



A database for management of gene expression data *in situ*

Ekaterina Poustelnikova¹, Andrei Pisarev¹, Maxim Blagov¹,
Maria Samsonova^{1,*} and John Reinitz²

¹St. Petersburg State Polytechnic University, 29, Polytehnicheskaya ul., St Petersburg, 195251, Russia and ²Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794-3600, USA

Received on January 21, 2004; revised and accepted on February 3, 2004

Advance Access publication April 1, 2004

ABSTRACT

Motivation: To create a spatiotemporal atlas of *Drosophila* segmentation gene expression at cellular resolution.

Results: The expression of segmentation genes plays a crucial role in the establishment of the *Drosophila* body plan. Using the IBM DB2 Relational Database Management System we have designed and implemented the FlyEx database. FlyEx contains 2832 images of 14 segmentation gene expression patterns obtained from 954 embryos and 2 073 662 quantitative data records. The averaged data is available for most of segmentation genes at eight time points. FlyEx supports operations on images of gene expression patterns. The database can be used to examine the quality of data, analyze the dynamics of formation of segmentation gene expression domains, as well as estimate the variability of gene expression patterns. We also provide the capability to download data of interest.

Availability: <http://urchin.spbcas.ru/flyex>, <http://flyex.ams.sunysb.edu/flyex>

Contact: samson@spbcas.ru

INTRODUCTION

Biological and biomedical research is in the midst of changes driven by the development of new high-resolution assays and technologies. Among technologies used in the studies of development are fluorescent *in situ* hybridization and immunofluorescence histochemistry. These methods are used to study the level of RNA and protein *in situ*, providing detailed information on gene expression in time and space. Such information will make it possible to elucidate the physiological functions of genes, understand the networks of genetic interactions that underlie the process of animal development, reveal the molecular mechanisms of many diseases and develop efficient therapies.

Currently, the observational information about gene expression *in situ* is being accumulated at a continuously increasing

rate. For example, high-throughput RNA *in situ* hybridization was applied to the systematic characterization of gene expression patterns during *Drosophila* embryogenesis (Tomancak *et al.*, 2002). The Gene Expression Database (GXD) developed by the Jackson Laboratory (Ringwald *et al.*, 1999) stores images of gene expression patterns during mouse embryonic development.

Extensive documentation of a gene expression pattern requires a precise estimate of the degree of its natural and experimental variation, as well as information about its location in relation to the expression domains of other network genes. This can be done by comparison of many images from individual embryos stained for the expression of genes in the network under study. One way to make this comparison is to construct digital reference images of development, to which these expression patterns can be mapped. Such a collection of reference images is called a spatiotemporal atlas. Such atlases are currently being designed in several model systems (Hartenstein *et al.*, 1995). For example, the Edinburgh Mouse Atlas Project (EMAP) contains three-dimensional (3D) models of early post-implantation embryos and forms a bioinformatic framework to which the expression patterns of several genes have been mapped (Davidson and Baldock, 2001; Hecksher-Sorensen and Sharpe, 2001).

To be used in an informatics context the spatiotemporal atlas should be in the form of a database and should support efficient management, retrieval and analysis of gene expression data. Besides reference information the atlas should contain data from individual embryos and provide means for integrating gene expression data with spatiotemporal information. Finally, all this information should be presented through an easy to use interface that will allow a user to perform comparative analyses of data and images. This will help biologists to find meaningful interrelations within the data and formulate new hypotheses that can be tested by experiment.

In the course of an investigation of segment determination in *Drosophila* we acquired a large number of images of gene expression *in situ*. From these we have amassed

*To whom correspondence should be addressed.

high-resolution quantitative data on the expression of all genes in this functional network (Myasnikova *et al.*, 1999, 2001; Kozlov *et al.*, 2000). To make this data available to the scientific community we have constructed a database, known as FlyEx. FlyEx is a quantitative atlas of segmentation gene expression at cellular resolution. In this database, the gene expression data are tightly integrated with spatiotemporal information. In addition to the images of gene expression patterns, FlyEx contains numerical data on the expression of genes in the segmentation genetic network in different parts of the embryo and at different times during early development.

In this paper, we present the data model, architecture, contents and user interface of FlyEx.

SYSTEM AND METHODS

Description of the knowledge domain

Like all other insects, the body of the fruit fly *Drosophila* is made up of repeated units called segments. The determination of the segmental pattern takes place early in *Drosophila* embryogenesis, a process which we now briefly describe. Immediately following fertilization, the newly formed zygotic nucleus undergoes a series of rapid and synchronous nuclear divisions. Following the eighth such division, the nuclei migrate to the outside, or cortex, of the egg, which they reach by the end of the ninth division. This begins the ‘syncytial blastoderm’ stage, which is denoted as stage 4 in the standard nomenclature (Campos-Ortega and Hartenstein, 1985). The 10th through 13th nuclear divisions occur during this stage, which lasts from 90 to 130 min after fertilization. Because it is useful to distinguish different periods during stage 4, we will also refer to ‘cleavage cycles’, where cleavage cycle N lasts from the end of division $N - 1$ to the end of division N (Foe and Alberts, 1983). During cleavage cycle 14A cell membranes begin to invaginate between the nuclei, which defines the beginning of stage 5. When cellularization is complete, gastrulation (stage 6) begins. Segments become determined at about the onset of gastrulation (Simcox and Sang, 1983).

