Multiscale scatterplot matrix for visual and interactive exploration of metabonomics data

Fabien Jourdan, Alain Paris, Pierre-Yves Koenig, Guy Melançon

Multiscale scatterplot matrix for visual and interactive exploration of metabonomics data.

Fabien Jourdan[1], Alain Paris[1], Pierre-Yves Koenig[2], and Guy Melançon[2]

[1] UMR1089 Xénobiotiques INRA-ENVT, Institut National de Recherche
Agronomique, France
`Fabien.Jourdan@toulouse.inra.fr`
[2] Laboratoire d'Informatique, de Robotique et de Micro-électronique de Montpellier,
LIRMM UMR CNRS 5506, France
`{Pierre-Yves.Koenig, Guy.Melancon}@lirmm.fr`

Abstract. We describe a method turning scatterplot matrix visualizations into malleable graphical objects facilitating interaction and selection of pixelized data elements. The method relies on density estimation techniques [1, 2] applied through standard image processing. A 2D scatterplot is considered as an image and is then transformed into nested regions that can be easily selected. Based on Wattenberg and Fisher [3], and as confirmed by our experience, we believe users have a good intuition interpreting and interacting with these multiscale graphical objects. Bio-molecular data serves here as a case study for our methodology. The method was discussed and designed in collaboration with experts in metabonomics and has proven to be useful and complementary to classical statistical methods.

1 Introduction

Information visualization is a useful approach for the analysis and exploration of large multidimensional data sets. One major contribution of visualization when compared with other non visual approaches is to offer direct interaction with the data providing immediate feedback. Visualization can thus feed the analysis process and guide the user in the exploration or refinement of scientific hypotheses. This is of utmost importance in the study of biochemical data mostly because of the vast amount of experimental data that need to be sorted, but also because of its inherent complexity. The present paper focuses on a visualization technique helping the analysis of experimental data measuring the impact of specific molecules on metabolism. Visualization has been introduced in the processing chain involving biologists, statisticians and chemists and aims at bringing new insights on the studied phenomenon as well as on the experimental methodology itself.

In this setting, the raw data we need to study has a multidimensional character as it gathers numerous attributes describing both the phenomenon under study and the experiment itself. As we shall see, our methodology enters the pixel-oriented paradigm [4], as we suggest to lay the experimental data on a

two-dimensional layout while offering the user the possibility to interact with this 2D representation. We form images where pixels correspond to elementary data elements positioned according to numerical attributes. The real benefit of the interaction is achieved by allowing users to dynamically inquire about the underlying information. The 2D view is thus transformed into a more malleable material while preserving its statistical properties. The selection of subsets of experimental measurements is thus made easier while remaining faithful of the statistical phenomenon under study.

This work partly relies on techniques that were originally described and used for the exploration of relational data [5]. The similarity with our previous work is that we allow the user not only to explore and manipulate the processed scatterplot data, but also to directly act on this representation to inquire about the underlying data. In the present case, however, the experiments require that we build a representation gathering a series of scatterplots enabling the user to cross-examine different experimental conditions on the one hand, and observe the effect of drugs as time evolves on the other hand. Hence, the user actually views and interacts on a series of scatterplot matrix (see Fig. 7 for instance) all residing in a 3D data cube. Our approach is inspired from the seminal work by Becker and Cleveland [6] and shares similarities with that of Martin and Ward [7]. The idea of interacting with scatterplots through multiscale images was triggered by the work of Wattenberg and Fisher [3].

As we will explain, our visualization came as a mean of exploring and interacting with experimental data. The idea of combining simple image processing techniques together with scatterplot representations emerged from discussions with users where the need to easily select and highlight elements in numerous scatterplots became clear. The visual exploration and ability to interactively select elements from the scatterplots proved useful and complementary to classical statistical methods such as principal component analysis (PCA). Indeed, some phenomenons could only be observed by visual inspection and were left unnoticed by PCA and factorial analysis. It seems that the confidence assessed by classical statistical methods is offered at the price of leaving aside details that are nevertheless of interest when visually exploring the data.

2 Metabonomics

Metabonomics is usually defined as the study of changes in metabolite profiles as a result of a biological perturbation (such as disease or physiological stress)[3]. This is precisely what the experiment we will concerned with is about.

---

[3] Compare with *metabolomics* which is the systematic study of the chemical fingerprints that specific cellular processes leave behind. That is metabolomics focuses on the metabolites found in a biological organism, which are the end products of its gene expression.

