BioEMR: an integrative framework for cancer research with multiple genomic technologies

Yu Rang Park[1], Yun Jung Bae[1], Ju Han Kim[1, 2*]

[1] Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine; [2] Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

Abstract

The rapid development of omic technologies facilitate cancer researchers to apply multiple genomic technologies simultaneously. In fact, the complex nature of cancer biology is the reason why we need tools for data integration. Given the complexity of managing multiple technologies and dataset formats, several projects have been introduced including cancer Biomedical Informatics Grid (caGRID) and the Biomedical Research Institute Domain Group (BRIDG) with limited applicability. We introduce an object-oriented data model, Cancer Genomics Object Model (CaGe-OM) for multiple genomics data and Xperanto-CaGe, a web-based application using CaGe-OM with hybrid object-relational mapping technique. The hybrid approach uses objectrelational mapping which is extended to include dynamic structure by using Entity-Attribute-Value (EAV) model. CaGe-OM and Xperanto-CaGe are an attempt to establish a comprehensive framework for integrated storage and interpretation of clinical and multiple genomics data and to facilitate model-level integration of other newly emerging data types. A pilot implementation for the integrated clinical, histo-pathological and genomic information systems is introduced.

Background

The emergence of a variety of high-throughput technologies produces overwhelming amount of heterogeneous genomic data in a quest to measure multi parts of a biological system simultaneously (mRNA, proteins, metabolites, etc)[1]. For managing and representing theses genomics data, several technology-specific data models have been proposed, including MAGE-OM for transcriptomics [2], PEDRo for proteomics [3], SMAR [4], ArMET [5], and MIAMET [6] for metabolomics, and Tissue MicroArray-Object Model (TMA-OM) [7] for tissue microarray.

Despite the increasing number of cancer studies using multiple genomic technologies, there is no integrated data model for multiple functional genomics experimental and clinical data. Several initial efforts have been introduced for solving this problem. The applications provided by National Cancer Institute (NCI) cancer Biomedical Informatics Grid (caGRID) and the Biomedical Research Institute Domain Group (BRIDG) are not yet fully completed and the large-scale architectures and some inter-dependency problems between modules can be prohibitively costly for a real-world application with limited purposes [8, 9]. The Chemical Effects in Biological Systems (CEBS) does not include cancer genomics data but focuses on functional genomic data in toxicology domain [10].

We proposed Cancer Genomics Object Model (CaGe-OM), for representing data from multiple omics technologies and clinico-histopathological domain in cancer research [11] along with TMA-OM [7]. In the present study, we implemented a web-based application, Xperanto-CaGe, using hybrid object-relational mapping technique in an attempt to establish a comprehensive framework for integrated storage and interpretation of clinical and multiple genomics data types with inclusively flexible design.

Result

Object model

To design an integrated data model for multiple functional genomics data in cancer research in CaGe-OM, we referenced four experimental data models (i.e. FuGe-OM, MAGE-OM, PEDRo and TMA-OM). For modeling clinical and histopathological data, we analyzed cancer management workflow and referenced document models of clinical and histopathological information like College of American Pathologist (CAP) Cancer Protocols (CPs) and National Cancer Institute (NCI) Common Data Emement (CDEs) [11].

CaGe-OM is a data model containing 183 classes grouped into 25 packages (Fig. 1). Most packages are categorized into 3 namespaces: the Common BioData, ClinicalData and TechnologySpecificData namespace. The remaining 6 packages are reused from the corresponding MAGE-OM packages. CaGe-OM is expressed in Class diagram of Unified Modeling Language (UML), which is a standard notation to represent the design and visualization of the system architecture.

