BIOINFORMATICS APPLICATIONS NOTE
Vol. 21 no. 14 2005, pages 3195–3197
doi:10.1093/bioinformatics/bti478

Databases and ontologies

MEPD: a resource for medaka gene expression patterns

Thorsten Henrich[1,*,†], Mirana Ramialison[1,†], Beate Wittbrodt[1], Beatrice Assouline[1,2], Franck Bourrat[3], Anja Berger[4], Heinz Himmelbauer[4], Takashi Sasaki[5], Nobuyoshi Shimizu[5], Monte Westerfield[6], Hisato Kondoh[7] and Joachim Wittbrodt[1]

[1]EMBL Heidelberg, Meyerhofstrasse 1, D-69117 Heidelberg, Germany, [2]INSERM E363, Faculty of Medicine Necker, 75730 Paris Cedex 15, France, [3]UPR CNRS 2197, Institut de Neurobiologie A. Fessard, 91198, Gif/Yvette, France, [4]Max-Planck-Institut für Molekulare Genetik, D-14195 Berlin-Dahlem, Germany, [5]Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan, [6]ZFIN, University of Oregon, Eugene, OR 97403-5274, USA and [7]Japan Science and Technology Agency, ERATO Kondoh Differentiation Signaling Project/SORST Kondoh Research Group, Kinki-chihou Hatsumei Center Building, Yoshida-Kawaracho 14, Sakyo-ku, Kyoto 606-8305, Japan

Received on February 4, 2005; revised on April 13, 2005; accepted on April 28, 2005
Advance Access publication May 6, 2005

ABSTRACT

Summary: The Medaka Expression Pattern Database (MEPD) is a database for gene expression patterns determined by *in situ* hybridization in the small freshwater fish medaka (*Oryzias latipes*). Data have been collected from various research groups and MEPD is developing into a central expression pattern depository within the medaka community. Gene expression patterns are described by images and terms of a detailed medaka anatomy ontology of over 4000 terms, which we have developed for this purpose and submitted to Open Biological Ontologies. Sequences have been annotated via BLAST match results and using Gene Ontology terms. These new features will facilitate data analyses using bioinformatics approaches and allow cross-species comparisons of gene expression patterns. Presently, MEPD has 19 757 entries, for 1024 of them the expression pattern has been determined.

Availability: A new version has been implemented at the EMBL in Heidelberg and is now accessible at http://www.embl.de/mepd

Contact: henrich@embl.de

Supplementary information: http://www.embl.de/mepd/supplement.html

INTRODUCTION

Medaka has developed into a standard developmental model organism over the past years (Schartl *et al*., 2004; Wittbrodt *et al*., 2002). Sequencing of Linkage Group 22 has nearly been completed [in the Keio group (N. Shimizu) using a new clone by clone (CBC) strategy on bacterial artificial chromosome clones combined with a whole genome shotgun approach (http://medaka.dsp.jst.go.jp/MGI/LG22/)] and a first draft of the genome sequence has been made publicly available by the Kohara project (Medaka Genome Project; http://dolphin.lab.nig.ac.jp/medaka/) last year.

Thanks to robotics, many laboratories can now afford a high-throughput *in situ* hybridization approach and expression pattern data are accumulating (Henrich and Wittbrodt, 2000; Nguyen *et al*., 2001; Quiring *et al*., 2004). The central collection and accessibility of these data, as provided by the databases for major model systems [e.g. ZFIN for zebrafish (Sprague *et al*., 2003) or FlyBase for *Drosophila* (FlyBase Consortium, 2003)] are of crucial importance to the efficient use and exchange of the resources and to bioinformatics approaches towards the evolution of gene expression and function. Ontologies facilitate bioinformatics approaches and interspecies comparisons, and hence are widely used in biological databases (Bard, 2003).

We developed new ontologies for anatomy, developmental stages and phenotypes in close collaboration with the zebrafish experts, which we provide to the scientific community. In MEPD (Henrich *et al*., 2003), we translated the description of expression patterns with keywords into an ontology-based description. Likewise, sequences have been annotated using Gene Ontology (GO) terms. We collect data from diverse sources (from marker gene collections as well as from high-throughput screens) resulting in an extensive increase of entries (∼20 000 compared with 711 in the first release). Within the Medaka Genome Initiative (Shima *et al*., 2003) we established MEPD as a central gene expression pattern depository for our model organism.

