Medical Glyphs in Personalized Medicine

Heimo Müller, Michael Kalkusch, Fritz Wiesinger, Doris Ulrich

Information Visualization, Glyphs

ABSTRACT

This paper describes an interactive data exploration system for the visualization of molecular and clinical data and the navigation in" 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 gene data and clinical data in order to categorise specific subgroups of patients with certain disease features. The huge amount of data provided by molecular analytical methods (genetic polymorphisms, gene expression data, proteomics) can only be analysed 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 structures of the data spaces are made visible by using innovative methods of visualisation based on multiple high resolution displays in combination with data projection technologies. As input parameter for analysis of data spaces we make use of the human visual capacity to grasp complex patterns to reveal hidden structures and correlations.

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 datasets in the application field of personalized medicine. Our systems aims to

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

One of the main challenges in this undertaking was the provision of a set of validated 3D glyphs for the medical problem domain and methods for layout and interaction between these visual data objects using large screen displays in combination with data projection technologies. In order to meet the user's requirements we support a broad range of medical data formats. Furthermore the visualization software has to be efficient in terms of the human-computer interaction and visualization methods have to be validated in the medical domain. In the application domain of personalized medicine the following goals will be met:

1. Medical experts will be able to have an overview of data sets, e.g. whole genome gene expression profiles and clinical data from several hundred patients. Such an overview will be 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, harmonisation of notations) and pre-processed for later analysis steps. Especially for the clinical data this is absolutely essential, as the "coupling" with molecular data is very much dependent on the original data quality.
2. 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. Using iterative clustering and computational steering an expert will be able to interact with the analysis process. Using a set of overlay layers additional information (e.g. gene pathways, tissue images, patient history data etc.) can be attached directly to the observed genes, experiments and patient subgroups.

Due to the huge number and different structures of molecular and medical parameters (e.g., genetic polymorphisms, gene expression levels, protein expression and protein modifications, diagnosis of disease, laboratory parameters, imaging data, treatment, outcome, accompanying diseases, life style etc.) the coupling of clinical metadata and molecular datasets is still an unsolved problem. While research on efficient visual data mining of very large datasets is a current topic of research, the particular approach of visualizing molecular and patient-specific clinical data together, combined with 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 is perfectly addressed by 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, an interdisciplinary approach is essential.

A method for the integrated visualization of microarray data is presented in (Grinstein, 2003) and (Smrtic and Grinstein 2005). Visual data exploration methods on large data sets, especially hierarchical data structures are described by (Hege et al., 2003; Keim and Kirgel, 1994; Grinstein and Meneses, 2001). Hinneburg, Keim and Wawryniuk from the University of Konstanz devolped a software solution (HD-Eye) for the visualisation of high-dimensional data and Fekete and Plaisant (University of Maryland) worked on “Interactive Information Visualization of a Million Items”. Visualization techniques for multivariate and multidimensional data can be found in (Hege et al., 2003; Santos et al., 2004). Our approach in focus and context interfaces and large screen displays will be particularly based on the research of (Lamping and Rao, 1996; Baudisch et al., 2001; Kosara et al., 2002) and falls into specific expertise of the group of D. Schmalstieg. 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 quantitative methods for the evaluation of the glyph generation process, similar to the approaches described in (Tory, 2004) and (Plaisant, 2005). For the special case of the evaluation of DNA and tissue micro-array analysis we build on the work of (Saraiya, 2005) and for the evaluation of high level tasks, e.g. perceiving relationships or making conclusions (Amar, 2005) can be a starting point.

2 From Data to Visual Information

Clinical metadata consist of a broad range of entities, e.g. numerical values of laboratory parameters, textual annotation, disease categories (staging and grading of tumours), images, hierarchical data (e.g. family history) etc., covering both qualitative and quantitative data.

We developed an architecture for generic data analysis coupled with specific visualization methods, see Fig- 1. The architecture was developed due to the necessity of dealing with several hundreds of experiments each with several thousand gene expression values and a rich set of clinical data at one time. First experiments were done with a breast cancer gene expression data set (150 experiments, each generating expression data from 36.000 genes).

For the experiments we developed a framework for linked views and prototypical visualization modules for large scale microarray data. In the pre-processing phase the following steps are carried out:

1. Data Import and attribute tagging
2. Value transformations
3. Creation of a hierarchical structure
4. Mapping to graphical attributes.

A very important task in the data pre-processing is the annotation of the data space describing 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 for the semi-automatic generation of medical glyphs.

