Visual Data Mining: Effective Exploration of the Biological Universe

David Otasek[1], Chiara Pastrello[1], Andreas Holzinger[2], and Igor Jurisica[1,3,*]

[1] Princess Margaret Cancer Centre, University Health Network, IBM Life Sciences Discovery Centre, and TECHNA for the Advancement of Technology for Health, TMDT Room 11-314, 101 College Street, Toronto, ON M5G 1L7, Canada
juris@ai.utoronto.ca
[2] Medical University Graz, Institute for Medical Informatics, Statistics and Documentation Research Unit HCI, IBM Watson Think Group, Auenbruggerplatz 2/V, A-8036 Graz, Austria
a.holzinger@hci4all.at
[3] Departments of Medical Biophysics and Computer Science, University of Toronto

Abstract. Visual Data Mining (VDM) is supported by interactive and scalable network visualization and analysis, which in turn enables effective exploration and communication of ideas within multiple biological and biomedical fields. Large networks, such as the protein interactome or transcriptional regulatory networks, contain hundreds of thousands of objects and millions of relationships. These networks are continuously evolving as new knowledge becomes available, and their content is richly annotated and can be presented in many different ways. Attempting to discover knowledge and new theories within this complex data sets can involve many workflows, such as accurately representing many formats of source data, merging heterogeneous and distributed data sources, complex database searching, integrating results from multiple computational and mathematical analyses, and effectively visualizing properties and results. Our experience with biology researchers has required us to address their needs and requirements in the design and development of a scalable and interactive network visualization and analysis platform, NAViGaTOR, now in its third major release.

Keywords: Visual Data Mining, Interactive Data Mining, Knowledge Discovery, Scalable Network Visualization, Biological Graphs, Networks.

1 Introduction and Motivation

1.1 The Need for Visual Data Mining

One of the grand challenges in our "networked 21st century" is in dealing with large, complex, and often weakly structured data sets [1], [2], and in big volumes of unstructured information [3].

This "big data" challenge is most evident in the biomedical domain [4]: the emergence of new biotechnologies that can measure many molecular species at once,

* Corresponding author.

A. Holzinger, I. Jurisica (Eds.): Knowledge Discovery and Data Mining, LNCS 8401, pp. 19–33, 2014.

large scale sequencing, high-throughput facilities and individual laboratories worldwide produce vast amounts of data sets including nucleotide and protein sequences, protein crystal structures, gene-expression measurements, protein and genetic interactions, phenotype studies etc. [5]. The increasing trend towards personalized medicine brings together data from very different sources [6].

The problem is that these data sets are characterized by heterogeneous and diverse features. Individual data collectors prefer their own different schema or protocols for data recording, and the diverse nature of the applications used results in various data representations. For example, patient information may include simple demographic information such as gender, age, disease history, and so on as non-standardized text [7]; results of X-ray examination and CT/MR scan as image or video data, and genomic or proteomic-related tests could include microarray expression data, DNA sequence, or identified mutations or peptides. In this context, heterogeneous features refer to the varied ways in which similar features can be represented. Diverse features refer to the variety of features involved in each distinct observation. Consider that different organizations (or health practitioners) have their own schemata representing each patient. Data heterogeneity and diverse dimensionality issues then become major challenges if we are trying to enable data aggregation by combining data from all sources [8], [9].

This increasingly large amount of data requires not only new, but efficient and most of all end-user friendly solutions for handling it, which poses a number of challenges [10]. With the growing expectations of end-users, traditional approaches for data interpretation often cannot keep pace with demand, so there is the risk of modelling artefacts or delivering unsatisfactory results. Consequently, to cope with this flood of data, *interactive* data mining approaches are vital. However, exploration of large data sets is a difficult problem and techniques from interactive visualization and visual analytics may help to assist the knowledge discovery process generally and data mining in particular [11], [12], leading to the approach of Visual Data Mining (VDM).

1.2 A Short History of Visual Data Mining

One of the first VDM approaches was in a telecommunications application. This application involved a graph-based representation and a user interface to manipulate this representation in search of unusual calling patterns. This approach proved extremely effective for fraud detection [13].

A further work by Alfred Inselberg (1998) [14] proposed the use of parallel coordinates for VDM, which transforms the search for relations into a 2-D pattern recognition problem. Parallel coordinates are a splendid idea for visualizing multi-dimensional geometry [15]; a good overview on parallel coordinates can be found in [16], however, to date they are still rarely used in biomedical applications.

The field of VDM started to expand to diverse domains, as highlighted in a special issue in issue 5 of the 1999 volume of IEEE Computer Graphics and Applications [17] including a work on visual mining of high-dimensional data [18]. A state-of-the

art analysis was provided by Keim et al. at the EUROGRAPHICS 2002 [19]. A good overview of VDM can be found in [20]. A recent overview on VDM for knowledge discovery, with a focus on the chemical process industry can be found in [21] and a recent work on VDM of biological networks is [12]. A very recent medical example for interactive pattern visualization in *n*-dimensional data sets by application of supervised self- organizing maps is [22]. A general overview on the integration of computational tools in visualization for interactive analysis of heterogeneous data in biomedical informatics can be found in [23].

