Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species

Frederic Bastian[1,2,*], Gilles Parmentier[1,2,*], Julien Roux[1,2], Sebastien Moretti[1,2], Vincent Laudet[3], and Marc Robinson-Rechavi[1,2]

[1] Department of Ecology and Evolution, University of Lausanne, quartier UNIL-Sorge, 1015 Lausanne, Switzerland
[2] Swiss institute of bioinformatics, Lausanne, Switzerland
[3] Université de Lyon, Institut de Génomique Fonctionnelle de Lyon, ENS Lyon, Université Lyon 1, CNRS, INRA, Institut Fédératif 128 Biosciences Gerland Lyon Sud, France
Marc.Robinson-Rechavi@unil.ch

Abstract. Gene expression patterns are a key feature in understanding gene function, notably in development. Comparing gene expression patterns between animals is a major step in the study of gene function as well as of animal evolution. It also provides a link between genes and phenotypes. Thus we have developed Bgee, a database designed to compare expression patterns between animals, by implementing ontologies describing anatomies and developmental stages of species, and then designing homology relationships between anatomies and comparison criteria between developmental stages. To define homology relationships between anatomical features we have developed the software Homolonto, which uses a modified ontology alignment approach to propose homology relationships between ontologies. Bgee then uses these aligned ontologies, onto which heterogeneous expression data types are mapped. These already include microarrays and ESTs. Bgee is available at http://bgee.unil.ch/

Keywords: gene expression pattern, homology, ontology, data integration.

1 Introduction

Gene expression patterns (when and where a gene is expressed) are a key feature that underlies the development of organisms and phenotypes of individuals. They are an important aspect of the study of gene function. Moreover, the study of the evolution of developmental processes, often called "evo-devo", has shown that the primary source of change in the evolution of phenotypes is changes in gene expression [1] rather than sequence.

Comparing gene expression patterns between animals is thus a major step in the study of gene function as well as of animal evolution, and also provides a link between genes and phenotypes.

* Corresponding author.

A. Bairoch, S. Cohen-Boulakia, and C. Froidevaux (Eds.): DILS 2008, LNBI 5109, pp. 124–131, 2008.

In biological research, results obtained in different organisms are routinely compared. A comparative approach may be chosen for practical reasons because the organism of interest (humans, farm animals) may be less amenable to experimentation than more or less distant model species (as mouse, rat, zebrafish, or fruit fly).

Another reason is that components of gene expression may vary for no obvious reason [2]; this introduces the problem of distinguishing this signal from the noise caused both by random evolution and the inaccurate data measurements. Comparative study of gene expression in several species may contribute to this distinction. For example, comparing multiple samples from humans and rodents gave sufficient statistical evidence for a functionally relevant component of gene expression [3], and allowed for significant improvement in tumour characterisation [4].

Transcriptome data have also been compared among species to gain direct insight into evolutionary processes. For instance, yeast microarray data provided evidence for divergence of expression after genome duplication [5], and further studies have succeeded in extracting some evidence for the evolution of new gene functions after genome duplication in yeast and human lineages [6, 7]. A comparative approach would allow to understand the mechanisms and the consequences of gene expression evolution.

We have developed Bgee (a dataBase for Gene Expression Evolution) to address these questions. Bgee must answer the following requirements, to enable large scale gene expression pattern comparison:

- Precise description of the anatomy and developmental stages of each species, stored in a computer-understandable way.
- Integration of expression data in order to know in which anatomical features (spatial mapping) and which developmental stages (temporal mapping) genes are expressed.
- Comparison criteria between anatomies, developmental stages, and genes.

To unambiguously describe anatomy and development of a species in a computer-understandable way, ontologies are required: they describe a domain of knowledge, by using well-defined concepts and designing relationships amongst them. Several databases provide species-specific ontologies that describe anatomical features for a species, such as ZFIN [8] for the zebrafish. But as far as we know, no database provides relationships between these ontologies to allow comparisons.

