BIOINFORMATICS

Vol. 18 Suppl. 1 2002
Pages S87–S95

Support vector regression applied to the determination of the developmental age of a Drosophila *embryo from its segmentation gene expression patterns*

E. Myasnikova [1], *A. Samsonova* [2], *M. Samsonova* [1,*] *and J. Reinitz* [3]

[1] *St.Petersburg State Technical University, 29 Polytechnitcheskaya ul., St.Petersburg, 195251, Russia,* [2] *European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton-Cambridge, CB10 1SD, UK and* [3] *University at Stony Brook, Stony Brook, NY, 11794-3600, USA*

Received on January 19, 2002; revised on April 1, 2002; accepted on April 1, 2002

ABSTRACT

Motivation: In this paper we address the problem of the determination of developmental age of an embryo from its segmentation gene expression patterns in *Drosophila*.

Results: By applying support vector regression we have developed a fast method for automated staging of an embryo on the basis of its gene expression pattern. Support vector regression is a statistical method for creating regression functions of arbitrary type from a set of training data. The training set is composed of embryos for which the precise developmental age was determined by measuring the degree of membrane invagination. Testing the quality of regression on the training set showed good prediction accuracy. The optimal regression function was then used for the prediction of the gene expression based age of embryos in which the precise age has not been measured by membrane morphology. Moreover, we show that the same accuracy of prediction can be achieved when the dimensionality of the feature vector was reduced by applying factor analysis. The data reduction allowed us to avoid over-fitting and to increase the efficiency of the algorithm.

Availability: This software may be obtained from the authors.

Contact: samson@fn.csa.ru

Keywords: gene expression patterns; development; embryo staging; support vector regression; segmentation genes; Drosophila.

INTRODUCTION

As the assembly of full nucleotide sequence is completed in an ever-growing list of organisms, the focus of genomic research is shifting from structural areas to functional ones. Functional genomics itself is expanding from the study of low level function like enzymatic specificity to higher level functions seen only in metazoa such as immune response, neural function, and development. For all metazoan functions, particularly development, spatio-temporal information on gene expression is essential data for bridging the gap between the DNA sequence of the gene and its physiological function. Moreover, the fundamental fact that the cells of a metazoan organism contain the same genetic material but express different genes indicates that such data must be obtained at cellular resolution in space, and at a temporal resolution that is close to the characteristic time for changes in gene expression. One useful way to obtain such data is by means of fluorescent *in situ* hybridization and immunofluorescence histochemistry. The analysis of such data raises many new questions. In this paper we address the problem of temporal resolution by providing new methods of temporal characterization.

The gene network responsible for segment determination in *Drosophila* was functionally characterized some years ago by entirely classical genetic methods (Nusslein-Volhard and Wieschaus, 1980; Jurgens *et al.*, 1984; Wieschaus *et al.*, 1984; Nusslein-Volhard *et al.*, 1984). The genes in this network act early in development. Following fertilization, 13 rapid and synchronous nuclear divisions take place under maternal control. Thus the early embryo can be coarsely staged by ‘cleavage cycle’, where cleavage cycle N lasts from the end of nuclear division $N - 1$ to the end of division N (Foe and Alberts, 1983). The cleavage cycles are 12 minutes or less in duration, except for cycle 14A, which is 50 minutes long and terminated by the onset of gastrulation, following which divisions are no longer synchronous. At cleavage cycle 10 the embryo becomes a hollow ellipsoid of cell nuclei which are not separated by

[*] To whom correspondence should be addressed.

© Oxford University Press 2002

membranes, called the syncytial blastoderm. Cell membranes invaginate and surround the nuclei in cycle 14A, and gastrulation begins when cellularization is complete. Zygotic segmentation genes are first transcribed at cycle 10. These genes are expressed in patterns which gain increasing spatial resolution over time (Akam, 1987; Ingham, 1988), and most of the gain in resolution happens during cycle 14A.

The segment determination process described above is a classical 'morphogenetic' field, in the sense that the nuclei of the blastoderm interact with one another so as to assign developmental pathways in a precise manner. No morphogenetic field has been completely mapped at the molecular level, but such maps are central to constructing a true functional genomics of development. We are engaged in an effort to characterize the functional genomics of development of the morphogenetic field controlling *Drosophila* segmentation by precisely mapping all of the regulators of the field at cellular resolution and describing the dynamics of the determination process. This effort involves both theory and experiment (Myasnikova *et al.*, 2001; Reinitz *et al.*, 1995a; Reinitz and Sharp, 1995b; Reinitz *et al.*, 1998; Sharp and Reinitz, 1998). The theoretical component uses mathematical modelling and simulations, while the experimental part employs automated image processing and data analysis methods to construct the integrated spatio-temporal map of gene expression at cellular resolution.