1.3 Interactivity and Decision Support

For data mining to be effective, it is important to include the human expert in the data exploration process, and combine the flexibility, creativity, and general knowledge of the human with the enormous computational capacity and analytical power of novel algorithms and systems. VDM integrates the human in the data exploration process; it aims to effectively represent data visually to benefit from human perceptual abilities, allowing the expert to get insight into the data by direct interaction with the data. VDM can be particularly helpful when little is known about the data and the exploration goals are ill-defined or evolve over time. The VDM process can be seen as a hypothesis generation process: the visualizations of the data enable the user to gain insight into the data, and generate new hypotheses to support data mining and interpretation [24], [25].

VDM often provides better results, especially in cases where automatic algorithms fail [11]. However, it is indispensable to combine interactive VDM with automatic exploration techniques; hence we need machine learning approaches due to the complexity and the largeness of data, which humans alone cannot systematically and comprehensively explore. Consequently, a central goal is to work towards enabling effective human control over powerful machine intelligence by the integration of both machine learning methods and manual VDM to enable human insight and decision support [26], the latter is still the core discipline in biomedical informatics [27].

2 Glossary and Key Terms

Biological Pathway Exchange (BioPAX): is a RDF/OWL-based language to represent biological pathways at the molecular and cellular level to facilitate the exchange of pathway data. It makes explicit use of relations between concepts and is defined as an ontology of concepts with attributes [28].

CellML: is an open standard XML, for describing mathematical models, originally created out of the Physiome Project, and hence used primarily to describe models relevant to the field of biology [29, 30].

Graph dRawing with Intelligent Placement (GRIP): is based on the algorithm of Gajer, Goodrich & Kobourov [31] and written in C++ and OpenGL, and uses an adaptive Tcl/Tk interface. Given an abstract graph, GRIP produces drawings in 2D

and 3D either directly or by projecting higher dimensional drawings into 2D or 3D space [32].

KEGG Markup Language (KGML): is an exchange format of the KEGG pathway maps, which is converted from the KGML+ (KGML+SVG) format. KGML enables automatic drawing of KEGG pathways and provides facilities for computational analysis and modeling of gene/protein networks and chemical networks [33].

Proteomics Standards Initiative Molecular Interaction XML format (PSI MI): was developed by the Proteomics Standards Initiative (PSI) as part of the Human Proteome Organization (HUPO) [34]. PSI-MI is the standard for protein–protein interaction (PPI), intended as a data exchange format for molecular interactions, not a database structure [35].

Protein-Protein Interactions (PPIs): are fundamental for many biological functions [36], [37]. Being able to visualize the structure of a protein and analyze its shape is of great importance in biomedicine: Looking at the protein structure means to locate amino acids, visualize specific regions of the protein, visualize secondary structure elements, determine residues in the score or solvent accessible residues on the surface of the protein, determine binding sites, etc. [38], [39].

Systems Biology Markup Language (SBML): is a language intended as future standard for information exchange in computational biology and especially within molecular pathways. The aim of SBML is to model biochemical reaction networks, including cell signaling, metabolic pathways and gene regulation [40].

Visual Data Mining (VDM): is an approach for exploring large data sets by combining traditional data mining methods with advanced visual analytics methods and can be seen as a hypothesis generation process [14], [11], [41].

3 Representing Biological Graphs

3.1 A Constantly Changing Understanding

The functions of life on a sub-cellular level rely on multiple interactions between different types of molecules. Proteins, genes, metabolites, all interact to produce either healthy or diseased cellular processes. Our understanding of this network of interactions, and the interacting objects themselves, is continuously changing; and the network itself is evolving as we age or as disease progresses. Our methods for discovering new relationships and pathways change as well.

NAViGaTOR 3 addresses these realities by having a very basic core rooted in graph theory, with the flexibility of a modular plugin architecture that provides data input and output, analysis, layout and visualization capabilities. NAViGaTOR 3 implements this architecture by following the OSGi standard (http://www.osgi.org/Main/HomePage). Available API enables developers to expand standard distribution by integrating new features and extending the functionality of the program to suit their specific needs.

3.2 Data Formats

The ability to share data effectively and efficiently is the starting point for successful analysis, and thus several attempts have been made to standardize formats for such data exchange: PSI-MI [35], BioPAX [42], KGML, SBML [40], GML, CML, and CellML [30].

Each of these formats has a different focus and thus uniquely affects the way a particular network can be described. Some formats, like PSI, focus on describing binary interactions. Others, such as BioPAX, can describe more complex topology, allowing for many-to-many interactions and concepts such as meta-graphs. However, the majority of biological data remains represented in tabular format, which can vary wildly in content and descriptiveness.

NAViGaTOR 3 was designed with the knowledge that a researcher may need to combine heterogeneous and distributed data sources. The standard distribution supports the loading, manipulation, and storage of multiple XML formats and tabular data. XML data is handled using a suite of file loaders, including XGMML, PSI-MI, SBML, KGML, and BioPAX, which store richly-annotated data and provide links to corresponding objects in the graph. Tabular data is stored using DEX [43], a dedicated graph database from Sparsity Technologies (http://www.sparsity-technologies.com/dex).