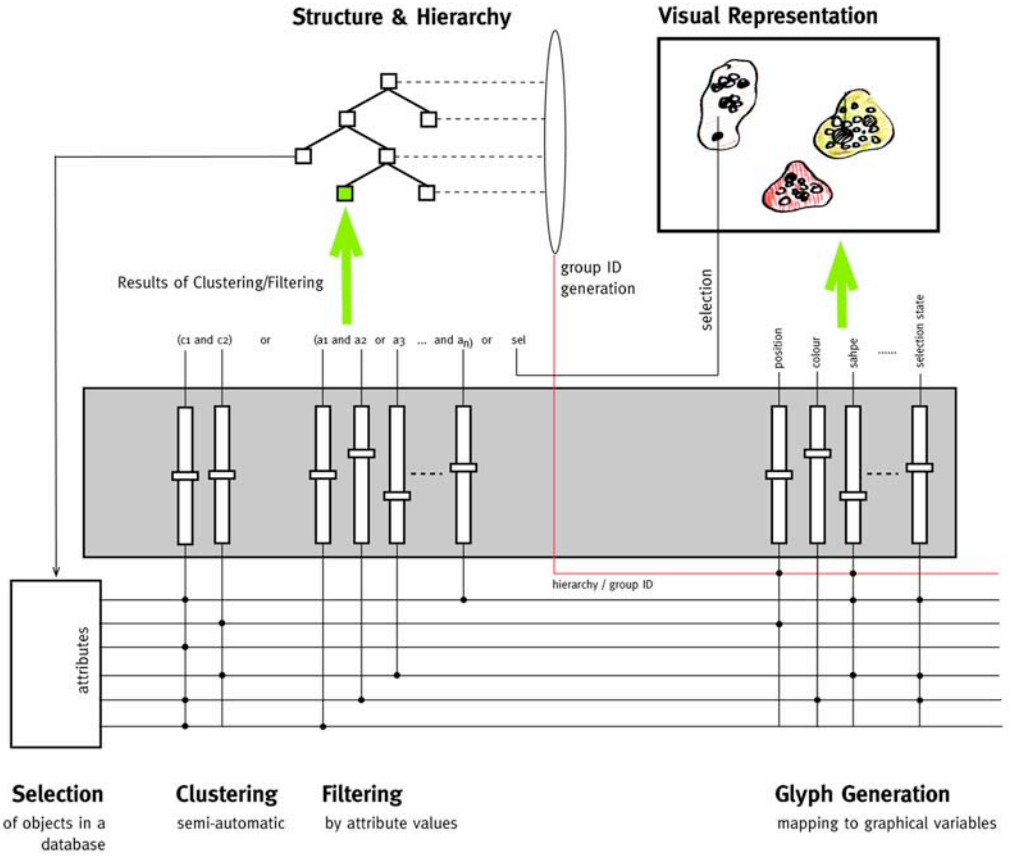

Fig. 1 - Interactive structure generation

In the data pre-processing step we identify *(hidden) structures* in the multidimensional data set. Therefore we describe a structure by the introduction of groups, e.g. all patients with a certain disease, and a membership function, which defines an affiliation of an object to a specific group. An object is characterized by a data vector, which is either a row in the multidimensional data set (e.g. all attributes of a patient record) or a column (the distribution of data values of a specific attribute). A structure can be *flat* - only objects belong to a group - or *hierarchical*, in this case groups can also contain other groups.

The creation of a hierarchical structure can be done either

- **automatically**, by a classification (clustering) algorithm,
- **semi-automatically** (machine aided) by an algorithm, which is controlled by user input in a closed feedback cycle, or
- through an **interactive** definition of object affiliations.

3 Graphical Variables

The final step in data pre-processing is the generation of data values, which can be directly mapped to graphical variables in glyph generation. Jacques Bertin systematically classified in his Semiology of Graphics (1983) the use of visual elements to display data and relationships. Bertin defined seven basic visual variables: shape, color, value, texture, position, size and orientation.