In the construction of a high resolution spatio-temporal atlas of gene expression, data are acquired from fixed embryos, each of which is stained by an immunofluorescence method to show the pattern of expression of three segmentation genes. Thus the temporal dynamics must be reconstructed from many samples, each at a different stage of development. A fundamental step in such reconstruction is to determine the developmental age of each embryo. At present a combination of experimental methods is used to solve this problem. Cleavage cycle is useful prior to cycle 14A, but during this cycle other markers must be used. For example, *in vivo* measurements (Merril *et al.*, 1988) of the degree of membrane invagination as a function of time give a standard curve that makes it possible to obtain developmental time from the degree of membrane invagination in fixed tissue. The highly dynamic nature of segmentation gene expression in cycle 14A suggests that a precise expression-based clock could be constructed by standardizing these patterns against membrane invagination.

Expert human observers can arrange expression patterns in temporal order by visual inspection, but cannot assign a precise clock time from egg deposition. Temporal reconstruction of the dynamics of gene expression is crucial for elucidating the physiological functions of genes and understanding the network of genetic interactions that underlie the process of normal ontogeny. Thus the development of an automated method for staging embryos will be a fundamental advance in interpreting gene expression information during early embryogenesis in *Drosophila*.

The solution of the temporal characterization problem described here is based on support vector (SV) regression. SV regression (Smola and Scholkopf, 1998) is a statistical method in pattern recognition theory, in which regression functions are created from a set of labelled training data. This method possesses certain advantages compared to classic regression. It has proved to be more flexible because it allows for the use of loss functions of various types and, in addition, nonlinearity is easily introduced into the model by applying kernels.

In this report, we show that SV regression provides an accurate determination of embryo age and enables us to attribute an age to any other embryo with a certain probability according to its gene expression pattern.

SYSTEM AND METHODS

The dataset

In our experiments gene expression was measured by a confocal scanning microscope using fluorescence tagged antibodies (Kosman *et al.*, 1998b). For each embryo a 1024 × 1024 pixel image with 8 bits of fluorescence data in each of 3 channels was obtained. Each embryo is scanned for the expression of three genes. The resulting image is segmented for nuclei, reducing the image to a table in which each nucleus is characterized by a unique identification number, the x and y coordinates of its centroid, and the average fluorescence levels of three proteins (Kosman *et al.*, 1997; Myasnikova *et al.*, 1999; Kosman, 1998a). The embryo is oriented so that the x axis corresponds to its anterior-posterior (A-P) axis and the y axis to its dorsal-ventral (D-V) axis. x and y coordinates are expressed as percent of the length of the embryo in the x and y directions. The quantitative gene expression data are further processed to yield averaged data on the expression of each segmentation gene, called integrated data. These processing steps include data registration and averaging (Myasnikova *et al.*, 2000).

At present, our dataset contains confocal scans of about 1400 embryos, of which 809 are wild type and belong to cycle 14A. Of these, all are stained for the pair-rule segmentation gene *even-skipped* (*eve*) and two other genes that vary among the dataset.

As the initial step of temporal characterization 809 wild type embryos from Oregon-R flies, which belong to cleavage cycle 14A, were divided by visual inspection of pair-rule gene expression patterns into 8 temporal equivalence classes (Myasnikova *et al.*, 2001). 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 expression pattern of *eve* is particularly important, since it is highly dynamic and each embryo is stained for that gene. We selected embryos within cycle 14A for scanning without regard for age, so we expect our dataset to be uniformly distributed in time. The 8 classes are approximately equally populated, and since cycle 14A is about 50 minutes long (Foe and Alberts, 1983), each class represents a little over 6 minutes.

The evolution of the expression patterns of *eve* during cleavage cycle 14A has the following features. Time classes 1, 2, and 3 do not have seven well-defined *eve* stripes and the number and location of stripes changes rapidly. The remaining groups (classes 4 to 8) do have seven well-defined stripes. After all the stripes are clearly visible their intensities increase in the posterior portion of the embryo. By the end of cycle 14A, all *eve* stripes have reached maximum and equal intensity and maximum sharpness.

In this work, we use SV regression to analyse the developmental ages of a set of 501 wild type cycle 14A embryos which belong to temporal classes 3 through 8. 103 of these embryos were rephotographed in Nomarski optics to visualize the morphology of the blastoderm. These data were used to measure the degree of membrane invagination for these embryos by means of manual separation of the region of interest and by texture analysis [Surkova et al., in preparation]. Using the standard curve giving membrane invagination as a function of developmental time (Merril *et al.*, 1988) the precise developmental age of each embryo was determined. The measured ages turned out to be distributed uniformly over the range from 20 to 60 minutes from the onset of cleavage cycle 14A. The 103 rephotographed and standardized embryos were used as a training set.

Methods

All the algorithms described in this paper are implemented in C running on Unix. The programs have been tested and used on SPARC/SunOS 5.0, PIII/Linux, and Alpha/DU 4.0E systems.

