years time the exact localisation of each single transcript will be known for the major model species. We hope that our resource will help to store them in an organised way, to compare different species expression patterns and to provide tools to analyse this data. We show that the organised storage of expression annotation data is sufficient to classify genes into clusters of similar expressed genes and thereby offers an entry point for cross species comparisons through computational biology.

As an example of such an approach we present an application of deciphering the code of cis-regulatory modules. In this context we take advantage of the information stored in 4D*Xpress* and analyse the correlation between regulatory input and the spatio-temporal expression pattern. More specifically, we systematically investigated whether a significant correlation between expression annotation and the occurrence of at least one common conserved transcription factor (TF) binding site exists. Using binding site information for 309 TFs we found a significant correlation for the predicted target genes of 4 TFs. This demonstrates that in some cases, genes predicted to be the common targets of at least one transcription factor have similar pattern of expressions.

4D*Xpress*

We have integrated gene expression data for drosophila (2), medaka (3), zebrafish (4) and mouse (5) so far. In table 1 we give an overview on the gene expression patterns that have been integrated for 4D*Xpress*. The best-annotated model species are drosophila and zebrafish at the moment with almost 6000 annotated genes each. Mouse has slightly less annotated genes reflecting the difficulty to yield large amounts of specimens when compared to egg-laying species like fish and fly. The annotated mouse genes represent a

large set of important developmental regulators and are annotated in great detail sometimes using a 3D virtual embryo (6).

	Source	Genes	Stages	Stages per Gene	Anatomy Terms	Anat. Terms per Gene	Anat. Terms per Stage	Distinct Anat. Terms
drosophila	bdgp	5951	21048	3.54	29867	5.02	1.42	288
medaka	mepd	882	2746	3.11	5047	5.72	1.84	338
zebrafish	zfin	5779	102671	17.77	178851	30.95	1.74	694
mouse	mgi	3893	12799	3.29	17291	4.44	1.35	1661
		16505	139264	8.44	231056	14.00	1.66	2981

Table 1: Content of 4D*Xpress*. Annotation status of gene expression patterns at present time.

Expression data has been gathered in different ways. For drosophila and medaka the major annotation results from a screen. Expression has been analysed at 3 or 4 distinct time points (table 1, stages per gene), whereas zebrafish expression patterns are additionally annotated from literature by a team of database curators. Annotation is done for continues developmental stages.

Anatomy ontologies are often huge, however only a limited fraction of the terms is actually used for expression annotation (table 1, distinct annotations). Again, ZFIN uses a rich vocabulary with almost 700 distinct terms. The values for mouse and medaka need to be treated with care, as the ontologies used for annotation here are the cross product of anatomy and stage ontologies and therefore overestimates vocabulary richness.

The web application is based on a MVC (Model-View-Controller) architecture using the Struts Framework, and enhanced with applets, JavaScript and AJAX (Asynchronous JavaScript and XML) technologies to build a powerful, interactive, user-friendly interface. 4D*Xpress* is available at http://ani.embl.de/4DXpress. The usage of the interface is documented in detail on our home page. Genes can be searched either by a range of external identifiers, symbol, name or by their expression pattern annotation. Ontologies that were used to annotate gene expression can be browsed. Anatomy ontologies are browse-able by a tree-based tool, which allows users to query terms and expand and collapse individual nodes. Developmental stage anatomies can be browsed by species and external links provide more information on stage definitions. Species-specific stage ontologies were mapped onto a common stage list and thereby establish temporal relationships, which can be accessed via web interface. Our annotation tool allows users to annotate gene expression patterns resulting from either whole mount in situ hybridisation experiments, transgenic reporter gene expression or antibody staining. The same tool can be used for all supported species (for now: zebrafish, mouse, medaka, drosophila, platynereis).

Expression data acquired through either in situ hybridisation, antibody or transgenic expression and microarrays can complement each other. The first provides high-resolution data; the latter can quickly give an overview on all genes in a genome. That is why we have decided to set up a data warehouse for microarray data at EBI: 4D

ArrayExpress Data Warehouse (4DDW). It is accessible at: http://www.ebi.ac.uk/microarray-as/4DDW_EMBL/. So far we have established 4737 reciprocal links for mouse, drosophila and zebrafish. When querying microarray data at ArrayExpress (7) users can quickly go to 4D*Xpress* and vice versa.

The 4D*Xpress* schema is based on the common MISFISHIE (8) standard allowing a straightforward integration of additional species and data exchange with other databases.

Cross Species Relationships

One of the major goals of our project is to be able to compare gene expression patterns between the different model species. For doing so relationships need to be established between genes (orthology), between time windows (developmental stages) and most challenging between anatomical structures.