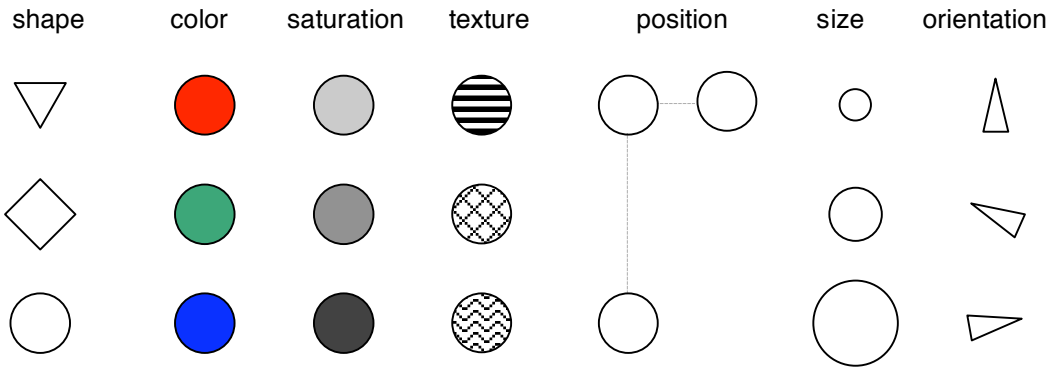

Fig. 2 - Graphical variables (Bertin)

Graphical variables combined with a visual semantics allow the linking of data attributes to glyphs representing an object. In visual communication graphical variables can encode different meanings. It is therefore essential that graphical variables match the data characteristics of the original data values. The choice of the graphical variable effects how the user will be able to perceive and recognise structures present in the data values. Graphical variables can be grouped according to the scale of measurement and the length of the scale they can visualize in an effective way. The following list summarizes the properties of graphical variables:

- Shape

  The shape of an 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”).

- Colour

  The perceptual variable colour (hue) is a nominal variable, even though the wavelength of light assigns an ordering to colours, 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-colour deficient person can distinguish between seven and ten million different colours; however, colour a deeply subjective attribute, therefore not more than 10 to 20, carefully chosen color values should be used in glyph generation.

- Saturation and Texture

  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 variable, i.e. roughly 10 values can be distinguished in an effective way.

- Position

  The position variable can be mapped to ordinal values, and is a very finely grained (long) variable.

- Size

  The size variable can be also be mapped to ordinal values, however it is “shorter” than the position variable. Please note, that quantitative differences in one dimension (i.e., length) are perceived better than in two dimensions (area).

- Orientation

  Orientation can be mapped to an ordinal data value, but is a very short graphical variable, i.e. only very few of different orientations can be perceived.

Meta-data is stored for all data types and value distributions. Therefore the system can propose a *graphical variable* based on this information, or for example raise a warning message when the user maps an ordinal value to a shape property in the setup of the final data transformation.

4 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, and attribute glyphs visualize the value distributions of a single parameter for either all objects or a subset of objects. Fig. 3 shows a sketch of a object glyph summarizing up to 10 clinical parameters and an attribute glyph for the distribution of a tri-state attribute.

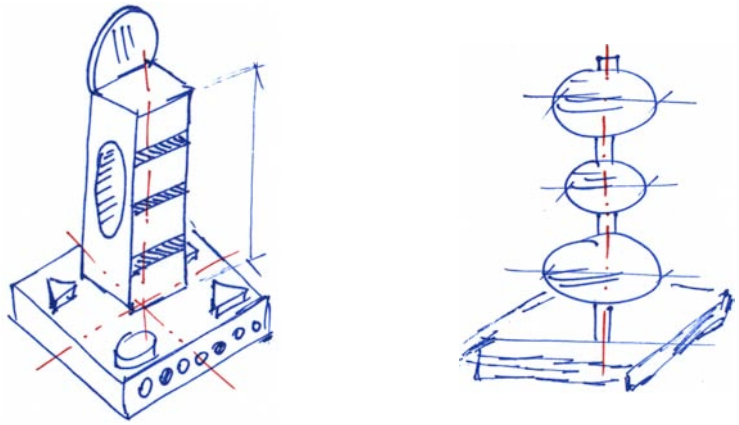

Fig. 3 - Object- and attribute glyphs

We decided to use 3D glyphs in order to provide on the one hand the necessary information density for a wide range of parameters, and on the other hand to differentiate between primary and secondary information (visualization hierarchies). For the 3D projection, we choosen an isometric view (no perspective distortion) because glyphs can be compared independent of their spatial position. The user interaction is done in a simpler way.

Both types of glyphs 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. All glyphs can be manipulated either manually or semi-automatically to define subgroups and hierarchies (clusters). They are linked, e.g. when selecting a subset in the object glyphs the attribute glyphs are updated immediately and when selecting a subset of attributes the visual appearance of all object glyphs is further on only determined by the selected attributes.