Algorithm In previous efforts in temporal analysis (Myasnikova *et al.*, 1999) we have considered the development of the gene expression in time as a discrete process and have sought merely to assign embryos to their appropriate temporal class. The time evolution of the patterns is, in fact, continuous, and the ages of embryos in the dataset are distributed uniformly over the whole of cycle 14A. To reconstruct the aging process in time we need to be able to detect the age of an embryo on the basis of knowledge about its gene expression patterns. The development of the method can be subdivided into two major stages, the first of which is the extraction of characteristic expression features of embryos of different age, and the second is standardization against morphological data.

Expression of segmentation genes is largely a function of position along the anterior-posterior axis of the embryo body, and so can be well represented in one dimension. The segmented data extracted from the central 10% horizontal strip running in an A-P direction on the midline of an embryo are presented as diagrams, demonstrating the variation of gene expression (Figure 1). We have shown that in the 1D case the most essential features of an expression pattern are the number and values of extrema of gene expression. We applied the quadratic spline approximation to extract these features from 1D patterns of *eve* expression as described (Myasnikova *et al.*, 2000). The method for feature extraction is illustrated in Figure 1 for representative expression patterns of embryos belonging to earlier (third) and later (eighth) temporal classes.

At the next stage we use the group of 103 embryos in which the precise developmental age was determined by measuring membrane invagination. For all these embryos the full set of exactly 13 extremal features was extracted. Each embryo of the group is thus characterized by a multidimensional vector containing as components the value of developmental age together with 13 parameters of the gene expression patterns. We then attribute an age to other embryos in the dataset on the basis of their gene expression patterns by SV regression.

Support vector regression The basic idea of SV regression is formulated in Vapnik (1995) and Smola and Scholkopf (1998). Suppose we are given training data presented by l observations (embryos). Each observation consists of a pair: a vector of characteristic features $\vec{x}_i \in \mathbb{R}^n$, $i = 1, \ldots, l$ and the associated 'truth' y_i (an embryo age), given us by a trusted source (membrane invagination). To generalize the SV algorithm to the regression case an analog of soft margin is constructed in the space of the target values $y \in \mathbb{R}$ by using Vapnik's ε-insensitive loss function $|\xi|_\varepsilon$ described by

$$|\chi|_\varepsilon = \begin{cases} 0 & \text{if } |\chi| \leq \varepsilon \\ |\chi| - \varepsilon & \text{otherwise.} \end{cases}$$

In other words, we do not take into consideration deviations of $f(\vec{x}_i)$ from the actually obtained targets y_i as long as they are less than ε, but will penalize any deviation larger than this. To estimate the linear regression $f(\vec{x}) = (\vec{w}, \vec{x}) + b$ one minimizes a regularized empirical risk functional

$$R_{reg}[f] = \frac{1}{l} \sum_{i=1}^{l} |y_i - f(\vec{x}_i)|_\varepsilon + \frac{1}{2}\|\vec{w}\|^2. \tag{1}$$

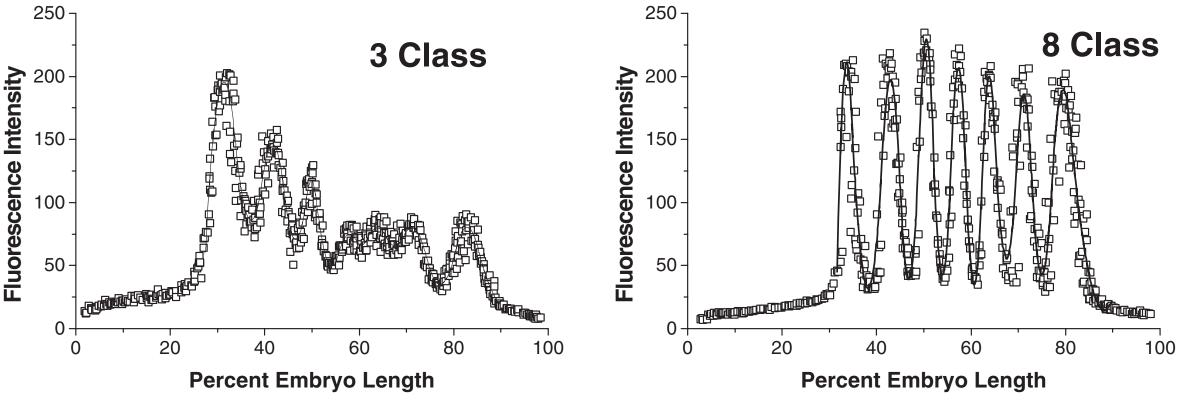

Fig. 1. Representative *eve* expression patterns in embryos belonging to 3rd (left panel) and 8th (right panel) temporal classes. The nuclear expression levels are shown as hollow squares; the solid line shows the approximating quadratic spline function.

The problem is transformed into the constrained convex optimization problem by employing slack variables ξ and ξ^*.

