SpRay: A Visual Analytics Approach for Gene Expression Data

Janko Dietzsch[*] ZBIT, University of Tübingen
Julian Heinrich[†] VISUS, University of Stuttgart
Kay Nieselt[‡] ZBIT, University of Tübingen
Dirk Bartz[§] ICCAS/VCM, Universität Leipzig

Abstract

We present a new application, SpRay, designed for the visual exploration of gene expression data. It is based on an extension and adaption of parallel coordinates to support the visual exploration of large and high-dimensional datasets. In particular, we investigate the visual analysis of gene expression data as generated by microarray experiments; We combine refined visual exploration with statistical methods to a visual analytics approach that proved to be particularly successful in this application domain. We will demonstrate the usefulness on several multidimensional gene expression datasets from different bioinformatics applications.

Keywords: Visual analytics, bioinformatics, gene expression experiments, microarray data, large-scale microarray

Index Terms: I.3.3 [Computer Graphics]: Line and Curve Generation, Display Algorithms I.3.6 [Computer Graphics]: Interaction Techniques J.3 [Life and Medical Sciences]: Biology and Genetics

1 Introduction

The investigation of large high-dimensional datasets generated by recently developed high-throughput methods is a very common task in bioinformatics. This is largely due to the use of these methods in a wide variety of applications in biology and medicine. The need for useful methods for such investigations will become even more important as they become a more and more common part of the daily work in the bioscience and bio-engineering laboratories and hospitals. For instance, microarray-based gene expression studies generate data for several thousands of genes (data samples) under numerous different conditions (dimensionality of the data). The data itself is stored in the *gene expression matrix* as the fundamental structure, which we use as the basis for our visual analysis. This matrix contains the expression values of one gene under the different conditions in its rows and the gene expression values of a certain condition in its columns. Conditions imply a large variety of different meanings, which can be external or internal stress factors (e.g., heat or chemical irritation) under which the cell is growing, pathological states of the cell, mutated cells, or time points of time series.

Note that the terminology of gene expression matrices is different from the standard terminology in the context of data visualization. The term samples – in the context of bioinformatics used to depict different conditions – is mapped to the different dimensions. In contrast, the individual genes are mapped to the data values (or data samples in the visualization terminology). For this reason, we try to avoid the term data sample when we depict the individual data points and call the gene expression values data values.

There is a strong need for adequate methods to reveal relevant effects that are latently contained in the data and to separate these from the noise attributed to the measuring procedure. Several statistical methods already exist that attempt to achieve this goal [1]. Nevertheless, the analysis of a microarray-based gene expression experiment is still a very challenging task. Often the application of only one method is not successful and it is necessary to employ a number of different methods [1]. This situation leads directly to the design of comprehensive, flexible, and extendable software systems like SpRay to analyze microarray data. Nevertheless, a consensus of the different analysis methods must be found to get reliable results. To address this issue and to profile the used statistical analysis methods, our novel contribution is the conjoined visual exploration of the original data together with the associated deduced statistical data in a common data space (see Fig. 1). This combination of automatic (statistical) and visual analysis leads to a visual analytics approach that provides more insights in the structure of the data and that prevents misleading impressions as much as possible at the same time. In this paper, we introduce such a visual analytics approach for the analysis of high-dimensional microarray data.

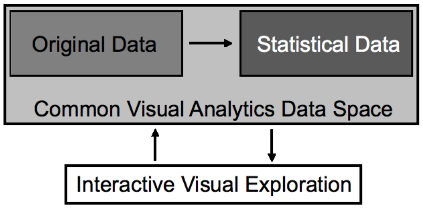
Original Data
Statistical Data
Common Visual Analytics Data Space
Interactive Visual Exploration

Figure 1: The Visual Analytics (VA) approach of SpRay: statistical data is derived (semi-)automatically from the original data and combined with it into the common visual analytics data space. This common VA data space is then visually explored to analyze the data.

The remainder of this paper is organized as follows. After briefly reviewing related work in Section 2, we introduce the used variation of the parallel coordinates plot to visually explore and interpret the provided data (Section 3). Section 4 presents the results of the visual analysis of the examined datasets, which in turn will be discussed in Section 5. Finally, we present our conclusions and point to future directions of research (Section 6).

2 Related Work

The importance of designing appropriate visualization methods for bioinformatics was already discussed in [21]. Furthermore, Gilbert et al. used in an early approach [8] heatmaps, dendrograms (for cluster hierarchies), and VRML models to represent the data, in particular the cyclicity of Spellman's yeast cell cycle (see Section 4). Since then, several papers have addressed the visualization for this application domain. Saraiya et al. [23] studied the applicability of five different visualization tools for microarray data. These tools focus on specific visualization techniques such as heatmaps (Clusterview), parallel coordinates (TimeSearcher), and a combination of several techniques such as scatterplots, histograms, parallel coordinates, and heatmaps (Hierarchical Clustering Explorer, Spotfire, GeneSpring). The dataset size of the explored datasets ranged from 170 genes to 1060 genes (data points), and the data dimensionality from three dimensions to 90 dimensions. Their basic conclusion was that tools designed with a specific context in mind do

[*] e-mail: dietzsch@informatik.uni-tuebingen.de
[†] e-mail: julian.heinrich@visus.uni-stuttgart.de
[‡] e-mail: nieselt@informatik.uni-tuebingen.de
[§] e-mail: dirk.bartz@medizin.uni-leipzig.de

IEEE Symposium on Visual Analytics Science and Technology
October 12 - 13, Atlantic City, New Jersey, USA
978-1-4244-5283-5/09/$25.00 ©2009 IEEE

not perform very well for other applications. Furthermore, they stressed the importance of the supported interaction techniques to derive knowledge from the data.

In 2003, Swayne et al. described the GGobi system that provides several linked visualization techniques, including parallel coordinates and scatterplots [26]. Flexible colormapping is provided through (automatic) brushing and statistical data can be generated through R, a framework for statistical computing [20]. One of the strengths of GGobi is the use of animations or ”tours” to provide a quick overview of the data. While GGobi works well with smaller datasets (e.g., Half-Marathon and Microarray validation in Table 1), its response time becomes significantly slower for larger ones due to its slow rendering. The actual problem of GGobi, however, is overplotting when too many data points are present. Furthermore, it does not focus on the conjoined analysis of original and statistical data (generated through R).