The appropriate criterion to make comparisons in an evolutionary context is homology: we need to compare features that derive from the same ancestral element. We have thus designed homology relationships between anatomies of different species. This is a difficult task, and Bgee implements computational methods to achieve it (section 2). Then, we need homology relationships between genes. This point has already been abundantly treated in bioinformatics, and will not be discussed in detail in this paper. Finally, we need relationships between developmental stages. As these stages are artificial features that help to describe the continuous process of development, homology cannot be defined in a rigorous manner. We have rather designed a mapping of "equivalent" developmental stages between species (section 3).

To describe gene expression patterns, Bgee requires large amounts of data. To this end, heterogeneous data types are used (ESTs, microarrays, and soon *in situ* hybridizations). The common information to gather is whether an experiment has

determined that a gene is expressed or not, and with which confidence. We have applied different statistical tests for each data type to obtain this information (section 4).

Thanks to the successful implementation of all these requirements (anatomical and developmental ontologies, comparison relationships between ontologies and genes, integration of heterogeneous expression data), Bgee allows the easy retrieval of gene expression data for different species, as well as the automated comparison of gene expression patterns.

2 Designing Homology Relationships between Anatomical Ontologies by an Ontology Alignment Approach

To study the evolution of gene expression patterns, comparisons have to be done between organs that evolved from a common ancestral structure. Thus designing relationships between anatomical ontologies consists in finding correspondences (homology relationships) between the concepts (organs) of these ontologies. This problem is a special case of "schema matching", or "ontology alignment".

Ontology alignment ([9] for a review) is the process of determining correspondences between ontology concepts. Usually, this technique is used to find the common concepts present in two ontologies. In the case of anatomical ontologies, the concepts to align are not strictly common, but rather, related: a homology relationship is not an equivalence relationship. For this reason, ontology alignment approaches developed for other applications cannot be applied as is: these methods would be misled by the existence of elements of same names and related to the same concept, but not homologous (eye of insects and of vertebrates for instance), or reciprocally, homologous elements with different names (caudal fin and upper limb for instance). This is why we apply modified ontology alignment techniques in order to find putative homologies between two species anatomies. An expert has to manually validate the putative homologs. This method is implemented by Homolonto, a software that we have developed in Java. Homolonto will be presented in detail elsewhere; we present here the outline of its algorithm.

Our process is a supervised one: at each step, some homology relationships are proposed to the expert, who may validate them or not. Computations are made based on these decisions, and new propositions are made to the expert.

The algorithm starts with a list of pairs, which have identical names. This is based on the assumption that two structures that have the same name are likely homologous. For example, "optic cup" of the ZFIN ontology (zebrafish) and "optic cup" of the EHDA ontology (human) will be paired, but "optic cup" of ZFIN will not be initially paired with "optic nerve" of EHDA. The score of similarity between terms is up weighted by the proportion of common words, and down weighted by the frequency of these words (frequent words are less informative, e.g. "endoderm"). Moreover, scores are propagated between pairs which are neighbors in both ontologies. For example, the score of the "optic cup" pair is added to the score of the "eye" pair, as "optic cup" is part of "eye". In the same way, the score of the "eye" pair is added to the "optic cup" one.

Each pair is proposed to the expert, in descending order of scores. The expert may validate or invalidate the hypothesis of homology, or delay decision. The expert may choose to evaluate any number of pairs before triggering an iteration, in which computations are performed. Computations create or extend homology groups. The new homology information is propagated through the ontologies. The underlying idea is that if two concepts A and B are homologous, then one of the sub-concepts of A is probably homologous to one of the sub-concepts of B even if they have different names. Of note, validated homology contributes a significantly higher score than name similarity. Propagation is down weighted by the number of sub-concepts, to avoid generating many false positives (e.g. all the children of "whole body").