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{l}\left(\xi_i + \xi_i^*\right) \\ &\text{subject to} \quad \begin{cases} \varepsilon + \xi_i^* - (f(\vec{x}_i) - y_i) \geq 0 \\ \varepsilon + \xi_i - (y_i - f(\vec{x}_i)) \geq 0 \\ \xi_i, \xi_i^* \geq 0 \end{cases} \\ &\qquad\qquad \text{for} \quad \forall \quad i \in [1 \ldots l]. \end{aligned} \tag{2}$$

The constant $C > 0$ determines the trade off between the 'flatness' of f and amount up to which deviations larger than ε are tolerated.

It has been shown (Smola and Scholkopf, 1998) that (2) can be solved in its dual formulation, which moreover provides the key for extending SV regression to nonlinear functions by introducing kernels. The dualization is implemented by the standard method utilizing Lagrange multipliers, which yields a problem in convex quadratic programming:

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{2}\sum_{i,j=1}^{l}\left(\alpha_i - \alpha_i^*\right)\left(\alpha_j - \alpha_j^*\right)k\left(\vec{x}_i, \vec{x}_j\right) \\ & + \sum_{i=1}^{l}\left(\alpha_i - \alpha_i^*\right)y_i - \varepsilon\sum_{i=1}^{l}\left(\alpha_i + \alpha_i^*\right) \\ \text{subject to} \quad & \sum_{i=1}^{l}\left(\alpha_i^* - \alpha_i\right) = 0; \qquad \alpha_i, \alpha_i^* \in [0, C]. \end{aligned} \tag{3}$$

The functional (3) is optimized now with respect to the dual parameters α_i and α_i^*. The function f is hence presented via dual parameters by

$$f(\vec{x}) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(\vec{x}_i, \vec{x}) + b. \tag{4}$$

The kernel functions $k(\vec{x}, \vec{x'})$ are introduced into the model to make the SV algorithm nonlinear. To do this the training patterns $\vec{x_i}$ are preprocessed by a map $\Phi : \mathbb{R}^n \mapsto \mathbb{F}$ from the input space into some feature space and then the standard SV regression algorithm is applied. The kernel is defined by a dot product $k(\vec{x}, \vec{x'}) = \langle\Phi(\vec{x}), \Phi(\vec{x'})\rangle$. It has been shown that to correspond to a dot product in some feature space, $\mathbb{F}$, the functions $k(\vec{x}, \vec{x'})$ must satisfy the Mercer conditions (Schölkopf *et al.*, 1998).

It is easy to show that the simplest polynomial kernels k with an integer parameter p and $a \geq 0$

$$k(\vec{x}, \vec{x'}) = (\langle\vec{x}, \vec{x'}\rangle + a)^p \tag{5}$$

are suitable SV kernels. Note that with $p = 1$ and $a = 0$, the kernel (5) reduces to the dot product and model (3) turns into the linear SV regression model (2).

The convex quadratic problem (3) is solved by the sequential minimal optimization (SMO) algorithm proposed by Platt (1998). This is a chunking algorithm which iteratively selects from the set of parameters subsets of size two and optimizes the target function with respect to them. The key point of the method is that for two parameters the optimization subproblem can be solved analytically without explicitly invoking a quadratic optimizer.

RESULTS

Non-reduced data

The training set consisted of the 103 embryos for which the precise developmental age was determined and whose

Table 1. Results of testing SV regression with scalar and polynomial kernels applied to the training set of 103 embryos. For each experiment the value of the cost function (6) is calculated for different values of the model parameters at $\varepsilon = 0.05$. In brackets we give the maximal differences between the observed and predicted values of embryo age in minutes. The optimal values of the cost function are marked out

	Scalar	Polynomial ($p = 2$)			
C		$a = 0.3$	$a = 0.5$	$a = 1$	$a = 2$
1	1.93 [7.36]	1.84 [7.63]	1.83 [6.33]	1.78 [5.94]	1.64 [6.30]
2	1.85 [7.25]	1.83 [5.83]	1.65 [6.47]	1.65 [5.78]	1.56 [6.94]
10	1.74 [6.87]	1.49 [8.47]	1.44 [7.88]	1.45 [6.40]	1.69 [7.59]
20	1.73 [6.59]	1.60 [7.71]	**1.31** [5.58]	1.45 [5.49]	1.66 [6.91]
30	**1.70** [6.91]	1.40 [6.67]	1.77 [7.44]	1.71 [7.07]	1.47 [7.22]

pattern was characterized by exactly 13 extremal features. The results of SV regression analysis are presented in Table 1 for the dot product (or scalar) kernel and polynomial kernel of second order ($p = 2$). We have tested the model (3) for choosing the optimal values of the model parameter C and the parameter a of the polynomial kernel (5).