Peeters et al. [18] presented a system that combines an interactive visualization of DNA sequences with provided annotation information. A more application-specific system is PQuad, which visualizes differential protein expression data from mass spectroscopy using colored horizontal line graphs to indicate the predicted positions of the peptides and proteins along DNA strands [9]. GVis [10] focusses on the scalable visual representation at different hierarchy and abstraction levels. A less abstract, more measurement specific system was presented by Linsen et al. [15], who used colored height fields of m/z-ratios and time.

A framework for the visual integration of additional meta-information of gene expression data was introduced in [7] and demonstrated in an application of the heat colormap. The enhanced heatmap showed the clear advantages of the integration of supplemental data from different sources for the visual exploration of microarray data.

An interesting application of the parallel coordinate plot (PCP) was presented in Rübel et al. [22]. The Berkeley Drosophila Transcription Network Project (*BDTNP*) developed a suite to aid the quantitative, computational analysis of three-dimensional gene expression patterns of early embryo states of Drosophila on a cellular resolution. Similar to our work presented here, the PCPs employed here were used to investigate the expression levels of a couple of genes of every cell. One cell is represented in the PCP by one polyline and the expression levels of the different genes are assigned to the dimensions. Some extensions are described to improve the reception of interesting effects displayed by the data, for instance opacity modulation for visual clustering and a three dimensional extension of the PCP.

Recently, Westenberg et al. presented a visualization system, GeneVis, that focuses on genome expression and regulatory networks dynamics [28]. Similar to our approach they use statstical testing to grasp the reliability of the gene expression data. However, they do not provide the process of an interactive conjoined analysis of original and derived data that is not restricted to a special kind of statistical evidence values. Instead, their visualization focus is on the network and interaction aspects deduced from the data rather than the data itself.

In summary, no system, but SpRay, provides a conjoined analysis of original and statistical/derived data in one common visual analytics data space (see Fig. 1), **which is the major contribution of the work, and** which proved to be very useful. Furthermore, many bioinformatics systems do not sufficiently address the case of overplotting, which reduces its utility for large datasets.

3 Extending Parallel Coordinates

A well-known traditional technique for the visual representation of multidimensional datasets is the parallel coordinate plot (*PCP*) introduced by Inselberg [11]. In the field of gene expression, the PCP is already established as a profile plot of the expression values of genes along the experimental conditions. In this commonly employed kind of plot the PCP remains restricted only to the visualization of the gene expression data itself.

Unfortunately, the conventional PCPs do not scale well with the number of data values. In particular, a large number of data points will cause an overdraw problem, so that common patterns and details are hidden by the clutter of lines. Several approaches are proposed to address this issue, such as sampling [5, 12], curved lines [16, 29], and clustering [6, 13, 17]. Closely related to the methods that we use are transparency/density or saturation modulation based approaches [27, 2, 19, 17]. We use variations of these methods combined with color-coding and linked additional data plots – such as scatterplots, histograms, and data tables, which we will discuss in more details below.

The transparency of the lines can be varied to uncover common traits of data items along the different dimensions of the data. In particular, the transparency can be modulated globally for all polylines of the data values. Alternatively, the transparency can be set specifically for each dimension, taking into account the number of polylines passing through a local area (bucket) of the axis.

The option to color-code the whole polyline of a data item according to the data values of one dimension supports the discovery of relations between the different data dimensions, depending on the overdraw and the noise. In our novel application SpRay, we support a number of different colormaps, including the rainbow (hue) map, the more isometric luminance and saturation maps, and a heat (temperature) map. Since we need additional cues to differentiate the polyline bundles, the perceptually preferable grey-level-map is not usable.

The visual exploration is further assisted by integrated linked simultaneous views like scatterplots (matrices) or table lenses between dimensions selected by the user and histogram plots of the individual dimensions, similar to Doleisch et al.'s SimVis system [4] aiming at flow simulation data. These plots also take advantage of the color- and opacity-coding specified in the parallel coordinate plot.

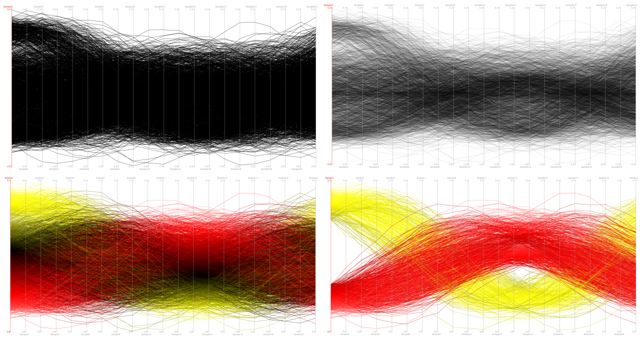

Figure 2: Top/Left: Straight forward parallel coordinate representation of the full Sinusoidal Demo dataset. Top/Right: Transparency weighted representation of the full Sinusoidal Demo dataset. Some structures are already visible through the transparency modulation. Bottom/Left: Application of the heat colormap to the first dimension. Bottom/Right: After culling of the data points of the base-line cluster, the high amplitude clusters become nicely visible.

In contrast to most other applications of parallel coordinates, microarrays can produce invalid data values for some dimensions (conditions/experiments), while generating valid data values for other dimensions. Therefore, our system has to deal with these situations. We address the invalid data values by replacing the invalid data value by imputation, e.g. the average of the valid data values for all experiments of the same gene. All imputed values can be marked in a specific color (e.g., red).

Note that while SpRay is aimed at gene expression data, it is not limited to this kind of data; every multi-dimensional data with or without invalid entries can be visualized with it. In [14], for example, SpRay is used to visualize a tissue classification process. It is implemented in C++, using OpenGL and QT to provide good portability and good performance.

In the following, we demonstrate the usage of SpRay for microarray data. For the purpose of illustration, we use an artificial 21-dimensional dataset that mimics phases of different cell cycle related gene transcript profiles based on three different clusters of noisy sine waves (Sinusoidal Demo, see Table 1). This test example is similar to the first real world example presented in the next section.

The Gaussian noise modulates both the phase shift and the amplitude (Fig. 2 top/left). The first sinusoidal structures become visible after tuning the transparency modulation to a higher number of data values (Fig. 2 top/right). The used colormap emphasizes these structures such that the noise level can be reduced by removing the noise data points (Fig. 2 bottom/left). In particular the cluster on the base-line (low amplitude) can now be easily removed to emphasize the remaining clusters with a high amplitude which deviate from the baseline (Fig. 2 bottom/right).

