**Dissertation**

# Visualization Support For In Silico Medicine

Fleur Jeanquartier

Graz, 2017

*7060 Institute of Interactive Systems and Data Science, Graz University of Technology*
*Graz University of Technology*


Supervisor/First reviewer: Mag.phil. Mag.rer.nat. Dr.phil. Univ.-Doz. Ing. Andreas Holzinger
Second reviewer: Univ.-Prof. Dr.rer.nat. Tobias Schreck, M.Sc.

*in cooperation with the*
*Research Unit HCI-KDD*
*Institute for Medical Informatics, Statistics and Documentation*
*Medical University Graz, Auenbruggerplatz 2/V, A-8036 Graz, Austria*

# Abstract (English)

Seven selected papers build this cumulative dissertation. The selected publications are included in the form as they have been originally published. The publications showcase the scientific introduction and examination of the topic ”Visualization Support For In Silico Medicine”. In particular, this cumulative thesis sums up works ranging from a case study in cell physiology, a general report of biomedical informatics approaches for visual analysis of biological data, looking into more detail into the visualization of RNA secondary structures, protein-protein-interaction and last but not least focusing on computational approaches to support tumor growth analysis. Finally this work concludes with prospect research strategies for making use of visualization to support ”In Silico Tumor Growth modeling”.

# Abstract (German)

Sieben ausgewählte Publikationen bilden diese kumulative Dissertation. Die Beiträge sind in jenem Format eingebunden, wie sie ursprünglich publiziert wurden. Die Papiere zeigen exemplarisch die wissenschaftliche Aufbereitung zum Thema „Visualisierung Zur Unterstützung Von In Silico Medizin". Genauer umfasst diese kumulative Arbeit eine einleitende Fallstudie zu computergestützten Analysemethoden in der Domäne der Zellphysiologie, einen Bericht über aktuelle Anwendungen in der medizinischen Informatik zur visuellen Analyse komplexer Daten, eine Studie zur visuellen Darstellung von RNA-Strukturen und ihren Wahrscheinlichkeiten, eine Studie über aktuelle Möglichkeiten zur visuellen Analyse von Protein-Protein-Interaktionen und führt schließlich in die Thematik der Tumorforschung ein. Die Arbeit berichtet abschließend von neuartigen Ansätzen für computergestützte Visualisierung zur Unterstützung von „In Silico Modellierung von Tumor Wachstum".

# Acknowledgement

First of all, I would like to thank my supervisor Prof. Andreas Holzinger for his guidance, patience, encouragement and support in finding my way to conducting scientific research.

I also would like to thank my second supervisor for providing helpful perspectives on the use of visualization, as well as the HCI-KDD team and my colleagues for providing feedback.

Moreover, I am deeply grateful for all the support and motivation I received by my family and friends.

In particular I thank my beloved mother for inspiring me to always keeping up with studying, exploring, reflecting, analyzing and never giving up, even after her lifetime.

I also would like to thank my twin sister for discovering life together with countless different experiences and helping me finding my personal approach to help conquering cancer.

Last but not least, I acknowledge support by the Graz University of Technology and the Medical University of Graz.

Fleur Jeanquartier
Graz, 2017

## Statutory Declaration

*I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.*

Graz, _____ _____

Place, Date Signature

## Eidesstattliche Erklärung

*Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.*

Graz, am _____ _____

Ort, Datum Unterschrift

# Contents

Chapter 1

# Introduction

Within my phd studies I was working on topics related to the working title of this PhD thesis ”Visualization Support For In Silico Medicine”, that has been assigned to the subject of HCI/Visualization. During the last years I managed to publish several conference and journal publications in the fields of HCI and Visualization, in particular on Visual Analytics (VA) within the medical domain.

Interested in how fundamental research is being applied in biomedical science and how computational approaches are used, I started by conducting a case study in the domain of biomedical science [2]. I visited laboratories dealing with research in preclinical and translational medicine to understand how visual analysis is applied to fundamental research. The first goal was to highlight opportunities how fundamental research may benefit from applying sophisticated visualization methods known in the domain of computer science. Therefore, by observing some experiments, documenting data analysis steps and highlighting developable computational approaches, a first case study indicated the need for computational assistance in both validation as well as data analysis and proposed improving the choice of visual analysis tools that are used towards an integrated visualization and analysis approach.

Since then I also looked into other sub-research areas of biomedical science and studied uses of visualization to support the analysis of heterogeneous data from different biological scales. First of all, by continuing with participating in a review publication on interactive visual analysis in biomedical informatics I could strengthen my understanding of visualization methods used in in silico medicine as well as consolidated current state of the art within this field [8]. The state-of-the-art paper reflected on 59 examples of related literature and categorized them into two perspectives, namely into three different levels of integration versus three different analytical tasks. The work concluded with listing several open problems and underlined that most top ranked problems in the biomedical domain are related to usability issues.

To widen the understanding of the use of visualization in in silico medicine I explored and surveyed integrated visualization features for analyzing protein-protein-interaction databases [4]. Graphs are often used for visualizing biological data. Protein-protein-interaction databases integrate several different graph visualization libraries. 53 online available protein-protein-interaction databases have been examined, 10 of them have been described in more detail and ranked, according to their effective and efficient application to interactive visual analysis. Significant differences in user interface quality and data quality have been shown. The work also lists some open problems when visualizing biological networks ranging from providing interactive features for exploration and the handling of large graphs and high levels of details.

Since the beginning of my phd research I have been interested in tracking visualization problems within the biological domain. Therefore I tracked listed challenges and topics of ongoing and upcoming BioVis meetings. BioVis meetings have the goal to foster visualization research in problems in biological data visualization, as well as bioinformatics and biomedical research in state-of-the-art visualization research. Therefore, I participated in a BioVis contest [1] on uncertainty visualization of specific RNA data and described three different approaches to visualizing RNA folding uncertainty [5]. Each entry was reviewed

by 5 reviewers from the biological and/or visualization domain, but none of the entries could solve the problem completely. While a lot of questions remained open, there were several ideas for narrowing down the problem. My proposed ideas included user interface metaphors for interacting with several possible configurations to setting thresholds or for selecting structures that should be compared. The interactive approach was well rated regarding its excellent and detailed presentation. However, certain shortcomings regarding hiding information behind interactions have been marked as well as unanswered questions regarding the possibility of cluttering when not being limited to a certain threshold. On the other handy, one reviewer already mentioned, that the latter problem may be solved by the proposed graph approach that also was reviewed to work best. Next to my interactive proposal there was also a static one making use of Arc Diagrams, the winner with a circular one called CS2-UPlot, another circularly arranged one called RNA-SequenLens, a combination of the traditional dot plot and the MFE structure in the background and one making use of nested concave hulls to highlight certain probabilities [1].

Last but not least, my latest works include a project dealing with visualization support for in silico cancer research. Together with a domain expert in molecular biomedical science I managed to implement a simulation tool for cancer growth visualization that has been published in a high ranked systems biology journal [3]. Further case studies and reviews [7, 6] round up this research and mark the beginning of a promising research opportunity for future projects to fight cancer.

Chapter 2

# Publications

1. Fleur Jeanquartier, Andreas Holzinger: On Visual Analytics and Evaluation in Cell Physiology: A Case Study. CD-ARES 2013: 495-502 (2013)
2. Cagatay Turkay, Fleur Jeanquartier, Andreas Holzinger, Helwig Hauser: On Computationally-Enhanced Visual Analysis of Heterogeneous Data and Its Application in Biomedical Informatics. Interactive Knowledge Discovery and Data Mining in Biomedical Informatics 2014: 117-140 (2014)
3. Fleur Jeanquartier, Claire Jean-Quartier, Andreas Holzinger: Integrated web visualizations for protein-protein interaction databases. BMC Bioinformatics 16: 195 (2015)
4. Fleur Jeanquartier, Claire Jean-Quartier, Andreas Holzinger: Visualizing Uncertainty of RNA Sequence Base Pairing Variants: `http://biovis.net/year/2015/design/update` (2015)
5. Fleur Jeanquartier, Claire Jean-Quartier, Tobias Schreck, David Cemernek, Andreas Holzinger: Integrating Open Data on Cancer in Support to Tumor Growth Analysis. ITBAM 2016: 49-66 (2016)
6. Fleur Jeanquartier, Claire Jean-Quartier, David Cemernek, Andreas Holzinger: In silico modeling for tumor growth visualization. BMC Systems Biology 10: 59 (2016)
7. Fleur Jeanquartier, Claire Jean-Quartier, Max Kotlyar, Tomas Tokar, Anne-Christin Hauschild, Igor Jurisica, Andreas Holzinger: Machine Learning For In Silico Modeling Of Tumor Growth. ML for Health Informatics, LNAI 9605, Chapter 21, 978-3-319-50477-3, Springer, Heidelberg (2016)

## 2.1. On Visual Analytics and Evaluation in Cell Physiology: A Case Study

Within this conference publication I report on field notes and observations including lessons learned while studying how research is conducted in analyzing imaging data from experiments in the domain of cell physiology. Therefore, I visited a laboratory dealing with research in molecular biology and medical science, observed how experiments are made while taking notes, accompanied data analysis process steps, and finally interviewed a domain expert. A key challenge was the inability to publish any details on experiments' results, because of privacy protection reasons. However, the study was a key motivator for conducting further studies in the domain of biomedical science, because the observation highlights the need for enabling collaboration, conducting evaluation and improving tools used so far.

# On Visual Analytics and Evaluation in Cell Physiology: A Case Study

Fleur Jeanquartier and Andreas Holzinger

Research Unit Human-Computer Interaction, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz
{f.jeanquartier,a.holzinger}@hci4all.at

**Abstract.** In this paper we present a case study on a visual analytics (VA) process on the example of cell physiology. Following the model of Keim, we illustrate the steps required within an exploration and sense-making process. Moreover, we demonstrate the applicability of this model and show several shortcomings in the analysis tools functionality and usability. The case study highlights the need for conducting evaluation and improvements in VA in the domain of biomedical science. The main issue is the absence of a complete toolset that supports all analysis tasks including the many steps of data preprocessing as well as end-user development. Another important issue is to enable collaboration by creating the possibility of evaluating and validating datasets, comparing it with data of other similar research groups.

**Keywords:** visual analytics, evaluation of visualization, human computer interaction, biomedical science.

## 1 Introduction

From the first data analysis attempts to exploratory data analysis, up to information visualization, today we are facing the possibilities of visual analytics (VA). With VA several analysis processes may be transformed and become more effective and efficient through integrating automated analysis results and reasoning [1]. There is ongoing research in a variety of application areas ranging from document analysis over network security to molecular biology. Applying visual analysis techniques within these areas bring up certain limitations [2]. Dealing with the complexity of biological data requires sophisticated visualization technologies. Prominent examples of visualization for exploration and analysis in the domain of biology come from systems biology and include, among many others, the visualization of biological networks and omics data [3] such as protein structures [4], visual analysis of gene expression data [5], but also visual analysis of cell signaling networks [6]. For populating such databases for network analysis biologists also deal with basic research in cell physiology.

In Fig. 1 we see a slightly modified version of the VA Process, first described by [7]. According to Keim, humans have to be included early in the data analysis

A. Cuzzocrea et al. (Eds.): CD-ARES 2013, LNCS 8127, pp. 495–502, 2013.

process. By using their background knowledge and being supported by processing, transformation and visualization tools the analysis process eventually brings up new insight. We illustrate the mapping of a VA process during one example cell physiological experiment. Life scientists especially in the domain of biomedical science may struggle with the fact, that the process starts with a first data analysis. As for the observed work process later described in the Section 2 the domain expert also started with describing the hypothesis and then choosing suitable materials and methods for data acquisition. According to [7] input for the data sets used in the VA process are of heterogeneous nature and can be results from scientific experiments. Therefore we included the prior results as input for the feedback loop. The hypothesis may be formed by a preceding exploration. The domain expert makes use of knowledge gained by preceding work.

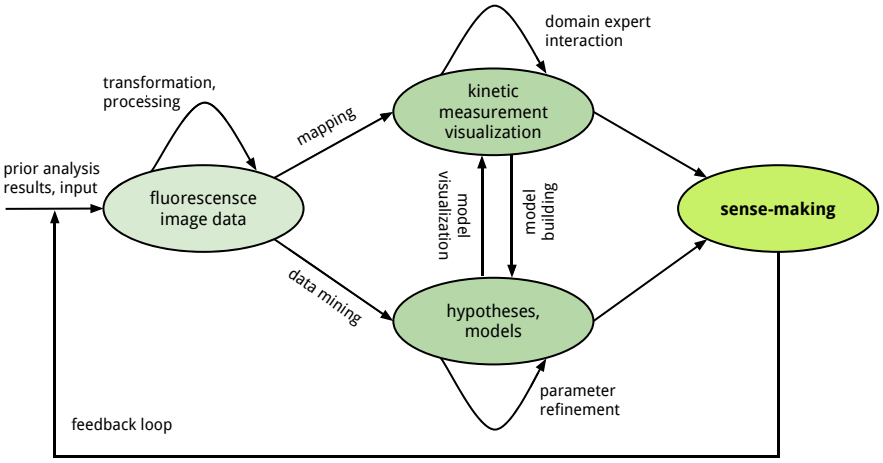


**Fig. 1.** Adapted Version of Keim's VA Process for the Application In Cell Physiology

Human computer interaction (HCI) and knowledge discovery (KDD) along with biomedical informatics are of increasing importance to effectively make sense out of data [8]. Biologists can benefit from a data deluge with the means of an integrated visualization approach, however, conducting evaluation and improving toolsets are still required to overcome certain hurdles on the way to new insights [9]. A domain expert as analyst often works alone while analysing data sets facing many problems, some of that have been illustrated by [10]. To foster sense-making and insights in VA systems it is essential to conduct studies and determine how people are using such systems [11, 12]. Case studies as field studies are a common approach to evaluating VA systems [13]. Qualitative evaluation such as observational studies can be conducted in a more realistic setting and allow improved understanding of existing practices for analysis and environmental constraints [14].

Consequently, we describe an observational study of a domain expert in cell physiology to present the current practice of VA in this domain.

## 2 Observation

The user, a domain expert within biomedical science, is part of the visual analysis and KDD process of a group of researchers dealing with cell physiology experiments. We accompanied the domain expert while investigating and analysing a set of experiments' results and observed the expert's analysis work. The analysis process includes visual analysis as part of the data processing, data analysis and KDD process as well as visual communication for dissemination.

A fluorescent biosensor [15] measures the concentration of certain molecules within cellular compartments. Fluorescent biosensors can be used for monitoring various processes and analytes such as metabolites, ions, target localization, gene expression and physiological relevant changes within subcelluar regions [16]. The biosensor allows to quantify variations in concentration or localization of the specific analyte within the cell by a change in fluorescence intensity. This quantification is further visualized as intensity signal over time in terms of kinetic curves. By that method, data in hundreds of columns and rows is recorded and has to be processed further. In summary, this method provides the measurement of biological signaling dynamics in vivo.

Experiments start with monitoring kinetics in signal transduction. The signal represents the fluorescence intensity [17]. First of all sequences of high-resolution fluorescent imaging of cells are acquired to capture dynamic changes. This action takes place in the lab's dark room. Fluorescence images are captured by a digital camera incorporating a CCD detector, connected to the fluorescence microscope. A commercial bioimaging software is used to communicate with the hardware, translating recorded signals to raw data. The software also provides some data/image processing functionality. Once the measurements are complete, the analysis process continues with data processing and image analysis. Noise (such as background lights within the dark room) reduction of images is supported by a ratio function. The domain expert marks specific regions of interest within the cell in order to monitor biological activites in healthy and pathological cells. Image segmentation is done manually insofar as the domain expert manually selects specific regions of interest on the image data for further comparison and analysis. Hence, regions of interest as polygon shapes are placed on every raw source to display the intensity value. The evaluation of whether the data and to what extent is accurate is done by manually comparing specific regions with a background region. The software allows the scientist to explore the data only in a very limited way. For not occupying the lab's dark room workplace for the time-consuming tasks of data processing and analysis, the expert moves to another workplace outside the dark room. Consequently, when the domain expert believes, that the data is sufficient, the raw data is exported to a commercial spreadsheet computation software via CSV for further processing and analysis.

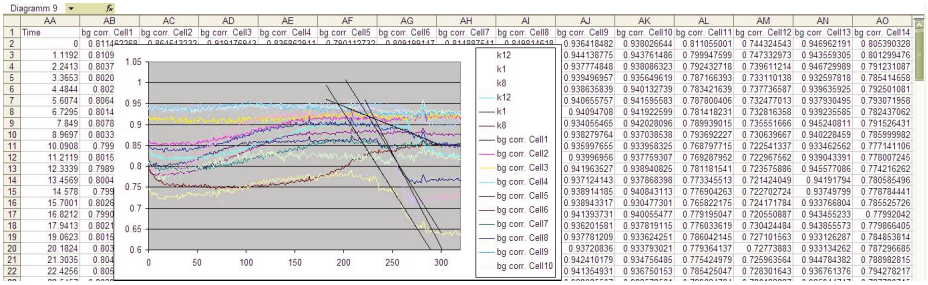


**Fig. 2.** First visual analysis of intensity signal over time

The domain expert creates a first visualization (compare Fig. 2) of the data, describing the kinetic changes in specific groups of healthy and pathological cells. This task is done semi-automatically by end-user development [18]. The following visual analysis shows, that the data has to be further filtered, corrected and transformed and finally improved in terms of readability and to be visualizable for the task of dissemination. It is up to the domain expert and the implicit knowledge of models, developed by the group of researchers within the lab, which transformation and manipulations are considered to be appropriate. Some of the processing tasks are automatic and some are again manual. The domain expert uses several tools for the various tasks and switches between them while advancing in the analysis process. While the process itself is occasionally being discussed in group, several smaller but complex actions are double checked by colleagues. Both the experiments and the visual analysis process are repeated many times until certain "surprising" [19] results get visible. This repetitive approach to gain new insights, also known as explorative data analysis (EDA), supports the process illustrated in Fig. 1 as it consists of a feedback loop. Finally, when the visual analysis results show surprising effects, the domain experts concludes with the dissemination (see Fig. 3) of the results, again with the means of visualization. The final visualizations are again being iteratively improved.

By further discussing the case study's process and comparing it with the VA process, we try to outline certain issues when dealing with the evaluation of scientific visualizations.

## 3 Discussion

The case study shows, that there are analysis processes in biomedical science which embody VA as a lived approach. At the same time, the case study also shows the need for improvements regarding HCI and end-user development. Experts in this domain are using their domain knowledge in combination with both automatic and visual analysis together, but need to be guided by computer science experts to improve the choice of tools that are used. There are certain tasks still done manually that could be automated or at least semi-automated, using

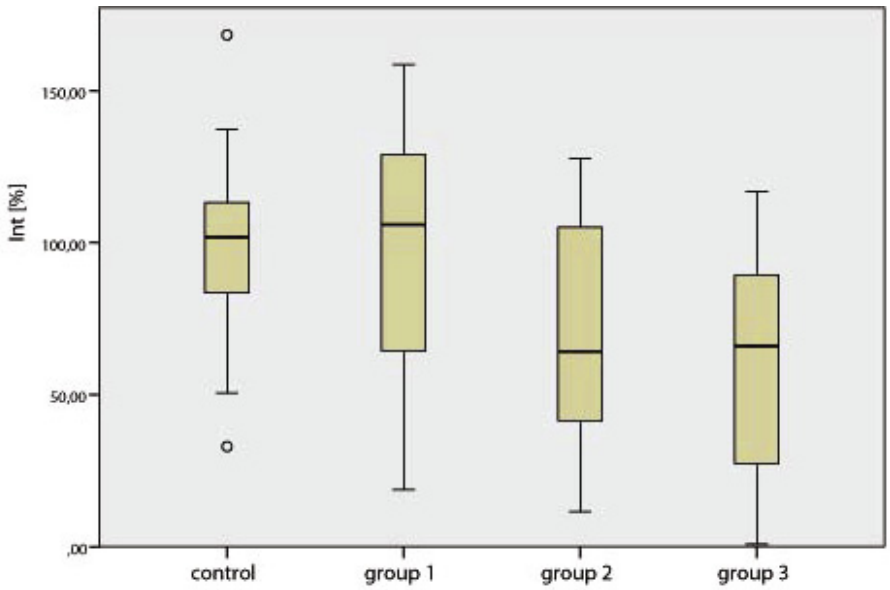


**Fig. 3.** Visualization of the kinetic parameters measured within the groups of healthy control to pathological cells for dissemination

the right tool, such as image segmentation, noise reduction as well as post-processing up to creating a set of fitting visualizations for dissemination. For instance, there already exist attempts to automatically estimate suitable background in fluorescence imaging [20] and automatic localization of cell nuclei [21]. The case study supports the statement, proposed by VA, that automated analysis often speeds up analysis tasks. It also shows that communication through visual representations is used for the dissemination of research results (see Fig. 3). Therefore, the case study shows, that visual analysis plays an important role for gaining new insights and dissemination. However, the case study also highlights that certain evaluation tasks are missing.

Lessons Learned include that there is a gap in free exploration of data and information due to the lacking usability and interaction possibilities of the tools used. The researchers within this field of study state, that they do not know about powerful VA solutions. At the same time, they are facing certain restrictions that hinder them to cooperate fully with computer science experts. During the observation the domain expert made complaints about shortcomings in the analysis tools' functionality and usability. Many tasks have to be repeated, not only due to data inconsistency, but also because most tools are hardly fault-tolerant and lack in supporting the user in certain data preprocessing steps as well as in the post-processing such as choosing the right visualization technique and improving the visualization's readability.

There are several possibilities to improve the end-user development and to minimize the interaction junk [22] within the observed process, such as simplifying

the creation of the effect curves for both visual analysis as well as dissemination. Furthermore, the visualizations in use are still limited to curve diagrams and bar charts. Alternative visualization metaphors such as multi-variate data visualizations [23] allow scientists to explore the data and it's various dimensions in other ways and may highlight certain effects that are not visible within the current effect curves.

The discussion after the observation further included improvements and suggestions to support the whole VA process. Due to the reason that both data as well as study results are confidential we are not allowed to go into detail in this respect. However, we already communicate general aspects of HCI and KDD and present general suggestions for improving VA within this domain. The domain expert agreed that there are several possibilities how evaluation could be integrated to support VA. Lam et al. already list some fitting evaluation goals and questions within the VDAR- and the CTV scenario [13]. However, the very idea of discussing the visual analysis process with a domain expert in HCI already brought up certain shortcomings within the visual analysis work. Suggestions include: Evaluating the dataset, comparing it to datasets of other similar groups of researchers, would help validating specific models as well as techniques and speed up the analysis work. Moreover, enabling and facilitating collaboration supports scientific problem solving [24]. The researchers in the group also agree on the fact, that evaluating software in use and furthermore, having the possibility to improve and extend the tools functionality would improve their daily research tasks. Incooperating the many steps of data examination and preprocessing into a single tool would be highly appreciated. The case study highlights the need for conducting evaluation and improvements in VA in the domain of biomedical science.

## 4 Conclusion

Every day scientists in many sub domains of life sciences such as biomedical science are facing the challenging task of VA with the goal of reaching new insights. Life scientists may benefit from a data deluge with the means of an integrated visualization approach. However, conducting evaluation and improving certain toolsets for exploratory data analysis and end-user development are prominent challenges on the way to new insights.

We described an observational study of VA in cell physiology. We compared the process to Keim's VA process. The case study shows, that there are analysis processes in biomedical science which embody VA. Further studies may include additional practice of VA related analysis work of various other approaches in biomedical science. The observation highlights the need for conducting evaluation and improvements in VA in the domain of biomedical science. We suggested evaluation possibilities and further noted challenges regarding its' application for visualization in life sciences. Among others, suggestions include incooperation and improvement of support for developing visualization in regard to analysis.

## References


[1] Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F.: Mastering The Information Age-Solving Problems with Visual Analytics. Florian Mansmann (2010)

[2] Kohlhammer, J., Keim, D., Pohl, M., Santucci, G., Andrienko, G.: Solving Problems with Visual Analytics. Procedia Computer Science 7, 117–120 (2011)

[3] Gehlenborg, N., O'Donoghue, S.I., Baliga, N.S., Goesmann, A., Hibbs, M.A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., et al.: Visualization of omics data for systems biology. Nature Methods 7, S56–S68 (2010)

[4] Doncheva, N.T., Assenov, Y., Domingues, F.S., Albrecht, M.: Topological analysis and interactive visualization of biological networks and protein structures. Nature Protocols 7(4), 670–685 (2012)

[5] Lex, A., Streit, M., Kruijff, E., Schmalstieg, D.: Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context. In: 2010 IEEE Pacific Visualization Symposium (PacificVis), pp. 57–64. IEEE (2010)

[6] Berger, S.I., Iyengar, R., Maayan, A.: Avis: Ajax viewer of interactive signaling networks. Bioinformatics 23(20), 2803–2805 (2007)

[7] 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)

[8] Holzinger, A.: On Knowledge Discovery and interactive intelligent visualization of biomedical data: Challenges in Human–Computer Interaction & Biomedical Informatics. In: DATA-International Conference on Data Technologies and Applications, pp. 5–16 (2012)

[9] Donoghue, S.I.O., Gavin, A.-c., Gehlenborg, N., Goodsell, D.S., Hériché, J.-k., Nielsen, C.B., North, C., Olson, A.J., Procter, J.B., Shattuck, D.W., Walter, T., Wong, B.: Visualizing biological data - now and in the future. Nature Publishing Group 7(3), S2–S4 (2010)

[10] 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)

[11] Kang, Y.A., Görg, C., Stasko, J.: How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation. IEEE Transactions on Visualization and Computer Graphics 17(5), 570–583 (2010)

[12] Wong, P.C., Shen, H.-W., Johnson, C.R., Chen, C., Ross, R.B.: The Top 10 Challenges in Extreme-Scale Visual Analytics. IEEE Computer Graphics and Applications 32(4), 63–67 (2012)

[13] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, S.: Empirical Studies in Information Visualization: Seven Scenarios.. IEEE Transactions on Visualization and Computer Graphics 18(9), 1–18 (2011)

[14] Carpendale, S.: Evaluating information visualizations. In: Kerren, A., Stasko, J.T., Fekete, J.-D., North, C. (eds.) Information Visualization. LNCS, vol. 4950, pp. 19–45. Springer, Heidelberg (2008)

[15] Morris, M.C.: Fluorescent biosensors of intracellular targets from genetically encoded reporters to modular polypeptide probes. Cell Biochemistry and Biophysics 56(1), 19–37 (2010)

[16] Okumoto, S., Jones, A., Frommer, W.B.: Quantitative imaging with fluorescent biosensors. Annual Review of Plant Biology 63, 663–706 (2012)

[17] Mehta, S., Zhang, J.: Reporting from the field: genetically encoded fluorescent reporters uncover signaling dynamics in living biological systems.. Annual Review of Biochemistry 80, 375–401 (2011)

[18] Lieberman, H., Paternò, F., Wulf, V.: End user development, vol. 9. Springer (2006)
[19] Beale, R.: Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. International Journal of Human-Computer Studies 65(5), 421–433 (2007)
[20] Chen, T.-W., Lin, B.-J., Brunner, E., Schild, D.: In situ background estimation in quantitative fluorescence imaging.. Biophysical Journal 90(7), 2534–2547 (2006)
[21] Song, Y., Cai, W., Huang, H., Wang, Y., Feng, D.D., Chen, M.: Region-based progressive localization of cell nuclei in microscopic images with data adaptive modeling. BMC Bioinformatics 14(1), 173 (2013)
[22] Endert, A., North, C.: Interaction junk. In: Proceedings of the 2012 BELIV Workshop on Beyond Time and Errors - Novel Evaluation Methods for Visualization - BELIV 2012, pp. 1–3. ACM Press, New York (2012)
[23] Fuchs, R., Hauser, H.: Visualization of Multi-Variate Scientific Data. Computer Graphics Forum 28(6), 1670–1690 (2009)
[24] Good, B.M., Su, A.I.: Games with a scientific purpose. Genome Biology 12(12), 135 (2011)

## 2.2. On Computationally-Enhanced Visual Analysis of Heterogeneous Data and Its Application in Biomedical Informatics

I participated in a joined work for a book chapter on integrative visual analysis in biomdeicine. My part was to review existing work in this domain, describe and classify visualization examples, identify challenges and extend the summary table.

# On Computationally-enhanced Visual Analysis of Heterogeneous Data and its Application in Biomedical Informatics

Cagatay Turkay[1], Fleur Jeanquartier[2]
Andreas Holzinger[2], and Helwig Hauser[3]

[1] giCentre, Department of Computer Science, City University, London, UK
Cagatay.Turkay.1@city.ac.uk
[2] Research Unit HCI, Institute for Medical Informatics, Statistics and Documentation Medical University Graz, Austria
{f.jeanquartier,a.holzinger}@hci4all.at
[3] Visualization Group, Department of Informatics, University of Bergen, Norway
Helwig.Hauser@uib.no

**Abstract.** With the advance of new data acquisition and generation technologies, the biomedical domain is becoming increasingly data-driven. Thus, understanding the information in large and complex data sets has been in the focus of several research fields such as statistics, data mining, machine learning, and visualization. While the first three fields predominantly rely on computational power, visualization relies mainly on human perceptual and cognitive capabilities for extracting information. Data visualization, similar to Human–Computer Interaction, attempts an appropriate interaction between human and data to interactively exploit data sets. Specifically within the analysis of complex data sets, visualization researchers have integrated computational methods to enhance the interactive processes. In this state-of-the-art report, we investigate how such an integration is carried out. We study the related literature with respect to the underlying analytical tasks and methods of integration. In addition, we focus on how such methods are applied to the biomedical domain and present a concise overview within our taxonomy. Finally, we discuss some open problems and future challenges.

**Keywords:** Visualization, Visual Analytics, Heterogenous Data, Complex Data, Future Challenges, Open Problems

## 1 Introduction and Motivation

Our society is becoming increasingly information-driven due to new technologies that provide data at an immense speed and scale. Even scientific practices are going under significant changes to adapt to this tremendous availability of data and data analysis is an important part in answering scientific questions. One of the fields where data analysis is especially important is biomedicine. In this domain, datasets are often structured in terms of both the scales they relate to,

e.g., from molecular interactions to how biological systems in the human body, and the inherent characteristics they carry, e.g., images from different medical devices. Such structures are both a challenge and a opportunity for scientists and significant efforts are put in several domains to understand these data. In this paper, we focus on how visualization, in particular those that incorporate computational analysis, approaches and enhances the analysis of structured information sources. We start with a section that discusses our goals and move on to more specific discussions on understanding information in data.

### 1.1 Goals

The best way of beginning such a paper, would be to start with the definition of Visualization and discuss the **goal of visualization**: A classical goal of visualization is, in an interactive, visual representation of abstract data, to amplify the acquisition or use of knowledge [1] and to enable humans to gain *insight* into complex data sets, either for the purpose of data exploration and analysis, or for data presentation [2], [3] (see section Glossary and Key Terms for more discussions). Visualization is a form of computing that provides new scientific insight through visual methods and therefore of enormous importance within the entire knowledge discovery process [4].

The **goal of this paper** is to provide a concise introduction into the visualization of large and heterogeneous data sets, in particular from the biomedical domain. For this purpose we provide a glossary to foster a common understanding, give a short nutshell-like overview about the current state-of-the-art and finally focus on open problems and future challenges. We base our taxonomy on a 2D structure on the different analytical tasks and on how computational methods can be integrated in visualizations. All the relevant works are then grouped under these categories. In addition to studies that do not have a specific application domain, we categorize visualization methods that specifically aimed at solving biomedical problems. Such subsets of work are presented under each category.

The **goal of this dual focus strategy** is to identify areas where visualization methods have shown to be successful but have not yet been applied to problems in the biomedical domain.

### 1.2 Understanding Information in Data

Understanding the relevant information in large and complex data sets has been in the focus of several research fields for quite a time; studies in statistics [5], data mining [6], machine learning [7], and in visualization [8] have devised methods to help analysts in extracting valuable information from a large variety of challenging data sets. While the first three fields predominantly rely on computational power, visualization relies mainly on the perceptual and cognitive capabilities of the human for extracting information. Although these research activities have followed separate paths, there have been significant studies to bring together the strengths from these fields [9–11]. Tukey [12] led the way

in integrating visualization and statistics with his work on **exploratory data analysis**. Earlier research on integrating statistics [13] and data mining [9] with information visualization have taken Tukey's ideas further.

This vision of integrating the best of both worlds has been a highly praised goal in visualization research [14–16] and parallels the emergence of *visual analytics* as a field on its own, which brings together research from visualization, data mining, data management, and human computer interaction [15]. In visual analytics research, the integration of automated and interactive methods is considered to be the main mechanism to foster the construction of knowledge in data analysis. In that respect, Keim [17] describes the details of a visual analysis process, where the data, the visualization, hypotheses, and interactive methods are integrated to extract relevant information. In their sense-making loop, based on the model introduced by van Wijk [18], the analytical process is carried out iteratively where the computational results are investigated through interactive visualizations. Such a loop aims to provide a better understanding of the data that will ultimately help the analyst to build new hypotheses. However, previously presented approaches still lack considering certain research issues to support a truly cross-disciplinary, seamless and holistic approach for the process chain of *data* > *information* > *knowledge*. Research needs to deal with data integration, fusion, preprocessing and data mapping as well as issues of privacy and data protection. These issues are being addressed in the HCI-KDD approach by Holzinger [19], [20] and is supported by the international expert network HCI-KDD (see hci4all.at).

### 1.3 Understanding Information in Biomedical Data

Interactive visual methods have been utilized within a wide spectrum of domains. In biomedicine, visualization is specifically required to support data analysts in tackling with problems inherent in this domain [20–22]. These can be summarized in three specific and general challenges:

*Challenge 1:* Due to the trend towards a data-centric medicine, data analysts have to deal with increasingly growing volumes and a diversity of highly complex, multi-dimensional and often weakly-structured and noisy data sets and increasing amounts of unstructured information.

*Challenge 2:* Due to the increasing trend towards precision medicine (P4 medicine: Predictive, Preventive, Participatory, Personalized (Hood and Friend, 2011)), biomedical data analysts have to deal with results from various sources in different structural dimensions, ranging from the microscopic world (systems biology, see below), and in particular from the "Omics-world" (data from genomics, proteomics, metabolomics, lipidomics, transcriptomics, epigenetics, microbiomics, fluxomics, phenomics, etc.) to the macroscopic world (e.g., disease spreading data of populations in public health informatics).

*Challenge 3:* The growing need for *integrative* solutions for interactive visualization of the data mentioned in challenge 1 and 2. Note that, although there are many sophisticated results and paradigms from the visualization community, integrated solutions, e.g. within business hospital information systems, are rare today.

An example from the biological domain can emphasize the aforementioned challenges: Biologists deal with data of different scale and resolution, ranging from tissues at the molecular and cellular scale ("the microscopic") up to organ scale ("the macroscopic"), as well as data from a diversity of databases of genomes and expression profiles, protein-protein interaction and pathways [23]. As understood by *systems biology*, the biological parts do not act alone, but in a strongly interwoven fashion, therefore biologists need to bridge and map different data types and analyze interactions [24]. Biomedicine has reached a point where the task of analyzing data is replacing the task of generating data [25]. At this point, visual analysis methods that support knowledge discovery in complex data become extremely important.

## 2 Glossary and Key Terms

In this section, we try to capture visualization and data analysis related terms that are only referenced explicitly within this paper. We do not cover the whole spectrum of visualization and analysis terms.

*Visualization:* is a visual representation of datasets intended to help people carry out some task more effectively according to Tamara Munzner [26]. Ward describes visualization as the graphical presentation of information, with the goal of providing the viewer with a qualitative understanding of the information contents [3].

*Space:* A set of points a $\in \mathbb{S}$ which satisfy some geometric postulate.

*Topological Visualization:* a prominent trend in current visualization research, driven by the data deluge. A topological abstraction provides a common mathematical language to identify structures and contexts [27], [28].

*Visual Analytics:* is an integrated approach combining visualization, human factors and data analysis to achieve a deep understanding of the data [14, 15].

*Interactive Visual Analysis (IVA):* is a set of methods that have overlaps with visual analytics. It combines the computational power of computers with the perceptive and cognitive capabilities of humans to extract knowledge from large and complex datasets. These techniques involve looking at datasets through different, linked views and iteratively selecting and examining features the user finds interesting.

*Heterogeneous data:* composed of data objects carrying different characteristics and coming from different sources. The heterogeneity can manifest itself in several forms such as different *scales of measure*, i.e., being categorical, discrete or continuous, or challenging to relate representations, e.g., genomic activity through gene expression vs. molecular pathways; a recent example of such data sets is described by Emmert-Streib et al. [29].

*Classification:* Methods that identify which subpopulation a new observation belongs on the basis of a training set of observations with known categories.

*Factor Analysis & Dimension Reduction:* is a statistical method that aims to describe the information in the data by preserving most of the variety. This process often leads to derived, unobserved variables called the factors [5]. Similarly, there exist dimension reduction methods, such as Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS) that project higher dimensional data onto lower dimensional spaces by preserving the variance in the data [5].

*Decision tree:* is a predictive statical model that enhances classification tasks [30]. It is often represented visually as a tree to support decision making tasks.

*Regression analysis:* is a statistical method that aims to estimate the relations between data variables. In other words, it tries to model how dependent certain factors are on others in the data [31].

## 3 State of the Art

There are a number of surveys that characterize how the integration of automated methods and interactive visualizations are accomplished. Crouser and Chang [32] characterize the human computer collaboration by identifying what contributions are made to the process by the two sides. In their survey, several papers are grouped according to these types of contributions. According to the authors, humans contribute to the analytical processes mainly by *visual perception*, *visuospatial thinking*, *creativity* and *domain knowledge*. On the other side, the computer contributes by *data manipulation, collection and storing*, and *bias-free analysis routines*. Bertini and Lalanne [16] categorize methods involving data mining and visualization into three: *computationally enhanced visualization*, *visually enhanced mining*, and *integrated visualization and mining*. Their categorization depends on whether it is the visualization or the automated method that plays the major role in the analysis.

In this state of the art analysis, we categorize the related literature in two perspectives. Our first perspective relates to the analytical task that is being carried out. After an investigation of literature from the computational data analysis domain [5, 33, 34], we identify a general categorization of the most common data analysis tasks as follows: *summarizing information*, *finding groups &*

*classification*, and *investigating relations & prediction*. We discuss these tasks briefly under each subsection in the following. Our second perspective relates to how the integration of computational tools in visual analysis is achieved. We identify three different categories to characterize the level of integration of computational tools in visualization, namely, *visualization as a presentation medium*, *semi-interactive use of computational methods* and the *tight integration of interactive visual and computational tools*. These levels are discussed in detail in Section 3.1.

In the following, we firstly organize the literature under the three analytical task categories and then group the related works further in sub-categories relating to the levels of integration. Before we move on to the literature review, we describe the three levels of integration introduced above. Even though we describe each analysis task separately, the categorization into the three common analysis tasks can be seen as a series of steps within a single analysis flow. Starting with summarizing information, proceeding with finding groups and last but not least finding relations and trends. One aspect that we do not cover explicitly is the consideration of outliers. Outlier analysis focuses on finding elements that do not follow the common properties of the data and needs to be part of a comprehensive data analysis process [35]. In this paper, we consider outlier analysis as an inherent part of summarizing information although there are works that are targeted at treating outliers explicitly [36].

Table 1 groups the investigated literature under the categories listed here. One important point to make with respect to the allocations to sub-groups in this table is that the borders within the categories are not always clear and there is rather a smooth transition between the categories. There are methods that try to address more than one analytical question. For such works, we try to identify the core questions tackled to place them in the right locations in this table. Similar smooth transitions also existent for the levels of integration, and our decision criteria is discussed in the following section.

### 3.1 Levels of Integration

On the first level of integration of computational tools within visual data analysis, visualization is used as a presentation medium to communicate the results of computational tools. These visualizations are either static representations, or only allow limited interaction possibilities such as zooming, panning, or making selections to highlight interesting parts of the data. A typical example for this category is the use of graphical plotting capabilities of statistical analysis software such as R [37]. In this system, users often refer to static visualizations to observe the results from computational procedures, such as clustering or fitting a regression line.

The second level of integration involves the use of the computational tool as a separate entity within the analysis where the tool's inner working is not transparent to the user. In this setting, the user interacts with the computational mechanism either through *modifying parameters* or *altering the data domain* being analyzed. The results are then presented to the user through different

| | Visualization as presentation | Semi-interactive Methods | Tight Integration |
|---|---|---|---|
| Summarizing Information | [38], [25] | [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50] [51] | [52], [53], [54], [55], [56], [57] |
| Groups & Classification | [58] [59], [60] | [61], [62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75] | [76], [77], [78], [79], [80], [81] |
| Dependence & Prediction | [82], [83], [47] | [84], [85], [86], [87], [88], [89], [90] | [91], [92], [93] |

**Table 1.** Analytical Tasks vs. Levels of Integration. This 2D structure is used to categorize the reviewed literature in this paper.

visual encodings that are often accompanied by interaction. One potential benefit here is that if problems are just too large so that a comprehensive computational approach is totally unfeasible, for ex., exhaustively searching a high-dimensional parameter space, then some directed steering by the intelligent expert user can help.

The third level constitutes mechanisms where a tight integration of interactive methods and computational tools is achieved. In these approaches, the automated methods are used seamlessly within interactive visual analysis. Sophisticated interaction mechanisms make the automated tools an integral part of the visualization. Methods in this category also interfere with the inner working of the algorithms and the results of automated tools are communicated immediately to the user.

When the second and the third levels are considered, we observe that categorizing a paper is not straightforward since the boundaries between these levels are smooth rather than discrete. In that respect, our classification criteria for level three is whether the integration allows for flexibility and done in a seamless way. If the integration is done at a manner where the automated method exists explicitly as a black-box that allows interaction to a certain level, we categorize the method under level two.

### 3.2 Summarizing Information

Data sets are becoming large and complex both in terms of the number of items and the number of modalities, i.e., data measured/collected from several sources, they contain. In order to tackle with the related visualization challenges, methods that are based on the summarization of underlying information are widely used in both automated and interactive visual data analysis [94]. Methods in this

category involve the integration of descriptive statistics, dimension reduction, and factor analysis methods in general.

### Visualization as presentation

For this category, we focus only on visualization tools in the biomedical context where there are many examples for visualization as presentation. As databases have become an integral part of dissemination and mining in biomedicine, the consolidation of such experiments data already brought up comprehensive tools for managing and sharing data. To name one, the Cell Centered Database [38] is a public image repository for managing and sharing (3D) imaging data. Next to image databases there is also a wide variety of different visualization tools, including interaction networks, pathway visualizations, multivariate omics data visualizations and multiple sequence alignments that have been reviewed recently by others [24, 25, 95]. In this context, visualization is most commonly used for exploration (hypothesis generation). Common visualization methods in addition to network visualization include scatter plots, profile plots/parallel coordinates and heatmaps with dendograms, while many tools provide combinations of those as linked views. Comprehensive summaries of visualization tools exist for certain areas. Nielsen et al. [25] present a review on tools for visualizing genomes, in particular tools for visualizing sequencing data, genome browsers and comparative genomics. Gehlenborg et al. [24] present a table of visualization tools in the area of systems biology, categorized by the different focusses of omics data. While most tools still lack in usability and integration, some of the listed tools already provide sophisticated interactive possibilities like annotating, comparing and showing confidence measures and prediction results next to view manipulations such as navigating, zooming and filtering. There is also a trend towards implementing web-based solutions to facilitate collaboration.