2.1 The experimental procedure

The experiment concerns mouse populations which are either normally fed, or go through a change from their usual diet. Another factor is also studied, where mice are administered a drug or a toxic chemical (a molecule or group of molecules), in addition to their diet. The challenge is to understand how their organism reacts to these perturbations (diet change and/or drug taking). The whole experiment follows a patented protocol [8] we shall briefly describe here.

The effect of the diet, with or without drug taking, is observed through samplings (blood or urine samplings). Mice are sampled a few times (once a month for two months and then more regularly, e.g.). Each sample is carefully processed and follows a series of steps to measure the presence of certain molecules (concentration) in reaction to the physiological stress the mice experience.
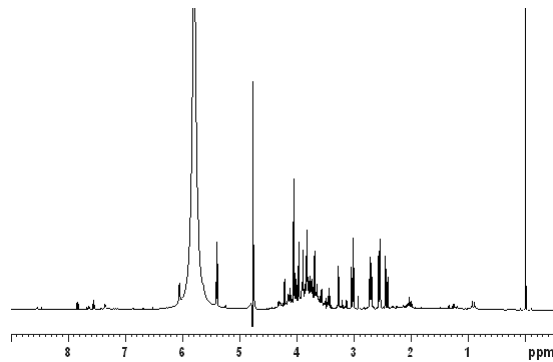

Fig. 1. The spectrum diagram describes the concentration (peaks) at which yet unidentified molecules appear in mice's biofluids.

The samples are then analyzed to output a spectrum diagram measuring how much given molecules are present in the animal's organism (see Fig. 1) — one for each set of experimental conditions. At this step however, indirect measures only reveal the concentration at which some molecules are present in the organism. Extra work is needed to actually identify which molecules are present. Both steps have to be accomplished separately because concentrations are revealed through indirect measurements involving ion rays. Roughly speaking, the horizontal axis of the spectrum diagram (Fig. 1) is divided into sub-intervals according to the peak (concentration) structure. The largest peak on the left part of the diagram might for instance give rise to a series of sub-intervals. The ppm value associated with each sub-interval provides information that will help the identification of the associated molecule in a subsequent step. Sub-intervals are then considered as statistical variables that undergo a series of tests.

After collecting this numerical data, and after statistical analysis, the whole protocol leads back to the biological question, which is to identify pathways in-

volved in the organism's reaction to physiological stress. Indeed, the presence and concentration of molecules in the samples relate to specific metabolic reactions that need to be discovered and studied. As the concentration of molecules vary from one situation to another, the biologist might infer hypotheses about how the organisms fights this stress (drug and/or diet change).

2.2 The data and task

The variables (related to molecule concentration) extracted from the spectrum diagram in Fig. 1 then undergo statistical tests (PCA) capable of identifying (with known statistical errors) groups of experimental conditions showing clearer contrast in terms of diet, drug taking and time elapsed after the experiment was initiated. A factorial analysis then allows the identification of the most significant variables with respect to the dominant PCA axis. In other words, the analysis is able to detect which molecules goes through the most notable variations under one or more experimental conditions.

Hence, the protocol produces a vast amount of figures. All mice populations (normal diet or unusual diet, with or without drug taking) are sampled six times, and approximately 750 molecules are traced in each of these situations. The complexity does not however come from this moderate data volume but from the need to simultaneously compare numerous pairs of situations. Indeed, because we are interested in studying molecules showing profiles that differ in distinct situations, we build scatterplots gathering profile information collected from two situations. Highly concentrated molecules in a both situations are obviously of interest. The question however is to see whether this high profile of expression is observed in all situations, for instance, before enquiring about the pathways involved in the production of that molecule. Another scenario of interest is one where two molecules might systematically show opposite profiles (high concentration for one and low concentration for the other, or vice-versa, in all situations).

3 The visualization

3.1 2D scatterplots

The visualization we designed gathers 2D scatterplots mixing two sets of experimental conditions. More precisely, each axis spreads over the interval [0,1] where $p \in [0, 1]$ corresponds to a normalized concentration (of a molecule in the organism). Hence, given a set of conditions S_1 (diet, drug taking and time elapsed) we assign each molecule m the concentration $p_{S_1}(m)$ at which it was observed. Now, given another set of conditions S_2 each molecule is mapped onto the 2D point $(p_{S_1}(m), p_{S_2}(m))$. Fig. 2 shows a typical 2D scatterplot. In this example, the two situations correspond to the same diet and drug taking conditions, but to different time intervals[4]. The x-axis correspond to molecule concentration after