The genetic network which controls segmentation in *Drosophila* is well characterized (Wieschaus *et al.*, 1984; Nusslein-Volhard *et al.*, 1984). The initial determination of the segments is a consequence of the expression of 16 genes. Several of these genes (known as maternal coordinate genes) are expressed from the maternal genome to provide asymmetric initial conditions. The others are zygotic and are expressed in patterns that become more spatially refined over time.

Expression of segmentation genes is largely a function of position along the anterior–posterior (A–P) axis of the embryo, and so can be well represented in one dimension. Hence, to investigate the mechanism of segment determination it is sufficient to construct reference data for the central 10% strip of an 2D lateral projection of an embryo. We call this data one-dimensional integrated data or integrated data for 10% strip (Myasnikova *et al.*, 2001).

There are several steps in the construction of an 1D spatiotemporal atlas of segmentation gene expression:

Step 1: Confocal imaging. In the experiments, gene expression was measured using fluorescence tagged antibodies as described (Kosman *et al.*, 1998). The gain of the confocal microscope was adjusted as follows. Each gene is expressed at maximum intensity at a characteristic time, which corresponds to a distinct spatial pattern that can be recognized by an experienced human observer. The gain was adjusted for each gene product such that a few individual pixels from a domain expressed at maximum intensity were at 255 on the 8-bit scale. This permits the comparison of the levels of gene expression at different times, and the averaging of data between embryos stained with the same sera in different experiments performed on different days. Each gene product was detected in a single channel of a confocal microscope, and for each channel two raw images were made corresponding to two optical sections of an embryo separated by two microns. These images were averaged, cropped and rotated to yield an embryo image that displays the expression pattern of a single gene in a given embryo.

Step 2: Image segmentation (Kosman, 1999, http://urchin.spbcas.ru/FlyEx/proc_steps/dave.html). Each embryo was scanned for the expression of three genes at a time. Three embryo images were combined and the resultant image was segmented to construct a binary nuclear mask. The mask is a one bit deep image in which a pixel is equal to one if it is on nucleus and zero otherwise. This mask was used to determine the average x and y coordinates of each nucleus and to estimate the average fluorescence level of each of three gene products.

Step 3: Data normalization (E. Myasnikova *et al.*, submitted for publication). At this step quantitative gene expression data were rescaled in order to get rid of distortions caused by the presence of a background signal. The method for removal of the background signal was based on the observation that the level of a given gene expression in a null mutant embryo for that gene is well fit by a very broad 2D dimensional paraboloid. The background paraboloid was automatically determined from the areas of wild-type embryos in which a given gene is not expressed and used to remove background from the entire embryo.

Step 4: Temporal characterization (Myasnikova *et al.*, 2002). We acquired gene expression data from fixed embryos. Thus the temporal dynamics must be reconstructed from many samples, each at a different stage of development. A fundamental step in such reconstruction is the determination of the developmental age of each embryo. Cleavage cycle was used to stage embryos prior to cycle 14A, as each cleavage cycle from 9 to 13 is relatively short in duration. However, cleavage cycle 14A is about 50 min long and therefore during this cycle other markers were used for staging embryos. As

a preliminary step, we have divided all the embryos within cleavage cycle 14A into eight temporal equivalence classes on the basis of thorough visual inspection of the expression pattern of the *even-skipped (eve)* gene, which was scanned in all embryos and has a very dynamic pattern. The operational definition of a temporal equivalence class is that an experienced observer cannot see clear expression pattern differences among embryos belonging to a given class. The 8 classes are approximately equally populated, and each class represents a little over 6 min. We selected within cycle 14A for scanning without regard for age, so we expect our dataset to be uniformly distributed in time.

Step 5: Registration (Myasnikova et al., 2001; Kozlov et al., 2002, <http://www.bioinfo.de/isb/2002/02/0011>). To eliminate small individual differences quantitative gene expression data were subjected to registration. Two registration methods were used. Both methods are based on the extraction of Ground Control Points (GCPs). For GCPs the extrema of the expression pattern of the *eve* gene were used. The affine coordinate transformation was applied to make the corresponding GCPs in different images coincide as closely as possible. In one registration method (the spline or SpA method) GCPs were extracted by a quadratic spline approximation, while in the other method (FRDWT or wavelet) the fast dyadic redundant wavelet transform was used.

Step 6: Averaging (Myasnikova et al., 2001). To get 1D reference data at cellular resolution and at each time point we use registered expression data from a given segmentation gene in the central 10% strip. These data were grouped according to A–P position in 100 bins and averaged within each bin. Since a single nucleus is very close to 1% egg length in diameter, this procedure maps the averaged gene expression data onto one row of nuclei.

Though the expression of segmentation genes is largely a 1D function of position on the x -axis, the examination of 2D patterns can supply biologically important information about the spatiotemporal organization of segmentation gene expression domains. However due to individual variations and limitations of the confocal imaging procedure, which permits us to scan for only three genes at a time, it is impossible to perform a direct comparison of patterns from individual embryos. This problem can be solved by construction of a reference pattern for each segmentation gene and each time point. We call these patterns integrated patterns.

The method for the construction of integrated pattern is presented in detail (Kozlov et al., 2002). It is based on creating an averaged model that reproduces the spatial distribution of nuclei over an embryo image. This knowledge is used to construct a nuclear model of a pattern. The number of nuclei in this model is equal to the mean value of nuclei in embryo images. At the next stage, the average level of expression of each gene in each averaged nucleus is computed from a series of embryos of the same age.