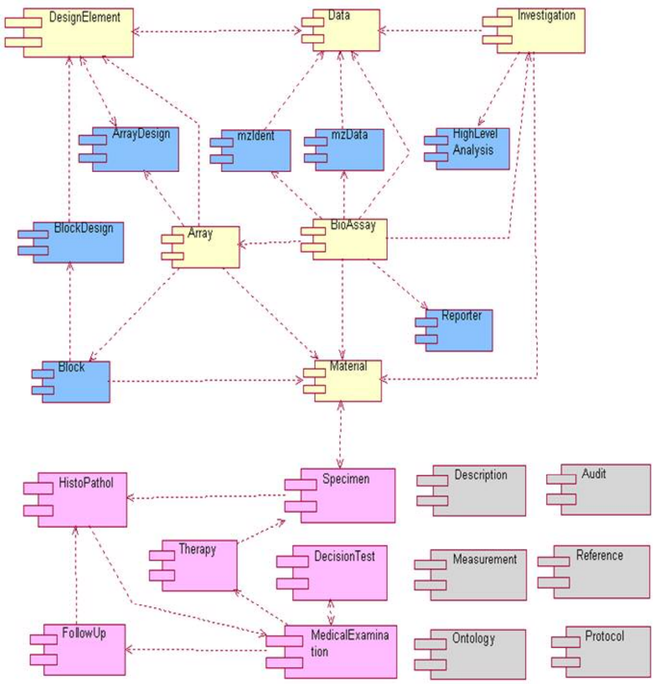

Fig. 1 – The relationships of the 25 packages in CaGe-OM. Most packages in this model are categorized into three namespaces; the CommonBioData (in yellow), ClinicalData (in pink) and TechnologySpecificData (in blue). Six packages (in gray) are adopted from MAGE-OM and remaining for general purposes.

Database design

The relational database schema of Xperanto-CaGe was derived from CaGe-OM using hybrid object-relational mapping approach. There are three fundamental object-mapping rules [12].

1) One table per an entire class hierarchy: all the attributes of all the classes in the hierarchy are stored.
2) One table per concrete class: each table includes both specific attributes of a class and any attributes it inherits.
3) One table per class including abstract super-classes: supports polymorphism and each attribute in the class inheritance tree is represented exactly once in a table.

We chose the second object-relational mapping rule because this rule provides efficient ad hoc reporting and dose not waist space; all the attributes for any single class are stored in one table. We also applied Entity-Attribute-Value (EAV) model to solve the problems associated with the storage of sparse attributes, attribute heterogeneity and flexibility in adding new attributes for a class. The EAV model, also called row modeling, stores the value of an attribute as a row with the names of its attribute in another column. The EAV model has been widely implemented in management systems for heterogeneous data sets [13, 14].

The detail mapping processes are as follow. Each table includes both the specific attributes of a class and any attributes it inherits, except for abstract classes. According to the multiplicity between classes, associations are defined as one of the normalization form. For instance, one or more (1..*) multiplicity is represented as second normalization form in relational database. Some classes belong to ClinicalData namespace, which have sparse and heterogeneous attributes, are mapped onto a table based on EAV model. Abstract classes are not captured. The associations of abstract classes are passed on to those of subclasses. Further information is available through the supplement web site (http://www.snubi.org/software/cage_om/).

Developing clinical pilot system: BioEMR

For the purpose of developing a pilot system for the evaluation of the practical utility of the integrated clinical, histopathological and high-throughput biological data in real clinical settings, we are establishing a pilot information system, named

BioEMR, with Xperanto-CaGe in the Breast Cancer Center at Seoul National University Hospital (Fig. 2).

Most of the clinical data from the legacy clinical information system can be represented in XML-based standard including HL-7, LOINC, DICOM and CDA. The biological data standards for the data from the genomics laboratory include BSML, MAGE-ML, MIAPE and TMA-OM. Both are extracted as XML files and deposited in an integrated document repository after layers of data processing. The integrated document repository is supported by clinical research and clinical trial knowledge database and analyzed by a set of analytical modules. Currently, pilot application modules are using this secondary information system, IDR, in parallel with the primary real-time hospital legacy system.

