Visualizing Uncertainty in Biological Expression Data

Clemens Holzhüter[a], Alexander Lex[b], Dieter Schmalstieg[b],
Hans-Jörg Schulz[b], Heidrun Schumann[a] and Marc Streit[b]

[a]University of Rostock, Department of Computer Graphics, Rostock, Germany;
[b]Graz University of Technology, Institute for Computer Graphics and Vision, Graz, Austria

ABSTRACT

Expression analysis of ∼omics data using microarrays has become a standard procedure in the life sciences. However, microarrays are subject to technical limitations and errors, which render the data gathered likely to be uncertain. While a number of approaches exist to target this uncertainty statistically, it is hardly ever even shown when the data is visualized using for example clustered heatmaps. Yet, this is highly useful when trying not to omit data that is "good enough" for an analysis, which otherwise would be discarded as too unreliable by established conservative thresholds. Our approach addresses this shortcoming by first identifying the margin above the error threshold of uncertain, yet possibly still useful data. It then displays this uncertain data in the context of the valid data by enhancing a clustered heatmap. We employ different visual representations for the different kinds of uncertainty involved. Finally, it lets the user interactively adjust the thresholds, giving visual feedback in the heatmap representation, so that an informed choice on which thresholds to use can be made instead of applying the usual rule-of-thumb cut-offs. We exemplify the usefulness of our concept by giving details for a concrete use case from our partners at the Medical University of Graz, thereby demonstrating our implementation of the general approach.

Keywords: Visualization, Uncertainty, Expression Data

1. INTRODUCTION

Uncertain data plays an increasingly important role in many application fields, as it becomes apparent that data of a borderline quality poses risks and offers chances at the same time. The risks are the ones associated with erroneous data, which may lead to potentially fatal conclusions, and the chances are those that lie in data which may not be perfect but good enough to potentially lead to a scientific break-through. Especially in a critical area such as biomedicine, where data errors can have fatal consequences, it is common to set rather conservative thresholds, discarding any data that is too uncertain and also some more, just to be on the safe side. Yet, in most cases this procedure will dispose of too much of the data, as data values right above the error threshold still hold the chance of being useful at a second glance. It is this borderline data that we term *uncertain*, and which describes the application-dependent and task-dependent margin between *valid data* and *invalid data*.

We strive to make the user aware of this interval of uncertainty, to allow him to explore it, and thus to make an informed interactive decision on where to put the cut-off threshold. Specifically, we do so for heatmap visualizations of ∼omics (e.g., genomics, proteomics, metabolomics) data derived from microarray experiments[1] – in our case for gene expression data, even though our solution can be applied to many kinds of ∼omics data sets. A gene's expression determines how much of a gene's functional products (mostly proteins) are being produced at a certain time in a specific tissue sample. This information is mainly of value when compared to other tissue samples, acquired either at another time, or taken from other organs or patients, for example. Comparing multiple "snapshots" of gene expression lets scientists reason about the causes or consequences of any observed changes. Gene expression heatmaps are an important tool for this analysis, but an integration of uncertainty information is so far largely unexplored. This may be due to the complexity and diversity of uncertainties in expression data, whereby uncertain data values can stem from a diverse set of sources:

Contact information: C. Holzhüter: cholz@informatik.uni-rostock.de, A. Lex: lex@icg.tugraz.at, D. Schmalstieg: schmalstieg@icg.tugraz.at, H.-J. Schulz: schulz@icg.tugraz.at, H. Schumann: schumann@informatik.uni-rostock.de, M. Streit: streit@icg.tugraz.at

- from *data acquisition* – e.g., introduced by the signal processing and given as a signal-to-noise ratio,
- from *data transformation* – e.g., generated by post-processing steps, such as statistical tests comparing different experiments and given in the form of p-values, and
- even from the *visualization* itself – e.g., produced by overplotting, where heatmap cells get drawn on the same pixel.

In this paper, we target all these types of uncertainty, which often occur in combination, and propose solutions to allow users to decide, how certain a data value has to be in order to include it in an analysis. An integration of these solutions along the lines of Shneiderman's *Visual Information Seeking Mantra*[2] ensures that the entire visual analysis workflow is covered. Our main design decision was to not invent new visualizations, as it is important to utilize the standard representations already offered in the application, in order to provide familiar handling and keep the learning curve low.[3] Instead, we chose to alter and enhance the clustered heatmap to provide the necessary uncertainty information directly within or alongside the heatmap.

2. RELATED WORK

One of the first definitions for uncertainty was given by Hunter[4] as a: "...lack of knowledge about the amount of error [which] is responsible for hesitancy in accepting results and observations without caution." In general, uncertainty is mostly described as a composite of different concepts, such as errors, accuracy, and subjectivity.[5] These types of uncertainty have been summarized by Skeels et al.[6] in the form of a three-level uncertainty classification, which distinguishes **measurement precision** from **completeness issues** such as missing values, and higher order uncertainty implied by modeling assumptions and termed **inference**. Other sources propose uncertainty levels that are more closely related to the common visualization pipeline. For example, Wittenbrink et al.[7] list three sources of uncertainty in data, namely: **data acquisition**, **data transformation**, and **visualization**. Besides differing in terms of their sources, uncertainties can also be differentiated by their types: **statistical uncertainty** and **bounded uncertainty**.[8] The former is given by a distribution with an infinite tail that encodes the most likely value measured for a data point, yet it cannot completely eliminate the small, but nevertheless still existing probability that the actual value is completely different. The latter gives a definite, bounded interval in which the value is guaranteed to lie.

