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	✓	✓

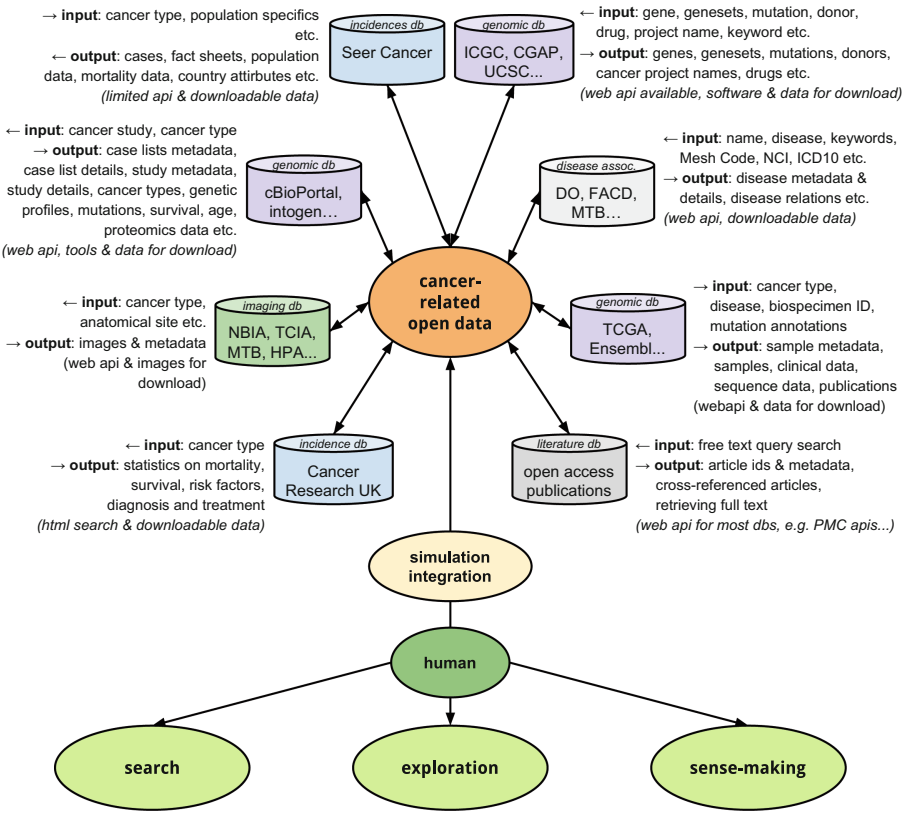

Fig. 1. Overview of cancer databases for integration

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].

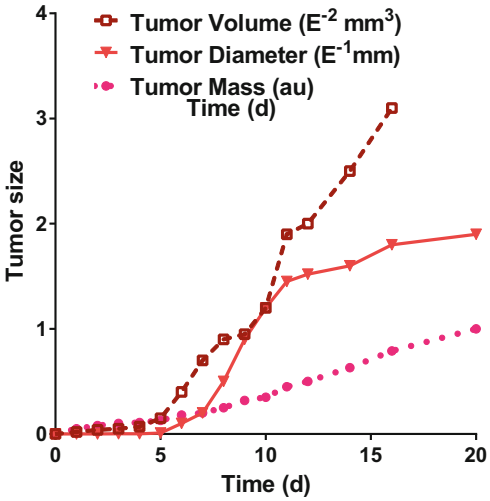

Fig. 2. Literature-curated discrete tumor growth - data samples: various tumor types, determined growth in tumor size, given in miscellaneous units, over time, presented in days.