---

[4] From now on, we shall refer to the condition "time elapsed after the experiment was initiated" by "time intervals" or "time elapsed".
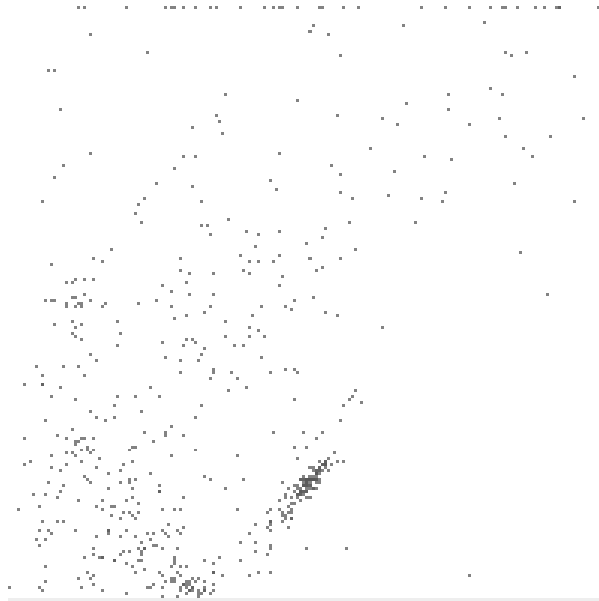
Fig. 2. Scatterplots are formed by crossing two sets of measurements (molecule concentration). In this case, the measurements concern the same mouse population, but were taken at two distinct moment in time. The x-axis corresponds to molecule concentration after one month, and the y-axis corresponds to molecule concentration after a bit more than three month.

a month, while the y-axis correspond to molecule concentration after a bit more than three months. The position assigned to a molecule reflects its concentration in the two different experimental (set of) conditions. Two molecules can be mapped to the same 2D point. We take this into account by assigning points on a grayscale in order to reflect frequencies. This is of importance for what follows.

It is in theory possible to build a scatterplot for all pairs of situations (set of conditions), leading to a set of (# diet types × # drug taking patterns × # of samplings) scatterplots like the one shown in Fig.2. However, we shall focus here on the evolution of molecule concentration over time. That is, we look at molecule concentrations for a given diet type and drug taking pattern (drug or no drug) after 32 days, 60 days, 88 days, and so on. We then form a (upper triangular) matrix where all scatterplots on row one map concentrations after day 32 on the horizontal axis, scatterplots on the second row map concentrations after day 60 on the horizontal axis, and so on. Conversely, scatterplots in column one map concentrations after day 32 on the vertical axis, etc. (See section 4.1.) We can form matrices for all possible pairs of diet type and drug taking patterns.

Since axes are mapped to time, one could be tempted to look at the data using time series or parallel coordinates. However, time series or parallel coordi-

nates did not offer a good readability of our data (which would typically require to be clustered) – see Fig. 3. Moreover, a time series visualization allows to easily read how the concentration of given molecules evolve throughout the whole experience, from day 1 to day 115 (mice are sampled 5 times). The interest of biologists however resides elsewhere. Their aim is not to model the variation of concentration of molecules, but rather to see whether the diet and drug pattern induce a clear change in the metabolism between two different samplings and, of course, understand why. Typically, these changes should appear as different spreading patterns of molecules on the scatterplots.
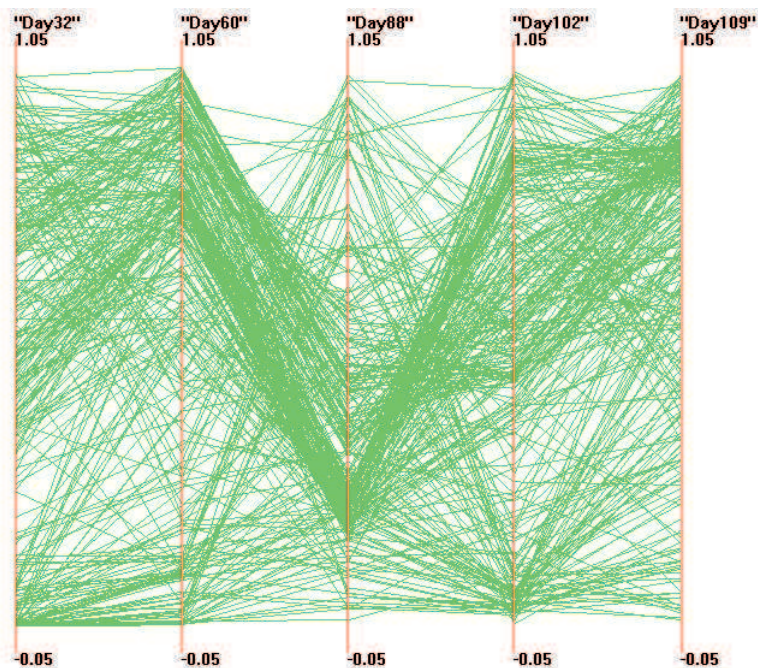