Visualizing bounded uncertainty is usually done by enlarging each data item in a plot to occupy the said interval on the value axis. This results for example in a linechart fanning out if the interval and thus the uncertainty gets larger, as it can be observed in the uncertainty visualization approaches by Streit et al.[9] For statistical uncertainty, statistics provides a large number of plots, which can be used to visualize it. Common examples of such statistical plots are boxplots, Range Plots, and histograms.[10] Overall, it is nowadays established to visualize the uncertainty of the data alongside the data itself.[11] For such an integrated visualization, a number of approaches are established, such as glyphs or animation/blinking.[12] Especially the field of geospatial visualization has developed a wide array of techniques to embed and show uncertainties directly on the map.[13] Griethe et al.[5] finally present a general approach on how uncertainty can influence a visualization on the different stages of the visualization pipeline.

In biomedical applications, uncertainty plays an immensely important role, as experimental error tends to affect the data.[14] Yet, this importance is so far not reflected in biomedical visualization. While there are plenty of statistical methods to deal with errors and establish sensible thresholds describing when to consider data as sufficiently certain,[15] they tend to be considered separately from the actual visualization. Data, which is not certain enough, is simply not passed on to a subsequent visualization and thus missing. Hence, efforts are called for to alleviate this situation and provide visual representations for biomedical data capturing their inherent uncertainty.[16] Current research endeavors in this direction tend to focus on medical applications, such as diagnosis and surgical procedure planning, in which complications can rise due to uncertainty.[17]

For uncertain expression data and its visualization, the situation is very similar: there exist a number of very sophisticated statistical methods to deal with uncertainties, such as Pearson et al.'s *puma* package for Bioconductors,[18] yet so far no integrated visualization method is known. We have explored some first ideas for heatmap visualizations of expression data and uncertainties in a poster presentation in 2010.[19] This paper describes solutions developed from these first ideas.

3. CONCEPTIONAL APPROACH TO VISUALIZE UNCERTAIN DATA

In order to make use of uncertain data instead of disregarding it, our approach consists of three rather straightforward steps, which are necessary to make an informed choice on a threshold as motivated in the introduction:

1. **identification** of uncertain data between an upper and lower threshold,
2. **visualization** of this uncertain data in the context of the certain data, and
3. **variation** of the thresholds between valid, uncertain, and invalid data.

Each of these three steps is introduced and discussed on a general level in the following as well as detailed and exemplified in Section 4.

3.1 Identification of Uncertainty in Expression Data

Gene expression data can be described as tabular data forming a matrix M with a set of genes G (or proteins in case of protein expressions) as rows and a set of experiments E as columns. A cell $(i, j) \in M$ contains thus the expression value of the i-th gene in the j-th experiment. The typical heatmap H precisely reflects this data setup by assigning each matrix cell $M(i, j)$ a corresponding cell $H(i, j)$ in view space and color-coding the gene regulation value onto it. Here, green indicates downregulated genes, black an intermediate regulation, and red indicates upregulated genes. Common expression data sets have up to several thousand genes and a few dozen replicated experiments, leading to a rather narrow but high heatmap.

Errors resulting in uncertain data are introduced in a variety of ways. To identify them, we adopt the approach of Wittenbrink[7] who considers different uncertainty sources along the data processing workflow. So we can identify and estimate/quantify the uncertainty independently for each step of the workflow and then in turn use it to determine its mentioned upper and lower threshold. The different error sources are described below.

By data acquisition: As any experimentation, the procedure of microarray experiments, as well as the digitalization and quantification of the experiment results, are inherently error-prone. This error of data acquisition can be measured by different means, ranging from single measures, such as the signal-to-noise ratio (SNR), to more complex aggregates of multiple measures, such as the *Quality Control Score* defined by Wang et al.[20] All of them have in common that they yield a local error quantification, which is defined per cell. To account for such errors, it is common practice to design experiments so that biological replicates can counter-balance this effect. Nevertheless, a conservative threshold separating the valid from the invalid cells does not only lead to the elimination of a few extra measurements, but to the omission of entire genes.

By data transformation: Additional errors may get introduced when data is processed. A typical example for this is an aggregation or averaging of multiple replicated experiments, commonly used to even out fluctuations of measurements and data acquisition errors. Yet, if the individual replicates differ too much, there is a high likelihood of a more serious flaw in the data, which would be hidden by averaging. This considerably impairs the validity of a subsequent gene expression analysis. Hence, statistical tests are used to compare the results from replicated experiments, and the error is then given in the form of a resulting confidence interval. These relative error measurements do not allow to reason which of the two replicates is the faulty one. Again, applying conservative thresholds on the confidence may lead to discarding potentially useful genes.

By visualization: This last kind of error is introduced by graphical artifacts that occur for example when trying to displaying more data than there are pixels available (overplotting) or when using 3D graphics with its natural limitations (occlusion, perspective distortion, etc.) While 3D heatmaps do exist – see for example the ViGeCo tool[21] – they are rarely used in day-to-day gene expression analysis, making overplotting the main concern for large heatmaps. For its quantification, a few measures exist, which work on a per-pixel basis, e.g., Ellis' and Dix'[22] *overplotted%* or Tufte's[23] *Data-Ink-Ratio* or *Data Density*. Yet, while the data density is the same for every part of a heatmap, *information density* is not: if two nearly identical rows get overplotted, half of the data but nearly no information is lost. We accommodate for this important distinction by proposing a new way of computing the visual uncertainty in terms of this notion of information loss: we scale up the shown overplotted heatmap to the size of the original unhampered heatmap H having one pixel per heatmap cell and compute their pixel-to-pixel differences. Thus, when both images are identical, because no overplotting occurred

or only identical rows were overplotted and thus no information was lost, there are no differences between both images. But when two differing cells are overplotted, scaling up this overplotted heatmap and comparing it with the original one allows for detecting this information loss.