3.3 Biological Scale

A sense of the scale biologists might contend with in attempting to model protein behavior can be seen in UniProt (http://www.uniprot.org), a database that documents protein sequences. In its 2013_10 release, UniProt contained 20,277 sequences for human proteins, while I2D (http://ophid.utoronto.ca/i2d) [44], a database that includes interactions between these proteins, in its 2.3 version contains 241,305 experimental or predicted interactions among 18,078 human proteins. If the protein interaction network is integrated with other data of similar size, such as transcriptome regulatory network, microRNA:gene regulation network, or drug:protein target network, the visualization can become challenging, not only because of the size of the graph, but due to rich annotation and underlying topology of these ‘interactomes’.

Often, even the best case layouts produce a gigantic ‘hairball’ in which a user is unable to trace paths between different objects in the network. It is important to keep in mind that such a network is still limited in scope; it doesn’t take into account genetic, metabolite or drug interactions. In a true ‘systems biology’ view, we need to integrate multiple layers of these individual networks into a larger, comprehensive, typed graph. Tools that attempt to analyse this data must take this scale into account. To be useful, visualization tools, and particularly interactive visualization tools must effectively handle networks of this size. In the case of NAViGaTOR, DEX can handle networks of up to 1 Billion objects. Visualizing networks is handled through JOGL (http://jogamp.org/jogl/www/), a library that speeds up large network rendering by taking advantage of the acceleration provided by GPU hardware whenever it is available.

4 Visualization, Layout and Analysis

Exploring data is often a combination of analysis, layout and visualization. We have found that being able to utilize and combine all three of these aspects quickly and efficiently simplifies and in turn enables effective research.

A central idea to NAViGaTOR 3's architecture is providing multiple views of the data. While the structure of the network and its annotations remains constant, NAViGaTOR 3 allows the user to view and manipulate it as a spreadsheet of nodes or edges, a matrix, or an OpenGL rendered graph. Individual views allow a user to make a selection of objects, which can then be transferred to another view. For example, a user can select the top 10 rows of a spreadsheet that sorts the network's nodes by a measurement such as gene expression, and then transfer that selection to the OpenGL view, allowing them to see those nodes in the context of their neighbors.

The most basic level of analysis supports access, search and data organization. The tabular data associated with a network can be searched using DEX queries, allowing for basic numeric searches (equals, greater than, less than, etc.) and text (exact match, regular expression, etc.). The spreadsheet view supports effective searching, data sorting and selecting. XML data benefits from rich data annotation, and can be searched using XPath, a dedicated XML query language.

XPath is extremely versatile, mixing logical, numeric and text queries in a single language. It also handles translation of XML data into tabular data.

Network structure provides additional insights, as it relates to the function of proteins that form it [45], [46]. Examining the network structure can range from searches for node neighbors and nodes of high degree to more mathematically complex operations such as all pairs shortest path calculations, flow analysis, or graphlets [47].

A key part of NAViGaTOR's tool set is the subset system, which enables the storage of selections from various graph views. Once they are stored, they can be manipulated with set arithmetic operations (union, difference, intersection). This allows the user to intelligently combine the results of searches and selections from other views.

Further strengthening the link between visualization and data is the concept of filters. Filters are visualization plugins for the OpenGL view that allow a user to map node or edge feature to a visual attribute, i.e., enabling interactive exploration of typed graphs by seamlessly combining analysis and human insight. For example, a confidence value for an edge can be translated into its width, color, shape or transparency. Similarly, node height, width, color, transparency, outline color, and outline size can be used to visualize gene, protein or drug characteristics and measurements. Thus, layout and presentation of rich, annotated networks can be easily modified, enabling new insight into complex data.

Graph layout is essential for effective visual analysis of complex graphs. NAViGaTOR 3 uses a combination of manual and automated layout tools. The standard distribution includes several versions of the GRIP (Graph dRawing with Intelligent Placement) [48], [49], [50] layout algorithm, which enables fast layouts of tens of thousands of nodes and edges. For example, visualizing protein interaction network topology changes in the presence or absence of specific receptors [51].

Besides GRIP, the user also has a selection of circular, arc and linear layouts, as well as moving, scaling and rotating tools to manually place nodes in a desired topology.

Usually, combinations of these layouts are necessary to effectively visualize the network [12]. Large and complex biological networks, even with the benefit of the GRIP layout, are usually tangled and poorly interpretable graphs. Analyzing network topology (hubs, graphlets, cliques, shortest path, flow, etc.) provides rich topological features that may aid in discovery and visualization of important insights. Researchers may have to map a data attribute to transparency to make areas of interest visible, or use an overlap of searches to color a selection of nodes or edges. Being able to use different analysis and layout methods combined with user's insight provides the flexibility to highlight desired results in the network, or discover novel insights and form hypotheses. Thus, NAViGaTOR extends the basic concept of network visualization to visual data mining.

To demonstrate the versatility of NAViGaTOR 3 we created an integrated network by combining metabolic pathways, protein-protein interactions, and drug-target data. We first built a network using metabolic data collected and curated in our lab, combining several steroid hormone metabolism pathways: androgen, glutathione, N-nitrosamine and benzo(a)pyrene pathway, the ornithine-spermine biosynthesis pathway, the retinol metabolism pathway and the TCA cycle aerobic respiration pathway. The reactions in the dataset are in the following format: metabolite A → enzyme → metabolite B. As shown in Figure 1, the different pathways are integrated and highlighted with different edge colours. The edge directionality highlights reactions and flow between the individual pathways.