Data model

Choice of a model The relational data model continues to be the most popular data model since the beginning of the eighties. Several reasons are responsible for its wide use (Codd, 1970; Kuznezov, 2003, <http://www.citforum.ru/database/osbd/contents.shtml>; Zelenkov, 1997, <http://alpha.netis.ru/win/db/toc.html>). First, the relational model is well grounded mathematically, providing predictability of data manipulation results. Second, the relational model is economic in basic concepts. The availability of a small set of abstractions allows accurate formal definitions and hence makes it possible to model the knowledge domain easily.

In addition there are practical advantages to this model because of the availability of relational database management systems (RDBMSs). High-quality RDBMSs are available commercially which support a variety of system functions automatically, such as data recovery after failure and the provision of simultaneous access of several users to shared data. Other database services, such as transaction support and parallelism are also well developed and implemented.

The expressiveness of the SQL query language, which was designed for the management of data stored in RDBMS, enables non-navigational data manipulation and does not require any knowledge about the physical organization of data in external memory. An additional advantage is the availability of a standard for SQL.

A relational approach to the modeling of gene expression data *in situ* is ideal both in terms of system development, and in terms of the structure of the data itself. As discussed above, most gene expression data at our disposal is quantitative in nature, and hence our knowledge domain can be naturally and completely described in terms of relations.

Although an important drawback of the relational model is poor representation of hierarchical structures and inadequate support of recursion, these drawbacks are insignificant for the modeling of gene expression data *in situ* as there is no need to use hierarchical structures or recursion to represent this information.

We have chosen to use the IBM DB2 RDBMS system to implement the database because it supports the necessary functionality and reliability of the system and requires minimum familiarization efforts from developers. Because our architecture is highly modular, it is possible to change to any other RDBMS in the future, should that be desirable.

Model of the subject domain There are two types of data in FlyEx: integrated data, which is reference data, and data from individual embryos.

Figure 1 presents the model of data from individual embryos. This model centers around the notion of an embryo. Each embryo is characterized by a unique name, developmental time (defined by cleavage cycle number and time class), genotype and fly line. In each embryo several genes are scanned. Data for each gene are collected in the different

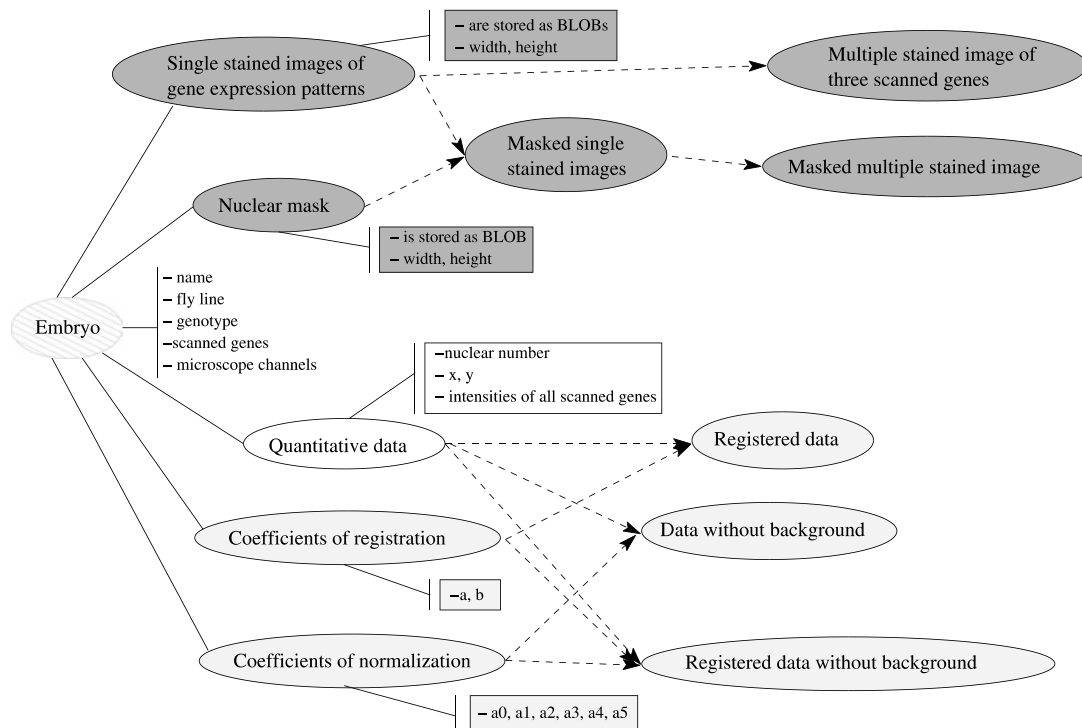


Fig. 1. Model of data from individual embryos. Dashed lines define operations performed on-the-fly. Different data types are shown in different colors: images of gene expression patterns (dark gray), quantitative data (white), processed data (light gray).

channels of a confocal microscope. The number of channels is always at least one. Each channel captures the fluorescence light emitted from a particular fluorophore.

Confocal imaging and image segmentation of an embryo generate several images. Of particular importance is an image that displays the expression pattern of a single gene in a given embryo (single stained image). This image presents a *Drosophila* embryo in its canonical orientation: anterior end to the left and dorsal side to the top.