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

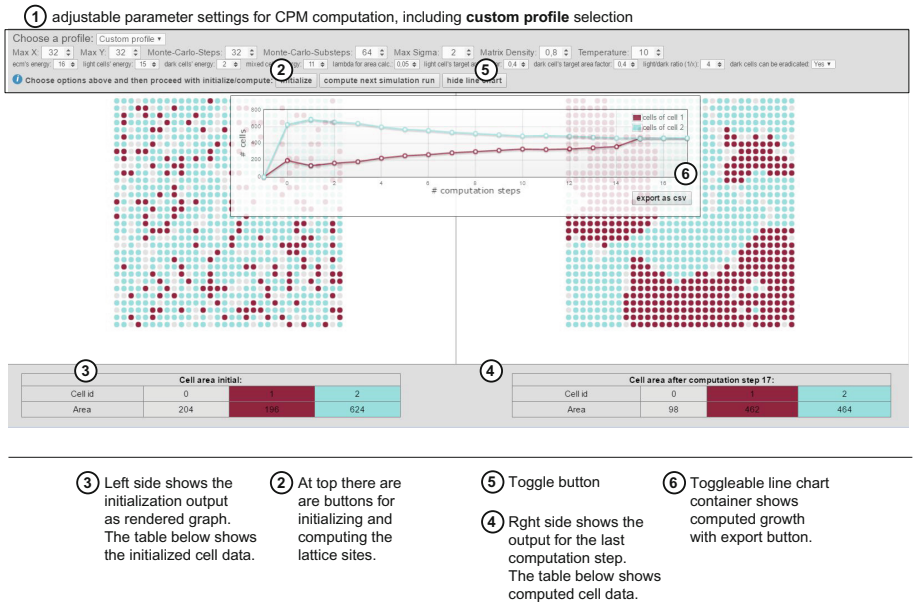

Fig. 3. Overview of User Interface with custom profile showing kinetics and cell sorting after several simulation steps

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.

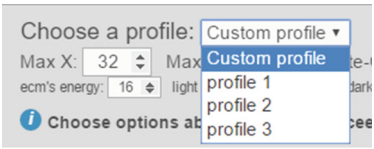

Fig. 4. Screenshot of profile selection possibilities

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 τ. 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

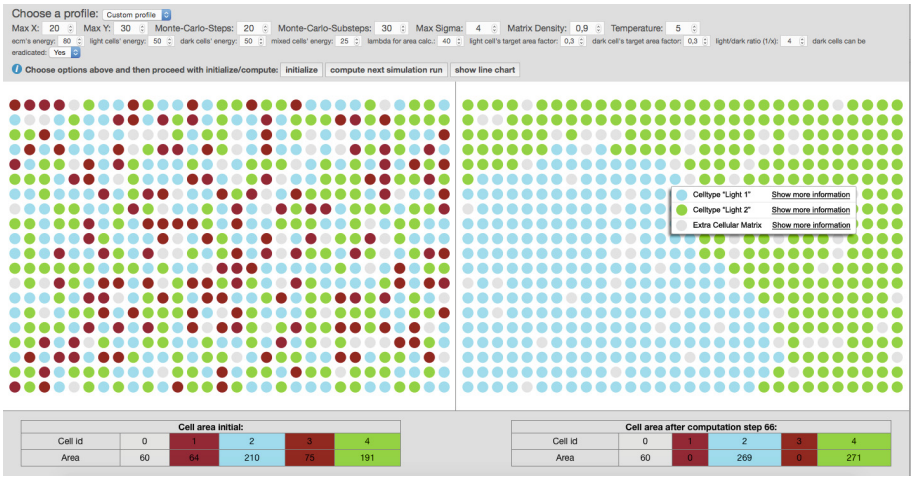

Fig. 5. Screenshot, showing additional information for cell nodes (Color figure online)

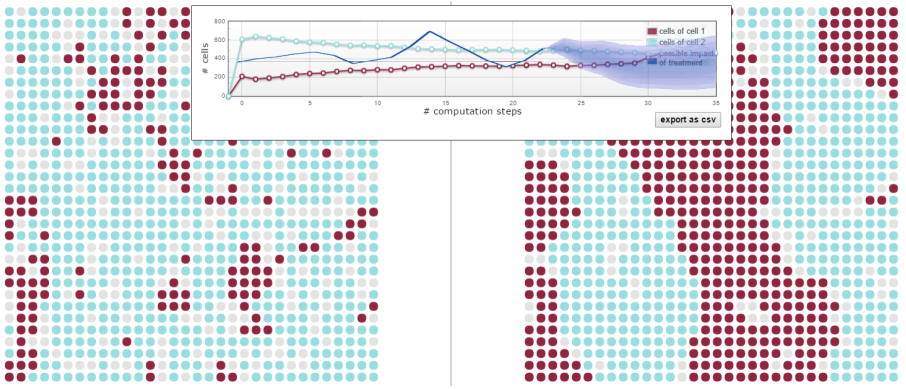

Fig. 6. Mock-Up of a time-line extension showing results of a computation taking additional information on treatment into account

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)