Orthology

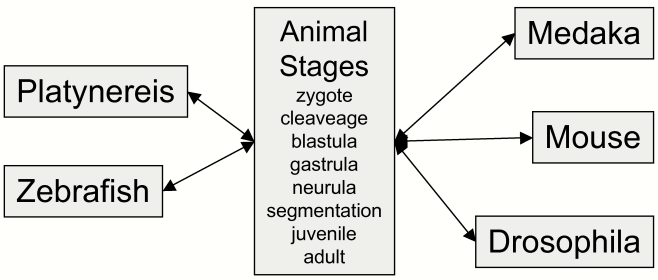
EnsEMBL compara (9) provides a reliable source of sequence homology relationships, which was computed using a tree-based approach. We have chosen to use this and update regularly upon new EnsEMBL releases. We assigned each gene to a cluster of orthologs using the one2one-, one2many and many2many orthology relationships. These clusters are visualised as a network in the gene view and used to sort the gene list retrieved from a query.

Developmental Stages

It is very difficult to identify corresponding developmental stages in two species, even when comparing two closely related fish like medaka and zebrafish. It is almost impossible to find an exact corresponding stage for one of the 46 medaka stages in zebrafish, because within the embryo different structures develop with different speed. E.g. the head and brain develops faster, whereas the tail and somites develops slower in medaka. So if one finds a matching stage in zebrafish regarding the number of somites, which is a very popular staging feature, the head would actually correspond to a later zebrafish stage.

However there is a list of 8 embryonic stages that is described in all developmental biology text books and is common to all bilaterian animals: zygote, cleavage, blastula, gastrula, neurula, organogenesis, juvenile and adult. By mapping each of the species stages onto one of the bilaterian stages the link between species stages can be done and combinatorial explosion can be prevented. A new species will only need to be mapped to the common stages (fig. 1) and not against all stages of all other species.

Fig 1: Mapping of developmental species was done through a list of stages common to all bilaterian animals.



Anatomical Structures

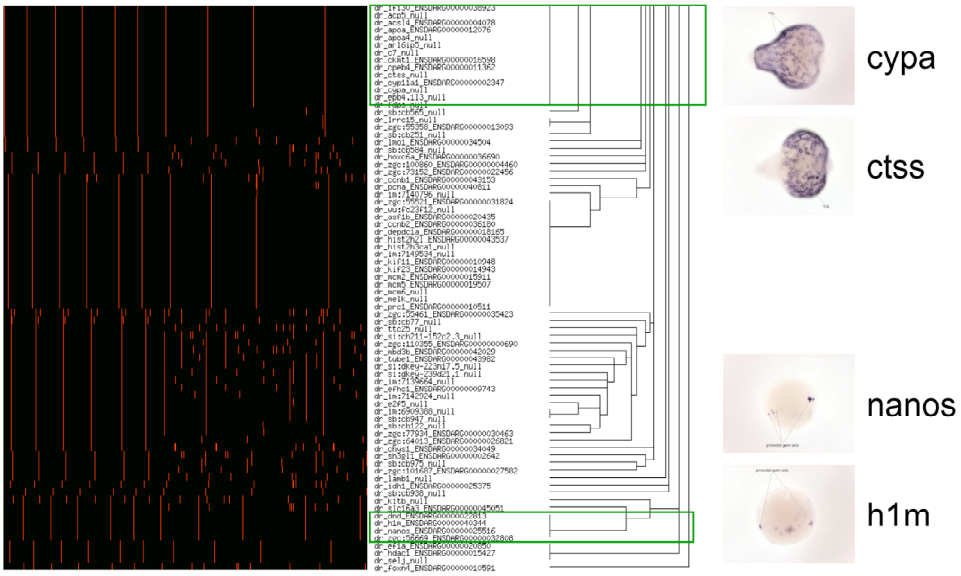
We have used lexical cues to start a simple anatomical mapping. 50% of all unique terms used for annotations could be mapped to high-level terms common to all species. Anatomy structure and co-expression cues will be used to refine these relationships. For co-expression we will use the integrated expression data of 4D*Xpress* as this data has the best spatial resolution.

The Common Anatomy Reference Ontology (CARO) is being developed to facilitate interoperability between existing anatomy ontologies for different species. It aims to provide a template for building new anatomy ontologies. We want to use CARO to build an anatomy ontology shared by all bilaterians. Similar to the stage mapping we then want to map species-specific anatomy terms onto the common ontology.