An image of nuclear mask (nuclear mask) displays where the individual nuclei of an embryo are located and can be obtained by different methods.

Both single stained image and nuclear mask have a unique height and width, which are measured in pixels. These images are stored in the database as binary large objects (BLOBs). The RDBMS does not have any information about the structure of a BLOB and interprets it as a stream of bits. All steps of BLOB processing (from insertion into the database to visualization) are done by in-house tools external to the database (see System architecture section). To accelerate information retrieval BLOBs are stored in separate relational tables.

The single stained images obtained for an individual embryo are used to generate an image on-the-fly, which displays the expression patterns of all the genes scanned in the embryo. This image is called a multiple stained image. Each expression pattern is displayed in different color.

Multiplication of the nuclear mask of an embryo by single stained image of a gene scanned in this embryo generates masked single stained image on-the-fly. This image displays embryo nuclei, in which the gene is expressed, and turns pixels between nuclei to black. To generate image, called as masked multiple stained image, masked single stained images of all the genes scanned in the embryo are combined on-the-fly.

The nuclear mask is used for the acquisition of the quantitative data on gene expression in one embryo. These data consist of a series of data records, one for each nucleus. Each nucleus is characterized by a unique identification number, the x and y coordinates of its centroid and the average fluorescence levels of expression of three genes scanned in the embryo. The x and y coordinates coincide with the A–P and dorso–ventral (D–V) axes of the embryo, respectively, and are expressed as a percentage of the maximum size of the embryo on these axes. This is important, because it compensates for size differences from embryo to embryo.

Data without background and registered data are generated from quantitative data on-the-fly by use of the coefficients of normalization and registration correspondingly. In general both data without background and registered data can be produced by different methods. These data types have one format which is similar to those of quantitative data.

FlyEx contains two types of reference data: integrated data for 10% strip and integrated pattern. Both data types

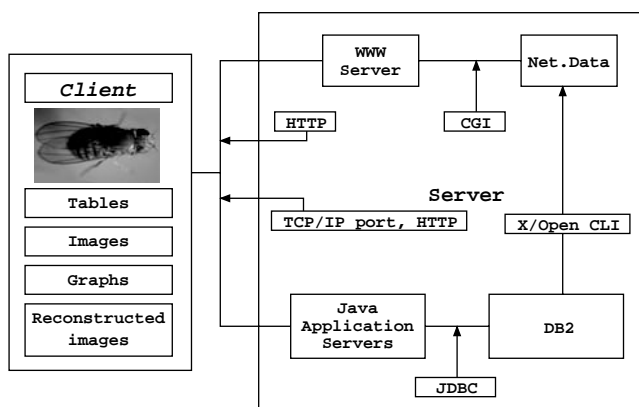


Fig. 2. The architecture of FlyEx.

are characterized by genotype and fly line, gene name, developmental time, the channel of confocal microscope and the list of embryos used for the construction of data.

The integrated data for 10% strip consist of a series of data records, one for each averaged nucleus. Each nucleus is characterized by a unique identification number, the x coordinate of its centroid, the averaged value of a given gene expression and the value of standard deviation, which characterizes the variability of gene expression in the nucleus. There are six types of integrated data for 10% strip in FlyEx, generated by different methods, which include different processing steps.

The integrated pattern consists of a series of data records, one for each nucleus of the nuclear model of a pattern. Each nucleus is characterized by a unique identification number, the x and y coordinates of its centroid and the average value of a given gene expression. Currently these data are generated by one method, which includes wavelet registration and removal of background signal.

System architecture

Figure 2 presents the architecture of FlyEx. The key component of it is IBM DB2 server (v 7.2), which runs under Linux Red Hat. HTML, Java scripts and Java applets were used to implement the web interface of a user.

There are several components between user interface and the database, namely WWW server, Net.Data and Java application servers. The Apache WWW server receives and processes requests from a user. Net.Data receives requests from WWW server via CGI interface, interacts with the database via OCLI interface and forms a HTML document. Java application servers interact with user interfaces and the database via HTTP protocol and JDBC respectively.

Java application servers Several Java application servers were implemented. The download server works to download gene expression patterns, quantitative, processed (data without background and registered data) and integrated data from the FlyEx database. The data can be downloaded in

different formats (ASCII, HTML, XML) and organized in different ways according to needs of a user.

The ImageServer is designed to process the images of gene expression patterns (A. Pisarev, E. Myasnikova, E. Poustelnikova and M. Samsanova). Several basic operations are supported, in particular, combination of two or three gray-scaled images into color multiple stained image, generation of absolute value of difference between two images, masking of one image by the other. Some operations, e.g. conversion of graphic formats, image scaling, presentation of an image as the stream of bytes and vice versa are implemented by use of the JMagick package, which represents a Java interface to the ImageMagick package. These packages are publicly available (Yeo, 2003, <http://www.yeo.nu/jmagick/>; Still, 2003, <http://www-106.ibm.com/developerworks/library/l-graf/?ca=dnt-428>). In addition the ImageServer supports cuts of rectangular area, the filtering of fluorescence intensity and contrast enhancement.

Besides standard procedures, which are applicable to the analysis of any type of images, the ImageServer supports the subject domain oriented methods for image processing. Currently background removal and image registration can be performed, the last operation brings images to unified scale for further comparison.