### Semi-interactive Methods

Perer and Shneiderman [46] discuss the importance of combining computational analysis methods, in particular statistics, with visualization to improve exploratory data analysis. Jänicke et al. [39] utilize a two-dimensional projection method where the analysis is performed on a projected 2D space called the attribute cloud. The resulting point cloud is then used as the medium for interaction where the user is able to brush and link the selections to other views of the data. Johansson and Johansson [40] enable the user to interactively reduce the dimensionality of a data set with the help of quality metrics. The visually guided variable ordering and filtering reduces the complexity of the data in a transparent manner where the user has a control over the whole process. The authors later use this methodology in the analysis of high-dimensional data sets involving microbial populations [41]. Fuchs et al. [42] integrate methods from machine learning with interactive visual analysis to assist the user in knowledge discovery. Performing the high-dimensional data analysis on derived attributes

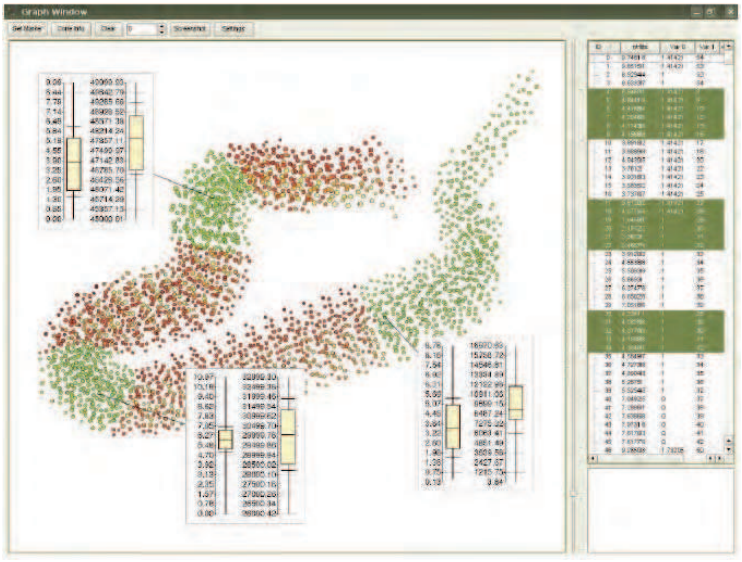

**Fig. 1.** Data can be visually analyzed on interactively created 2D spaces. (Image by Jänicke et al. [39])

is a strategy utilized in a number of studies. Kehrer et al. [50] integrate statistical moments and aggregates to interactively analyze collections of multivariate data sets. In the VAR display by Yang et al. [49], the authors represent the dimensions as glyphs on a 2D projection of the dimensions. A multidimensional scaling operation is performed on the glyphs where the distances between the dimensions are optimally preserved in the projection.

In Biomedicine there are only a few visualization tools that are being used to construct integrated web applications for interactive data analysis. Next to the UCSC Genome Browser [47], the IGV [48] is another common genome browser that integrates many different and large data sets and supports a wide variety of data types to be explored interactively. A few similar tools that are tightly integrated with public databases for systems biology are listed by Gehlenborg et al. [24].

In MulteeSum, Meyer et al. [51] used visual summaries to investigate the relations between linked multiple data sets relating to gene expression data. Artemis [44] supports the annotation and visual inspection, comparison and analysis of high-throughput sequencing experimental data sets. The String-DB [45] is a commonly used public comprehensive database for protein-protein interaction that supports visual data analysis by providing interactive network visualizations.

Otasek et al. [96] present a work on Visual Data Mining (VDM), which is supported by interactive and scalable network visualization and analysis. Otasek

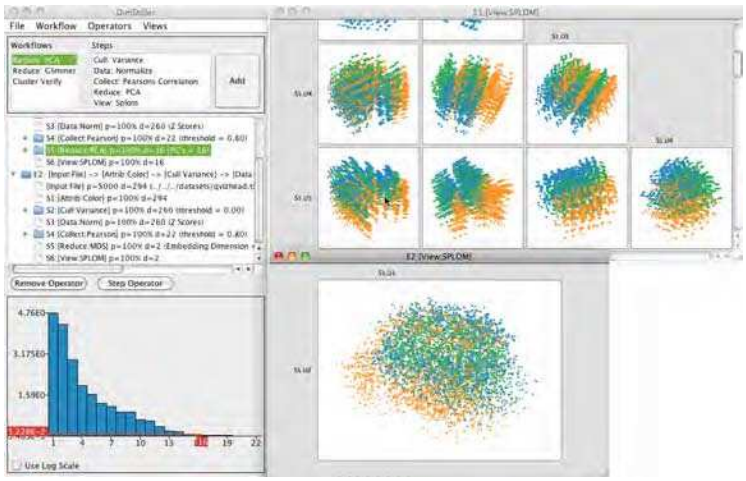


**Fig. 2.** A selection of data transformations are chained together interactively to achieve dimension reduction. (Image by Ingram et al. [55])

et al. emphasize that knowledge discovery within complex data sets involves many workflows, including 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.

Mueller et al. report in a recent work [97] on the successful application of data Glyphs in a disease analyser for the analysis of big medical data sets with automatic validation of the data mapping, selection of subgroups within histograms and a visual comparison of the value distributions.

### Tight Integration

Nam and Mueller [52] provides the user with an interface where a high-dimensional projection method can be steered according to user input. In MDSteer [53], an embedding is guided with user interaction leading to an adapted multidimensional scaling of multivariate data sets. Such a mechanism enables the analyst to steer the computational resources accordingly to areas where more precision is needed. Ingram et al. [55] present a system called DimStiller, where a selection of data transformations are chained together interactively to achieve dimension reduction. Endert et al. [54] introduce observation level interactions to assist computational analysis tools to deliver more reliable results. The authors describe such operations as enabling the *direct manipulation* for visual analytics [56]. Turkay et al. introduce the dual-analysis approach [57] to support analysis processes where computational methods such as dimension reduction [93] are used.

### 3.3 Finding groups & Classification

One of the most common analytical tasks in data analysis is to determine the different groups and classifications [5]. Analysts often employ cluster analysis methods that divide data into clusters where data items are assigned to groups that are similar with respect to certain criteria [98]. One aspect of cluster analysis is that it is an unsupervised method, i.e., the number of groups or their labels are not known a priori. However, when the analyst has information on class labels beforehand, often referred to as *the training set*, classification algorithms can be utilized instead. Below, we list interactive visualization methods where cluster analysis tools and/or classification algorithms are utilized.

#### Visualization as presentation

Parallel Sets by Kosara et al. [59] is a successful example where the overlaps between groups is presented with a limited amount of interaction. In the software visualization domain, Telea and Auber [60] represent the changes in code structures using a flow layout where they identify steady code blocks and when splits occur in the code of a software. Demvsar et al. [58] present a visualization approach for exploratory data analysis of multidimensional data sets and show it's utility for classification on several biomedical data sets.

#### Semi-interactive Methods

May and Kohlhammer [65] present a conceptual framework that improves the classification of data using decision trees in an interactive manner. The authors later proposed a technique called SmartStripes [66] where they investigate the relations between different subsets of features and entities. Interactive systems have also been used to help create decision trees [99]. Guo et al. [71] enable the interactive exploration of multivariate model parameters. They visualize the model space together with the data to reveal the trends in the data. Kandogan [72] discusses how clusters can be found and annotated through an image-based technique. Rinzivillo et al. [73] use a visual technique called progressive clustering where the clustering is done using different distance functions in consecutive steps. Schreck et al. [74] propose a framework to interactively monitor and control Kohonen maps to cluster trajectory data. The authors state the importance of integrating the expert within the clustering process in achieving good results. gCluto [75] is an interactive clustering and visualization system where the authors incorporate a wide range of clustering algorithms.

In *Hierarchical Clustering Explorer* [70], Seo and Shneiderman describe the use of an interactive dendogram coupled with a colored heatmap to represent clustering information within a coordinated multiple view system. Other examples include works accomplished within the Caleydo software for pathway analysis and associated experimental data by Lex et al. [61–63]. In a recent paper, the integrated use of statistical computations is shown to be useful to characterize

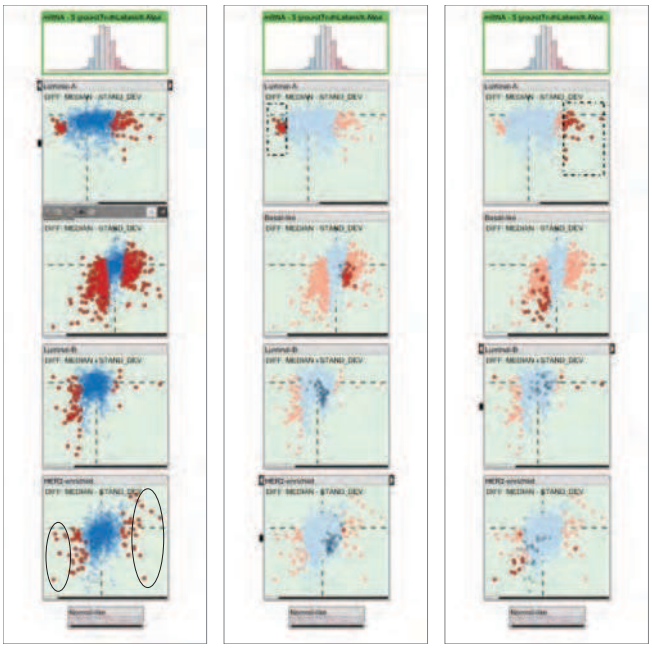

**Fig. 3.** Results of statistical test computations are communicated through visual encodings to support the identification of discriminative elements in subgroups. (Image by Turkay et al. [64])

the groupings in the data [64]. Gehlenborg et al. [24] identified that scatter plots, profile plots and heat maps are the most common visualization techniques used in interactive visualization tools for tasks like gene expression analysis. Younesy et al. [67] presents a framework where users have the ability to steer clustering algorithms and visually compare the results. Dynamically evolving clusters, in the domain of molecular dynamics, are analyzed through interactive visual tools by Grottel et al. [68]. The authors describe flow groups and a schematic view that display cluster evolution over time. Mayday is one framework example where a visual analytics framework supports clustering of gene expression data sets [69].

### Tight Integration

Turkay et al. presents an interactive system that addresses both the generation and evaluation stages in a clustering process [80]. Another example is the iVisClassifier by Choo et al. [81] where the authors improve classifitcation performance through interactive visualizations. Ahmed and Weaver [76] discuss how the clustering process can be embedded within an highly interactive system. Examples in biomedical domain are rare in this category. One example is by Rubel

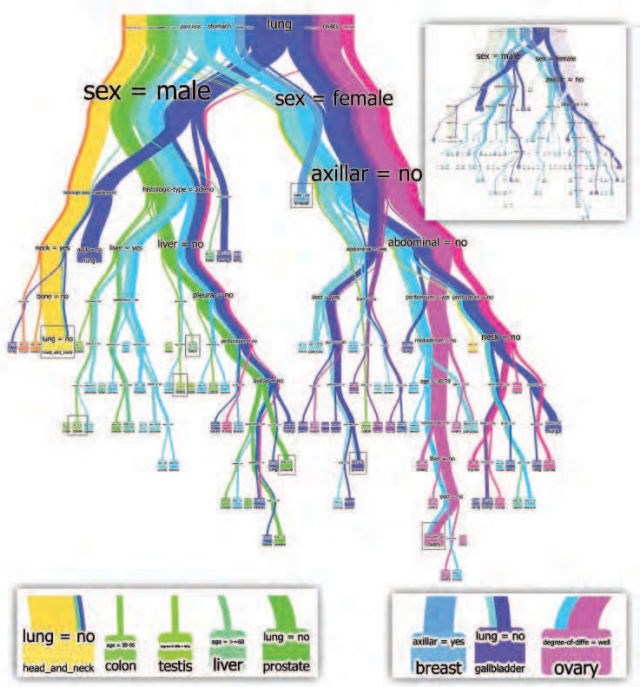


**Fig. 4.** Interactive systems have been used to help create and evaluate decision trees (Image by van den Elzen and van Wijk [99])

et al. [77], who present a framework for clustering and visually exploring (3D) expression data. In the domain of molecular dynamics simulation, there are some examples of tight integrations of interactive visualizations, clustering algorithms, and statistics to support the validity of the resulting structures [78], [79].

### 3.4 Investigating dependence

An often performed task in data analysis is the investigation of relations within different features in a data set [100]. This task is important to build cause and effect relations, understanding the level of dependence between features, and predicting the possible outcomes based on available information. In this category, we list interactive methods that incorporate computational tools to facilitate such tasks. Often employed mechanisms are: regression, correlation, and predictive analysis approaches. In the biomedical domain, Secrier et al. [101] present a list of tools that deal with the issue of time, however, they note that it is yet an open challenge in comparative genomics to find tools for analyzing time series data that can handle both the visualization of changes as well as showing trends and predictions for insightful inferences and correlations.


**Fig. 5.** Visualization helps analysts in making predictions and investigating uncertainties in relations within simulation parameters (Image by Booshehrian et al. [87])

### Visualization as presentation

In this category, we focus mainly on works from biomedical domain. Krzywinski et al. [82] presents a tool for comparative genomics by visualizing variation in genome structure. Karr et al. [83] present a promising topic, namely computing comprehensive whole-cell model and presenting model predictions for cellular and molecular properties.

Nielsen et al. [25] reviews tools for the visual comparison of genomes. The list of referenced tools includes Circos [82], a visualization presentation method for visualizing synteny in a circular layout. One example referenced is the already mentioned UCSC genome browser [47] that also provides simple phylogenetic tree graphs. The list also includes tools that integrate computational methods and support the visual analysis of comparative genomics more interactively, which are discussed in the next level of integration.

### Semi-interactive Methods

Visualization has shown to be effective in validating predictive models through interactive means [85]. Mühlbacher and Piringer [86] discuss how the process of building regression models can benefit from integrating domain knowledge. In the framework called Vismon, visualization has helped analysts to make predictions and investigate the uncertainties that are existent in relations within simulation parameters [87]. Interaction methods facilitate the investigation of multivariate

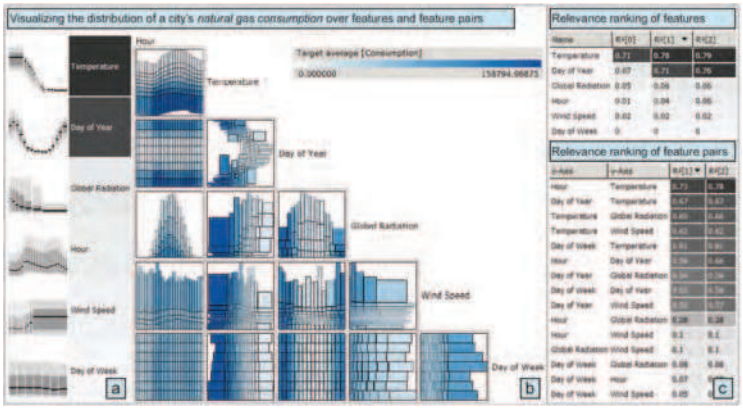


**Fig. 6.** The process of building regression models can benefit from integrating domain knowledge through interactive visualizations. (Image by Mühlbacher and Piringer [86])

relations in multi-variate data sets [89]. Yang et al. [90] analyze the relations between the dimensions of a data set to create a hierarchy that they later use to create lower-dimensional spaces.

Within biomedical applications, Meyer et al. [84] present a synteny browser called MizBee, that provides circular views for the interactive exploration and analysis of conserved synteny relationships at multiple scales. In a later paper, they investigate the dependencies within signals coming from related data sets and present a comparative framework [88].

### Tight Integration

Berger et al. [91] introduce an interactive approach that enables the investigation of the parameter space with respect to multiple target values. Malik et al. [92] describe a framework for interactive auto-correlation. This is an example where the correlation analysis is tightly coupled with the interactive elements in the visualization solution. Correlation analysis have been integrated as an internal mechanism to investigate how well lower-dimensioal projections relate to the data that they represent [93].

## 4 Open Problems

Chaomei Chen (2005) [102] raised a list of top 10 unsolved information visualization problems, interestingly on top are usability issues, which are particularly relevant for the biomedical domain, as a recent study has shown [103]. This is mostly due to the fact that usability engineering methods are still considered

as nice add-on and not yet an integrated part in the software development process [104]. Here we list a number of open problems in relation to the literature we cover in this report.

*Problem 1.* A topic that needs further attention is to address the uncertainty within the analysis process. The explorative nature of interactive visual analysis creates a vast amount of analysis possibilities and often leads to several plausible results. It is thus of great importance to reduce this space of possibilities and inform the user about the certainty of the results.

*Problem 2.* Although we have seen several works that involve a tight integration between computational methods and visualization, examples of *seamless integrations* are rare. With this term, we refer to interaction mechanisms where the support from appropriate sophisticated computational tools are provided to the user without the analyst noticing the complexities of the underlying mechanisms. One example to clarify this term could be: applying regression analysis locally on a selection within a 2D scatterplot and presenting the result immediately with a regression line.

*Problem 3.* One aspect that needs to be investigated further in the integration of interactive and automated methods is the *issue of usability*. Most of the solutions introduced here require significant literacy in statistics and skills in using different computational methods – which can lead to a demanding learning curve.

*Problem 4.* We have seen that most of the visual analysis methods are focussed at particular data types. However, given the current state of data collection and data recording facilities, there are often several data sets related to a phenomenon. There is the need for advanced mechanisms that can harness these various sources of information and help experts to run analysis that stretches over several data sets. This issue relates to the goal of developing an integrated visualization environment spanning several biological dimensions, from micro to macro towards an integrated approach. The recent survey by Kehrer and Hauser [105], which illustrates the many different axes along which data complexity evolves and how visualization can address these complexities, is a starting point to identify suitable approaches.

*Problem 5.* One observation we make is that the visualization methods often use the support from a single, specific computational mechanism. However, in order to achieve a comprehensive data analysis session, one needs to address all of the analysis tasks we present in our discussions above from summarizing information up to finding cause and effect [23, 101]. Especially, when works relating to biomedical applications are considered, we notice that studies that involve the tight integration of computational tools are rare. Given the successful application of such methods in other domains, it is expected that biomedical applications can also benefit significantly from these approaches.

## 5 Future Outlook

As stated within the open problems above, there is a certain need for mechanisms to improve the interpretability and usability of interactive visual analysis techniques. Possible methods could be to employ *smart labeling and annotation*, creating *templates that analysts can follow* for easier progress, and *computationally guided interaction* mechanisms where automated methods are *integrated seamlessly*. Such methods need to utilize computational tools as underlying support mechanism for users, one aspect that needs attention in this respect is to maintain the interactivity of the systems. Appropriate computation and sampling mechanisms needs to be developed to achieve such systems.

In order to address the uncertainties in visual data analysis, mechanisms that communicate the reliability of the observations made through interactive visualizations need to be developed, e.g., what happens to the observation if the selection is moved slightly along the x-axis of a scatter plot? If such questions are addressed, interactive and visual methods could easily place themselves in the everyday routine of analysts that require precise results.

The ability to define features interactively and refine feature definitions based on insights gained during visual exploration and analysis provides an extremely powerful and versatile tool for knowledge discovery. Future challenges lie in the integration of alternate feature detection methods and their utilization in intelligent brushes. Furthermore, integrating IVA and simulations, thus supporting computational steering, offers a wide range of new possibilities for knowledge discovery [106].

An interesting direction for future research relates to improving the usability of analysis processes. Current usability studies often focus on specific parts of a technique. However in order to evaluate the effectiveness of the whole analysis process, there is the need to perform comprehensive investigations on the interpretability of each step of the analysis and study the effects of using computational tools interactively. Such studies can be carried out in forms of controlled experiments where the analysts are given well-determined tasks and are asked to employ particular types of analysis routes. These routes can then be evaluated and compared against non-interactive processes where possible.

A challenging future research avenue for effective HCI is to find answers to the question "What is interesting?" as *Interest* is an essentially human construct [107], a perspective on relationships between data that is influenced by context, tasks, personal preferences, previous knowledge (=expectations) and past experience [108]. For a correct semantic interpretation, a computer would need to understand the *context* in which a visualization is presented; however, comprehension of a complex context is still beyond computation. In order for a data mining system to be generically useful, it must therefore have some way in which one can indicate what is interesting, and for that to be dynamic and changeable [109].

A very recent research route in HCI is *Attention Routing*, which is a novel idea introduced by Polo Chau [110] and goes back to models of attentional mechanisms for forming position-invariant and scale-invariant representations

of objects in the visual world [111]. Attention routing is a promising approach to overcome one very critical problem in visual analytics, particularly of large and heterogeneous data sets: to help users locate good starting points for their analysis. Based on *anomaly detection* [112], attention routing methods channel the end-users to interesting data subsets which do not conform to standard behaviour. This is a very promising and important research direction for Knowledge Discovery and Data Mining [**?**].

Top end research routes encompassing uncountable research challenges are in the application of computational topology [27], [113], [114] approaches for data visualization. Topology-based methods for visualization and visual analysis of data are becoming increasingly popular, having their major advantages in the capability to provide a concise description of the overall *structure* of a scientific data set, because subtle features can easily be missed when using traditional visualization methods (e.g. volume rendering or isocontouring), unless correct transfer functions and isovalues are chosen. By visualizing a topology directly, one can guarantee that no feature is missed and most of all solid mathematical principles can be applied to simplify a topological structure. The topology of functions is also often used for feature detection and segmentation (e.g., in surface segmentation based on curvature) [115].

In this state-of-the-art report, we investigated the literature on how visualization and computation support each other to help analysts in understanding complex, heterogeneous data sets. We also focused on to what degree these methods have been applied to biomedical domain. When the three different levels of integration are considered, we have observed that there are not yet many works falling under the third integration level. We have seen that existing applications in this category have significant potential to address the challenges discussed earlier in the paper. However, there exist several open problems, as discussed above, which can motivate the visualization and knowledge discovery community to carry out research on achieving a tight integration of computational power and capabilities of human experts.

## References

1. Card, S.K., Mackinlay, J.D., Shneiderman, B.: Information Visualization: Using Vision to Think. Morgan Kaufmann, San Francisco (1999)
2. Moeller, T., Hamann, B., Russell, R.D.: Mathematical foundations of scientific visualization, computer graphics, and massive data exploration. Springer (2009)
3. Ward, M., Grinstein, G., Keim, D.: Interactive data visualization: foundations, techniques, and applications. AK Peters, Ltd. (2010)
4. 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) (2014) I1
5. Johnson, R., Wichern, D.: Applied multivariate statistical analysis. Volume 6. Prentice Hall Upper Saddle River, NJ: (2007)
6. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley Longman Publishing Co., Inc. (2005)

7. Alpaydin, E.: Introduction to machine learning. MIT press (2004)
8. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Visual Languages, 1996. Proceedings., IEEE Symposium on, IEEE (1996) 336–343
9. Keim, D.: Information visualization and visual data mining. IEEE transactions on Visualization and Computer Graphics **8**(1) (2002) 1–8
10. Shneiderman, B.: Inventing discovery tools: combining information visualization with data mining1. Information Visualization **1**(1) (2002) 5–12
11. Ma, K.L.: Machine learning to boost the next generation of visualization technology. Computer Graphics and Applications, IEEE **27**(5) (2007) 6–9
12. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley (1977)
13. Cleveland, W.S., Mac Gill, M.E.: Dynamic graphics for statistics. CRC Press (1988)
14. Thomas, J.J., Cook, K.A.: Illuminating the Path: The Research and Development Agenda for Visual Analytics. National Visualization and Analytics Ctr (2005)
15. Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F.: Mastering The Information Age-Solving Problems with Visual Analytics. Florian Mansmann (2010)
16. Bertini, E., Lalanne, D.: Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. SIGKDD Explor. Newsl. **11**(2) (2010) 9–18
17. Keim, D., Andrienko, G., Fekete, J., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: Definition, process, and challenges. Information Visualization (2008) 154–175
18. van Wijk, J.J.: The value of visualization. In: Visualization, 2005. VIS 05. IEEE, IEEE (2005) 79–86
19. Holzinger, A.: Human-computer interaction and knowledge discovery (hci-kdd): What is the benefit of bringing those two fields to work together? In Cuzzocrea, A., Kittl, C., Simos, D., Weippl, E., Xu, L., eds.: Availability, Reliability, and Security in Information Systems and HCI. Volume 8127 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2013) 319–328
20. Holzinger, A., Jurisica, I. In: Knowledge Discovery and Data Mining in Biomedical Informatics: The future is in Integrative, Interactive Machine Learning Solutions. Springer, Heidelberg, Berlin (2014) in print
21. Holzinger, A.: On knowledge discovery and interactive intelligent visualization of biomedical data - challenges in humancomputer interaction and biomedical informatics. In: DATA 2012, INSTICC (2012) 9–20
22. Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J., Altman, R.B.: Bioinformatics challenges for personalized medicine. Bioinformatics **27**(13) (2011) 1741–1748
23. O'Donoghue, S.I., Gavin, A.C., Gehlenborg, N., Goodsell, D.S., Hériché, J.K., Nielsen, C.B., North, C., Olson, A.J., Procter, J.B., Shattuck, D.W., et al.: Visualizing biological datanow and in the future. Nature methods **7** (2010) S2–S4
24. Gehlenborg, N., O'Donoghue, S., Baliga, N., Goesmann, A., Hibbs, M., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., et al.: Visualization of omics data for systems biology. Nature methods **7** (2010) S56–S68
25. Nielsen, C.B., Cantor, M., Dubchak, I., Gordon, D., Wang, T.: Visualizing genomes: techniques and challenges. Nature methods **7** (2010) S5–S15
26. Munzner, T.: Visualization principles, Presented at VIZBI 2011: Workshop on Visualizing Biological Data (2011)
27. Hauser, H., Hagen, H., Theisel, H.: Topology-based methods in visualization (Mathematics+Visualization). Springer, Berlin Heidelberg (2007)

28. Pascucci, V., Tricoche, X., Hagen, H., Tierny, J.: Topological Methods in Data Analysis and Visualization: Theory, Algorithms, and Applications (Mathematics+Visualization). Springer, Berlin, Heidelberg (2011)
29. Emmert-Streib, F., de Matos Simoes, R., Glazko, G., McDade, S., Haibe-Kains, B., Holzinger, A., Dehmer, M., Campbell, F.: Functional and genetic analysis of the colon cancer network. BMC Bioinformatics **15**(Suppl 6) (2014) S6
30. Olshen, L.B.J.F.R., Stone, C.J.: Classification and regression trees. Wadsworth International Group (1984)
31. Cohen, J., Cohen, P., West, S.G., Aiken, L.S.: Applied multiple regression/correlation analysis for the behavioral sciences. Lawrence Erlbaum (2003)
32. Crouser, R.J., Chang, R.: An affordance-based framework for human computation and human-computer collaboration. Visualization and Computer Graphics, IEEE Transactions on **18**(12) (2012) 2859–2868
33. Brehmer, M., Munzner, T.: A multi-level typology of abstract visualization tasks. Visualization and Computer Graphics, IEEE Transactions on **19**(12) (2013) 2376–2385
34. Kerren, A., Ebert, A., Meyer, J.: Human-centered visualization environments. Springer-Verlag (2006)
35. Filzmoser, P., Hron, K., Reimann, C.: Principal component analysis for compositional data with outliers. Environmetrics **20**(6) (2009) 621–632
36. Novotný, M., Hauser, H.: Outlier-preserving focus+context visualization in parallel coordinates. Visualization and Computer Graphics, IEEE Transactions on **12**(5) (2006) 893–900
37. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2013)
38. Martone, M.E., Tran, J., Wong, W.W., Sargis, J., Fong, L., Larson, S., Lamont, S.P., Gupta, A., Ellisman, M.H.: The cell centered database project: an update on building community resources for managing and sharing 3d imaging data. Journal of structural biology **161**(3) (2008) 220–231
39. Jänicke, H., Böttinger, M., Scheuermann, G.: Brushing of attribute clouds for the visualization of multivariate data. IEEE Transactions on Visualization and Computer Graphics (2008) 1459–1466
40. Johansson, S., Johansson, J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. Visualization and Computer Graphics, IEEE Transactions on **15**(6) (2009) 993–1000
41. Fernstad, S., Johansson, J., Adams, S., Shaw, J., Taylor, D.: Visual exploration of microbial populations. In: Biological Data Visualization (BioVis), 2011 IEEE Symposium on. (2011) 127 –134
42. Fuchs, R., Waser, J., Gröller, M.E.: Visual human+machine learning. IEEE TVCG **15**(6) (2009) 1327–1334
43. Oeltze, S., Doleisch, H., Hauser, H., Muigg, P., Preim, B.: Interactive visual analysis of perfusion data. Visualization and Computer Graphics, IEEE Transactions on **13**(6) (2007) 1392 –1399
44. Carver, T., Harris, S.R., Berriman, M., Parkhill, J., McQuillan, J.A.: Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics **28**(4) (2012) 464–469
45. Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., et al.: String v9. 1: protein-protein interaction networks, with increased coverage and integration. Nucleic acids research **41**(D1) (2013) D808–D815

46. Perer, A., Shneiderman, B.: Integrating statistics and visualization for exploratory power: From long-term case studies to design guidelines. Computer Graphics and Applications, IEEE **29**(3) (2009) 39 –51
47. Kuhn, R.M., Haussler, D., Kent, W.J.: The ucsc genome browser and associated tools. Briefings in bioinformatics **14**(2) (2013) 144–161
48. Thorvaldsdóttir, H., Robinson, J.T., Mesirov, J.P.: Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. Briefings in bioinformatics **14**(2) (2013) 178–192
49. Yang, J., Hubball, D., Ward, M., Rundensteiner, E., Ribarsky, W.: Value and relation display: Interactive visual exploration of large data sets with hundreds of dimensions. Visualization and Computer Graphics, IEEE Transactions on **13**(3) (2007) 494 –507
50. Kehrer, J., Filzmoser, P., Hauser, H.: Brushing moments in interactive visual analysis. Computer Graphics Forum **29**(3) (2010) 813–822
51. Meyer, M., Munzner, T., DePace, A., Pfister, H.: Multeesum: A tool for comparative spatial and temporal gene expression data. Visualization and Computer Graphics, IEEE Transactions on **16**(6) (2010) 908–917
52. Nam, J., Mueller, K.: Tripadvisorn-d: A tourism-inspired high-dimensional space exploration framework with overview and detail. Visualization and Computer Graphics, IEEE Transactions on **19**(2) (2013) 291–305
53. Williams, M., Munzner, T.: Steerable, progressive multidimensional scaling. In: Proceedings of the IEEE Symposium on Information Visualization, Washington, DC, USA, IEEE Computer Society (2004) 57–64
54. Endert, A., Han, C., Maiti, D., House, L., North, C.: Observation-level interaction with statistical models for visual analytics. In: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, IEEE (2011) 121–130
55. Ingram, S., Munzner, T., Irvine, V., Tory, M., Bergner, S., Möller, T.: Dimstiller: Workflows for dimensional analysis and reduction. In: Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on. (2010) 3 –10
56. Endert, A., Bradel, L., North, C.: Beyond control panels: Direct manipulation for visual analytics. Computer Graphics and Applications, IEEE **33**(4) (2013) 6–13
57. Turkay, C., Filzmoser, P., Hauser, H.: Brushing dimensions – a dual visual analysis model for high-dimensional data. Visualization and Computer Graphics, IEEE Transactions on **17**(12) (2011) 2591 –2599
58. Demšar, J., Leban, G., Zupan, B.: Freeviz - an intelligent multivariate visualization approach to explorative analysis of biomedical data. Journal of biomedical informatics **40**(6) (2007) 661–671
59. Kosara, R., Bendix, F., Hauser, H.: Parallel sets: interactive exploration and visual analysis of categorical data. Visualization and Computer Graphics, IEEE Transactions on **12**(4) (2006) 558 –568
60. Telea, A., Auber, D.: Code flows: Visualizing structural evolution of source code. Computer Graphics Forum **27**(3) (2008) 831–838
61. Lex, A., Streit, M., Schulz, H.J., Partl, C., Schmalstieg, D., Park, P.J., Gehlenborg, N.: StratomeX: visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. Computer Graphics Forum (EuroVis '12) **31**(3) (2012) 1175?1184
62. Lex, A., Streit, M., Partl, C., Kashofer, K., Schmalstieg, D.: Comparative analysis of multidimensional, quantitative data. IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2010) **16**(6) (2010) 1027–1035

63. Partl, C., Kalkofen, D., Lex, A., Kashofer, K., Streit, M., Schmalstieg, D.: enroute: Dynamic path extraction from biological pathway maps for in-depth experimental data analysis. In: Biological Data Visualization (BioVis), 2012 IEEE Symposium on, IEEE (2012) 107–114
64. Turkay, C., Lex, A., Streit, M., Pfister, H., Hauser, H.: Characterizing cancer subtypes using dual analysis in caleydo stratomex. IEEE Computer Graphics and Applications **34**(2) (2014) 38–47
65. May, T., Kohlhammer, J.: Towards closing the analysis gap: Visual generation of decision supporting schemes from raw data. In: Computer Graphics Forum. Volume 27., Wiley Online Library (2008) 911–918
66. May, T., Bannach, A., Davey, J., Ruppert, T., Kohlhammer, J.: Guiding feature subset selection with an interactive visualization. In: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, IEEE (2011) 111–120
67. Younesy, H., Nielsen, C.B., Möller, T., Alder, O., Cullum, R., Lorincz, M.C., Karimi, M.M., Jones, S.J.: An interactive analysis and exploration tool for epigenomic data. In: Computer Graphics Forum. Volume 32., Wiley Online Library (2013) 91–100
68. Grottel, S., Reina, G., Vrabec, J., Ertl, T.: Visual verification and analysis of cluster detection for molecular dynamics. IEEE Transactions on Visualization and Computer Graphics **13**(6) (2007) 1624–1631
69. Dietzsch, J., Gehlenborg, N., Nieselt, K.: Mayday-a microarray data analysis workbench. Bioinformatics **22**(8) (2006) 1010–1012
70. Seo, J., Shneiderman, B.: Interactively exploring hierarchical clustering results. IEEE Computer **35**(7) (2002) 80–86
71. Guo, Z., Ward, M.O., Rundensteiner, E.A.: Model space visualization for multivariate linear trend discovery. In: Proc. IEEE Symp. Visual Analytics Science and Technology VAST 2009. (2009) 75–82
72. Kandogan, E.: Just-in-time annotation of clusters, outliers, and trends in point-based data visualizations. In: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, IEEE (2012) 73–82
73. Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., Andrienko, G.: Visually driven analysis of movement data by progressive clustering. Information Visualization **7**(3) (2008) 225–239
74. Schreck, T., Bernard, J., Tekusova, T., Kohlhammer, J.: Visual cluster analysis of trajectory data with interactive Kohonen Maps. In: IEEE Symposium on Visual Analytics Science and Technology, 2008. VAST'08. (2008) 3–10
75. Rasmussen, M., Karypis, G.: gCLUTO–An Interactive Clustering, Visualization, and Analysis System., University of Minnesota, Department of Computer Science and Engineering, CSE. Technical report, UMN Technical Report: TR (2004)
76. Ahmed, Z., Weaver, C.: An Adaptive Parameter Space-Filling Algorithm for Highly Interactive Cluster Exploration. In: Procedings of IEEE Symposium on Visual Analytics Science and Technology (VAST). (2012)
77. Rubel, O., Weber, G., Huang, M.Y., Bethel, E., Biggin, M., Fowlkes, C., Luengo Hendriks, C., Keranen, S., Eisen, M., Knowles, D., Malik, J., Hagen, H., Hamann, B.: Integrating data clustering and visualization for the analysis of 3D gene expression data. Computational Biology and Bioinformatics, IEEE/ACM Transactions on **7**(1) (2010) 64 –79
78. Turkay, C., Parulek, J., Reuter, N., Hauser, H.: Interactive visual analysis of temporal cluster structures. Computer Graphics Forum **30**(3) (2011) 711–720
79. Parulek, J., Turkay, C., Reuter, N., Viola, I.: Visual cavity analysis in molecular simulations. BMC Bioinformatics **14**(19) (2013) 1–15

80. Turkay, C., Parulek, J., Reuter, N., Hauser, H.: Integrating cluster formation and cluster evaluation in interactive visual analysis. In: Proceedings of the 27th Spring Conference on Computer Graphics, ACM (2011) 77–86
81. Choo, J., Lee, H., Kihm, J., Park, H.: ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In: Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on, IEEE (2010) 27–34
82. Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A.: Circos: an information aesthetic for comparative genomics. Genome research **19**(9) (2009) 1639–1645
83. Karr, J.R., Sanghvi, J.C., Macklin, D.N., Gutschow, M.V., Jacobs, J.M., Bolival Jr, B., Assad-Garcia, N., Glass, J.I., Covert, M.W.: A whole-cell computational model predicts phenotype from genotype. Cell **150**(2) (2012) 389–401
84. Meyer, M., Munzner, T., Pfister, H.: Mizbee: a multiscale synteny browser. Visualization and Computer Graphics, IEEE Transactions on **15**(6) (2009) 897–904
85. Piringer, H., Berger, W., Krasser, J.: Hypermoval: Interactive visual validation of regression models for real-time simulation. In: Proceedings of the 12th Eurographics / IEEE - VGTC Conference on Visualization. EuroVis'10, Aire-la-Ville, Switzerland, Switzerland, Eurographics Association (2010) 983–992
86. Muhlbacher, T., Piringer, H.: A partition-based framework for building and validating regression models. Visualization and Computer Graphics, IEEE Transactions on **19**(12) (2013) 1962–1971
87. Booshehrian, M., Möller, T., Peterman, R.M., Munzner, T.: Vismon: Facilitating analysis of trade-offs, uncertainty, and sensitivity in fisheries management decision making. In: Computer Graphics Forum. Volume 31., Wiley Online Library (2012) 1235–1244
88. Meyer, M., Wong, B., Styczynski, M., Munzner, T., Pfister, H.: Pathline: A tool for comparative functional genomics. In: Computer Graphics Forum. Volume 29., Wiley Online Library (2010) 1043–1052
89. Elmqvist, N., Dragicevic, P., Fekete, J.: Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. Visualization and Computer Graphics, IEEE Transactions on **14**(6) (2008) 1539–1148
90. Yang, J., Ward, M.O., Rundensteiner, E.A., Huang, S.: Visual hierarchical dimension reduction for exploration of high dimensional datasets. In: VISSYM '03: Proceedings of the symposium on Data visualisation 2003, Eurographics Association (2003) 19–28
91. Berger, W., Piringer, H., Filzmoser, P., Gröller, E.: Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. Computer Graphics Forum **30**(3) (2011) 911–920
92. Malik, A., Maciejewski, R., Elmqvist, N., Jang, Y., Ebert, D.S., Huang, W.: A correlative analysis process in a visual analytics environment. In: Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, IEEE (2012) 33–42
93. Turkay, C., Lundervold, A., Lundervold, A., Hauser, H.: Representative factor generation for the interactive visual analysis of high-dimensional data. Visualization and Computer Graphics, IEEE Transactions on **18**(12) (2012) 2621–2630
94. Mirkin, B.: Core Concepts in Data Analysis: Summarization, Correlation and Visualization: Summarization, Correlation and Visualization. Springer (2011)
95. Procter, J.B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., Barton, G.J.: Visualization of multiple alignments, phylogenies and gene family evolution. Nature methods **7** (2010) S16–S25

96. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I. In: Visual Data Mining: Effective Exploration ofthe Biological Universe. Springer, Heidelberg, Berlin (2014) in print
97. Mueller, H., Reihs, R., Zatloukal, K., Holzinger, A.: Analysis of biomedical data with multilevel glyphs. BMC Bioinformatics **15**(Suppl 6) (2014) S5
98. Tan, P., Steinbach, M., Kumar, V.: Introduction to data mining. Pearson Addison Wesley Boston (2006)
99. van den Elzen, S., van Wijk, J.J.: Baobabview: Interactive construction and analysis of decision trees. In: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on, IEEE (2011) 151–160
100. Hair, J., Anderson, R.: Multivariate data analysis. Prentice Hall (2010)
101. Secrier, M., Schneider, R.: Visualizing time-related data in biology, a review. Briefings in bioinformatics (2013) bbt021
102. Chen, C.: Top 10 unsolved information visualization problems. Computer Graphics and Applications, IEEE **25**(4) (2005) 12–16
103. Jeanquartier, F., Holzinger, A. In: On Visual Analytics And Evaluation In Cell Physiology: A Case Study. Springer, Heidelberg, Berlin (2013) 495–502
104. Holzinger, A.: Usability engineering methods for software developers. Communications of the ACM **48**(1) (2005) 71–74
105. Kehrer, J., Hauser, H.: Visualization and visual analysis of multifaceted scientific data: A survey. Visualization and Computer Graphics, IEEE Transactions on **19**(3) (2013) 495–513
106. Matkovic, K., Gracanin, D., Jelovic, M., Hauser, H.: Interactive visual steering-rapid visual prototyping of a common rail injection system. Visualization and Computer Graphics, IEEE Transactions on **14**(6) (2008) 1699–1706
107. Beale, R.: Supporting serendipity: Using ambient intelligence to augment user exploration for data mining and web browsing. International Journal of Human-Computer Studies **65**(5) (2007) 421–433
108. Holzinger, A., Kickmeier-Rust, M., Albert, D.: Dynamic media in computer science education; content complexity and learning performance: Is less more? Educational Technology & Society **11**(1) (2008) 279–290
109. Ceglar, A., Roddick, J.F., Calder, P.: Guiding knowledge discovery through interactive data mining. Managing data mining technologies in organizations: techniques and applications (2003) 45–87
110. Chau, D.H., Myers, B., Faulring, A.: What to do when search fails: finding information by association. In: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM (2008) 999–1008
111. Olshausen, B.A., Anderson, C.H., Vanessen, D.C.: A neurobiological model of visual-attention and invariant pattern-recognition based on dynamic routing of information. Journal of Neuroscience **13**(11) (1993) 4700–4719
112. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys (CSUR) **41**(3) (2009) 15
113. Carlsson, G.: Topology and data. Bulletin of the American Mathematical Society **46**(2) (2009) 255–308
114. Edelsbrunner, H., Harer, J.L.: Computational Topology: An Introduction. American Mathematical Society, Providence (RI) (2010)
115. Bremer, P.T., Pascucci, V., Hamann, B. In: Maximizing Adaptivity in Hierarchical Topological Models Using Cancellation Trees. Springer (2009) 1–18

## 2.3. Integrated web visualizations for protein-protein interaction databases

Within this journal publication I present results from studying visualization features provided by PPI analysis tools that are freely available via Internet. By observing that fundamental research in biomedicine in particular in the domain of proteomics needs tools that support identifying information on relations within protein interaction databases i created the idea of providing a review of existing visualization features. By that I started with supervising two computer science studies to set up a first comparison table. Together with my sister, a domain expert in molecular biomedical science, we validated first results and further extended the table. We present a comprehensive table of PPI databases and describe evaluation results of a sub-group of the identfied tools that are both openly available as well as provide visualization features.