Fig. 3. Parallel coordinates (obtained using XMDVTool [9]) can be used to see molecule concentrations evolve along time.

3.2 Gaussian kernel and blur

The fine-grained structure of the scatterplots makes them difficult to explore and manipulate. We borrow an idea from our previous work [5] initially inspired from Wattenberg and Fisher [3]. We consider the scatterplot as an image and blur it following a standard image processing technique. Each pixel, seen as a numerical value, is replaced by the result of a matrix convolution applied to

its neighborhood. As the matrix encodes a gaussian kernel the pixel's neighbors contribute differently to the resulting value. This standard technique in image processing [10] actually performs an estimation of the density function encoded by the scatterplot [1, 2]. The result of a gaussian convolution on the scatterplot of Fig. 2 is illustrated in Fig. 4. Parameters of the gaussian kernel can be set to vary the blurred image.
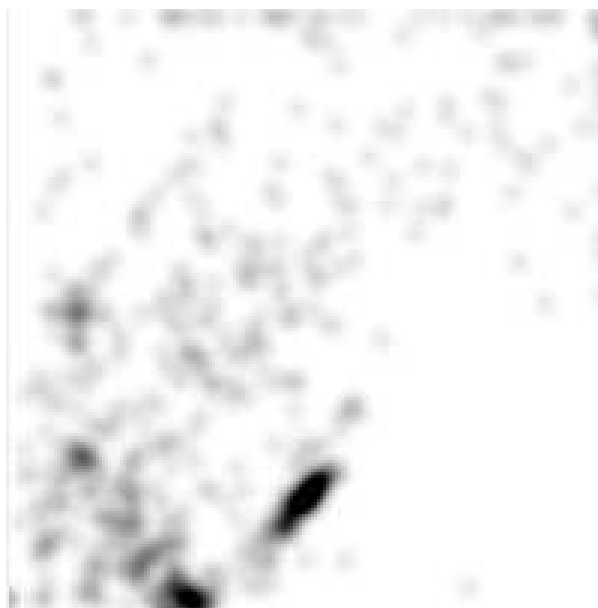

Fig. 4. The scatterplot in Fig. 2 is considered as an image and is blurred by applying a gaussian convolution. The blur is obtained from the scatterlot showed in Fig. 2.

3.3 Identifying pre-selected areas

The last step is to compute areas of relatively homogeneous grayscale in the blurred image. Molecules mapped to neighbor pixels in the scatterplot are merged into relatively homogeneous gray regions. This will help us identify and *select* regions gathering molecules behaving similarly. The regions are computed by performing image segmentation on the blurred scatterplot. This again is a standard image processing technique computing regions of neighbor pixels of similar grayscale value. The segmentation outputs regions delimited by closed curves, in a way similar to the level curves on a geographical map — we capture molecules having similar profiles, where similarity decreases as regions grow larger. The number of levels can be controlled in order to output larger or thinner regions. The right choice partially depends on the morphology of the original scatterplot.
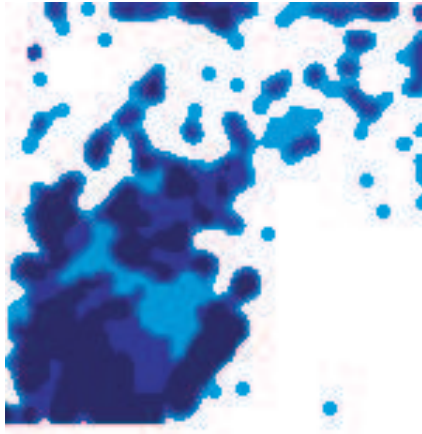
Fig. 5. Regions are computed from the blurred scatterplot by applying image segmentation techniques. The segmented image showed here is obtained from the blurred scatterplot in Fig. 4.