Interestingly, for all three classes of error, their lower limits, meaning the threshold between valid data and uncertain data, can be defined relatively straightforwardly. For errors in signal processing and image acquisition, there exist thresholds that reflect the commonly achievable measurement precision and which where thus established as a quasi standard. An example for such a common threshold is a SNR of 3, even though more fine-grained measures and more precise thresholds have already been proposed.[24] The same holds true for statistical testing, where confidence intervals are defined by setting the cut-off for the p-value to 0.05 as a rule of thumb, which is proven to work well.[25] As for the visualization error, the lower threshold at which the view is definitely valid is the point where no overplotting occurs and each data item is individually visible.

More complicated is the upper error limit, beyond which data is definitely considered invalid and no longer uncertain. The need for upper thresholds is reflected in some publications, where they are used to separate true negatives from false negatives.[26] Yet, they are far less common and there are usually no established values for them. Until such values are established the biomedical expert has to set these upper limits manually according to his experience with the used lab equipment and other factors that are hard to generalize. Examples for upper limits used by the experts involved with our use case are given in Section 4.

3.2 Visualization of Uncertain Expression Data

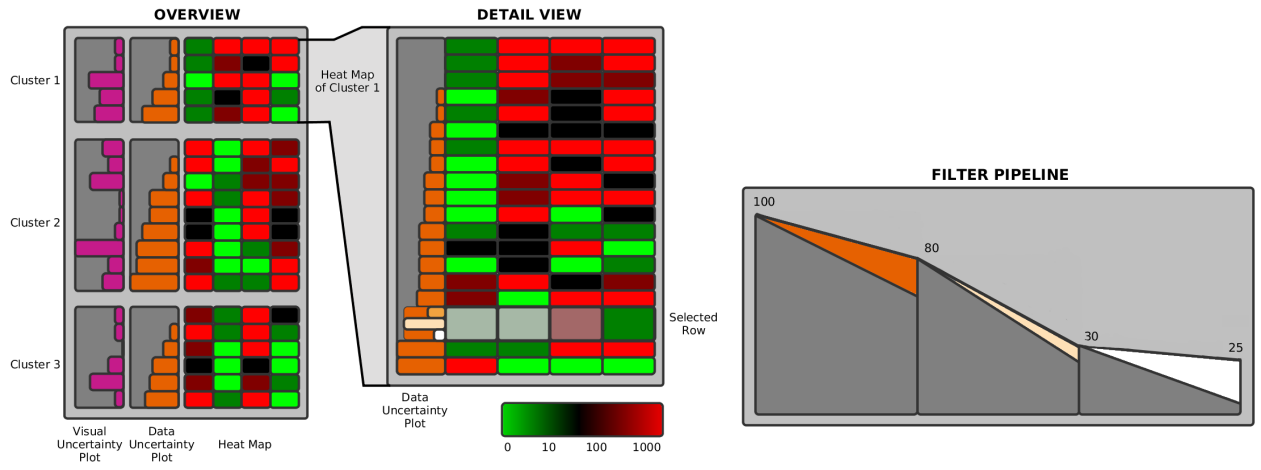
Before visualizing the data, it is sensible to discard all invalid data items (genes) which are beyond the upper error thresholds and thus are definitely of no use. Hence, only the valid and the uncertain data are displayed in the following. When communicating uncertainty in the data, three aspects are important:

- **qualitative aspect:** making the user aware that data is affected by uncertainty
- **quantitative aspect:** showing the extent of the uncertainty affecting the data
- **provenance aspect:** indicating which error(s) in the data have lead to its classification as uncertain

Besides these aspects regarding the uncertainty in the data, it is also mandatory to faithfully display the actual underlying data, as its uncertainty is merely secondary information. To make it easier for biomedical experts to make use of our visualization approach, a central design decision is to preserve the overall look & feel of the standard clustered heatmap display for the actual values and add the uncertainty information onto it. For visualizing larger heat maps we use an overview and a separate detail view. While the overview always shows the whole valid data set, we split the overview into disjoint clusters (this follows the assumption that heatmaps are only valuable when the data is clustered). The detail view shows only the values belonging to one of the clusters. To make it obvious which cluster from the overview is shown in the detail, we use ribbons to connect the cluster and the detail view. In general, our approach enhances heatmaps by supplying additional information per row in an attached extra column at the left of the heatmap. This fits well with the overall layout of a heatmap, which usually extends more in the vertical direction than it spreads out horizontally. In contrast to other uncertainty visualization techniques, it also uses reordering of the rows of the heatmap to subtly encode their relative uncertainty in their position without changing the overall form of the heatmap.

Overview: The overview aims to give a complete view of the potentially very large data set. It provides only a low granularity, which precludes the mapping of detailed uncertainty values. For the visual encoding of the uncertainty in the overview, we distinguish between visual uncertainty, which by design occurs only in the overview, and data uncertainty (summarizing both data acquisition and data transformation uncertainties). We encode the uncertainty using row-wise barplots, where a high bar indicates a high uncertainty. Figure 1(a) illustrates the two barplots side by side: on the far left in purple, the encoding for the visual uncertainty is shown. To its right, in orange are the plots for the data uncertainty, and next to it, the actual overview of the heatmap clusters.

