

Interactive Exploration of Medical Data Sets

Heimo Müller¹, Kurt Zatloukal¹, Marc Streit², Dieter Schmalstieg²

(1) Medical University Graz, (2) Graz University of Technology

{heimo.mueller|kurt.zatloukal}@meduni-graz.at, {streit|schmalstieg}@icg.tugraz.at

Abstract

This paper describes an interactive data exploration system for molecular and clinical data in the field of personalized medicine. It addresses the essential but to date unsolved problem of how to identify connections between genetic variants and their corresponding diseases or the response to certain drugs and treatments, respectively. It is therefore necessary to connect genetic with clinical data in order to categorize specific subgroups of patients with certain disease features. The huge amount of data provided by molecular analytical methods (e.g. data on genetic alterations, proteomic or metabolomic data) can only be analyzed by applying statistical methods and bioinformatics. However, even standard methods of statistics and bioinformatics fail when the data is inhomogeneous – as is the case with clinical data – and when data structures are obscured by noise and dominant patterns. The structure of large medical data sets is made visible by using so called object- and attribute-glyphs, which can be arranged in a two dimensional space and synchronized with a set of visualization views.

1. Introduction

We developed a data exploration system for the “visualization of” and “navigation in” huge molecular and medical data spaces using a specifically designed physical workplace for collaborative analysis of huge inhomogeneous data sets in the application field of personalized medicine. Our system aims to

- support hypothesis driven data analysis,
- support different contextual views on the data, and
- identify hidden correlation in unconnected databases

In order to fulfill these goals, the main challenges in this undertaking were the provision of a set of 3D

glyphs for the medical problem domain and methods for interactive exploration of medical databases. With the ability to arrange the glyphs in a two-dimensional space utilizing different spatial grammars and to synchronize different visualization through linked views, a medical expert can in the truest sense of the word, travel through his data space.

2. Related Work

Due to the huge number and different structures of molecular and medical parameters (e.g. mutations, genetic polymorphisms, epigenetic alterations, gene expression data, data on protein expression and protein modifications, data on metabolites, diagnosis of disease, laboratory parameters, imaging data, treatment, outcome, accompanying diseases, life style etc.) the coupling of clinical metadata and molecular data sets is still an unsolved problem. While research on efficient visual data mining of very large data sets is a hot topic in current research, the particular approach of simultaneous visualization of molecular and patient-specific clinical data and the use of high-throughput low-latency user interface techniques are novel.

Basic research on information visualization and user interfaces typically focuses on isolated aspects of the visual appearance of the user interface, but many of the approaches face problems when attempting to apply these techniques to real world problems. This has led to a stronger demand of research in visualization, which addresses the literally huge (in terms of data size) problems of medical data analysis. Conversely, medical research is concerned with the comprehension of the hidden meaning of medical data, by any means available. The efficiency of a new analysis tool can only be assessed by the expert working with this tool. As the research question is strongly determined by the requirements of medical experts and the visualization algorithms rely on real world data, a workflow was developed within an interdisciplinary approach.