As the dataset is centred on steroid hormone metabolism, we decided to include data from hormone-related cancers [52]. In particular, we retrieved the list of FDA-approved drugs used for breast, ovarian and prostate cancer from the National Cancer Institute website (http://www.cancer.gov/). We then searched in the DrugBank (http://www.drugbank.ca [53]) for targets for each drug and integrated them in the network.

Three targets are enzymes that are part of the original network (HSD11B1, CYP19A1, CYP17A1). Polymorphisms in CYP19 have been associated with increased risk of breast cancer [54], while polymorphisms in CYP17 have been linked to increased risk of prostate cancer [55].

CYP17 inhibitors are considered key drugs for castration resistant prostate tumours, due to their ability to block the signaling of androgen receptors even when the receptor expression is increased [56].

Thanks to the ability of NAViGaTOR to include various types of nodes, we can also see how frequently DNA is a target. In fact, many of the drugs used for breast and ovarian cancer treatment are DNA intercalating agents [57].

To further investigate whether drug targets are directly connected to the metabolic network we searched for direct interactions between the two types of nodes using protein interactions from I2D and identified three such targets (TUBA1, TOP1 and EGFR).

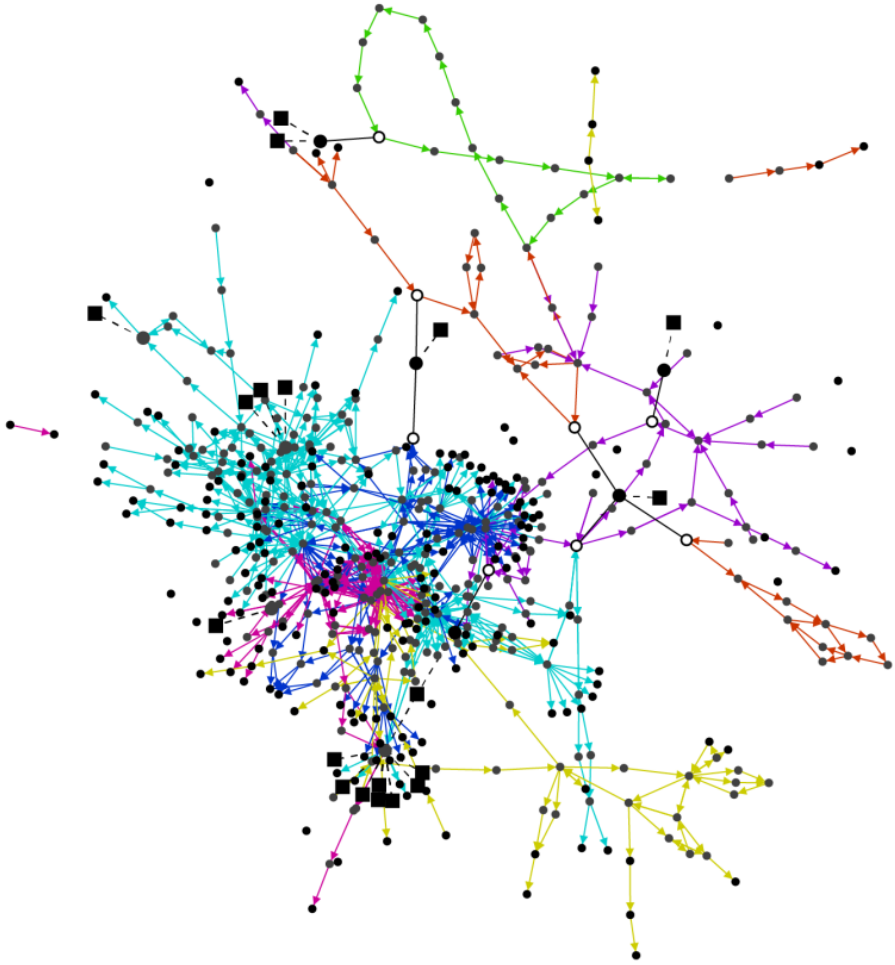

Fig. 1. Partially explored network – connecting drugs and metabolism. A network comprising metabolites, enzymes, and drugs in the early stages of exploration, colored according to the pathway (see complete overview in Figure 2).

EGFR overexpression appears in breast cancer, especially in triple-negative and in inflammatory breast cancer, and is associated with large tumor size, poor differentiation, and poor clinical outcomes [58]. EGFR inhibitor treatments (e.g., Erlotinib or Cetuximab) have been suggested for triple-negative breast cancer patients, and a few clinical trials showed promising results [59].

It would be interesting to study the effect of EGFR mutations in this network, to evaluate if they can have an effect on the patient's response to inhibitors similar to response to Erlotinib in non-small-cell-lung cancer patients [60].

Interestingly, several CYP and UGT proteins are key connectors of different metabolic pathways (highlighted in green in Figure 2), and have a biologically important role in the network. Both families of proteins have important roles in metabolic pathways (CYP450 are ubiquitously expressed in the body as they catalyze the fundamental carbon–oxidation reaction used for unnumbered metabolic reactions, while UGTs are used in reactions that form lipophilic glucuronides from a high variety of non–membrane-associated substrates, either endogenous or xenobiotics and has evolved as a highly specialized function in higher organisms) but they have mainly been associated with drug metabolism, in their wild-type or polymorphic forms [61], [62], [63].