Additionally to the encoding using barplots, the rows are sorted, so that all uncertain rows are placed together in one region of the heatmap. This is possible, as the actual order of rows within a heatmap is not fixed. In the case that clusters must be kept together, the sorting is done locally for each cluster. Furthermore, sorting is a



(a) (b)

Figure 1. The overview and detail view of our uncertainty visualization approach for heatmaps are shown in (a). The purple barplot indicates the visual uncertainty due to overplotting, the orange barplot the maximum acquisition or transformation uncertainty. In the detail view, no overplotting occurs and thus no visual uncertainty is shown. On demand more information is shown for selected rows in the detail: the extent of the individual uncertainties, as well as from which cell they originate by desaturating the cells according to their uncertainty. The filter pipeline concept is illustrated in (b) with three different filters set to individual thresholds. It gives information on how many items have been discarded as invalid and how many are classified as uncertain by the current cut-offs. The colors of the filters in the pipeline correspond to the colors of the bars in the detail row in (a), so that they can easily be identified and matched.

necessity, as within the overview, rows are heavily overplotted, because the available screen space is insufficient to assign each row at least a single pixel line within the heatmap. This massive overplotting would completely disguise individual uncertain rows in the heatmap, and the user would not be aware of them, until he zooms in. Yet, when sorting all uncertain rows into the same region, they can still be identified – at least in bulk – when being collapsed into a single line of pixels.

Due to the space limitations of the overview and the data density being already very high for the scaled down heatmap, the provenance aspect is not encoded in it, but can be viewed on demand in the detail view.

Detail view: As this view shows only a selection of rows from the overview (usually a single cluster), it offers enough space to not only eliminate visual uncertainty by overplotting, but also to display additional uncertainty information for each row in the heatmap. This additional information is mostly concerned with the provenance aspect that tells why exactly a row of the heatmap exhibits uncertainty. The detail view gives this information on two levels of granularity: per row and per cell.

The display per row is a rather seamless extension of the barplot on the side of each row in the overview. It splits up into multiple bars per row – each one detailing one of the different error sources, of which there can be multiple per class, such as different acquisition errors and different transformation errors (illustrated in the third row from the bottom in Figure 1(a)). The interval, in which the bars are shown, is exactly the interval of uncertainty. No visible bar means here that the error value is at least as low as the lower error threshold below which all data is assumed to be valid. Whereas a full bar means the error is close to the upper error threshold above which all data is assumed to be invalid. This indicates to the user which error(s) render a row uncertain and how much the conservative error threshold must be raised, so that a row would be considered valid and be included in subsequent analytics.

In order to identify which cell(s) of a row are responsible for a row's overall error, further details can be brought up on demand. Selecting a row results in an enlargement of the selected element, where the certainty on a cell level is encoded using saturation. Since we have multiple levels of uncertainty, the uncertainty that contributes the most to the entire row's uncertainty (and therefore corresponds to the highest bar for the row) is chose by default. Exploring the cell-wise uncertainties for the other levels can be achieved by clicking on the

corresponding bar, upon which the saturation is calculated based on the uncertainty of the cells for the selected level. This way, an error value close to the validity threshold would leave the color of the heatmap cell almost untouched, whereas an error value close to the invalid region would white-out the cell completely. Being able to investigate the concrete source of an uncertainty in this way is very valuable: If, for example, many genes are highly uncertain and it turns out that this uncertainty stems in all cases from a single faulty replicate, it would make sense to exclude this one faulty column instead of many rows.

3.3 Variation of the Uncertainty Thresholds for Expression Data

This last part of our approach goes beyond the pure communication of the uncertainty, its underlying error sources, and the concrete measurement it stems from. It allows the user to interactively adjust the thresholds used for classifying the data into valid, uncertain, and invalid. It is carried out directly within the diagram of the filter pipeline used for the error handling. This pipeline is shown in a separate but linked view, which extends our interactive filter pipeline[27] – see Figure 1(b). It depicts the thresholds, as well as in which order they are applied, and how many genes are classified as valid or uncertain in each of the filtering steps. Manipulating the thresholds by sliding them up or down renders more rows valid or uncertain respectively, and visual feedback is given by reordering the rows of the heatmap accordingly. In this way, after investigating the heatmap and its associated uncertainty, the biomedical expert can readjust the threshold for each error source according to his findings. This can be done by either including more genes than were originally deemed valid by the conservative threshold, or maybe even excluding more genes in the case of overall imprecise measurement results due to experimental conditions.

While this filtering method covers adjustments in the data space with respect to acquisition and transformation errors, zooming allows the user to also manipulate the uncertainty in view space. As the user controls the generation of the visualization, he does not have to stop at the level of adjusting a threshold for what is still acceptable, but can instead reduce the error source of overplotting by zooming into the heatmap and thus giving more space to a region of interest. The effect of zooming into the overview is directly reflected in the barplot display, where the visual error plays a decreasing role for the overall uncertainty as the user zooms in – all the way to the detail view where no more overplotting occurs.

Together with the overview and the detail view, these filter and zoom mechanism form an approach that covers Shneiderman's mantra and can also be used in its spirit starting from the overview, applying zoom and filter, and retrieving details on demand. This is exemplified for a concrete use case in the following section.