BASIC DATA UPDATE

Data submission to MEPD is an ongoing progress. At the time of writing, we have 19 757 sequence entries and expression data on 1024 expressed sequence tags (ESTs) (representing 623 genes) described by 3863 images (at various developmental stages; whole mount as well as sections) or anatomy ontology. MEPD contains gene expression information on ESTs from different screens: high-throughput analysis of an embryonic library '631' [767 clones (Quiring *et al*., 2004)], *in situ* screen in the developing tectum 'Hd' [113 clones (Nguyen *et al*., 2001)] as well as cloned marker gene set, which has been collected over years and contains important genes involved in developmental processes (e.g. Pax, Six, Fgf, Bmp and Wnt; 48 clones, Wittbrodt J, unpublished data). In addition, MEPD contains sequence entries of a Medaka unigene library ('McF', described below) and gene expression analysis of clones of this library has been started (96 clones).

[*]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.

Medaka unigene library

Within the Medaka Genome Initiative (Shima *et al*., 2003) a unigene library was established as a shared resource, to be described in detail elsewhere (Berger A., Hennig S., Sasaki T., Furutani-Seiki M., Kondoh H., Mitani H., Shima A., Janitz M., Herwig R., Lehrach H., Wittbrodt J., Shimizu N. and Himmelbauer H., manuscript in preparation). Briefly, libraries representing different embryonic stages of the medaka (gastrula, neurula and organogenesis) and medaka ovary, all from Cab-strain, were constructed by Makoto Furutani-Seiki (Japan Science and Technology Corporation, Kyoto), arrayed and normalized using the oligonucleotide fingerprinting technology (OFP) (Meier-Ewert *et al*., 1998) by A.Berger and H.Himmelbauer and sequenced by T.Sasaki and N.Shimizu. OFP is insensitive to the presence of related repeats in transcripts from unrelated genes, because fingerprints are calculated on the basis of entire clone inserts. Also, clustering parameters can be adjusted to distinguish members of multigene families (Fuchs *et al*., 2002). In MEPD, we have 18 364 sequence entries of this library, 10 461 of which represent the unigene set. Those have been re-arrayed and systematic high-throughput *in situ* hybridization analysis has been started.

EXPRESSION PATTERN ANNOTATION

To enable cross-species comparisons of gene expression patterns and to facilitate data analysis, we developed three different ontologies, hierarchical sets of developmental terms, anatomical terms and phenotypic terms, and submitted them to ‘Open Biological Ontologies’ (OBO) (http://obo.sourceforge.net/) where it can be downloaded in DAGedit format.

Each term has a unique identifier: MFO (Medaka Fish Ontology). We annotate individual database records by defining an intersection of the developmental and anatomical terms. The phenotypic terms are used to describe mutant phenotypes (Henrich *et al*., 2004).

The anatomy ontology has 4173 terms for the 46 developmental stages. The graphs were developed using DAGedit (http://www.geneontology.org/doc/GO.tools.html) and self-written tools. As a source, we used the anatomical description of Medaka embryos in the developmental staging paper by Iwamatsu (1994) and compared it with the detailed zebrafish anatomy ontology (M.Westerfield). In a second step during expression pattern annotation, we considered the feedback of fish anatomy experts (F.Bourrat, B.Assouline and J.Wittbrodt) for improving this ontology.

During production of these ontologies, we took great care to use the same terms for corresponding structures in Medaka and Zebrafish, and, wherever possible, in mouse and *Drosophila*. We have also developed a set of translation tables that define the relationships among terms that differ among these species.

SEQUENCE ANNOTATION

We used publicly available and self-written tools to process and annotate the Medaka cDNA sequences. The sequence processing flowchart, the parameters used for the decision tree and the programs are described in detail on the web page (http://www.embl-heidelberg.de/mepd/chart.gif).

Sequence filtering and cluster analysis

Prior to analysis, the EST sequences were trimmed to remove vectors and poly(A) tails using the TIGR Gene Indices Sequence Cleaning and Validation script (SeqClean), which was primarily designed to clean EST databases (http://www.tigr.org./tdb/tgi/software), and sequences <100 bp were discarded. Out of 19 757 entries 19 359 were processed for sequence clustering.

In order to relate expression data to genes rather than ESTs, a cluster analysis was performed using the TIGR gene indices clustering tools (Pertea *et al*., 2003). The 19 359 ESTs cluster into 10 474 genes: 3245 of them derived from singletons and the remaining fell into 7229 clusters comprising 2 to 129 ESTs.

Homology searches

Sequence similarity searches (blastn, tblastx and blastx) against public databases are performed regularly. The public databases are stored locally at the EMBL and are updated on a daily basis. We blast against the nucleotide (EMBL + GenBank + DDBJ, without ESTs or STSs), expressed sequence tag (EMBL + GenBank + DDBJ EST divisions) and protein (EMBL CDS translations + PDB + Swiss-Prot + PIR) databases.

In addition, we blasted the MEPD entries against all known zebrafish EnsEMBL proteins to establish links to the best hits in EnsEMBL (Birney *et al*., 2004) as well as to ZFIN (Sprague *et al*., 2003). This allows a direct comparison of the expression patterns of medaka and zebrafish orthologues.