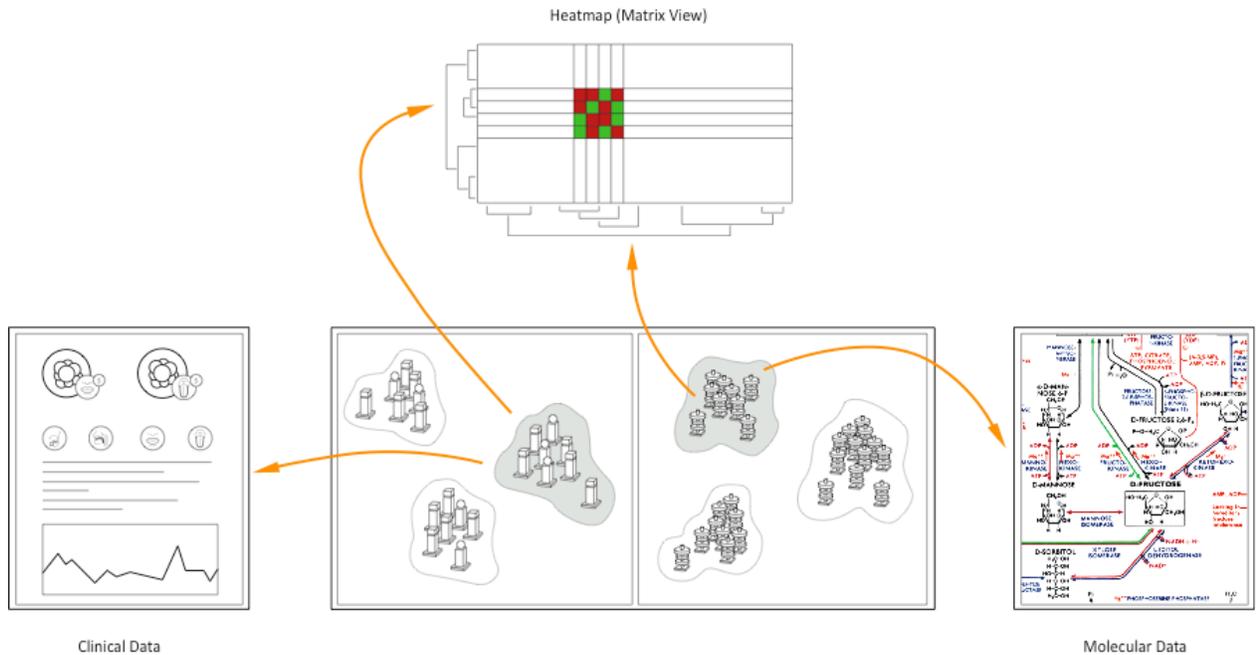


Figure 1 - Overall Architecture

A method for the integrated visualization of microarray data is presented by Grinstein and Smrtic [1][2]. Our approach in focus and context interfaces and large screen displays will be particularly based on the research of Lamping and Rao [3] Baudisch et al. [4] and Kosara et al. [5].

Visual data exploration methods on large data sets, especially hierarchical data structures are described by Hege et al. [6], Keim and Kirgel [7], Grinstein and Meneses [8]. Hinneburg, Keim and Wawryniuk from the University of Konstanz developed a software solution (HD-Eye) for the visualization of high-dimensional data and Fekete and Plaisant worked on “Interactive Information Visualization of a Million Items”. Visualization techniques for multivariate and multidimensional are described by Santos et al. [9].

3. Visualization Workbench

The starting point for the overall architecture, see figure 1, was the requirement to overview huge data sets (several ten thousand medical records), to group and select experiments and disease records and analyze several hundreds of experiments each with several thousand gene expression values and a rich set of clinical data in a fast and intuitive way.

3.1. Data Import and Attribute Description

Before we start with the visual data analysis the following data pre-processing steps are carried out:

1. Data integration and unification of the medical terminology
2. Description of the data structure and attribute scales
3. Value normalizations
4. Mapping to graphical attributes.

In the data description step we specify for each attribute the scale of measurement (discrete, continuous, categorical, ordinal, interval, nominal) and the range of values covered by an attribute. This information is later on used in the automatic verification of the glyph generation process. After the normalization of all attribute values the user determines which attribute of the input data is mapped to a specific graphic variable (shape, color, value, texture, position, size and orientation) of a glyph.

Graphical variables as introduced by Bertin [10] can be characterized according to the scale of measurement and their length.

The *shape* of a glyph is purely nominal and should therefore never be mapped to ordinal data values. However we can recognize a nearly infinite variety of shapes (the shape variable is “very long”). The perceptual variable *color (hue)* is a nominal variable, even though the wavelength of light assigns an ordering to colors, the human perceptual system takes no notice of it. There is some cultural ordering imposed on hue (red is “hotter” than blue), but it is weak because not all hues are related. A non-color deficient person can distinguish between seven and ten

million different colors; however, color is a deeply subjective attribute, therefore not more than 10 to 20 carefully chosen color values should be used in glyph generation. *Saturation* (the brightness of a glyph) and the *texture* (with respect to the grain size of the texture) are ordered and can be mapped to an ordinal scale. Saturation and texture are short variables, i.e. roughly 10 values can be distinguished in an effective way.