Evaluation of pairs, ordered by total score (base score + propagated score), and iteration, are repeated until the expert decides to terminate, or no more pairs are proposed. Compared to manual alignment of the ontologies, Homolonto reduces time considerably, with high sensitivity. Thus aligning the zebrafish (ZFIN; 2087 terms) and Xenopus (Xenbase; 480 terms) ontologies took one month by hand, but 2 days using Homolonto. The first 213 pairs proposed to the expert were valid at 80%, and contained 91% of all true positives.

To design homology relationships between several species, we merge the homology groups obtained by pair-wise alignment.

Finally, Homolonto generates an OBO [10] file containing the homology relationships. Bgee then parses this file to integrate the homologies into the database.

3 Mapping of the Developmental Ontologies

In relationship with the anatomical ontologies, Bgee uses for each species an ontology which describes its developmental stages, and links them using an *is_a* relationship by key states (e.g. embryo, hatching, larval).

To compare expression patterns, the comparisons have to be done both between homologous organs (see section 2), and at an equivalent developmental stage. But it is not possible to "simply" identify stages between species for which the state of the development is identical: organs do not develop at the same speed and with the same sequence, development is heterochronous (e.g. [11]).

A solution could be to identify, for each organ involved in a homology relationship, the different key states of their formation, and to design, organ by organ, equivalence relationship between these states in different species. This solution is difficult to implement, as it would imply manual definition for each organ separately, without any guiding principle in the data (i.e. we cannot use shared names and ontology structures as for anatomical homology).

Although there is no direct equivalence between the stages of two species because of heterochrony, it is instead possible to identify key events of development, common to all bilaterian animals. We have developed a small ontology of these common "metastages": embryo – including zygote, cleavage, blastula, gastrula, organogenesis –, post-embryonic development, adult. Then we have mapped the developmental stages of each species to these "metastages". This approach results in a loss of accuracy regarding the developmental ontologies, but allows to compare gene expression patterns taking into account the time dimension.

4 Integrating Heterogeneous Data on Anatomical and Developmental Ontologies

Integrating heterogeneous expression data is challenging, as it is difficult to compare the results of different types of techniques (e.g. ESTs, microarrays, *in situ* hybridizations) [12, 13], and even for a same type, to compare results between experiments (e.g. compare two microarray experiments made on different platforms). But as we want to be able to precisely describe expression patterns of genes, we need data as complete as possible. We also want to obtain data for all the species studied, and some techniques cannot be applied to all species, for instance *in situ* hybridizations on human. The information we want to collect is in which organs, and at which developmental stages, a gene is expressed. It means that for each experiment, we have to map the data to anatomical and developmental ontologies, and to apply statistical analyses, depending on the data type, to identify genes significantly expressed.

4.1 Mapping Expression Data to Ontologies

The main problem to map the data to ontologies is that annotations are often inconsistent between data sources: for instance, the description of the organs on which an experiment has been performed can be provided as free text, controlled vocabularies, or ontologies. Therefore, we have manually annotated each experiment stored in Bgee to determine the unique identifiers (ID) in the anatomical ontologies of the organs studied, and the ID of the developmental stages.

The granularity of the data is also highly variable. For instance, experiments can be reported on the organ “brain” or on the organ “forebrain”, at the stage “embryo” or at the stage “free blastocyst”. This is why ontologies are essential both for anatomy and for development: just listing the developmental stages would not have been sufficient.

4.2 Statistical Analyses

Bgee currently uses EST data from Unigene [14] and Affymetrix data retrieved from ArrayExpress [15]. For each data type, Bgee applies statistical tests to identify genes that are significantly expressed, with two levels of confidence: low and high.

For experiments based on tag counting, such as EST, SAGE, or MPSS, a statistical test [16] shows that a gene is expressed with a 95% confidence if 7 tags are mapped to this gene (the number of tags is statistically different from 0). So for EST data, we have considered a gene as expressed with a high confidence if an experiment has found at least 7 EST related to this gene, and with a low confidence from 1 to 6 EST.