Jeanquartier *et al. BMC Bioinformatics* (2015) 16:195
DOI 10.1186/s12859-015-0615-z

BMC Bioinformatics

RESEARCH ARTICLE Open Access

# Integrated web visualizations for protein-protein interaction databases

Fleur Jeanquartier[1*], Claire Jean-Quartier[1] and Andreas Holzinger[1,2]

## Abstract

**Background:** Understanding living systems is crucial for curing diseases. To achieve this task we have to understand biological networks based on protein-protein interactions. Bioinformatics has come up with a great amount of databases and tools that support analysts in exploring protein-protein interactions on an integrated level for knowledge discovery. They provide predictions and correlations, indicate possibilities for future experimental research and fill the gaps to complete the picture of biochemical processes. There are numerous and huge databases of protein-protein interactions used to gain insights into answering some of the many questions of systems biology. Many computational resources integrate interaction data with additional information on molecular background. However, the vast number of diverse Bioinformatics resources poses an obstacle to the goal of understanding. We present a survey of databases that enable the visual analysis of protein networks.

**Results:** We selected M= 10 out of N= 53 resources supporting visualization, and we tested against the following set of criteria: interoperability, data integration, quantity of possible interactions, data visualization quality and data coverage. The study reveals differences in usability, visualization features and quality as well as the quantity of interactions. StringDB is the recommended first choice. CPDB presents a comprehensive dataset and IntAct lets the user change the network layout. A comprehensive comparison table is available via web. The supplementary table can be accessed on `http://tinyurl.com/PPI-DB-Comparison-2015`.

**Conclusions:** Only some web resources featuring graph visualization can be successfully applied to interactive visual analysis of protein-protein interaction. Study results underline the necessity for further enhancements of visualization integration in biochemical analysis tools. Identified challenges are data comprehensiveness, confidence, interactive feature and visualization maturing.

**Keywords:** Visualization, Visual analysis, Network visualization, Protein-protein interaction, Systems biology

## Introduction and Motivation

Both, wet and dry scientists in the domains of Bioinformatics and Life Sciences have to deal with huge amounts of data on protein-protein interactions (PPIs) to understand human life. They have to rely on comprehensive data from web resources. Getting an overview is crucial. Visualization supports this complex task. There are numerous web resources and databases. But assessments of individual strengths and weaknesses of the available resources are scarce. In this paper, we evaluate identified resources in regard to the support of integrated visualization and highlight promising examples. To our knowledge there is no such up-to-date comparative study.

Proteins are the building blocks of life. Interactions between proteins determine cellular communication. Signal transduction cascades process information of various stimuli for a cell to respond to external signals. Cell signaling is based on molecular circuits consisting of receptor proteins, kinases, primary and secondary messengers. Together, they modulate gene transcription or the activity of other proteins [1].

Studies on these complex interaction networks give insight into life-determining processes and can be used for combating disease. Therefore, large datasets are used that contain information on PPIs gained from experiments

*Correspondence: f.jeanquartier@hci-kdd.org
[1] Research Unit HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria
Full list of author information is available at the end of the article

using yeast two-hybrid systems as well as affinity-bait systems [2]. Computational tools for uncovering PPIs are based on the comparison of large-scale experiments, literature curation, text-mining and computational prediction results of protein interactions. These tools are available to the public via online databases [3]. There are numerous software tools and huge databases of PPIs used to gain new insights into systems biology. While many Bioinformatics resources integrate interaction data with other types of information, visualization plays a major role in the process of understanding and sense-making [4–6].

In the last decade, experts started to integrate possibilities for visualization of PPI networks to facilitate exploration and analysis tasks. Visualizations of interaction networks are mostly rendered graphs providing an overall picture of pathways mapping biological functions [7–10].

Some of the many available resources lack maintenance and input of updates. Most of all, they lack usability [4, 5, 11]. The question remains: Which tool is the best choice for the analysis task at hand? Many analysts in the field of Biochemistry manually mine text. They try to find information on related studies and search for appropriate tools. Many researchers do not know which resources are available and which one is best suited to support their analysis. From a computer science perspective there are many possibilities to facilitate the analysis process, particularly making use of visualization features to fully exploit the human capabilities of information processing and pattern perception [12]. To support analysts in Biochemistry it is crucial to pick the right tool for the task at hand [6, 11]. We, therefore, highlight a small set of tools, available on the web, that integrate auxiliary visualization features. The study focuses on web page integrated visualization software that uses the most common technologies supported by current standard web browsers. Online solutions offer fast and easy utilization characteristics compared to client standalone tools. By making use of web visualization tools we overcome issues with standalone solutions including the complicated task of finding and installing third-party solutions, appropriate plugins, difficulties in retrieving biological data, finding appropriate information when searching in default databases that are too generic within local standalone solutions, lack of central storage, interchange and collaboration possibilities [10, 13]. Web visualization represents a field of research on its own finding solutions for limitations in speed, interoperability and navigation. Hence, interdisciplinary scientists improve Bioinformatics databases and tools by adding biological content as well as integrating pervasive web applications featuring graph-based information representation. Interaction and export options are integrated into online tools for further processing of graphs with standalone tools including Cytoscape or Navigator for high computing analysis tasks [9, 10, 14–16]. Standalone tools offer the possibility of individual upgrades in form of add-ons and plugins, numerously available online. Changes to web tools have to be implemented by the provider. Computing power and capacity constitute limiting factors for both web and standalone products. Cytoscape represents a software, most commonly used by bioinformaticians. Still, covering this topic goes beyond the scope of this work. We focus on software that can be easily accessed and used by all experimentalists who deal with PPI analysis. We focus on web software, that neither requires any particular system, nor any root rights, any user's knowledge of system administration or how to install a particular software.

We start with giving some background on visualization in PPI analysis. Then present the comparison study and summarize comparison results of identified tools that suite the task of interactive visual analysis. At last we present its' discussion and identified challenges.

## Background

The human genome contains over 20000 protein-coding genes, while the total number of different proteins is still unknown and estimated to be much higher [17, 18]. Comprehensive knowledge of protein interactions represents the key to understanding the underlying functional network. The molecular organization can be visualized as a network of differentially connected nodes. Each node stands for a protein and edges represent dynamic interactions. Nodes thereby receive input and output values as mathematical functions [19].

Computational results can be analyzed by interactive visualizations. The integrated process of Visual Analytics is essential to sensemaking in Life Sciences. Analyzing a problem in a visual way allows to highlight certain features that are not perceptible otherwise [4, 5, 11, 12].

There are several tools for PPI visualization that not only deal with the general questions of PPI analysis but focus on structural analysis of particular protein domains and peptide sequences (e.g. PDB that archives a large amount of macromolecular structural data that can be visualized). Furthermore, many resources are domain specific and do not support the analysis of the entire interactome (e.g. "NIA", a Mouse PPI Database, or PFAM, a collection of protein domains). The interactome incorporates proteins as well as other chemical compounds as ions, nucleic acids, in sum all interacting elements. In this work, we focus on general resources for PPI analysis that integrate tools for visualizing parts of the human proteinogenic interactome as PPI network.

Graph drawing represents the traditional way of visualizing interactions. Graph visualizations constitute a well-known, sophisticated method in computer science [14]. There are many different well-established and evaluated

layout algorithms for node arrangement in graphs. Force-directed layouts are the main algorithms used for graph drawing. As a result related nodes are placed closer to each other, and highly connected protein interactors as well as clusters of interactors are easily identifiable. Current network visualization resources make use of visualization libraries. One example is the Flash version of Cytoscape [20], that is used n the tool IntAct [21] among others. Additionally, JavaScript (JS) based visualization libraries are currently emerging, including BioJs [22], that is used in PINV [23]. Cytoscape.js is a successor of Cytoscape Web and there is also a wrapper for using cytoscape in BioJs [22].

However, there are several issues and open problems when visualizing biological networks [24, 25]. Nodes are connected through edges representing underlying interactions and should provide interactivity for supporting exploration [26]. Standalone tools like Gephi, Navigator or Cytoscape include various modifications and settings for such purposes. In case of (web-based) graph rendering there are several challenges regarding the handling of large graphs, when dealing with high levels of details and interaction features [16, 26, 27].

Figure 1 summarizes the visual analysis process. Current available biological databases contain huge quantities of different proteomic data that are used by tools to support the analysis process [3, 28]. Droit et al. [29] present an overview of different experimental and Bioinformatics methods to elucidate PPIs. Ben-Hur et al. [30] present computational approaches for prediction of PPIs to help experimentalists in the search for novel interactions. Mosca et al. [31] describe necessary steps towards a complete map of all human PPIs and list a set of currently available methods and resources for PPI analysis. There are several reviews and meta-databases of currently available interaction databases and tutorials on analyzing interaction data including [32–36], but none of these summaries depicts visualization features. Mora et al. [37] presents an analysis of some currently available software tools for PPI network visualization. However,

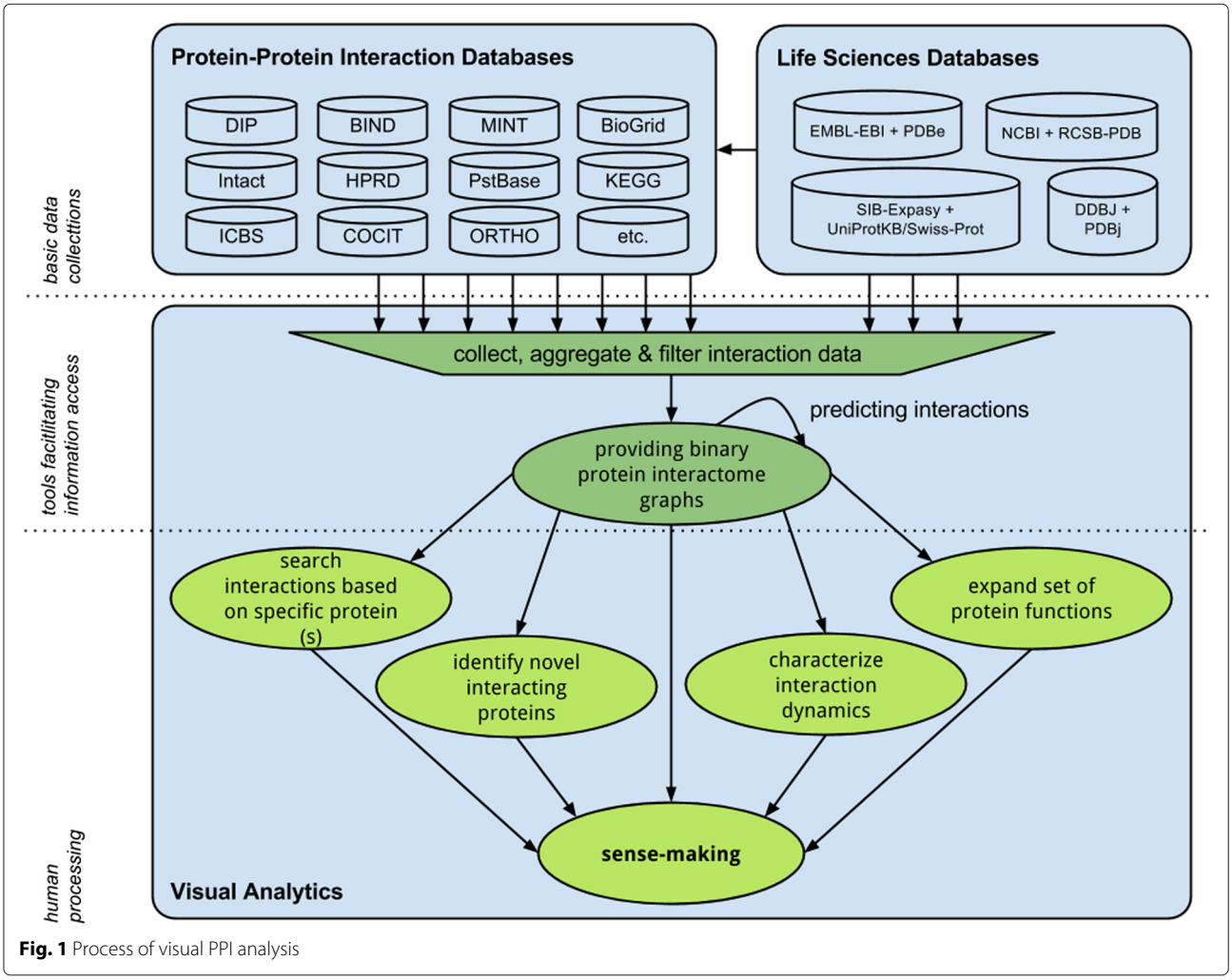


**Fig. 1** Process of visual PPI analysis

the authors only focus on standalone software tools and do not include the analysis of web-based tools. Oveland et al. [38] review different proteomics software and depict exemplified visualization features for a wide range of proteomics data. The authors give a broad overview, but neither focus on PPI network analysis, nor provide a comprehensive overview of online available resources. There are also works that describe how to visualize protein interactions in three-dimensional space [39–42]. Regarding efficiency and effectiveness there are already some ongoing evaluations and efforts [4, 11, 15]. Several works also emphasize the importance of collaboration between computer science and biology [11]. For instance, PPI analysts would benefit from deepening studies not only in organizing and processing data, but also in text mining for protein function prediction as well as for enriching and combining different data and tools for extending association networks etc.

Computational systems biology assesses biological networks to analyze and visualize their complex connections computationally at a system-wide level [43]. In silico models have the purpose of replacing costly and time-consuming experiments with reconstruction and prediction by integration of the vast amount of biological information into multiscale computational modeling [44]. Modeling cellular networks in the context of physiological processes as well as diseases, including proteins as their major effectors, remains an exciting, open-ended domain [45]. Filling the gaps of missing data input by addition of literature-curated functional protein annotations poses a major task. Text-mining tools should help to analyze the overwhelming amount of literature [46]. Still, in regard to reliability and universality, tools require continuous improvements, for instance recognition of variable nomenclature and the implementation of ortholog-based annotations from conserved protein interaction graphs [47]. Biological management systems aim to provide user-friendly work-flows, shared to scientists, with integrated real-time visualization [5, 48].

To our knowledge there is no up-to-date comparative study of current tools that facilitate the interactive visual analysis of protein systems.

## Methods

We compare web-based resources for PPI analysis. 4 analysts take part in the evaluation. The interdisciplinary team consists of 3 domain experts from Computer Science and 1 from Biochemistry. 2 of the analysts are mentioned in the Acknowledgments. The other domain experts are the first 2 authors of this manuscript. We test the Bioinformatics resources by examining search user interfaces as well as visualization abilities. A checklist is completed during the test that includes qualitative meta-data and notes on usage. Additionally, several quantitative parameters are evaluated such as the number of links to different PPI sources, the total amount of PPIs, the number of search results for the specific query and other data if available.

We conduct a search for the "G Protein-Coupled Receptor Associated Sorting Protein 1" (GPRASP1), also known as "gasp1" with its UniProt ID "Q5JY77". The example protein is chosen as input determinant due to its known involvement in G-protein coupled receptor (GPCR) signaling which constitutes a major cellular signal transduction cascade [49]. The cytosolic protein GPRASP1 is a validated tumor marker and, therefore, associated with cancer.[50]. Thus, we review the availability of information on disease associations. Additionally, we test for a set of proteins including GPRASP1 plus some of its putative interaction partners, namely cannabinoid 1 receptor CNR1 (P21554), calcitonin receptor CALCR (P30988), dopamine D2 receptor D2DR (P14416), bradykinin 1 receptor BDKRB1 (P46663) [49]. Results on the PPI searches regarding a single and multi-protein input are listed in Table 2.

We examine the presentation of results as well as visualization and interaction features. Quantitative and qualitative characteristics as well as notes are collected within spreadsheets. The results are summarized in a comprehensive comparison table (see link http://tinyurl.com/PPI-DB-Comparison-2015).

### Comparison Criteria

Evaluations of visualization tools have to be prepared carefully. It is essential to choose an appropriate baseline for comparison and metrics by evaluating efficiency, effectiveness, visualization quality and insights. There are quantifiable factors such as speed (e.g. task performance), accuracy, latency, number of results, or insights. Additionally, there are standards for measuring qualitative factors that are currently used for the evaluation of research in clinical data visualization [51–54]. Some of these criteria are taken into account and are summarized for comparison. The review focuses on the following 5 criteria:

- **Support of Multi-Platform:** Nowadays research is conducted on miscellaneous devices, several operating systems and various browsers. Therefore, it is necessary to assess the requirements of a particular tool. Javascript and SVG are generally slower than Java applets or proprietary browser plugins such as Flash or Silverlight [55, 56]. None of the tested tools makes use of Silverlight at the frontend. Although Javascript often has shown performance problems in past, browser performance is rapidly evolving. Therefore, Javascript and SVG solutions can be used for graph rendering [20, 56–58].

Next to a modern browser, end users often need to install plugins, including fFash. Java applets often need additional adjustments to the client's security settings. Thus, Java applets but also Flash frontends (regardless whether based on Java or not) may pose a hurdle in making use of a visualization tool. Thus, Javascript and SVG visualization get the highest score for evaluating this criteria.

- **Service in General:** Determines the quality of the user interface (UI) in general. The UI determines the simplicity and efficiency of the search and its visualization characteristics.
- **Interoperability (Import, Export, Formats, Plugins):** Summarizes a tool's network export options (e.g. textual, graphics, individual format), it's interaction possibilities, manual import or similar options. This is particularly crucial when starting an analysis with one specific tool or one specific platform but continuing with another one.
- **Visualization Quality (Speed, Clarity, Usability):** Describes the visualization itself. Main focus lies on speed, clarity, and ease to use. This section also identifies items for possible improvement. In Fig. 2 all network views are compared to each other visually.
- **Visualization Features:** There are interactive visualization features that are crucial to exploration interfaces [12]. This section examines and lists available features like drag-and-drop, move background, area-selection a.o.
- **Data Coverage:** Represents the number of hits from the single and multi-protein search for PPIs as well as further information on associated diseases.

Each of the ten identified PPI web resources are tested against these criteria and the extent to which requirements are met for supporting the interactive visual analysis of PPI networks is evaluated. The evaluation summary comprises quantitative results such as the number of linked databases as well as the number of interactions found. Evaluation results also include last updates as important factor of comprehensiveness.

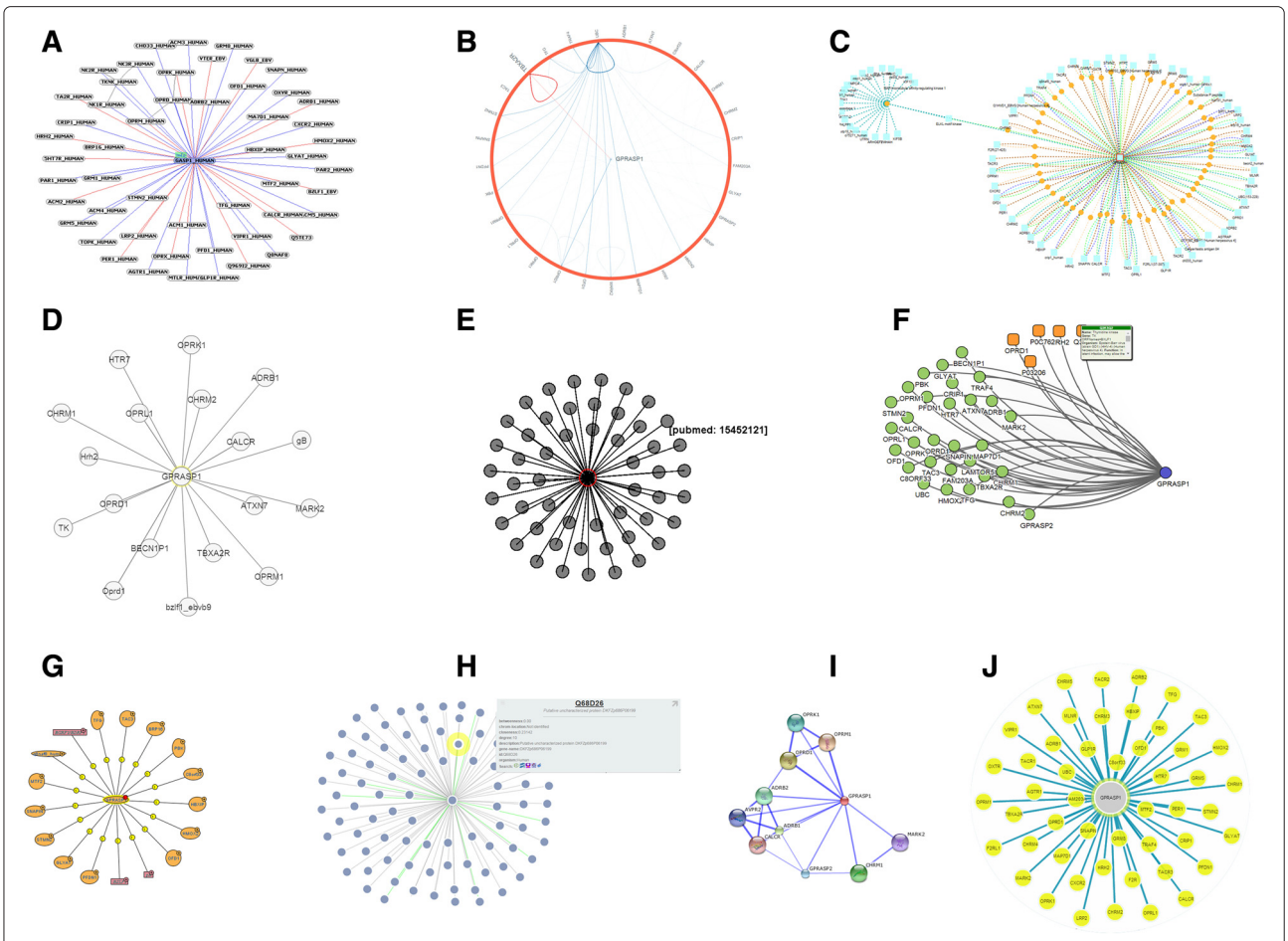


**Fig. 2** Graphical Comparison of all tools showing interactions networks for Q5Yj77: [A] APID, [B] Biogrid, [C] CPDB, [D] IntAct, [E] I2D, [F] Mentha, [G] MINT, [H] PINV, [I] String, [J] UniHI

## Results

We specifically describe the most promising web resources. The visualization features of the selected resources are summarized in Table 1. Quantitative results are summarized in Table 2. We conclude with highlighting the top rated three resources that integrate the most promising interactive visualization features as well as integrate data comprehensively.

The identified resources are: Agile Protein Interaction DataAnalyzer (APID) [59], BioGrid [10], ConsensusPathDB (CPDB) [60], IntAct - Molecular Interaction Database [21, 61], Interologous Interaction Database (I2D) [62], Mentha - The Interactome Browser [63], Molecular INTeraction database (MINT) [64], or more specific its' separate annotation of human PPIs called HomoMINT [65], Protein Interaction Network Visualizer (PINV) [23], StringDB - Search Tool for the Retrieval of Interacting Genes/Proteins [66] and Unified Human Interactome (UniHI) [67].

### Agile Protein Interaction DataAnalyzer (APID)

**Support of Multi-Platform:** APID allows a protein's interactions to be visualized as graph within a separate Java applet called ApinBrowser. Due to the usage of an embedded Java applet, the tool itself is multi-platform ready.

**Service in General:** APID allows queries of several input names. Results are presented in a concise way. Clicking on the number of interactions presents a more detailed overview of the PPIs including the number of experiments and information on sources of the various interactions. By clicking on the 'graph' labeled button the Java applets are loaded into a separate window.

**Interoperatibility:** The tabular data can be exported. The graph itself can be stored as an image. Import possibilities are limited to searches throughout linked databases. The creators also provide a Cytoscape plugin for APID called APID2NET.

**Visualization Quality:** The visualization is dynamic and makes use of a simple force-based layout for graph drawing. It lacks anti-aliasing and other modern rendering techniques for visualization.

**Visualization Features:** APinBrowser provides options for zoom, filter and limiting details on demand. There are minor adjusting possibilities such as background color and edge thickness. Still, this resource lacks several features as visual clustering or highlighting certain nodes and edges.

**Data Coverage:** A single protein query quickly returns a mid-range number of interactions. Unfortunately, there is no direct option to include more than one protein name or ID into the search. However, after searching for one protein and visualizing the graph, it is possible to add additional proteins by using the "add" and "import" functionality within the applet. By further clicking on paint the additional proteins are included into the graph visualization. Associations to diseases are not available.

**Evaluation Summary:** The user interface of queries includes a concise tabular overview of results. Yet, antialiasing and options for adjusting nodes are missing. The web resource itself might be outdated due to the fact that last updates have been added in 2006.

### BioGrid

**Support of Multi-Platform:** This Bioinformatics resource can be opened in all current browsers. Therefore, installation of a specific plugin is not required.

**Service in General:** Biogrid provides a simple search option offering a quick glance on results in addition to filter and sorting features. The presentation of the results shows basic information.

**Table 1** Summary of identified PPI resources' visualization control features

| Tool ID/ Control Feature | Apid | BioGrid | CPDB | IntAct | I2D | Mentha | Mint | Pinv | String | UniHI |
|---|---|---|---|---|---|---|---|---|---|---|
| Zoom | y | - | y | y | y | y | - | y | y | y |
| Select neighbors | - | - | y | y | y | y | - | - | - | - |
| Toggle labels | y | - | y | y | y | y | - | y | - | - |
| Fix/Unfix | - | - | - | - | y | y | y | y | y | - |
| Shrink/Grow | - | - | - | - | - | y | y | - | - | - |
| Toggle node shape | - | - | - | - | y | - | - | - | - | - |
| Select hubs | y | y | - | y | y | y | y | y | - | y |
| Select tree | - | y | - | - | - | y | - | - | - | - |
| Fit to screen | y | - | y | - | - | - | - | - | - | y |
| Clustering | - | - | y | - | - | - | - | y | y | y |
| Expand network | y | - | y | y | - | y | y | y | y | - |

**Table 2** Summary of the quantitative results concerning data integration

| Tool ID/ Quantity aspect | Apid | BioGrid | CPDB | IntAct | I2D | Mentha | (Homo)Mint | Pinv | String | UniHI |
|---|---|---|---|---|---|---|---|---|---|---|
| binary interactions of Q5JY77 | 52 | 35 | 149, (60 distinct) | 22 | 53 | 35 | 17 | 95 | 201 (default 37) | 50 |
| max. PPIs | 322 579 | 543 666 | 368 654 | 473 426 | 1 539 758 | 480 517 | 330 377 (Mint) | n/a | 332 235 675 | 374 833 |
| human PPIs | 83 670 | 173 728 | 221 328 | 154 338 | 318 717 | 157 932 | 241 458 (HomoMint) | 2 942 636 | 942 636 | n/a |
| predicted PPIs | 44 040 | n/a | n/a | n/a | 635 488 | n/a | 6 782 | n/a | n/a | n/a |
| experimental PPIs | 278 539 | n/a | n/a | n/a | 922 617 | n/a | 323 595 | n/a | n/a | n/a |
| group PPIs (Q5JY77, P21t4, P30988, P14416, P46663) | 91 | n/a | 4192 | 818 | 106 | 67 | 93 | 1894 | 470, 2 internal | 284 |
| disease associations | n/a | n/a | n/a | 0-2 | n/a | 0 | n/a | n/a | 13 | 0 |
| links to DBs | 29 | 12 | 32 | 27 | 29 | 6 | 6 | 1 | 23 | 15 |

**Interoperatibility:** The visualized graph can not be be exported. It can be downloaded as a simple textual list only. Additional download options can be found outside of the visualization view. However, a specific graph format for Cytoscape or similar tools is not included.

**Visualization Quality:** The button for opening the graphical viewer is placed non-intuitively. The graph view loads quickly and does not require any plugin by making use of a modern circular layout that can be seen in Fig. 3. The radial view is not as intuitive as traditional graph presentations and the small labels are hard to read. Still, additional information is found quickly during the exploration process. There are no interactive features connected to the graph's edges. By selecting a node, edges connected to this node are highlighted. During this process, the font size of the interacting nodes increases, that results in overlapping neighbors, rendering the text hardly readable. In terms of usability, the graph visualization provides features for basic analysis. Settings to adjust color and shape are missing.

**Visualization Features:** The visualization is static. The use of filtering options or other features forces the page to reload, which requires some computational time. Only exceptions are some hover effects. Rearrangement can be accomplished by clicking on a node. There are some features as highlighting, searching, filtering by the use of check-boxes and a field for input of text. Details are shown on mouse-over, also indicating the connected partners. Additional mouse-over details are options to search/follow interactions and download interaction data as text file. However, the visualization lacks zooming and scaling options.

**Data Coverage:** The single-protein query resulted in a low to mid-range number of interactions. Input options for a multi-protein search are not available, neither is information on disease associations.

**Evaluation Summary:** BioGrid supports visual analysis in a limited way.

#### ConsensusPathDB (CPDB)

**Support of Multi-Platform:** Dynamic rendering of SVG visualization is possible in all modern browsers.

**Service in General:** CPDB offers an intuitive search combined with short computational loading times for the presentation of results. In addition, mapping criteria for filtering makes this resource a supportive PPI analysis tool.

**Interoperatibility:** CPDB is supported by only a small number of institutions unlike the other resources. Yet, it makes use of most important databases and offers features such as manual upload.

**Visualization Quality:** The network's SVG based visualization is not as fancy as modern Flash based frontend presentations. Nevertheless, it already integrates anti-aliasing and interactiveness. CPDB provides many possibilities and includes many information sources. The graphs are largely and densely packed due to automatic stretching. The thickness of nodes does not correlate to the amount of visualized nodes. Their scale correlates with the zoom level, thus, the visualization becomes hard to read at a high zoom-level. The utilization of different colors and shapes facilitates a distinction between specific interaction- and node-types.

**Visualization Features:** Filter functions are not integrated into the visualization but have to be defined before mapping of interactions. The resource provides several criteria for mapping such as choosing particular databases to be integrated into the results. The dataset is visualized

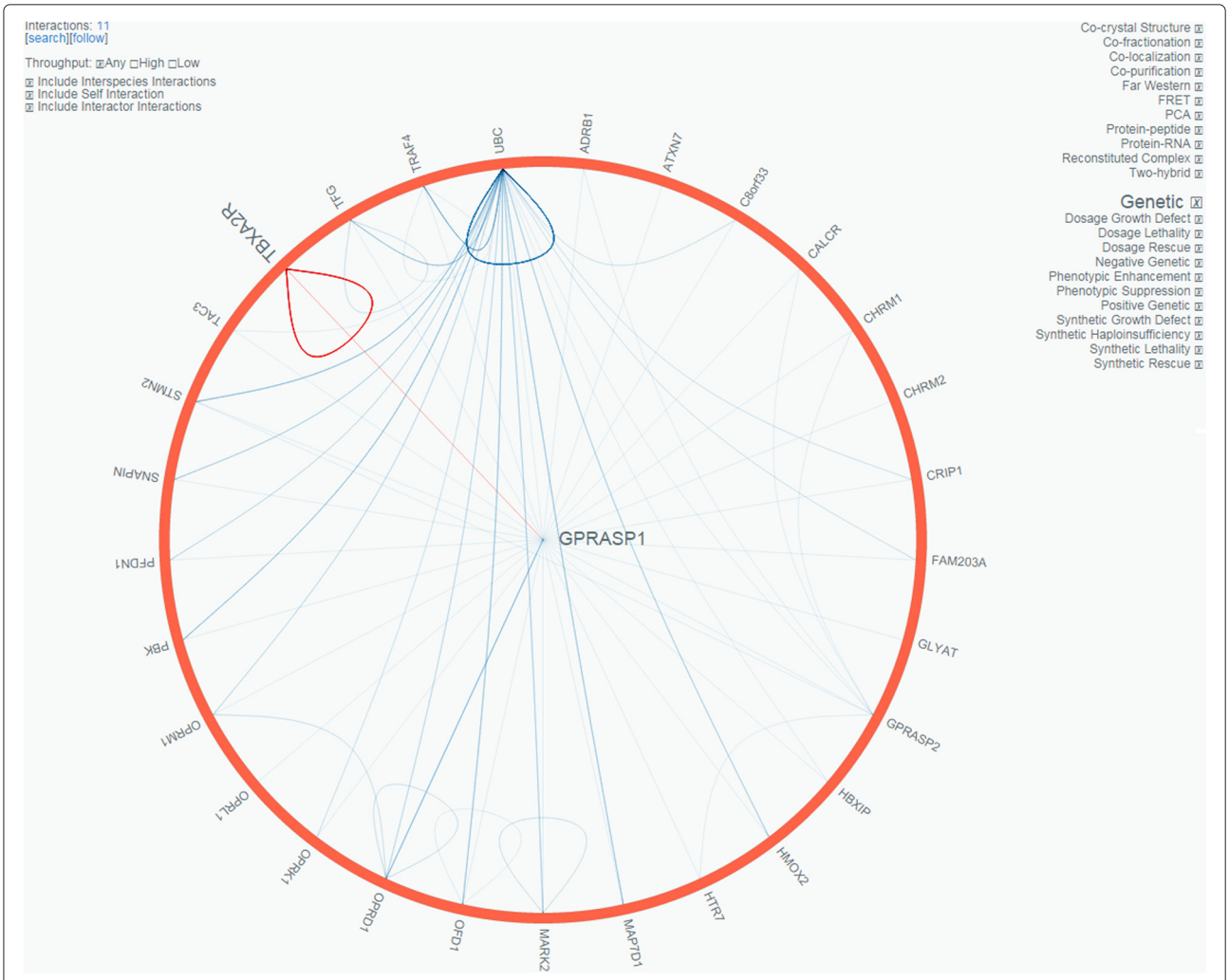


**Fig. 3** Screenshot of Biogrid's graph view

comprehensively. Additional information on nodes are shown by hovering and clicking on them. The network view makes use of zoom and repositioning options as well as color and shape differences of nodes and edges for highlighting certain attributes. The characters of shape and color are described in a concise and informative way within a legend. Edges can be merged and demerged. Network statistics can be retrieved and there is also a search option within the graph.

**Data Coverage:** CPDB shows the highest number of possible hits for both the single and multi-protein search. Information on associated diseases are not implemented.

**Evaluation Summary:** CPDB holds the key benefit for supporting exploration by making use of PPI data obtained from literature curation, computational text-mining, orthology-based prediction as well as manual upload. Figure 4 presents a CPDB graph including interaction data, integrated in a merged manner. The developers try to avoid redundancies, still, the network visualization shows much more protein interactions compared to the other tools examined. On the one hand, CPDB's graph presentation encourages exploration. On the other hand, there are difficulties of getting an overview.

### IntAct - Molecular Interaction Database

**Support of Multi-Platform:** The graph visualization is implemented via Flash. Flash has multi-platform support and is usable in all modern web browsers with installed Flash plugin.

**Service in General:** The search function is simple and intuitive. No preselection of attributes is necessary. Search results are presented as set of several subcategories.

**Interoperatibility:** PPI data within search results can be exported as tabular text. Additionally, the user can export export a network to the format of Cytoscape for further analysis and manipulation in the standalone tool.

**Visualization Quality:** The layout can be changed between force directed, radial and circular views. IntAct

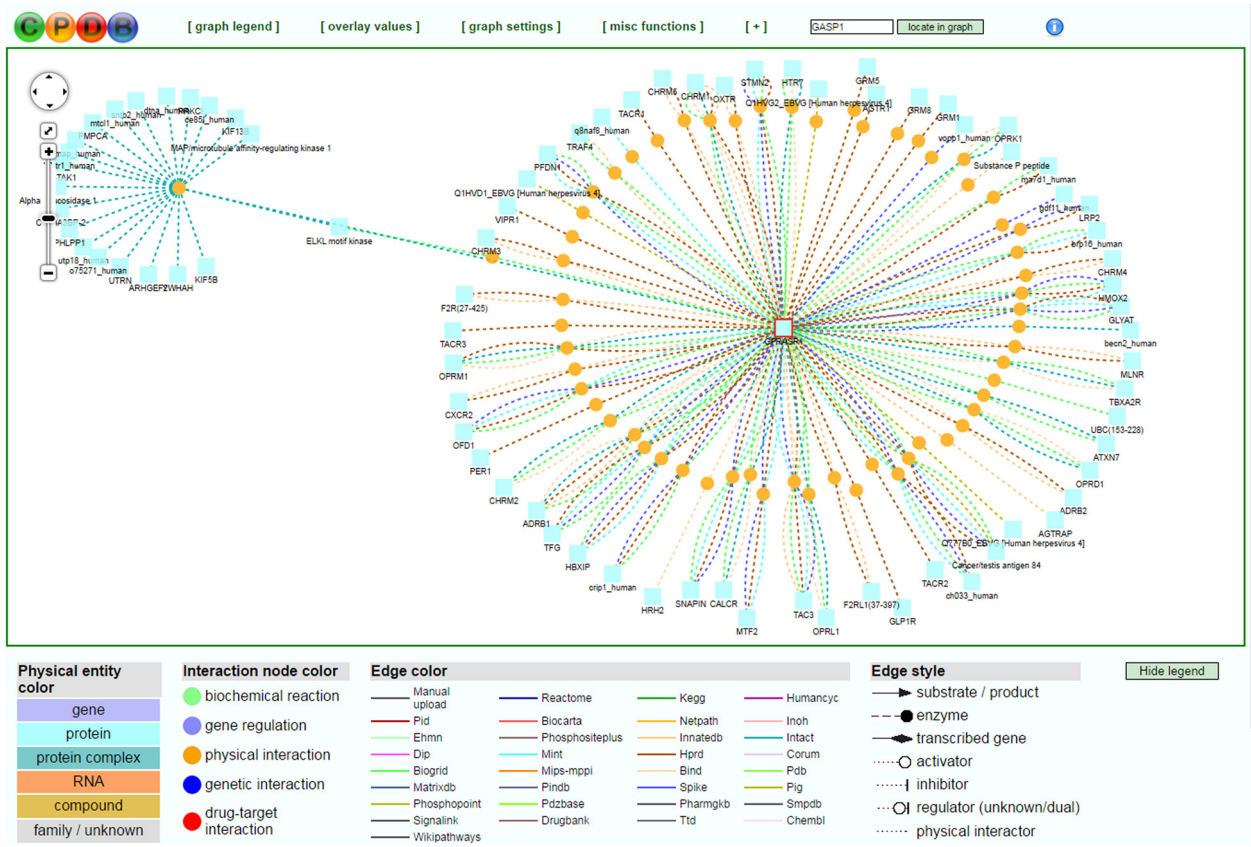


**Fig. 4** Screenshot of CPDB's UI of interaction mapping and visualization

offers additional features as merging/splitting groups of nodes and zooming with modern anti-aliasing. However, IntAct lacks options for adjusting color and shape. There is a clear need for visual clustering, since every node looks the same. Titles of nodes are too large and occupy more area than the nodes themselves. Nodes overlap edges even in small graphs.

**Visualization Features:** There are several features as simple zoom and repositioning. Limited details are shown on demand by clicking on a node. The graph layout can be interactively adjusted. The user can switch between the list and the graph tab. Edges can be merged and demerged. Specific interactions can be filtered. Yet, there is no integration of detailed variations and highlighting specific variables.

**Data Coverage:** The single-protein query returns the low number of 22 possible PPIs, in case of protein ID as input, or 23 possible interactions in case of name abbreviation. IntAct presents one of the highest number in PPIs for the protein-group query. The feature of connecting to further EMBL-EBI resources reports associations of diseases in case of abbreviated name query.

**Evaluation Summary:** IntAct is supported by EBI and updated regularly. The integrated Flash based graph provides different export options including a translation to Cytoscape. However, the integrated visualization lacks important features such as filtering, adjustment of color and shape attributes.

### Interologous Interaction Database (I2D)

**Support of Multi-Platform:** I2D's graph viewer needs Java installed and activated.

**Service in General:** The search option does not provide any auto-suggest and correction suggestions. The user has to search precisely. Other resources include such features. The table of results is very limited in information content, which only links to other meta-information on different platforms. No filter or sorting options are provided. It would be helpful to know the type of interaction at first sight.

**Interoperatibility:** There is only one possibility of inter-operating, as the graph can be exported as tabular text.

**Visualization Quality:** Due to the usage of an old fashioned Java applet the visualization lacks anti-aliasing and

visualization quality. Nodes are covered by edges also in graphs with low numbers of nodes and edges. Rescaling options are missing.

**Visualization Features:** There are many hidden features that require parallel or cumulative actions with multiple input devices. A legend on key usage can be found on the right side within the network view. The legend is large and one example to the non-intuitive visualization approach.

**Data Coverage:** I2D presents a mid-range number of possible interactions for the single and multi-protein search. An option for disease association was not available.

**Evaluation Summary:** This resource links to many databases and therefore steadily expands its comprehensiveness. Still, the tool itself does not facilitate the process of visual analysis due to the outdated visualization integration.

#### Mentha

**Support of Multi-Platform:** Mentha's so called 'interactome browser' is implemented by Java. A newer but also limited SVG version is additionally provided as an alternative to Java.

**Service in General:** This Bioinformatics resource offers an intuitive search field but a less intuitive presentation of the results. The 'browse' button starts the network view. The 'list' button itemizes interaction results and meta-information.

**Interoperatibility:** The new version does not provide export or import. The Java version supports export as textual tabular data and png graphics.

**Visualization Quality:** The SVG version is intuitive but still limited in optional features. Promising updates are already planned.

**Visualization Features:** The dynamic network viewer features zoom, filter details on demand and provides a flexible layout. Moreover, the Java version offers possibilities for coloring and highlighting.

**Data Coverage:** The interactome browser presents a low to mid-range number of possible interactions in case of the single-protein search and the lowest count in PPIs using the multi-protein input. Results can be easily filtered by confidence for a fast overview. The list is supplemented with meta-information from e.g. KEGG database and could offer associations to diseases but without any results from the particular evaluated search.

**Evaluation Summary:** There are several differences between the old and new visualization that are being integrated into Mentha. One comes with better compliance to the browser, the other one offers a higher degree of interaction possibilities. If being combined and steadily updated, the two visualization possibilities would definitely support the sense-making process. Future updates will include further enhancements to the new visualization.

#### Molecular INTeraction database (MINT) / HomoMINT

**Support of Multi-Platform:** (Homo)Mint requires a browser with Java installed.

**Service in General:** The search UI provides a concise overview of results as well as includes an overview of the various databases used.

**Interoperatibility:** No import and export functions are integrated.

**Visualization Quality:** The resource is based on an old Java version does not integrate state of the art rendering techniques such as anti-aliasing. Most important interaction features are offered and performance is sufficient. A graphical legend is missing for a quick glance at means of color or shape.

**Visualization Features** Interaction possibilities include zoom, filter and details on demand. The user can change the size of nodes in order to improve speed and clarity. An adjustable threshold is available for filtering the output and number of displayed nodes. Drag and drop is possible (as in most other Java applets, too). Some features require a long computing time. One example is the option 'connect' on a newly selected node for adding edges to it's neighbors. Others are the MITAB and PSI functions. In this case, there are no notifications to the user. According to Nielsen's response times, feedback should be provided after one second.

