Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Bio-Nano-Informatics: An Integrated Information Management System for Personalized Oncology

Todd H. Stokes[1,*], John Phan[2], C.F. Quo[2], Shuming Nie[2,3], May D. Wang[1-3]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology
[2]The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University
[3]Winship Cancer Institute, Emory University
Atlanta, GA 30332

ABSTRACT

The Emory-Georgia Tech Nanotechnology Center for Personalized and Predictive Oncology is one of the eight national centers funded by the National Cancer Institute (NCI). Its overall goal is to combine nanotechnology and biocomputing with clinical oncology for personalized detection, diagnosis and treatment of human cancer. Within this large-scale and multifaceted center, a key challenge is how to integrate and manage data and resources. Here we have developed an "intelligent" information system for data management, interpretation, and for translation of new results to clinical applications.

Index Terms – LIMS, Intelligence-Based, Information Integration

1. INTRODUCTION

In biomedical informatics research, a major goal is to translate lab-based research to bed-side patient care. Key technologies include genomic and proteomic data mining, sequence analysis, biomarker discovery, and molecular pathway modeling. In 2005, the National Cancer Institute supported Emory-Georgia Tech to establish a national Center of Cancer Nanotechnology Excellence (CCNE). The mission of our CCNE is to combine cancer biology with nanotechnology and informatics to deliver novel molecular imaging probes, nanotherapeutics, and biocomputing tools to detect, diagnose and treat cancer. More than seventy faculty members and clinicians from biology, engineering, and clinical oncology are involved in this large-scale center. Research results from these individual groups are shared with each other and are reported to the NCI periodically. To manage this information flow, it is important to develop advanced laboratory information management systems (LIMS). Current information management systems aim to achieve the following objectives [1-4]: (i) to reduce administrative costs and redundancy; (ii) to develop laboratory methods to facilitate the usage of laboratory instruments and reagents; (iii) to integrate different instruments into an automated workflow architecture; (iv) to facilitate data production, storing, mining, visualization; (v) to ensure data quality and accessibility to other scientists for information dissemination; and (vi) to report research status to funding agencies as required. The current LIM systems, however, are not amenable to extension to large research centers with 10s to 100s of projects and researchers. Furthermore, the nature of multi-disciplinary and collaborative projects specifically requires systems to address data heterogeneity, diverse professional cultures, and traceability at various organizational levels.

2. SYSTEM DEVELOPMENT

1-4244-0033-3/06/$20.00 ©2006 IEEE.

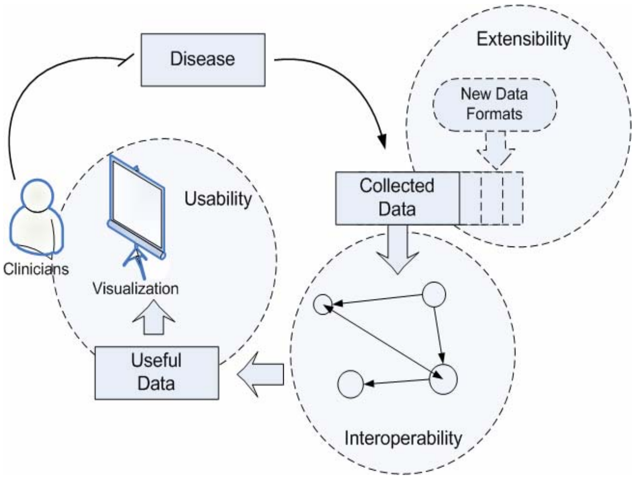


Figure 1: Schematic diagram of the intelligent information processing system showing three key functions: usability, interoperability, and extensibility.

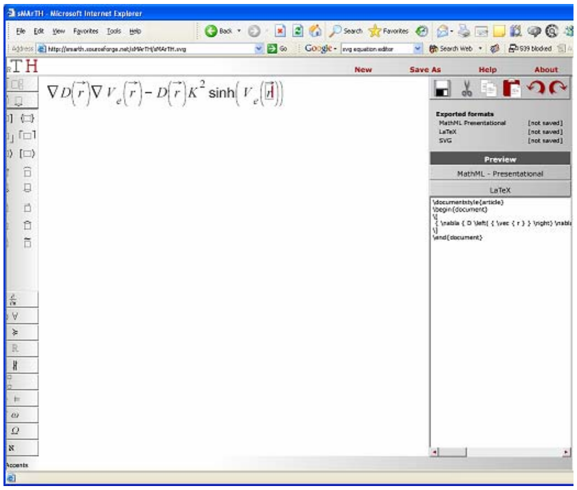
Faced with the specific needs as described above, we have developed an intelligence-based, readily scalable information processing and management system. This system will assist data screening and integration by comparing with existing patents and literature, and will provide a prioritized list based on requirement from users. We have evaluated three factors in our design: usability, interoperability, and extensibility (Figure 1). Usability refers to how easily users from different backgrounds interact with the system based on their specific requirements. Interoperability refers to how the information presented in different data and tool formats can be integrated. Extensibility ensures that the architecture can be extended to new data formats or new user bases without major modifications. The primary approach to extensibility is to make each piece of the design modular and to use common standards to represent inputs and outputs to each module.