The *position* of a glyph can be mapped to ordinal values, and is a very finely grained (long) variable. The *size* of a glyph element can also be mapped to ordinal values; however it is “shorter” than the position variable. Finally the *orientation* of a glyph can be mapped to an ordinal data value, but is a very short graphical variable, i.e. only very few different orientations can be perceived.

3.2. Object and Attribute Glyphs

Object glyphs represent a subset of different clinical parameters, e.g. age, tumor staging, survival data, etc. within a complex graphical sign, whereas attribute glyphs visualize the value distributions of a single parameter for either all objects or a subset of objects. Figure 2 shows an outline of an object glyph summarizing up to 6 clinical parameters and an attribute glyph for the distribution of a tri-state attribute.

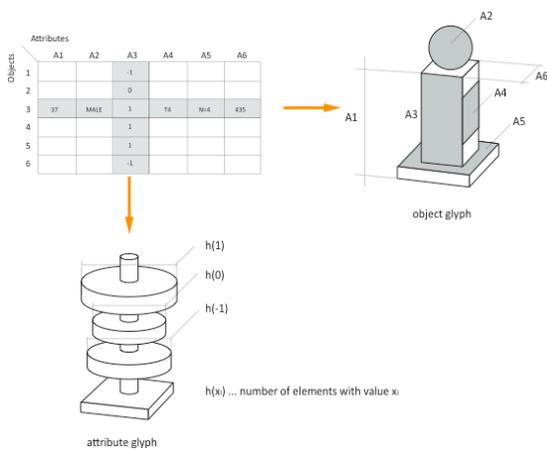


Figure 2 - Object and Attribute Glyph

Figure 3 shows an object glyph, which represents a tissue sample with the attributes sample age, tissue localization, as well as the tissue T, N, M and R grading attributes.

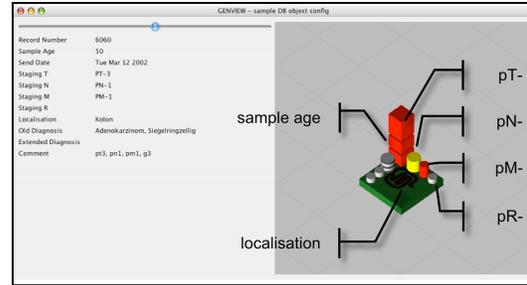


Figure 3 - Object Glyph of a Tissue Sample

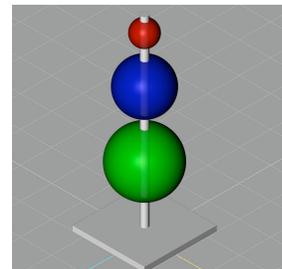


Figure 4 - Attribute Glyphs showing Gene Expressions

Figure 4 shows an attribute glyph representing the distribution of 3 stage attributes (e.g. gene expression) of a series of experiments. The size of the lower sphere shows the number of inactive / deregulated elements the middle sphere of neutral elements and the upper sphere the number of active elements. In figure 5 a clustering result for approx. 100 experiments and 180 genes is shown.

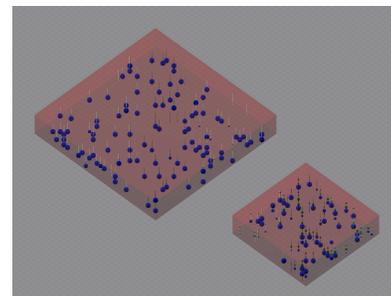


Figure 5 - Clustering of Attribute Glyphs

3.3. Pathway Visualization

After the researchers have identified relevant clinical cases, associated genetic analysis data from these cases is investigated. This investigation uses microarray data from tissue samples along with genetic pathway information to relate the activations of

multiple genes. The pathway visualization module has been developed to facilitate this task [11].