Dataset	#Data Points	#Dim's	#Mined Data Points
Sinusoidal Demo	2850	21+0	850
Yeast Cell Cycle	6178	18+3	approx. 800
Half-Marathon	345	8+10	13/6
Microarray Validation	1921	6+3	17

Table 1: Overview of examined artificial (Sinoidal Demo) and real world datasets. For each dataset, we list the number of data points, the number of dimensions (conditions + statistically derived conditions), and the relevant data points.

4 Analyzing Microarray Data

In the course of this study, we visually explored a number of multidimensional data from gene expression experiments. Three of these datasets are discussed in this section and are also summarized in Table 1. They represent different exploration types that represent confirming scenarios (dataset two and three) and an explorative scenario (dataset three). The first real world example (dataset two) focuses on the genes that are active during the yeast cell cycle, and hence expose a similar cyclic pattern, which in turn is used for the model-based analysis. For the second example, SpRay is used as a tool to guide the statistical analysis of differential expression. It is furthermore used to explore the effect of different model-free statistical correction methods to support the selection of the most appropriate one. The third example illustrates how SpRay can be applied as a quality tool for the validation of a new custom-made microarray. All three examples demonstrate typical daily use applications of microarray analysis. Furthermore, their respective analysis employs our visual analytics approach, where statistical data is automatically computed – or its computation is steered based on an employed model – and visually explored together with the original data (see Fig. 1).

Yeast Cell Cycle: Finding Periodic Patterns in Microarray Data

The first dataset is well-known in bioinformatics and describes genes of the yeast *Saccharomyces cerevisiae* that are influenced by the cell cycle (cycle-regulated) [25]. Spellman et al. investigated the periodical variation of gene transcript levels in association with the cell cycle in a comprehensive microarray-based analysis. To get reliable gene expression signals, the cells from yeast cultures were first synchronized by an arrest-release synchronization method resulting in three different gene sets (α, CDC15, elutriation). mRNA was extracted at consecutive time points following synchronization, and gene expression values of more than 6000 genes were measured using two-color cDNA microarrays. The arrays were scanned and the basic analysis was done with common methods for background correction, normalization, and quality filtering of the spot signals. On top of this analysis, cyclicity, correlation, and clustering procedures were employed to quantify and characterize the association of the gene transcript levels with the cell cycle phases.

Spellman et al. found 800 genes which satisfy the minimum criterion for cell cycle regulation that was defined. Follow up analysis of the response of these genes to induce a certain cell phase and the analysis of promoter sites of these genes showed further evidence for a cell cycle association for a subset of these genes.

The yeast cell cycle dataset was closely examined in many papers. Shedden and Cooper [24] re-analyzed the data and derived a more specific conclusion. They found that the randomization of data showed less strong periodic patterns than the experimental data. Therefore, noise and random data fluctuations could be ruled out to contribute to the cyclicity of the data[1].

In this paper, we only used the data of the α factor arrested cells. We re-analyzed the data in a similar way to Shedden and Cooper [24] with a sinusoidal regression fit of cell cycle genes. The expression values $y(t_j)$ of a gene at time points[2] t_j were least square fitted against a linear model with the two harmonic basic curves:

$$y(t_j) = \beta_s \sin\left(\frac{2\pi}{T} t_j\right) + \beta_c \cos\left(\frac{2\pi}{T} t_j\right) + r(t_j). \tag{1}$$

To detect the sinusoidal expression pattern of genes according to the cell development, the period T was set to the nominal interdivision time of 66 minutes specified by Shedden and Cooper [24]. The value $y(t_j)$ is decomposed using Equation (1) into the interesting harmonic part:

$$h(t_j) = \beta_s \sin\left(\frac{2\pi}{T} t_j\right) + \beta_c \cos\left(\frac{2\pi}{T} t_j\right), \tag{2}$$

and the residual part $r(t_j)$ that quantifies the aperiodic content of y_j or oscillations with a significantly different period in comparison to the selected value of T. To evaluate the quality of the fit to the model for every single gene, we calculated the coefficient of determination:

$$R^2 = \frac{SS_R}{SS_t} = \frac{\sum\left(\hat{y}_j - \bar{y}\right)^2}{\sum\left(y_j - \bar{y}\right)^2}, \tag{3}$$

which measures the proportion of variability that is explained by the model SS_R (regression sum of squares) and the total variability SS_t (total sum of squares). The values of R^2 lie between 0 and 1 ($0 \le R^2 \le 1$), where $R^2 = 1$ implies a perfect fit and $R^2 = 0$ no fit. For the visualization, we expressed the harmonic part $h(t_i)$ in a

[1] The first two synchronization methods (α, CDC15) produced good reproducibility of the results, in contrast to the third method (elutriation). Shedden and Cooper suggested in [24] that this may be rather due to the stress response of the cells to the first two synchronization methods, than normal variation of the transcript levels inside an undisturbed yeast cell. This, however, does not limit the use of visual analytics methods to extract information from the α dataset.

[2] While the time point samples in this example itself – not all gene expression data is taking samples at different time points – can be represented as a kind of time-varying datasets, the combination with the statistical derived values for the visual analysis is more flexible in a parallel coordinate representation.