Fig. 5 shows the resulting multilevel image after segmentation is computed on the blurred image (b) (all entries of the matrix in Fig. 8 also illustrate the result of image segmentation on various scatterplots). The final segmented image indeed reveals a multiscale structure as regions are all nested according to the different grayscales in the blurred image.
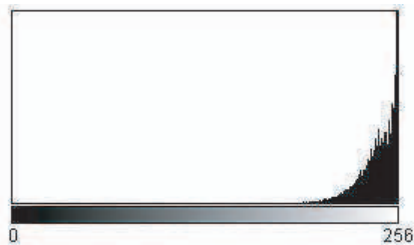

Fig. 6. Histogram describing how grayscale distributes over the pixels of the scatterplot in Fig. 2. The darker pixel has value 166. The histogram clearly indicates that a majority of pixels are pale gray.

The segmentation is performed based on a partition of the grayscale range following ideas we shall now describe. First note that the grayscale values vary over the interval $[0, 256[$ (black pixels have a value of 0, and white pixels have a maximum value of 255). Fig. 6 shows how the grayscale distributes in the blurred image (Fig. 4). Observe that the gray scale does not distribute uniformely over

the full range $[0, 256[$, suggesting that a division into interval of equal length should be avoided.

Assume we want to segment the image into regions on k different levels. We want to find sub-intervals $[a_1, a_2[$, $[a_2, a_3[$, ..., $[a_{k-1}, a_k[$ (with $a_1 = 0$ and $a_k = 256$), such that the proportion of points lying in $[a_i, a_{i+1}[$ is $1/k$. This can be easily done using the inverse image of the density function (integral of the distribution in Fig. 6). When $k = 4$ this amount to finding what is often called the *quartile* of the grayscale distribution. This idea is commonplace in visualization and has been applied in other situations such as when defining a colormap over a set of data elements [11].
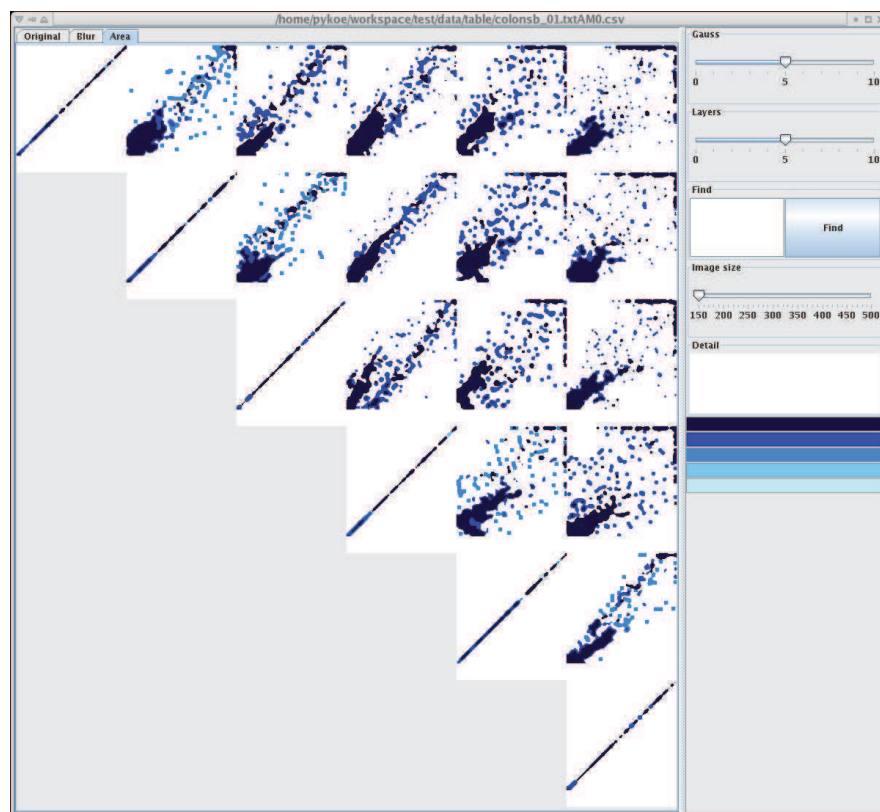

Fig. 7. Application snapshot showing the various interaction that can be performed. The matrix shows (processed) 2D scatterplots of pairs of situations. Each line/column correspond to a different time interval elapsed after initiating the experiment — time intervals increase from left to right.