Pathways are abstract graphs that model cellular functions. The graph nodes represent genes while the edges depict signals or chemical reactions. KEGG [12] and BioCarta (www.biocarta.com) are widely used pathway databases that are currently integrated in the system. While current research is mainly focused on the re-routing of the graphs by using automated layout algorithms our system takes the pathway images as they are and just superimposes the graph nodes [13]. As a result important meta-information contained in the hand-crafted pathways can be preserved while getting a fully interactive pathway representation.

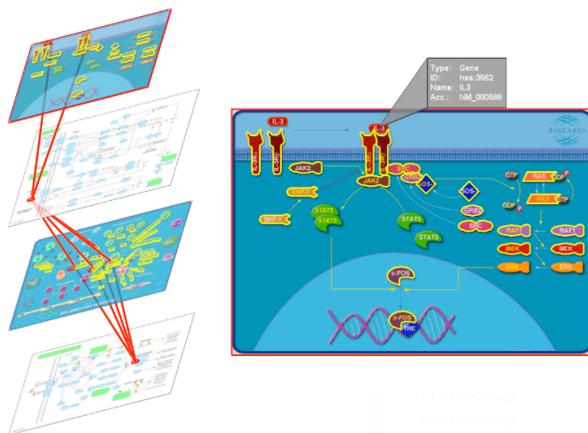


Figure 6
Stacked 2.5D Pathway Visualization

Due to the fact that genes can be part of multiple reactions and processes they often occur in several pathways. Figure 6 shows how the same gene is highlighted and connected in a stacked 2.5D pathway representation. The user selects a gene in one pathway and the system dynamically loads all related graphs to the stacked representation. Line connections among the pathway planes help users to visually identify relationships. By selecting the pathway texture in the stack the graph is moved to the focus position. This interactive data presentation is particularly important to gain a better understanding of the overall network of biological reactions and its relationships.

The interactive pathway visualization is of limited use without being related to the microarray data reflecting expression values of thousands of genes. Therefore it is essential to map the gene expression data directly onto the pathways. In figure 7 a time series experiment is color coded onto the pathway

nodes. In this case a specific gene is up-regulated, i.e. activated, at time point t_0 (red) and down-regulated in t_2 (green). This enables the user to investigate genes in the context of biological processes.

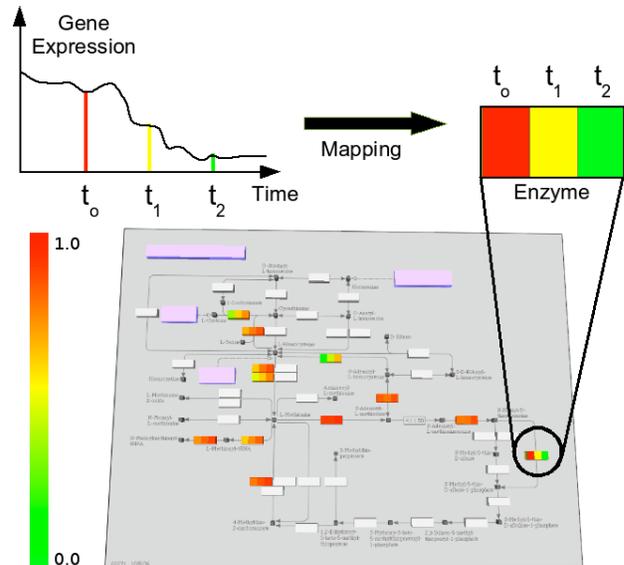


Figure 7
Gene Expression Mapping Onto Pathways

This approach closes the workflow starting from the clinical patient data over the patients' microarray data to the abstract pathway models.

4. Case Study

In the following case study we demonstrate the methods for the exploration of a disease database, containing information about 1.5 million disease records. For this huge number of elements we have chosen a simple object glyph representing the T staging (color), survival (color flag) and the disease free survival (height of the glyph). In the first step the expert chooses a specific organ, e.g. colon, and gets a visual overview of about 11.000 cases together with the distribution (histogram) of the most important attributes. Figure 8 shows the result of an organ selection, where the corresponding glyphs are arranged in within a rectangle ordered by the T staging and disease free survival. This initial arrangement gives an overview of about 11.000 cases and just gives an impression about the number of colon cases and the T staging by the primary color of the glyphs.