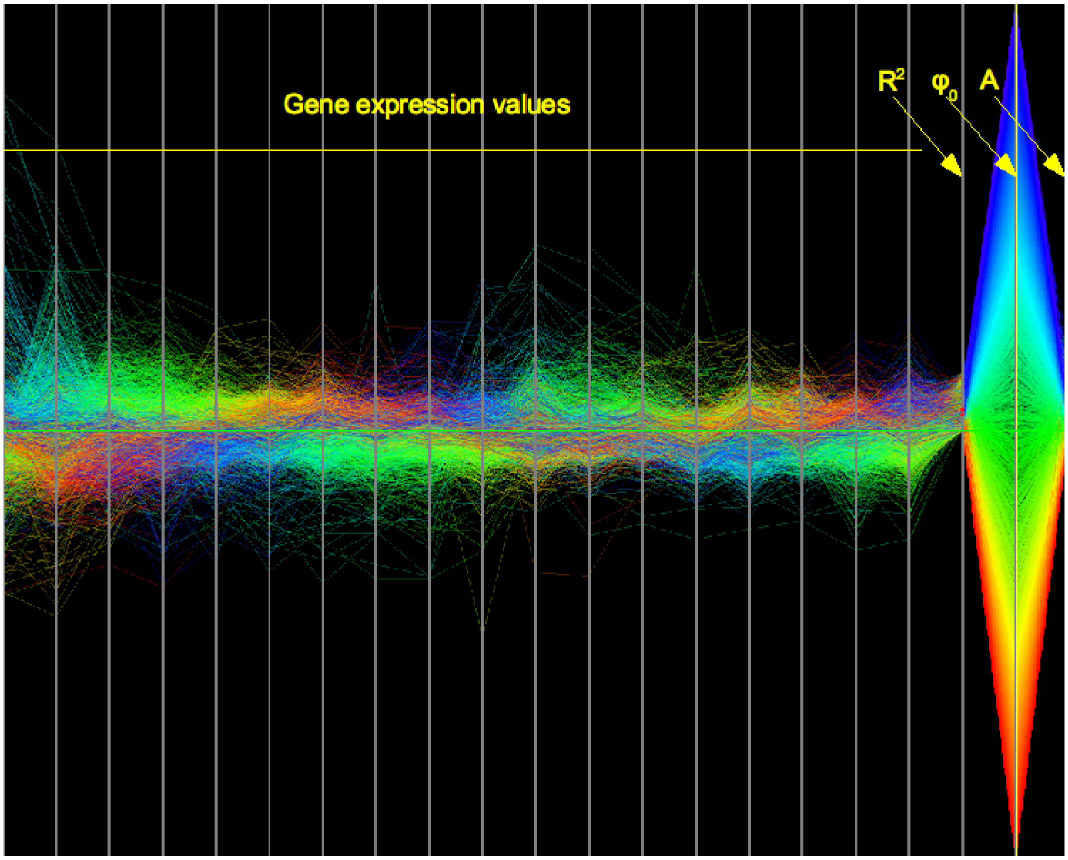

Figure 3: Yeast cell cycle dataset. The first 18 dimensions of this parallel coordinate plot (PCP) correspond to the gene expression values measured by Spellman et al. at the 18 time points for the α factor arrested cells [25]. The last three dimensions correspond to the values obtained by the harmonic regression analysis (HRA): the coefficient of determination R^2, the zero-phase angle ϕ_0, and the amplitude A of the estimated curve (see Section 4 for details). The polylines of the PCP are colored by the zero-phase dimension, so that the periodic changes of the transcript levels of groups of genes can be very easily identified (blue, green).

more descriptive way as a single modulated and shifted sine wave:

$$h(t_i) = A * \sin\left(\frac{2\pi}{T_k} t_i + \phi_0\right). \tag{4}$$

The amplitude A and the zero-phase angle ϕ_0 are determined by the coefficients β_s and β_c and can be calculated with the help of the commonly known addition theorems of trigonometry.

Figure 3 shows the visual representation of the gene expression values in SpRay; all time points of the dataset are presented together with the deduced values R^2, ϕ_0, and A. The coloring of all genes is defined by the zero-phase angle ϕ_0. The periodic changes in gene activity can be very easily identified by visual examination of colored polylines, eg. the visible cyclic patterns of cyan and green polylines. The visualization determined by the derived value ϕ_0 guides the user to a deeper insight of the data, for example, which genes are active in the cell cycle. Through user-interaction with SpRay, it is possible to emphasize further different aspects of the data, for instance to search genes that show a nearly anti-correlation pattern along the cell cycle. Figure 4 shows the result of a specific zero-phase selection on the ϕ_0-dimension, exposing two antisymmetric cycles.

More questions can be answered with the help of other additional views that provide a more detailed view, such as a scatterplot of appropriate dimensions, in particular if the dimensions are not positioned next to each other. The scatterplot of R^2 against the amplitude A colored by phase shift ϕ_0 can be used to determine if only expression profiles of a high amplitude A achieve good R^2 values (close to 1) and how these are related to the zero-phase angle ϕ_0 (see Fig. 5). As we can see in this figure, R^2 and A are highly correlated (both grow in the same direction), while no pattern can be observed from the ϕ_0 colors.

Half-Marathon Dataset: Emphasize Relevant Expression Patterns

The second dataset is taken from a study [30] that investigated the effects of an exhausting endurance exercise on the immune system. It is generally believed that a strong influence exists, which is attributed to both a cellular shift in the composition of the peripheral blood and to changes in gene expression levels. That study used a custom-made cDNA-microarray of immune and stress response related genes to investigate these different aspects in a systematic way. Blood samples were taken from eight well-trained male half-marathon runners in rest before the run (t_0), immediately (up to 15 min) after the run (t_1), and 24 hours after the run (t_2).

The most interesting effects were seen between the status before the run (t_0) and immediately after the run (t_1), hence only these time points are included in this investigation. The study indicated interesting changes in the transcript level of inflammatory genes and even more interesting evidence for an association with the anti-oxidative defense. Both indicate the higher stress level of the body. Here, however, we are interested in evaluating the behavior of the ten different p-value correction methods, as we will discuss below.

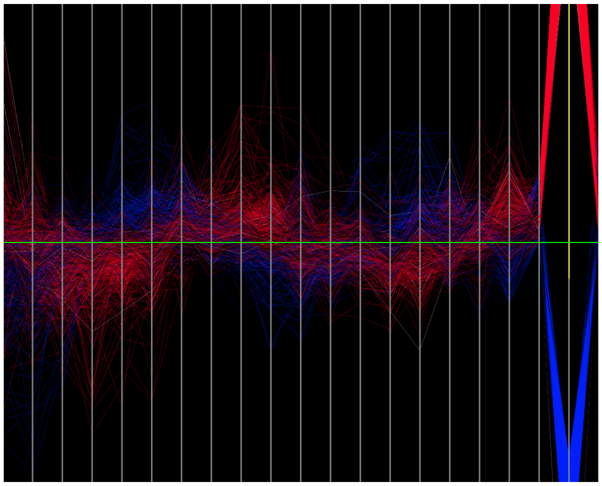

Figure 4: Selection of two groups of genes which show an anti-correlated gene expression pattern along the cell cycle.