4. USE CASE

For this case study we collaborated with our partners from the Medical University in Graz. They are interested in explaining differences between two genetic variations of mice: one shows symptoms when exposed to poison, while the other variation does not. The symptoms are similar to a well known disease in humans, so that they speculate that the genetic susceptibility might have similar causes in mice and in men. To find out about the activity of the mice's genes, they conducted gene expression microarray experiments at different points in time (0 days, 7 days, and 8 weeks). Each experiment was repeated three times in order to account for both: measurement errors and individual traits found in only one of the mice.

Identification of uncertainty: The expert starts his analysis by loading the data set, which consists of the actual expression values and a signal-to-noise ratio for each value. Since an analysis based on too noisy data is invalid, typically those values with a signal-to-noise ratio of 3 or less are filtered.[24] Data above this threshold is considered certain. By not strictly removing those values, but instead specifying a lower bound, the expert can extend the data to be considered for an analysis significantly. However, the data, which lies between these bounds, is uncertain to varying degrees, thus introducing the first level of uncertainty.

To first find the genes which change in expression, fold-change filters are applied. Again, an established threshold of a 3-fold change marks the border to certain values, while genes in that 2-fold to 3-fold change region can be considered uncertain. By setting such a fold-change filter for both genetic variations of mice, the second and third level of uncertainty is introduced.

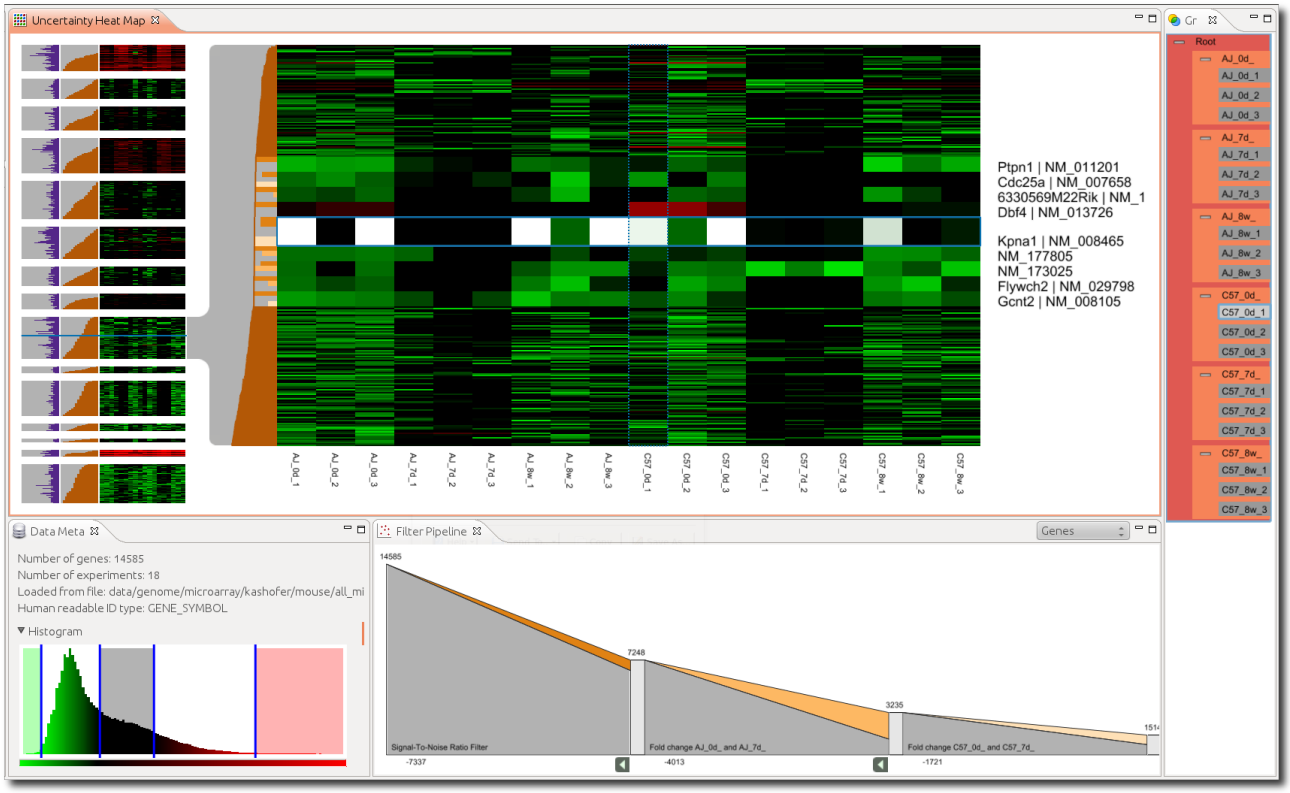


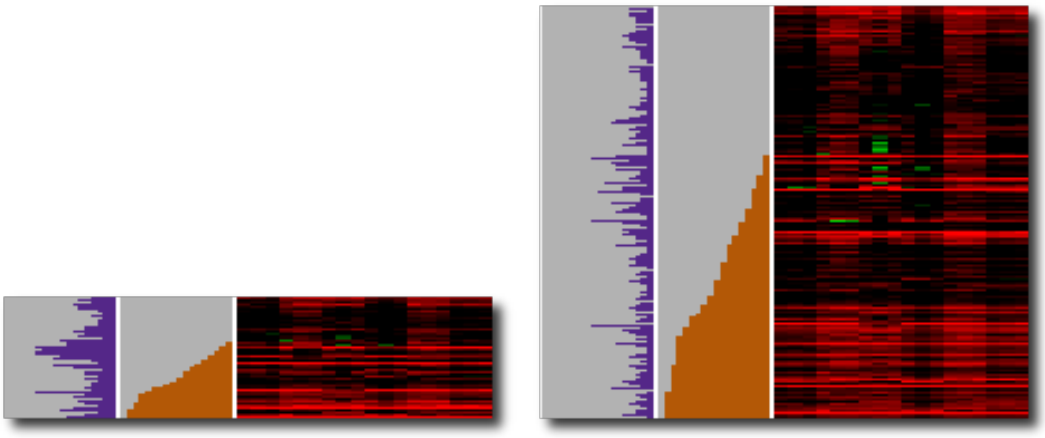
Figure 2. A screenshot of the entire system showing overview (left), detail view (right), and the filter pipeline (bottom) in our implementation.

