IN SILICO ANALYSIS OF GENE EXPRESSION PATTERNS DURING EARLY DEVELOPMENT OF *XENOPUS LAEVIS*

N. POLLET, H. A. SCHMIDT, V. GAWANTKA, C. NIEHRS and M. VINGRON.
Department of Theoretical Bioinformatics, Department of Molecular Embryology, Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, D-69120, GERMANY

The information as to where and when a mRNA is present in a given cell is essential to bridge the gap between the DNA sequence of a gene and its physiological function. Therefore, a major component of functional genomics is to characterize the levels and the spatio-temporal domains of gene expression. Currently, there is just a few specialised public databases available storing the data on gene expression while they are needed as a resource for the field. Moreover, there is a need to develop and assess computational tools to compare and analyse expression profiles in a suitable way for biological interpretation. Here we describe our recent work on developing a database on gene expression for the frog *Xenopus laevis*, and on setting up and using new tools for the analysis and comparison of gene expression patterns. We used histogram clustering to compare expression profiles at both gene and tissue levels using a set of data coming from the characterization of the expression of genes during early development of *Xenopus*. This enabled us to draw a tree of tissue relatedness and to identify coexpressed genes by i*n silico* analysis.

1 Introduction

During embryonic development, a single cell, the fertilized egg, gives rise to a number of tissues, each comprised of many cell types organised in a characteristic spatial arrangement. The processes controlling cell fate, morphogenetic movements and pattern formation involve the control of expression of different sets of genes at different levels. Each differentiation state of a cell can thus be characterized by a specific set of gene transcripts.

The description and analysis of gene expression patterns is crucial to elucidate the physiological functions of genes and to understand the network of genetic interactions that underlies the process of normal development. Knowing the patterns of expression of large numbers of genes should be a means allowing the identification of promoters having particular specificity, of marker genes used to monitor cells in a specific state and of genes which are tightly coregulated.

Recent technical advances in gene expression monitoring have permitted global surveys of gene activity at high resolution during development and differentiation. Surveys done in yeast, rat, *Xenopus* and human cells produced the kind of data

necessary to analyse gene expression patterns and revealed whole groups of genes which are both coregulated and functionally interacting[1,2,3].

Embryonic gene expression patterns in vertebrates are currently compiled using existing data in mouse, in a collaborative effort between the MRC Edinburgh and the Jackson Laboratories[4]. The frog *Xenopus laevis*, while being a model organism in embryology since decades, is lacking a centralized database like Flybase for *Drosophila*, ACEDB for the nematode or MGD for the mouse. During our large-scale in situ hybridization screen, we developed a database using ACEDB as the software engine[3,5]. Such a database should provide a major source of information to allow the easy identification of genes based on their expression patterns, facilitating the elucidation of gene interactions during development.

We wanted to apply a gene expression clustering method to classify genes based on their expression patterns and to compute the relatedness of tissues using gene expression profiles on our dataset[3]. This would enable to tackle the question : What are those genes responsible for the making of this given tissue?

Here we describe our recent work on developing a database on gene expression for the frog *Xenopus laevis* and on using histogram clustering for the analysis of gene expression patterns.

The clustering of genes or tissues based on mRNA levels is an instance of the general problem of how to cluster arbitrary objects. Such objects may come as vectors in some high dimensional space or one is merely given distances among the objects. In the case of the quantified mRNA levels we are given a matrix of values which we take at face value for our current purposes, although the scale has been arbitrary. The matrix has one row per gene and one column per tissue. Thus, each row can be seen as a vector associated to a gene and each column can be seen as a vector of much larger dimension associated to a tissue. Consequently, both genes and tissues can be clustered.

2 Axeldb

2.1 System environment

Axeldb was implemented using ACEDB (A Caenorhabditis elegans database) as the database software[5]. ACEDB version 4.5e for Unix was used along with Perl 5.004, CGI.pm 2.42, AcePerl 1.54 and a tailored AceBrowser 2.01. The whole database is accessible on the World Wide Web at <http://www.dkfz-heidelberg.de/abt0135/axeldb.htm>.