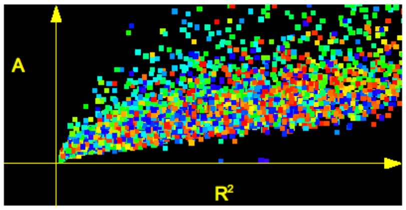

Figure 5: Scatter plot of dimensions R^2 (x-axis) against A (y-axis) colored by dimension ϕ_0. Every point represents one gene profile. The cluster shows an existing relation between R^2 and A, but none with ϕ_0.

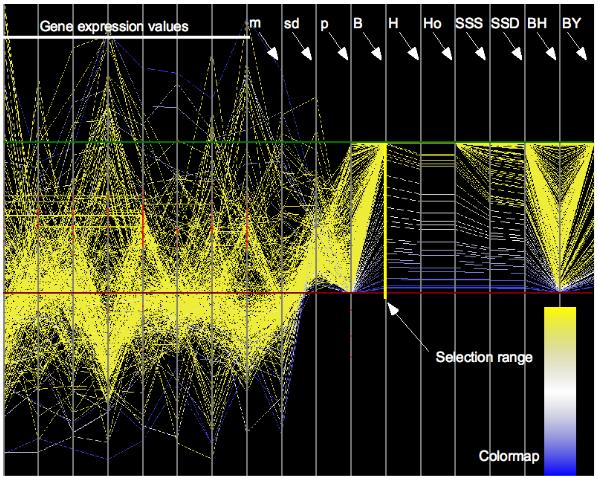

Figure 6: Half-marathon dataset. The PCP displays in the first 8 dimensions the log-ratios of the gene expression values of eight well-trained half-marathon runners before (t_0) and after the run (t_1). The following 10 dimensions correspond to deduced statistical values (mean (m) and standard deviation (sd) of the log-ratios, the raw p-value of a t-test (p), and seven p-values corrected for multiple testing, see Section 4 for details). The green line shows the 1.0-level for the various p-values (no significance) and the red line shows the 0-level (full significance). This figure shows nicely that the p-value of the majority of samples is of very low (yellow) significance. The view of the PCP (with a color-coding according to the p-value corrected after Bonferroni – marked with vertical yellow line) shows very nicely the influence of the different correction methods.

The data of the eight male runners are interpreted as biological replicates and are assigned to 8 dimensions of the PCP (see also Tab. 1). They are represented as the log-ratios of the gene expression values at times t_0 and t_1: $log\left(\frac{expr_{t_1}}{expr_{t_0}}\right)$. 10 added dimensions represent deduced statistical values; mean and standard deviation of the log-ratios of all runners, the p-value of a t-test against the null hypothesis of no difference in gene expression between the points in time t_1 and t_0, and seven p-values corrected by different methods to address the multiple testing problem (Bonferroni (B), Holm (H), Hochberg (Ho), Sidak (SSS, SSD), Benjamini-Hochberg (BH), and Benjamini-Yekutieli (BY)), which are standard methods in bioinformatics[3]. Overall, this results in a PCP with 345 polylines for all genes and 18 dimensions.

The specific choice of an adequate correction method is a non-trivial problem in the context of microarray data analysis. If the correction method is too rigorous, many interesting gene expression changes could be missed (high false negative rate). Also, if the method is not strict enough, too many false positives render the follow up investigations time-consuming, extensive, and expensive. An applicable trade-off must be found based on the goals of the study. Figure 6 gives a good impression of the eight measured values and the ten deduced statistical parameters. The isomorphic luminance-based two-color-coding is defined by the dimension that represents the Bonferroni corrected p-values. This method is the most rigorous and was selected for the study to get very reliable results and a very low false positive rate. It furthermore exhibits a very regular spread over the whole *significance interval*. The most interesting genes are genes whose corrected p-values fall below the defined level of statistical significance, which was predefined to 0.05 in this study. As we can see in Figure 6, the most interesting genes are largely hidden by the large amount of other gene data values. Hence, it is necessary to prevent the irrelevant genes (see Fig. 7 from being represented in the plots, near the green horizontal line) and to emphasize the most interesting genes (see Fig. 8, red colored samples). Note that the red colored expression values here show the 0.2 significance level (13 data points in Tab. 1). The highly significant 6 data points with a p-value (correct after Bonferroni) of 0.05 is the lower sub-section close to the 0-level-line.

Validation of Custom-made Microarray: Outlier Detection

The third example analyzes a dataset that was generated to validate a new custom-made microarray. This microarray was developed as an enhanced successor of the array which was used in the half-marathon study above. The dataset consists of five *self-self* experiments (the first five dimensions of the PCP) and a real experiment (the sixth dimension of the PCP). *Self-self* experiments use the same biological material for both channels of one microarray slide and address the technical sources of the signal error. Consequently, no differential gene expressions should be detected and all measured log-ratios are supposed to be near zero. Nevertheless, the microarray analysis exposes a few extreme outliers which suggest a difference in gene expression (see Fig. 9). The real experiment compares the material from two saliva samples which were taken immediately before and after an extensive *physical training*. In this case, it was expected to see real changes in the expression levels of different genes.

The comparison of the *physical training* experiment to the *self-self* experiments shows a slightly smaller range of all measured val-

[3] Although all these methods are standard, the question which methods are the most appropriate for a specific situation is still disputed.

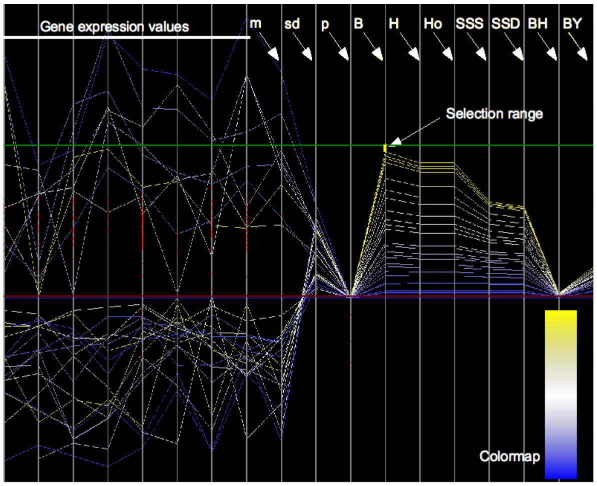

Figure 7: The same data is visualized as in Fig. 6, but all genes with an insignificant difference between gene expression values (corrected (B) p-value larger than 0.99, vertical yellow line) are now culled.