Following these three principles, we have designed and developed a new information system (1) that is simple, flexible, and adaptable to new discoveries; and (2) that has a multi-scale information reporting hierarchy ranging from detailed technical data to high-level executive summaries. Also, it is a web-based system that collects research progress updates from each researcher and allows research leaders, directors, and managers to customize and generate status reports based on predefined templates, reporting hierarchies, and intelligent text mining and scoring methods. This system can be configured to deal with a variety of dynamic collaborative efforts and yet is simple enough to be used by project teams with little computational expertise.

Information processing at the user end is designed with an analogy of a newspaper office. Researchers take the role of reporters working on a variety of stories. They submit regular, categorized, and proof-ready updates to the Editor (the center director). When one of their projects reaches a planning period, they submit a project planning update. In this way, researchers give frequent and uniform updates, leaving the pressure of deciding how to format the front page of each edition to the Editor. This saves time and allows researchers to focus more on their research. At the same time, the quality of their reports should be improved because updates are submitted at the moment when researchers are focused on that project rather than recalling and organizing them from memory later.

In our system, content cleaning, expansion or focusing will often take place on the side of the Editor, and this is to ensure that they have plenty of high-quality contents to work with. For example, figures may be required for certain update types on a per-member basis. The information will be scored based on human knowledge such as scientific importance, level of detail, urgency, etc. These scores can then be used to generate interpretation reports which help the center director to understand the discovery and its scope.

The key step for data integration is developed in the LAMP (Linux Apache mySQL PHP) web platform. Cascading Style Sheets (CSS) are used so that the look and feel of the system can be easily modified to integrate within existing LIMS. The email reminder system is dependent on being hosted on the Linux/Unix operation system as cron and sendmail are used. The LaTeX formatting standard is used to enhance the readability of reports. This formatting language has long been used in the scientific community to format papers for submission to conferences and journals and is thus a *de facto* standard for this type of system. Our team has evaluated the HTML standard, but found that it did not provide important features such as automated equation formatting and handling figure references and citations. Other tools and utilities incorporated into the system are LaTeX2rtf (http://latex2rtf.sourceforge.net/) for creating an output format that is readable by Microsoft Word and sMArTH (http://smarth.sourceforge.net/), which gives users a friendly interface for creating complex equations without having to learn LaTeX syntax. The flexibility of our system is built into the data model used in the reporting database. Currently, the database is composed of less than 20 tables and maintains a history of updates, figures, and reports as they are generated. The result is shown in Figure 2.



$$\nabla D\left(\vec{r}\right)\nabla V_e\left(\vec{r}\right) - D\left(\vec{r}\right)K^2 \sinh\left(V_e\left(\vec{r}\right)\right)$$

```
\documentstyle{article}
\begin{document}
\[
 { \nabla { D \left( { \vec { r } } \right) \nabl
\]
\end{document}
```

Figure 2. Screen shot of the sMArTH interface. Many medical researchers are not familiar with the LaTeX syntax and thus would have trouble representing equations in their status updates. This friendly web interface will automatically import the LaTeX representation of equations entered visually by the user.

In addition to text integration, we also use visualization to represent progress for different projects. All project results are designed to have a web-based SVG representation. Each tool window is implemented using HTML frames, which allows for resizing and reorganizing the windows. Figure 3 shows the new prototyping user-interface. The upper left frame is the starting point, where data sets are uploaded or identified. For biomarker selection, these data sets are most likely oligonucleotide microarrays, mass spectrometry data, or protein chips. In the analysis window (top center), the user can run high performance codes on our computing cluster. The example in Figure 3 shows a visualization of a genetic algorithm that uses pattern recognition to find markers in microarray data. The results interpretation window (top right) shows a web service built on top of the Gene Ontology that will be updated to use a new gene significance indicator. This new indicator will replace the p-value statistics, which has been found by our work with collaborators to be somewhat misleading because of the incomplete state of the Gene Ontology. P-value is also ill-suited for comparative studies.