This example shows only one of the several possible integrated networks that can be built using NAViGaTOR 3, and highlights the role of the analysis of the network structure in pointing out major biological players.

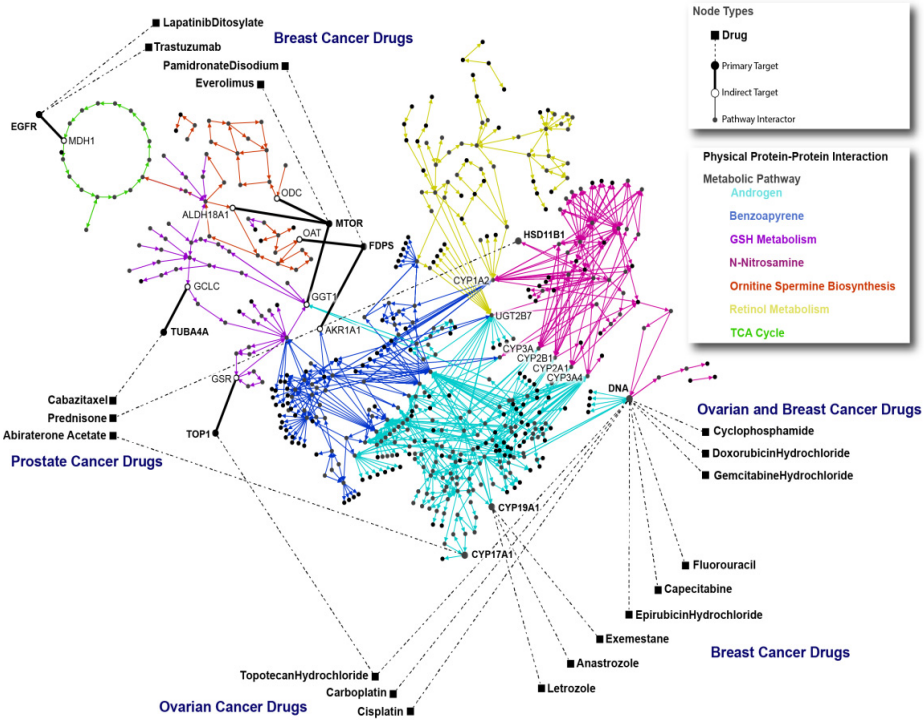

Fig. 2. Completed network – the same network as in Figure 1 with drugs and biologically relevant components emphasized

Biological networks will continue becoming larger and more complex thanks to the higher throughput of novel technologies and increased data integration. This highlights the need for tools that scale up to large and complex networks. Moreover, and maybe more importantly, this highlights the necessity for tools with the ability to integrate different -omics data collections, to discover cross-talk and to build an

increasingly more complete representation of the real cell or an organism. NAViGaTOR fits perfectly in this context and provides the researcher with the functionality needed to advance data discovery at the same speed of high-throughput data production.

5 Open Problems and Future Work

The deployment of VDM techniques in commercial products remains sparse – and in today's traditional hospital information systems such approaches are completely missing. Future work must involve the tight integration of sophisticated interactive visualization techniques with traditional techniques from machine learning with the aim to combine fast automatic data mining algorithms with the intuitive power and creativity of the human mind [64]. A further essential aspect at the clinical workplace is to improve both the quality and speed of the VDM process. VDM techniques also need to be tightly integrated with available systems used to manage the vast amounts of relational, semi-structured and unstructured information such as the typical patient records [3] and omics data [9]. The ultimate goal is to broaden the use of visualization technologies in multiple domains, leading to faster and more intuitive exploration of the increasingly large and complex data sets. This will not only be valuable in an economic sense but will also enhance the power of the end user, i.e. the medical professional.

There are several reasons for slower commercial acceptance of VDM [65], including multi-disciplinarity and the resulting lack of expertise, and resistance to changing system architectures and workflows. While so-called guided data mining methods have been produced for a number of data mining areas including clustering [66], association mining [67] and classification [68], there is an architectural aspect to guided data mining, and to VDM in general, which has not been adequately explored so far, and which represents a rich area for future research.

Another area of future work for the VDM community is quantification. Since VDM methods can be more time-consuming to develop and special expertise is needed for their effective use, successful deployment requires proper metrics that demonstrates time improvement or quality improvement over non-visual methods.

Technological challenges are present in problem solving, decision support and human information discourse; according to Keim et al. (2008) [65], the process of problem solving supported by technology requires the understanding of technology on the one hand, and comprehension of logic, reasoning, and common sense on the other hand. Here the danger lies in the fact that automatic methods often fail to recognize the context, if not explicitly trained.

A grand challenge is to find the most appropriate visualization methodology and/or metaphor to communicate analytical results in an appropriate manner. A recent example on Glyph-based visualizations can be seen in [69], while noting that most often such approaches are limited to a certain domain.

User acceptability, which is also on Keim's 2008 list is an additional grand challenge: many sophisticated visualization techniques have been introduced, but they

are not yet integrated in the clinical workplace, mainly due to end users' refusal to change their routine – this is most apparent in the medical domain [70]; an often ignored aspect in that respect is the previous exposure to technology [71]; in particular elderly end users are not so enthusiastic in adopting new technologies to their daily routine. Consequently, it is very important that advantages of VDM tools are presented and communicated to future users to overcome such usage barriers, taking usability engineering into full account [72].