While by default a resultant image is in JPEG format, a special parameter allows to specify any format of the image (GIF, TIFF, PNG, BMP, etc.), if necessary.

The NLP server (Samsonova *et al.*, 2003) processes natural language queries to the database and currently supports Russian and English. The query in natural language can contain different operations, e.g. relation ('more than', 'less than', etc.), intervals ('from n to m '), logic ('and', 'or', 'not'), aggregation ('how many ...?'), etc. The processing of the natural language query includes: the search of word forms in dictionaries, the identification of semantic components, SQL query generation and optimization, as well as formatting of query result as a table, graph or reconstructed image.

The ABBAS server is designed to identify synonyms and higher or lower level concepts in the natural language query and transform these terms to the logical level terms of the subject domain. Besides, this server has additional functions. First, it processes queries, which were constructed visually by interactive selection of concepts on the conceptual schema (Samsonova *et al.*, 2003). Second, it periodically monitors the functioning of system services and forms diagnostic messages in case of failure.

The DBA server is implemented to process SQL queries to the database and has several functions. It processes SQL queries for alphanumeric information and data in CLOB and BLOB formats, as well as administrator's requests for design and modification of tables; formats the alphanumeric data requested by a user as a table; processes requests for the insertion of new alphanumeric data, as well as new data in CLOB and BLOB formats into the database.

It should be stressed that the architecture of the system provides for a user interface, which is based on standard web browsers and HTTP protocol, and allows to use both FireWall and Proxy servers.

Visualization of quantitative and processed data We have designed Java applet to visualize quantitative and processed data (i.e. registered data, data without background or registered data without background). Quantitative gene expression data are loaded on a client side and used to generate processed data on-the-fly.

To display registered data, the x coordinate value of each nucleus i of a given embryo is subjected to the affine transformation $x'_i = \rho x_i + \Delta$ where ρ and Δ are registration coefficients for the embryo.

The data without background are generated from quantitative gene expression data by linear mapping, which transforms fluorescence at or below background level to zero and maximum possible fluorescence to itself:

$$a_i^{(n)} = \max \left(\frac{a_i - S(x_i, y_i)}{255 - S(x_i, y_i)} 255, 0 \right), \quad i = 1, \dots, N \quad (1)$$

where $S(x, y) = a_0x^2 + a_1y^2 + a_2x + a_3y + a_4xy + a_5$.

The registered data without background are generated on a client side by a combination of two processing procedures described above.

Visualization of the variation of gene expression during cycle 14A FlyEx contains reference data on expression of segmentation genes at 9 time points, namely at cycle 13 and 8 time points of cleavage cycle 14A. The latter cycle is of particular importance as the period of determination of segments. We have designed the Java applet, which displays the smooth variation of gene expression during cycle 14A. For each segmentation gene the temporal variation of expression is interpolated by the Lagrange method from the values of integrated data for 10% strip or from the values of integrated pattern at 8 time points of cycle 14A. A total of 10 additional time points are generated for each time interval. The time delay between the sequential display of each of 80 datasets may be varied from 1 ms to 2 s.

IMPLEMENTATION

Database contents

The addition of data to FlyEx is ongoing; currently the database contains data for 954 embryos. We store 2832 single stained images, which display the expression patterns of 14 segmentation genes. The nuclear mask is available for each embryo. All embryos contained in FlyEx belong to blastoderm cleavage cycles from 10 to 14A (Foe and Alberts, 1983). At present, FlyEx contains 2 073 662 quantitative data records.

FlyEx stores the coefficients of registration and normalization to generate processed data from quantitative gene expression data on-the-fly. Two registration coefficients are

available for each registration method and for each embryo belonging to temporal classes from 2 to 8 of cleavage cycle 14A. The normalization coefficients are stored for each gene scanned in an embryo.

FlyEx contains reference data of two types: one-dimensional data for the central 10% strip of y -values along the mid-line of an embryo in the A–P direction and data for the 2D projection of an embryo. The integrated data are stored in FlyEx for each temporal class and only for those genes which have been stained in more than nine embryos. Further experimental work will fill in sparsely populated groups.

The web interface

There are several ways to access information stored in FlyEx. First, query forms are designed to select each data type, download data from the database and analyze gene expression information. Second, the information can be extracted by sequential browsing.

Two kinds of search forms are provided. One of them allows retrieval of data by embryo name, while the other permits selection of data by metadata information (e.g. gene name(s), cleavage cycle, time class, data types, etc.). In this case the embryo list is dynamically generated, and from this list the data for individual embryos can be easily accessed or downloaded from the database.

Besides, queries to the FlyEx database can be formulated in natural language or constructed visually by use of the conceptual schema (Samsonova *et al.*, 2003). To formulate and execute the natural language query, the HTML form ‘Natural Language Interface’ is to be filled by the user.

Similar to sequential browsing, the visual construction of queries is designed to support work of a beginner, who uses the conceptual schema of information on the expression of segmentation genes in *Drosophila*. This schema presents basic concepts in the domain, as well as relations between them as a graph. When a query is constructed visually, the user selects the concepts of interest by a mouse click. Upon selection the user submits the query by pressing the Send Query button. After submission of the query the user is presented with a list of predefined queries, which are sorted according to their relevance. These queries can be used to retrieve additional information from the database, which will refine the term meaning and make completely explicit structure and content of the information.