The information behind these blocks can be integrated to support knowledge-based text processing and mining. The final frame along the bottom shows an assortment of validation methods for biomarker discovery: confocal microscopy, tissue staining, and fluorescent nanoparticles. These results can be quantified, stored using the general storage format mentioned in a previous section, and can further contribute to the knowledge base available to the analysis (or modeling) package.

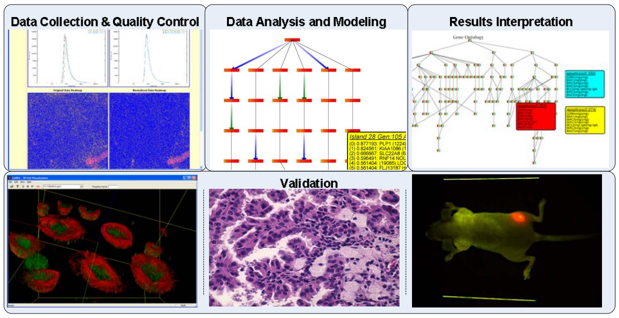


Figure 3. Visual representation of research results in a large-scale and multidisciplinary center. Behind each panel in the integrated web application is a suite of integrated web services. The user interface uses Javascript to support quick Drag-and-Drop queries of data values between major components.

3. DISCUSSION

The design and implementation of this fully integrated data interpretation and integration system has its origins in knowledge-based text mining and SVG-based visualization. This system has proven an important accomplishment in our cancer nanotechnology center. We are also developing an interface to allow the Editor to do more high-level reorganization before the report compile and generation step is executed. For future work specific to CCNE, the plan is to integrate this reporting system with GForge, an open-source software bug tracking system that will organize problems reported by users of our software and report on the progress of our developers in resolving those problems.

Currently, the usability study metric is being developed to quantitatively analyze the system accessibility to non-computer savy users. We have designed and developed this this to be NCI caBIG compatible. After extensitive internal testing, this system will be deployed from websites at Georgia Tech and Emory University (http://www.bio-miblab.org, http://www.wciccne.org) for other Centers of Cancer Nanotechnology Excellence to download and use. Also, it will be interfacing with NCI caBIG (cancer Biomedical Informatics Grid) for it to distribute to wider cancer research community.

4. ACKNOWLEDGMENT

This work was supported by grants from the National Institutes of Health (U54CA119338, R01CA108468 to M.D.W. and S.N.) and the Georgia Cancer Coalition Distinguished Cancer Scholar Award (to M.D.W.).

5. REFERENCES

[1] M. Steinlechner, W. Parson, “Automation and high through-put for a DNA database laboratory: development of a laboratory information management system,” Croat Med J, vol. 42, no. 3, pp. 252-255, 2001.

[2] H. Sanchez-Villeda, S. Schroeder, M. Polacco, M. McMullen, S. Havermann, G. Davis et al, “Development of an integrated laboratory information management system for the maize mapping project,” Bioinformatics, vol. 19, no. 16, pp. 2022-2030, May 2003.

[3] M-M. Cordonnier-Pratt, C. Liang, H. Wang, D.S. Kolychev, F. Sun, R. Freeman, R. Sullican, L.H. Pratt, “MAGIC database and interfaces: an integrated package for gene discovery and expression, ” Comp Funct Genom, vol. 5, pp. 268-275, 2004.

[4] K. Thurow, B. Gode, U. Dingerdissen, N. Stoll, “Laboratory information management systems for life sciences applications,” Org Proc Res and Dev, vol. 8, pp. 970-982, 2004.

[5] Chang F. Quo, B. Wu, May D. Wang. Development of a Laboratory Information System for Cancer Collaboration Projects. Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. Shanghai, China. Sept. 2005.

[6] P. A. Pevzner, *Educating biologists in the 21st century: bioinformatics scientists versus bioinformatics technicians*, Bioinformatics, 20 (2004), pp. 2159-2161.

[7] P. Shafer, T. Isganitis and G. Yona, Hubs of knowledge: using the functional link structure in Biozon to mine for biologically significant entities, Bmc Bioinformatics, 7 (2006)

[8] Landau, R. H., D. Vediner, et al. (2002). "Future scientific digital documents with MathML, XML, and SVG." Computing in Science & Engineering 4(2): 77-85.