After the user has set the limits that define the valid, uncertain, and invalid intervals for all levels of uncertainty, the invalid genes are filtered from the data set. The remaining set of certain and uncertain genes is then visualized as a heatmap. The overview visualization also introduces the visual uncertainty, measured by the amount of overplotting and weighted by the spread of the overplotted values. This way, an upregulated expression (red) overplotting a downregulated expression (green) results in a higher visual uncertainty than overplotting matrix cells with similar expression.

After identifying the different sources of uncertainty and before the user can start the actual analysis, he runs additionally a partitional clustering on the data, as otherwise the 14,500 genes in the data set are hard to overview. Interactively interwoven visualization and clustering techniques, such as the Matchmaker,[28] can help to find a good clustering, which partitions the data well. Figure 2 shows the resulting heatmap, which reflects the generated clusters in its overview representation.

Overview visualization of uncertainty: By inspecting the overview, the user gets a feeling for the data set itself as well as the contained uncertainty. Initially, the user judges the quality of the resulting clusters. Usually, a high quality clustering yields homogeneous clusters. If this is not the case, this might be an indication for a non-optimally chosen cluster algorithm or additional settings like the distance measure. In case the user is not satisfied with the clustering result, he can re-run the algorithm with altered settings. The plots attached to the left hand side of the heatmap in the overview show the visual and data uncertainty, see Figure 3.

Plotting the overall data uncertainty for each row in sorted order yields curves of different shape and form for each cluster. For our use case, the most interesting shapes are those that are concave and have a long tail, as for example in the fifth cluster from the bottom in Figure 2. This means that for this particular cluster, uncertainty above the threshold does not increase drastically. Instead, the uncertainty grows very slowly and has a dramatic increase only at the end. Clusters exhibiting such a distribution of uncertainty are most likely to contain a number of genes that are just marginally more uncertain than those below the error threshold and



(a) (b)

Figure 3. The overview of a single cluster rendered at two different zoom levels. The plots visualizing the visual uncertainty and the data uncertainty are attached on the left of the heatmap. The visual uncertainty plot in (a) indicates a strong variation within a particular gene's expression and serves as a pointer for further investigation on a closer zoom level or within the detail view, thereby exposing an inhomogeneous area within an otherwise homogeneous cluster. (b) shows the same cluster as (a), but rendered on a bigger portion of the screen. Now, the downregulated genes (green) that caused the peaks are clearly visible in the heatmap.

are thus good candidates to be included with the data.

The visual uncertainty plot is a valuable feature for the discovery of outliers in the clusters that are otherwise not visible in the overview due to the information loss introduced by overplotting. Peaks in the plot indicate sub-regions or single outliers that cannot be detected in the overview heatmap because of the aggregation of multiple expression values to a single pixel. A zoom feature allows the user to investigate what data in the cluster causes the peaks. The visual uncertainty is recalculated on the fly when the size of the rendered overview changes. Reasons for this are a zoom action or the resizing of the window. Figure 3 shows the overview at two different zoom levels for the same cluster, so that the effect of zooming on the visual uncertainty is clearly visible. In addition to the zoom feature, the user can always select the cluster to inspect its data in the detail view.

Variation of uncertainty: The heatmap in the overview enhanced with the uncertainty of each row already gives a good enough impression for adjusting the threshold if necessary. To that end, the filters that were used on the data are shown in their order of application in a linear pipeline, which can be seen at the bottom of Figure 2. It does not only depict how many genes are left after each filter operation, but also how many of them are uncertain with respect to each filter. This makes it clearly visible, which filter would discard (too) many uncertain values and may thus give leeway to adjust the threshold accordingly. The adjustment of the thresholds can be done interactively and is guided by a histogram indicating exactly how many items get included or eliminated by raising or lowering the thresholds, respectively. After a change has been made to the thresholds in the pipeline, it becomes mandatory to re-cluster the newly filtered data set, as additional or missing data items may have an influence on the clustering result. This readjustment of the thresholds can be applied as often as necessary, until good cut-off values are found for each filter. In general, it makes sense to only alter one filter at a time and then evaluate the outcome, as otherwise the effects of multiple adjustments mix and it becomes hard to discern which threshold adjustment was responsible for a certain outcome.

For manipulating the uncertainty thresholds on the data, a proxy such as the pipeline view is needed to interact upon. In contrast, the visual uncertainty can be manipulated directly by zooming in or out of the overview. The higher the zoom level, the fewer data items must be displayed on the screen, leading to reduced overplotting, and hence the visual uncertainty drops. This effect is also visible in Figure 3, where the visual uncertainty (purple barplot) is significantly lower overall for the zoomed-in figure than for the zoomed-out figure.