**Data Coverage:** Mint shows the lowest number of interactions for the single protein. Only 3 out of 5 proteins from the group input are detected and result into 93 PPIs after connecting the single graphs to each one of them. Information on associated diseases are available showing 3 interacting proteins out of 93 to be involved in pathological processes.

**Evaluation Summary:** Both quantitative (number of databases linked or number of interactions found) and qualitative results (old-fashioned visualization without anti-aliasing) underline the limitations of the Bioinformatics resource MINT. Since it is produced and provided by Uniroma, it is recommended to switch to the newer PPI tool supported by Uniroma: Mentha, which offers new visualization features, not limited to Java anymore.

#### Protein Interaction Network Visualizer (PINV)

**Support of Multi-Platform:** The graph visualization runs in current browsers having Javascript installed and activated.

**Service in General:** The user interface for a query is intuitive. The idea of using the BioJS and D3 framework to create an HTML5 application, as it has been applied to this tool, offers interesting possibilities for supporting

visual analysis online. However, performance limitations for large and dense graphs are still an issue when using the tool more intensely. Feedback often is missing at the right point and interaction possibilities could be smoother.

**Interoperability:** There are several possibilities to exporting the graph, both graphically and as textual tables.

**Visualization Quality:** Due to the increasing prospects of JS, the graph is rendered dynamically as SVG using anti-aliasing. This mode allows the user to interact with nodes and edges including smooth transitions. The default graph layout is a standard force-based view. In addition, PINV offers a circular layout, a heatmap as well as a simple table view.

**Visualization Features:** The tool features several interaction possibilities, foremost zoom, filter and some details on demand. Next to the zoom option there are several possible manipulations to the visualization by defining rules for filtering, highlighting, coloring and options for uploading expression data. The screenshot in Fig. 5 illustrates that exploration is based on the process of defining rules.

**Data Coverage:** A suitable data-set has to be chosen from a list of online available sources before conducting protein search. By choosing the 'human' data-set the single-protein input results into a higher count of 95 PPIs. One of the highest counts of 1894 PPIs follow from the multi-protein input. Further information on disease associations are not available.

**Evaluation Summary:** The visual analysis tool provides features for exploration and sensemaking in a modern fashion. Wizard-like usage and adding rules for manipulation can be recommended for other tools. Performance issues as well as not caught JS errors hinder the task of visual analysis of PPIs.

### StringDB - Search Tool for the Retrieval of Interacting Genes/Proteins

**Support of Multi-Platform:** StringDB's interactive network viewer requires a modern browser including the Flash plugin.

**Service in General:** The query option is simple and includes data from several databases including multiple organisms.

**Interoperability:** The graph can be exported as several file formats, both as graphic and as text.

**Visualization Quality:** Graphs are rendered dynamically as PNG ore implemented as interactive Flash visualization that offers numerous interaction possibilities. In addition to the network view, there are options for simple visualizations such as the occurrence view. Figure 6 illustrates some of StringDB's UI capabilities. Further information as well as structural data are included if available. Details are displayed within the context menus upon clicking on individual nodes.

**Visualization Features:** The resource provides a variation of four different designs, namely confidence,

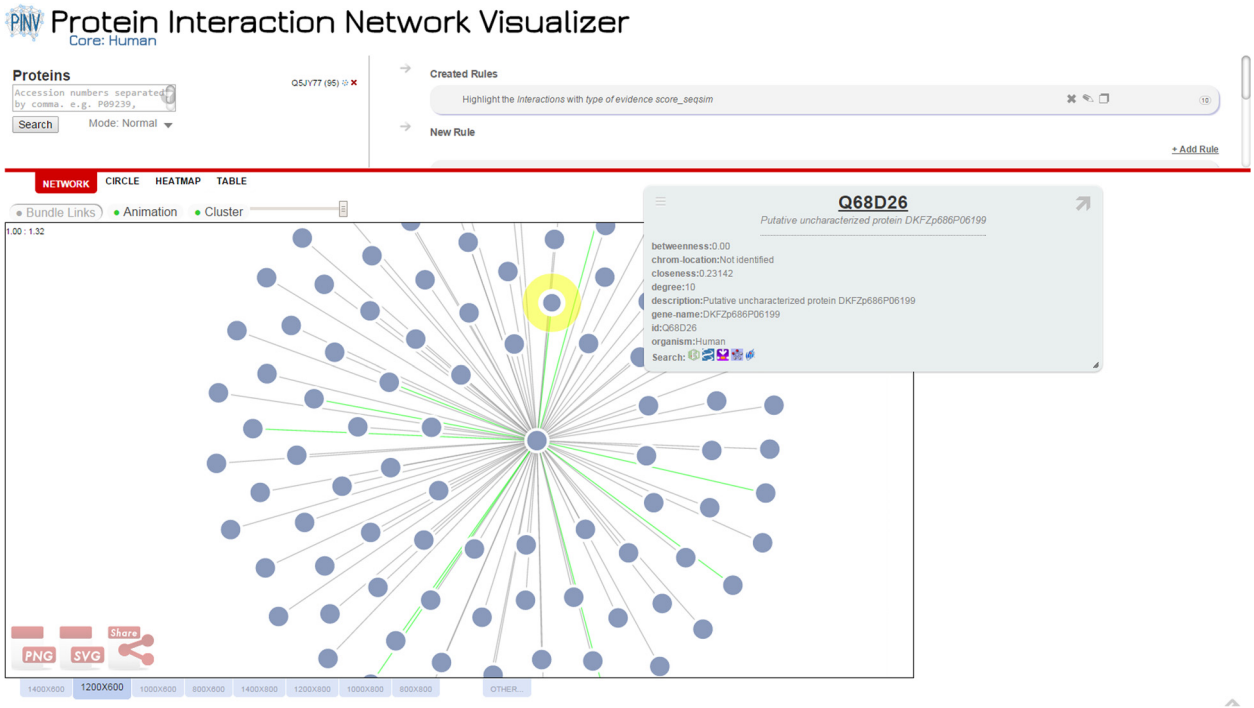


**Fig. 5** Screenshot of PINV UI showing search results for Q5Yj77

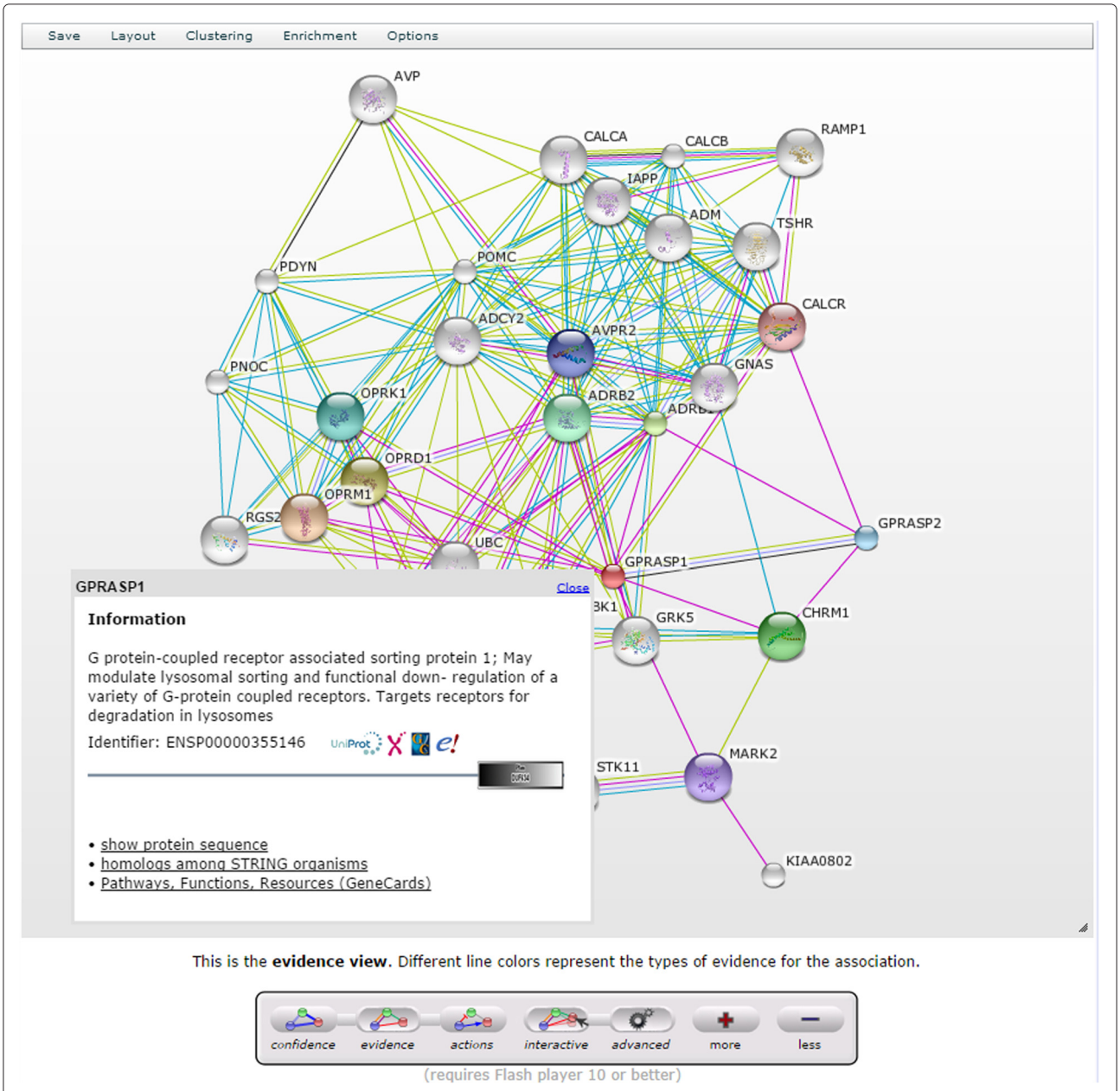


**Fig. 6** Screenshot of STRING UI showing the evidence view

evidence, actions and interactive view. The view can be changed from a simple default to an advanced mode. The interactive view allows the user to adapt the layout. The UI provides many different filter and control features next to simple zoom and scaling functionality. StringDB offers visualization options, such as node/label hide/show, and functional options of clustering or enrichment. The nodes and edges are colored. Node colors represent direct associations but are not adjustable. Line Colors are mapped to types of evidence. Line thickness represents confidence. These presentation presets are not customizable. The view does not allow zoom and is not adjustable in an arbitrary fashion. It provides options to grow and shrink the rendered image.

**Data Coverage:** The single protein query returns a mid-range number of hits, as does the multi-protein query. The default limit of reported interactions is set to 10 and has to be increased accordingly. Possible interactions are easily filtered by confidence. StringDB provides the option to get further information on disease associations. 13 associations are found within the 37 interacting proteins.

**Evaluation Summary:** StringDB combines comprehensiveness with state-of-the-art visualization features. It supports PPI visualization and analysis.

#### Unified Human Interactome (UniHI)

**Support of Multi-Platform:** The graph visualization runs with Adobe Flash.

**Service in General:** The Java-based implementation needs to be improved regarding loading perfomance. The search UI is intuitive and easy to use. Still, tabs cannot be changed easily due to the UI's implementation without hovering effects. UniHI links to several databases as common to most PPI resources. The graph visualization is rendered within the network tab.

**Interoperatibility:** Export options include text files, png and pdf.

**Visualization Quality:** The network visualization makes use of the common Cytoscape Web. This tool provides a modern but also simple Flash interface as frontend. The visualization encloses basic layout and filtering features that are capable of smoothly rendering large graphs. Unfortunately, the graph does not include any visual details. The visualization is rendered within a separate window. Selected or highlighted nodes are indicated by a lighter circle around the node. A separate menu at the right side of the resource includes filter and analysis features. Textual information is hard to read due to its' small font-size. UniHI makes use of basic clustering or enrichment functions. Types of connections are colored differently within the visualization (red and blue). However, version 7.1 lacks functional layout palettes.

**Visualization Features:** The resource includes common control features such as zoom, repositioning and scaling to fit the page. It is possible to filter interactions (e.g. regarding source of interactions or amount of evidence). Details are provided in separate windows by clicking on a node. Analysis options are also provided. There are 'Help' links and a reset button for reconstructing the original graph setup.

**Data Coverage:** The single protein query yields a mid-range count in PPIs as does the multi-protein search. Information on target proteins are received from the KEGG database. In case of our query no implication on pathological associations could be detected.

**Evaluation Summary:** The old Java applet frontend has been upgraded to making use of Cytoscape's Flash version. Yet, the resource does not meet the needs for exploration. Most of all, UniHI lacks performance and often throws irreproducible server errors that force the user to restart the query. Thus, UniHI cannot be recommended to support exploration as a step towards sensemaking.

## Discussion

We conducted an extensive web research and scanned through a list of more than 300 tools for PPI analysis. 53 are available online and suite the basic needs of protein system analysis within the human interactome. Only a small subset of the examined online tools (10 out of 53) offers integrated visualization. Interactive visualization features are summarized in Table 1. Quantitative metrics are summarized in Table 2.

At first glance, the primary goal of a search within web resources is to receive the largest amount of data. We quantified data retrieval by the number of possible interactions with a specific input variable. Therefore, web resources have to integrate data from several databases, and they have to be updated regularly. Ideally, data is obtained from several sources at once including literature curation, computational text-mining and prediction methods. A great amount of data does not equal a great deal of information. The search field and input options have to be easy to use. The user will stop his/her search at the initial stage if query options are not properly presented in the resource. Moreover, the presentation of data is crucial for its interpretation.

An ideal software tool for PPI analysis would possess the following features: At default results should be available as concise overview. Detailed information should become apparent on demand. Options for filtering and adjusting the confidence level are essential for a successful data translation. Graph visualization should be scalable and include features for manipulation. Nodes and edges exemplary should be adjustable in color, shape, size and position. Resources should offer various options to graph export and import. Results should be both complemented and downloadable as tabular text, graphics and also in other standardized file formats used by standalone tools. Above all, Bioinformatics web resources have to provide a modern interface. They have to comply with multi-platform standard browsers avoiding performance issues, outdated proprietary software, annoying software update requests or server errors.

In summary, the ideal web-based Bioinformatics resource features comprehensiveness, an intuitive user interface, as well as a modern visualization.

Each of the evaluated software has its respective strengths and weaknesses:

APID provides intriguing entry points such as a concise overview and a Cytoscape plugin. On the other hand, it lacks state-of-the-art rendering and modern visualization features like visual clustering.

Biogrid would benefit from improvements regarding readability and interactive features. Visualization would be ameliorated by making use of color and shape variations to visualize specific attributes. None of the test users found the option for opening the graphical viewer in Biogrid at first sight. This fact indicates the need for usability improvements.

CPDB presents a comprehensive dataset, while its visualization's overview could be improved.

IntAct features an option for changing the network layout. However, it is only suitable to represent simple networks due to the lack of tagging and additional information.

I2D lacks state-of-the-art visualization quality and an intuitive and effective user interface. I2D's user interface hinders exploration and sense-making.

(Homo)Mint provides interesting interactive visualization features like an adjustable threshold and drag and drop. Unfortunately, a graphical legend on feature description is missing. Some features require long computation times, and visualization quality is not state-of-the-art.

The idea of using JS frameworks such as BioJS and D3 in PINV is promising. However, PINV does not fully comply with the task of visual analysis of PPIs due to occurring performance issues as well as not caught JS errors.

StringDB's presentation presets are not customizable yet. However, StringDB is our first choice of Bioinformatics resources due to its comprehensiveness, the use of confidence scores and state-of-the-art visualization features.

UniHI comes with two versions, a network view based on Java and another one running with Adobe Flash. The Java-based implementation needs to be improved regarding loading performance. Performance limitations are more likely to arise due to issues on server- and not client-side.

Force-directed layout is the main algorithm used in this kind of visualization tools. 2D graphs are the preferred solution for integrated visual analysis of PPI online. None of the tested tools features 3D views.

Only a few resources reasonably support exploration and sense-making. All identified web resources differ from standard graph visualization tools, mostly standalone software. Resources dedicated to PPI analysis also vary from graph analysis applications in other domains like link, social network or market analysis. Differences are observed in visualization quality and interaction possibilities. Therefore, export/import options are commonly implemented.

While conducting the evaluation of several online network visualization tools for PPI analysis we identified the following prominent challenges:

## Challenges

- **Challenge 1:** Current tools vary strongly in terms of comprehensiveness. Thus, it is still a crucial issue to link to all PPI databases available, finding suitable update mechanisms and providing a good overview in the distinct presentation of PPI networks.
- **Challenge 2:** Another only little-touched issue is dealing with confidence levels. Only a few tools provide the possibility to manipulate the graph drawing by adjusting the confidence of the various interactions as well as computing common metrics for graph network analysis. This is not only due to incompleteness of the underlying data used, but also because interactive features for visualization manipulation have long not been point of interest in the tool's development.
- **Challenge 3:** A more general but also clear challenge deals with maturing visualization integration within the Biochemistry domain. There is a clear need to foster usage of modern visualization features such as easily changing layout settings, deleting nodes or adding group annotations, integrating richer possibilities for interactive visual clustering and extending layout palettes. The evaluation also highlights the need to also integrate, next to force-based algorithms, multi-level algorithms to overcome issues of assessing certain differences in networks and providing possibilities for presenting large graphs as both visually appealing and readable.

## Conclusions

The top three rated resources are String, IntAct and CPDB. They integrate graph visualization and can be successfully applied to interactive visual analysis of PPI. We also identified significant differences both in the UI as well as in the amount of hits on PPIs. Web-based resources are best used as starting point in research. Detailed analysis is still more efficient, effective and satisfying by making use of standalone graph visualization tools. This fact clearly reveals the necessity of further enhancing visualization integration in analysis tools in the domain of Biochemistry.

Closing, we encourage greater collaboration amongst the two scientific research fields of Systems Biology and Computer Science regarding visualization techniques.

**Competing interests**
The authors declare that they have no competing interests.

**Authors' contributions**
FJ, CJ and AH devised the idea of a comparison study for supporting the analysis of PPIs. FJ planned and accompanied the whole comparison study. CJ applied her knowledge of Protein biochemistry and took part in the evaluation. AH pointed to specific related work and contributed valuable feedback. All authors read and approved the final manuscript.

**Acknowledgments**
This work is based on research and comparison studies that firstly have been conducted by two HCI students, Philipp Neidhöfer and Rainer Hofmann-Wellenhof have been supervised accompanied by the paper's authors. The research work was later on improved and narrowed down by the authors.

**Author details**
[1] Research Unit HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Auenbruggerplatz 2/V, 8036 Graz, Austria. [2] Institute for Information Systems & Computer Media Graz University of Technology, Inffeldgasse 16c, 8010, Graz, Austria.

Received: 9 January 2015 Accepted: 15 May 2015
Published online: 16 June 2015

**References**


1. Berg JM, Tymoczko JL, Stryer L, Clarke ND, Vol. 2002. Biochemistry. Lubert: Stryer; 2002.
2. Das J, Mohammed J, Yu H. Genome-scale analysis of interaction dynamics reveals organization of biological networks. Bioinformatics. 2012;28:1873–1878.
3. Mathivanan S, Periaswamy B, Gandhi T, Kandasamy K, Suresh S, Mohmood R, Ramachandra Y, Pandey A. An evaluation of human protein-protein interaction data in the public domain. BMC bioinformatics. 2006;7:S19.
4. O'Donoghue SI, Gavin AC, Gehlenborg N, Goodsell DS, Hériché JK, Nielsen CB, North C, Olson AJ, Procter JB, Shattuck DW, et al. Visualizing biological data - now and in the future. Nat Methods. 2010;7:S2–S4.
5. Turkay C, Jeanquartier F, Holzinger A, Hauser H. On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Berlin Heidelberg: Springer; 2014. p. 117–140.
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. 2014;15(Suppl 6):I1.
7. Han K, Park B, Kim H, Hong J, Park J. Hpid: The human protein interaction database. Bioinformatics. 2004;20:2466–2470.
8. Brown KR, Jurisica I. Online predicted human interaction database. Bioinformatics. 2005;21:2076–2082.
9. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 2011;39:D561–D568.
10. Chatr-aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, et al. The biogrid interaction database: 2013 update. Nucleic Acids Res. 2013;41:D816–D823.
11. Jeanquartier F, Holzinger A. On visual analytics and evaluation in cell physiology: a case study. In: Availability, Reliability, and Security in Information Systems and HCI. Berlin Heidelberg: Springer; 2013. p. 495–502.
12. Ware C. Information Visualization: Perception for Design, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2012.
13. Holzinger A, Errath M. Mobile computer web-application design in medicine: some research based guidelines. Universal Access in the Information Society. 2007;6:31–41.
14. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27:431–432.
15. Gehlenborg N, O'Donoghue S, Baliga N, Goesmann A, Hibbs M, Kitano H, Kohlbacher O, Neuweger H, Schneider R, Tenenbaum D, et al. Visualization of omics data for systems biology. Nat Sethods. 2010;7: S56–S68.
16. Leong GW, Lee SC, Lau CC, Klappa P, Omar MSS. Comparison of computational tools for protein-protein interaction (ppi) mapping and analysis. Jurnal Teknologi. 2013;63:.
17. Consortium EP, et al. The encode (encyclopedia of dna elements) project. Science. 2004;306:636–640.
18. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, et al. A draft map of the human proteome. Nature. 2014;509:575–581.
19. Berggŕd T, Linse S, James P. Methods for the detection and analysis of protein-protein interactions. PROTEOMICS. 2007;7:2833–2842.
20. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD. Cytoscape web: an interactive web-based network browser. Bioinformatics. 2010;26: 2347–2348.
21. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, et al. The intact molecular interaction database in 2012. Nucleic Acids Res. 2011. gkr1088.
22. Salazar GA, Meintjes A, Mulder N. Ppi layouts: Biojs components for the display of protein-protein interactions. F1000Research. 2014;3:50.
23. Salazar GA, Meintjes A, Mazandu G, Rapanoël HA, Akinola RO, Mulder NJ. A web-based protein interaction network visualizer. BMC bioinformatics. 2014;15:129.
24. Albrecht M, Kerren A, Klein K, Kohlbacher O, Mutzel P, Paul W, Schreiber F, Wybrow M. On open problems in biological network visualization In: Eppstein D, Gansner E, editors. Graph Drawing. Volume 5849 of Lecture Notes in Computer Science. Berlin Heidelberg: Springer; 2010.
25. Krzywinski M, Birol I, Jones SJ, Marra MA. Hive plots - rational approach to visualizing networks. Brief Bioinformatics. 2012;13:627–644.
26. Agapito G, Guzzi PH, Cannataro M. Visualization of protein interaction networks: problems and solutions. BMC Bioinformatics. 2013;14:S1.
27. Holzinger A, Ofner B, Dehmer M. Multi-touch graph-based interaction for knowledge discovery on mobile devices: State-of-the-art and future challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Berlin Heidelberg: Springer; 2014. p. 241–254.
28. Herbert KG, Spirollari J, Wang JT, Piel WH, Westbrook J, Barker WC, Hu ZZ, Wu CH. Bioinformatic databases: Wiley Encyclopedia of Computer Science and Engineering; 2008.
29. Droit A, Poirier GG, Hunter JM. Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function. J Mol Endocrinol. 2005;34:263–280.
30. Ben-Hur A, Noble WS. Kernel methods for predicting protein-protein interactions. Bioinformatics. 2005;21:i38–i46.
31. Mosca R, Pons T, Céol A, Valencia A, Aloy P. Towards a detailed atlas of protein–protein interactions. Curr Opin Struct Biol. 2013;23:929–940.
32. Atias N, Sharan R. Comparative analysis of protein networks: hard problems, practical solutions. Commun ACM. 2012;55:88–97.
33. Tuncbag N, Kar G, Keskin O, Gursoy A, Nussinov R. A survey of available tools and web servers for analysis of protein–protein interactions and interfaces. Brief Bioinformatics. 2009;10:217–232.
34. Koh GC, Porras P, Aranda B, Hermjakob H, Orchard SE. Analyzing protein protein-interaction networks. J Proteome Res. 2012;11:2014–2031.
35. Klingström T, Plewczynski D. Protein-protein interaction and pathway databases, a graphical review. Brief Bioinformatics. 2011;12:702–713.
36. Bader GD, Cary MP, Sander C. Pathguide: a pathway resource list. Nucleic Acids Res. 2006;34:D504–D506.
37. Mora A, Michalickova K, Donaldson IM. A survey of protein interaction data and multigenic inherited disorders. BMC Bioinformatics. 2013;14:47.
38. Oveland E, Muth T, Rapp E, Martens L, Berven FS, Barsnes H. Viewing the proteome: How to visualize proteomics data PROTEOMICS. 2015;15: 1341–1355.
39. Pavlopoulos GA, O'Donoghue SI, Satagopam VP, Soldatos TG, Pafilis E, Schneider R. Arena3d: visualization of biological networks in 3d. BMC Syst Biol. 2008;2:104.
40. Han K, Ju BH, Park J. Interviewer: Dynamic visualization of protein-protein interactions In: Goodrich M, Kobourov S, editors. Graph Drawing. Volume 2528 of Lecture Notes in Computer Science. Berlin Heidelberg: Springer; 2002. p. 364–365.
41. Byun Y, Han K. Visualization of protein-protein interaction networks using force-directed layout In: Sloot P, Abramson D, Bogdanov A, Gorbachev Y, Dongarra J, Zomaya A, editors. Computational Science? ICCS 2003. Volume 2659 of Lecture Notes in Computer Science. Berlin Heidelberg: Springer; 2003.
42. Wang Q, Tang B, Song L, Ren B, Liang Q, Xie F, Zhuo Y, Liu X, Zhang L. 3dscapecs: application of three dimensional, parallel, dynamic network visualization in cytoscape. BMC Bioinformatics. 2013;14:322.
43. Wang Y, Zhang XS, Chen L. Computational systems biology: integration of sequence, structure, network, and dynamics. BMC Syst Biol. 2011;5:S1.
44. Prokop A, Csukas B, Vol. 1. Systems Biology: Integrative Biology and Simulation Tools. Berlin Heidelberg: Springer; 2013.
45. Omenn GS. Grand challenges and great opportunities in science, technology, and public policy. Science. 2006;314:1696–1704.
46. Huang M, Ding S, Wang H, Zhu X. Mining physical protein-protein interactions from the literature. Genome Biol. 2008;9:S12.
47. Jaeger S, Gaudan S, Leser U, Rebholz-Schuhmann D. Integrating protein-protein interactions and text mining for protein function prediction. BMC Bioinformatics. 2008;9:S2.
48. Gopalacharyulu PV, Lindfors E, Bounsaythip C, Kivioja T, Yetukuri L, Hollmén J, Orešić M. Data integration and visualization system for enabling conceptual biology. Bioinformatics. 2005;21:i177–i185.
49. Moser E, Kargl J, Whistler JL, Waldhoer M, Tschische P. G protein-coupled receptor-associated sorting protein 1 regulates the postendocytic sorting of seven-transmembrane-spanning g protein-coupled receptors. Pharmacology. 2010;86:22–29.

50. Zheng X, Chang F, Zhang X, Rothman VL, Tuszynski GP. G-protein coupled receptor-associated sorting protein 1 (gasp-1), a ubiquitous tumor marker. Exp Mol Pathol. 2012;93:111–115.
51. Brooke J. Sus-a quick and dirty usability scale. Usability Eval Ind. 1996;189:194.
52. North C. Toward measuring visualization insight. IEEE Comput Graph Appl. 2006;26:6–9.
53. Carpendale S. Evaluating information visualizations. In: Information Visualization. Berlin Heidelberg: Springer; 2008. p. 19–45.
54. Lam H, Bertini E, Isenberg P, Plaisant C, Carpendale S. Empirical studies in information visualization: Seven scenarios. IEEE Trans Vis Comput Graph. 2012;18:1520–1536.
55. Lammarsch T, Aigner W, Bertone A, Miksch S, Turic T, Gartner J. A comparison of programming platforms for interactive visualization in web browser based applications. In: Information Visualisation, 2008. IV '08. 12th International Conference. IEEE: IEEE Computer Society; 2008. p. 194–199.
56. Andrews K, Wright B. Fluiddiagrams: Web-based information visualisation using javascript and webgl. In: Proceedings of the Eurographics Conference on Visualization (EuroVis 2014 Short Paper); 2014. p. 91–95.
57. Ono K, Demchak B, Ideker T. Cytoscape tools for the web age: D3.js and cytoscape.js exporters. F1000Research. 2014;3:143–143.
58. Harger JR, Crossno PJ. Comparison of open-source visual analytics toolkits. In: IS&T/SPIE Electronic Imaging, International Society for Optics and Photonics; 2012. p. 82940E–82940E.
59. Prieto C, De Las Rivas J. Apid: agile protein interaction dataanalyzer. Nucleic Acids Res. 2006;34:W298–W302.
60. Kamburov A, Stelzl U, Lehrach H, Herwig R. The consensuspathdb interaction database: 2013 update. Nucleic Acids Res. 2013;41: D793–D800.
61. Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, Campbell NH, Chavali G, Chen C, Del-Toro N, et al. The mintact project-intact as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 2014;42:D358.
62. Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol. 2007;8:R95.
63. Calderone A, Castagnoli L. Cesareni, G: mentha: a resource for browsing integrated protein-interaction networks. Nat Methods. 2013;10:690–691.
64. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardozza AP, Santonico E, et al. Mint, the molecular interaction database: 2012 update. Nucleic Acids Res. 2012;40:D857–D861.
65. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, Cesareni G. Homomint: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC Bioinformatics. 2005;6:S21.
66. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, et al. String v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41:D808–D815.
67. Kalathur RKR, Pinto JP, Hernández-Prieto MA, Machado RS, Almeida D, Chaurasia G, Futschik ME. Unihi 7: an enhanced database for retrieval and interactive analysis of human molecular interaction networks. Nucleic Acids Res. 2014;42:D408–D414.

## 2.4. Visualizing Uncertainty Of RNA Sequence Base Pairing Variants

I participated in this design contest regarding visualization uncertainty in the domain of RNA biomedicine. The topic caught my attention because I am particularly interested in deepening my understanding in biomedicine and also believe that visualization is key to knowledge transfer as well as gaining new insights.

# Visualizing Uncertainty of RNA Sequence Base Pairing Variants

Fleur Jeanquartier, Claire Jean-Quartier and Andreas Holzinger

**Abstract**—This work describes a design oriented approach to visualizing uncertainty of RNA secondary structure probabilities. We address the challenge of finding an intuitive visual representation of encoding uncertainty in RNA secondary structures. We highlight certain limitations and present three different but not exclusive approaches for tackling this challenge.

## 1 Introduction

In molecular biology researchers have to deal with a decreasing certainty when predicting secondary structures of RNA sequences. Practical testing is limited, computational methods fill the gap in the data with predicted and hence uncertain data. Computational biologists have developed methods to predict the secondary structures (2D folding views of RNA) from a primary sequence of RNA. The outputs of this calculation includes the minimum free energy structure (MFE), the thermodynamically favored and most likely structure, and equilibrium base pairing probabilities. These outputs are typically visualized as a "dot plot", where a box on a square grid of $nxn$ ($n$ is the sequence's length) encodes the base pair binding probability in its area on a logarithmic scale. In addition, the predicted MFE structure is often represented as a secondary structure graph.

## 2 Background

Dot plots (base pair probability matrices) are a common way for visualizing secondary structure calculations. The squares in the plot area represent a pair $(x,y)$, while either color, transparency, blur effects or size of a dot is used to indicate the probability of a base pair [13]. For today, conservation consensus dot plots can even be interactively controlled to some extent: For example, Sorescu et al. [12] describes a mechanism to specify a threshold probability for dynamic visualization adaptation. However, dot plot representations for base pair probabilities are also said to be confusing when complexity rises, and therefore alternative representations exist too. Base pairings visualization can also be found as linear and circular representations. Alberts et al. [1] introduced so called "RNAbow" diagrams. Hofacker [6] described a software package for analyzing secondary structures and rendering structures as mountain plot and other representations.
When speaking of uncertainty, uncertain data sets may have diverse sources, including data acquisition (signal-to-noise ratio), data mapping (pre-processing and post-processing) and the visualization method itself. Uncertainty can be described as a composite of different concepts, such as errors, accuracy, and subjectivity [4]. Visualizing uncertainty is a difficult problem in all kinds of scientific domains too [5, 11, 2, 8]. Potter et al. [10] already identified uncertainty representations commonly used in visualization and presented a taxonomy of visualization approaches.
None of the mentioned research already dealt with visually encoding uncertainty of the complete set of folding possibilities into one single visualization.
Therefore, we submit this entry to the BioVis 2015 Design Contest [3], that addresses the challenge of visualizing uncertainty of RNA secondary structures. In the following, we describe our visual approaches to the challenge of visualizing uncertainty.

- *Fleur Jeanquartier, Claire Jean-Quartier and Andreas Holzinger are with the Research Unit HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz. E-mail:* {f.jeanquartier,c.jeanquartier,a.holzinger}@hci-kdd.org

## 3 Visual Approach to Challenge 1

We address the first contest's challenge, namely visualizing uncertainty. The problem is defined as follows:

### 3.1 Problem:

Design an intuitive visual representation of RNA secondary structure to encode the uncertainty within all the possible base pairing possibilities. The top-right triangle of a dot plot encodes base pairing probabilities and the bottom-left triangle represents the MFE structure. The RNA sequence of $n$ nucleotides is shown on the edge of the $nxn$ square grid. The MFE secondary structure is visualized as a graph, where the color of each nucleotides depicts the strength of base pairing. The challenge is to design a structural representation that is in line with the uncertainty.
To deal with this challenge, however, using the right visualization technique is a question of scaling: An unanswered question remains: What is the limit of possible base pairing probability matrices that can be visualized within one single visualization? Since the number ob potential secondary structures is exponential to the rna sequence's length $n$ [9]. Therefore, we present the following three different approaches for (interactive) visual analysis of rna base pair configurations:

### 3.2 Approach 1:

One possible interactive visualization approach is sketched in Fig. 1:

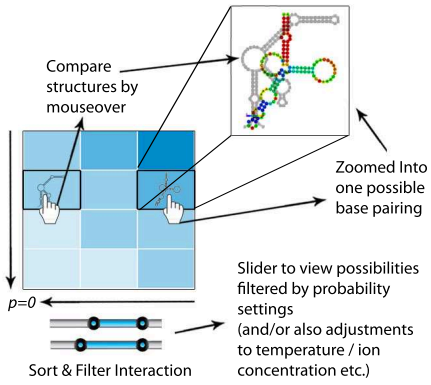


Fig. 1. Visualizing encoded uncertainty of RNA secondary structure possibilities as interactive heatmap including detail view

Holzhuter et al. [7] have shown that particularly heat maps can be dangerous as they can be over-plotted. It is possible, up to a certain amount, to visualize the ensemble organized in a heatmap. But, as common to information visualization, there will be the necessity to integrate interactive exploration features for zoom and filter. We also sketched such interaction integrations. The slider filter at the bottom supports viewing only those rectangles that are related to the most probable configurations but also allows for highlighting the unusual ones. Different perspectives support the interactive visual analysis approach. Additional interactions should be taken into account, like a

slider for filtering specific temperature areas and/or ion concentration settings and adding a switch for sorting not only by probability but also other data variables (i.e. number of base pairs, hairpins, free energy).

### 3.3 Approach 2:

To overcome some of the heatmap's limitations, another additional or alternative approach is visualizing the complete set of dot plot representations as interactive visual analysis approach making use of the "Rolodex"-art metaphor (also known from window manager in operating systems, apple's time machine or windows exposé), illustrated in Fig. 2. All possible structures are visualized as matrices one after another, while the most probable, the MFE, is the first one on top and behind lay the less probable ones. Interaction allows for toggling through all the possible structures seamlessly while clicking on upper right part of the dot plot all secondary structures are shown in a details view the following manner: All the possible configurations are shown at once, while the most probable is on top. Below all other configurations are shown but with increased transparency values. The most likely is therefore 100% opaque, while the less likely ones are more translucently renderer.
Additionally, Eterna's animation metaphor can be used: Single bases and base pairs within the details view can be animated insofar, as the base pairs movement in pixel per second is related to the structure's folding stability and probability.

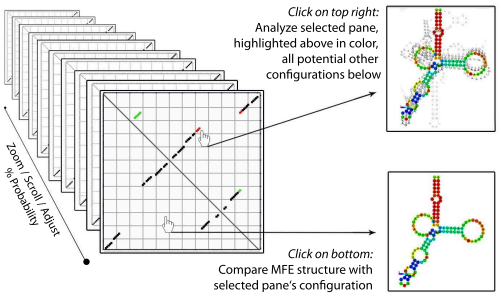


Fig. 2. Visualizing encoded uncertainty of RNA secondary structure base pairings by exploring complete set/ensemble at once

### 3.4 Approach 3:

Last but not least, another possible approach could be visualizing all possibilities not as box but as part of a network graph, sketched in Fig. 3. The graph is composed by the complete ensemble of structures as follows: Each node represents one possible folding structure, each edge stands for a user defined number $x$ similar base pairs between two structures, while the whole graph integrates the complete "picture". Thereby, similar base pair areas can be marked with another color (compare sketched red area in Fig. 3)
The nodes' transparency (or color/contrast variance) depicts the probability of the particular structure. The node that stands for the MFE is highlighted (in darkest contrast or special color) as the root or center of the graph as the most probable base pairing combination. If the MFE is not the most probable configuration, the visualization can be adapted to distinguish between root, as most probable one, and MFE, as a node somewhere else within the graph highlighted by another color.
According to the dynamic programming algorithm for all subsequences $(i, j)$ of a dot plot, the less probable folding possibilities can be traced back too. Less probable configurations are marked in a translucent manner: The more like configurations are represented by nodes with higher opacity while the more unlikely ones are rendered with less opacity.
Regarding the interaction: By adjusting $x$ certain isles are highlighted, where the configurations represented by the nodes within an isle are more similar to each other. Additional network analysis approaches may further suite the rna analysis process.

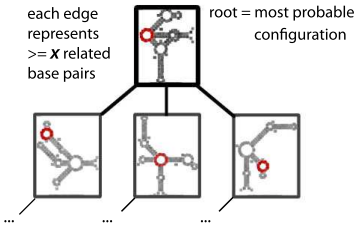


Fig. 3. Visualizing encoded uncertainty of RNA secondary structure by putting focus on the configurations' related base pairs as network graph

## 4 Material And Methods

Due to the fact, that the submission should be no more than 2 pages we include only a few figures into it. We also recommend watching a short animation, that depicts some details about the three different visualization approaches and the structural representation that is in line with the uncertainty: `http://youtu.be/PZp5GNpNZX4`.

## 5 Terms and Conditions

By submitting this entry, we give the BioVis 2015 organizers permission to publish it in conference-related materials. Any usage or reference to any submission will include full credit to its authors.


## Acknowledgments

We gratefully acknowledge the dataset provided by Maria Beatriz Walter Costa, Henrike Indrischek, Katja Nowick and Christian Hner zu Siederdissen at The University of Leipzig for the purposes of the BioVis 2015 Contest.


## References


[1] D. P. Aalberts and W. K. Jannen. Visualizing rna base-pairing probabilities with rnabow diagrams. *RNA*, 19(4):475–478, 2013.

[2] M. Albrecht, A. Kerren, K. Klein, O. Kohlbacher, P. Mutzel, W. Paul, F. Schreiber, and M. Wybrow. On open problems in biological network visualization. In *Graph Drawing*, pages 256–267. Springer, 2010.

[3] BioVis. Design contest, 2015.

[4] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *SimVis*, pages 143–156, 2006.

[5] C. D. Hansen, M. Chen, C. R. Johnson, A. E. Kaufman, and H. Hagen. *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*. Springer, 2014.

[6] I. L. Hofacker. Rna secondary structure analysis using the vienna rna package. *Current protocols in bioinformatics*, pages 12–2, 2009.

[7] C. Holzhüter, A. Lex, D. Schmalstieg, H.-J. Schulz, H. Schumann, and M. Streit. Visualizing uncertainty in biological expression data. In *IS&T/SPIE Electronic Imaging*, pages 82940O–82940O. International Society for Optics and Photonics, 2012.

[8] A. Holzinger, M. Schwarz, B. Ofner, F. Jeanquartier, A. Calero-Valdez, C. Roecker, and M. Ziefle. Towards interactive visualization of longitudinal data to support knowledge discovery on multi-touch tablet computers. In *Availability, Reliability, and Security in Information Systems*, pages 124–137. Springer, 2014.

[9] G. Pavesi, G. Mauri, M. Stefani, and G. Pesole. Rnaprofile: an algorithm for finding conserved secondary structure motifs in unaligned rna sequences. *Nucleic acids research*, 32(10):3258–3269, 2004.

[10] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer, 2012.

[11] J. Smith, D. Retchless, C. Kinkeldey, and A. Klippel. Beyond the surface: current issues and future directions in uncertainty visualization research. In *Proceedigs of the 26th International Cartographic Conference*, pages 1–10, 2013.

[12] D. A. Sorescu, M. Möhl, M. Mann, R. Backofen, and S. Will. Carnaalignment of rna structure ensembles. *Nucleic acids research*, page gks491, 2012.

[13] A. Wilm, K. Linnenbrink, and G. Steger. Construct: improved construction of rna consensus structures. *BMC bioinformatics*, 9(1):219, 2008.

## 2.5. Integrating Open Data on Cancer in Support to Tumor Growth Analysis

Within this conference publication I present preliminary results from studying the availability of open cancer data with the goal to identify some pointers for data that can be used for in silico modeling of tumor growth as well as tumor growth visualization and simulation.

# Integrating Open Data on Cancer in Support to Tumor Growth Analysis

Fleur Jeanquartier[1( )], Claire Jean-Quartier[1], Tobias Schreck[3], David Cemernek[1], and Andreas Holzinger[1,2]

[1] Holzinger Group, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria
{f.jeanquartier,c.jeanquartier,d.cemernek,a.holzinger}@hci-kdd.org
[2] Institute of Information Systems and Computer Media, Graz University of Technology, Graz, Austria
[3] Institute of Computer Graphics and Knowledge Visualisation Graz, University of Technology, Graz, Austria
tobias.schreck@cgv.tugraz.at

**Abstract.** The general disease group of malignant neoplasms depicts one of the leading and increasing causes for death. The underlying complexity of cancer demands for abstractions to disclose an exclusive subset of information related to the disease. Our idea is to create a user interface for linking a simulation on cancer modeling to relevant additional publicly and freely available data. We are not only providing a categorized list of open datasets and queryable databases for the different types of cancer and related information, we also identify a certain subset of temporal and spatial data related to tumor growth. Furthermore, we describe the integration possibilities into a simulation tool on tumor growth that incorporates the tumor's kinetics.

**Keywords:** Open data · Data integration · Cancer · Tumor growth · Data · Visualization · Simulation

## 1 Introduction

Interactive data integration, data fusion and, first and foremost, the selection of datasets is a key research direction to enable knowledge discovery in health informatics generally, and bioinformatics and computational biology specifically [1].