After retrieval from the database, images of gene expression patterns (single stained images, multiple stained images, masked single stained images, masked multiple stained images) can be scaled, subjected to contrast enhancement or filtered by intensity. Cuts of rectangular area are also supported. These operations improve the visualization of images and permit selection of the regions where the expression level of at least one (or each) gene exceeds a predefined threshold or lies in a predefined interval.

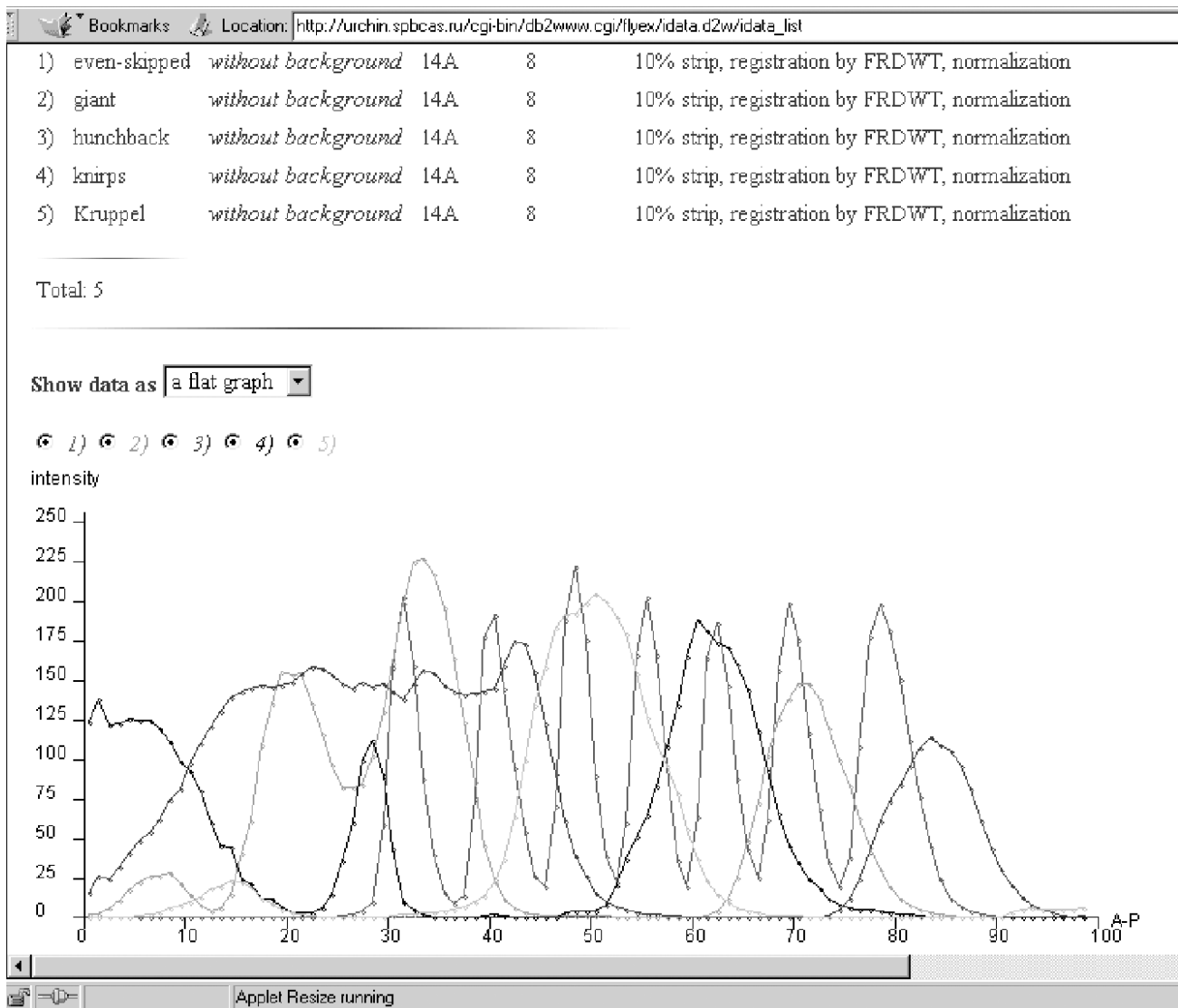


Fig. 3. The 1D atlas built from integrated data for 10% strip displays the spatial variation of segmentation gene expression at temporal class eight.

The quantitative and processed data on gene expression in individual embryos are retrieved from FlyEx in the form of dynamically generated tables, 3D graphs or reconstructed images. The graphs display spatial information about the level of expression of each of three scanned genes in the 2D projection of a given embryo, while the reconstructed image shows expression patterns of these genes. Special options are available for interactive modification and adjustment of graphical displays, namely, the selection of a gene expression pattern for visualization, cuts of a strip along the A–P axis of an embryo, intensity filtering and zoom in the x direction.

The integrated data for 10% strip can be displayed as a table or flat graph (Fig. 3). We provide an opportunity to display up to eight gene expression patterns as one flat graph.

The integrated patterns can be presented to a user as a table, flat graph, 3D graph or reconstructed image, in which not more than three patterns can be visualized on the reference embryo. Special options permit to interactively display each of three patterns in different colors. This facilitates the comparison of relative positions of segmentation gene expression domains in the reference embryo.

The variation of gene expression in the reference 10% strip during cycle 14A is visualized as a flat graph. Two additional methods are used to visualize this variation in the reference embryo, namely 3D graph and reconstructed image.