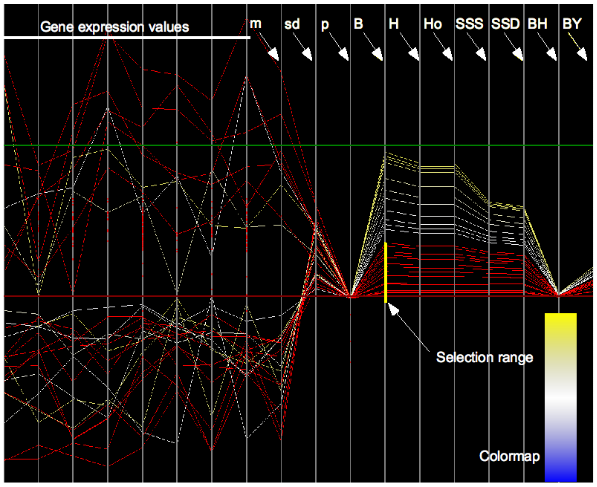

Figure 8: This plot emphasize the significant genes (red) of the study (corrected (B) p-value smaller than 0.2, vertical yellow line).

ues for the first one, but these values are somewhat broader distributed over this range. This can be seen easily in PCP (Fig. 9), the scatterplots (Fig. 10), and the histogram plots (Fig. 11). It clearly depicts the difference of the technical and biological signals.

To get a better understanding of the observed effects, three statistical parameters were added (the last three dimensions). For the first one we computed for each gene the fraction of flags set across all six experiments. These flags are the result of an image analysis (performing spot detection and signal extraction) of the microarray slides and indicate a problematic signal quality. A gene with no flag set across all experiments has the most reliable signal quality, while the genes that had a flag set in each experiment (altogether 12[4]) have the worst reliable signal qualities. Figure 12 shows the PCP after keeping only the very unreliable (high flag rate, yellow) and very reliable expression values (low flag rate, blue). This figure shows the success of the image analysis flagging; reliable (blue) values of the *self-self* experiments consistently deviate only lightly

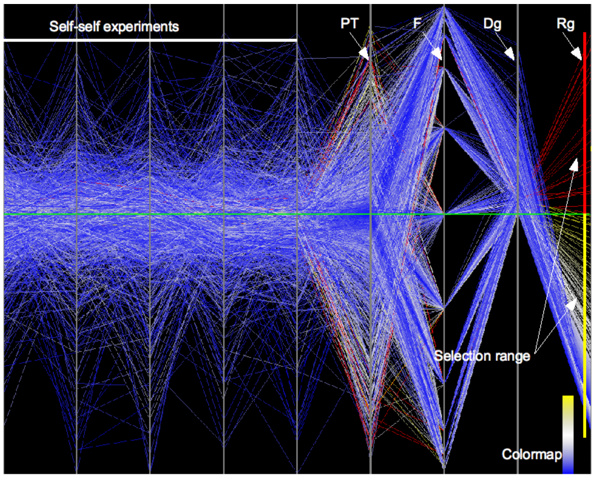

Figure 9: Microarray validation dataset. The first five dimensions represent the self-self experiments (SE_1, ..., SE_5), the sixth dimension the *physical training* (PT), followed by the statistical dimensions flags (F), mean deviation (D_g), and relevance (R_g). The green line indicates the zero level and the vertical red and yellow lines indicate the brushing selections. Note that the vertical axis is logarithmically scaled.

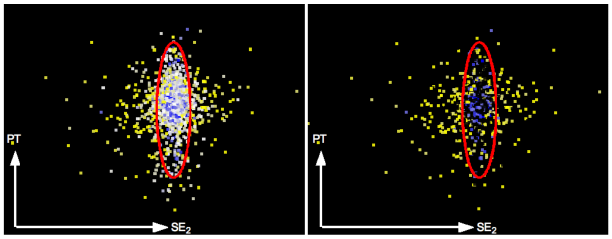

Figure 10: Scatter plot of *self-self* SE_2 experiment (x-axis) and *physical training* experiment (PT, y-axis) before (left) and after removal of the log-ratios of the expression values with medium reliability flags set (right and Fig.12). The *physical training* experiment shows a more compact (red ellipsoid) distribution.

from the zero level, while the majority of unreliable values (yellow) deviate significantly.

The second statistical dimension, D_g, is the mean of the log-ratios $|M_{i,g}|$ for a certain gene g along all *self-self* experiments i (with $N = 5$):

$$D_g = \frac{1}{N}\sum_{i=1}^{N}|M_{i,g}| \tag{5}$$

Genes with D_g deviating substantially from the expected value of zero indicate here a problematic quality.

The final statistical dimension combines the previous deviation metric with the values of the *physical training* experiment such that genes with a high log-ratio $|M_g^{PT}|$ in *physical training* and a low D_g are emphasized:

$$R_g = \frac{|M_g^{PT}|}{D_g}. \tag{6}$$

Here, large values R_g indicate interesting biological signals.

Figure 13 demonstrates the effect of this relevance metric. The irrelevant expression values are removed (yellow selection), while the relevant ones (red) are maintained. This figure shows also nicely how the values classified as relevant expose a high deviation from the zero level for the *physical training* experiment (PT), and a small

[4] To avoid a measurement bias, each experiment is actually performed twice (”dye swap”).

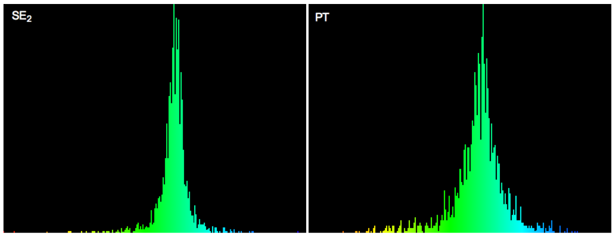

Figure 11: Histograms of *self-self* experiment (SE_2, left) and *physical training* experiment (PT, right). The PT shows a broader distribution as SE_2 or the other (not shown) *self-self* experiments.