Our aim is to link publicly and freely available data on cancer to an enhanced version of our recently presented tool on tumor growth [2]. Thereby, we list open databases providing datasets on the different types of cancer and collect related information. The datasets are examined for growth-related parameters and subsequently integrated into a simulation tool on modeling neoplasms. This simulation on neoplasia comprises abnormal tissue growth such as benign and malignant tumors. Additional text-based information and non-growth-relevant data is scanned and revised for accessory visualization features.


M.E. Renda et al. (Eds.): ITBAM 2016, LNCS 9832, pp. 49–66, 2016.
DOI: 10.1007/978-3-319-43949-5_4

We further describe and sketch possibilities for integration and visualization of cancer-related data into our recently presented simulation and visualization tool on tumor growth [2]. The Web tool is based on the implementation of the Cellular Potts Model (CPM) and Cytoscape, that is available at https://github.com/davcem/cpm-cytoscape. We present an integrative approach to cancer research. The study rests upon the idea of enhancing the tumor growth simulation by integrating multiple genuine data.

First, we introduce the topic of open data for research in general and on cancer in detail. Further, we recap the biological settings for cancer modeling. We approximate and appoint open datasets on cancer involving tumor growth information by considering temporal and spatial aspects. And, we discuss their feasible incorporation into an online simulation. We proceed with a summary on the key challenges for embedding open data to our cancer simulation. We thereby suggest that an integrative approach is key to understanding cancer.

## 2 Related Work

### 2.1 Open Data for Scientific Research

There is a strong trend towards an increasing number of freely available datasets becoming available in many domains, including scientific research. The idea of open data is to provide unrestricted access for sharing, validating, reusing and merging relevant data to advance scientific research. Several works already show that new opportunities arrive with the increasing amount of open data. The so-called *Fourth Paradigm* [3] envisions data-driven research by widened access to open data for common good.

While open data provides opportunities, there are challenges associated with the provision, discovery and usage of open data. Typically, relevant content needs to be retrieved by researchers. Then, data from different sources of possibly heterogeneous data regarding data type, quality, and resolution need to be integrated for joint analysis.

Interactive visualization can help to explore and related data during the discovery process. Domain- as well as application-specifics need to be taken into account to choose the right visualization tool for supporting search and exploration in general data exploration [4–6]. In previous work, approaches for discovery of relevant data in research data repositories based on exploration and visual querying have been proposed. The VisInfo system [7] allows to query for content in large time series databases. Often, content needs to be related to metadata. In [8] data patterns are correlated with metadata, for enhanced exploration. Visual search for bivariate data has been addressed in [9] using features obtained from scatter plot representations of input data. In absence of example queries from real data, user sketching of patterns can be useful, if appropriate similarity functions can be obtained [10]. Besides exploration, visual-interactive approaches can also be useful for the effective semi-interactive integration of heterogeneous data sources, which is a primary requirement in many open data analysis projects [11].

More specifically regarding the medical domain, we recently compared methods for visualizing and analyzing data in online proteomics databases. Only a few available tools meet the needs for interactive visual analysis [12].

Increasing data availability is not only considered as an opportunity but also new issues arise. Challenges of data integration in the biomedical sciences include determining available and usable data, completeness, re-use for novel approaches for data discovery and exploitation [1,13].

### 2.2 Open Data in Cancer Research

Biomedical data comes in many guises [1]. Initiatives are already fostering open-access research for improving patient care. There are several freely accessible web portals, yet, providing exploration support for cancer genomics due to increasing efforts in the area of Bioinformatics regarding genomic data handling [14–22]. For example, challenges in normalizing clinical drug data have been illustrated while using open access druggable genome datasets for target discovery in the context of cancer therapeutics [23].

With regard to imaging data there are several online resources providing several million cancer images, which are partly public, partly protected. Available imaging data includes computed tomography, magnetic resonance and other images. De-identification scripts support moving more and more images on public servers [24].

Text mining for literature curation is common for omics data [25]. Summaries of fundamental concepts for text mining in cancer research are mainly concerned on relation extraction mechanisms such as identifying protein-protein, gene-gene or gene-disease relations [26]. Text mining has already been combined with manually curated data for data integration in the context of disease-gene associations [27]. Several open access literature resources exist to apply text mining for finding suitable disease data. However, text mining in biomedical literature is more sophisticated than for clinical data [28]. Only a few databases provide information on cancer incidences and statistics. Movements come from the American Cancer Society and the World Health Organization [29–31]. Data protection regulations and privacy is one of the obstacles to tackle to providing open data for biomedical research [33,34] There are approaches for space-time analysis and visualization related to cancer, but they deal with population data such as location and age [35].

Sophisticated integrative analysis tools for cancer are yet to be found [36]. Online available disease ontologies help understanding the relationships of cancer terms and foster communication and exchange [37,38].

To our knowledge, there is no approach to identifying tumor growth related open data. We therefore focus on identifying temporal and spatial entities within available cancer data.

### 2.3 Biological Background

There are two basic biological phenomenons which play essential roles in the disease of cancer. First, spontaneous mutations occur naturally and frequently within all cells [39]. Secondly, normal cells can undergo programmed cell death, so-called apopotosis, with time. In some cases however, such mutations can have an effect on cellular functions. Tumor cells are characterized by a change in the proliferative capacity. Malignancy can be developed if mutations lead to the inhibition of apoptosis or excessive proliferation and could further end in differentiation. Tumors can look and function similar to normal cells. Benign masses of tumor cells are normally localized. They only become problematic if space is limited or keep producing hormones in excess [40]. Malignant tumor cells become more serious. They do not only grow more rapidly but they can also invade other tissues and parts throughout the body. Parameters that relate to the specified aspects in tumor growth are of particular importance for modeling cancer. Since mutations are the onset of cancer, open data is concentrated on genetic data. Still, in order to combat the disease relational information has to be retained.

## 3 Approach

Our approach is to study open datasets for querying and relating interaction data to (gene classified) cancer diseases. The goal is to extend an existing framework for simulating and visualizing tumor growth [2] by integrating a selected subset of spatial and temporal data for supporting exploration and sense-making. To achieve this goal several data integration steps are necessary. Most important, available data has to be identified and examined for relevance.

### 3.1 Relevance to Tumor Growth?

We focus on summarizing and picking specific information on tumor growth. Presently, there are no web-resources providing exclusive data on tumor growth. So, relevant information has to be isolated from an abundance of data in matters of cancer research. We aim to gather cancer-relevant data in regard to spatial and temporal criteria in particular.

Temporal and spatial characteristics on tumor growth can be influenced by several factors, such as gene regulation or mutations as well as drugs and other inhibitors or promoters. In cancer, the balance between growth promoting and inhibiting factors is shifted towards proliferation. The underlying signal-transduction pathways are complex biological processes involving several key steps as well as mediators which are dynamically and differentially regulated. The influencing factors have to be recognized and parameterized in order to be integrated into the simulation.

We are equally interested in statistical assessment of growth kinetics from various tumors and cancer subtypes, as well as incidence reports on isolated case

reports. Notably, entity relationship descriptions and interaction data in regard to tumor growth characteristics are of relevance and primary focus.

Previous studies on tumor growth prediction could be likewise included. In order to enhance the cancer modeling tool, we aim to provide a comprehensive simulation comprising growth characteristics of various kinds of tumors. Most studies on predictive cancer modeling focus on the kinetics of various cancer diseases. We try to collect and capture the specifics of several tumor types and to likewise broaden and refine the visualization approach tumor growth analysis.

## 4 Results

We present an overview of available cancer-related open data. We categorize identified datasets corresponding to the content types that can be found with respect to cancer research. The study shows that genomic data as well as imaging data is increasingly available. But, explicit information on temporal and spatial aspects are hardly found. Text mining in incidence reports and open access publications have to be taken into account in order to find suitable data for tumor growth simulation. Furthermore, we describe the integration of a subset of open data related to tumor kinetics, temporal and spatial data in particular, into an existing tumor growth simulation user interface that is freely online available via github.

### 4.1 Overview of Available Data

We categorize online available information from cancer research under 5 different categories. First, many datasets provide **genomic data**. Secondly, **incidence data** can be analyzed and downloaded from several portals. Third, there are large archives consisting of **imaging data**. Fourth, there are several databases that consist of **disease associations** such as disease ontologies. Last but not least, open access databases provide a comprehensive list of **literature data** for text mining.

By considering content quality, license information and access possibilities for each of the listed entries, we chose a subset that satisfied the needs for free non-commercial usage as well as data relevance. Table 1 lists facts about the identified databases regarding its data category relation.

Starting with a review of currently available cancer genomic databases for research [41], our search strategy included systematically examining lists of databases of cancer-related data presented at metasites found via online search. Therefore, we iteratively extended a table of cancer related databases until we arrived at a comprehensive list of databases that we are summarizing below. We examined available databases and included information about access possibilities as well as descriptions about the provided data type/category, the data's coverage, whether download of data as well as a web API is provided, license information and last but not least studied optional input and output entities.

**Table 1.** Statistics about list of non-filtered databases

| Category | # Identified databases | # Chosen databases | Possibilities for spatial data | Possibilities for temporal data |
|---|---|---|---|---|
| Genomic data | 15+ | 9 | – | – |
| Imaging data | 6+ | 5 | ✓ | – |
| Incidence data | 6 | 4 | – | ✓ |
| Disease associations | 6 | 3 | ✓ | ✓ |
| Literature data | 2+ | 2 | ✓ | ✓ |

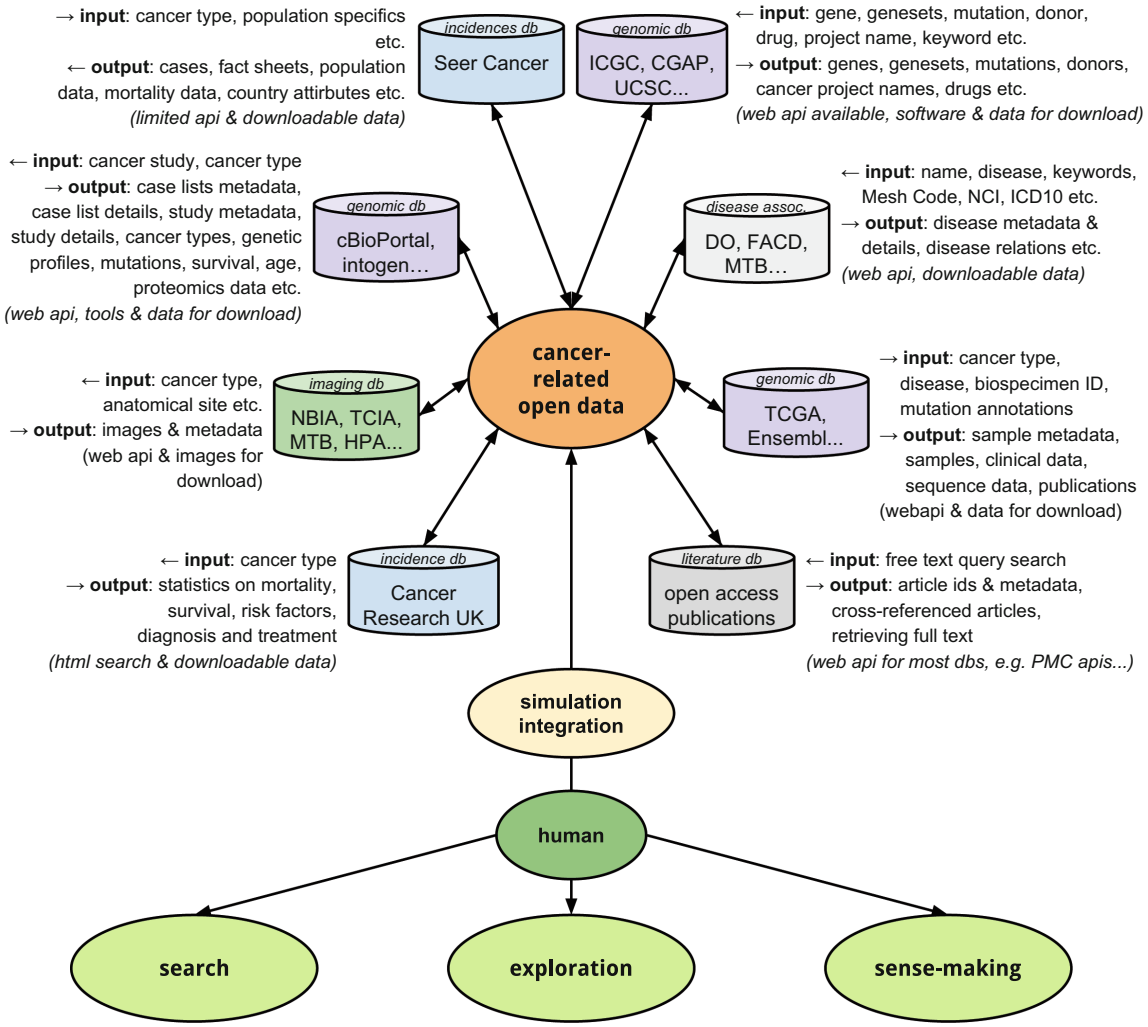


**Fig. 1.** Overview of cancer databases for integration

Therefore, next to the availability of spatial and temporal data, we further differentiate between possible input and output. Figure 1 shows an overview of our approach. The input and output is being summarized. The node's color corresponds to the data's category.

**Table 2.** Summary of examined databases that may be suitable for the task of data integration

| Category/Name | Abbreviation | Data access | Ref. |
|---|---|---|---|
| **Genomic data** | | | |
| The Cancer Genome Atlas - Data Portal | TCGA | REST, download | [16] |
| cBio Cancer Genomics Portal | cBioPortal | REST, download | [15] |
| NCI's Cancer Genome Anatomy Project | CGAP | download | [43] |
| International Cancer Genome Consortium - Data Portal | ICGC | REST, download | [19] |
| United States Cancer Statistics - Cancer Genomics Browser | UCSC | download | [16] |
| Catalogue of somatic mutations in cancer | COSMIC | REST, download | [18] |
| Integrative Onco Genomics | INTOGEN | download | [20] |
| Integrative Genomics Viewer | IGV | download | [21] |
| Many more general genome databases such as Ensembl | ENSEMBL | REST, download | |
| **Imaging data** | | | |
| The Cancer Imaging Archive | TCIA | REST, download | [24] |
| CancerData.org - Sharing data for cancer research | CancerData | download | [45] |
| Mouse Tumor Biology - Database | MTB | download | [44] |
| National Biomedical Imaging Archive | NBIA | REST, download | [24] |
| Many more such as the Human Protein Atlas | HPA | download | |
| **Incidence data** | | | |
| WHO Cancer Mortality Database | WHOdb | download | [46] |
| Center for Disease Control and prevention - Cancer Data and Statistics | CDC | download | |
| Surveillance, Epidemiology, and End Results - Program | SEER | download | [30] |
| Cancer Incidence in Five Continents | CI5 | download | [31] |
| **Disease associations** | | | |
| Diseases Ontology | DO | REST, download | [37] |
| Mouse Tumor Biology - Database | MTB | download | [44] |
| NCI Thesaurus | NCIt | REST, download | [38] |
| **Literature data** | | | |
| PubMed Central | PMC | REST, download | [26] |
| Europe PubMed Central | Europe PMC | REST, download | [32] |

Table 2 lists all examined databases providing cancer-related content as download that is free for non-commercial, scientific purposes, sorted by category.

The summarizing table shows only a small subset of examined resources due to the fact that several licensing issues as well as quality issues such as deprecated data that has not been maintained for years have been identified during our research. We also observed that several data portals make use of others, e.g. the Disease Ontology's cancer project includes several mappings from other databases, especially genomic data. The "+" in the column of identified data-

bases within Table 1 implies that more databases could be found but are already included within other databases. To that effect, the databases' peculiarities also include data coverage such as databases that cover other databases' contents as well. Due to that reason, we chose to use only the largest two archives of biomedical literature data for further literature mining.

### 4.2 Literature Mining

We conducted a search for some tumor growth related terms to test the suitability of literature databases for finding data to be integrated. PubMed has been reported to be one of the best biomedical publication archives [26]. Therefore, we chose to conduct some mining within the two public archives of biomedical and life sciences literature, "Europe PMC" and "Pubmed Central" (PMC). Additionally, we made use of an information retrieval tool for biological literature called "Textpresso" [42]. Example queries are summarized below.

**Table 3.** Example queries for text mining

| Database or tool | Query for "abnormal cell growth" | Query for "tumor growth" | Query for "tumor cell growth" | Query for "neoplasm" |
|---|---|---|---|---|
| Textpresso | 111 matches, 33 documents | 3891 matches, 926 documents | 37072 matches, 6519 documents | 3990 matches, 2000 documents |
| Europe PMC | 1399 matches, 277 open access | 121435 matches, 35174 open access | 12555 matches, 4089 open access | 4076094 matches, 436216 open access |
| PMC | 1389 matches | 98822 matches | 13557 matches | 2837065 matches |

Making use of specific text mining tools is favored over literature mining for finding most relevant results and presenting sets of results. E.g. highlighting matching sentences is crucial to a fast scan through results and the identification of relevant information.

### 4.3 Data Processing

Most online portals provide free access to the data available as downloadable content, some accompany web interfaces such as web services for direct access too. In each case further data processing steps are necessary to respond to the needs of (visual) data mining and integration into the existing user interface.

Most genomic data portals already provide entity relationship (ER) diagrams for documentation of available data entities and relations. However, we focused on finding temporal as well as spatial tumor growth data and were not able to identify explicit information about those aspects within available cancer genomic data. Further mining techniques have to be taken into account to accomplish

the task of finding suitable information about specific growth impact on cancer disease-gene associations.

As a starting point for data integration we created a set of different growth functions by literature curation. We collected data points for comparing discrete growth functions for tumor growth, vascularization inhibition and cell density inhibition on growth. Data points come from three different publications found via PMC and is summarized in Fig. 2 [47–49].

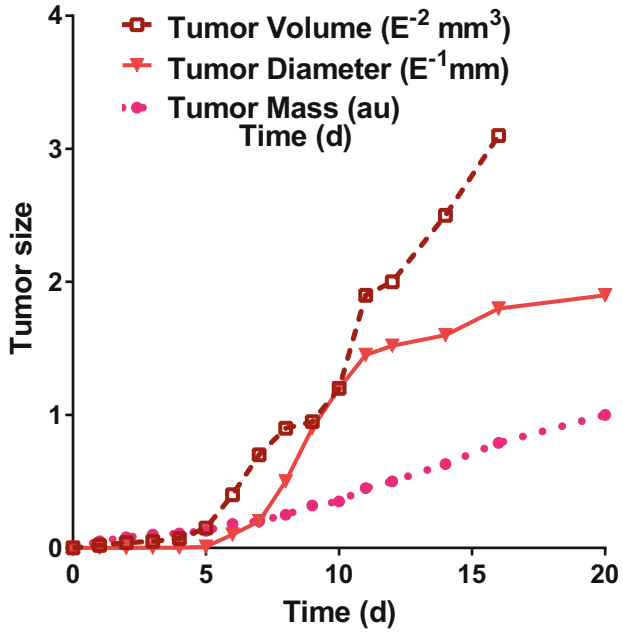


**Fig. 2. Literature-curated discrete tumor growth - data samples:** various tumor types, determined growth in tumor size, given in miscellaneous units, over time, presented in days.

### 4.4 User Interface Extensions

Cpm-cytoscape is a tool for scientific simulation and visual analysis of tumor growth. The web application makes use of the CPM for modeling tumor growth. The CPM is a popular lattice-based, multi-particle cell-based model that has been used for modeling tumor growth in a wide area. The tool incorporates a novel graph-based visualization approach [2]. Figure 3 shows an annotated screenshot of the existing user interface, describing the different interaction and visualization possibilities of the tool's user interface.

The tool's framework integrates visualization features for analysis via JavaScript and HTML. A Converter Class allows for extending the data objects

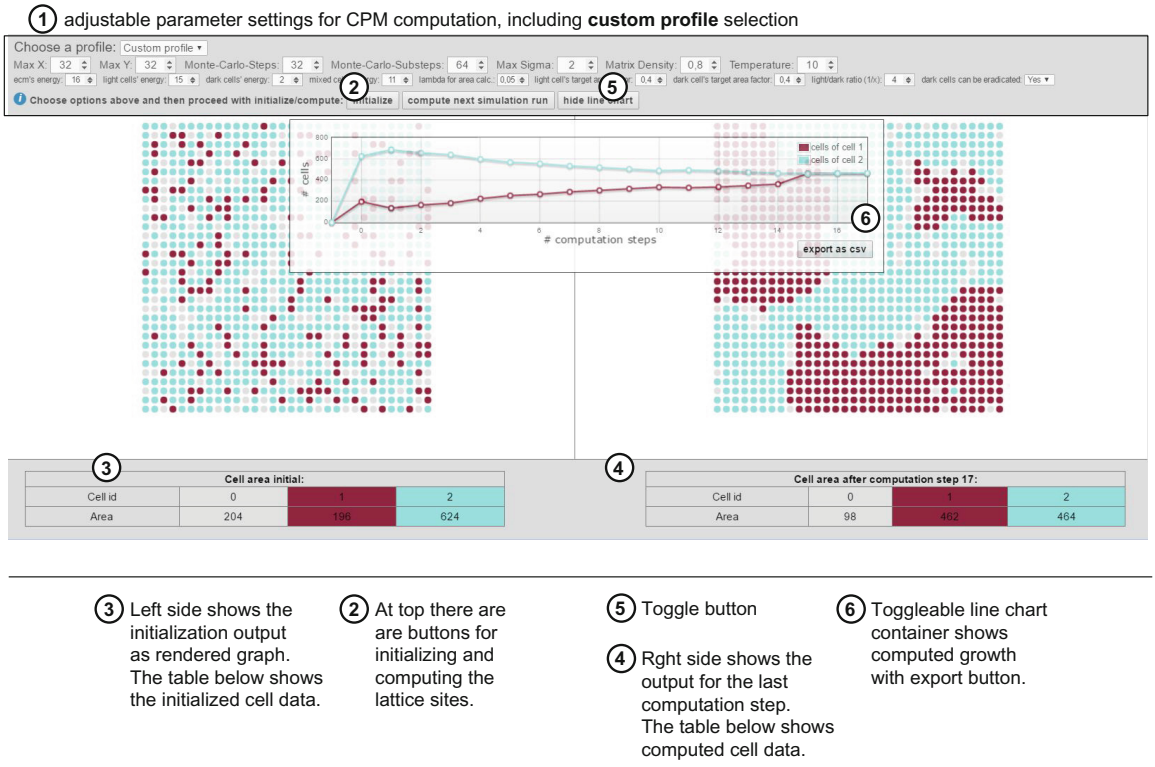


**Fig. 3.** Overview of User Interface with custom profile showing kinetics and cell sorting after several simulation steps

that represent simulated cell sorting and kinetics. Another Converter allows for processing data to communicate between backend and frontend. This Java Class maps the graph data from the modeling computation to the format needed by the visualization renderer in the frontend. Such converter classes are easily extendable and support integrating additional information. The simulation and its several computation steps are started via Representational State Transfer (REST) calls, while the user interface displays response information both within the graph visualization as well as in an overlay as simple Line diagram. Details on its usage and implementation can be found on the project's github page [2].

**Profile Specific Simulation and Visualization.** The first implemented extension to the user interface is the ability to provide "profiles" for running simulations under different configurations. The simulation can be started with the help of choosing a profile or specifying a custom profile. Figure 3 shows a completed simulation for a custom profile. The profile extension is a good example of extending the user interface neatly and encapsulated. A separate JavaScript function call via changeProfile() is located in an separate extension. Each profile for selection is represented as JSON file for easy maintenance. The profile can be selected via a dropdown (Fig. 4). The parameter settings that are available via JSON files can be replaced with a dynamic function that communicates with another server to get all the various parameter settings.

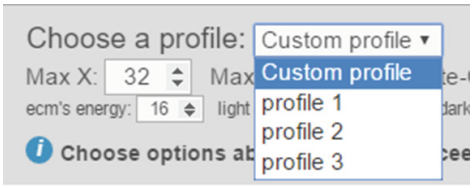


**Fig. 4.** Screenshot of profile selection possibilities

Until now we did not find any database that holds all the data needed to have different complete configurations to run a simulation, therefore we are providing static configuration files to try out different settings that have found via manual literature mining. However, this extension is a good example to start the task of data integration and can be further extended as soon as a suitable dataset is available.

**Presenting Details on Cell Nodes.** The visualization of cell sorting and kinetics is based on a graph. Each node is representing a so called "cellular brick" of a cell. A cell is a set of 0 to $n$ cellular bricks with the same cell-index, while each cell *sigma* is of a specific cell-type $\tau$. Until now, we only differentiate between proliferating tumor cells and healthy cells as distinct cell types, with different growth rates and volume constraints for each type, rendered as colored nodes. Thirdly, we use grey nodes to represent the extracellular matrix (ECM). Additional information on nodes can be provided via context menu. According to the node's cell-index $\sigma_{i,j}$ additional information about the associated cell-type can be shown, while proliferating tumor cells are called "dark" cells and the other healthy cells are called "light" cells. Cells with $\sigma_{i,j} = 0$ represent the ECM, visualized as grey nodes. Cells with odd $\sigma_{i,j}$ represent the "dark" cells and are visualized as dark red colored nodes. The other cells with an even $\sigma_{i,j}$ show "light" cells and can be recognized by the lighter blue to green colored nodes.

**Search for Reports on Related Diagnosis and Treatment.** Text-based search within an existing incidence data provides exploration of similar cases, diagnosis, treatment as well as other possible relations. Figure 5 shows a mock-Up of a simple integration. As starting point we just link to additional information. However, a tight integrative approach would be adding further data to the computation of the several simulation's steps. Taking additional information into account such as drug information that has impact on growth could then be presented as uncertainty visualization as sketched in Fig. 6.

**Direct Inclusion of Time-oriented Data for Growth Simulation.** An ultimate goal is to include information not only on existing related incidences but far more information on drugs and other inhibitors or promoters to be integrated directly into the computation process. In particular time-oriented data as we see in the simple line diagram showing the growth of different celltypes supports integration of additional information to be visualized for further exploration and

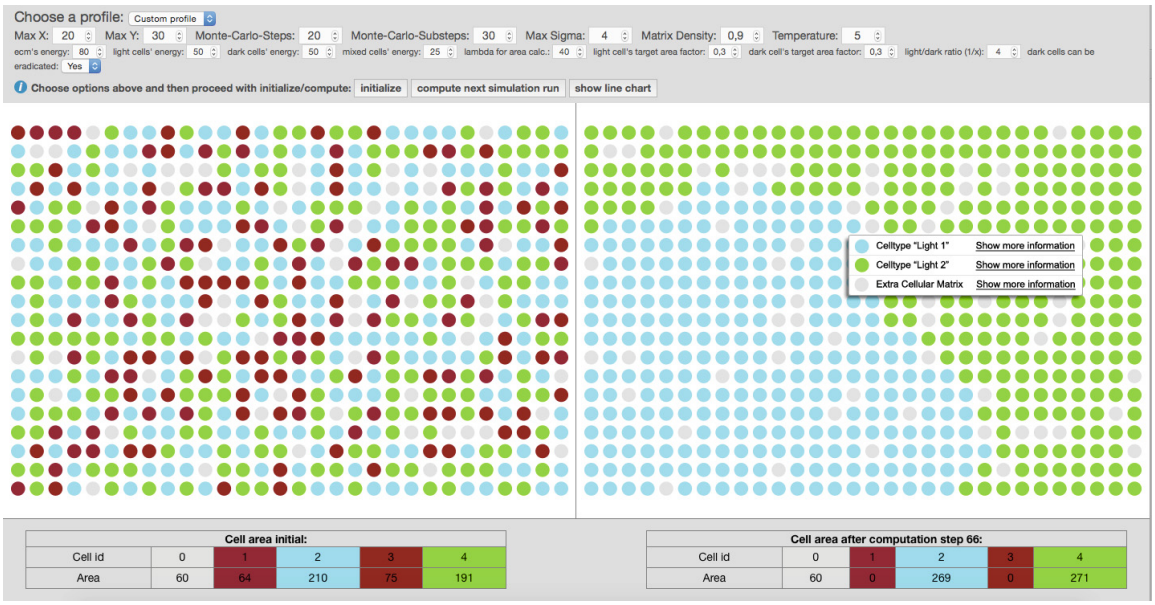


**Fig. 5.** Screenshot, showing additional information for cell nodes (Color figure online)

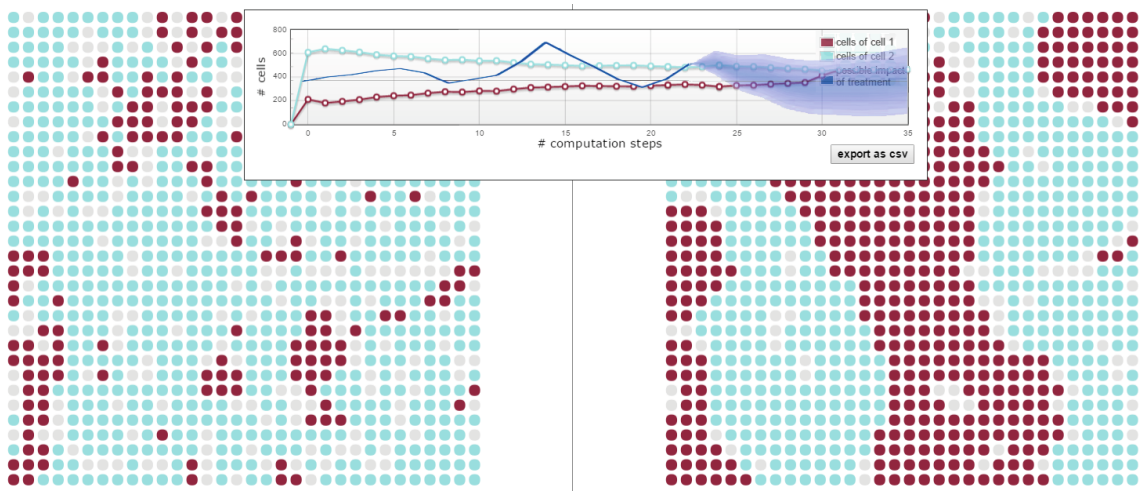


**Fig. 6.** Mock-Up of a time-line extension showing results of a computation taking additional information on treatment into account

analysis. Regarding the carcigonesis we have to include information about several attributes of tumor progression as well as genetic theory. Genomic databases also provide data in biotab format that includes temporal data such as "days to death" [50]. The possibilities are numerous. Comparing progress is possible with visualization metaphors such as making use of a Layer Area graph, Braided graph, Stream-graph or even parallel coordinates as well as many others [51].

## 5 Challenges

Our work is an intermediate step in extending cancer research using a specific tool and feeding it with additionally enhanced data. A number of challenges has

to be addressed. There are many open issues for data integration, in particular to cancer data. We summarize and explain the most important ones.

**Relevance.** A key challenge is finding suitable relations in a domain-specific manner. Are relevant data such as growth rates explicitly available via open data sources or hidden within text retrieval of open access publications (literature curation)? How can relevant data sets be successfully retrieved?

**Data Quality.** Regarding data quality, aspects of accuracy and completeness have to be taken into account. Several genomics databases show associations between diseases and genes for several reasons, sometimes only because of the fact that queried terms occurred in the same publication. Further data processing steps have to be taken into account to decrease retrieval of false-positive or false-negative associations.

**Tight Integration of Visualization.** Integration for visual data analysis is possible on different levels. Moving beyond visualization as simple presentation of computation results, several interaction possibilities have to be included seamlessly to foster understanding of the underlying processes [5].

Specifically in the case of simulations, experts need to set many parameters but it is often not clear what the effect of the different parameters will be. Hence, there is a need for representing sensitivity and also, uncertainty of the analysis results. The latter is particularly relevant in case of incomplete data, or data of varying levels of resolution. Moreover, the integration of the knowledge of a domain expert can sometimes be indispensable, and the interaction of a domain expert with the data would greatly enhance the whole knowledge discovery process pipeline, i.e. interactive machine learning puts an human-into-the-loop to enable what neither a human nor a computer could do on their own [52].

**Ease of Use.** Incorporating a human computer interaction perspective into cancer simulation and visual analysis, we have to face the danger of user interface overload due to the complexity of data integration. Integrating various multi-dimensional result-sets of different databases in a consistent and concise way to maintain an intuitive user interface. While our approach is to provide tumor growth simulation and visual analysis via an intuitive user interface that is online available, questions to be answered still remain: How to facilitate exploration and discovery and how to make complex cancer data easily accessible.

## 6 Discussion and Conclusion

Cancer research is a data-intensive application domain that, on the one hand, raises many challenges for researchers, technicians and clinicians. On the other one in silico modeling may benefit from the many possibilities that come with accessible data related to the disease of cancer.

We implemented an easily extendable user interface using open-source components, with the ultimate goal of supporting in silico modeling by dissemination

and contribution throughout the Computational Biology community for cancer research. Visualization for scientific simulations can have a positive impact on exploration, comparison and understanding. Therefore we are iteratively extending a visualization approach to tumor growth simulation and describe some examples as a starting point, how publicly available data can be used to further enhance the analysis of tumor kinetics.

We believe that it is essential to exploit and integrate data to achieve the goal of supporting clincians' decision making. The tool's extensions have been co-designed and validated by a domain-expert, but have not been evaluated by clinicians so far. Future plans are to conduct iterative testing and validating.

This contribution is preliminary work and aims to facilitate integration of heterogeneous data sources for tumor simulation and analysis by providing a categorized list of databases and describing integration possibilities. Open Data for cancer research can be disposed on a large scale: Incidence reports can be used to enhance a statistical and probabilistic approach to prediction regarding population data such as age, sex, etc. Imaging archives can be exploited for input testing. Further, profiles can be created and utilized. First attempts are discussed in [53]. Databases provide information about mutation probabilities regarding specific cancer types. Subsequently, genomic information can be used for biomarker discovery, for targeting strategies regarding novel drugs. Moreover, the comparison of biopsies with other incidence reports may foster personalized medicine. Data can be used for parameter refinement not only for extending the set of profiles but also including more variables according multicellular structures.

In general, the sheer abundance of data, derived from multiple experiments in cancer research, asks for a more comprehensive approach to data retrieval, analysis and application [36].

The progress of sophisticated biochemical and biomedical methods may not outrank the development of bioinformatic methods in order to salvage the often multi-dimensional information There is a general need to readily access cancer data from public repositories. Data integration resembles one promising option to this task.

So far, Web repositories on cancer information focus genomic and mutational data in particular. We experienced that one can easily get sunk within this magnitude of information in search of completely different readings. We aim to pick and choose details of growth-relevance in order to refine and improve kinetic models within field of computational biology in cancer. In anticipation of future development, in terms of personalized medicine, individual mutational profiles could be compared to those from repositories and integrated by determining the scope of the specific tumor growth. This approach could be equally employed for proteomic material. For that matter, further information on spatial and temporal changes due to genetic changes have to be allocated to online repositories. Ultimately, such an approach will predict the outcome of the disease and the patient's survival possibilities.

Concluding, we believe that the key to understanding the concept of cancer lies within the integrative translation and multi-dimensional connection of open data.

## References


1. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinform. **15**(Suppl. 6), I1 (2014)
2. Jeanquartier, F., Jean-Quartier, C., Cemernek, D., Holzinger, A.: In silico modeling for tumor growth visualization. BMC Syst. Biol. (2016)
3. Hey, T., Tansley, S., Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research (2009)
4. Ward, M.O., Grinstein, G., Keim, D.: Interactive Data Visualization: Foundations, Techniques, and Applications. CRC Press, Natick (2010)
5. 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.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 117–140. Springer, Heidelberg (2014)
6. Unger, A., Schumann, H.: Visual support for the understanding of simulation processes. In: IEEE Pacific Visualization Symposium, PacificVis 2009, pp. 57–64. IEEE (2009)
7. Bernard, J., Daberkow, D., Fellner, D., Fischer, K., Koepler, O., Kohlhammer, J., Runnwerth, M., Ruppert, T., Schreck, T., Sens, I.: VisInfo: a digital library system for time series research data based on exploratory search - a user-centered design approach. Int. J. Digit. Libr. **1**, 37–59 (2015). Springer
8. Bernard, J., Ruppert, T., Scherer, M., Kohlhammer, J., Schreck, T.: Content-based layouts for exploratory metadata search in scientific research data. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 139–148. ACM, June 2012
9. Scherer, M., von Landesberger, T., Schreck, T.: Visual-interactive querying for multivariate research data repositories using bag-of-words. In: Proceedings of ACM/IEEE Joint Conference on Digital Libraries, pp. 285–294 (2013)
10. Shao, L., Behrisch, M., Schreck, T., von Landesberger, T., Scherer, M., Bremm, S., Keim, D.: Guided sketching for visual search and exploration in large scatter plot spaces. In: Proceedings of EuroVA International Workshop on Visual Analytics, pp. 19–23 (2014)
11. Kandel, S., Paepcke, A., Hellerstein, J., Wrangler, J.H.: Interactive visual specification of data transformation scripts. In: ACM Human Factors in Computing Systems (CHI) (2011)
12. Jeanquartier, F., Jean-Quartier, C., Holzinger, A.: Integrated Web visualizations for protein-protein interaction databases. BMC Bioinform. **16**(1), 195 (2015). doi:10.1186/s12859-015-0615-z
13. Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Tegnér, J.: Data integration in the era of omics: current and future challenges. BMC Syst. Biol. **8**(Suppl. 2), I1 (2014)
14. Angrist, M., Cook-Deegan, R.: Distributing the future: the weak justifications for keeping human genomic databases secret and the challenges and opportunities in reverse engineering them. Appl. Transl. Genomics **3**(4), 124–127 (2014)

15. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Antipin, Y.: The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. **2**(5), 401–404 (2012)
16. Cline, M.S., Craft, B., Swatloski, T., Goldman, M., Ma, S., Haussler, D., Zhu, J.: Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. Sci. Rep. **3**, 2652 (2013)
17. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Mc Henry, K.T.: The landscape of somatic copy-number alteration across human cancers. Nature **463**(7283), 899–905 (2010)
18. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Kok, C.Y.: COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. **43**(D1), D805–D811 (2015)
19. Zhang, J., Baran, J., Cros, A., Guberman, J.M., Haider, S., Hsu, J., Wong-Erasmus, M.: International Cancer Genome Consortium Data Portala one-stop shop for cancer genomics data. Database (Oxford) (2011) bar026
20. Rubio-Perez, C., Tamborero, D., Schroeder, M.P., Antoln, A.A., Deu-Pons, J., Perez-Llamas, C., Lopez-Bigas, N.: In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell **27**(3), 382–396 (2015)
21. Thorvaldsdttir, H., Robinson, J.T., Mesirov, J.P.: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings Bioinform. **14**(2), 178–192 (2013)
22. Dietmann, S., Lee, W., Wong, P., Rodchenkov, I., Antonov, A.V.: CCancer: a birds eye view on gene lists reported in cancer-related studies. Nucleic Acids Res. **38**(Suppl. 2), W118–W123 (2010)
23. Jiang, G., Sohn, S., Zimmermann, M.T., Wang, C., Liu, H., Chute, C.G.: Drug normalization for cancer therapeutic and druggable genome target discovery. AMIA Summits Transl. Sci. Proc. **2015**, 72 (2015)
24. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L., Prior, F.: The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J. Digit. Imaging **26**(6), 1045–1057 (2013)
25. Ongenaert, M., Van Neste, L., De Meyer, T., Menschaert, G., Bekaert, S., Van Criekinge, W.: PubMeth: a cancer methylation database combining text mining and expert annotation. Nucleic Acids Res. **36**(Suppl. 1), D842–D846 (2008)
26. Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Shen, B.: Biomedical text mining and its applications in cancer research. J. Biomed. Inform. **46**(2), 200–211 (2013)
27. Pletscher-Frankild, S., Pallej, A., Tsafou, K., Binder, J.X., Jensen, L.J.: DISEASES: text mining and data integration of diseasegene associations. Methods **74**, 83–89 (2015)
28. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical text mining: state-of-the-art, open problems and future challenges. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 271–300. Springer, Heidelberg (2014)
29. Torre, L.A., Siegel, R.L., Ward, E.M., Jemal, A.: Global cancer incidence and mortality rates and trendsan update. Cancer Epidemiol. Biomark. Prev. **25**(1), 16–27 (2016)
30. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. CA: A Cancer J. Clin. **66**(1), 7–30 (2015)

31. Bray, F., Ferlay, J., Laversanne, M., Brewster, D.H., Gombe Mbalawa, C., Kohler, B., Soerjomataram, I.: Cancer incidence in five continents: inclusion criteria, highlights from Volume X and the global status of cancer registration. Int. J. Cancer **137**(9), 2060–2071 (2015)
32. Europe PMC Consortium: Europe PMC: a full-text literature database for the life sciences and platform for innovation. Nucleic Acids Res. **43**(D1), D1042–D1048 (2015)
33. Holzinger, A., Jurisica, I.: Knowledge discovery and data mining in biomedical informatics: the future is in integrative, interactive machine learning solutions. In: Holzinger, A., Jurisica, I. (eds.) Knowledge Discovery and Data Mining. LNCS, vol. 8401, pp. 1–18. Springer, Heidelberg (2014)
34. Kieseberg, P., Weippl, E., Holzinger, A.: Trust for the doctor-in-the-loop. In: European Research Consortium for Informatics and Mathematics (ERCIM) News: Tackling Big Data in the Life Sciences, vol. 104(1), pp. 32–33 (2016)
35. Greiling, D.A., Jacquez, G.M., Kaufmann, A.M., Rommel, R.G.: Space-time visualization and analysis in the Cancer Atlas Viewer. J. Geogr. Syst. **7**(1), 67–84 (2005)
36. Wei, Y.: Integrative analyses of cancer data: a review from a statistical perspective. Cancer Inform. **14**(Suppl. 2), 173 (2015)
37. Wu, T.J., Schriml, L.M., Chen, Q.R., Colbert, M., Crichton, D.J., Finney, R., Mitraka, E.: Generating a focused view of disease ontology cancer terms for pan-cancer data integration and analysis. Database (2015) bav032
38. Sioutos, N., de Coronado, S., Haber, M.W., Hartel, F.W., Shaiu, W.L., Wright, L.W.: NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J. Biomed. Inform. **40**(1), 30–43 (2007)
39. Drake, J.W., Charlesworth, B., Charlesworth, D., Crow, J.F.: Rates of spontaneous mutation. Genetics **148**(4), 1667–1686 (1998)
40. Lodish, H., Berk, A., Zipursky, S.L., et al.: Molecular Cell Biology, 4th edn. W.H. Freeman, New York (2000)
41. Yang, Y., Dong, X., Xie, B., Ding, N., Chen, J., Li, Y., Fang, X.: Databases and web tools for cancer genomics study. Genomics Proteomics Bioinform. **13**(1), 46–50 (2015)
42. Müller, H.M., Kenny, E.E., Sternberg, P.W.: Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol. **2**(11), e309 (2004)
43. Schaefer, C., Grouse, L., Buetow, K., Strausberg, R.L.: A new cancer genome anatomy project web resource for the community. Cancer J. **7**(1), 52–60 (2001)
44. Bult, C.J., Krupke, D.M., Begley, D.A., Richardson, J.E., Neuhauser, S.B., Sundberg, J.P., Eppig, J.T.: Mouse Tumor Biology (MTB): a database of mouse models for human cancer. Nucleic Acids Res. **43**(D1), D818–D824 (2015)
45. Roelofs, E., Dekker, A., Meldolesi, E., van Stiphout, R.G., Valentini, V., Lambin, P.: International data-sharing for radiotherapy research: an open-source based infrastructure for multicentric clinical data mining. Radiother. Oncol. **110**(2), 370–374 (2014)
46. WHO cancer mortality database (IARC). http://www-dep.iarc.fr/WHOdb/WHOdb.htm. Accessed 01 May 2016
47. Eyler, C.E., et al.: Glioma stem cell proliferation and tumor growth are promoted by nitric oxide synthase-2. Cell **146**(1), 53–66 (2011)
48. Herman, A.B., Savage, V.M., West, G.B.: A quantitative theory of solid tumor growth, metabolic rate and vascularization. PLOS One **6**, e22973 (2011)

49. Kisker, O., Becker, C.M., Prox, D., Fannon, M., D'Amato, R., Flynn, E., Fogler, W.E., Kim Lee Sim, B., Allred, E.N., Pirie-Shepherd, S.R., Folkman, J.: Continuous administration of endostatin by intraperitoneally implanted osmotic pump improves the efficacy and potency of therapy in a mouse xenograft tumor model. Cancer Res. **61**, 7669 (2001)
50. Mroz, E.A., Tward, A.M., Hammon, R.J., Ren, Y., Rocco, J.W.: Intra-tumor genetic heterogeneity and mortality in head and neck cancer: analysis of data from the cancer genome atlas. PLoS Med. **12**(2), e1001786 (2015)
51. Aigner, W., Miksch, S., Schumann, H., Tominski, C.: Visualization of Time-oriented Data. Springer Science & Business Media, New York (2011)
52. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform. **3**(2), 119–131 (2016). Springer
53. Jean-Quartier, C., Jeanquartier, F., Cemernek, D., Holzinger, A.: Tumor growth simulation profiling. In: Renda, M.E., Bursa, M., Holzinger, A., Khuri, S. (eds.) ITBAM 2016. LNCS, vol. 9832, pp. 208–213. Springer, Heidelberg (2016)

## 2.6. In silico modeling for tumor growth visualization

Within this journal publication we describe an in silico modeling approach to tumor growth with the help of a visualization and simulation tool. I created the idea of this tool together with my sister, expert in molecular biomedical science. By being supervising a computer science student we achieved to create a first implementation. I further extended the user interface according to feedback provided by my sister. My sister and I also designed the first experiments and analyzed data together.