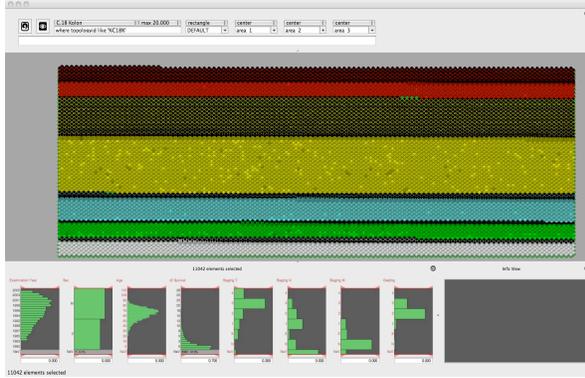


Figure 8 - 11.000 Colon Cancer Cases

In the histogram view (lower part of the user interface) the user can select a group of objects by their value range (red slider) and immediately sees the distribution of the selected elements (green area) in relation to the overall distribution of elements (blue area). In figure 9 we selected cancers from a specific examination period (left slider) and see the subgroup histograms of the patients' sex, age, disease free survival and the tumor stage (T staging).

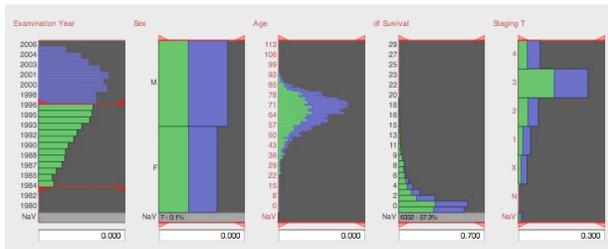


Figure 9 - Attribute Distributions and Selection

The selected subgroup can now be arranged according to different spatial grammar. In figure 10 the objects were grouped by the examination year in two periods: 1988-1993 and 1998-2003. Within a group male patients are on the left side and female patients on the right side. Finally the vertical position of an object is determined by the patients' age and the horizontal position by the T staging. The disease free survival is visualized using the height of the glyph.

With the help of iterative clustering algorithms, - we use a simple version of ant-based clustering, - the user can separate objects into subgroups. Figure 11 depicts the result of such a clustering with the distance

function determined by equal shares through the T staging and the disease free survival attributes.

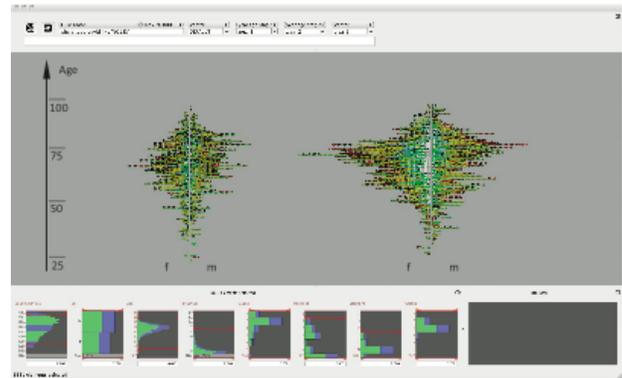


Figure 10 – Comparison of colon cancer diagnosed in the two time periods: 1988-1993 and 1998-2003

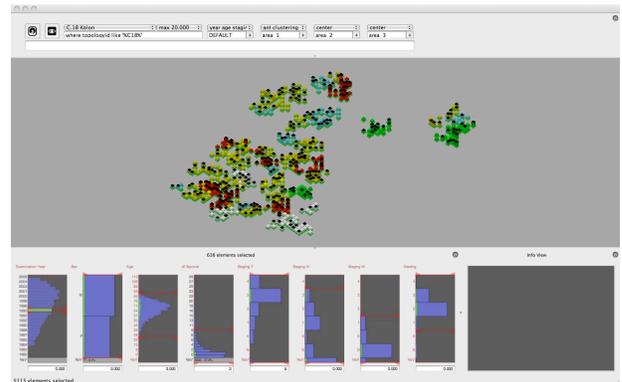


Figure 11 - Clustering Results

The user can zoom in and out in a very fast way, as the glyph rendering is fully hardware accelerated and the application was further optimized for short response times. With a mouse over operation the full database record of the covered object is shown in the right window and within the histogram views the values of the active object are highlighted (red color).