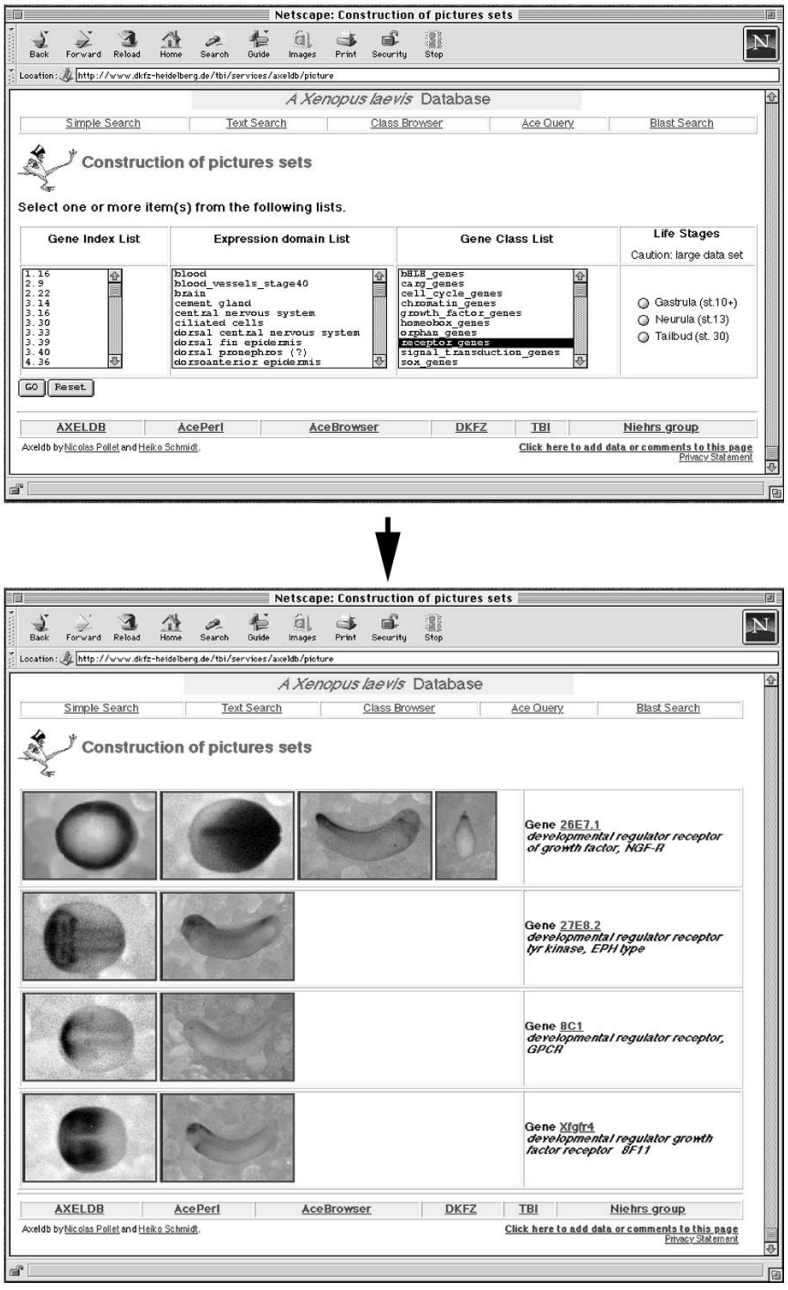
Netscape: Construction of pictures sets
Back Forward Reload Home Search Guide Images Print Security Stop
Location: http://www.dkfz-heidelberg.de/tbi/services/axeldb/picture
A Xenopus laevis Database
Simple Search Text Search Class Browser Ace Query Blast Search
Construction of pictures sets
Select one or more item(s) from the following lists.
Gene Index List
Expression domain List
Gene Class List
Life Stages
Caution: large data set
1.16
2.9
2.22
3.14
3.16
3.30
3.33
3.39
3.40
4.36
blood
blood_vessels_stage40
brain
cement gland
central nervous system
ciliated cells
dorsal central nervous system
dorsal fin epidermis
dorsal pronephros (?)
dorsoanterior epidermis
bHLH_genes
cang_genes
cell_cycle_genes
chromatin_genes
growth_factor_genes
homeobox_genes
orphan_genes
receptor_genes
signal_transduction_genes
sox_genes
Gastrula (st.10+)
Neurula (st.13)
Tailbud (st. 30)
GO
Reset
AXELDB AcePerl AceBrowser DKFZ TBI Niehrs group
Axeldb by Nicolas Pollet and Heiko Schmidt.
Click here to add data or comments to this page
Privacy Statement
Netscape: Construction of pictures sets
Back Forward Reload Home Search Guide Images Print Security Stop
Location: http://www.dkfz-heidelberg.de/tbi/services/axeldb/picture
A Xenopus laevis Database
Simple Search Text Search Class Browser Ace Query Blast Search
Construction of pictures sets
Gene 26E7.1
developmental regulator receptor of growth factor, NGF-R
Gene 27E8.2
developmental regulator receptor tyr kinase, EPH type
Gene 8C1
developmental regulator receptor, GPCR
Gene Xfgfr4
developmental regulator growth factor receptor 8F11
AXELDB AcePerl AceBrowser DKFZ TBI Niehrs group
Axeldb by Nicolas Pollet and Heiko Schmidt.
Click here to add data or comments to this page
Privacy Statement