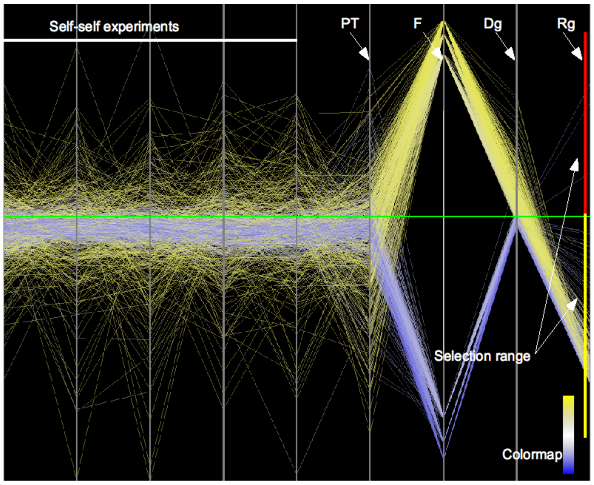

Figure 12: The same data is visualized as in Fig. 9. Log-ratios of the expression values with a high set flag (not reliable) are colored with a yellow luminance variation, and log-ratios with a low (or no) set flag (reliable) are colored with a blue luminance variation. All other log-ratios are removed.

deviation for the *self-self* experiments. Some of these relevant values, however, are also classified as not reliable by the image analysis metric. These outliers are not uncommon for measured microarray data.

5 Discussion

SpRay supports the visual exploration of high-dimensional data, such as microarray data, using parallel coordinates and other information visualization methods. Trends and clusters can be explored through the application of specific transparency modulations and colormaps. However, often the raw data does not provide enough structure to allow a comprehensive analysis. Therefore, we combine visual exploration with statistical analysis methods for a visual analytics approach. This combination allows to discover relations that were difficult to reveal with visual methods alone, since it allows the identification of irrelevant data, which can henceforth be removed from the visual representation.

Another valuable advantage of this combination is the possibility of visualizing the effect of the various analysis methods, as we have shown with the half-marathon dataset. Reliability or instability of the individual methods can be examined and considered for a specific application and allows this way a better understanding of them. The essence of the different correction methods is nicely depicted in all plots of the Half-Marathon dataset (Figs. 6 - 8). The lower span of the raw p-values (p) delivered by the t-test are spread by all correction methods over a greater area. The most rigorous method, and therefore the largest spread of the lower area, is produced by the Bonferroni correction (B). All other methods cause an increasingly smaller spread (H, Ho, SSS, SSD) or a significantly different spread for less rigorous correction methods such as Benjamini-Hochberg (BH) and Benjamini-Yekutieli (BY), which (for our study) emphasize too much on the samples with too little significance (yellow coloring). This significant difference is also visible in the scatterplots of B against BH, and B against BY (Fig. 14). Note, however, that the relationship (vertical sorting) of the expression values between the conditions has not changed through-out the correction methods (Fig. 7).

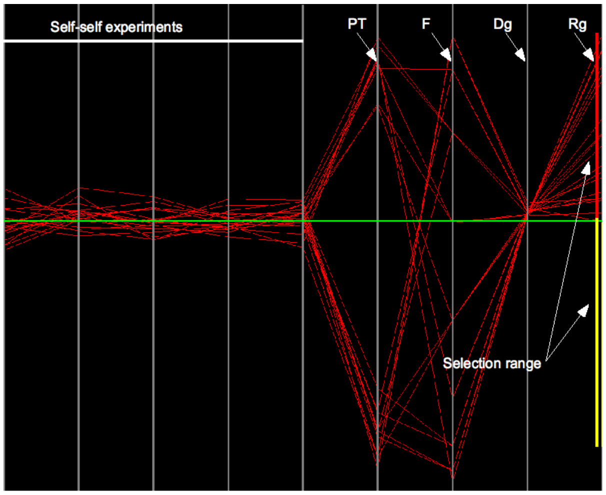

Figure 13: The same data is visualized as in Fig. 9, but all log-ratios of the expression values with a low relevance are removed.

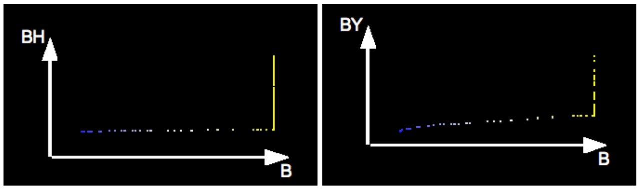

Figure 14: Scatter plot to compare the Bonferroni with the Benjamini-Hochberg (left) and the Benjamini-Yekutieli (right) correction methods for this set of expression profiles. In the plots, the BH correction is obviously much less stringent than the B correction and only somewhat less stringent than BY.

The third example showed how a new custom-made microarray can be validated using SpRay. We showed that virtually all outliers of the *self-self* experiments could be detected by the reliability flags and all relevant expression values were detected by the relevance metric. A visual exploration on the experiment data alone would have probably indicated (wrongly) a dysfunctional microarray.

As mentioned in Section 3, SpRay provides a diverse set of colormaps to be applied to the different dimensions of the parallel coordinates. One of them is the rainbow (hue) map. Although the use of the rainbow map in visualization is highly disputed [3], since it may suggest different closeness or distance to equally distant values, we found that it provides a good mechanism to differentiate the different gene expression values over the many conditions (yeast cell cycle dataset). Other, perceptually ordered or even perceptually isometric colormaps – eg. various temperature maps – provide less qualitative differentiation, hence we opted for the rainbow map. For the other two studies, however, we used luminance variations between two colors (blue/yellow). Note that for all experiments, we are only looking for the qualitative differentiation, not for a quantitative one, hence a perceptual ordered colormap is not required.

Overall, the visual analytics approach of combining original and (semi-)automatically derived data, and the visual exploration of the combined dataset proved quite successful in all three examples.

6 Conclusion and Future Work

In this paper, we presented an approach that combines visual exploration techniques with statistical analysis methods to extract meaningful information from microarray data. In particular this tight integration of statistical analysis with interactive visual exploration – now integrated into the emerging approach of visual analytics – proved to be very powerful and useful. This approach provides an integrated visualization of the original data and the statistically derived value. By visualizing the effect on the data and the derived values at the same time, it allows also the quick validation and evaluation of statistical methods on their appropriateness, which may lead to a more standardized approach to the analysis of microarray data.

SpRay provides numerous statistical analysis methods, which are used in this paper to combine visual exploration and statistical analysis to visual analytics. If, however, more methods are required, a direct link between SpRay and *R* (an important analysis system in bioinformatics) opens up the full statistical functionality of *R*.

Although overplotting has been addressed through the use of colormaps and opacity modulation, large datasets will still suffer from it. Hence, our future work will particularly focus on solutions to this issue.