The minimal value of the cost function, defined by the empirical risk functional (1) with no regularization term

$$R_{emp}[f] = \frac{1}{l}\sum_{i=1}^{l} |y_i - f(\vec{x}_i)|_{\varepsilon}, \qquad (6)$$

is achieved at $C = 30$ with the scalar kernel and at $C = 20$ and $a = 0.5$ for the polynomial kernel. The results of regression estimation are presented in the left panel of Figure 2. We should note that the rate of convergence of the optimization procedure (3) is strongly dependent on the parameter C. At large values of C the convergence is slow, and so, for example, given $\varepsilon = 0.05$, the problem with the polynomial kernel at $C = 20$ requires more than 1000 iterations to converge. Although the polynomial model provides a better fit to the observed data than the scalar one, we cannot be sure that this model guarantees better prediction of ages for embryos not included in the training set. We do not possess any information about their precise developmental age, and hence we cannot judge the reliability of the predictive method solely from the quality of the regression fit.

In the usual procedure of cross-validation the training set is divided into two subsets, with the first one (the working set) used for regression estimation, and the second one for the prediction. Then the predicted values of the elements of the second set are compared to their observed values. As our training set is of a small size, we need to test the accuracy of the prediction by a more economical procedure. We consequently exclude, one by one, a single item from the training set and use all the rest as a working set, thus predicting the age of the excluded embryo. As a criterion of the quality of prediction the risk function (6) is used with the entries computed for the excluded items

$$\frac{1}{l}\sum_{i=1}^{l} |y_i - f_i(\vec{x}_i)|_{\varepsilon}. \qquad (7)$$

f_i here are the different functions each time newly estimated for the training set with the exclusion of ith embryo. We calculate the criterion (7) taking the optimal values of parameters from Table 1. The criterion takes the value 1.94 for the scalar kernel and 2.11 for the polynomial kernel. Thus while better fit to the measured data is achieved with the polynomial kernel, superior prediction is afforded by the model with the scalar kernel. Taking into account that the developmental ages vary over a range of 20 to 60 minutes from the onset of cycle 14A, the error value of about 2 minutes is small enough to permit one to conclude that the predictions are reliable, especially those of the linear model.

Next we used the optimal SV regression function (4) for the prediction of the expression-based age of embryos not included in the training set. The results are displayed in Figure 3 for 398 embryos. Their developmental ages were not experimentally measured, but in our previous work all these embryos were assigned to one of 8 temporal classes on the basis of visual inspection of their gene expression patterns, in particular that of *eve*. It is clear that such classification is not exact and thus can serve only as an indirect criterion of the quality of age prediction. Figure 3 shows good correlation with the predefined temporal classes, but it is evident from the figure that the prediction results for embryos belonging to late temporal classes are less accurate than for the early ones. This is true in particular for embryos from temporal classes 7 and 8, which are poorly discriminated by the prediction method. The possible reasons for this discrepancy will be discussed below.

To check the correspondence of predicted ages to the given temporal classes in a more precise manner we can conduct an analysis of variance (ANOVA). The purpose of this method is to test the statistical significance of differences in the means of groups, which in our case are temporal classes. By this test we can compare the estimated variance of embryo age due to the between-groups variability with the within-group variability. We can compare those two estimates of variance via the F test, which tests whether the ratio of the two variance estimates is significantly greater than 1. For our predicted data the F statistic with 5 and 392 degrees of freedom takes the value 508.72 for the scalar kernel and 446.61 for the polynomial one. These tests are both highly significant, with the p-level less than 10^{-6}, and so both provide a good

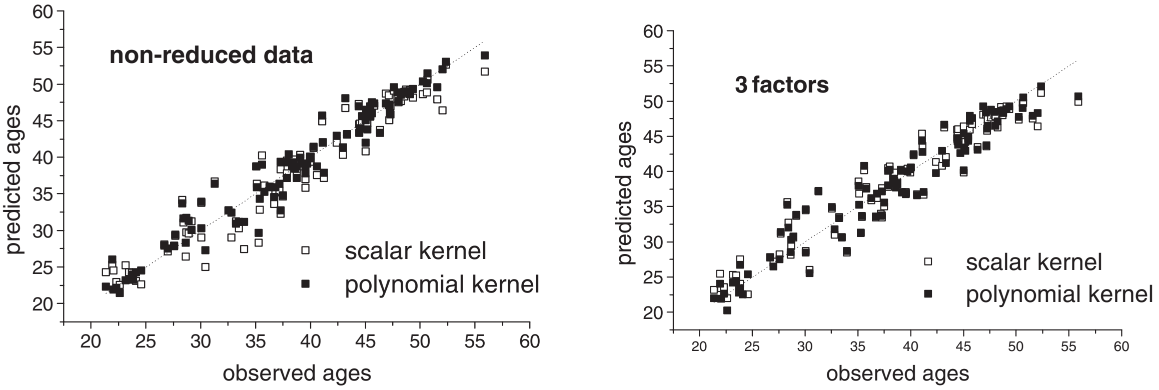

Fig. 2. The embryo ages (measured in minutes from the onset of cycle 14A) for the training set predicted by SV regression are given versus their observed values determined by measuring the degree of membrane invagination. The dotted lines is a locus of points corresponding to ideal cases where predicted and observed values coincide. On the left panel the non-reduced data are presented; on the right panel the results are given for data reduced to the 3 factor case.