Figure 1

Figure 1. (Previous page) Browsing pictures in Axeldb. Example of a script enabling the query of Axeldb based on gene names, expression domains, gene class or life stage. The output of a search on genes encoding receptors is shown. Both pictures and gene names are hyperlinks.

2.2 Data sources

Actually, all informations available in Axeldb are coming from the study performed in our laboratory, with the notable exception of litterature references extracted from Medline. Since our large scale gene expression study combined mRNA in situ hybridization and cDNA sequencing, we had to cope with sequence data, gene expression patterns and picture documentation.
The expression patterns are described by using a comprehensive list of 17 anatomical terms for the tailbud stage embryos, and by a short textual description for all stages studied. A score between 0 (no expression detected) and 4 (very strong expression detected) was attributed for each structure.

2.3 Representation of expression patterns

The basic of Axeldb is to link expression pattern information to pictures, cDNA clones, genes and sequence. On top of that, information concerning genes, such as their usefulness as differentiation markers, their functional category based on sequence informations and their appartenance to a synexpression group is integrated. This allow to display all informations available for any given gene (Fig. 1). We incorporated modifications in the original description of object classes in ACEDB to meet our special needs, our models.wrm file is available at http://www.dkfz-heidelberg.de/abt0135/axeldb/misc.html. At present comprehensive informations on 273 genes is available.

2.4 Query interfaces

The combination of AcePerl and AceBrowser interfaces allows complex queries to be be passed on Axeldb server transparently for the end-user, and to display information in a comprehensive way[6]. On top of simple searches and ACEDB-like queries, we developed a script to display only the pictures documenting the expression pattern of a given gene fulfilling a search criterias, enabling to select any set of gene and see the results of whole-mount in situ hybridization (Fig. 1).

3 Distance calculation

3.1 Clustering of the tissues

Clustering data into a hierarchical tree rests upon the idea that there is some kind of either a hierarchical structure in the data or that there is a branching process that has given rise to the data and that can be depicted as a tree. In the case of the tissues there is no apparent reason to believe in a hierarchic structure. There is, however, good reason to believe that a developmental, branching process has given rise to these data. Thus it seems reasonable to aim at reconstructing a tree structure for the tissue vectors.

To this end, a distance matrix needs to be computed for the tissue vectors. Since these data are not really numerical data it is dangerous to naively compute some kind of geometric or correlation distance. Instead we base our distance computation on an intermediate step of the computation of a similarity matrix which allows us to define a similarity score in a way that is appropriate for the data. Our similarity matrix assigns scores to matching the different intensities. For example, when a gene is highly expressed in two tissues this should yield a good similarity score. When the same gene is absent in two tissues, this seems to be less convincing a signal of a similarity between the two tissues although not a signal for their being different either. The most dissimilar situation would be when a gene is strongly expressed in one tissue and absent in the other. Based on this intuition we have used the following (symmetric) similarity matrix:

	0	1	2	3	4
0	4	3	2	1	0
1		5	4	3	2
2			6	5	4
3				7	6
4					8

The similarity scores range between 0 and 8, with 0 standing for strong dissimilarity while 8 signifies strong similarity. Once the 0 is fixed as the lowest score, the scores in the first row are assigned in increasing order in steps of 1 all the way to the left where the (0/0) entry is then assigned a 4. Scores further rise in steps of 1 along the main diagonal while decreasing to the right along each row.