Faced with unsustainable costs and enormous amounts of under-utilized data, health care needs more efficient practices, research, and tools to harness the full benefits towards the concept of personalized medicine [73].

A major challenge lies in the development of new machine learning methods for knowledge discovery in protein-protein interaction sites, e.g. to study gene regulatory networks and functions. However, when applied to such big data, the computational complexities of these methods become a major drawback. To overcome such limitations Extreme Learning Machines provide a trade-off between computational time and generalization performance [74].

Acknowledgements. We would like to thank Dr. M. Kotlyar and G. Morrison for their help on data retrieval.

References

1. Holzinger, A.: On Knowledge Discovery and Interactive Intelligent Visualization of Biomedical Data - Challenges in Human–Computer Interaction & Biomedical Informatics. In: DATA 2012, Rome, Italy, pp. 9–20. INSTICC (2012)
2. Holzinger, A.: Weakly Structured Data in Health-Informatics: The Challenge for Human-Computer Interaction. In: Proceedings of INTERACT 2011 Workshop: Promoting and Supporting Healthy Living by Design. IFIP, pp. 5–7 (2011)
3. Holzinger, A., Stocker, C., Ofner, B., Prohaska, G., Brabenetz, A., Hofmann-Wellenhof, R.: Combining HCI, Natural Language Processing, and Knowledge Discovery - Potential of IBM Content Analytics as an assistive technology in the biomedical domain. In: Holzinger, A., Pasi, G. (eds.) HCI-KDD 2013. LNCS, vol. 7947, pp. 13–24. Springer, Heidelberg (2013)
4. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
5. Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O., Rhee, S.Y.: Big data: The future of biocuration. Nature 455(7209), 47–50 (2008)
6. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics 15(suppl. 6), I1 (2014)
7. Kreuzthaler, M., Bloice, M.D., Faulstich, L., Simonic, K.M., Holzinger, A.: A Comparison of Different Retrieval Strategies Working on Medical Free Texts. J. Univers. Comput. Sci. 17(7), 1109–1133 (2011)
8. Wu, X.D., Zhu, X.Q., Wu, G.Q., Ding, W.: Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering 26(1), 97–107 (2014)

9. Huppertz, B., Holzinger, A.: Biobanks – A Source of large Biological Data Sets: Open Problems and Future Challenges. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics. LNCS, vol. 8401, pp. 317–330. Springer, Heidelberg (2014)
10. Jeanquartier, F., Holzinger, A.: On Visual Analytics And Evaluation In Cell Physiology: A Case Study. In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 495–502. Springer, Heidelberg (2013)
11. Keim, D.A.: Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics 8(1), 1–8 (2002)
12. Pastrello, C., Otasek, D., Fortney, K., Agapito, G., Cannataro, M., Shirdel, E., Jurisica, I.: Visual Data Mining of Biological Networks: One Size Does Not Fit All. PLoS Computational Biology 9(1), e1002833 (2013)
13. Cox, K., Eick, S., Wills, G., Brachman, R.: Brief Application Description; Visual Data Mining: Recognizing Telephone Calling Fraud. Data Min. Knowl. Discov. 1(2), 225–231 (1997)
14. Inselberg, A.: Visual data mining with parallel coordinates. Computational Statistics 13(1), 47–63 (1998)
15. Inselberg, A., Dimsdale, B.: Parallel coordinates: A tool for visualizing multi-dimensional geometry, pp. 361–378. IEEE Computer Society Press (1990)
16. Heinrich, J., Weiskopf, D.: State of the Art of Parallel Coordinates. In: Eurographics 2013-State of the Art Reports, pp. 95–116. The Eurographics Association (2012)
17. Wong, P.C.: Visual data mining. IEEE Computer Graphics and Applications 19(5), 20–21 (1999)
18. Hinneburg, A., Keim, D.A., Wawryniuk, M.: HD-eye: Visual mining of high-dimensional data. IEEE Computer Graphics and Applications 19(5), 22–31 (1999)
19. Keim, D., Müller, W., Schumann, H.: Information Visualization and Visual Data Mining; State of the art report. In: Eurographics (2002)
20. de Oliveira, M.C.F., Levkowitz, H.: From visual data exploration to visual data mining: A survey. IEEE Transactions on Visualization and Computer Graphics 9(3), 378–394 (2003)
21. Stahl, F., Gabrys, B., Gaber, M.M., Berendsen, M.: An overview of interactive visual data mining techniques for knowledge discovery. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery 3(4), 239–256 (2013)
22. Rosado-Munoz, A., Martinez-Martinez, J.M., Escandell-Montero, P., Soria-Olivas, E.: Visual data mining with self-organising maps for ventricular fibrillation analysis. Computer Methods and Programs in Biomedicine 111(2), 269–279 (2013)
23. Turkay, C., Jeanquartier, F., Holzinger, A., Hauser, H.: On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. LNCS, vol. 8401, pp. 117–140. Springer, Heidelberg (2014)
24. Blandford, A., Attfield, S.: Interacting with Information. Synthesis Lectures on Human-Centered Informatics 3(1), 1–99 (2010)
25. Holzinger, A., Scherer, R., Seeber, M., Wagner, J., Müller-Putz, G.: Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation. In: Böhm, C., Khuri, S., Lhotská, L., Renda, M.E. (eds.) ITBAM 2012. LNCS, vol. 7451, pp. 166–168. Springer, Heidelberg (2012)
26. Holzinger, A.: Interacting with Information: Challenges in Human-Computer Interaction and Information Retrieval (HCI-IR). In: IADIS Multiconference on Computer Science and Information Systems (MCCSIS), Interfaces and Human-Computer Interaction, pp. 13–17. IADIS, Rome (2011)