All information stored in FlyEx can be downloaded for further analysis. To do it the user has to specify data types, mode

of information output, output file formats (ASCII, XML, or HTML for data and JPEG, GIF, TIFF or BMP for images), as well as the email address for confirmation of the completion of download.

Application

The main function of FlyEx is to answer the user's questions about the dynamics of formation of segmentation gene expression domains.

The Analysis tools form provides access to a variety of operations on images of gene expression patterns and data. Currently these operations enable a user to address the following questions:

- How are the domains of segmentation genes localized in relation to each other at different developmental times? The answer is provided by the visualization of a spatiotemporal atlas segmentation gene expression. Two different atlases can be constructed from data stored in FlyEx, namely the 1D atlas, which is built from integrated data for 10% strip (Fig. 3), and the second atlas composed of the integrated patterns of segmentation genes.
- What is the quality of gene expression data? The quality of quantitative gene expression data can be estimated by examination of nuclear masks. The registration methods can be evaluated by comparing gene expression data before and after registration for embryos belonging to the same temporal class. The level of background signal can be assessed by visualization of quantitative data and data without background for a given embryo as one graph or by examination of the image, which was obtained by subtraction of the image without background from the corresponding single stained image.
- How are expression domains of a given gene formed with time? This question can be resolved by the comparison of integrated data for 10% strip or integrated patterns for a given gene at different time points.
- How variable is the expression of a given gene? Does this variability change with time or in space? These problems can be addressed in three ways. First, 1D quantitative data for a given gene and for individual embryos of the same age can be compared with each other and with corresponding 1D reference data. Second, a user can examine the image, which was obtained by subtracting one single stained image of a given gene from another image of the same age, or by subtracting any single stained image from the corresponding reference pattern. Third, three single stained images of a given gene and of one age, or two single stained images of a given gene and of one age and the corresponding reference pattern can be combined to reveal the difference in position of expression domains.

DISCUSSION

We describe the construction of the FlyEx database containing spatiotemporal information on the expression of *Drosophila* segmentation genes at cellular resolution.

In FlyEx gene expression is measured at the protein level and is considered to be proportional to protein concentrations. While at present the majority of studies of differences in gene expression measure mRNA abundance, the protein-based gene expression data is especially important, because it bears information about the final expression product, rather than about an intermediate. However, there is no question that protein- and RNA-based measurements are complementary and we plan to insert into FlyEx data on segmentation gene expression measured at mRNA level.

Currently FlyEx contains images and quantitative data on expression of 14 segmentation genes. In future, we plan to collect data for *en*, *wg* and other segmentation genes.

All the embryos contained in FlyEx have been divided into eight temporal equivalence classes. We expect our dataset to be uniformly distributed in time and each class to represent a little over 6 min of development. However, in fact the ages of embryos could be scattered non-uniformly over the whole cycle 14A. The experimental detection of embryo ages by measuring a degree of membrane invagination, as well as the development of an automatic method for the age detection of an embryo are now in progress and this work will form a basis for the more accurate temporal characterization of embryos.

FlyEx is designed conceptually as a quantitative atlas of segmentation gene expression at cellular resolution. Proceeding from the observation that the expression of segmentation genes is largely a function of position along the A–P axis of the embryo body, we have constructed two spatiotemporal atlases. The first 1D atlas is built from integrated data for 10% strip; the second is composed from the integrated patterns of segmentation genes. Our strategy is to store along with the reference data, the data derived from individual embryos; images of gene expression patterns, quantitative gene expression data and other types of processed data underlying averaged information. Besides its scientific significance this information will enable other researchers to scrutinize the quality of methods employed to obtain reference information on gene expression.

The most critical issue in a database design is the information access. We provide several ways to access the information in FlyEx. Both sequential browsing and search forms can be applied. When a search form is used, queries can be formulated either by interactive specification of search parameters or in the natural language. In addition, queries can be constructed visually by the selection of one or several concepts of interest on the conceptual schema. Different methods for information retrieval adapt the interface of FlyEx to professional experiences of various users.

FlyEx provides a variety of approaches to visualize data. Of particular importance is its capability to support operations on the images of gene expression patterns. We have implemented

both the standard procedures, which are applicable to analysis of any type of image, and the subject domain oriented methods for image processing. These operations improve the visualization of images and allow analysis of gene expression information.

Recently, we studied the applicability of the array DBMS RasDaMan to the management of gene expression data *in situ* (Pisarev et al., 2003). This DBMS extends standard SQL92 with multidimensional expressions, which support high-level operations on images. However, according to our overall experience, the formulation of new queries to images in RasDaMan requires an extensive programming work. The system architecture, which we propose in this study, easily allows addition of any new image processing methods in future. These methods can be applied to on-line analysis of any image information, provided that ImageMagick and JMagick are installed on the server.

Quantitative gene expression data, processed and integrated data are presented as graphs or reconstructed images. A user can interactively modify and adjust the graphical displays in various ways to facilitate the analysis of data. Evidently, both graphs of gene expression patterns and embryo images permit to nominate FlyEx as an atlas of gene expression *in situ*.

The main function of FlyEx is to answer questions about data. Currently a user can examine the quality of data, analyze the dynamics of formation of segmentation gene expression domains, and estimate the degree of variability of each gene expression.