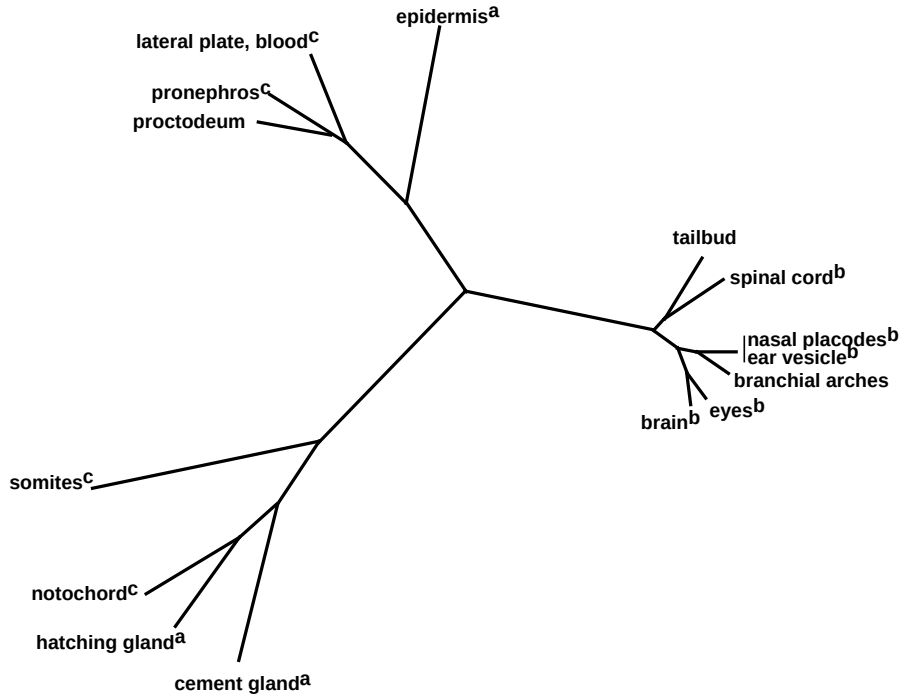

Figure 2

Figure 2. Tree of tissue relatedness. The distance matrix computed as described in the text was used to draw a tree using the kitsch program. Tissues marked with an [a] are derived from the ectoderm, those marked with a [b] are derived from the neurectoderm and those marked with a [c] are derived from mesoderm.

Once the individual similarity scores are chosen one can compute a similarity score for a pair of tissue vectors by adding up the individual similarity scores for each gene. This results in a similarity matrix. This matrix, however, has mathematically very awkward features since the row-sums differ widely. This is due to different tissues showing different overall amounts of expressed genes. Using this matrix directly for clustering purposes is very difficult but it can be converted into a distance matrix with comparatively little effort. Each tissue has a certain similarity value to any other tissue. The overall weight of a row is irrelevant when one focusses on how a single tissue distributes its affinities to the others. Thus, it makes sense to normalize the rows by their sum, i.e. to convert each row into a histogram.

While this operation makes the matrix asymmetric it now allows to define distances among the tissues as the distances among these histograms which each characterizes a tissue. We subsequently apply a standard distance measure to these histograms, namely relative entropy (also known as Kulback-Leibler distance).

Since relative entropy is an asymmetric measure it needs to be symmetrized. This is achieved by first introducing a midpoint defined by the arithmetic average between two histograms. Then the distance between the two histograms is defined as the sum of the two relative entropies between either of the histograms and the midpoint. In the end this results in a distance matrix which has 0 on the main diagonal, otherwise positive and is symmetric.