4 Case study

4.1 Interaction

Fig. 7 shows a snapshot of our application. When starting the exploration, the user will set various experimental conditions defining the scatterplot matrix to be visualized. As mentioned earlier, we assume the user decides to let time intervals vary while keeping the other experimental conditions fixed (same diet/drug taking schema for all scatterplots). The visualization we built finds its full utility when the user is able to act on it and select various regions. To this end, the user may:

– vary the parameters of the gaussian kernel;
– vary the number of layers in the segmented image;
– click on a region and have access to the underlying data;
– activate regions containing any given molecule by performing a search.
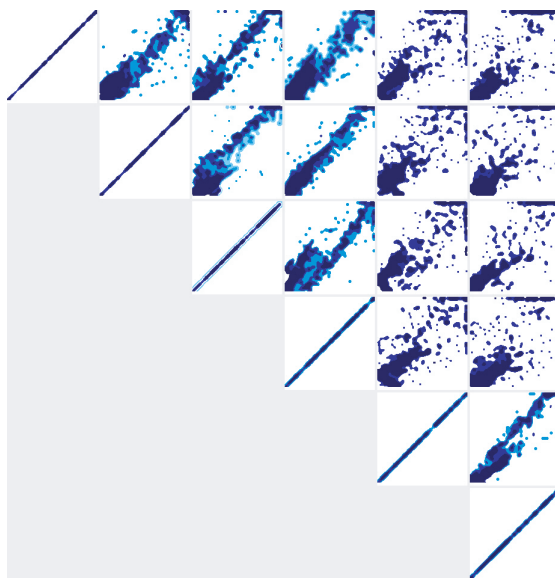

Fig. 8. Scatterplot matrix of mice experiencing a diet change. Observe that the data in the last two column spread further away from the diagonal.

Fig. 8 shows the scatterplot matrix gathering data collected from a population of mice that have *not* been administered *any drug*, but have only experienced a change from their usual diet. One thing is immediately observed when looking at the scatterplot images: those lying on the last two columns spread much further away from the diagonal than do the first columns. This observation agrees

with the PCA analysis (see Fig. 9); using only PCA however requires additional work in order to determine the reasons underlying this cut. Following this observation, biologists were able to trace back methodological errors in mice diets. We see here one advantage when using the scatterplot matrix. Indeed, any statistical analysis of the data will be obscured by the diet event during the last two samplings. Using the scatterplot matrix however, users can simply decide to ignore the data from the last two columns.
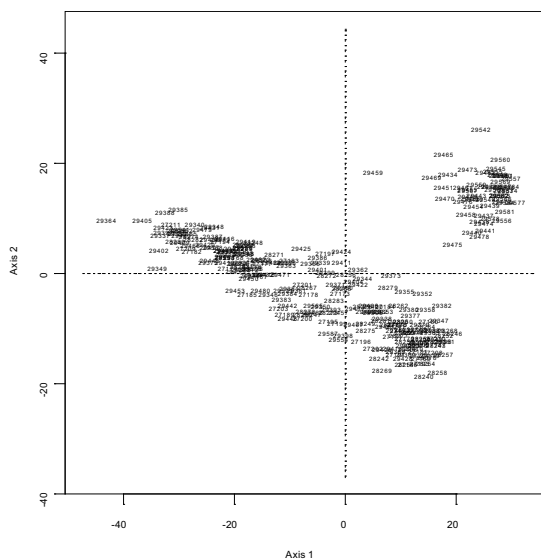

Fig. 9. PCA analysis clearly separates the variables into two groups indicating that mice diet experienced an unusual change.

Fig. 10 provides another example. In this case however, mice have all been administered a drug but have *not* experienced a change from their usual diet.