Detail visualization of uncertainty:

As the user selects an interesting cluster, its content is presented in a larger detail view for further inspection. Since visual uncertainty does not longer occur, only data uncertainty is visualized left hand to the heatmap. A fish-eye enlargement affects not only the selected, but closer rows within a small range as well. Within this range all data uncertainty barplots are fragmented into multiple sub-bars, making it obvious from which source the uncertainties in the selected area are coming from. These sub-bars can be individually selected to show the amount of individual uncertainty in each experiment for the selected gene, revealing weather all experiments or only one outlier was responsible for the gene to be uncertain. This is achieved by desaturating the cell color to white, to accentuate highly uncertain experiments. As shown in Figure 2, the selected row yields multiple highly uncertain experiments for the first uncertainty source, identifying it for an invalid gene. To easier pinpoint interesting sections from the overview within the detail, e.g., the source of a high visual uncertainty peak, the user selected row is marked within the overview with a blue line. The obligatory labeling of genes and experiments are also provided within the detail view. Due to the amount of genes, labels are only shown for genes which are part of the fish-eye selection.

The proposed uncertainty visualization concept is realized as a part of the Caleydo information visualization framework[29]*. Caleydo is written in Java and uses the Java OpenGL binding (JOGL) for rendering.

5. CONCLUSION

Our approach proves to be highly flexible, as it enables the user not only to pursue the task we have mainly focused on – the minimization of false negatives – but also to target a number of other interesting questions. The most obvious is probably its ability to minimize false positives, in case of a too optimistic error threshold. A scenario, in which this has to be done, would be the identification of conjectured cancer patients, which is also determined from gene expression profiles. To be on the safe side, in this scenario the threshold is set optimistically for determining patients who are possibly affected to rather include more results from uncertain genes than to miss out on one. Here, our method can be applied to cut down on the false positives by raising the threshold instead of lowering it, after investigating a patient's gene expression in a heatmap. This can be achieved with the same setup. It would merely be the workflow that changes, as now, for instance, the convex uncertainty curves are of interest, which signal a high overall uncertainty and thus genes which should potentially be dismissed and not be incorporated in the profiling.

Another interesting question, which can be explored with our method, is how many and which of the error sources must actually be included to yield a representative overall uncertainty. This is hard to answer in general and depends a lot on the actual data, for which a potentially large number of uncertainties can be measured and derived by considering additional error models or simply performing the same statistical test with varying subsets of data. Since the filter pipeline also allows the user to set error thresholds all the way to the maximum value, which corresponds to not applying a threshold for a given error at all, error sources can thus be interactively incorporated or excluded from the overall filter. Investigating the changing effect of switching filters on and off is a powerful exploratory tool for the user to get an understanding of which error sources actually affect the data and thus the outcome of the expression analysis. Knowing the important error measures can reduce the cost of experimentation, as no longer the whole range of possible errors has to be measured, and also the preprocessing time, as probably only a few statistical tests are meaningful and thus needed.

A last problem would be to not only find the right error measures, but also the appropriate order in which to consider them. For example, many measures are based on statistical properties of the data set they are computed for. Hence, it is important whether one measure is computed first and used to filter the data set before another measure is computed and used for subsequent filtering, or the other way around. Our approach permits to alter the order of the filters by dragging axes onto different positions – very much like a parallel coordinates plot does. This way, different orders of filters can be tried and explored for their effects to the clustered heatmap.

Overall, these points illustrate the adaptability and usefulness of our approach, as it apparently does not only cover the specific use case of our collaboration partners, but also a wider range of questions associated with uncertainty in expression data. Hence, in future work we will expand our approach into these other application

*http://www.caleydo.org

scenarios, which will hopefully further illuminate the aspect of uncertainty in data and visualizations, as well as its role in visual analysis of biological data.

One particular interesting concrete idea to follow up on is aimed to make additional use of visual uncertainty: By using the calculated information loss it is feasible to develop heterogeneity-driven distortion techniques for the detail view, which only shrink homogeneous regions of the visualization to a degree that directly correlates to their homogeneity – similar to the approach used by Oelke et al.[30] Furthermore, it is also possible to combine this approach with user-defined importance weights for different features of the data. These importance weights could be assigned with respect to the task at hand, e.g., for the identification of differently regulated genes, equally regulated genes could by shrunk or vice versa. This approach can be applied locally by using distortion lenses, or globally to provide a compact overview.

ACKNOWLEDGMENTS

This work was funded in part by the graduate school *dIEM oSiRiS* of the German Research Foundation (DFG), the *InGenious* project (385567) granted by the Austrian Research Promotion Agency (FFG), and the *Caleydoplex* project (P22902) granted by the Austrian Science Fund (FWF). The authors thank Karl Kashofer from the Institute of Pathology at the Medical University of Graz for valuable discussions and the data set featured in Section 4. Moreover, we are grateful to Christian Partl for his help with the implementation.

REFERENCES

[1] Mayer, B., ed., [*Bioinformatics for Omics Data - Methods and Protocols*], no. 719 in Methods in Molecular Biology, Springer (2011).

[2] Shneiderman, B., "The eyes have it: A task by data type taxonomy for information visualizations," in [*Proceedings of the IEEE Symposium on Visual Languages (VL '96)*], 336–343 (1996).

[3] O'Donoghue, S., Gavin, A., Gehlenborg, N., Goodsell, D., Hériché, J., Nielsen, C., North, C., Olson, A., Procter, J., Shattuck, D., Walter, T., and Wong, B., "Visualizing biological data-now and in the future," *Nature Methods* **7**(3s), 2–4 (2010).

[4] Hunter, G. J. and Goodchild, M. F., "Managing uncertainty in spatial databases: Putting theory into practice," *URISA Journal* **5**(2), 55–62 (1993).