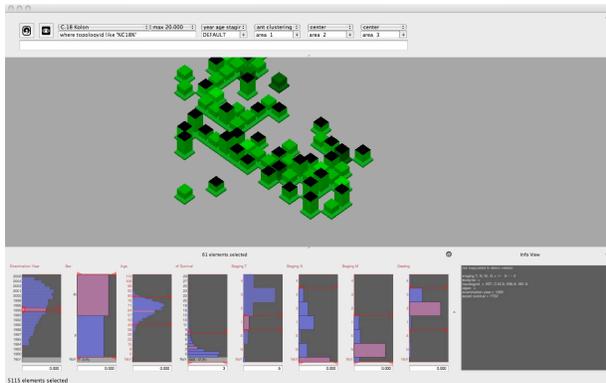


Figure 12 - Detail View

5. Discussion

In the application domain of personalized medicine the following goals are covered:

Medical experts can simultaneously overlook different huge data sets, e.g. whole genome gene expression profiles and comprehensive clinical data. Such an overview is given in the object domain – the expert can observe from several thousand up to one million objects at the same time – and also on the temporal domain, i.e. different subgroups can be generated and compared in a “fluid way”. With this functionality the data sets can be depurated (removal of faulty insertions, harmonization of notations) and pre-processed for later analysis steps. Especially for the clinical data this is essential, as the “coupling” with molecular data is very much dependent on the original data quality.

Through the “coupling” of molecular data and a broad spectrum of clinical data, the medical expert can, for instance, identify connections between genetic parameters, patient subgroups, and drug responses. Furthermore, effects of changes in the health care system on disease outcome can be easily detected (e.g., effect of improved diagnosis on early detection of cancer; effect of new treatments or follow up measures on patient survival) (see also figure 10). Using iterative clustering and computational steering, an expert is able to interact with the analysis process and may identify new medically relevant disease sub-entities. Using a set of linked views additional information (e.g. gene pathways, tissue images, patient history data, etc.) can be attached directly to the observed genes, experiments and patient subgroups.

6. Conclusions and Future Work

The analysis of huge inhomogeneous data sets is one of the main challenges in personalized medicine, which aims at more specific diagnosis and treatment of diseases of individual patients. The diseases of individual patients are characterized on the basis of several parameters including molecular data (e.g. genetic, proteomics or metabolomic data) as well as a broad spectrum of medical data (e.g. family history, laboratory parameters, clinical phenotypes, histopathological diagnoses or treatments). Currently, no suitable tool is available to cope with the increasing demands of data integration in personalized medicine. To address this need we develop new data visualization and interaction methods, which facilitate the detection of correlations between clinical and a variety of molecular data. With the help of large scale displays and multi-layer visualizations experts can perform a hypothesis-driven interactive data analysis.

We developed 3D glyphs for the visualization of a subset of different clinical parameters, e.g. age, staging, survival data etc. and simple 2D glyphs and 3D glyphs for the visualization of value distributions. Both types of glyphs will have symmetry axes (like a crystal) in order to provide an efficient way of spatial arrangement for the interactive definition of hierarchical structures and subgroups.

Currently, we are working on a validation study for an intuitive medical icon library and prepare a benchmark data set of interrelated clinical and molecular data together with several medical hypotheses. In the future, we plan to set up an infrastructure for long-term evaluation activities across the multiple medical application domains within the Biobank initiative of the Medical University in Graz.

For the evaluation of medical glyphs we use state-of-the art methods for qualitative analysis of the graphical user interface (interviews, questionnaires, heuristic evaluations) and will apply methods for the evaluation of the visual analytic applications, as described by Tory [14] and Plaisant et al. [15, 16]. For the special case of the evaluation of DNA and tissue microarray analysis we build on the work of Saraiya [17] and for the evaluation of high level tasks, e.g. perceiving relationships or making conclusions on the work of Amar [18] and Dastani [19].