Co-Expression Analysis

Having gene expression annotation available in an organised way allows us to analyse gene expression annotation computationally. Using the same clustering tools as used for micro array analysis such as the TIGR Multi Experiment Viewer (10) simple hierarchical clustering can identify genes with similar expression patterns (fig 2).

Fig 2: Genes can be classified upon their expression pattern annotation. Genes were clustered using the binary distance and a hierarchical clustering algorithm with the TIGR MEV package.



The method can be validated by looking at a few examples of genes that have been clustered together and by examining their in situ images. Indeed simple expression

patterns such as cypa and ctss or nanos and h1m would have also been described as similar by researchers.

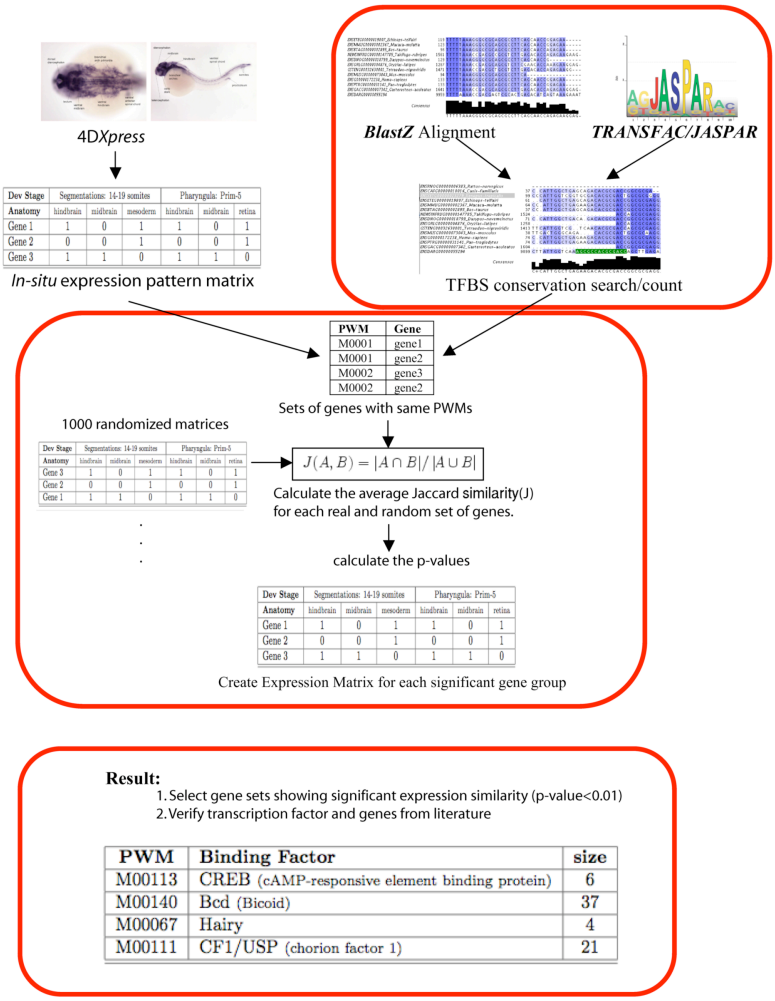
More complex expression patterns like those of developmental regulators do not cluster well with other genes though. However there is still room for improvements. I.e. the ontology relations have not yet been exploited by the clustering algorithm. Semantic similarities measures are able to account for that. We will apply this method on gene expression annotation.

When comparing genes across species we need maximum overlap. Zebrafish and drosophila are the most complete data sets. We have 964 ortholog groups annotated in the two species and 336 for three species (including mouse). They also overlap temporally in the developmental stages: blastula, gastrula, neurula, organogenesis and juvenile.

We will use semantic similarities to generate co-expression networks for each species individually and then use orthology relationships to identify conserved patterns in these networks. Looking at the terms used in different species for annotating conserved patterns we hope to find candidates for the cross species anatomy mapping. In addition it will be interesting to study the regulatory sequences of genes appearing in the same conserved network pattern (see below).

Application

To demonstrate the value of 4D*Xpress* for computational approaches, we analysed the correlation between regulatory inputs of genes and their corresponding expression patterns derived from the zebrafish in-situ annotation taken from 4D*Xpress*. In a first step, the human target genes of all transcription factors with known binding sites in TRANSFAC (11) were predicted using a similar approach as described previously (12) (fig. 3). This approach is based on extreme conservation of the binding site from human to fish, only the predicted most conserved target genes for each transcription factor are further selected. The corresponding ortholog genes in zebrafish are then retrieved and the in-situ expression pattern information is mapped to the predicted target genes.