Acknowledgements

This research has been partially funded through the DFG Priority Program 1335 ”Scalable Visual Analytics”. We like to thank Dr. Derek Zieker (Institute for Clinical and Experimental Transfusion Medicine Tübingen) for providing the half-marathon and the microarray validation datasets and Stephan Symons for proofreading the manuscript.

References

[1] D. Allison, X. Cui, G. Page, and M. Sabripour. Microarray Data Analysis: From Disarray to Consolidation and Consensus. *Nature Reviews Genetics*, 7(1):55–65, Jan 2006.

[2] A. Artero, F. de Oliveira, and H. Levkowitz. Uncovering Clusters in Crowded Parallel Coordinates Visualization. In *Proc. of IEEE Symposium on Information Visualization*, pages 81–88, 2004.

[3] D. Borland and R. Taylor II. Rainbow Color Map (Still) Considered Harmful. *IEEE Computer Graphics and Applications*, 27(2):14–17, 2007.

[4] H. Doleisch, M. Gasser, and H. Hauser. Interactive Feature Specification for Focus+Context Visualization of Complex Simulation Data. In *Proc. of EG/IEEE VGTC Symposium on Visualization*, pages 239–248, 2003.

[5] G. Ellis and A. Dix. Enabling Automatic Clutter Reduction in Parallel Coordinate Plots. *IEEE Transactions on Visualization and Computer Graphics (Proc. of InfoVis)*, 12(5):717–723, 2008.

[6] Y. Fua, M. Ward, and E. Rundensteiner. Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets. In *Proc. of IEEE Visualization*, pages 43–50, 1999.

[7] N. Gehlenborg, J. Dietzsch, and K. Nieselt. Framework for Visualization of Microarray Data and Integrated Meta Information. *Information Visualization*, 4(3):164–175, 2005.

[8] D. Gilbert, M. Schroeder, and J. van Helden. Interactive Visualization and Exploration of Relationships between Biological Objects. *Trends in Biotechnology*, 18(12):487–494, 2000.

[9] S. Havre, M. Singhal, D. Payne, and B. Webb-Robertson. PQuad: Visualization of Predicted Peptides and Proteins. In *Proc. of IEEE Visualization*, pages 473–480, 2004.

[10] J. Hong, D. Jeong, C. Shaw, et al. GVis: A Scalable Visualization Framework for Genomic Data. In *Proc. of EG/IEEE VGTC Symposium on Visualization*, pages 191–198, 2005.

[11] A. Inselberg. The Plane with Parallel Coordinates. *The Visual Computer*, 1:69–92, 1985.

[12] J. Johansson and M. Cooper. A Screen Space Quality Method for Data Abstraction. *Computer Graphics Forum (Proc. of EuroVis)*, 27(3):1039–1046, 2008.

[13] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing Structure within Clustered Parallel Coordinates Displays. In *Proc. of IEEE Symposium on Information Visualization*, pages 125–132, 2005.

[14] F. Li, D. Bartz, L. Gu, and M. Audette. An Iterative Classification Method of 2D CT Head Data Based on Statistical and Spatial Information. In *Proc. of International Conference on Pattern Recognition*, 2008.

[15] L. Linsen, J. Löcherbach, M. Berth, and J. Bernhardt. Differential Protein Expression Analysis via Liquid-Chromatography/Mass-Spectrometry Data Visualization. In *Proc. of IEEE Visualization*, pages 447–454, 2005.

[16] K. McDonnell and K. Mueller. Illustrative Parallel Coordinates. *Computer Graphics Forum (Proc. of EuroVis)*, 27(3):1031–1038, 2008.

[17] M. Novotný and H. Hauser. Outlier-preserving Focus+Context Visualization in Parallel Coordinates. In *Proc. of IEEE Visualization*, pages 893–900, 2006.

[18] T. Peeters, H. van de Wetering, M. Fiers, and J. van Wijk. Case Study: Visualization of Annotated DNA Sequences. In *Proc. of EG/IEEE VGTC Symposium on Visualization*, pages 109–114, 2004.

[19] K. Pradhan, D. Bartz, and K. Mueller. SignatureSpace: A Multidimensional, Exploratory Approach for the Analysis of Volume Data. Technical Report WSI-2005-11, ISSN 0946-3852, Dept. of Computer Science (WSI), University of Tübingen, 2005.

[20] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[21] T. Rhyne, T. Dunning, G. Calapristi, et al. Panel 4: Evolving Visual Metaphors and Dynamic Tools for Bioinformatics Visualization. In *Panel 4, IEEE Visualization*, pages 579–582, 2002.

[22] O. Rübel, G. Weber, S. Keränen, et al. PointCloudXplore: Visual Analysis of 3D Gene Expression Data Using Physical Views and Parallel Coordinates. In *Proc. of EG/IEEE VGTC Symposium on Visualization*, pages 203–210, 2006.

[23] P. Saraiya, C. North, and K. Duca. Visualizing Biological Pathways: Requirements Analysis, Systems Evaluation and Research Agenda. In *Proc. of IEEE Symposium on Information Visualization*, pages 191–205, 2005.

[24] K. Shedden and S. Cooper. Analysis of Cell-Cycle Gene Expression in Saccharomyces Cerevisiae Using Microarrays and Multiple Synchronization Methods. *Nucleic Acids Research*, 30(13):2920–2929, 2002.

[25] P. Spellman, G. Sherlock, M. Zhang, et al. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[26] D. Swayne, D. Lang, A. Buja, and D. Cook. GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization. *Computational Statistics & Data Analysis*, 43:423–444, 2003.

[27] E. Wegman and Q. Luo. High Dimensional Clustering Using Parallel Coordinates and the Grand Tour. *Computing Science and Statistics*, 28:361–368, 1997.

[28] M. Westenberg, S. van Hijum, O. Kuipers, and J. Roerdink. Visualizing Genome Expression and Regulatory Network Dynamics in Genomics and Metabolic Context. *Computer Graphics Forum (Proc. of EuroVis)*, 27(3):887–894, 2008.

[29] H. Zhou, X. Yuan, H. Qu, et al. Visual Clustering in Parallel Coordinates. *Computer Graphics Forum (Proc. of EuroVis)*, 27(3):1047–1054, 2008.

[30] D. Zieker, E. Fehrenbach, J. Dietzsch, et al. cDNA Microarray Analysis Reveals Novel Candidate Genes Expressed in Human Peripheral Blood Following Exhaustive Exercise. *Physiological Genomics*, 23(3):287–294, 2005.