Furthermore, we will focus our research on automatic glyph generation, interactive definition of spatial grammar (arrangements, levels of details,

blending of visual stimuli) and the development of further visualization modules for clinical data and time series.

7. Acknowledgments

This work was supported by grants from Styrian Zukunftsfonds, the FIT-IT program of the Austria research funding agency FFG and the Austrian Genome Program (GEN-AU).

8. References

- [1] Grinstein G., Integrated, Tightly-Coupled, High-Dimensional Analysis and Visualization for Microarray Expression Data in CHI's Data Visualization and Interpretation Conference Proceedings, 2003.
- [2] Smrtic MB., Grinstein G., Interactive Visualization of Microarray Data on Pathways, Proceedings of the 2005 BioIT Conference, Boston, MA., 2005.
- [3] Lamping, J. and Rao, R., The hyperbolic browser: A focus+context technique for visualizing large hierarchies. *Journal of Visual Languages and Computing*, 7(1):33–35., 1996.
- [4] Baudisch, P., Good N., and Stewart P., Focus Plus Context Screens: Combining Display Technology with Visualization Techniques. In *Proceedings of UIST 2001*, Orlando, FL., 2001.
- [5] Kosara R., Miksch S., Hauser H., Focus and Context Taken Literally, *IEEE Computer Graphics and its Applications*, Special Issue: Information Visualization, pp. 22-29, 22(1), Jan.-Feb., 2002.
- [6] Hege H., Hutanu A., Kähler R., Merzky A., Radke T., Seidel E., Ullmer B., Progressive retrieval and hierarchical visualization of large remote data, *Proceedings of the Workshop on Adaptive Grid Middleware*, 2003.
- [7] Keim D., Kirgel H.P., *VisDB: Database Exploration Using Multidimensional Visualization*, *IEEE Computer Graphics and Applications*, Volume 14, Issue 5, 1994.
- [8] Grinstein G., Meneses C., *Visual Data Exploration in Massive Data Sets*, in *Information Visualization in Data Mining and Knowledge Discovery*, Morgan-Kaufmann Publishers, 2001.
- [9] Santos S, Brodlie K., Gaining understanding of multivariate and multidimensional data through visualization, *Computers and Graphics*, Vol 28, Issue 3, June 2004.
- [10] Bertin, J., *Graphische Semiologie. Diagramme, Netze, Karten*, Walter de Gruyter, Berlin 1974.
- [11] Streit M., Kalkusch M., Kashofer K., Schmalstieg D., Navigation and Exploration of Interconnected Pathways, To appear in: *Proceedings of EuroVis2008*, Eindhoven, Netherlands, May 2008.
- [12] Kaneshia et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic acids research* 34, D354–7, 2006.
- [13] Streit M.: *Metabolic Pathway Visualization Using Gene-Expression Data*. Master's thesis, Graz University of Technology, 2007.
- [14] Tory M., Möller, Human factors in visualization research, *IEEE Transactions on Visualization and a. Computer Graphics*, Vol. 10. No. 1, 2004.
- [15] Plaisant C., The challenge of information visualization evaluation, *Proceedings of the working conference on Advanced visual interfaces*, 2004.
- [16] Plaisant, C., Grinstein, G., Scholtz, J., Whiting, M., O'Connell, T., Laskowski, S., Chien, L., Tat, A., Wright, W., Gorg, C., Liu, Z., Parekh, N., Singhal, K., Stasko, J. *IEEE Computer Graphics and Applications* 28, 2, March-April 2008, pp.12-21 *IEEE Digital Library* 2008.
- [17] Saraiya P., North C., Duca K., An insight-based methodology for evaluating bioinformatics visualizations, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 11. No. 4.
- [18] Amar, R, Stasko J., Knowledge precepts for design and evaluation of information visualizations, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 11. No. 4, 2005.
- [19] Dastani M., The role of visual perception in data visualization, *Journal of Visual Languages and Computing* vol. 13(6), 2002.