[5] Griethe, H. and Schumann, H., "The visualization of uncertain data: Methods and problems," in [*Proceedings of the Simulation and Visualization (SimVis '06)*], 143–156, SCS Publishing House (2006).

[6] Skeels, M., Lee, B., Smith, G., and Robertson, G., "Revealing uncertainty for information visualization," in [*Proceedings of the Conference on Advanced Visual Interfaces (AVI '08)*], 376–379, ACM Press (2008).

[7] Wittenbrink, C. M., Pang, A. T., and Lodha, S. K., "Glyphs for visualizing uncertainty in vector fields," *IEEE Transactions on Visualization and Computer Graphics* **2**(3), 266 –279 (1996).

[8] Olston, C. and Mackinlay, J. D., "Visualizing data with bounded uncertainty," in [*Proceedings of the IEEE Symposium on Information Visualization (InfoVis '02)*], 37–40, IEEE Computer Society Press (2002).

[9] Streit, A., Pham, B., and Brown, R., "A spreadsheet approach to facilitate visualization of uncertainty in information," *IEEE Transactions on Visualization and Computer Graphics* **14**(1), 61–72 (2008).

[10] Potter, K., Kniss, J., Riesenfeld, R., and Johnson, C., "Visualizing summary statistics and uncertainty," *Computer Graphics Forum* **29**(3), 823–832 (2010).

[11] Zuk, T. and Carpendale, S., "Theoretical analysis of uncertainty visualizations," in [*Proceedings of the Conference on Visualization and Data Analysis (VDA '06)*], **6060**, 66–79, IS&T/SPIE (2006).

[12] Pang, A. T., Wittenbrink, C. M., and Lodh, S. K., "Approaches to uncertainty visualization," *The Visual Computer* **13**(8), 370–390 (1997).

[13] MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., and Hetzler, E., "Visualizing geospatial information uncertainty: What we know and what we need to know.," *Cartography and Geographic Information Science* **32**(3), 139–161 (2005).

[14] Zbilut, J. P. and Giuliani, A., "Biological uncertainty," *Theory in Biosciences* **127**(3), 223–227 (2008).

[15] Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S. B., and Bassett, D. E., "Rosetta error model for gene expression analysis," *Bioinformatics* **22**(9), 1111–1121 (2006).

[16] Nielsen, C. B., Cantor, M., Dubchak, I., Gordon, D., and Wang, T., "Visualizing genomes: techniques and challenges," *Nature Methods* **7**(3s), S5–S15 (2010).

[17] Lundström, C., Ljung, P., Persson, A., and Ynnerman, A., "Uncertainty visualization in medical volume rendering using probabilistic animation," *IEEE Transactions on Visualization and Computer Graphics (Vis '07)* **13**(6), 1648–1655 (2007).

[18] Pearson, R. D., Liu, X., Sanguinetti, G., Milo, M., Lawrence, N. D., and Rattray, M., "puma: a bioconductor package for propagating uncertainty in microarray analysis," *BMC Bioinformatics* **10**(211) (2009).

[19] Holzhüter, C., Schulz, H., and Schumann, H., "Enriched heatmaps for visualizing uncertainty in microarray data," in [*Poster at the Eurographics Workshop on Visual Computing for Biomedicine (VCBM '10)*], (2010).

[20] Wang, X., Ghosh, S., and Guo, S. W., "Quantitative quality control in microarray image processing and data acquisition," *Nucleic Acids Research* **29**(15), e75 (2001).

[21] Tominski, C. and Schumann, H., "Visualization of gene combinations," in [*Proceedings of the Conference on Information Visualisation (IV '08)*], 120–126, IEEE Computer Society Press (2008).

[22] Ellis, G. and Dix, A., "The plot, the clutter, the sampling and its lens," in [*Proceedings of the Conference on Advanced Visual Interfaces (AVI '06)*], 266–269, ACM Press (2006).

[23] Tufte, E. R., [*The Visual Display of Quantitative Information*], Graphics Press, Cheshire, Coneeticut, 2nd ed. (1983).

[24] He, Z. and Zhou, J., "Empirical evaluation of a new method for calculating Signal-to-Noise ratio for microarray data analysis," *Applied and Environmental Microbiology* **74**(10), 2957–2966 (2008).

[25] Hall, P. and Selinger, B., "Statistical significance: Balancing evidence against doubt," *Australian Journal of Statistics* **28**(3), 354–370 (1986).

[26] Pounds, S. and Morris, S. W., "Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values," *Bioinformatics* **19**(10), 1236–1242 (2003).

[27] Geymayer, T., Lex, A., Streit, M., and Schmalstieg, D., "Visualizing the effects of logically combined filters," in [*Proceedings of the Conference on Information Visualisation (IV '11)*], IEEE Computer Society Press (2011).

[28] Lex, A., Streit, M., Partl, C., Kashofer, K., and Schmalstieg, D., "Comparative analysis of multidimensional, quantitative data," *IEEE Transactions on Visualization and Computer Graphics (InfoVis '10)* **16**(6), 1027–1035 (2010).

[29] Lex, A., Streit, M., Kruijff, E., and Schmalstieg, D., "Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context," in [*Proceeding of the IEEE Symposium on Pacific Visualization (PacificVis '10)*], 57–64, IEEE Computer Society Press (2010).

[30] Oelke, D., Janetzko, H., Simon, S., Neuhaus, K., and Keim, D. A., "Visual Boosting in Pixel-based Visualizations," *Computer Graphics Forum* (2011).