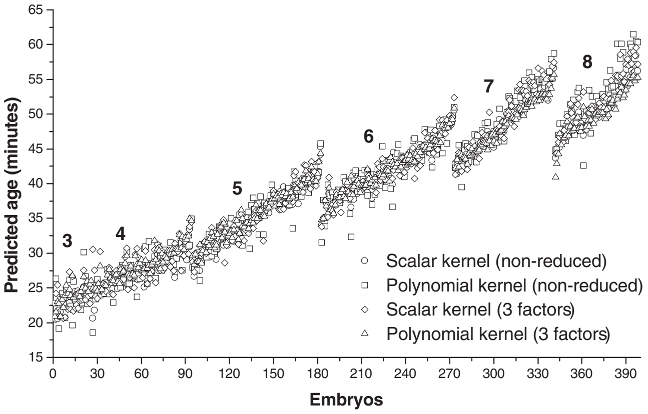

Fig. 3. Predicted values of ages of the embryos not belonging to the training set. The data are grouped according the predefined temporal classes from 3 to 8.

overall discrimination between the classes. Moreover, the *F* test may be used to compare to what extent the age predicted by each of the two methods correlates with the preliminary temporal classification. The *F* statistic for the scalar model takes on a value greater than for the polynomial one and this is also indirect evidence of the better accuracy of the scalar kernel model.

Proceeding from all of the above we can conclude that SV regression provides quite satisfactory results in age prediction. In what follows we shall show that at least the same accuracy of prediction can be achieved even if we restrict ourselves to a very small number of ‘new’ variables characterizing the embryo gene expression pattern instead of the 13 extremal features used so far. This will increase the efficiency of the algorithm as the rate of convergence grows very fast with decreases in the dimensionality of the problem. There exists one more very strong argument in favour of data reduction. As it has already been mentioned, the training set is of a small size relative to the dimension of the feature vector. This may cause the so-called over-fitting effect which in turn may cause unreliable prediction and lack of robustness with respect to changes in the algorithm parameters. Moreover, the 13 extremal features of the *eve* gene expression patterns are strongly correlated and hence redundant. All this indicates that it is necessary to reduce the dimension of the feature vector by constructing a new set of features and a good tool for this purpose is factor analysis.

Data reduced by factor analysis

Factor analysis as a method for the reduction of data dimension is based on the ideas of principal component analysis. Several correlated variables are linearly combined into one factor so as to maximize the variance of the ‘new’ variable (factor), while minimizing the variance orthogonal to the new variable. After the first factor, which maximizes the overall variance of the data, has been extracted we define another factor that contains the maximum amount of the remaining variability, and so on. In this manner, consecutive factors are extracted. Because each consecutive factor is chosen to maximize the variability that is not captured by the preceding factor, consecutive factors are orthogonal to one another, and hence independent. This procedure implies that the new factors are linear combinations of the initial variables.

We have extracted the full set of 13 new factors from the whole dataset of 501 embryos. For the 103 embryos belonging to the training set the first two factors have significant correlation with the observed ages, of 92%

and 25% respectively. This is a good reason to expect satisfactory results from SV regression even when only the first few factors are considered.

The model was tested by first taking into account only a single factor, then the first two factors, then three and so on, in order to find the optimal values of parameters. For up to 7 or 8 factors, the value of the cost function (6) proved to be weakly dependent on the parameter changes. The best results were achieved with C in the range from 0.8 to 3 for both kernel models and at a in the range from 0.5 to 3. As the number of factors was increased past 7, behaviour became more unstable. In Table 2 we present the results of SV regression for 1 to 13 factors, using $C = 1$ and $a = 2$. In the upper part of the table the values of the cost function (6) are given.

Close inspection of Table 2 reveals a number of interesting features. First, let us consider the full set of factors. Comparing the cost function values with those obtained for the non-reduced data from Table 1 we can see that for the scalar model the cost function calculated for 13 factors coincides with that from the previous section. This is natural because the new factors are linear combinations of the old variables and hence the old and new linear regression models coincide in this case. For the model with the polynomial kernel the value of the cost function is smaller than for non-reduced data (0.59 vs 1.31). But inspection of the lower part of the table, which presents the results of testing the accuracy of prediction within the training set, shows that for the high dimensional cases (more than 8 factors) the criterion of prediction quality takes on large values for the polynomial model. This is particularly true for the maximal deviations given in brackets, which are quite huge. In this case over-fitting is manifest and it is clear that not more than 8 factors should be used for regression estimation and prediction. Analyzing the goodness of prediction part of the table we can state that both models provide the best results when not more than 3 factors are included. Moreover the criterion (7) takes on approximately the same value of about 2.0[†]for both models.

The above results show that only three factors are required for the prediction of ages of the embryos not included in the training set. To discriminate among the models we take into account that the problem with the scalar kernel converges much faster (e.g., for 3 factors in 59 vs 197 iterations) and hence appears preferable.