This distance matrix can now be used to analyze it for a possible tree structure. We applied several programs and looked for consistent features in their respective output. Splitstree (Fig.3) is a very conservative program in that it only shows splits in the data that are very strong. In terms of actual tree construction programs we used results generated by the Phylip[7] programs fitch, kitsch and neighbor, where the latter program was used to compute both a neighbor joining tree and a UPGMA tree. Splitstree[8] which models the distance data as a so-called split-metric, suggests a fairly linear structure of the data. They can be enumearted starting with the neurectodermal derivatives through a group consisting of pronephros, proctodeum, epidermis, lateral plate and blood to the other end of the spectrum which is formed first by somites, then notochord and hatching gland and finally cement gland.

The actual tree building programs tend to follow this outline in that they roughly divides the data into the three groups of i) brain, spinal cord, ear vesicle, nasal placode, branchial arches, tailbud and ii) epidermis, blood, lateral plate, pronephros, proctodeum and iii) notochord, somites, hatching gland, cement gland. Tree-likeness is indeed not perfect since, e.g., Neighbor Joining Algorithm results in some negative edges.

3.2 Clustering of the genes based on their expression patterns

The difficulty in analyzing the tissue data lies largely in the fact that some tissues have many genes active at high levels while others are "weakly populated". However, it seemed a reasonable assumption to normalize each gene by the sum of its own intensities across all tissues. This immediately lets us transform the rows of our data matrix into histograms such that the same algorithms for derivation of a distance matrix can be used.
For a distance matrix of that size it is impractical to apply any sophisticated tree building algorihtm, all the more since there is no interpretation of these data as having developed by some kind of branching structure. Therefore we used the simple hierarchical clustering method UPGMA to compute a tree for the genes. Here, the tree serves merely to group more similarly behaving genes together. The result is shown in Fig.3.

4. Discussion

In vertebrates, some hundred expression patterns have been characterised in each of the embryos from the mouse, zebrafish or frog. There is some laboratories engaged in large-scale in situ hybridization screening[9]. Due to these and other non-systematic efforts we expect a dramatic increase in the number of characterised gene expression patterns in the future. The development of specialised databases will therefore be imperative to manage and interpret these biological informations. Axeldb was developed to be one such example for the frog *Xenopus laevis*. We consider that this kind of knowledge base for gene expression patterns will be more important in the future.
We used ACEDB for the software engine of Axeldb. ACEDB appeared as an excellent compromise between simplicity of set-up, evolvability and capabilities. Moreover, a powerful scriptable interface is available[6]. For those outside the field of *Xenopus laevis*, the database would provide expression patterns of homologous genes from other species. The scheme of Axeldb would allow to build gene expression databases, and to cross-reference the data produced by other large-scale in situ hybridization screen. This would enable to link the data of expression patterns concerning known orthologous genes. These have proven predictable power for the other species, including man, where such analyses are not easily feasible but highly desirable. As a long term goal Axeldb may be made open for outside entries to serve as a resource for the *Xenopus* field, acquiring, managing and releasing data about the biology of the frog.