The left image in Fig. 4 shows numerical and textual paramaters of a tissue sample (tumor staging, sample age), and the localisation and on the right side shows a visual summary by means of an object glyph.

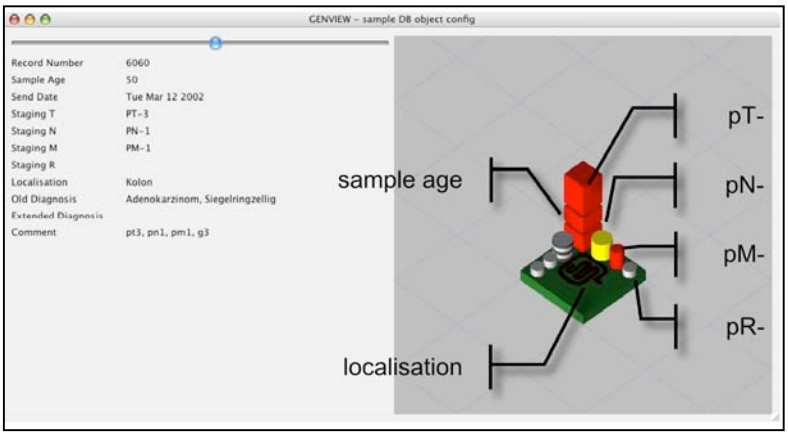

Fig. 4 Object glyph of a tissue sample

5 Application Examples

In the presentation layer the user can interact with multidimensional data using different visualization views. The user can determine which parameter of the input data (or some intermediate step in the pre-processing phase) is mapped to graphic variables (shape, color, value, texture, position, size and orientation). If the position is a result of a clustering algorithm or indirectly given by the object hierarchy, the object hierarchy can be manipulated interactively. After a hierarchical structure is determined, sub-groups of objects can be used as input for further (linked) views such as heatmaps, scatterplots or parallel coordinates.