Notice the red patches that have been activated through all scatterplots. This activation has actually been triggered by the selection of a small red patch in the scatterplot sitting at position (4,5) on the fourth row and fifth column — the selected patch is pointed at by the arrow. This small red patch corresponds to molecules that are found at rather high concentration after the fourth time interval after the experiment has been initiated (high x-coordinate) but that show low concentration after the next time interval has elapsed (low y-coordinate). A search through all regions of all other scatterplots allow to automatically activate regions containing any molecule present in the selected patch. In a sense, the automatic activation of patches actually performs brushing [6, 7]. The use of image blurring and segmentation however provides brushes specific to the dataset under study.
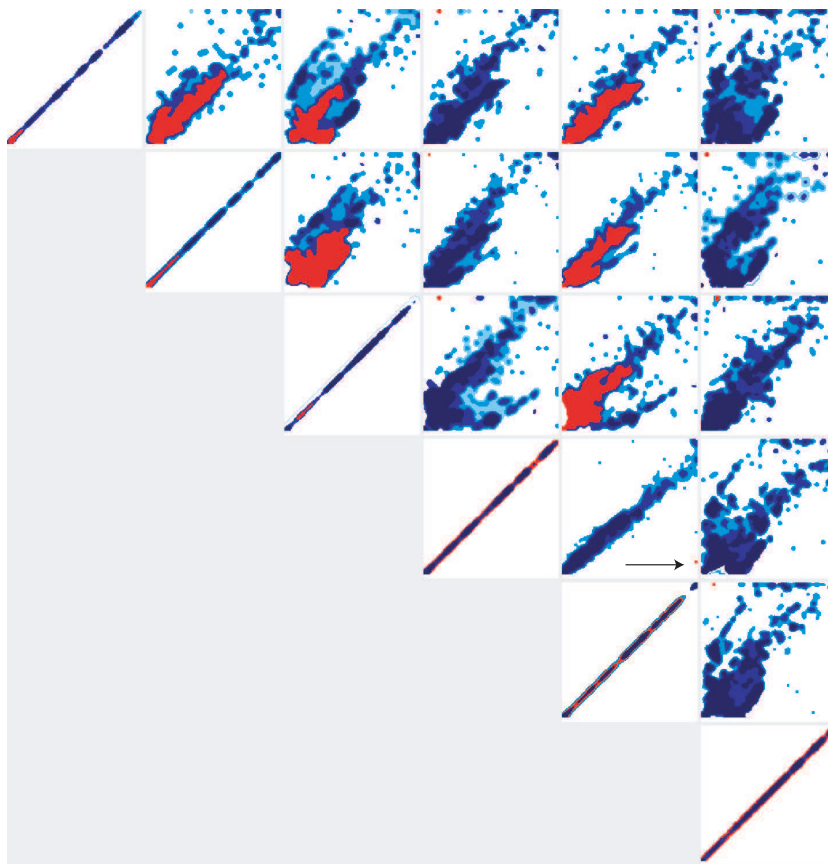
Fig. 10. A small red patch indicated by a pale blue oval has been selected in the scatterplot sitting on the first row and second column. The other red regions have been activated as they contain molecules underlying the selected patch.

The visual inspection of the automatically activated regions is informative and can potentially bring knowledge about how the molecules relate to one another, or how they relate to metabolism. The red patch we selected (the one pointed at by the arrow) contain molecules showing a drop in concentration from a time interval to the next. The activated regions captures other types of variations. Regions close to the diagonal indicate similar concentration at both time intervals. Regions located in the left upper part show profiles dual to the one we selected and correspond to an increase in concentration. These variations might indicate a change of phase regulated by molecule concentration.

In other case, the initially selected patch divide into two distinct subset of molecules showing distinct behaviors in some conditions. This observation can not be inferred directly from statistical analysis, and on the contrary, is rather

immediate when visually exploring the data. Note also that these observations seem more or less obvious because of their enhanced readability (as a result of the blur and segmentation processes).

4.2 Combining classical statistics with visual interaction

We shall provide evidence showing that visual exploration of the scatterplots comes as a complementary approach to classical statistical analysis.

PCA is used first to detect whether certain experimental conditions have a stronger and more determinant effect. Factorial analysis then splits variables according to these experimental conditions.
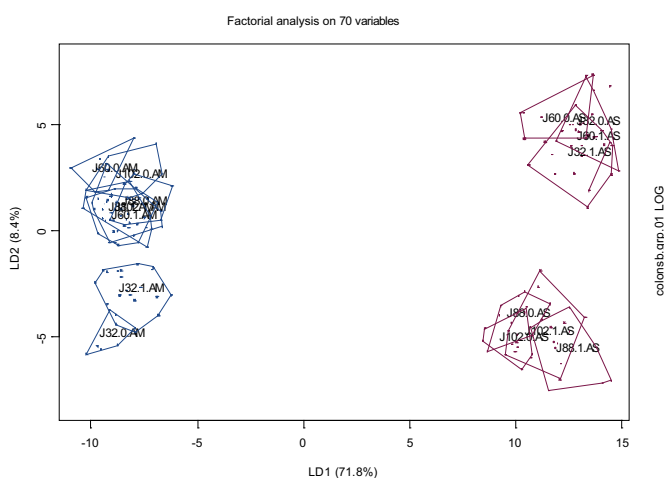

Fig. 11. Factorial analysis splits variables into 4 groups.