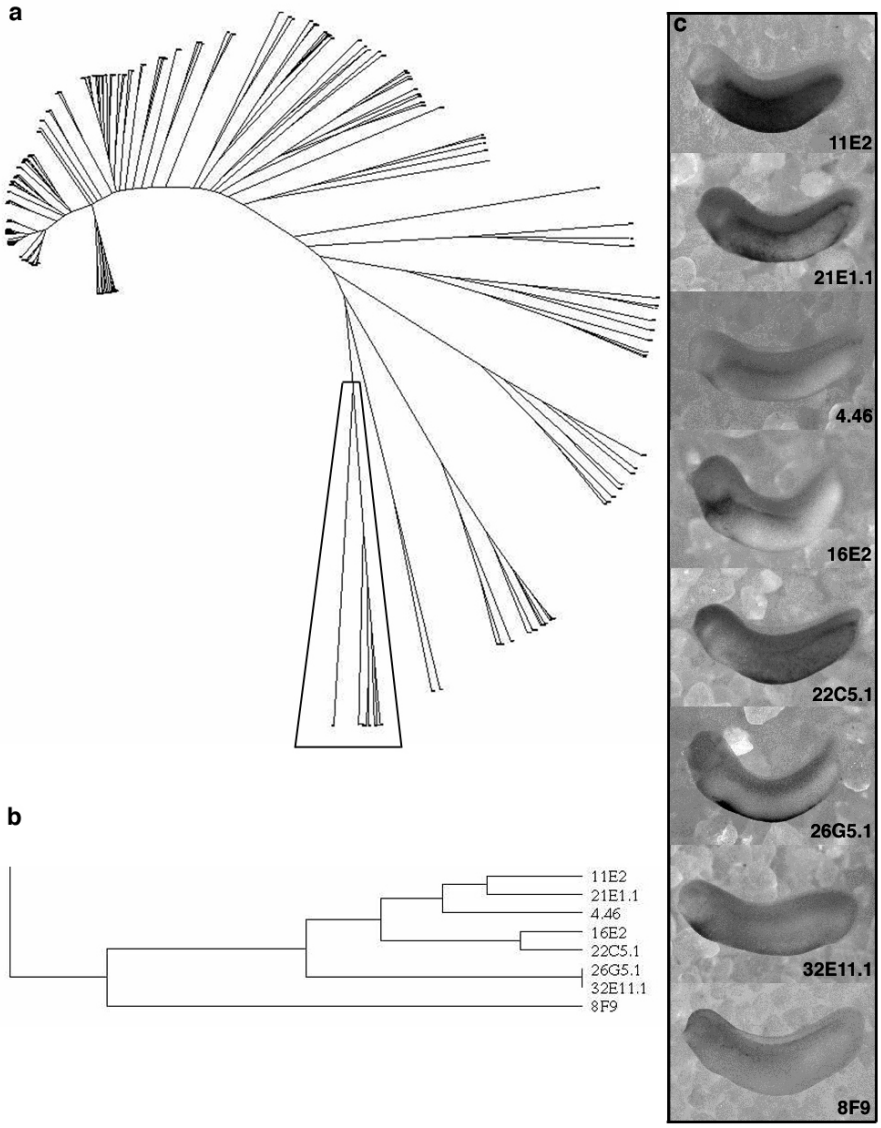
a
b
11E2
21E1.1
4.46
16E2
22C5.1
26G5.1
32E11.1
8F9
c
11E2
21E1.1
4.46
16E2
22C5.1
26G5.1
32E11.1
8F9

Figure 3

Figure 3. (Previous page) Clustering of genes. a: Tree representing the clustering of 268 genes based on their spatial expression profiles in the tailbud stage embryo of Xenopus. The region drawn is shown in more details in b. c: In situ hybridization results on tailbud stage embryos viewed laterally, anterior to the left. The probe used are from the clustered genes shown in b. 11E2, 21E1.1, 4.46 are expressed in pronephros, lateral plate mesoderm and blood islands. 16E2, 22C5.1 are expressed in lateral plate and blood. 26G5.1 and 32E11.1 are expressed only in the blood region. 8F9 is expressed mainly in cells from the lateral plate.

Ontogenesis is largely governed by lineage relationships between cell populations in the body parts. This means that a tree can be traced depicting this relationship between body parts and their embryological origins. Cellular differentiation occurs by the differential expression of the genetic programme, a mechanism known to be mainly due by the regulation of transcription. We wanted to compute the relatedness of tissues using gene expression profiles, to reconstruct this kind of tree. The result of the splitstree analysis indicates that the lineage relationships between the tissues are not strong enough to light-up a tree-like structure in the data. The linear order between the tissues suggests more the advance in the differentiation program of the respective tissues. On one extreme, the brain and most neuroectodermal derivatives are not as advanced in differentiation than the cement gland, which is fully differentiated since stage 28 of development. The inspection of the tree produced by the tree building programs suggests a superimposition of the lineage relationships between tissues. Therefore the observed tree represents an average between the lineage relationships and the advance in differentiation. The input from genes acting as differentiation markers counterbalance the ontogenetic relationships between tissues.