Jeanquartier *et al. BMC Systems Biology* (2016) 10:59
DOI 10.1186/s12918-016-0318-8

BMC Systems Biology

RESEARCH ARTICLE Open Access

# In silico modeling for tumor growth visualization

Fleur Jeanquartier[1†], Claire Jean-Quartier[1†], David Cemernek[1] and Andreas Holzinger[1,2*]

## Abstract

**Background:** Cancer is a complex disease. Fundamental cellular based studies as well as modeling provides insight into cancer biology and strategies to treatment of the disease. In silico models complement in vivo models. Research on tumor growth involves a plethora of models each emphasizing isolated aspects of benign and malignant neoplasms. Biologists and clinical scientists are often overwhelmed by the mathematical background knowledge necessary to grasp and to apply a model to their own research.

**Results:** We aim to provide a comprehensive and expandable simulation tool to visualizing tumor growth. This novel Web-based application offers the advantage of a user-friendly graphical interface with several manipulable input variables to correlate different aspects of tumor growth. By refining model parameters we highlight the significance of heterogeneous intercellular interactions on tumor progression. Within this paper we present the implementation of the Cellular Potts Model graphically presented through Cytoscape.js within a Web application. The tool is available under the MIT license at https://github.com/davcem/cpm-cytoscape and http://styx.cgv.tugraz.at:8080/cpm-cytoscape/.

**Conclusion:** In-silico methods overcome the lack of wet experimental possibilities and as dry method succeed in terms of reduction, refinement and replacement of animal experimentation, also known as the 3R principles. Our visualization approach to simulation allows for more flexible usage and easy extension to facilitate understanding and gain novel insight. We believe that biomedical research in general and research on tumor growth in particular will benefit from the systems biology perspective.

**Keywords:** Cancer, Tumor growth, In silico, In silico medicine, Visualization, Visual analysis, Computational biology, Cellular Potts model, Glazier and Graner model, Cell proliferation

## Background

Around 13 % of all deaths worldwide are due to cancer [1]. Cancer depicts a group of diseases which refer to abnormal new growth of cells which can spread and invade different areal parts throughout the body [2]. A tumor is most commonly described as an abnormal growth of clustered cells which can be either benign (well-structured and non-harmful) or malignant (cancerous) [3]. Treatment against cancer directly relates to the growth-behaviour rendering the onset of therapy critical for its outcome. As a matter of fact, oncology is primarily based on prediction aspects [4]. In this regard, we focus on the assessment and prediction of tumor growth. The growth of tumors depends on their supply of oxygen, nutrients as well as survival factors and is influenced by growth factors as well as its local surroundings [5]. Characteristics are individually based on the different types of tumors [6]. The mathematical basis for tumor growth has been described in the mid of the last century not to be exclusively exponential but to be following a continuous deceleration as presented by the Gompertz function [7, 8]. Modern approaches, for example, take the heterogeneous subclonal mixtures [9] of tumor cells into account or even its interdependency to cellular motility [10]. Our model includes basic ideas of tumor growth, set for further enhancement through multiple expansion possibilities. We apply in-silico modeling of tumor-growth as a primary

*Correspondence: a.holzinger@hci-kdd.org
†Equal contributors
[1] Holzinger Group, Research Unit HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Auenbruggerplatz 2/V, 8036 Graz, AT, Austria
[2] Institute of Information Systems and Computer Media, Graz University of Technology, Inffeldgasse 16c, 8010 Graz, AT, Austria

tool, and further advance it to a novel Web-based simulation, evenhandedly available for biomedical scientists and clinicians with a focus on feature visualization. Features are key to learning and understanding. Thus, features are of enormous importance for knowledge discovery.

#### Computational modeling in biomedical science

These days, biomedical science heavily relies on computerized support for analyzing big data, quantifying dynamic and multiscale events, or likewise for simulating complex models. Computational models have been applied for intra- and inter-cellular, tissue- and organ-specific aspects [11]. Additionally, there is the ongoing project of creating a virtual physiological human [12] in order to support clinical decision-making. The project includes multi-level modeling of a wide range of information dealing with patient-specific signaling and genetic data up to whole-organ physiological mechanisms.

There are two main advantages of the bioinformatic approach in computational modeling of disease. First, simulations can be used for predictions in regard to the basic idea of alternative testing methods in addition to or instead of laborious experimentation. Alternative testing methods comprise the categories of replacement, reduction as well as refinement of in-vivo experimentation, that are summed up by the 3R principle [13, 14]. Thereby, in-silico methods are applied to in-vivo and in-vitro extrapolation [15, 16]. Secondly, prediction models overcome the lack of experimental methods for insufficient or nonexistent early screening tests. In general, models can be used to gain insight into complex biological systems and may address the gaps in literature as well as form the foundation for future research [17–19]. Simplification and approximation of the numerous detailed information gained from biomedical science offers the possibility to patient-personalized prediction, avoids hard-to-measure variables or compensates non-measurable factors [20]. Still, models are, so far, inflexible to simple extensions or even rescaling. Furthermore, we have to overcome the conflict between complexity and oversimplification. For instance, global mapping of cell community is computationally too laborious while the averaged approach lacks detailed description of molecular variables [21]. Still, in silico modeling and other computational techniques help answering key questions in cancer research [22–25].

We emphasize the approach of computational modeling of biological systems and developing computational modeling tools for simulation and reproducibility of experiments in biologic research. Fisher et al. [26] coin the term Executable Biology which highlights the difference between mathematical and computational biomodels in regard to their representation. Executable Biology describes computational algorithms in support to reproducible results in biomedical research as well as efficient simulation and analysis of biological systems. In this regard, Executable Biology is recommended to be integrated as standard method into bio-science.

Regarding the dynamics of tumor growth, computational models for various types of tumors exist, from animal models and the human body, dealing with the individual stages of tumor development [27]. In silico cancer modeling provides significant opportunities, however, Edelman et al. [28] argue that it is yet in its infancy.

Understanding the tumor heterogeneity with respect to personalized cancer treatment represents the ultimate goal of computational tumor-growth modeling. For that matter, multiple groups of scientist have to work together, accentuating the need for interchangeable infrastructure of linking big data and adoptable specialized models [29].

#### Mathematical modeling of tumor growth

Tumor growth kinetics follow relatively simple laws that can be mathematically described [30]. Such mathematical models could forecast individual phases of tumor growth [31]. In general, there are basic modeling approaches of cancer kinetics [28], that include exponential growth, the Gompertz model [32], metabolic models [33], the so-called universal model [34] and hybrid models [35]. Various mathematical models have been developed for the description and prediction of tumor growth. Each model, available so far, is optimized for specific scales of time and size plus certain aspects of metabolism or interactions [28, 35]. In regard to different biological scales, Deisboeck et al. [36] discuss innovative multi-scale cancer modeling approaches, ranging from atomic and molecular up to macroscopic scale. However, there is no universal law yet. Simple models have prediction rates less than 70 %, while some models used for specialized simulations achieve $\geq$ 80 % prediction rates [30]. Cancer models can be categorized based on their basic mechanisms to calculate tumor growth, but several additional factors have to be considered. Tumors originate from differentiating cells exhibiting the behavior of excessive proliferation up to migration [20, 37]. Tumors can be either dormant or growing [38, 39]. After reaching a critical mass, primary tumor growth stops and migration through metastasis will occur. From a biological perspective, tumor growth also depends on the underlying network structure [40–42].

#### Cellular Potts modeling of tumor-growth

The Cellular Potts model (CPM) poses a most widely used example of agent-based models which are feasible for research regarding cell-based phenomena and, therefore, are favorable for cancer research [43, 44]. The CPM was first presented by Graner and Glazier [45, 46]. The CPM or also named Glazier-Graner-Hogeweg (GGH) model

is based on individual cells in contrast to continuum models which summarize cell populations to tissues and continuous materials [47, 48]. It represents a modeling approach on tissue level with the main focus on intracellular and intercellular events as well as the cellular microenvironment. It has been implemented for tumor progression and invasion before [43]. The model includes single-cell characteristics of cellular geometry and interactions, rendering the simulation more efficient for questions on a detailed level than for a general overview. Glazier and Graner's model was originally developed for simulating the rearrangement of individual cells and cell sorting [46]. They upgraded the model to a compartmental view of cellular subelements. In principle, various cells are described as objects covering multiple shifting nodes on a 2D or 3D lattice while moving and changing their size. Thereby, CPM simulations support studies on type-specific cellular morphology and interaction [49]. The model describes different cell states and allows for additional parameters such as cell division and migration [50] as well as chemical diffusion and the extracellular matrix (ECM) [51]. Graner et al. [45] showed that differential cell adhesion and chemotaxis can be controlled through CPM, while the model is robust in regard to certain parameter choices. Glazier et al. [47] revise several development steps of the CPM and Szabo et al. [43] summarize the usefulness of CPM for simulating multi-cellular processes related to cancer. Boas et al. [52] recently conducted a global sensitivity analysis of the CPM, taking model extensions for angiogenesis into account, and showed that introducing a dynamic parameter for chemoattraction has the highest impact, being followed by the diffusion coefficient and cell-cell adhesion.

CPM has been used in a wide range of applications and there are extensions in terms of kinetics also referred to as extended CPM as well as hybrid CPM models [49]. The background of CPM modeling on cell sorting for various cell-types has been successfully used for the simulation of benign tumor growth [53] and cancer invasion [54]. Moreover, multiscale-models based on CPM have been implemented for various cancer-related studies [43, 51, 55–60].

#### Visualization for computational modeling

Visualization supports the understanding of biological data and provides insight into biological systems [61]. Visualization and computation mutually contribute to the sense-making process of biomedical analysts [62]. It is advised to provide integrated frameworks for biological studies. Graphical representations used for biological data visualization need to be adjusted to an appropriate level of detail. Graphs, in which each node represents a biological object and each edge a relation between these nodes, are often found in visualizations of biological data. While it has been primarily used for large interaction networks so far, graph visualization offers several user-friendly layout algorithms and is applicable for a wide are of application areas, ranging from social networks, finance to biology [61, 63]. Our recent study [64] on integrated visualization of biological networks highlights current possibilities for using Web technologies to support analysts in exploring biological relations.

The field of computational cancer biology lacks visualization types apart from network visualization. The "cBioPortal" with its focus on cancer genomics offers interactive visualization of pathway networks, mutations in protein domains, statistical information and trends on gene sets and clinical patient data of 10 published cancer studies [65]. Besides, there are only a few attempts on integrating visualization in computational modeling tools for cancer biology. Simulation results of a multiscale model for glioma growth have been visualized by the use of the software SciRun [66]. Specific cell growth processes can be simulated and visualized with the tool CellSys [67]. CompuCell3D [68] and the Tissue Simulation Toolkit [69] are exemplary frameworks for testing and extending computational models, integrating visualization features on cell interactions for simulation and analysis. Last but not least, there have been efforts in developing a virtual biobank [70] and a cancer modeling community [36] to exchange data and to facilitate visualization integration.

Though computational modeling has become a feasible tool for tumor growth research, simulation tools are rare. There is a step by step tutorial available how to simulate collective cell behavior based on Cellular Potts modeling [71]. CompuCell3D is one of these tools which has been used for in silico modeling of cellular and multi-cellular behaviors [68]. The latter research group introduces a tutorial for building cell-based simulations for visualizing tumor growth by making use of an open source library for simulating the CPM, written in C++. Though providing step-by-step instructions, basic knowledge of the use of the terminal and a C++ compiler are required. This technical know-how is often a limitation to clinicians and researchers in biomedical sciences. Moreover, they do not describe how to create iterative computations and how to differentiate between cell-types.

However, despite the availability of many different tumor growth models on the one hand and many Web-based visualization libraries on the other hand, adequate and usable simulation tools are still rare. To our knowledge, there have been no efforts in creating easy to use, Web-based computational cancer modeling tools that integrate visualization features. Our main idea is creating usable and extendable implementations of tumor models to foster ease of use of simulations and support knowledge discovery.

## Methods

### Mathematical basis of tumor growth

In general, tumor growth is mathematically summarized by the Gompertz function [7, 8, 32, 43]:

$$\frac{V_t}{V_0} = e^{\frac{a}{x*\left(1-e^{-at}\right)}}$$

with tumor size at variable time $V_t$ and the initial tumor size $V_0$, $a$ and $b$ being tumor-type characteristic constants, for cell clone division [7, 8]. In detail, we choose to describe tumor growth using the CPM by GGH where the probability for a spin copy and therefore cell proliferation is expressed as:

$$p\left(\sigma_{i,j} \rightarrow \sigma_{i',j'}\right) = \begin{cases} e^{\frac{-\Delta H}{T}} & \text{if } \Delta H > 0; \\ 1 & \text{if } \Delta H \leq 0; \end{cases} \quad (1)$$

The CPM is a time-discrete markov chain and its transitions $\sigma_{i,j} \rightarrow \sigma_{i',j'}$ are calculated by a Hamiltonian (or energy) function $\Delta H$, a sum of several terms [46, 47]. We further describe details on its implementation within the next subsection.

### Implementation of the CPM

The Potts model is based on the differential adhesion hypothesis which states that motile cells rearrange themselves according to the lowest energy configuration along the potential energy landscape [46, 72]. Within the CPM by GGH, cells are assigned certain spin states. Cells are build up by multiple cellular bricks, likewise termed (cellular) lattice nodes, sites or points. A multi-scale growth is accomplished through surface adhesion and space competition of *cellular bricks* scattered through the discrete lattice. *Cellular bricks* are associated with spins at lattice sites. Spins can be flipped between spin states allocating a celluar brick to another cell. These spin-copy attempts are calculated through Monte Carlo Steps (MCS). MCS are the mathematical basis for the probability simulation. The key parts of the computation are the Hamiltonian function $\Delta H$, also referred to as configuration energy [47], shown in Eq. 2, and the temperature $T$ shown in Eq. 3.

$$H = J \sum_{i,j} \left(1 - \delta_{\sigma_{i,j}\sigma_{i',j'}}\right) \quad (2)$$

If $\Delta H < 0$ the new spin state is always accepted because the system's energy will be decreased. If $\Delta H \geq 0$ the new spin state is accepted with a certain probability. While the cell is growing its target volume increases too. A *cell* in the CPM is the set of all *cellular bricks* with the same cell-index. Each cell relates to a certain cell-type. The cell-types are defined by the set $\tau$.

$\Delta H$ constitutes the energy of interactions between cellular bricks $i$ with the neighbour $j$. The discrete version of the Kronecker delta $\delta = 1$ if two neighbouring bricks are from the same cell, otherwise $\delta = 0$.

A cell will reach a critical point for division upon minimum $\Delta H$. Each cellular brick is assigned a $\sigma_{i,j}$ with type-dependent interaction energies, the spin-spin coupling energy constants $J(\sigma_{i,j})$ to neighbouring cells. $J$ effects a cell to be inclined to comprise a formation of connected cellular bricks over loose entities.

*MCS* is a series of $n$ spin-copy attempts for a lattice consisting of $n$ lattice sites. Each MCS step resembles the rearrangement of cells and, therefore, the time. The calculation shown beneath includes the temperature $T$ which resembles a cellular motility factor [47]. The MCS calculates a change in configuration of $H_0$ to $H_1$ for:

$$\Delta H = H_1 - H_0 \leq 0 \text{ or otherwise } p = e^{\frac{-\Delta H}{T}} \quad (3)$$

The CPM Hamiltonian $H$ is the sum of a series of terms that are related to different cell attributes such as interaction energy as well as volume. Extended versions exist that include other addends [49]. The original CPM includes a second term next to the first term of all surface energies $J$. $H$ also includes a $\lambda$ as cellular constraint as function of elasticity, shown in Eq. 4.

$$H = J \sum_{i,j} \left(1 - \delta_{\sigma_{i,j}\sigma_{i',j'}}\right) + \lambda \sum_{\sigma} (\nu(\sigma) - V_t(\sigma))^2 \quad (4)$$

In more detail, $H$ includes the number of lattice sites $\nu(\sigma)$ in a given domain with the spin $\sigma$, and the target number $V_t(\sigma)$ within that domain. The second term confines a cell's volume $\nu$ to the range of a specific target volume $V$, while the variable $\sigma i'j'$ sums up the number of neighbours. We focus on a schematic two-dimensional cellular grid. A cell's volume $\nu$ and target volume $V_t$ is thereby reduced to area $a$ and $A_t$.

### Web-based model implementation

The implementation for the purpose of visual analysis of tumor growth includes:

- CPM implementation, based on Glazier et al. 1993 [46]
- Servlet for client-server communication
- Network visualization based on Cytoscape.js [73]
- Line chart visualization based on Flot [78]
- HTML5 frontend
- Tests
- Documentation

The CPM is implemented as server side backend. Thereupon a cross-browser user interface integrates client side visualization libraries for multiple visualization outputs (Fig. 1).

The presented tool *cpm-cytoscape* offers an HTML5 based graphical user interface that makes use of JavaScript (JS) libraries, first and foremost *Cytoscape.js*. Below the frontend, the backend is implemented in JAVA and

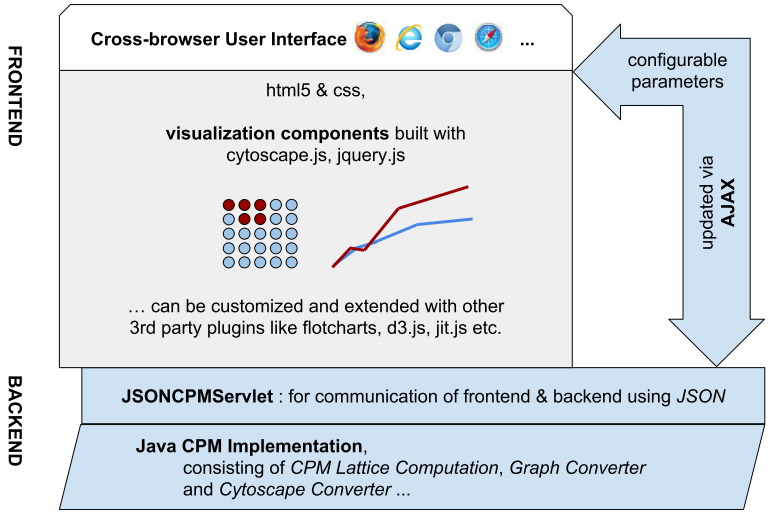


**Fig. 1** Architectural Representation of cpm-cytoscape: The architecture is composed of two distinct layers: The *frontend* layer contains a Cross-browser web presentation layer. It contains the customizable visualization components as well as a dynamic user interface for interacting with the CPM implementation via AJAX (with GET/POST). At the *backend* layer the JSONCPMServlet serves as interface for the actual CPM Lattice computation in the backend

information between frontend and backend is exchanged in JavaScript Object Notation (JSON), a common data exchange format that is used by Cytoscape.js. The JSON data holds a reference for the output container as well as several elements. The elements further contain child elements such as the complete set of edges and nodes, while each node again contains data about id, position, color, neighbour, parent, selection and other parameters. Moreover, the JSON structure includes information about the graph's layout and style parameters. By making use of a Java implementation of the CPM computation, a set of Java Servlets are requested asynchronously and delivering the data needed both for the computation in the backend and for the visualization rendering in the frontend.

### Visualization

We developed an HTML5 frontend that can be easily adjusted by means of modern web design via editing markup, JS and presentation stylesheets. The frontend can further be extended by integrating additional control elements as well as by making use of additional JS-based visualization libraries. For the visualization we searched for a library capable of rendering nodes along a lattice, and we found Cytoscape.js to be the graph visualization library of our choice. We use visualization libraries to create and update the visualization during a simulation run. The rendering method requests the *JSONCPMServlet*, a Java servlet that delivers data needed for the frontend rendering. Therefore, the JSONCPMServlet first receives JSON data, parses it, maps it and sends it back as JSON, that is then used for the graph rendering. For now, the frontend rendering parts include a graph visualization and a simple line chart. We use *Cytoscape.js* to plot the lattice-based graph visualization as well as *Flot*, a *Jquery.js* extension, to draw simple line charts.

### Usage of cpm-cytoscape

Based on a study on a brain cancer type modelled by CPM [51] and our ongoing work on tumor growth profiles for simulation [74] we introduce the tool through a short tutorial at https://github.com/davcem/cpm-cytoscape. We encourage readers to use GitHub for having a closer look at our implementation, explore its features and suggest enhancements as well as participate in the development. Design and implementation of the presented tool took place in an iterative manner. Informal validations have been conducted by several discussions with a domain expert. The basic idea up to the model's implementation and the tool's user interface have been co-designed and reviewed by a domain-expert.

## Results

We present a new 2D visualization approach for a dynamic cellular model simulation that accounts for lattice size, cell size, environment parameters and interactions between cells. The tool developed and used for the simulations has been published in the GitHub repository, saved as cpm-cytoscape. It can be obtained via the url address: https://github.com/davcem/cpm-cytoscape. Further, we provide a demo version that is online available on: http://styx.cgv.tugraz.at:8080/cpm-cytoscape/.

We created the tool to allow for easy manipulation by its user. The upper region offers a number of variables which can be set by the user in order to discriminate and process various experiments. The CPM is computed solely in the Java backend, while initialization parameters can be adjusted in the frontend and are communicated by requesting the servlet. By varying several parameters the user is allowed to simulate a wide range of conditions. These parameters are the lattice's size (*x,y*), the count of monte carlo steps, its' substeps, *max* $\sigma$, the *matrix density*, *interaction parameters* as well as the *temperature*. The Java packages consist of the implementation of the CPM itself, a graph converter to convert the CPM lattice into a graph structure, a more specific cytoscape converter to represent the graph enrichment needed for the visualization library as well as the servlet to provide the communication interface between backend and frontend via JSON.

Individual cells are visualized as group of nodes, we refer to as *cellular bricks*, on a grid. Cytoscape.js provides a grid layout rendering algorithm that arranges the nodes in a square grid whereby the circular nodes represent subcompartments of cells. We differentiate between *light* cells that represent normal cells, *dark* cells that represent mutated cells and the ECM that surrounds cells. The ECM is represented as grey nodes. The other nodes with $\sigma \geq 1$ are represented by the colored, either dark or light nodes. For now, we only differentiate between a light and a dark cell-type. Nodes which are not indexed as light or dark cells are attributed to the ECM. They resemble the cellular surroundings without peculiar growth variables.

The *growth rate* can be visualized as line chart for $\sigma = 2$ by using the button "show line chart". The line chart shows the amount of computation steps on the x-axis and the amount of cellular bricks on the y-axis. Experimental data can be exported as spreadsheet in the format of comma-separated values. This option offers the possibility of making the data available offline for further analysis.

**Initialization and lattice settings:** The lattice is created on the left side of the browser window by pressing the button *initialize* (Fig. 2). Thereby, the size and likewise the number of nodes is determined by the input of variables *x* and *y*. This allows to adjust the experimental area. Nodes are indexed randomly to *light* and *dark* cells or *ECM* according to the input of the number of cellular clusters $\sigma$, *matrix density* and the *light/dark ratio*. After initializing a random graph according to the user interface's settings the computation possibilities with the button "compute next simulation run" and "compute next two simulation runs" are enabled (Fig. 2).

Our implementation of the CPM currently consists of *maxSigma* cells relating to 3 different cell-types, while $\sigma = 0$ attributes to the ECM, the odd numbers refer to dark cells and the even numbers to light ones. Therefore, by making use of the **max** $\sigma$ parameter one can also define more than two different cells, also referred to as cellular clusters. *Max* $\sigma$ defines the quantity of individual cell components or respectively cellular clusters. If max $\sigma$ is set to 2 we use the color lightblue for light (normal/healthy) cellular bricks and darkred for the dark (tumor/mutated) ones (Fig. 2). If max $\sigma$ is set to $> 2$ we use a colorscheme for coding dark and light cell nodes slightly differently to better distinguish between different $\sigma$, shown in Fig. 3. The factor $\sigma$ can be redefined to resemble the number of cell-types. The cell-types are represented by $\tau$, in some papers also referred to as cell or medium. We currently distinguish between three cell-types, namely ecm, light and dark cells as denoted in the original paper by Graner et al. [46]. A cell-type is referred to as $\tau_i$, while $\tau = \{0, 1, 2\}$ with $\tau_{i=0} = ECM$, $\tau_{i=1} = dark$, $\tau_{i=2} = light$.

The *matrix density* defines the number of cellular bricks indexed as light or dark cells in relation to the given number of nodes. Setting *matrix density* $= 1$ uses all lattice sites for cellular bricks. Setting *matrix density* $= 0$ represents a lattice site filled only with ECM.

The parameters *MCS* and *#substeps* represent units of time, while a substep is related to a random copy attempt. We implemented the number of *MCS* and *substeps* as variables and allow the parameters to be defined and adjusted by the user. Each *MCS* is divided into a specified amount of *substeps* for simulating different time settings.

The temperature $T$ functions as cellular motility factor since high $T$ leads to frequent spin-copies, thus, an increase in the number of cellular bricks and an increase in cellular invasive radius. The impact of $T$ on the overall run is highlighted in Fig. 4 (panel A). The default temperature is set to 10 degrees as suggested in [46, 75]. A comparison of our default settings with values, previously published by others, are summarized in Table 1.

The *parameter for area energy* $\lambda$ represents a limiting factor to cell growth, also termed cellular elasticity $\lambda$. Panel B in Fig. 4 demonstrates the impact of $\lambda$. High $\lambda$ values more strongly constrain cell growth while low $\lambda$ leads to frequent spin-copies. The target area $A_t$ is related to the lattice's size parameters *x* and *y*, while the target area factors for light and dark cells can be adjusted.

The *energy interaction parameter J* is the basis to the overall Hamiltonian and spin-copy attempts. This so-called boundary energy coefficient determines cell growth as multiplicative degree of freedom [47]. Panels C to F in Fig. 4 illustrate the impact of low and high interaction values for different cells as light and dark cells and ECM on the overall simulation outcome and the underlying Hamiltonian and spin-copy attempts. The impact on the simulation by the parameter variables are presented within Fig. 4.

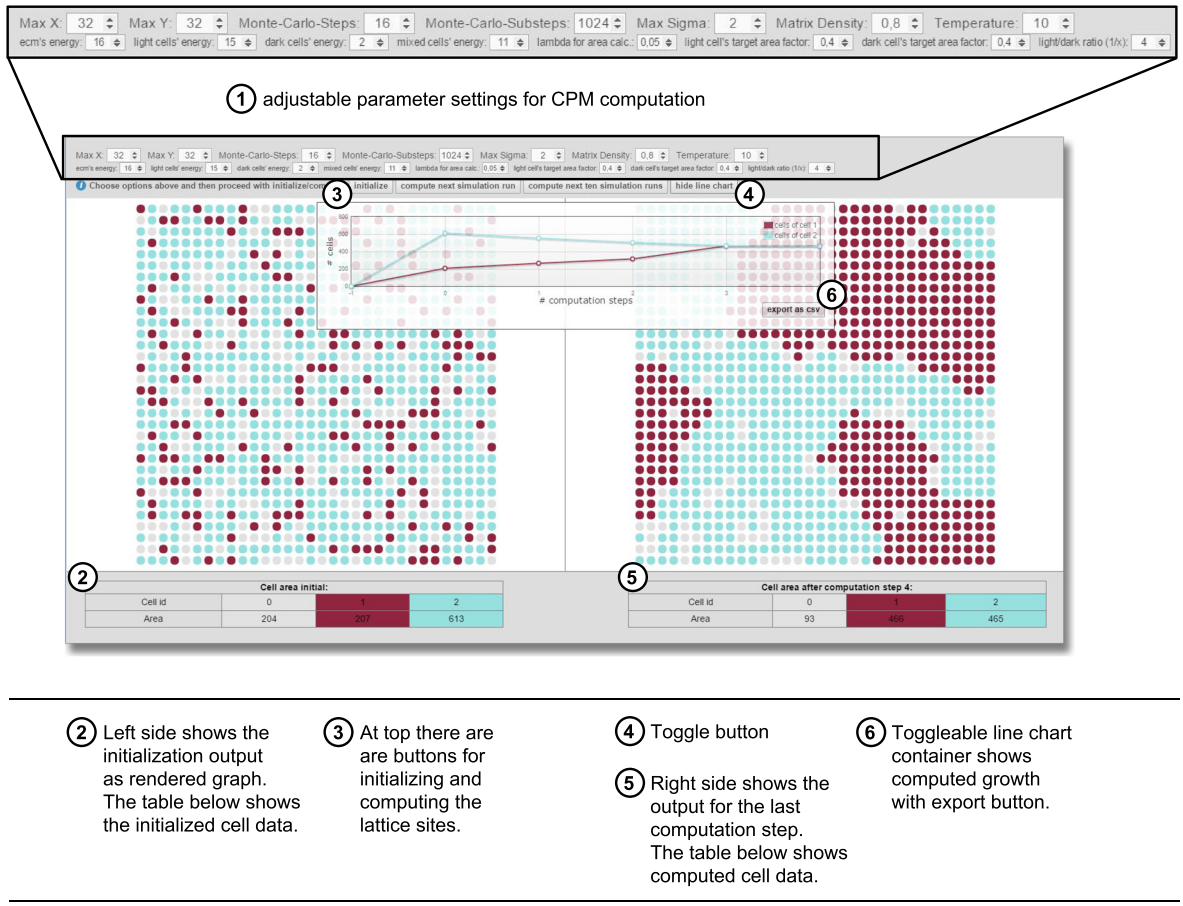


**Fig. 2** Overview of the tool's user interface: ① At top there are adjustable parameter settings for CPM computation and ③ buttons for initializing and computing the lattice sites. ② The left side shows the initialization output as rendered graph with grey nodes representing parts of ECM, while colored nodes representing cellular bricks corresponding either to light blue colored (normal/healthy) or dark red colored (tumor/mutated) cells. The table below shows information about the initialized cell data. ⑤ The right side shows the output for the last computation step, while the table below contains computed cell data. ④ A toggle buttons controls the ⑥ lightbox in the middle that provides line chart visualization and export

### Application example of cpm-cytoscape

We created a step-by-step tutorial on the presented tool using a tumor growth example based on parameters from a study on a brain cancer type modelled by CPM [51], available under https://github.com/davcem/cpm-cytoscape [74]. This example results in cellular growth of dark cells, representing tumor cells, showing a trend similar to Gompertz law. The simulated cancer cells thereby imitate 2D cultured glioma cells or likewise tumor-spheroids implanted in animals [51].

## Discussion

We present a web-based solution to allow for simple access to such a tumor growth visualization tool via Internet. By making use of the CPM implementation, we describe a potential use case for the cpm-cytoscape tool. The manipulable tool offers the advantage of adjustable settings for several input variables. By correlating various growth parameters we highlight the importance of heterogeneous cell interactions regarding its impact on tumor growth.

**Options to visualization:** There are many JS-based visualization libraries which can be used to foster the goals of visualization, namely to facilitate understanding and to gain novel insight, in our case into one of the many questions of biomedical research [76]. We make use of *Cytoscape.js* since it features user-friendly presentation of interaction data and supports several common browsers like Chrome, Firefox and Safari, while the first is the fastest one. It represents an open-source library on graph theory that was written in JS and developed for analysis and visualisation [73]. Thereby, layouts of the display area can be altered while graph elements can be accessed offering several possible operations including sorting and filtering as well as graph querying. These options can be exploited


**Fig. 3** Screenshot of a graph rendering with $\sigma_{max} = 14$: The graph consists of 14 distinct cells (also denoted to as cellular clusters). Each cell is represented by a certain amount of nodes that we call cellular bricks. Cellular bricks with dark red or purple color tones correspond to the dark (tumor/mutated) cells, while nodes in a light shade of blue to green are referencing the so called light (normal/healthy) cells. The amount of these are listed in the table below the graph visualization. For this example, the cell with $id = 9$, represented by the purple nodes, consists of 39 cellular bricks at the initialization phase. After 2 computation steps, we see at the right side, that the cell with $id = 9$ has grown and now holds 465 cellular bricks. Grey nodes represent the ECM

for future extensions to the tool. Moreover, Cytoscape.js [77] is regularly updated and supports directed as well as undirected, mixed or multi-graphs.

Furthermore, Cytoscape.js layouts can be easily changed by just specifying another graph layout for the layout parameter in the *cytoscapeRender* method. There are also alternative visualization libraries that can be used in the frontend [77–79]. Possible alternatives to Cytoscape's layout algorithm would be using a bubble chart layout or even a three dimensional surface plot layout that can be created with another JS library such as D3.js.

Cytoscape.js offers different layout rendering options out of the box. We chose to use the grid layout that fits into traditional CPM visualization. In general, tumor growth kinetics and effects of cell growth can be visualized as line chart with the two dimensions of volume/size or cell number over time [80]. Therefore, we use the extension of simple line charts. Time series visualization may help users from the fact that time spans and iterations can dynamically be adjusted and are neither restricted by sensory constraints nor by experiment and animal costs.

**Lattice-based visualization of cells:** The lattice is organized in two dimensions, since 2D-modeling reduces the computational load just as well as visualization comprehensiveness. Still, in terms of numbers, the model could be manually transcribed and extended to a third dimension as the need arises.

In a figurative sense, the lattice represents tissue in the biological context. Cellular bricks are translated as textural compartments of a biological cell-layer. By way of example, the two-dimensional cellular grid can then be described as representative cross-section translated from the possible style of tissue slices. In a conceptional matter of speaking, cellular bricks represent variable compartmental states of a cell that can be translated to several criteria such as the impact of genes or likewise proteins, effects by modulators, inhibitors as well as promoters, or localized phenomena in general. The specific factors can be applied and extended in regard to the individual focus of research in a problem-directed manner.

**Initialization and lattice-site settings:** The variable number of lattice sites offers the possibility to adjust the computational workload according to the requirements of individual questions. In difference to general computational models, the Web-based implementation is attempted to be computed with low latency. Good rendering performance of computation results is needed to create dynamic output for smaller lattice sizes at once, as well as to enable animation for multiple computation steps at once. Still, some experiments concerning specific timing problems will have to be conducted using a high number of nodes. Thus, the variables can be be chosen in compliance with the requirements.

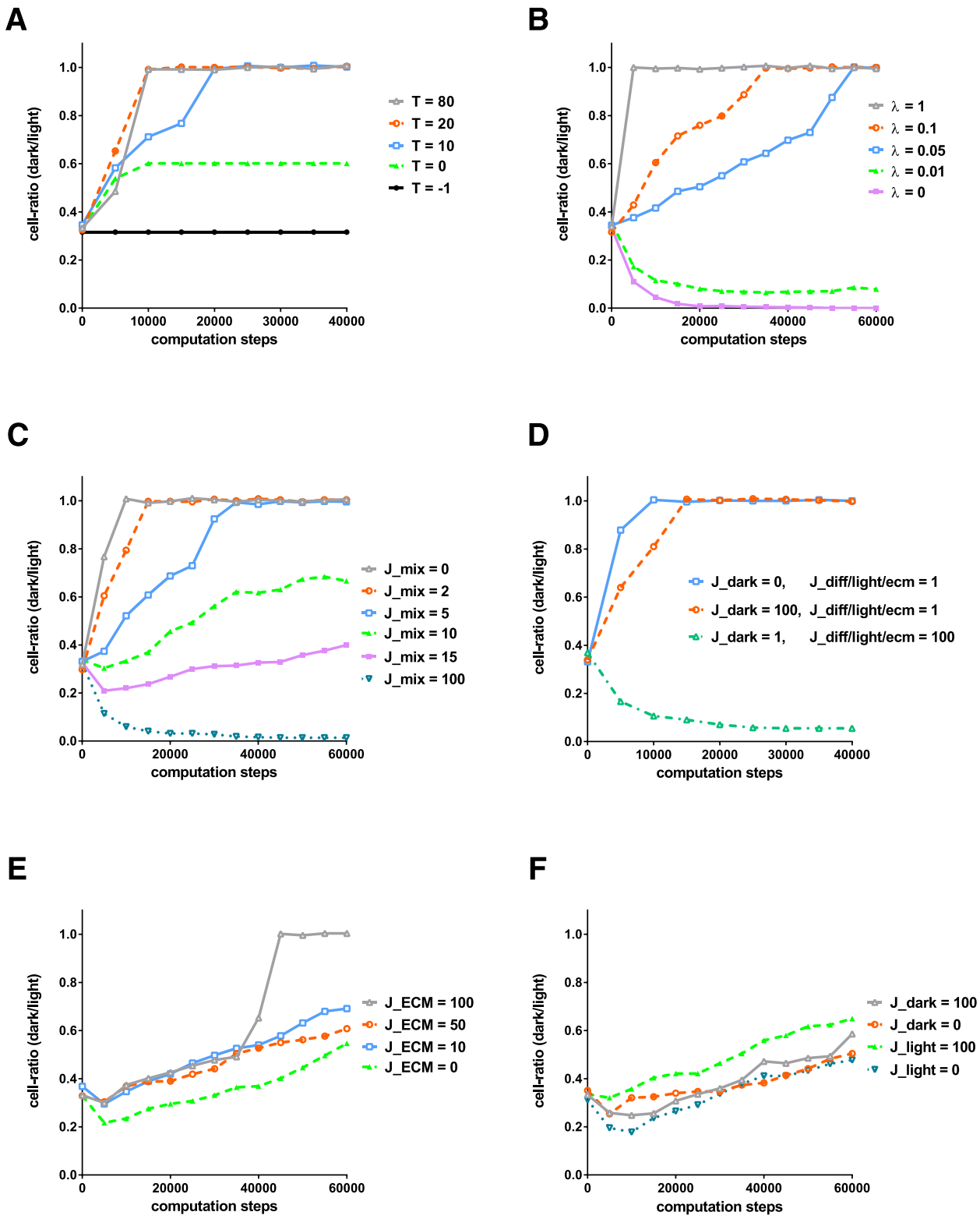


**Fig. 4** Cell growth in relation to varying parameters: line chart showing representative ratios between numbers of dark and light cellular bricks over computed steps. Comparison of varying parameters, for temperature $T = 80, 20, 10, 0, -1$ (panel A), $\lambda = 1, 0.1, 0.05, 0.01, 0$ (panel B), $J_{mix} = 0, 2, 5, 10, 15, 100$ (panel C), comparison of various *Js* as indicated for $J_{dark}$, $J_{light}$, $J_{mix}$, $J_{ecm}$ (panel D), $J_{ecm} = 100, 50, 10, 0$ (panel E), $J_{dark}$ and $J_{light}$ each 0 or 100 (panel F). Adjusted to default settings of *nodes* $= 32 * 32$, *mcs* $= 32$, *mcssubsteps* $= 64$, $\sigma_{max} = 2$, $\lambda = 0.05$, *targetAreas* $= 0.4$, initial *dark/lightratio* $= 1/4$

**Table 1** CPM parameter settings: comparison of presented default settings and values from literature [45, 46, 51, 75]

| | Max X * Y | MCS, substeps | max $\sigma$ | matrix density | T | $J_{ECM}$ | $J_{light}$ | $J_{dark}$ | $J_{mixed}$ | $\lambda$ | $A_{t(light)}$ | $A_{t(dark)}$ | $ratio_{light/dark}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| default settings | 32 * 32 | 32, 64 | 2 | 0.8 | 10 | 16 | 15 | 2 | 11 | 0.05 | 0.4 | 0.4 | 1/4 |
| GGH 1992 | 40 * 40 | 100, 1 | 2 | 1 | 10 | 16 | 14 | 2 | 11 | 1 | 0 | 0 | 1 |
| GGH 1993 | $\leq \sqrt{40*1000} * \sqrt{40*1000}$ | 16, max X * Y | 1000 | 1 | 5 | 8–16 | 14 | 2 | 11 | 1 | 40 | 40 | 1 |
| Ouchi 2003 | 128 * 128 | 1 , 1 | 16 | 1 | 10 | –/0 | –5 | –25 | –3 | 10 | 64 | 64 | 1 |
| Rubenstein 2008 | 500 * 500 | 400, Max X * Y | 65 | <0.1 | 0 | 0 | 2 | 2 | 9 | 1 | 40/2 | 50/2 | 1 |

The random distribution, to a certain degree, simulates environmental behaviour and the random occurrence of mutations within cells. Spheroid models start from an initial mass of proliferative cells only. Still, in nature, mutated cells showing abnormal growth are intermixed with "normal" cells. Thereby, our tool allows to set various cell-types. Tumor cells are set to grow by means of proliferation and further invasion. The ECM can be set as background or individual cells to be equally or inhomogeneously in size and distribution [51]. For the future, we plan to implement extensions that will include additional initialization settings, such as the introduction of a dynamically configurable cell-type or another dimension. Further variations could include the option of spheroid models. Another elaborate feature could even offer pre-defined cellular mixtures corresponding to uploaded images from treated tissue-slices.