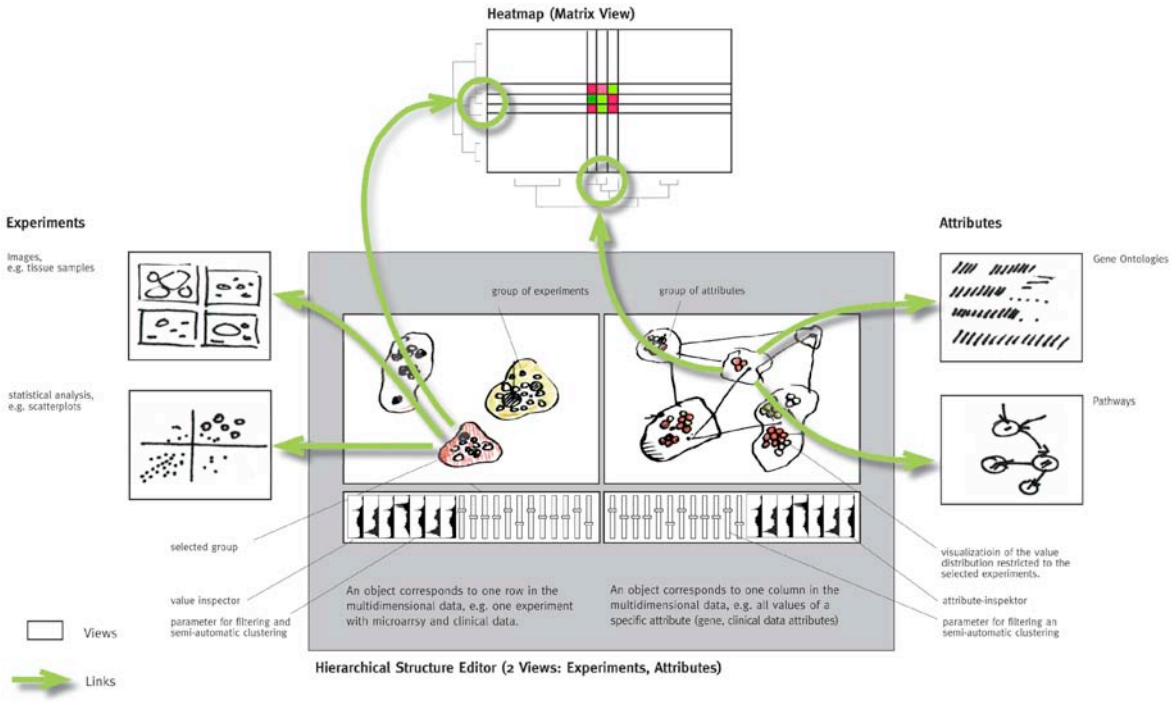

Fig. 5 Linked Views

Heatmaps, introduced by Michael Eisen in 1998, are the most common visualization techniques in microarray analysis. Heatmaps can be seen as a special case of Bertin Matrices which were introduced 1977 as a display and an analysis strategy for multivariate data. Today's data processing potentials and computer graphics power open new application areas. A Bertin matrix is a matrix of glyphs, which allows rearrangements of the row and column permutations to a more homogeneous structure. Bertin matrices were designed for interactive use with an automatic initial pre-processing of the data, sorting and arranging according to some specified clustering criterion. However, aspects as the appropriate formulation of hypothesis and usage of lateral information that need direct interaction.

Fig. 6 shows attribute glyphs showing the distribution of gene expression values for a big number of tissue samples. Each glyph represents a specific gene, and the semitransparent red boxes depict a cluster of genes.

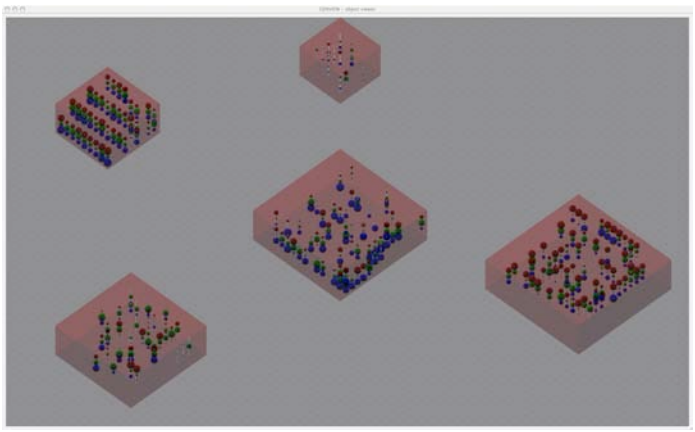

Fig. 6 Attribute gylphs with automatic clustering

Fig. 7 shows simple object glyphs representing the tumor staging of over 10,000 patients through colour coding and a special spatial grammar, visualizing the improvements in the diagnosis and changes in the age distributions of patients.

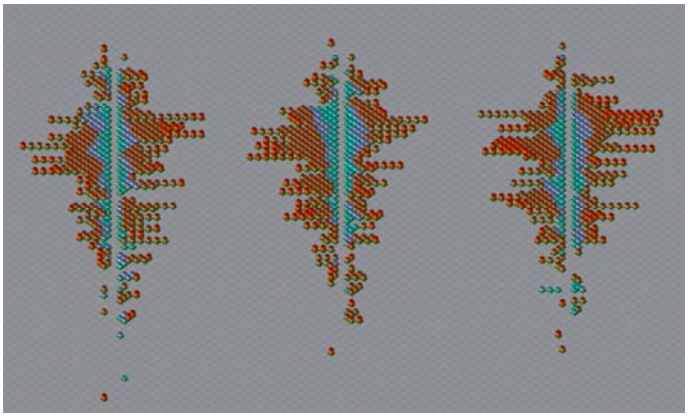

Fig. 7 -Object glyphs arranged by a spatial grammar

6 Conclusion

The analysis of huge inhomogeneous datasets is one of the main challenges in personalised medicine, which aims at more specific diagnosis and treatment of diseases. In personalized medicine the diseases of individual patients are characterized on the basis of several parameters including molecular data (e.g. genetic polymorphisms, gene expression or

proteomics data) as well as a broad spectrum of medical data (e.g. laboratory parameters, clinical phenotypes or pathological alterations). 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 gene-expression data. With the help of large scale displays and multi-layer visualizations experts can perform a hypothesis-driven 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. In our future work we plan to focus on automatic glyph generation, animated glyphs and the definition of spatial grammars (arrangements, levels of details, blending of visual stimuli).

7 Acknowledgements

Our thanks are due to all partners of the Project GenView, especially Kurt Zatloukal, Dieter Schmalstieg, Martin Asslaber and Karoline Pickel for their contributions and critical reviews and various discussions.

8 References

Amar, R, Stasko J. (2005). Knowledge precepts for design and evaluation of information visualizations, IEEE Transactions on Visualization and Computer Graphics, Vol. 11. No. 4, 2005.