Functional annotation using Gene Ontology Annotation (GOA)

Based on the sequence similarity results (see flowchart for criteria), MEPD entries are annotated according to the GO guide for computational annotations (http://www.geneontology.org/GO.annotation.html). We run an automatic annotation (IS: ‘inferred by sequence similarity’) refined by a manual one (IC: ‘inferred by curator’). Annotations and corresponding evidences are stored within the database and displayed on the search result interfaces.

Furthermore, in order to classify the ESTs into the GO biological process, molecular function and cellular component categories (Ashburner *et al*., 2000), we used GO assignments for the UniProt database, kindly provided by the GOA project at EBI (Camon *et al*., 2004), and stored them locally. So far, 7055 entries have been described in total, 5316 by ‘biological process’, 6280 by ‘molecular function’ and 4422 by ‘cellular localization’.

FUTURE PERSPECTIVE

In a collaborative project, we are aiming to complete expression analysis of the medaka transcriptome within the next three years.

Using similar ontologies for Zebrafish and Medaka we want to establish links from one entry to the expression pattern of the homologous gene in the ‘other’ fish and aim to develop interfaces that can search expression patterns in both fish species.

Currently, we are developing the technique for recording three-dimensional (3D) expression information. We used a Selective Plane Illumination Microscope, which has been developed at EMBL (Huisken *et al*., 2004) to capture *in situ* hybridization patterns as well as a framework for a virtual 3D embryo. We mapped some of the anatomy ontology terms onto this framework and we will use the virtual embryo as a 2D and 3D expression pattern annotation tool.

We are currently clustering expression patterns (Tomancak *et al*., 2002) using their anatomy ontology term description to check for synexpression groups (Niehrs and Pollet, 1999). The unigene set will be available on microarray and can be used to complement these studies.

ACKNOWLEDGEMENTS

We would like to thank MEPD users for their constructive feedback on the interfaces and the Medaka ontologies. We thank Geo Pertea from TIGR for his help with the tgicl software. We thank Daniel Barrell and Wolfgang Fleischmann for their help with the UniProt GO annotation. This work was supported by grants from the HFSPO (J.W., H.H. and H.K.) and the EU (J.W.).

REFERENCES

Ashburner,M., *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bard,J. (2003) Ontologies: formalising biological knowledge for bioinformatics. *Bioessays*, **25**, 501–506.

Birney,E. *et al*. (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.

Camon,E. *et al*. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**(Database issue), D262–D266.

FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.

Fuchs,T. *et al*. (2002) DEFOG: a practical scheme for deciphering families of genes. *Genomics*, **80**, 295–302.

Henrich,T. and Wittbrodt,J. (2000) An *in situ* hybridization screen for the rapid isolation of differentially expressed genes. *Dev. Genes Evol.*, **210**, 28–33.

Henrich,T. *et al*. (2003) MEPD: a Medaka gene expression pattern database. *Nucleic Acids Res.*, **31**, 72–74.

Henrich,T. *et al*. (2004) GSD: a genetic screen database. *Mech. Dev.*, **121**, 959–963.

Huisken,J. *et al*. (2004) Optical sectioning deep inside live embryos by selective plane illumination microscopy. *Science*, **305**, 1007–1009.

Iwamatsu,T. (1994) Stages of normal development in the Medaka *Oryzias latipes*. *Zool. Sci.*, **11**, 825–839.

Meier-Ewert,S. *et al*. (1998) Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Res.*, **26**, 2216–2223.

Nguyen,V. *et al*. (2001) An *in situ* screen for genes controlling cell proliferation in the optic tectum of the medaka (*Oryzias latipes*). *Mech. Dev.*, **107**, 55–67.

Niehrs,C. and Pollet,N. (1999) Synexpression groups in eukaryotes. Nature, **402**, 483–487.

Pertea,G. *et al*. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Quiring,R. *et al*. (2004) Large-scale expression screening by automated whole-mount *in situ* hybridization. *Mech. Dev.*, **121**, 971–976.

Schartl,M. *et al*. (2004) Current status of medaka genetics and genomics. The Medaka Genome Initiative (MGI). *Methods Cell. Biol.*, **77**, 173–199.

Shima,A. *et al*. (2003) Fish genomes flying. Symposium on Medaka Genomics. *EMBO Rep.*, **4**, 121–125.

Sprague,J. *et al*. (2003) The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res.*, **31**, 241–243.

Tomancak,P. *et al*. (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol.*, **3**, RESEARCH0088.

Wittbrodt,J. *et al*. (2002) Medaka—a model organism from the far East. *Nat. Rev. Genet.*, **3**, 53–64.