27. Holzinger, A.: Biomedical Informatics: Computational Sciences meets Life Sciences. BoD, Norderstedt (2012)
28. Strömbäck, L., Lambrix, P.: Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. Bioinformatics 21(24), 4401–4407 (2005)
29. Lloyd, C.M., Halstead, M.D., Nielsen, P.F.: CellML: Its future, present and past. Progress in biophysics and molecular biology 85(2), 433–450 (2004)
30. Miller, A.K., Marsh, J., Reeve, A., Garny, A., Britten, R., Halstead, M., Cooper, J., Nickerson, D.P., Nielsen, P.F.: An overview of the CellML API and its implementation. BMC Bioinformatics 11(1), 178 (2010)
31. Gajer, P., Goodrich, M.T., Kobourov, S.G.: A multi-dimensional approach to force-directed layouts of large graphs. In: Marks, J. (ed.) GD 2000. LNCS, vol. 1984, pp. 211–221. Springer, Heidelberg (2001)
32. Gajer, P., Kobourov, S.G.: GRIP: Graph dRawing with Intelligent Placement. In: Marks, J. (ed.) GD 2000. LNCS, vol. 1984, pp. 222–228. Springer, Heidelberg (2001)
33. http://www.kegg.jp/kegg/xml/
34. Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C.: The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nature Biotechnology 22(2), 177–183 (2004)
35. Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A., Vinod, N., Bader, G., Xenarios, I., Wojcik, J., Sherman, D.: Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biology 5(1), 44 (2007)
36. Jones, S., Thornton, J.M.: Principles of protein-protein interactions. Proceedings of the National Academy of Sciences 93(1), 13–20 (1996)
37. Zhang, A.: Protein Interaction Networks: Computational Analysis. Cambridge University Press, Cambridge (2009)
38. Wiltgen, M., Holzinger, A.: Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: Zara, J., Sloup, J. (eds.) Central European Multimedia and Virtual Reality Conference (available in EG Eurographics Library), pp. 69–74. Czech Technical University (CTU), Prague (2005)
39. Wiltgen, M., Holzinger, A., Tilz, G.P.: Interactive Analysis and Visualization of Macromolecular Interfaces Between Proteins. In: Holzinger, A. (ed.) USAB 2007. LNCS, vol. 4799, pp. 199–212. Springer, Heidelberg (2007)
40. Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A.: The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. Bioinformatics 19(4), 524–531 (2003)
41. Wong, B.L.W., Xu, K., Holzinger, A.: Interactive Visualization for Information Analysis in Medical Diagnosis. In: Holzinger, A., Simonic, K.-M. (eds.) USAB 2011. LNCS, vol. 7058, pp. 109–120. Springer, Heidelberg (2011)
42. Demir, E., Cary, M.P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J.: The BioPAX community standard for pathway data sharing. Nature Biotechnology 28(9), 935–942 (2010)
43. Martinez-Bazan, N., Gomez-Villamor, S., Escale-Claveras, F.: DEX: A high-performance graph database management system. In: IEEE 27th International Conference on Data Engineering (ICDEW), pp. 124–127 (2011)
44. Brown, K.R., Jurisica, I.: Online predicted human interaction database. Bioinformatics 21(9), 2076–2082 (2005)