Baudisch, P., Good N., and Stewart P., (2001) Focus Plus Context Screens: Combining Display Technology with Visualization Techniques. In Proceedings of UIST 2001, Orlando, FL.

Beesley, J., Roush, C., and Baker, L. (2004) High-throughput molecular pathology in human tissues as a method for driving drug discovery. DDT, 9, pp. 182-189.

Bouchie, A. (2004) Coming soon: a global grid for cancer research. Nature Biotechnology, 22.

Buckhaults, P. (2006) Gene expression determinants of clinical outcome. Curr. Opin. Oncol. ,18 (1), pp. 57-61.

Craig P., Kennedy J., Cumming A. (2005). Animated interval scatter-plot views for the exploratory analysis of large-scale micro-array time-course data in Information Visualization, Vol. 4, No. 3.

Dastani M. (2002) The role of visual perception in data visualization, Journal of Visual Languages and Computing vol. 13(6).

Gehlenborg N., Dietzsch J., Nieselt K, A. (2005). Framework for Visualization of Micro-array Data and Integrated Meta Information, in Information Visualization, Vol. 4, No. 3. (2005)

Grinstein G., Meneses C. (2002) Visual Data Exploration in Massive Data Sets, in Information Visualization in Data Mining and Knowledge Discovery, Morgan-Kaufmann Publishers, 2001.

Grinstein G. (2003) Integrated, Tightly-Coupled, High-Dimensional Analysis and Visualization for Microarray Expression Data in CHI's Data Visualization and Interpretation Conference Proceedings, 2003.

Hege H., Hutanu A., Kähler R., Merzky A., Radke T., Seidel E., Ullmer B. (2003) Progressive retrieval and hierarchical visualization of large remote data, Proceedings of the Workshop on Adaptive Grid Middleware.

Kaiser, J. (2002) Population databases boom, from Iceland to the U.S. Science, 298, pp. 1158-1161.

Keim D., Kirgel H.P. (1994) VisDB: Database Exploration Using Multidimensional Visualization,

IEEE Computer Graphics and Applications, Volume 14 , Issue 5 (September 1994)

Kincaid R., Ben-Dor A., Yakhini Z. (2005). Exploratory visualization of array-based comparative genomic hybridization in Information Visualization, Vol. 4, No. 3.

Kosara R., Miksch S., Hauser H. (2002) Focus and Context Taken Literally, IEEE Computer Graphics and its Applications, Special Issue: Information Visualization, pp. 22-29, 22(1), Jan.-Feb., 2002.

Lamping, J. and Rao, R. (1996). The hyperbolic browser: A focus+context technique for visualizing large hierarchies. Journal of Visual Languages and Computing, 7(1):33–35.

Nevins, J.R., Huang, E.S., Dressman, H., Pittman, J., Huang, A.T., and West, M. (2003) Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. Hum. Mol. Genet., 12, Rev. Issue 2, R153-R157.

Pantazi S.V., Arocha J. F. Moehr J. R. (2004). Case-based medicine informatics, BMC Medical Informatics and Decision Making 2004, 4:19.

Plaisant C. (2004). The challenge of information visualization evaluation, Proceedings of the working conference on Advanced visual interfaces, 2004.

Sadee, W., Dai, Z. (2005) Pharmacogenetics/genomics and personalized medicine. Hum. Mol. Genet., 14 Spec No. 2, R207-214.

Saraiya P., North C., Duca K. (2005) An insight-based methodology for evaluating bioinformatics visualizations, IEEE Transactions on Visualization and Computer Graphics, Vol. 11. No. 4, 2005.

Santos S, Brodlie K. (2004). Gaining understanding of multivariate and multidimensional data through visualization, Computers and Graphics, Vol 28, Issue 3, June 2004.

Sauter, G., Simon, R., and Hillan, K. (2003) Tissue microarrays in drug discovery. Nature Reviews, 2, pp. 962-972.

Smrtic MB., Grinstein G. (2005) Interactive Visualization of Microarray Data on Pathways, Proceedings of the 2005 BioIT Conference, Boston, MA.

Thomas J., Cook K. A., (eds.) (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics, J. J. , IEEE CS Press, 2005.

Tolman K.G., Fonseca V., Tan MH, Dalpiaz A. (2004) Narrative review: hepatobiliary disease in type 2 diabetes mellitus. Ann Intern Med. 141:946-956.

Tory M., Möller T. (2004) Human factors in visualization research, IEEE Transactions on Visualization and Computer Graphics, Vol. 10. No. 1, 2004.