PCA underlined the fact that the most predominant factors explaining differences between molecule profiles were the last two samplings. The factorial analysis identified variable number 347 as being clearly involved in the metabolic changes. Using our tool we performed a search to locate variable 347 through all pairs of situations. As an output of the search, we also obtained the list of all variables lying in a segmented patch with variable 347. We find here a second advantage of using our multiscale scatterplot matrix together with calssical analysis. Indeed, because factorial analysis can only be computed on a limited subsets of variables (this is a consequence of the factorial analysis which requires that between 30% and 50% of all variables be used), our tool can indeed be used to discover variables having a profile similar to that of variable 347, which were not part of the factorial analysis. This is a typical example of a successful combination of classical statistics together with visual interaction.

5   Conclusion and future work

We have described a method turning scatterplot matrix visualizations into malleable graphical objects facilitating interaction and selection of pixelized data elements. The method relies on density estimation techniques [1, 2] applied through standard image processing. A 2D scatterplot is considered as an image and is then transformed into multiscale graphical objects where nested regions can be easily selected. Based on Wattenberg and Fisher [3], and as confirmed by our experience, we believe users have a good intuition interpreting and interacting with these multiscale graphical objects. It may well be that most people now have experience using geographical level maps.

This method has been introduced within an experimental protocol after experts have recognized its usefulness and complementarity to classical statistical methods. Our research did not focus on the design of new visualization techniques. Indeed, discussions with experts made it clear that we only had to assemble classical image processing techniques with scatterplot matrices together with basic interaction. An ongoing collaboration with all experts involved in the whole experiment (biologists, statisticians and chemists) will help the integration of this visual exploration method together with their existing practice.

Access to the underlying data is not yet completely satisfactory, mainly because of the inherent complexity of the experimental protocol. Also, more user feedback is needed in order to decide of various parameters that could be preset (radius of gaussian kernels, thickness of regions, etc.). Extension of our tool will include other types of kernels to let the final processed scatterplot vary. Again, user feedback should help measure the impact of these design choices on the usefulness and usability of the method.

We also plan to develop a view gathering all experimental conditions into a "datacube". The set of all scatterplots matrices obtained by letting all conditions vary indeed form a 3D or even 4D cube. The user might find useful playing with the cube in order to select particular experimental conditions when defining the initial scatterplot matrix.

References

1. Silverman, B.: Density Estimation for Statistics and Data Analysis. Chapman & Hall (1986)
2. Scott, D.W.: Multivariate Density Estimation : Theory, Practice, and Visualization. Wiley Series in Probability and Statistics. Wiley-Interscience (1992)
3. Wattenberg, M., Fisher, D.: A model of multi-scale perceptual organization in information graphics. In North, S.C., Munzner, T., eds.: IEEE Symposium on Information Visualization, IEEE Computer Society (2003)
4. Keim, D.A.: Designing pixel-oriented visualization techniques: theory and applications. IEEE Transactions on Visualization and Computer Graphics **6**(1) (2000) 59–78
5. Chiricota, Y., Jourdan, F., Melançon, G.: Metric-based network exploration and multiscale scatterplot. In Ward, M., Munzner, T., eds.: IEEE International Symposium on Information Visualization, IEEE Computer Society (2004) 135–142

6. Becker, R.A., Cleveland, W.S.: Brushing scatterplots. Technometrics **29**(2) (1987) 127–142
7. Martin, A.R., Ward, M.O.: High dimensional brushing for interactive exploration of multivariate data. In: IEEE Conference on Visualization '95, IEEE Computer Society (1995) 271–278
8. Dumas, M., Canlet, C., Debrauwer, L., Martin, P., Paris, A.: Selection of biomarkers by a multivariate statistical processing of composite metabonomic data sets using multiple factor analysis. Journal of Proteome Research **4**(5) (2005) 1485–92
9. Ward, M.O., Rundensteiner, E.A., Yang, J., Doshi, P.R., Rosario, G.: Interactive poster: Xmdvtool: Interactive visual data exploration system for high-dimensional data sets. In: IEEE Symposium on Information Visualization. (2002) 52–53
10. Russ, J.C.: The Image Processing Handbook. 3rd edn. CRC Press (1998)
11. Herman, I., Marshall, M.S., Melançon, G.: Density functions for visual attributes and effective partitioning in graph visualization. In Roth, S.F., Keim, D.A., eds.: IEEE Symposium on Information Visualization, IEEE Computer Society (2000) 49–56