45. Pržulj, N., Wigle, D.A., Jurisica, I.: Functional topology in a network of protein interactions. Bioinformatics 20(3), 340–348 (2004)
46. Ghersi, D., Singh, M.: Disentangling function from topology to infer the network properties of disease genes. BMC Systems Biology 7(1), 1–12 (2013)
47. Memišević, V., Pržulj, N.: C-GRAAL: Common-neighbors-based global GRAph ALignment of biological networks. Integrative Biology 4(7), 734–743 (2012)
48. Gajer, P., Kobourov, S.G.: GRIP: Graph drawing with intelligent placement. J. Graph Algorithms Appl. 6(3), 203–224 (2002)
49. Gajer, P., Goodrich, M.T., Kobourov, S.G.: A multi-dimensional approach to force-directed layouts of large graphs. Computational Geometry 29(1), 3–18 (2004)
50. Ma, K.-L., Muelder, C.W.: Large-Scale Graph Visualization and Analytics. Computer 46(7), 39–46 (2013)
51. Lissanu Deribe, Y., Wild, P., Chandrashaker, A., Curak, J., Schmidt, M.H., Kalaidzidis, Y., Milutinovic, N., Kratchmarova, I., Buerkle, L., Fetchko, M.J.: Regulation of epidermal growth factor receptor trafficking by lysine deacetylase HDAC6. Science Signaling 2(102), ra84 (2009)
52. Henderson, B.E., Feigelson, H.S.: Hormonal carcinogenesis. Carcinogenesis 21(3), 427–433 (2000)
53. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V.: DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. Nucleic Acids Research 39(suppl. 1), D1035–D1041 (2011)
54. Ma, X., Qi, X., Chen, C., Lin, H., Xiong, H., Li, Y., Jiang, J.: Association between CYP19 polymorphisms and breast cancer risk: Results from 10,592 cases and 11,720 controls. Breast Cancer Research and Treatment 122(2), 495–501 (2010)
55. Douglas, J.A., Zuhlke, K.A., Beebe-Dimmer, J., Levin, A.M., Gruber, S.B., Wood, D.P., Cooney, K.A.: Identifying susceptibility genes for prostate cancer—a family-based association study of polymorphisms in CYP17, CYP19, CYP11A1, and LH-β. Cancer Epidemiology Biomarkers & Prevention 14(8), 2035–2039 (2005)
56. Reid, A.H., Attard, G., Barrie, E., de Bono, J.S.: CYP17 inhibition as a hormonal strategy for prostate cancer. Nature Clinical Practice Urology 5(11), 610–620 (2008)
57. Brana, M., Cacho, M., Gradillas, A., de Pascual-Teresa, B., Ramos, A.: Intercalators as anticancer drugs. Current Pharmaceutical Design 7(17), 1745–1780 (2001)
58. Masuda, H., Zhang, D., Bartholomeusz, C., Doihara, H., Hortobagyi, G.N., Ueno, N.T.: Role of epidermal growth factor receptor in breast cancer. Breast Cancer Research and Treatment 136(2), 331–345 (2012)
59. Gelmon, K., Dent, R., Mackey, J., Laing, K., McLeod, D., Verma, S.: Targeting triple-negative breast cancer: Optimising therapeutic outcomes. Annals of Oncology 23(9), 2223–2234 (2012)
60. Tsao, M.-S., Sakurada, A., Cutz, J.-C., Zhu, C.-Q., Kamel-Reid, S., Squire, J., Lorimer, I., Zhang, T., Liu, N., Daneshmand, M.: Erlotinib in lung cancer—molecular and clinical predictors of outcome. N. Engl. J. Med. 353(2), 133–144 (2005)
61. Tukey, R.H., Strassburg, C.P.: Human UDP-glucuronosyltransferases: Metabolism, expression, and disease. Annual Review of Pharmacology and Toxicology 40(1), 581–616 (2000)
62. Haining, R.L., Nichols-Haining, M.: Cytochrome P450-catalyzed pathways in human brain: Metabolism meets pharmacology or old drugs with new mechanism of action? Pharmacology & Therapeutics 113(3), 537–545 (2007)

63. Kilford, P.J., Stringer, R., Sohal, B., Houston, J.B., Galetin, A.: Prediction of drug clearance by glucuronidation from in vitro data: Use of combined cytochrome P450 and UDP-glucuronosyltransferase cofactors in alamethicin-activated human liver microsomes. Drug Metabolism and Disposition 37(1), 82–89 (2009)
64. Holzinger, A.: Human–Computer Interaction & Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Cuzzocrea, A., Kittl, C., Simos, D.E., Weippl, E., Xu, L. (eds.) CD-ARES 2013. LNCS, vol. 8127, pp. 319–328. Springer, Heidelberg (2013)
65. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual Analytics: Scope and Challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 76–90. Springer, Heidelberg (2008)
66. Anderson, D., Anderson, E., Lesh, N., Marks, J., Perlin, K., Ratajczak, D., Ryall, K.: Human-guided simple search: Combining information visualization and heuristic search. In: Proceedings of the 1999 Workshop on New Paradigms in Information Visualization and Manipulation in Conjunction with the Eighth ACM Internation Conference on Information and Knowledge Management, pp. 21–25. ACM (1999)
67. Ng, R.T., Lakshmanan, L.V., Han, J., Pang, A.: Exploratory mining and pruning optimizations of constrained associations rules. In: ACM SIGMOD Record, pp. 13–24. ACM (1998)
68. Ankerst, M., Ester, M., Kriegel, H.-P.: Towards an effective cooperation of the user and the computer for classification. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 179–188. ACM (2000)
69. Mueller, H., Reihs, R., Zatloukal, K., Holzinger, A.: Analysis of biomedical data with multilevel glyphs. BMC Bioinformatics 15(suppl. 6), S5 (2014)
70. Holzinger, A., Leitner, H.: Lessons from Real-Life Usability Engineering in Hospital: From Software Usability to Total Workplace Usability. In: Holzinger, A., Weidmann, K.-H. (eds.) Empowering Software Quality: How can Usability Engineering Reach These Goals?, pp. 153–160. Austrian Computer Society, Vienna (2005)
71. Holzinger, A., Searle, G., Wernbacher, M.: The effect of Previous Exposure to Technology (PET) on Acceptance and its importance in Usability Engineering. Universal Access in the Information Society International Journal 10(3), 245–260 (2011)
72. Holzinger, A.: Usability engineering methods for software developers. Communications of the ACM 48(1), 71–74 (2005)
73. Chawla, N.V., Davis, D.A.: Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. J. Gen. Intern. Med. 28, S660–S665 (2013)
74. Wang, D.A., Wang, R., Yan, H.: Fast prediction of protein-protein interaction sites based on Extreme Learning Machines. Neurocomputing 128, 258–266 (2014)