We have used three factors to predict ages for 398 embryos not belonging to the training set. The results are presented in Figure 3. The method shows a good correlation of the predicted ages with the preliminary temporal classification, with the best discrimination achieved for the early temporal classes. The $F(5, 392)$ statistics calculated for the models with the scalar and polynomial kernels are equal to 522.85 and 474.48 respectively. We again obtain results very similar to those obtained for non-reduced data: the scalar model discriminates a bit better between the classes, while both models provide a very high significance on F-tests.

CONCLUSIONS

In this paper we address the problem of the determination of the developmental age of a *Drosophila* embryo from its segmentation gene expression patterns. At present a combination of expensive and time-consuming experimental methods is used to solve this problem. Knowledge of a precise embryo age is absolutely necessary to reconstruct temporal dynamics of gene expression and to decipher the network of genetic interactions that underlies early development in *Drosophila*.

We have developed a fast method for the automated staging of an embryo on the basis of its gene expression pattern by applying SV regression. The SV regression is a statistical method for creating regression functions of arbitrary type from a set of training data. In comparison with classic regression this method possesses certain advantages. Use of ε-insensitive loss function provides more flexible regression, as only deviations greater than ε are taken into account. In addition, nonlinearity is easily introduced into the model by applying kernels.

The training set is composed of those 103 embryos for which the precise developmental age was determined by measuring the degree of membrane invagination. The SV regression function was estimated by solving a problem in convex optimization programming. Testing the quality of regression on the training set showed good prediction accuracy. The optimal regression function was then used for the prediction of the expression based age of 398 embryos, the precise age of which had not been measured.

Due to the high dimensionality of the optimization problem in comparison with the size of the training set, we are unsurprised by the poor reliability of the SV regression method. To avoid over-fitting and to increase the efficiency of the algorithm we reduced the number of variables by applying factor analysis. The analysis of the reduced data showed that the best results are achieved when a model with a scalar kernel is applied. The most reliable prediction is provided by using up to three factors in the regression estimation.

The prediction of ages of embryos not included in the training set, i.e., to those for which we do not possess any information about their precise developmental age, show good correlation with the predefined temporal classes. However the prediction results for embryos belonging

[†] We should keep in mind that as the value of ε in the ε-insensitive loss function in (1) is taken equal to 0.05, it makes no sense to consider errors less than 0.05 and so all the table entries are taken into account only up to the first decimal place

Table 2. Results of SV regression estimation and prediction accuracy computed on the training set using the data reduced by factor analysis. The entries of the upper part of the table are the values of the criterion (6) and of the lower part of (7). In brackets the maximal differences between the observed and predicted values of embryo age are given. The columns correspond to the number of factors considered. The parameters were taken with $C = 1$ for both models and $a = 2$ for the polynomial model

SV regression

factors	1	2	3	4	5	6	7	8	9	10	11	12	13
Scalar	1.93	1.93	1.91	1.91	1.87	1.82	1.82	1.82	1.78	1.79	1.78	1.76	1.69
	[7.7]	[7.6]	[7.3]	[7.3]	[7.1]	[6.9]	[6.8]	[6.9]	[6.6]	[6.8]	[6.5]	[6.3]	[7.8]
Poly-	1.95	1.85	1.79	1.81	1.62	1.53	1.24	1.13	0.94	0.97	0.74	0.59	0.59
nomial	[7.7]	[7.5]	[6.9]	[6.9]	[7.0]	[6.9]	[6.1]	[6.4]	[6.2]	[7.1]	[7.4]	[11.0]	[8.7]

Prediction

factors	1	2	3	4	5	6	7	8	9	10	11	12	13
Scalar	1.96	1.99	2.06	2.10	2.12	1.99	2.13	2.19	2.15	2.31	2.39	2.35	2.07
	[7.7]	[7.6]	[7.6]	[7.3]	[7.2]	[7.1]	[7.0]	[7.1]	[6.7]	[8.0]	[8.6]	[7.3]	[8.2]
Poly-	1.98	2.04	1.97	2.12	2.21	2.63	1.88	2.37	3.05	2.99	3.85	3.91	5.12
nomial	[7.6]	[7.6]	[7.0]	[7.0]	[7.3]	[8.4]	[7.1]	[8.7]	[11.0]	[10.7]	[28.8]	[22.8]	[21.8]

to late temporal classes are less accurate than for early ones. We can suggest two categories of explanations for this apparent discrepancy, concerning respectively the quality of experimental data and that of pattern recognition algorithms. Because we are incrementally improving both, these issues are convolved.

With respect to experimental data, we note that embryos were attributed to temporal classes on the basis of visual inspection of the *eve* gene expression pattern. It is evident that such classification is not exact and thus can not serve as an absolute criterion of the quality of age prediction. A fundamental difficulty is that boundaries between classes always involve the separation of almost identical embryos. Moreover, when standardizing against membrane invagination we are currently limited to published experiments (Merril *et al.*, 1988), which, while carefully performed, were not originally intended for the use they are put to here. We are undertaking the acquisition of our own *in vivo* membrane observations, which we anticipate will provide a more precise standard of age. One question of fundamental biological significance is the inherent temporal variability of the system, a temporal resolution beyond which expression patterns do not provide a suitable clock. This work provides a preliminary upper limit of biological temporal variability of about 2 minutes.