Affymetrix data are measurements of fluorescence intensity. Labelled cDNAs prepared from samples are hybridized with oligonucleotide probes. All probes mapping to the same transcript constitute a probeset. Identifying genes significantly expressed consists in finding genes for which the signal of the probeset is significantly different from the background signal. This method is implemented by the MAS5 software [17]; based on these statistical analyses, probesets are flagged as "present", "marginal", or "absent". This allows us to classify genes expressed with a high confidence when their probeset is flagged as "present", and with a low

confidence when "marginal". Although MAS5 classification is efficient [18], the estimation of the background signal can be biased depending on probe sequence affinity [19]. We are currently implementing another method of detection [19], which uses the gcRMA algorithm [20] to normalize the signal taking into account probe sequences, and uses a subset of weakly expressed probesets for estimating the background. A Wilcoxon test is then applied to compare the normalized signal of the probesets with the background signal. Genes will be considered expressed with a high confidence if the p-value is lower than 1%, and with a low confidence if the p-value is between 1 and 5 %.

Bgee will soon include *in situ* hybridization data. For data based on image analyses, statistical tests cannot be applied easily. Determining if a gene is expressed is usually done manually by an expert. A quality annotation can also be provided, summarizing the quality of the image, the hybridization, and the probes design. Such information is already present in several databases (e.g. ZFIN [8]), and Bgee will rely on them.

5 Database and Web-Interface of Bgee

The database of Bgee is developed with MySQL, and currently includes anatomical ontologies, developmental ontologies, and expression data for four species: human, mouse, zebrafish, and Xenopus:

- The anatomical ontologies come from eVoc [21] for human, Xspan [22] for human and mouse, MGD [23] for adult mouse, ZFIN [8] for zebrafish, and Xenbase [24] for Xenopus.
- EST data come from Unigene [14] and Affymetrix data from ArrayExpress [15]. *In situ* hybridization will be collected from specialized databases, as ZFIN or BGEM [25].
- Gene ontology [26] annotations and homology relationships between genes are recovered from Ensembl [27].
- Bgee currently includes a total of 104,881 genes. 51,277 have expression data, in 587 anatomical structures and 93 developmental stages.

The web interface of Bgee is developed in Java using the servlet container Tomcat, with a Model-View-Controller architecture. The user experience is improved by the use of AJAX technologies (Asynchronous Javascript And XML). The website of Bgee, available at http://bgee.unil.ch/, proposes several ways to easily retrieve or compare expression data:

- Querying the database: data can be queried for genes, gene families, anatomical structures, or developmental stages, based on their names, synonyms, abbreviations, identifiers, or descriptions.
- Browsing the ontologies: anatomical and developmental ontologies can be browsed as a tree structured view. Information about the genes expressed is displayed for each anatomical structure or developmental stages. The display of these expression data can be adjusted by selecting data type and data quality, or by entering a list of gene identifiers or of GO terms.

- Retrieving the expression pattern of a gene: the expression pattern of a gene is also displayed as a tree structured view of the organs where it is expressed, at the selected developmental stage. The data used to define the pattern can be modified by selecting the data type or data quality.
- Comparing the expression patterns of homologous genes: the expression patterns of a gene family can be compared choosing the species studied, and as for the ontology browsing, by selecting data type and quality, list of genes or of GO terms.

The homology relationships and developmental ontologies, both in OBO format, the Homolonto software and source code, and the Bgee database and source code, will soon be available on our website.

6 Conclusions

We have developed pipelines to integrate ontologies and expression data to Bgee, and automatically perform statistical analyses. We also have developed the Homolonto software to facilitate the design of homology relationships. We have paid great attention to make the Java code of Bgee easy to evolve, with a clean architecture and reusable components. We have thus implemented all the requirements to add more species and more data types into Bgee in the future. We plan to add in the short-term *in situ* hybridization data.