It is widely believed that the application of mathematical and computational methods to the analysis of gene expression data will help to understand the mechanisms of cellular function and dynamical behavior. FlyEx contains quantitative gene expression data obtained in large scale in one laboratory and by one method. These features makes FlyEx a valuable resource for theoretical studies available as a on-line freeware. To facilitate work with data stored in FlyEx we provide the capability to download data of interest in different formats.

ACKNOWLEDGEMENTS

We thank David Kosman, Ekaterina Myasnikova, Konstantin Kozlov, Svetlana Surkova and Anastasia Samsonova for key contributions to this work. We are grateful to Alexander Samsonov for his criticisms on various versions of this paper. This work is supported by NIH grant R01 RR07801 and GAP award RBO-1286.

REFERENCES

- Blagov,M., Pisarev,A., Myasnikova,E., Poustelnikova,E. and Samsonova,M. A system for on line processing and analysis of the images of gene expression patterns. (*in preparation*).
- Campos-Ortega, J.A. and Hartenstein,V. (1985) The Embryonic Development of *Drosophila melanogaster*. Springer-Verlag, Berlin.
- Codd,E.F. (1970) A relational model of data for large shared data banks. *Commun. ACM*, **13**, 377–387.
- Davidson,D. and Baldock,R. (2001) Bioinformatics beyond sequence: mapping gene function in the embryo. *Nat. Rev. Genet.*, **2**, 409–417.
- Foe,V.A. and Alberts,B.M. (1983) Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in *Drosophila* embryogenesis. *J. Cell Sci.*, **61**, 31–70.
- Hartenstein,V., Lee,A. and Toga,A. (1995) A graphical digital database of *Drosophila* embryogenesis. *Trends Genet.*, **11**, 51–58.
- Hecksher-Sorensen,J. and Sharpe,J. (2001) 3D confocal reconstruction of gene expression in mouse. *Mech. Dev.*, **100**, 59–63.
- Kosman,D., Reinitz,J. and Sharp,D.H. (1998). Automated assay of gene expression at cellular resolution. In *Proceedings of the 1998 Pacific Symposium on Biocomputing*. World Scientific Press, Singapore, pp. 6–17.
- Kosman,D. (1999) Acquisition of quantitative data on gene expression in individual embryos.
- Kozlov,K., Myasnikova,E., Samsonova,M., Reinitz,J. and Kosman,D. (2000) Method for spatial registration of the expression patterns of *Drosophila* segmentation genes using wavelets. *Comput. Technol.* **5**, 112–119.
- Kozlov,K., Myasnikova,E., Pisarev,A., Samsonova,M. and Reinitz,J. (2002) A method for two-dimensional registration and construction of the two-dimensional atlas of gene expression patterns *in situ*. *In Silico Biol.*, **2**, 125–141.
- Kuznezov,S.D. (2003) Principles of modern data bases.
- Myasnikova,E., Kosman,D., Reinitz,J. and Samsonova,M. (1999) Spatio-temporal registration of the expression patterns of *Drosophila* segmentation genes. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 195–201.
- Myasnikova,E., Samsonova,A., Kozlov,K., Samsonova,M. and Reinitz,J. (2001) Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods. *Bioinformatics*, **17**, 3–12.
- Myasnikova,E., Samsonova,A., Samsonova,M. and Reinitz,J. (2002) Support vector regression applied to the determination of the developmental age of a *Drosophila* embryo from its segmentation gene expression patterns. *Bioinformatics*, **18**(Suppl. 1), s87–s95.
- Myasnikova,E., Samsonova,M. and Reinitz,J. Removal of background signal from *in situ* data on the expression of segmentation genes in *Drosophila* (submitted for publication).
- Nusslein-Volhard,C., Wieschaus,E. and Kluding,H. (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. I. Zygotic loci on the second chromosome. *Roux's Arch. Dev. Biol.*, **193**, 267–282.
- Pisarev,A., Poustelnikova,E., Samsonova,M. and Baumann,P. (2003) Mooshka: a system for management of multidimensional gene expression data *in situ*. *Inform. Technol.*, **28**, 269–285.
- Ringwald,M., Mangan,M.E., Eppig,J.T., Kadin,J.A. and Richardson,J.E. (1999) Gxd: a gene expression database for the laboratory mouse. *Nucleic Acids Res.*, **27**, 106–112.
- Samsonova,M., Pisarev,A. and Blagov,M. (2003) Processing of the natural language queries to a relational database. *Bioinformatics*, **19**, (Suppl. 1), i241–i249.
- Simcox,A.A. and Sang,J.H.(1983) When does determination occur in *Drosophila* embryos. *Dev. Biol.*, **97**, 212–221.

- Still,M. (2003) Graphics from the command line. IBM developerWorks, July 16, 2003.
- Tomancak,P., Beaton,A., Weiszmam,R., Kwan,E., Shu,S.Q., Lewis,S.E., Richards,S., Ashburner,M., Hartenstein,V., Celniker,S.E. and Rubin,G.M. (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biology*, **3**, research0088.1–research0088.14.
- Wieschaus,E., Nusslein-Volhard,C. and Jurgens,G. (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. III. Zygotic loci on the X-chromosome and fourth chromosome. *Roux's Arch. Dev. Biol.*, **193**, 296–307.
- Yeo,E. (2003) JMagick.
- Zelenkov,Y.A. (1997) Introduction into data bases.