With respect to pattern recognition considerations, we use a feature vector which is formed by spline approximation of the 1D expression pattern of the *eve* gene and is represented by values of extrema of gene expression. Other essential features of a pattern such as amplitude and number of peaks, as well as value of the first derivative at the points where a gradient of fluorescence intensity between adjacent nuclei reaches maximum were not taken into account. These additional features are especially essential to characterize gene expression patterns of embryos belonging to temporal classes 7 and 8, as representatives of these classes have very similar values of extrema of *eve* gene expression. Thus to achieve more reliable temporal characterization of our dataset we need to increase a dimension of feature vector by including the features mentioned above and/or extracting features from a two-dimensional image. The exploration of these methods on improved experimental data will be the subject of future work. As this work is completed, we plan to make our method publicly available to the scientific community by providing on-line access to the application, which will predict the developmental age of the embryo of interest from its segmentation gene expression pattern.

ACKNOWLEDGEMENTS

This work is supported by NIH grants 4 RO3 TW01147-01 and RO1 RR07801-11, NATO Collaborative Linkage Grant LST 975850 and GAP award RBO-1238. We thank David Kosman and Svetlana Surkova for key contributions to this work.

REFERENCES

Akam,M. (1987) The molecular basis for metameric pattern in the *Drosophila* embryo. *Development*, **101**, 1–22.

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.

Ingham,P.W. (1988) The molecular genetics of embryonic pattern formation in *Drosophila*. *Nature*, **335**, 25–34.

Jurgens,G., Wieschaus,E., Nusslein-Volhard,C. and Kluding,H. (1984) Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. II. Zygotic loci on the third chromosome. *Roux's Arch. Dev. Biol.*, **193**, 283–295.

Kosman,D. (1998a.) Method for acquisition of quantative gene expression data *in situ*. http://www.csa.ru/flyex/proc_steps/dave.html.

Kosman,D., Reinitz,J. and Sharp,D.H. (1997) Automated assay of gene expression at cellular resolution. In Altman,R., Dunker,K., Hunter,L. and Klein,T. (eds), *Proceedings of the 1998 Pacific Symposium on Biocomputing*. World Scientific Press, Singapore, pp. 6–17. http://www.smi.stanford.edu/projects/helix/psb98/kosman.pdf.

Kosman,D., Small,S. and Reinitz,J. (1998b) Rapid preparation of a panel of polyclonal antibodies to *Drosophila* segmentation proteins. *Dev. Genes Evol.*, **208**, 290–294.

Merrill,P.T., Sweeton,D. and Wieschaus,E. (1988) Requirements for autosomal gene activity during precellular stages of *Drosophila melanogaster*. *Development*, **104**, 495–509.

Myasnikova,E., Samsonova,A., Samsonova,M. and Reinitz,J. (2000) Spatial registration of *in situ* gene expression data. *Computation in Cells: Proceedings of an EPSRC Emerging Computing Paradigms Workshop*, University of Hertfordshire, UK, 2000. Department of Computer Science, Technical report No: 345, pp. 21–26.

Myasnikova,E., Kosman,D., Reinitz,J. and Samsonova,M. (1999) Spatio-temporal registration of the expression patterns of *Drosophila* segmentation genes. In Lengauer,T., Schneider,R., Bork,P., Brutlag,D., Glasgow,J., Mewes,H.-W and Zimmer,R. (eds), *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.

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.

Nusslein-Volhard,C. and Wieschaus,E. (1980) Mutations affecting segment number and polarity in *Drosophila*. *Nature*, **287**, 795–801.

Platt,J.C. (1998) Sequential minimal optimization: A fast algorithm for training support vector machines. Report MSR-TR-98-14, Microsoft Research Tech., Microsoft, Redmond, USA.

Reinitz,J., Kosman,D., Carlos,E., Vanario-Alonso and Sharp,D. (1998) Stripe forming architecture of the gap gene system. *Dev. Genet.*, **23**, 11–27.

Reinitz,J., Mjolsness,E. and Sharp,D. (1995) Cooperative control of positional information in *Drosophila* by bicoid and maternal hunchback. *J. Exp. Zool.*, **271**, 47–56.

Reinitz,J. and Sharp,D. (1995) Mechanism of formation of *eve* stripes. *Mech. Dev.*, **40**, 133–158.

Schölkopf,B., Smola,A. and Müller,K.-R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

Sharp,D. and Reinitz,J. (1998) Prediction of mutant expression patterns using gene circuits. *BioSystems*, **47**, 79–90.

Smola,A. and Scholkopf,B. (1998) A tutorial on support vector regression, NeuroCOLT2 Technical Report Series, NC2-TR-1998-030, http://www.neurocolt.com, 1998.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

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.