The comparison of gene expression patterns using clustering methods should help to identify those genes called differentiation markers used to monitor specific cell types or developmental mechanisms and should highlight those genes which are tightly coregulated. To our knowledge, this is the first study clustering spatial gene expression patterns. The spatial gene expression profiles cluster mainly according to the number of tissues showing expression. Genes that can act as markers are far from those expressed in a number of tissues (Fig. 2). The comparison between our previous categorization of genes from a manual inspection of data[3] and this clustering indicates the usefulness of this latter approach to identify genes coexpressed in one or a few structures. Sets of genes that are tightly coexpressed in many structures, that we previously called synexpression groups[3] are not nicely showing up in the tree, although they cluster together. We think that it is essentially because they are expressed in one of the neurectodermal derivatives that alone are domains of expression of many genes. Moreover, regionalized expression in the embryonic structures we scored is not apparent in the data, while it holds important informations to put together genes in a synexpression group.

Many large-scale studies of gene expression relies on the use of DNA arrays. While

these methods produce numerical data reflecting the levels of mRNA in homogenates of cells, they do not capture spatial information. To decipher the processes occuring during early development requires the knowledge of the spatial distribution of transcripts in situ on small subsets of cells, or eventually at a cellular resolution. The method of choice to study spatial distribution of transcripts is in situ hybridization to whole-mounts embryos, or isolated organs. The results of such experiments is hard to express by numerical values reflecting objective measurements, and hence needs other algorithms to be analysed, different from those used to look the data from gene expression monitoring using DNA arrays.

With clustering being one of the main approaches to expression data analysis, the corresponding methods of distance computation are becoming centrally important. In time series-like data correlation coefficients adequately grasp what is important, namely the relative changes in expression strength along the time axis. Other kinds of data or other kinds of experiments do not lean themselves to this interpretation. In the in situ hybridization data we have observed a situation where the overall amount of signal is not important. What counts is the distribution over the different objects or domains. This has led us to convert the data into histograms. Having done this, (symmetrized) relative entropy takes the role of the correlation coefficient in order to derive a distance matrix.

Similarly, one also needs to distinguish to which end one wishes to compute a clustering. Is the clustering intended to simply yield an overview of the data by sorting things into similar bins, or are we aiming at reconstructing some tree-like process? Classical data analysis methods provide ways of clustering objects into bins and also provide error measures describing the quality of this clustering. If one suspects a tree-like process having generated the data, one can look for tree-likeness in the distance matrix. A program like Splitstree[8] can give the user a first impression of what are the strong, obvious features of a data set. A dedicated tree construction program will enforce a tree no matter whether it is in the data or not. In our case, split decomposition has shown the strong trend, while tree construction programs refined this information. On the other hand, where there is no tree-like structure at all there is still the possibility of low-dimensional representation of the data.

We are currently exploring the application of correspondence analysis to the in situ expression data.

References

1. J.L. DeRisi, R.I. Vishwanath and P.O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**:680-686 (1997).
2. X. Wen et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci USA* **95** 334-339 (1998).

3. V. Gawantka, N. Pollet, H. Delius, M. Vingron, R. Pfister, R. Nitsch, C. Blumenstock and C. Niehrs. Gene expression screening in *Xenopus* identifies molecular pathways, predicts gene function and provides a global view of embryonic patterning. *Mech. Dev* **77**:95-141 (1998).
4. M. Ringwald, R. Baldock, J. Bard, J.T. Eppig, M. Kaufmann, J.H. Nadeau, J.E. Richardson and D. Davidson. A database for mouse development. *Science* **265**:2033-2034 (1994).
5. R. Durbin and J. Thierry-Mieg. A *C. elegans* database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov (1991).
6. L. Stein and J. Thierry-Mieg. Scriptable access to the *Caenorhabditis elegans* genome sequence and other ACEDB databases. *Genome Res.* **8**:1308-1315 (1999).
7. W.M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science* **155**:279-284 (1967).
8. H.J. Bandelt and A.W.M. Dress. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol*. **1**:242-252(1992).
9. C. Niehrs. Gene expression screens in vertebrate embryos: more than meet the eyes. *Genes Funct*. **1**:229-231 (1997).