**The impact and translation for *MCS* and *#substeps*:** A MCS's series of random copy attempts is equal to the total amount of cellular bricks. Graner and Glazier [45, 46] proposed *MCS* to be $16 \times x \times y$ while $x \times y \approx 1000$ and $x = y \approx< 40$ and did not make use of defining substeps. They suggested this setting for observing gradual movement behaviour. Later works define one *MCS* to consist of as many index-change attempts as the number of pixels in the lattice $x \times y$. If the setting for $MCS \times \#substeps$ is lower than $x \times y$, then unintended results are observed.

The time, by means of *MCS* steps, is an abstraction and relates to tumor specifics. The various kinds of tumor cells proliferate and divide more frequently than normal cells, depending on the localities and their differentiation status. Thereby, tumors can be classified by their spatial occurrence, and further, be characterized by their temporal growth dynamics. For each case, MCS steps can be translated to either hours, days or years. Future extensions to our tool will include pre-defined initialization settings of growth rates and time units corresponding to exemplary tumor types.

**Temperature *T*:** In general, temperature affects movement, and in our case, cell growth. In more detail, $T$ functions like a cellular motility factor since high $T$ will lead to frequent division of cells, thus, an increase in the number of cellular bricks and an increase in cellular invasive radius as shown in Fig. 4 (Panel A). If the interaction energy, represented by the several J parameters, is much greater than T, cells will shed into loose bricks at the boundaries. If T is too large, relative to J, boundaries will become stiff. Low temperatures inhibit proliferation. Subzero temperatures stop changing spin values and therefore kinetics and growth. At very low subzero temperatures, any biological activity is effectively stopped but cells could also take damage through freezing, that could be taken into account as additional factor in future studies.

**The energy interaction parameter *J*:** The range of the individual interaction energies is defined by the original cell-types as well as the manifested mutations responsible for the excessive proliferation by tumor cells. Thereby, these factors correlate with the class of tumor and it's tissue-residency. Individual cells exhibit heterogeneous tendencies towards growth correlating to tumor aggressivity, thus, interaction energies can vary over time. This phenomenon can be manually emulated by adjusting the individual interaction parameters after a specified number of MCS. Future extensions could include this adjustment as an automatic option in correlation to underlying relations of further variables.

In our case, default parameters of cpm-cytoscape implicate low values within the first term for the Hamiltonian computation, consisting of the interaction parameters $J$, in comparison to the second term, factoring values of area calculation such as $\lambda$, $a$ and $A_t$ (see details to Eq. 4). As can be seen in Fig. 4 (panel C) a change in $J_{mix}$, the interaction energy between different cells, impacts growth of dark cells considerably. However, there are no significant differences if the $J$ parameter of dark or light cells is changed selectively (panel F). Changes of $J_{ECM}$, the interaction energy between parts of ECM, result in similar insignificance, though high values can lead to sudden changes in the ratio between dark and light cells through dark cells migrating to and taking over former ECM space (panel E). Rather high values are needed to manipulate ratios. Figure 4 (panel D) demonstrates three cases of combined changes in the interaction parameters $J_{dark}$, the interaction energy of dark cells, in comparison to the interaction energy of light cells $J_{light}$, as well as $J_{mix}$ and $J_{ecm}$. The ratio between dark and light cells is only slightly decreased upon an 100-fold increase of $J_{dark}$. However, the number of dark cells over light cells is completely reduced upon increasing $J_{mix}$ and $J_{ecm}$. At the same time, the relation between $J_{dark}$ and $J_{light}$ plays a minor role in determining the probability of spin-copy attempts rather to their measure in proportion to $J_{mix}$ and $J_{ecm}$. This fact can be translated to the biological importance of heterogeneous interactions between cells and their environment. Further refinement will include the integration of additional parameters such as $J_{dark-ecm}$, $J_{light-ecm}$ or other $J_{diff}$ as well as the search for suitable realistic values to relate to different cell-types, a factor to be taken into account in future studies.

**The target area and the parameter for area energy λ:** The factor $\lambda$ is considered a constraint, in our case, for limiting cell growth. The so-called cellular elasticity $\lambda$

attaches the value of area calculation within the Hamiltonian computation. Differences between current and target area will likely have more effect on spin-copy attempts if $\lambda$ is high. If $\lambda$ becomes too high relative to the residual calculation parameters, any spin-copy attempt should be refused. This is true as long as the cellular area is different from the target area. The quadratic function does not distinguish whether the cellular area is larger or smaller than the target area. In terms of cell size, cellular elasticity will play a major role for rigid cells which tend to stay within the range of their target volume. Cell growth and division are correlated so that cells of unequal size will divide at a given speed and even out to a mean cell size. This is true only, if cell growth rate is constant. An abnormal increase in cell size is possible under the influence of excessive discharge of growth hormones or similar pathological circumstances such as hypertrophy. Other cases of instant changes in cell size include the natural processes of cellular differentiation and enlargement or shrinkage according to the metabolic state.

Generally, various cell-types are differently sized. Some cancers are known to manifest giant cells. Even normal cells exhibit different dimensions according to their origin. Cell diameters range from $1\mu m$ to $1mm$ and more, for instance nerve cells can reach a length over $1m$ [81]. Furthermore, cell-sizes vary within one cell-type. Still, cells have medial sizes specific to their type. This constraint is thereby necessary to limit cellular growth to an underlying biological scale.

For future matters, the discrete view of cellular area can have a completely different meaning. Cellular bricks resemble conceptional factors that occur or are replaced, distributed or accumulated within individual cells. These factors will be assigned by the researcher depending on a given task and scope of work.

**The ECM** occupies space which is not attributed to cellular clusters. Its energy area is initially suppressed, but if reprogrammed to a positive number within the source code, the ECM will grow and spread like light and dark cells. This could simulate gap-filling after cell-death and be the case of radiation procedures, cellular starvation or exposure effects of chemicals. This variation will be of importance in future studies introducing multiple affectors of cell growth by integration of biomedical databases, including drug, protein and genetic information related to tumor growth.

**The matrix density** was introduced as factor for simulating various cell densities within the area of interest. For instance, tissue slices could show distinct cellular colonization in locally fragmented patterns. Moreover, different cell-types as well as organelles can exhibit various densities. In general, varying cell densities can be attributed to the water content relative to the mass of proteins, nucleotides, carbohydrates or lipids within and around the cells.

Cell density often resembles the proliferative state of cells controlling protein expression. Consequently, the change in matrix density can be used for future studies focusing its effect on tumor growth, dormancy or metastasis. Further, matrix density can be interpreted in a more formalized manner, such as the variable abundance and occurrence of discrete factors within cellular regions.

**The role of fostering in silico modeling:** There is a trend towards computational simulations of biological processes making use of different mathematical models [82]. In particular simulation-based experiments in the field of bioinformatical cancer research can save resources in terms of time and costs. Collaboration between experimentalists and modelers has to be promoted and extended. This fact is most interesting for fostering cooperation of researchers from the interdisciplinary fields of computer science, mathematics, human-computer interaction, life sciences and biomedicine [83].

The tool represents a basic instrument to supporting biomedical researches and a preliminary step towards supporting clinical scientists. Until now, the tool has not been evaluated by clinicians. Future plans are to conduct further iterative testing and verification and to experiment with machine learning approaches [84].

## Conclusion

Recent advances in Computational Biology show high potential to deepen the understanding of origin and progression of cancer. Our general aim is to enrich cancer research by providing a tool that will make Computational Biology applicable to both researchers and clinicians. We focus on the fundamental pathological processes of cancer which are represented by tumor growth. Since abnormal cell growth involves chaotic, heterogeneous and highly differentiated structures, we chose to investigate cellular growth on the single-cell level. By refining model parameters of the cellular potts model, we highlight the impact of heterogeneous interceullular interactions on tumor growth.

Herein, we describe the implementation of the CPM for the purpose of simulation and visual analysis of tumor growth and provide its sources on github. We chose the lattice-based visualization style as primary approach to present and display tumor growth for research purposes. The graph computation allows for multiple different visualization approaches. The user interface is highly adjustable and its implementation is designed to be extended. The possibilities and accessibility of our simulation and visualization approach might ultimately promote

researchers and practitioners to progressing the field of tumor research towards personalized medicine.

Our approach offers several potential future applications of studying tumor dynamics. First, we plan to implement more simplistic models in order to offer fast computations and visualizations. Secondly, we plan to integrate various profiles into the tool, to offer exemplary simulations on different types of tumors [74]. Next to iterative testing, profiles lead to the task of verification. Furthermore, the implementation of additional dynamic parameters may enhance the simulation's possibilities. Multiple optional features to modeling as well as visualization styles will provide preferential outcomes in regard to detailed information or fast overview performance. Another interesting step towards supporting researchers and clinicians is providing image loading and size detection of regions of interests as input parameter for the simulation. Future integrations will include biomolecular networks such as drug-protein impact or genetic alteration patterns. Harnessing tumor growth data and related gene data as well as providing an open source database for tumor growth related data [85] are big steps forward to supporting science collaborations and clinical applications, and finally help contributing to fight cancer.

We believe that our approach is a motivator for fostering in silico modeling towards 3R and a better understanding of tumor dynamics.

**Abbreviations**
CPM, cellular Potts model; ECM, extracellular matrix; GGH, Glazier-Graner-Hogeweg; JS, javascript; MCS, Monte Carlo step

**Acknowledgements**
We thank the Institute of Computer Graphics and Knowledge Visualization at the Graz University of Technology for providing the demo server.

**Availability of data and materials**
Source code is available at https://github.com/davcem/cpm-cytoscape.

**Author's contributions**
Conceived and designed the experiments: FJ, CJ, DC. Performed the experiments: FJ, CJ, DC. Analyzed the data: FJ, CJ, DC. Contributed to theoretic background and competitive research: FJ, CJ, AH. Wrote the paper: FJ, CJ, DC, AH. All authors read and approved the final manuscript.

**Competing interests**
The authors declare that they have no competing interests.

**Consent for publication**
Not applicable.

**Ethics approval and consent to participate**
Not applicable.

Received: 10 May 2016 Accepted: 12 July 2016
Published online: 08 August 2016

**References**

1. GLOBOCAN (IARC). Better understanding of the biology of cancer cells. Lakartidningen. 2000;97(28–29):3260–64.
2. Klein G. Better understanding of the biology of cancer cells. Ugeskr Laeger. 2000;162(39):5199–204.
3. Bloemena E. Cancer and oncogenesis. Ned Tijdschr Tandheelkd. 2008;115(4):180–5.
4. Vickers AJ. Prediction models in cancer care. CA Cancer J Clin. 2011;61(5):315–26.
5. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.
6. U.S. National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, U.S. Department of Health and Human Services, National Institutes of Health, Update Aug. 17, 2014 by Martin LJ. https://www.nlm.nih.gov/medlineplus/ency/article/001310.htm. Accessed 4 Aug 2016.
7. Laird AK. Dynamics of Tumour Growth. Br J Cancer. 1964;18(3):490–502.
8. Loeb L. Tissue Growth and Tumor Growth. J Cancer Res. 1917;2(135).
9. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, Marjoram P, Siegmund K, Press MF, Shibata D, Curtis C. A Big Bang model of human colorectal tumor growth. Nat Genet. 2015;47(3):209–16.
10. Waclaw B, Bozic I, Pittman ME, Hruban RH, Vogelstein B, Nowak MA. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. Nature. 2015;525(7568):261–4. doi:10.1038/nature14971. Epub 2015 Aug.
11. Holzhütter HG, Drasdo D, Preusser T, Lippert J, Henney AM. The virtual liver: a multidisciplinary, multilevel challenge for systems biology. Wiley Interdiscip Rev Syst Biol Med. 2012;4(3):221–35.
12. Hunter P, et al. A vision and strategy for the virtual physiological human. Interf Focus. 2010;368(1920):2595–2614.
13. Russell WMS, Burch RL. The Principles of Humane Experimental Technique. London: Methuen; 1959, pp. 69–154.
14. Tannenbaum J, Bennett BT. Russell and Burch's 3Rs Then and Now: The Need for Clarity in Definition and Purpose. J Am Assoc Lab Anim Sci (JAALAS). 2015;54(2):120–132.
15. Hunt CA, Ropella GEP, Ning Lam T, Tang J, Kim SHJ, Engelberg JA, Sheikh-Bahaei S. At the Biological Modeling and Simulation Frontier. Pharm Res. 2009;26(11):2369–2400.
16. Hosea NA, Jones HM. Predicting pharmacokinetic profiles using in silico derived parameters. Mol Pharm. 2013;10(4):1207–15.
17. Gong H, Clark EM. Computational Modeling and Verification of Signaling Pathways in Cancer. ANB. 2010;6479:117–135.
18. Hanin L. Seeing the invisible: how mathematical models uncover tumor dormancy, reconstruct the natural history of cancer, and assess the effects of treatment. Adv Exp Med Biol. 2013;734:261–82.
19. Salz T, Baxi SS, Raghunathan N, Onstad EE, Freedman AN, Moskowitz CS, Dalton SO, Goodman KA, Johansen C, Matasar MJ, de Nully Brown P, Oeffinger KC, Vickers AJ. Are we ready to predict late effects? A systematic review of clinically useful prediction models. Eur J Cancer. 2015;51(6):758–66.
20. Choe SC, Zhao G, Zhao Z, Rosenblatt JD, Cho H-M, Shin S-U, Johnson NF. Model for in vivo progression of tumors based on co-evolving cell population and vasculature, Scientific reports, Massachusetts, 2nd edition, Scientific reports 1: Nature Publishing Group; 2011.
21. Coveney PV, Fowler PW. Modelling biological complexity: a physical scientist's perspective. J R Soc Interface. 2005;2(4):267–80.
22. Du W, Elemento O. Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies. Oncogene. 2015;34(25):3215–25.
23. Wolkenhauer O, Fell D, De Meyts P, Blüthgen N, Herzel H, Le Novere N, Höfer T, Schürrle K, van Leeuwen I. SysBioMed report: advancing systems biology for medical applications. IET Syst Biol. 2009;3(3):131–6.
24. Friedman R, Boye K, Flatmark K. Molecular modelling and simulations in cancer research. Biochim Biophys Acta. 2013;1836(1):1–14.
25. Gago F. Modelling and simulation: a computational perspective in anticancer drug discovery. Curr Med Chem Anticancer Agents. 2004;4(5):401–3.
26. Fisher J, Henzinger TA. Executable cell biology. Nat Biotechnol. 2007;25(11):1239–49.
27. Enderling H, Rejniak KA. Simulating Cancer: Computational Models in Oncology. Front Oncol. 2013;3:233.
28. Edelman LB, Eddy JA, Price ND. In silico models of cancer. Wiley Interdiscip Rev Syst Biol Med. 2010;2(4):438–59.
29. Sakkalis V, et al. Web-based workflow planning platform supporting the design and execution of complex multiscale cancer models. IEEE J Biomed Health Inform. 2014;18(3):824–31.

30. Benzekry S, Lamont C, Beheshti A, Tracz A, Ebos JML, Hlatky L, Hahnfeldt P. Classical mathematical models for description and prediction of experimental tumor growth. PLoS Comput Biol. 2014;10(8): e1003800. doi:10.1371/journal.pcbi.1003800.
31. Hanin L, Seidel K, Stoevesandt D. A "universal" model of metastatic cancer, its parametric forms and their identification: what can be learned from site-specific volumes of metastases. J Math Biol. 2015;72(6):1633–62.
32. Gocka EF, Reed LJ. A method of fitting non-symmetric Gompertz functions for characterising malignant growth. Int J Biomed Comput. 1977;8(4):247–54.
33. Resendis-Antonio O, González-Torres C, Jaime-Munoz G, Hernandez-Patiño CE, Salgado-Muñoz CF. Modeling metabolism: A window toward a comprehensive interpretation of networks in cancer. Cancer modeling and network biology - Accelerating toward personalized medicine. Semin Cancer Biol. 2015;3:79–87.
34. Guiot C, Degiorgis PG, Delsanto PP, Gabriele P, Deisboeck TS. Does tumor growth follow a "universal law"? J Theor Biol. 2003;225(2):147–51.
35. Rejniak KA, Anderson ARA. Hybrid models of tumor growth. Interdiscip Rev Syst Biol Med. 2011;3(1):115–125.
36. Deisboeck TS, Wang Z, Macklin P, Cristini V. Multiscale Cancer Modeling. Annu Rev Biomed Eng. 2011;13:127–55. NIH Public Access.
37. Deisboeck TS, Mansury Y, Guiot C, Degiorgis PG, Delsanto PP. Insights from a novel tumor model: Indications for a quantitative link between tumor growth and invasion. Med Hypotheses. 2005;65(4):785–90.
38. Folkman J, Hochberg M. Self-regulation of growth in three dimensions. J Exp Med. 1973;138(4):745–53.
39. Enderling H, Hahnfeldt P, Hlatky L, Almog N. Systems biology of tumor dormancy: linking biology and mathematics on multiple scales to improve cancer therapy. Cancer Res. 2012;72(9):2172–5.
40. Lu P, Weaver VM, Werb Z. The extracellular matrix: a dynamic niche in cancer progression. J Cell Biol. 2012;196(4):395–406.
41. Chen Y, Lowengrub JS. Tumor growth in complex, evolving microenvironmental geometries: a diffuse domain approach. J Theor Biol. 2014;361:14–30.
42. Sciumè G, Santagiuliana R, Ferrari M, Decuzzi P, Schrefler BA. A tumor growth model with deformable ECM. Phys Biol. 2014;11(6):065004.
43. Szabó A, Merks RM. Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution. Front Oncol. 2013;3:87.
44. Wang Z, Butner JD, Kerketta R, Cristini V, Deisboeck TS. Simulating cancer growth with multiscale agent-based modeling. Semin Cancer Biol. 2015;30:70–8.
45. Graner F, Glazier JA. Simulation of biological cell sorting using a two-dimensional extended Potts model. Phys Rev Lett. 1992;69(13): 785–790.
46. Glazier JA, Graner F. Simulation of the differential adhesion driven rearrangement of biological cells. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics. 1993;47:2128–54.
47. Glazier JA, Balter A, Poplawski NJ. Magnetization to Morphogenesis: A Brief History of the Glazier-Graner-Hogeweg Model. In: Single cell-based models in Biology and Medicine. Birkhäuser, Basel: Mathematics and Biosciences in Interaction; 2007. p. 79–106.
48. Balter A, Merks RM, Poplawski NJ, Swat M, Glazier JA. The Glazier-Graner-Hogeweg model: extensions, future directions, and opportunities for further study. In: Single-Cell-Based Models in Biology and Medicine. Birkhüuser Basel; 2007. p. 151–167.
49. Voss-Bőhme A. Multi-Scale Modeling in Morphogenesis: A Critical Analysis of the Cellular Potts Model. PLoS ONE. 2012;7(9):e42852.
50. Scianna M, Preziosi L, Wolf K. A Cellular Potts Model simulating cell migration on and in matrix environments. Math Biosci Eng. 2013;10(1):235–261.
51. Rubenstein BM, Kaufman LJ. The Role of Extracellular Matrix in Glioma Invasion: A Cellular Potts Model Approach. Biophys J. 2008;95(12): 5661–5680.
52. Boas SE, Jimenez MIN, Merks RM, Blom JG. A global sensitivity analysis approach for morphogenesis models. BMC Syst Biol. 2015;9(1):1.
53. Stott EL, Britton NF, Glazier JA, Zajac M. Stochastic simulation of benign avascular tumour growth using the Potts model. Math Comput Model. 1999;30(5–6):183–198.
54. Turner S, Sherratt JA. Intercellular adhesion and cancer invasion: a discrete simulation using the extended Potts model. J Theor Biol. 2002;216(1):85–100.
55. Ghaemi M, Shahrokhi A. Combination of the cellular Potts model and lattice gas cellular automata for simulating the avascular cancer growth. In: Cellular Automata. Berlin, Heidelberg: Springer-Verlag; 2006. p. 297–303.
56. Liu C, Lu B, Li C. A Parameter Selection Model for Avascular Tumor Growth. Internat J Control Automation. 2014;7(12):155–64.
57. Giverso C, Scianna M, Preziosi L, Lo Buono N, Funaro A. Individual cell-based model for in-vitro mesothelial invasion of ovarian cancer. Math Model Nat Phenom. 2010;5(1):203–23.
58. Osborne JM. Multiscale Model of Colorectal Cancer Using the Cellular Potts Framework. Cancer Informat. 2015;14(Suppl 4):83.
59. Scianna M, Preziosi L. A cellular Potts model for the MMP-dependent and -independent cancer cell migration in matrix microtracks of different dimensions. Comput Mech. 2014;53(3):485–97.
60. Sottoriva A, Vermeulen L, Tavare S. Modeling evolutionary dynamics of epigenetic mutations in hierarchically organized tumors. PLoS Comput Biol. 2011;7(5):e1001132.
61. O'Donoghue SI, Gavin A-C, Gehlenborg N, Goodsell DS, Hériché J-K, Nielsen CB, North C, Olson AJ, Procter JB, Shattuck DW, Walter T, Wong B. Visualizing biological data—now and in the future. Nat Methods. 2010;7(3 Suppl):2–4.
62. Turkay C, Jeanquartier F, Holzinger A, Hauser H. On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Springer Berlin Heidelberg; 2014. p. 117–140.
63. Von Landesberger T, Kuijper A, Schreck T, Kohlhammer J, van Wijk JJ, Fekete J-D, Fellner D. Visual Analysis of Large Graphs: State-of-the-Art and Future Research Challenges. Comput Graph Forum Wiley. 2011;30(6): 1719–49.
64. Jeanquartier F, Jean-Quartier C, Holzinger A. Integrated Web visualizations for protein-protein interaction databases. BMC Bioinforma. 2015;16(1):195. doi:10.1186/s12859-015-0615-z.
65. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):11.
66. Engwer C, Knappitsch M, Surulescu C. A multiscale model for glioma spread including cell-tissue interactions and proliferation. Mathematical Biosciences and Engineering. 2016;13:443–460.
67. Hoehme S, Drasdo D. A cell-based simulation software for multi-cellular systems. Bioinformatics. 2010;26(20):2641–2.
68. Swat MH, Thomas GL, Belmonte JM, Shirinifard A, Hmeljak D, et al. Multi-scale modeling of tissues using CompuCell3D. Methods Cell Biol. 2012;110:325–366.
69. Merks RMH, Glazier JA. A cell-centered approach to developmental biology. Physica A: Statistical Mechanics and its Applications. 2005;352(1): 113–30.
70. Fenstermacher D, Street C, McSherry T, Nayak V, Overby C, Feldman M. The cancer biomedical informatics grid (caBIG TM). In: Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual Conference. IEEE; 2005. p. 743–746.
71. Daub JT, Merks RM. Cell-Based Computational Modeling of Vascular Morphogenesis Using Tissue Simulation Toolkit. Methods Mol Biol. 2015;1214:67–127.
72. Steinberg M. On the mechanism of tissue reconstruction by dissociated cells, III. Free energy relations and the reorganization of fused heteronomic tissue fragments. PNAS. 1962;48:1769–76.
73. The Cytoscape Consoritum. http://js.cytoscape.org/. Accessed 4 Aug 2016.
74. Jeanquartier F, Jean-Quartier C, Cemernek D, Holzinger A. Tumor Growth Simulation Profiling In: LNCS, editor. Information Technology in Bio- and Medical. Springer; 2016;9832.
75. Ouchi NB, Glazier JA, Rieu JP, Upadhyaya A, Sawada Y. Improving the realism of the cellular Potts model in simulations of biological cells. Physica A: Statistical Mechanics and its Applications. 2003;329(3–4): 451–8.
76. Ware C. Information visualization: perception for design. Amsterdam: Morgan Kaufmann; 2012, p. 4.

77. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape. js: a graph theory library for visualisation and analysis. Bioinformatics. 2016;32(2):309–11.
78. Laursen O. Flot: Attractive javascript plotting for jquery. 2014. http://www.flotcharts.org/. Accessed 4 Aug 2016.
79. Ono K, Demchak B, Ideker T. Cytoscape tools for the Web age. D3.js and cytoscape.js exporters. F1000Research. 2014;3:143.
80. Sweeney TJ, Mailänder V, Tucker AA, Olomu AB, Zhang W, Cao Ya, Negrin RS, Contag CH. Visualizing the kinetics of tumor-cell clearance in living animals. Proc Natl Acad Sci. 1999;96(21):12044–9.
81. Lloyd AC. The Regulation of Cell Size. Cell. 2013;154(6):1194–205.
82. Johnson D, Connor AJ, McKeever S, Wang Z, Deisboeck TS, Quaiser T, Shochat E. Semantically Linking In Silico Cancer Models. Cancer Informat. 2014;13(Suppl 1):133–43.
83. 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. 2014;15(Suppl 6):I1.
84. Holzinger A. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics. 2016;3(2):119–131.
85. Jeanquartier F, Jean-Quartier C, Schreck T, Cemernek D, Holzinger A. Integrating Open Data on Cancer in Support to Tumor Growth Analysis. Information Technology in Bio- and Medical Informatics, Lecture Notes in Computer Science, LNCS 9832: Springer; 2016.

## 2.7. Machine Learning For In Silico Modeling Of Tumor Growth

This book chapter provides a practical overview of pointers to machine learning methods applied to tumor growth modeling research. This work is an essential step towards understanding the different possibilities of applying ML techniques to support in silico modeling of tumor growth. A future goal is to possibly use ML for validation and later on also to integrate novel insights into our tumor growth modeling visualization tool. Therefore, I set up the paper's structure, started with the review work, identified challenges and opportunities and last but not least, invited the other authors to contribute to describing application examples and finalized the paper.

# Machine Learning for *In Silico* Modeling of Tumor Growth

Fleur Jeanquartier[1( )], Claire Jean-Quartier[1], Max Kotlyar[2], Tomas Tokar[2], Anne-Christin Hauschild[2], Igor Jurisica[2], and Andreas Holzinger[1]

[1] Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics and Documentation, Medical University Graz, Graz, Austria
{f.jeanquartier,c.jeanquartier,a.holzinger}@hci-kdd.org
[2] Princess Margaret Cancer Centre, University Health Network, Toronto, Canada
juris@ai.utoronto.ca

**Abstract.** The various interplaying variables of tumor growth remain key questions in cancer research, in particular what makes such a growth malignant and what are possible therapies to stop the growth and prevent re-growth. Given the complexity and heterogeneity of the disease, as well as the steadily growing set of publicly available big data sets, there is an urgent need for approaches to make sense out of these open data sets. Machine learning methods for tumor growth profiles and model validation can be of great help here, particularly, discrete multi-agent approaches.

In this paper we provide an overview of current machine learning approaches used for cancer research with the main focus of highlighting the necessity of *in silico* tumor growth modeling.

**Keywords:** Tumor growth · Cancer modeling · Machine learning · Computational biology

## 1 Introduction

Cancer prognosis and prediction is advancing by making use of data that has been mined and interpreted with the help of machine learning techniques. Machine Learning (ML) also aids the process of interpreting and understanding the complexity in big data sets [1].

Johnson *et al.* describe cancer informatics as hybrid discipline; although, even with the latest ML advances, there is still a gap to fill in fostering mathematical modeling and computer simulation of cancer [2].

Modeling tumor growth is a very challenging problem because, besides from being highly complex, it involves dynamic interactions spanning multiple scales both in time and space. This involves both continuous and discrete variables that call for hybrid approaches [3]. Araujo and Mcelwain [4] historically summarize how mathematical modeling has contributed to elucidating tumor growth.


A. Holzinger (Ed.): ML for Health Informatics, LNAI 9605, pp. 415–434, 2016.
DOI: 10.1007/978-3-319-50478-0_21

### 1.1 Glossary and Key Terms

*In Silico* refers to being performed on a computer instead of a wetlab and stands opposite to *in vivo* or *in vitro* [5]. Naturally, integration and interplay of all three approaches is essential for research advances.

*Machine Learning (ML)* addresses the question of how to design algorithms that improve automatically through experience [6]. Besides primary goal of learning useful models, scalability of these algorithms play an increasingly important role in the the era of "big data analytics".

*Interactive Machine Learning (iML)* defines learning algorithms that can interact with both computational agents and human agents, and can optimize their learning behavior through these interactions [7], by bringing in a human-in-the-loop [8].

*Agent-Based Modeling (ABM)* depicts a computational method for simulating a system, which is based on individual units, calculated by a given rule-set on a discrete level [9].

*Cellular Potts Modeling (CPM)* defines a stochastic process of simulating the collective behavior of cellular structures [10].

*Cellular Automata (CA)* are representations for modeling complex systems dynamics [11–13].

*Support Vector Machines (SVM)* are supervised learning algorithms to solve primarily classification and regression problems [14,15].

*Electronic Health Records (EHR)* are longitudinal electronic records of patient health information with the ability to generate complete records of clinical patient encounters [16].

*Protein-Protein Interactions (PPI)* comprise the concurrence and the effect of proteins on each other based on surface properties as well as local features [17]. PPIs form the basic concept of biological communication and the specificity in signal transduction [18–20].

## 2 Motivation for Applying ML to Cancer Research

There are different entry points for ML to tumor growth research. Within this paper, we summarize possible approaches to using ML in the field of cancer research and the various kinds of models of tumor growth in computational or systems biology.

Cancer research started around 250 years ago [21]. There are several methods to study the disease, still, basic research comes down with animal experimentation. *In vitro* cell systems and the comparison of cellular processes help to understand the complexity of uncontrolled cell growth.

*In silico* models complement traditional *in vitro* and *in vivo* animal models. While ML is not new to cancer research the full potential of diverse ML algorithms has not been realized yet. In fact, *in silico* techniques are often underrated but can be vital to fundamental questions to beat cancer [22]. Knowledge discovery with ML outperforms bio assays [23] and image analysis could outperform human [24]. The principles of the 3Rs - replacement, reduction and refinement - can be used for the reduction of animal research, saving resources as well as reducing costs spent on clinical and wet-lab experiments in cancer research. In this regard, computerized experiments, meeting the terms of 3R, offer new possibilities for biomedical research. *In silico* suits the task of refinement as well as knowledge discovery. Recently, we presented an *in silico* approach for tumor growth simulation that holds the advantage of data visualization over multiple implementation possibilities [25,26]. It is clear that ML techniques will give new insights into tumor growth modeling. Thereby, the goal is to increase the basic understanding of tumor progression as well as the onset of cancer.

## 3 In Silico Modeling of Cancer

*In silico* models involve various disciplines of mathematics, biology, medical and computer science. The underlying data is computationally processed from biomedical literature sources, based on wet-lab and clinical investigations, and extended or refined through hypothesises and theoretical characterizations [22].

There are different kinds of models in biology, such as spatial ones, space free ones but also cell descriptive models based on density, cell-based, sub-cellular or molecular, relating to their scale of phenomenon, and so far, various models for cancer have been described [10,27]. Models can also be differentiated by their biological scale, ranging from the cellular and molecular level up to the genetic macro scale. On the other hand, there are also diverse computational modeling approaches, such as statistical, network-based as well as models on tissue-level. Regarding the cell-cell interactions there are discrete/agent-based to continuum-based modeling approaches. This leads us to the term agent that is shortly discussed in the next paragraph.

### 3.1 Agents in Modeling and ML

Agents play an important role both in Agent-based modeling (ABM) as well as in Machine Learning (ML). As described by Russell *et al.*, "an agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors" [28]. According to [29] ABM is used to model phenomena as dynamical systems of interacting agents. Thereby, agents individually assess a situation and make decisions on the basis of a set of rules [30]. So far, agents can be robot or human [7].

New *agent-based models of tumor growth* have been developed to foster the understanding of cancer, while agents can be used to model different parts of tumor growth to understand peculiarities such as factors that influence a tumor

becoming malignant etc. [9,25,31]. Followingly, we shortly describe aspects of tumor growth for ABM.

*Tumor growth kinetics* follow simple laws that can be mathematically modeled [32]. Among them, the Gompertz law describes growth following a continuous deceleration [33–35].

*Cellular Potts Model (CPM)* is an agent-based modeling approach that has been introduced and described by Graner and Glazier [36]. It is used to simulate the collective behavior of cellular structures and has been used in a wide range of applications, among them, tumor dynamics [10].

*Spatial & temporal* scales are key descriptors in ABM in general and in modeling tumor growth in particular [10,25,32]. Regarding the description of spatial aspects, different topologies are used in ABM, such as spatial grids. Grids have been implemented as CA, i.e. Conway's Game of Life [37]. We [25] use also the term lattice as a group (not partially ordered set) to describe the topology and therefore the connectedness of several cellular bricks. The agent's neighborhood is described by an agent only interacting with its neighbors located close-by. However, agents may also interact with their environment, therefore environmental parameters can be taken into account. Regarding the temporal aspects, ABM follows discrete event cues, in particular a sequential schedule of interactions, computed by Monte Carlo steps (MCS).

*Cellular Automata (CA)* is a concept introduced by Stanislaw Ulam and John von Neumann in 1940s [11–13]. A typical CA includes a spatial lattice comprising units, called cells, where each cell can reside in one of finite number of pre-defined states. State of each cell in the lattice is updated according to the transition rules, so that the state of the cell in the given time depends on its own previous state and on the previous state if its close neighbors. The overall state of the entire lattice is evolving in discrete times steps, either synchronously, when all cells are updated at once, or asynchronously, when single randomly selected cell is updated in each time step. The Concept of CA was later popularized by Stephen Wolfram, who showed that even simple transition rules allow CA to exhibit variety of complex behaviors including phenomena of "self-organization" [38]. CA have been then extensively utilized in model dynamics of complex systems across diverse fields, including cancer biology. CA have been successfully adopted to realistically model tumor growth [39–46], as well as angiogenesis [47–49] and immune evasion [50,51].

Transition rules governing the behavior of the automaton, are sometimes formulated directly according to the available experimental knowledge [39,44,48], but more often are subject of inference using numerical optimization with respect to desired macroscopic qualities, e.g., transient dynamics of the tumor growth, or its geometric properties [42,46]. Alternatively, transition rules and associated quantitative parameters are varied in order to reveal association between microscopic properties of the single cell and macroscopic properties of the tumor [40,41].

Ideally, a model gives emergence to phenomena that could not be *a priori* deduced, and can be tested against experimental data.

ABM is not inductive, that means models are not based on a set of data and do not make inferences that lead to that data, but rather describe a system's mechanisms of rules and seek to reconstruct observations. This leads to ML, that is suitable to find patterns in existing data as well as can be used for validation, to extend *in silico* modeling tools.

## 4 ML Applications Areas in Cancer Research

ML approaches for cancer research have been reviewed before [1,52–56]. These reviews deal both with biological questions as well as on algorithmic details. While most ML reviews in this domain cover genomic studies and image based analysis, some also tackle the question how to support the understanding of tumor kinetics in particular. But there is a clear lack of new results in this area. An advanced search within EuropePMC with the query:

($TITLE$ : "*cancer*" AND "*machine learning*") AND ($OPEN\_ACCESS : y$) yielded 671 results.

The search query: ($KW$ : "*machine learning*" AND $KW$ : "*cancer*") AND ($OPEN\_ACCESS : y$) delivered only 41 results.

Regarding the term "tumor growth" there are hardly any works. The query: ($TITLE$ : "*tumor growth*") AND ($KW$ : "*machine learning*") even resulted in no results at all.

This work is not aimed at providing a comprehensive list of all studies that can be found on machine learning methods related to tumor growth research, even, if there are hardly any found. It is rather thought to provide a practical overview of pointers to machine learning methods applied to tumor growth modeling research with identifying challenges and opportunities.

In order to understand the different possibilities of applying ML techniques to cancer research, we first differentiate between specific application areas and later continue on describing research on tumor growth in particular. An overview of ML applications in cancer research is presented in Fig. 1.

Most reviews on ML for cancer focus on discussing existing cancer research that applies ML methods for predicting susceptibility, recurrence and survival [1,53]. Next to prediction, ML methods are applied to identification and diagnosis [57]. A classification of ML application areas in bioinformatics shows partially overlapping areas of genomics, proteomics and metabolomics but also evolutionary developmental biology, text mining, systems biology other advanced modeling applications [58]. Computational prediction approaches based on computer algorithms, allow for multivariate analysis in cancer diagnosis and comprise several methods such as linear or penalized discriminant analyses, logistic regression, learning vector quantization, decision trees, random forest, support vector machines, Bayesian networks and artificial neural networks [59,60]. These computational approaches overcome the lack of sensitivity and selectivity that, still, are often found in conventional methods based on univariate factors such as single biomarkers [60]. To evaluate prediction accuracy of these models the data is

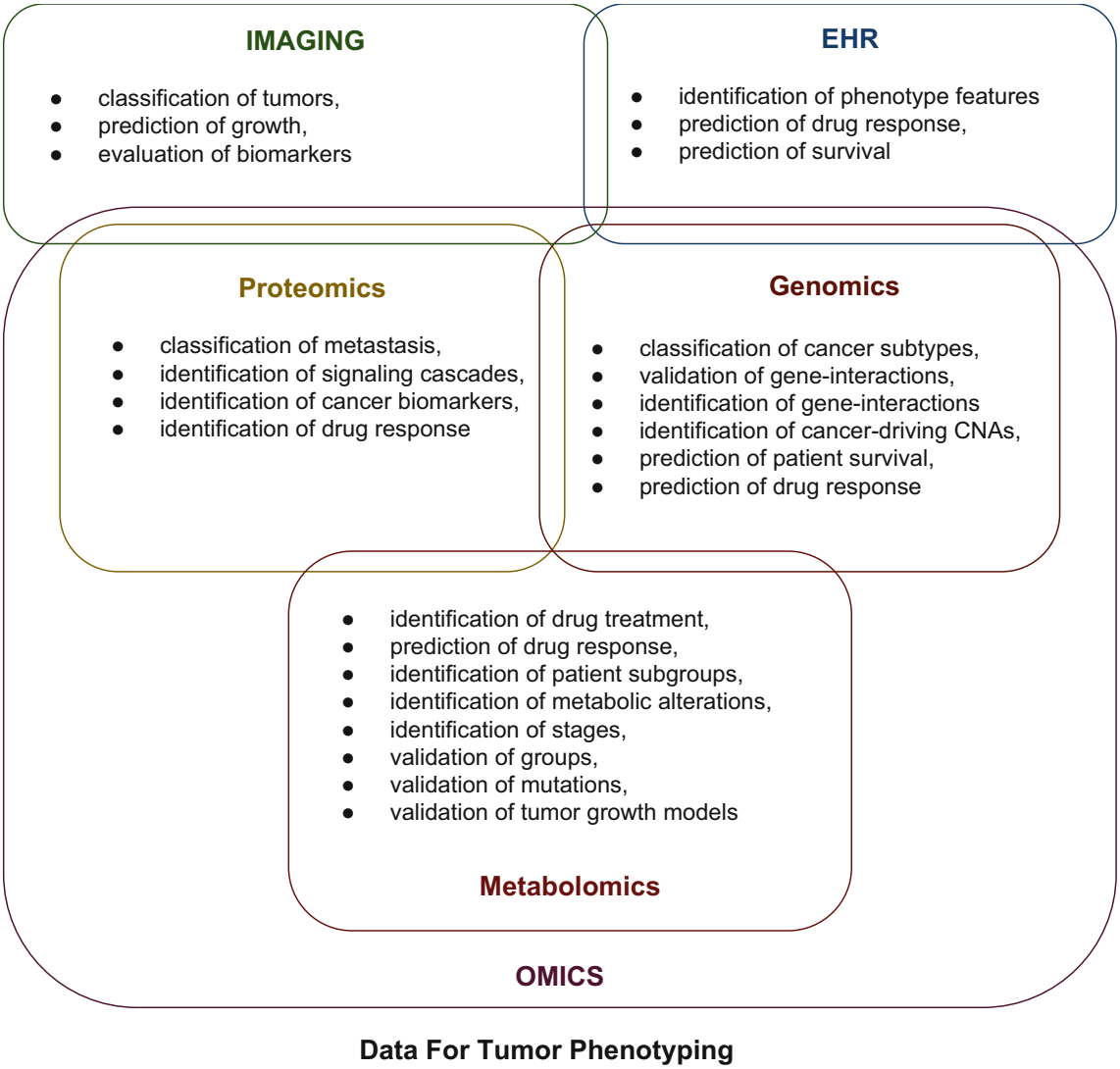


**Fig. 1.** Overview of ML approaches in cancer research regarding data type

randomly separated into training-, validation-, and test-sets. However, this gold standard method is solely feasible for large data sets. Cross-validation, a simple and commonly applied approach, splits data into subsets, while each subset is left out once for testing, the model is trained on the remaining data. Independent of univariate or multivariate methodology, permutation-based evaluation is recommended to assess the superiority of the model compared to a model trained on a randomized outcome variable [61].

ML approaches for cancer research can also be organized according to their algorithmic approach as well as the type of data used, ranging from imaging, genomics up to pathologic and demographic [53]. We next list works sorted by data approach to provide pointers for using ML on open cancer data [26].

### 4.1 Processing Imaging Data

ML can be used to detect and classify tumors in medical images [53]. For example, Morris *et al.* model glioma tumor growth using magnetic resonance (MR) scans for learning the parameters of a diffusion model [62]. Thereby they use patient data and preprocessing of images such as noise reduction and segmentation for feature extraction and consecutively prediction of glioma growth through classification and diffusion.

HealthAgents is another interesting project implementing a multi-agent system (MAS) for classifying brain tumors by applying pattern recognition methods on MR images [63].

Moreover, ML methods have been used for the evaluation of different radiomic features for predicting survivability [64]. Results highlight the several features' utility as radiomic biomarkers [64].

Cancer imaging, in particular image analysis of MR scans, already provides many possibilities for biomarkers [55]. But, images not only allow for measurements of the dynamics of shape and size. Fluorescence microscopy is also used to monitor small parts of cells [65]. Understanding complex diseases also requires identifying interactions among different components which leads us to the world of "Omics". Processing additional data such as combing picture archives with genomic profiles and even more, with electronic health records (EHR), brings us one step closer towards personalized medicine. Next, we summarize main concepts in Omics data and further proceed with examples in processing electronic healthcare records and hybrid data approaches:

### 4.2 Processing Omics Data

The molecular etiology of cancer is not well understood. Although numerous molecular cancer biomarkers have been identified, they are often ineffective for tasks such as cancer diagnosis, classification of cancer subtypes, prediction of cancer recurrence, or prediction of response to treatment [66]. One of the most promising strategies for addressing these problems is analysis of molecular networks, combined with machine learning and graph theory algorithms. These approaches lead to better predictions across diverse samples, and identify molecular mechanisms underlying cancer [67].