The multi-species computer coding and storage of expression patterns was an essential key to perform high throughput analyses. We will now be able to design analysis tools dedicated to the comparison of expression patterns, and to address open biological questions, such as the relationships between evolution of development and of gene expression, or the identification of candidate genes for diseases.

Acknowledgements. We thank Frederic Ricci for data annotation. Funding was provided by Etat de Vaud, the program Crescendo, the SIB, the Decrypthon program.

References

1. Carroll, S.: Endless Forms Most Beautiful: The New Science of Evo Devo and The Making of the Animal Kingdom. W. W. Norton & Company, New York (2005)
2. Yanai, I., Graur, D., et al.: Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. Omics 8, 15–24 (2004)
3. Jordan, I.K., Marino-Ramirez, L., et al.: Evolutionary significance of gene expression divergence. Gene 345, 119–126 (2005)
4. Schlicht, M., Matysiak, B., et al.: Cross-species global and subset gene expression profiling identifies genes involved in prostate cancer response to selenium. BMC Genomics 5, 58 (2004)
5. Gu, Z., Nicolae, D., et al.: Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet 18, 609–613 (2002)
6. Gu, X., Zhang, Z., et al.: Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc. Natl. Acad. Sci. USA 102, 707–712 (2005)

7. He, X., Zhang, J.: Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169, 1157–1164 (2005)
8. Sprague, J., Clements, D., et al.: The Zebrafish Information Network (ZFIN): the zebrafish model organism database. Nucleic Acids Res. 31, 241–243 (2003)
9. Shvaiko, P., Euzenat, J.: Ontology Matching. Springer, Heidelberg (2007)
10. Smith, B., Ashburner, M., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25, 1251–1255 (2007)
11. Jeffery, J.E., Bininda-Emonds, O.R., et al.: A new technique for identifying sequence heterochrony. Syst. Biol. 54, 230–240 (2005)
12. Lee, C.K., Sunkin, S.M., et al.: Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. Genome biology 9, R23 (2008)
13. Kuo, W.P., Liu, F., et al.: A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. Nat. Biotechnol. 24, 832–840 (2006)
14. Wheeler, D.L., Barrett, T., et al.: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 36, 13–21 (2008)
15. Parkinson, H., Kapushesky, M., et al.: ArrayExpress–a public database of microarray experiments and gene expression profiles. Nucleic Acids Res. 35, 747–750 (2007)
16. Audic, S., Claverie, J.M.: The significance of digital gene expression profiles. Genome Res. 7, 986–995 (1997)
17. Liu, W.M., Mei, R., et al.: Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics 18, 1593–1599 (2002)
18. Choe, S.E., Boutros, M., et al.: Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. Genome biology 6, R16 (2005)
19. Schuster, E.F., Blanc, E., et al.: Correcting for sequence biases in present/absent calls. Genome biology 8, R125 (2007)
20. Wu, Z., Irizarry, R.A., et al.: A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association 99, 909–917 (2004)
21. Kruger, A., Hofmann, O., et al.: Simplified ontologies allowing comparison of developmental mammalian gene expression. Genome biology 8, R229 (2007)
22. Aitken, S.: Formalizing concepts of species, sex and developmental stage in anatomical ontologies. Bioinformatics 21, 2773–2779 (2005)
23. Eppig, J.T., Blake, J.A., et al.: The mouse genome database (MGD): new features facilitating a model system. Nucleic Acids Res. 35, 630–637 (2007)
24. Bowes, J.B., Snyder, K.A., et al.: Xenbase: a Xenopus biology and genomics resource. Nucleic Acids Res. 36, 761–767 (2008)
25. Magdaleno, S., Jensen, P., et al.: BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system. PLoS Biol. 4, e86 (2006)
26. Ashburner, M., Ball, C.A., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 25, 25–29 (2000)
27. Hubbard, T.J., Aken, B.L., et al.: Ensembl 2007. Nucleic Acids Res. 35, 610–617 (2007)