Result:

1. Select gene sets showing significant expression similarity (p-value<0.01)
2. Verify transcription factor and genes from literature

PWM	Binding Factor	size
M00113	CREB (cAMP-responsive element binding protein)	6
M00140	Bcd (Bicoid)	37
M00067	Hairy	4
M00111	CF1/USP (chorion factor 1)	21

Figure 3: Analysis Pipeline for correlating expression and binding site occurrence.

For each target gene set, the average expression distance is computed using the Jaccard metric. In order to evaluate if the average distance is significantly lower than expected under a random model, randomizations of the expression matrix is done by shuffling the gene IDs and the random distances are computed and compare to the real distance. Out of 309 predicted target gene sets, 4 show a significantly lower ($p<0.01$) average expression distances (fig. 3). This list includes the transcription factor Hes1 where all the predicted target genes with expression information (Her6, Atoh1, ide2a and Neurog1) are expressed in the diencephalon, hindbrain, midbrain and telencephalon of the developing zebrafish embryo. This result demonstrates that in situ data as stored in 4D*Xpress* can be used for the identification of new regulatory sequences in the genome.

Future Perspective

Now that we have a stable database schema and a efficient web interface to access data, in the future we will focus on three points:

1. Integrating more data and species
   There is more gene expression pattern annotation available in the public domain. The next species we are aiming for is Xenopus laevis with 17.000 images (Naoto Ueno, NIBB in Okazaki), Ciona intestinalis (7000 genes) and C. elegans.
2. Developing tools for data analysis
   We will calculate distances between genes based on their expression annotation vector. This will enable users to easily find genes with similar expression patterns as the gene of interest. We are planning to establish tools integrated in the web interface, that allows users to cluster expression data an correlate clusters of genes with other sources of data like chromosomal location, GO, KEGG or occurrence of binding sites.
3. Mapping Anatomy Ontologies
   Our strategy is described above. Establishing such a resource is as challenging as it will be valuable when achieved.

Acknowledgement
This work was carried out in the Centre for Computational Biology at EMBL. We are grateful to the model organism database crews for providing us their expression pattern data. In particular we thank Martin Ringwald and Susan McClatchy to give us direct access to the MGI database; thanks to Monte Westerfield and Judy Sprague for helping us with the ZFIN data reports; we thank Pavel Tomancak for helping us to understand the BDGP mySQL schema. We thank Mirana Ramialison for extending the medaka annotation and Francois Spitz for mouse in situ images. We are grateful to the MISFISHIE team for initiating an expression pattern data exchange format.

References

1. Haudry, Y., Berube, H., Letunic, I., Weeber, P., Gaugneur, J., Girardot, C., Kapushesky, M., Arendt, D., Bork, P., Brazma, A. *et al.* (in press NAR 2008) 4DXpress: A database for cross species expression pattern comparisons. *Nucleic Acids Res*.
2. Tomancak, P., Berman, B.P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S.E. and Rubin, G.M. (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*, **8**, R145.
3. Henrich, T., Ramialison, M., Wittbrodt, B., Assouline, B., Bourrat, F., Berger, A., Himmelbauer, H., Sasaki, T., Shimizu, N., Westerfield, M. *et al.* (2005) MEPD: a resource for medaka gene expression patterns. *Bioinformatics*, **21**, 3195-3197.
4. Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D.G., Mani, P., Ramachandran, S. *et al.* (2006) The

Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res*, **34**, D581-585.

5. Smith, C.M., Finger, J.H., Hayamizu, T.F., McCright, I.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. and Ringwald, M. (2007) The mouse Gene Expression Database (GXD): 2007 update. *Nucleic Acids Res*, **35**, D618-623.
6. Christiansen, J.H., Yang, Y., Venkataraman, S., Richardson, L., Stevenson, P., Burton, N., Baldock, R.A. and Davidson, D.R. (2006) EMAGE: a spatial database of gene expression patterns during mouse embryo development. *Nucleic Acids Res*, **34**, D637-641.
7. Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M. *et al.* (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, **35**, D747-750.
8. Deutsch, E.W., Ball, C.A., Bova, G.S., Brazma, A., Bumgarner, R.E., Campbell, D., Causton, H.C., Christiansen, J., Davidson, D., Eichner, L.J. *et al.* (2006) Development of the Minimum Information Specification for In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE). *Omics*, **10**, 205-208.
9. Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res*, **35**, D610-617.
10. Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374-378.
11. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, **29**, 281-283.
12. Del Bene, F., Ettwiller, L., Skowronska-Krawczyk, D., Baier, H., Matter, J.-M., Birney, E. and Wittbrodt, J. (2007) In vivo Validation of a Computationally Predicted Conserved Ath5 Target Gene Set. *Plos Genetics*.