**Protein-Protein Interaction (PPI)** networks were the first type of molecular network used for identifying cancer biomarkers. Chuang *et al.* [68] identified PPI subnetworks that could serve as biomarkers for classifying breast cancer metastasis. Their approach combined PPI data with gene expression data from patients with and without breast cancer metastasis. The approach searched for protein subnetworks whose corresponding gene expression levels could distinguish metastatic and non-metastatic patients. The average expression of all genes in a subnetwork was used as a biomarker, unlike previous approaches, where biomarkers were individual molecules. The identified subnetworks had significant associations with hallmarks of cancer, and indicated novel relationships between

signaling cascades (functional networks or pathways) and tumor progression. Furthermore, subnetwork biomarkers outperformed single-gene biomarkers in two important aspects: reproducibility across data sets and classification performance. Reproducibility considers whether the same biomarkers can be identified using different data sets: subnetwork biomarkers from different expression data sets overlapped by 12.7 %, whereas single-gene biomarkers overlapped by only 1.3 %. Classification performance - the ability of biomarkers to predict metastatic status - was assessed by using biomarkers as inputs to classifiers (logistic regression and support vector machines), that were tested through cross-validation. Subnetwork biomarkers significantly outperformed sets of single-gene biomarkers with all classifiers and data sets tested. Subsequent studies used PPI networks to identify subnetwork biomarkers of bladder, colorectal, gastric, liver, and lung cancers [69,70], and single-protein biomarkers of brain, breast, liver, lung, and skin cancers [71–75]. PPI networks have also been used to identify biomarkers of response to cancer treatment [76,77]. Cancer-related biomarkers cannot only be described in Proteomics but also in Genomics.

**Genomic Data** has brought up several biomarkers for measuring therapeutic response and validating drug treatment of cancer [78]. Moreover, genomic data such as gene expression samples can be used for identifying cancer subtypes [79] but also for predicting evolution even including response to drugs [53]. For example, gene expression data [79] and molecular profiling [80] have been used to improve glioma classification. Genomic data has also been used for the prognosis of possible relapse after treatment of prostate cancer [81].

Upstill *et al.* describe ML approaches for discovering gene-gene interactions in sequencing data [57]. While They call the type of data “disease data”. They also underline that most studies report on applying ML for validating results rather than on identifying new disease-related interactions. The Matchmaker Exchange API [82] is a tool for cohort discovery and variant disease causal validation that also makes use of so called “disease data” from different genomic databases.

In general, networks based on gene expression data have been used to identify biomarkers predictive of patient drug response and prognosis. Two types of networks are typically constructed from gene expression data: co-expression networks, where edges connect pairs of genes that have correlated expression across samples, and gene regulatory networks, where edges indicate regulatory effects between pairs of genes. Both types of networks have helped identify cancer biomarker genes and gene modules. These biomarkers were used as inputs to statistical or machine learning methods for various disease prediction and classification tasks. Biomarkers from co-expression networks have been used to predict patient prognosis [83–85] and response to treatment [85]. Applications of gene regulatory networks have included biomarker discovery for prostate cancer [86] and breast cancer [87], and modeling of ovarian cancer progression [88].

As genomic alterations are a fundamental feature of cancer, several network-based methods have been developed for analyzing these alterations, and identifying subsets that are cancer biomarkers. Jörnsten *et al.* developed causal

network models to understand how DNA copy number alterations in glioblastoma affect gene expression [89]. These models, based on regression and bootstrapping methods, predict key cancer-related alterations, their effects on gene expression, and patient survival. Shi *et al.* developed an alternative network model, using Laplacian shrinkage, to analyze the effects of copy number alterations on gene expression [90]. Leung *et al.* introduced a method for identifying frequently mutated gene modules in molecular networks associated with patient drug response, patient survival and other clinical or phenotypic data [91]. Similar approaches identify cancer-deregulated subnetworks [92].

There is a vast amount of publicly available heterogeneous genomic data, making data mining and ML well suited to solve key problems in the world of genomic medicine [93]. Complementing genetic studies leads us to the field of Metabolomics.

**Metabolomics** has been introduced to cancer "omics" studies relatively recently. It opened new opportunities towards biomarker discovery, identification of signaling molecules associated with cell growth, cell death, cellular metabolism [101]. Metabolomics is therefore frequently used for studies aiming at the detection of cancer even in early stages. Most commonly used analytical technologies comprise NMR spectroscopy, LC/MS, GC/MS and MCC/IMS [101,102]. In order to meet the demands of cellular proliferation and the required uptake and conversion of nutrients into biomass, cancer cells modify their metabolism during tumor development. Many of these key metabolic alterations are similar across tumor cells. A prominent example are the changes in the glucose metabolism leading to an increase of the described biosynthetic activities, and to the 'Warburg' effect, an inevitable adaptation to cope with the lack of ATP generation [103].

Metabolomics technology can be used to identify clinically relevant subgroups of cancer patients. For instance, O'Shea *et al.* analyzed the metabolites in sputum from patients with lung cancer and age-matched volunteers smoking controls using flow infusion electrospray ion mass spectrometry and found potential marker using artificial neural networks [104]. A sequential application of recursive feature elimination on linear-SVM and orthogonal partial least squares discriminant analysis (PLS-DA) was used to find the minimum set of discriminant features separating early-stage ovarian cancer patients samples from controls. Permutation testing was performed to validate the results [105]. Another study analyzed the metabolom of exhaled air by MCC/IMS within normal, COPD and lung cancer patients. A variety of supervised ML methods, e.g., linear-SVM or random forest, were applied to evaluate their capabilities to differentiate the three groups [106].

A second group of studies focus on validation. G12C k-RAS mutation has been suspected to be a key player in promoting metabolic rewiring, in isogenic non-small cell lung caner (NSCLC) cell line. Brunelli *et al.* applied OPLS-DA models and discovered a robust separation between G12C and WT k-RAS isoforms both *in vitro* and *in vivo*. Authors further validated their findings by

mapping the quantified metabolites to the KEGG pathway database. Furthermore, they identify a list of most likely enriched metabolic pathways associated with the given metabolites [107].

The third application focuses on the prediction of disease outcome. Metabolomic NMR fingerprinting was utilized to assess the survival of patients with metastatic colorectal cancer (mCRC). A combination of partial least squares and support vector machines (PLS-SVM) was first applied to discriminate patients with mCRC and healthy subjects. In a second step, PLS-SVM was successfully used to evaluate whether patients with short or long overall survival can be identified by metabolomic profiling using NMR [108]. Wei *et al.* utilized a metabolomics approach to predict the effectiveness of treatments. In particular, PLS-DA is applied to model the response to neoadjuvant chemotherapy for breast cancer [109].

These findings show that metabolomics data can be used to differentiate not only tumor from control samples but also identify different stages of the growing tumor. Thereby, these technologies could be used for continuous monitoring of tumor growth and development in order to validate and optimize presented approaches *in silico* tumor growth models. Processing healthcare records forms another example in need of computerized support within the field of personalized cancer therapy and research, that is discussed next.

### 4.3 Processing Healthcare Records and Combined Data

When dealing with medical records, its anonymization is an important topic that can be supported through the use of ML [7]. Learning from various data sets opens up novel possibilities for cancer research.

So far, several works have described different ML techniques for the classification of patient cohorts [94]. Standardized multi-scale information models of cancer phenotypes provide information in computable form that are important for complementary approaches such as tumor growth modeling [95].

Delen *et al.* [96] describe a comparative study of neural networks, decision trees as well as logistic regression for mining a data set of more than 200,000 cases provided by SEER [97] for testing prediction of breast cancer survivability.

Menden *et al.* describe an approach for predicting how cancer cells respond to drugs based on combined data analysis, genomic features of cell lines as well as chemical features of drugs [98].

EHR have also been used for predicting cancer survival with the help of support vector machines (SVMs) [99]. Weighted Bayesian networks have been developed on the combination of EHR and PubMed data to predict pancreatic cancer [100].

Hybrid methods provide effective means to detect and quantify a broad range of small molecules for studying complex biological networks.

## 5 ML Towards Extending *In Silico* Modeling

Lisboa *et al.* [55] highlight how modeling of biological processes related to cancer may benefit from data mining approaches. They summarize main concepts found in literature as on the one hand, mining data from experiments to better understand parts of signaling pathways, and on the other hand to predict the evolution of dynamical systems.

Integration of data can be used to extend the descriptive part of compartmental states [26], such as by relating information on inhibitors and promotors to tumor growth curves, but also by making use of cancer classifications to create cancer profiles [110].

ML methods can be used on open cancer data for several possibilities, i.e., identifying tumor suppressing and inhibiting genes and further advancing a tumor growth related interaction network [111] that may help find and select precisely targeted treatments [92,112,113]. Existing treatment data can be further integrated into simulation tools to validate both tool and model and improve the tumor growth prediction rates. Such predictions gained via ML approaches can be combined with the ABM approach for further analysis. Other subjects of interest can be described further, such as specific cells or parts of it, that are again remodeled as discrete entities or agents, and iteratively validated to support sense-making in tumor growth analysis.

Additionally, visualization supports interaction with data and models [114]. Visualization in ABM is needed to visually convey the behavior of the model [31,115]. We have recently introduced a novel visualization approach of simulating and analyzing cell-related variables regarding tumor growth kinetics [25]. Thereby, visualization is used to show patterns of tumor growth. The graph-based visualization approach makes use of nodes, representing cellular bricks. These cellular bricks are related to compartmental states, including localized phenomena.

Last but not least, ML can be used to include image analysis in two ways: First, images can be used as input for the modeling, while the classification of images can be supported by ML techniques. Second, by analyzing a set of existing images related to tumor growth, the model can be compared to ML results and further validated.

## 6 Challenges in Network-Based ML Approaches

Network analysis combined with machine learning has proven to be an effective approach for identifying biomarkers and molecular mechanisms of cancer [116]. This approach is likely to further increase in popularity, but continued progress will require addressing multiple challenges:

*Challenge 1.* foremost, we need to increase coverage and annotation of diverse networks, to include tissue and process specificity;

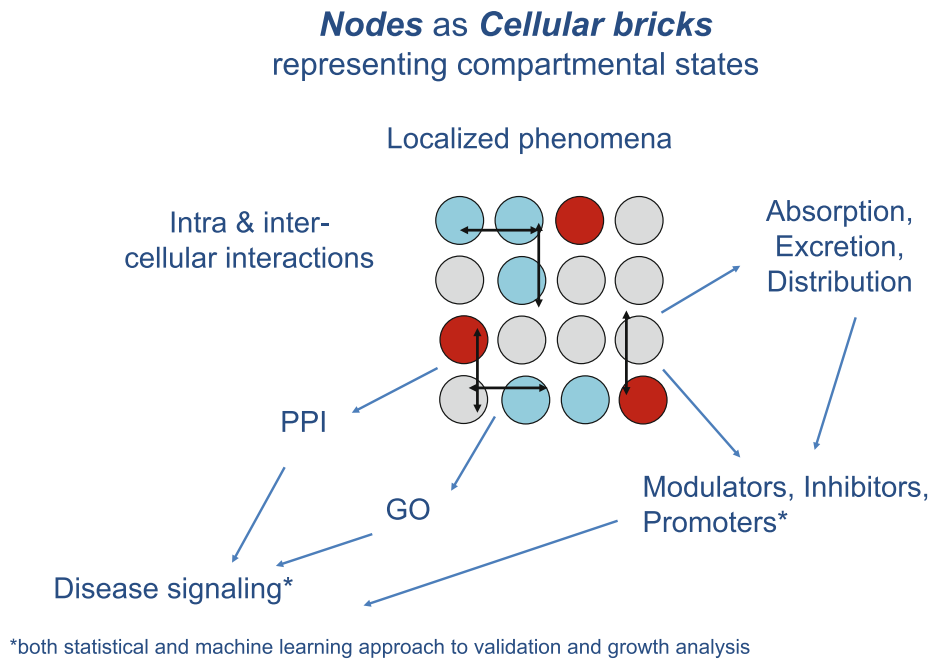


**Fig. 2.** The "big picture" in the modeling and visualization of tumor growth [25,26]

*Challenge 2.* we need to improve scalability of algorithms to addressed increased size and complexity of networks;

*Challenge 3.* biomarker performance will need to be measured by standardized unbiased methods;

*Challenge 4.* multiple types of omics data will need to be combined into unified network models; and

*Challenge 5.* networks may need to be tailored to individuals to facilitate personalized medicine.

## 7 Challenges in Modeling Tumor Growth Dynamics

*In Silico* models complement the lack of *in vitro* and *in vivo* models. However, tumor growth modeling also brings up many questions concerning specific aspects of the various kinds of benign and malignant neoplasms. The main challenges in modeling tumor growth kinetics include:

*Challenge 1.* **There is no universal tumor growth model.** As Benzekry *et al.* [32] described, dormancy phases creates challenges for finding a generic growth law. The Gompertz or power law has been used to predict tumor growth; however, with a very low prediction rate. A so called "Universal Law of tumor growth" has to be found yet. However, [117] proposed to classify tumor growth patterns into fundamentally different categories. Therefore, cancer classification and profiling has to be taken into account.

*Challenge 2.* **The disease's complexity** poses a big open problem to tumor growth modeling. According to Edelman *et al.* [27] modeling the heterogeneous nature of tumor growth needs to take various characteristics into account. These characteristics are to be comprehensively discovered, as well as, in the latter modeled.

*Challenge 3.* **Data heterogeneity** challenges integration and fusion. Data fusion poses significant challenges. While diverse data sets exist, data comes from different laboratories, with different type and quality controls, different representation and processing [27,118–120]. Integrating current bioinformatics workflows with knowledge engineering provides the necessary step in the right direction.

*Challenge 4.* **Visualizing evidence and uncertainty** with aggregation and display of specific information is required to make informed decisions. However, visualization still poses a big challenge. Offering reproducable, transparent and interactive visual analysis output of learned patterns is one of the many challenges for applying Visual Analytics methods to the biomedical domain [121–124].

*Challenge 5.* Finally, the question remains of **how to infer knowledge from existing data**. Machine learning may be used to infer graphical models from data [118], but there are difficult learning tasks to infer graphical models, yet to be solved.

## 8 Conclusion and Future Outlook

Combining ML and ABM can be used on various biological scales, as shown in Fig. 2: The lattice's nodes are represented as cellular bricks, which can be related to localized phenomena such as intra- & intercellular interactions, information on absorption, excretion, distribution as well as modulators, inhibitors and promoters, but also protein interactions and gene ontology. The overall goal remains to understand properties and peculiarities regarding cancer disease signaling.

In summary, ML can be used to improve *in silico* modeling, ranging from model validation to identifying novel insights. Future studies may involve the integration of proteomic and metabolomic networks behind ABM in order to simulate drug effects on tumor growth towards personalized medicine. Further exploration on genomic information regarding disease-driving mutations could be embedded within a multi-agent approach to simulating tumor growth in more detail. This may include studies on evolutionary dynamics of tumor growth and the underlying cellular heterogeneity of tumors using *in silico* environments.

## References

1. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2015)

2. Johnson, D., Osborne, J., Wang, Z., Marias, K.: Computer simulation, visualization, and image processing of cancer data and processes. Cancer Inf. **14**(Suppl 4), 105 (2015)
3. Tzedakis, G., Tzamali, E., Marias, K., Sakkalis, V.: The importance of neighborhood scheme selection in agent-based tumor growth modeling. Cancer Inf. **14**(Suppl 4), 67–81 (2015)
4. Araujo, R.P., McElwain, D.S.: A history of the study of solid tumour growth: the contribution of mathematical modelling. Bull. Math. Biol. **66**(5), 1039–1091 (2004)
5. Sieburg, H.B.: Physiological studies in silico. Stud. Sci. Complex. **12**(2), 321–342 (1990)
6. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)
7. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf. (BRIN) **3**(2), 119–131 (2016)
8. Holzinger, A., Plass, M., Holzinger, K., Crişan, G.C., Pintea, C.-M., Palade, V.: Towards interactive machine learning (iML): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In: Buccafurri, F., Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-ARES 2016. LNCS, vol. 9817, pp. 81–95. Springer, Heidelberg (2016). doi:10.1007/978-3-319-45507-5_6
9. Wang, Z., Butner, J., Kerketta, R., Cristini, V., Deisboeck, T.: Simulating cancer growth with multiscale agent-based modeling. Semin. Cancer Biol. **30**, 70–78 (2015)
10. Szabó, A., Merks, R.M.: Cellular potts modeling of tumor growth, tumor invasion, and tumor evolution. Front. Oncol. **3**, 87 (2013)
11. Von Neumann, J.: The general and logical theory of automata. Cereb. Mech. Behav. **1**(41), 1–2 (1951)
12. Neumann, J.V., Burks, A.W.: Theory of self-reproducing automata (1966)
13. Ulam, S.: Some ideas and prospects in biomathematics. Ann. Rev. Biophys. Bioeng. **1**(1), 277–292 (1972)
14. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
15. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (2000). doi:10.1007/978-1-4757-3264-1
16. Mantas, J.: Electronic health record. Stud. Health Technol. Inf. **65**, 250–257 (2002)
17. Waugh, D.F.: Protein-protein interactions. Adv. Protein Chem. **9**, 325–437 (1954)
18. Pawson, T., Nash, P.: Protein-protein interactions define specificity in signal transduction. Genes Dev. **14**, 1027–1047 (2000)
19. Przulj, N., Wigle, D., Jurisica, I.: Functional topology in a network of protein interactions. Bioinformatics **20**(3), 340–348 (2004)
20. Jeanquartier, F., Jean-Quartier, C., Holzinger, A.: Integrated web visualizations for protein-protein interaction databases. BMC Bioinf. **16**, 195 (2015)
21. Wagoner, J.K.: Occupational carcinogenesis: the two hundred years since percivall pott. Ann. N. Y. Acad. Sci. **271**(1), 1–4 (1976)
22. Trisilowati, Mallet, D.G.: In silico experimental modeling of cancer treatment. ISRN Oncol. **2012**, 828701 (2012)
23. Kotlyar, M., Pastrello, C., Pivetta, F., Sardo, A.L., Cumbaa, C., Li, H., Naranian, T., Niu, Y., Ding, Z., Vafaee, F., et al.: In silico prediction of physical protein interactions and characterization of interactome orphans. Nat. Methods **12**(1), 79–84 (2015)

24. Snell, E.H., Lauricella, A.M., Potter, S.A., Luft, J.R., Gulde, S.M., Collins, R.J., Franks, G., Malkowski, M.G., Cumbaa, C., Jurisica, I., et al.: Establishing a training set through the visual analysis of crystallization trials. Part II: crystal examples. Acta Crystallogr. Sect. D: Biol. Crystallogr. **64**(11), 1131–1137 (2008)
25. Jeanquartier, F., Jean-Quartier, C., Cemernek, D., Holzinger, A.: In silico modeling for tumor growth visualization. BMC Syst. Biol. **10**(1), 1 (2016)
26. Jeanquartier, F., Jean-Quartier, C., Schreck, T., Cemernek, D., Holzinger, A.: Integrating open data on cancer in support to tumor growth analysis. In: Renda, M.E., Bursa, M., Holzinger, A., Khuri, S. (eds.) ITBAM 2016. LNCS, vol. 9832, pp. 49–66. Springer, Heidelberg (2016). doi:10.1007/978-3-319-43949-5_4
27. Edelman, L.B., Eddy, J.A., Price, N.D.: In silico models of cancer. Wiley Interdisc. Rev. Syst. Biol. Med. **2**(4), 438–459 (2010)
28. Russell, S., Norvig, P.: Artificial Intelligence. Prentice-Hall, Englewood Cliffs (1995)
29. Macal, C.M., North, M.J.: Tutorial on agent-based modelling and simulation. J. Simul. **4**(3), 151–162 (2010)
30. Bonabeau, E.: Agent-based modeling: methods and techniques for simulating human systems. Proc. Nat. Acad. Sci. **99**(suppl 3), 7280–7287 (2002)
31. Starruß, J., de Back, W., Brusch, L., Deutsch, A.: Morpheus: a user-friendly modeling environment for multiscale and multicellular systems biology. Bioinformatics **30**(9), 1331–1332 (2014)
32. Benzekry, S., Lamont, C., Beheshti, A., Tracz, A., Ebos, J.M., Hlatky, L., Hahnfeldt, P.: Classical mathematical models for description and prediction of experimental tumor growth. PLoS Comput. Biol. **10**(8), e1003800 (2014)
33. Laird, A.K.: Dynamics of tumour growth. Br. J. Cancer **18**, 490–502 (1964)
34. Loeb, L.: Tissue growth and tumor growth. J. Cancer Res. **2**, 135 (1917)
35. Gocka, E.F., Reed, L.J.: A method of fitting non-symmetric gompertz functions for characterising malignant growth. Int. J. Biomed. Comput. **8**, 247–254 (1977)
36. Glazier, F., Glazier, J.A.: Simulation of biological cell sorting using a two-dimensional extended potts model. Phys. Rev. Lett. **69**(13), 2013–2016 (1992)
37. Gardner, M.: Mathematical games: the fantastic combinations of John Conway's new solitaire game "life". Sci. Am. **223**(4), 120–123 (1970)
38. Wolfram, S.: Statistical mechanics of cellular automata. Rev. Modern Phys. **55**(3), 601 (1983)
39. Qi, A.S., Zheng, X., Du, C.Y., An, B.S.: A cellular automaton model of cancerous growth. J. Theor. Biol. **161**(1), 1–12 (1993)
40. Smolle, J., Stettner, H.: Computer simulation of tumour cell invasion by a stochastic growth model. J. Theor. Biol. **160**(1), 63–72 (1993)
41. Smolle, J.: Cellular automaton simulation of tumour growth-equivocal relationships between simulation parameters and morphologic pattern features. Anal. Cellular Pathol. **17**(2), 71–82 (1998)
42. Kansal, A.R., Torquato, S., Harsh, G., Chiocca, E., Deisboeck, T.: Simulated brain tumor growth dynamics using a three-dimensional cellular automaton. J. Theor. Biol. **203**(4), 367–382 (2000)
43. Patel, A.A., Gawlinski, E.T., Lemieux, S.K., Gatenby, R.A.: A cellular automaton model of early tumor growth and invasion: the effects of native tissue vascularity and increased anaerobic tumor metabolism. J. Theor. Biol. **213**(3), 315–331 (2001)

44. Alarcón, T., Byrne, H.M., Maini, P.K.: A cellular automaton model for tumour growth in inhomogeneous environment. J. Theor. Biol. **225**(2), 257–274 (2003)
45. Gerlee, P., Anderson, A.R.: An evolutionary hybrid cellular automaton model of solid tumour growth. J. Theor. Biol. **246**(4), 583–603 (2007)
46. Brutovsky, B., Horvath, D., Lisy, V.: Inverse geometric approach for the simulation of close-to-circular growth. The case of multicellular tumor spheroids. Phys. A Stat. Mech. Appl. **387**(4), 839–850 (2008)
47. Chaplain, M., Anderson, A.: Mathematical modelling, simulation and prediction of tumour-induced angiogenesis. Invasion Metastasis **16**(4–5), 222–234 (1995)
48. Anderson, A.R., Chaplain, M.: Continuous and discrete mathematical models of tumor-induced angiogenesis. Bull. Math. Biol. **60**(5), 857–899 (1998)
49. Markus, M., Böhm, D., Schmick, M.: Simulation of vessel morphogenesis using cellular automata. Math. Biosci. **156**(1), 191–206 (1999)
50. de Pillis, L.G., Mallet, D.G., Radunskaya, A.E.: Spatial tumor-immune modeling. Comput. Math. Methods Med. **7**(2–3), 159–176 (2006)
51. Mallet, D.G., De Pillis, L.G.: A cellular automata model of tumor-immune system interactions. J. Theor. Biol. **239**(3), 334–350 (2006)
52. Vidyasagar, M.: Machine learning methods in the computational biology of cancer, vol. 470. The Royal Society (2014)
53. Cruz, J.A., Wishart, D.S.: Applications of machine learning in cancer prediction and prognosis. Cancer Inf. **2**, 59–77 (2006)
54. Madhukar, N.S., Elemento, O., Pandey, G.: Prediction of genetic interactions using machine learning and network properties. Front. Bioeng. Biotechnol. **3**, 172 (2015)
55. Lisboa, P.J., Vellido Alcacena, A., Tagliaferri, R., Napolitano, F., Ceccarelli, M., Martín Guerrero, J.D., Biganzoli, E.: Data mining in cancer research. IEEE Comput. Intell. Magaz. **5**(1), 14–18 (2010)
56. Vellido, A., Biganzoli, E., Lisboa, P.J.: Machine learning in cancer research: implications for personalised medicine. In: ESANN, pp. 55–64 (2008)
57. Upstill-Goddard, R., Eccles, D., Fliege, J., Collins, A.: Machine learning approaches for the discovery of gene-gene interactions in disease data. Brief. Bioinf. **14**(2), 251–260 (2013)
58. Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armañanzas, R., Santafé, G., Pérez, A., et al.: Machine learning in bioinformatics. Brief. Bioinf. **7**(1), 86–112 (2006)
59. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2014)
60. Hu, X., Cammann, H., Meyer, H.A., Miller, K., Jung, K., Stephans, C.: Artificial neural networks and prostate cancer-tools for diagnosis and management. Nat. Rev. Urol. **10**, 174–182 (2013)
61. Eckel, S.P., Baumbach, J., Hauschild, A.C.: On the importance of statistics in breath analysis-hope or curse? J. Breath Res. **8**(1), 012001 (2014)
62. Morris, M., Greiner, R., Sander, J., Murtha, A., Schmidt, M.: Learning a classification-based glioma growth model using MRI data. J. Comput. **1**(7), 21–31 (2006)
63. González-Vélez, H., Mier, M., Julià-Sapé, M., Arvanitis, T.N., García-Gómez, J.M., Robles, M., Lewis, P.H., Dasmahapatra, S., Dupplaw, D., Peet, A., et al.: Healthagents: distributed multi-agent brain tumor diagnosis and prognosis. Appl. Intell. **30**(3), 191–202 (2009)

64. Parmar, C., Grossmann, P., Bussink, J., Lambin, P., Aerts, H.J.: Machine learning methods for quantitative radiomic biomarkers. Sci. Rep. **5**, 13087 (2015)
65. Kherlopian, A.R., Song, T., Duan, Q., Neimark, M.A., Po, M.J., Gohagan, J.K., Laine, A.F.: A review of imaging techniques for systems biology. BMC Syst. Biol. **2**, 74 (2008)
66. Buchen, L.: Cancer: missing the mark. Nature **471**(7339), 428–432 (2011)
67. Wang, J., Zuo, Y., Man, Y., Avital, I., Stojadinovic, A., Liu, M., Yang, X., Varghese, R.S., Tadesse, M.G., Ressom, H.W.: Pathway and network approaches for identification of cancer signature markers from omics data. J. Cancer **6**(1), 54–65 (2015)
68. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D., Ideker, T.: Network-based classification of breast cancer metastasis. Mol. Syst. Biol. **3**, 140 (2007)
69. Liu, N., Liu, X., Zhou, N., Wu, Q., Zhou, L., Li, Q.: Gene expression profiling and bioinformatics analysis of gastric carcinoma. Exp. Mol. Pathol. **96**(3), 361–366 (2014)
70. Wong, Y.H., Chen, R.H., Chen, B.S.: Core and specific network markers of carcinogenesis from multiple cancer samples. J. Theor. Biol. **362**, 17–34 (2014)
71. Sanz-Pamplona, R., Aragüés, R., Driouch, K., Martín, B., Oliva, B., Gil, M., Boluda, S., Fernández, P.L., Martínez, A., Moreno, V., Acebes, J.J., Lidereau, R., Reyal, F., Van de Vijver, M.J., Sierra, A.: Expression of endoplasmic reticulum stress proteins is a candidate marker of brain metastasis in both ErbB-2+ and ErbB-2- primary breast tumors. Am. J. Pathol. **179**(2), 564–579 (2011)
72. Wang, Y.C., et al.: A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. BMC Med. Genom. **4**(1), 2 (2011)
73. Luo, T., Wu, S., Shen, X., Li, L.: Network cluster analysis of protein-protein interaction network identified biomarker for early onset colorectal cancer. Mol. Biol. Rep. **40**(12), 6561–6568 (2013)
74. Schramm, S.J., Li, S.S., Jayaswal, V., Fung, D.C.Y., Campain, A.E., Pang, C.N.I., Scolyer, R.A., Yang, Y.H., Mann, G.J., Wilkins, M.R.: Disturbed protein-protein interaction networks in metastatic melanoma are associated with worse prognosis and increased functional mutation burden. Pigment Cell Melanoma Res. **26**(5), 708–722 (2013)
75. Zhang, Y., Yang, C., Wang, S., Chen, T., Li, M., Wang, X., Li, D., Wang, K., Ma, J., Wu, S., Zhang, X., Zhu, Y., Wu, J., He, F.: Liveratlas: a unique integrated knowledge database for systems-level research of liver and hepatic disease. Liver Int. Off. J. Int. Assoc. Study Liver **33**(8), 1239–1248 (2013)
76. Ahn, J., Yoon, Y., Yeu, Y., Lee, H., Park, S.: Impact of TGF-b on breast cancer from a quantitative proteomic analysis. Comput. Biol. Med. **43**(12), 2096–2102 (2013)
77. Oh, J.H., Deasy, J.O.: A literature mining-based approach for identification of cellular pathways associated with chemoresistance in cancer. Brief. Bioinf. **17**(3), 468–478 (2016)
78. Majewski, I.J., Bernards, R.: Taming the dragon: genomic biomarkers to individualize the treatment of cancer. Nat. Med. **17**(3), 304–312 (2011)
79. Li, A., Walling, J., Ahn, S., Kotliarov, Y., Su, Q., Quezado, M., Oberholtzer, J.C., Park, J., Zenklusen, J.C., Fine, H.A.: Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. Cancer Res. **69**(5), 2091–2099 (2009)
80. Ceccarelli, M., Barthel, F.P., Malta, T.M., Sabedot, T.S., Salama, S.R., Murray, B.A., Morozova, O., Newton, Y., Radenbaugh, A., Pagnotta, S.M., et al.: Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell **164**(3), 550–563 (2016)

81. Lalonde, E., Ishkanian, A.S., Sykes, J., Fraser, M., Ross-Adams, H., Erho, N., Dunning, M.J., Halim, S., Lamb, A.D., Moon, N.C., et al.: Tumour genomic and microenvironmental heterogeneity for integrated prediction of 5-year biochemical recurrence of prostate cancer: a retrospective cohort study. Lancet Oncol. **15**(13), 1521–1532 (2014)
82. Mungall, C.J., Washington, N.L., Nguyen-Xuan, J., Condit, C., Smedley, D., Köhler, S., Groza, T., Shefchek, K., Hochheiser, H., Robinson, P.N., et al.: Use of model organism and disease databases to support matchmaking for human disease gene discovery. Hum. Mutat. **36**(10), 979–984 (2015)
83. Clarke, C., Madden, S.F., Doolan, P., Aherne, S.T., Joyce, H., O'Driscoll, L., Gallagher, W.M., Hennessy, B.T., Moriarty, M., Crown, J., Kennedy, S., Clynes, M.: Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. Carcinogenesis **34**(10), 2300–2308 (2013)
84. Yang, Y., et al.: Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat. Commun. **5**, 262–272 (2014)
85. Liu, R., Lv, Q.L., Yu, J., Hu, L., Zhang, L.H., Cheng, Y., Zhou, H.H.: Correlating transcriptional networks with pathological complete response following neoadjuvant chemotherapy for breast cancer. Breast Cancer Res. Treat. **151**(3), 607–618 (2015)
86. Yeh, H.Y., et al.: Identifying significant genetic regulatory networks in the prostate cancer from microarray data based on transcription factor analysis and conditional independency. BMC Med. Genom. **2**(1), 70 (2009)
87. Remo, A., et al.: Systems biology analysis reveals NFAT5 as a novel biomarker and master regulator of inflammatory breast cancer. J. Trans. Med. **13**(1), 138 (2015)
88. Akutekwe, A., Seker, H.: Inference of nonlinear gene regulatory networks through optimized ensemble of support vector regression and dynamic Bayesian networks. In: Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference 2015, pp. 8177–8180 (2015)
89. Jörnsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T.E.M., Nordlander, B., Sander, C., Gennemark, P., Funa, K., Nilsson, B., Lindahl, L., Nelander, S.: Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. Mol. Syst. Biol. **7**, 486 (2011)
90. Shi, X., Zhao, Q., Huang, J., Xie, Y., Ma, S.: Deciphering the associations between gene expression and copy number alteration using a sparse double Laplacian shrinkage approach. Bioinformatics (Oxford, England) **31**(24), 3977–3983 (2015)
91. Leung, A., Bader, G.D., Reimand, J.: HyperModules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery. Bioinformatics **30**(15), 2230–2232 (2014)
92. Wong, S.W., Cercone, N., Jurisica, I.: Comparative network analysis via differential graphlet communities. Proteomics **15**(2–3), 608–617 (2015)
93. Leung, M.K., Delong, A., Alipanahi, B., Frey, B.J.: Machine learning in genomic medicine: a review of computational problems and data sets. Proc. IEEE **104**(1), 176–197 (2016)
94. Shivade, C., Raghavan, P., Fosler-Lussier, E., Embi, P.J., Elhadad, N., Johnson, S.B., Lai, A.M.: A review of approaches to identifying patient phenotype cohorts using electronic health records. J. Am. Med. Inf. Assoc. **21**(2), 221–230 (2014)

95. Hochheiser, H., Castine, M., Harris, D., Savova, G., Jacobson, R.S.: An information model for computable cancer phenotypes. BMC Med. Inf. Decis. Making **16**(1), 121 (2016)
96. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. **34**(2), 113–127 (2005)
97. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2016. CA Cancer J. Clin. **66**(1), 7–30 (2016)
98. Menden, M.P., Iorio, F., Garnett, M., McDermott, U., Benes, C.H., Ballester, P.J., Saez-Rodriguez, J.: Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS one **8**(4), e61318 (2013)
99. Gupta, S., Tran, T., Luo, W., Phung, D., Kennedy, R.L., Broad, A., Campbell, D., Kipp, D., Singh, M., Khasraw, M., et al.: Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ Open **4**(3), e004007 (2014)
100. Zhao, D., Weng, C.: Combining pubmed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. J. Biomed. Inf. **44**(5), 859–868 (2011)
101. Shanmugasundaram, P., Viswanath, V., Sankar, A., Ravichandiran, V.: Metabolomics: a cancer diagnostic tool. J. Pharm. Res. **5**(12), 5210 (2012)
102. Handa, H., Usuba, A., Maddula, S., Baumbach, J.I., Mineshita, M., Miyazawa, T.: Exhaled breath analysis for lung cancer detection using ion mobility spectrometry. PloS one **9**(12), e114555 (2014)
103. Cairns, R.A., Harris, I.S., Mak, T.W.: Regulation of cancer cell metabolism. Nat. Rev. Cancer **11**(2), 85–95 (2011)
104. O'Shea, K., Cameron, S.J., Lewis, K.E., Lu, C., Mur, L.A.: Metabolomic-based biomarker discovery for non-invasive lung cancer screening: a case study. Biochimica et Biophysica Acta **1860**(11, Part B), 2682–2687 (2016). Systems Genetics - Deciphering the Complex Disease with a Systems Approach
105. Gaul, D.A., Mezencev, R., Long, T.Q., Jones, C.M., Benigno, B.B., Gray, A., Fernández, F.M., McDonald, J.F.: Highly-accurate metabolomic detection of early-stage ovarian cancer. Sci. Rep. **5**, 16531 (2015)
106. Hauschild, A.C., Baumbach, J.I., Baumbach, J.: Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification. Genet. Mol. Res. **11**(3), 2733–2744 (2012)
107. Brunelli, L., Caiola, E., Marabese, M., Broggini, M., Pastorelli, R.: Comparative metabolomics profiling of isogenic KRAS wild type and mutant NSCLC cells in vitro and in vivo. Sci. Rep. **6** (2016). doi:10.1038/srep28398. Nature Publishing Group
108. Bertini, I., Cacciatore, S., Jensen, B.V., Schou, J.V., Johansen, J.S., Kruhøffer, M., Luchinat, C., Nielsen, D.L., Turano, P.: Metabolomic NMR fingerprinting to identify and predict survival of patients with metastatic colorectal cancer. Cancer Res. **72**(1), 356–364 (2012)
109. Wei, S., Liu, L., Zhang, J., Bowers, J., Gowda, G.N., Seeger, H., Fehm, T., Neubauer, H.J., Vogel, U., Clare, S.E., Raftery, D.: Metabolomics approach for predicting response to neoadjuvant chemotherapy for breast cancer. Mol. Oncol. **7**(3), 297–307 (2013)
110. Jean-Quartier, C., Jeanquartier, F., Cemernek, D., Holzinger, A.: Tumor growth simulation profiling. In: Renda, M.E., Bursa, M., Holzinger, A., Khuri, S. (eds.) ITBAM 2016. LNCS, vol. 9832, pp. 208–213. Springer, Heidelberg (2016). doi:10.1007/978-3-319-43949-5_16

111. Koch, L.: Genetic screen: a network to guide precision cancer therapy. Nat. Rev. Genet. **17**, 504–505 (2016)
112. Kotlyar, M., Fortney, F., Jurisica, I.: Network-based characterization of drug-regulated genes, drug targets, and toxicity. Methods **57**(4), 477–485 (2012)
113. Fortney, K., Griesman, G., Kotlyar, M., Pastrello, C., Angeli, M., Tsao, M.S., Jurisica, I.: Prioritizing therapeutics for lung cancer: an integrative meta-analysis of cancer gene signatures and chemogenomic data. PLoS Comp. Biol. **11**(3), e1004068 (2015)
114. Sacha, D., Sedlmair, M., Zhang, L., Lee, J.A., Weiskopf, D., North, S., Keim, D.: Human-centered machine learning through interactive visualization: review and open challenges. In: Proceedings of the 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (2016)
115. Kornhauser, D., Wilensky, U., Rand, W.: Design guidelines for agent based model visualization. J. Artif. Soc. Soc. Simul. **12**(2), 1 (2009)
116. Savas, S., Geraci, J., Jurisica, I., Liu, G.: A comprehensive catalogue of functional genetic variations in the EGFR pathway: protein-protein interaction analysis reveals novel genes and polymorphisms important for cancer research. Int. J. Cancer **125**(6), 1257–1265 (2009)
117. Rodriguez-Brenes, I.A., Komarova, N.L., Wodarz, D.: Tumor growth dynamics: insights into evolutionary processes. Trends Ecol. Evol. **28**(10), 597–604 (2013)
118. Blair, R.H., Trichler, D.L., Gaille, D.P.: Mathematical and statistical modeling in cancer systems biology. Front. Physiol. **3**, 227 (2012). doi:10.3389/fphys.2012.00227. Frontiers Research Foundation
119. Holzinger, A.: Biomedical Informatics: Discovering Knowledge in Big Data. Springer, New York (2014)
120. Holzinger, A., Dehmer, M., Jurisica, I.: Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinf. **15**(Suppl 6), I1 (2014)
121. Sturm, W., Schreck, T., Holzinger, A., Ullrich, T.: Discovering medical knowledge using visual analytics: a survey on methods for systems biology and *-omics data. In: Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine, pp. 71–81. Eurographics Association (2015)
122. 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: State-of-the-Art and Future Challenges, vol. 8401, pp. 117–140. Springer, Heidelberg (2014)
123. Otasek, D., Pastrello, C., Holzinger, A., Jurisica, I.: Visual data mining: effective exploration of the biological universe. In: Holzinger, A., Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics: State-of-the-Art and Future Challenges. LNCS, vol. 8401, pp. 19–33. Springer, Heidelberg (2014)
124. Pastrello, C., Pasini, E., Kotlyar, M., Otasek, D., Wong, S., Sangrar, W., Rahmati, S., Jurisica, I.: Integration, visualization and analysis of human interactome. Biochem. Biophys. Res. Commun. **445**(4), 757–773 (2014)

Chapter 3

# Conclusion

The herewith presented papers showcase the scientific introduction and examination of the topic "Visualization Support For In Silico Medicine". Results highlight certain challenges for using and extending existing computational approaches, but also challenges when it comes to introducing novel approaches.

The state-of-the-art report on methods for visual analysis of heterogeneous data in biomedical informatics [8] listed several open problems, including the rarity of tightly integrated tools and the balance between focused and comprehensive approaches. The list continues with the necessity for improving usability of analysis processes, but also proposed the idea of attention routing for helping users to find good starting points for analysis. Additionally, it has been stated that several research challenges existed in the application of topology-based methods for visualization. Yet, the work mentioned that these methods were becoming more and more popular. Existing challenges have not been specific to the biomedical domain but rather general. Several works regarding this topic have been published also showing successful applications for the visual analysis of many types of data of different domains. Challenges of topology-based methods for visualization remain.

Data fusion has been and remains another recurring topic in biomedical research, mentioned in [6, 7, 8].

Uncertainty visualization and uncertainty awareness in analysis processes was mentioned as an additional big open problem listed in the state-of-the-art work [8], partially approached in [5] and mentioned again in [6, 7]. This topic will most probably remain interesting within the next years, too.

Above all, there are two central and recurring themes: First, it is essential to further enhance visualization integration in analysis tools in the biomedical domain. Second, we have to foster greater collaboration amongst the two scientific research fields of biomedical science and computer science.

Last but not least, many interesting challenges remain regarding visualization related to tumor growth. Topics include, next to visualizing probability and uncertainty, comparative visualization and model visualization.

There has been an increase in the availability of visualization tools for systems biology. I believe that biomedical research in general and research on tumor growth in particular will benefit from the systems biology perspective. By facilitating visualization integration and further making use of computational techniques we can help answering key questions in fundamental biomedical research.

...

In memoriam to our beloved friends and family we lost by cancer: We owe them to always keep up with studying, exploring, reflecting and analyzing. We shall never give up the search for a suitable approach to help conquering a merciless disease.

# Bibliography


[1] BIOVIS. 2017. 5th symposium on biological data visualization. http://biovis.net/year/2015/design/update.html.

[2] JEANQUARTIER, F. AND HOLZINGER, A. 2013. *On Visual Analytics and Evaluation in Cell Physiology: A Case Study*. Springer Berlin Heidelberg, Berlin, Heidelberg, 495–502.

[3] JEANQUARTIER, F., JEAN-QUARTIER, C., CEMERNEK, D., AND HOLZINGER, A. 2016. In silico modeling for tumor growth visualization. *BMC Systems Biology 10,* 1, 59.

[4] JEANQUARTIER, F., JEAN-QUARTIER, C., AND HOLZINGER, A. 2015a. Integrated web visualizations for protein-protein interaction databases. *BMC bioinformatics 16,* 1, 1.

[5] JEANQUARTIER, F., JEAN-QUARTIER, C., AND HOLZINGER, A. 2015b. Visualizing uncertainty of rna sequence base pairing variants. Design contest submission to BioVis 5th Symposium on Biological Data Visualization, Dublin, Ireland, 2015.

[6] JEANQUARTIER, F., JEAN-QUARTIER, C., KOTLYAR, M., TOKAR, T., HAUSCHILD, A.-C., JURISICA, I., AND HOLZINGER, A. 2016. Machine learning for in silico modeling of tumor growth. In *Machine Learning for Health Informatics*. Springer International Publishing, Cham, 415–434.

[7] JEANQUARTIER, F., JEAN-QUARTIER, C., SCHRECK, T., CEMERNEK, D., AND HOLZINGER, A. 2016. Integrating open data on cancer in support to tumor growth analysis. In *International Conference on Information Technology in Bio-and Medical Informatics*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 49–66.

[8] TURKAY, C., JEANQUARTIER, F., HOLZINGER, A., AND HAUSER, H. 2014. On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. In *Interactive knowledge discovery and data mining in biomedical informatics*. Springer, Berlin, Heidelberg, 117–140.