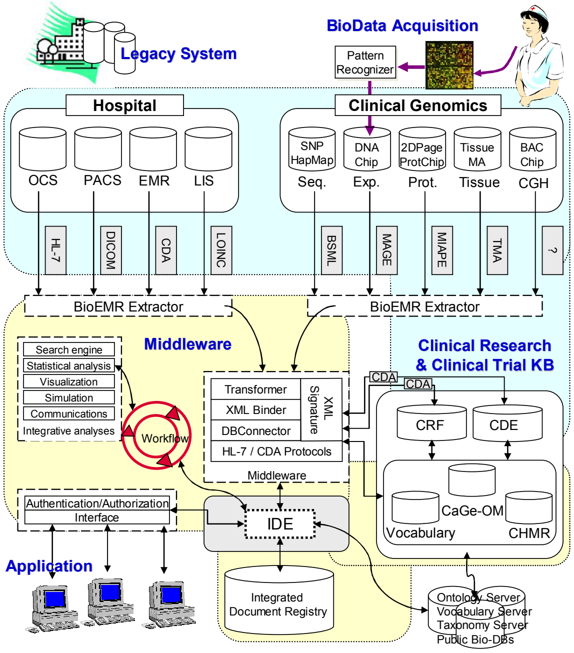

Fig. 2. BioEMR. Architecture of the pilot information system of integrated clinical, histo-pathological and genomic information.

Integrating external resources

We use the MGED Ontology for the description of common experimental procedure and array information. For describing TMA-specific and clinico-histopathological data, we use the controlled vocabulary defined in TMA-OM [7]. We also implemented an interface to add new user-defined terms.

For analyzing high-throughput functional genomic data, integration with statistical analysis tools is required. Xperanto-CaGe is linked to statistical analysis packages, BioChip Analysis and Data Integration (BioCANDI), which pipelines genomic data analysis modules implemented in R statistical language [15]. BioCANDI is composed of 15 normalizations and 54 high-level analysis protocols.

After the statistical analysis using BioCANDI, the genes with the significant expression change are represented with integrated annotation through Genome Research Informatics Pipeline (GRIP) system. The GRIP is an integrated annotation database of genes that includes genomics and proteomics as well as ontology and disease information [16].

Integration of data annotation and analysis systems help cancer researchers in biomarker discovery from multiple heterogeneous genomic datasets. Fig. 3 demonstrates the overall structure of Xperanto-CaGe.

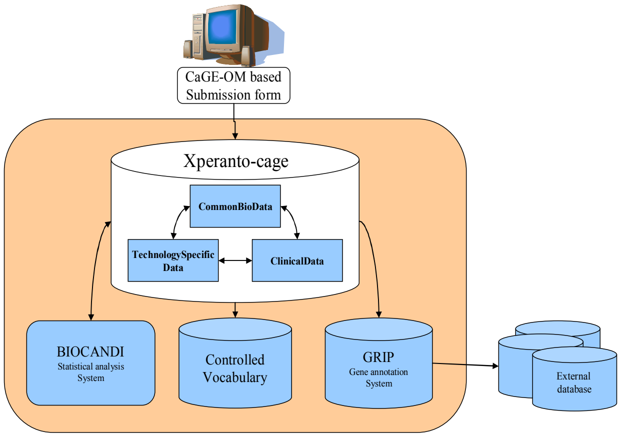

Fig. 3. System architecture of Xperanto-CaGe.

Conclusion

We developed Xperanto-CaGe based on CaGe-OM for representing and managing clinical and histo-pathological data as well as high-throughput biological experimental data covering most of the cancer types. They are developed considering the extensibility for newly emerging data types.

CaGe-OM and Xperanto-CaGe are attempts to establish a comprehensive framework for integrated storage and analysis of clinical and multiple genomic data and to facilitate model-level integration of unseen data types. The pilot system for breast cancer research (Fig. 2) using the CaGe-OM and Xperanto-CaGe described may serve as a test platform for future clinical genomics research and many translational bioinformatics applications.

Acknowledgments

This study was supported by a grant from the Ministry of Health & Welfare, Korea (A040163). Y.P.'s educational training was partly supported by a grant from Ministry of Science and Technology, Korea (2007-03992). *To whom correspondence should be addressed. Tel: +82 2 740 8320; Fax: +82 2 742 5947; Email: juhan@snu.ac.kr

References

1. Shar AR, Sinqhal M, Klicker KR, et al. Enabling high-throughput data management for systems biology the bioinformatics resource manager. Bioinformatics. 2007;23(7):906-9.
2. Spellman PT, Miller M, Stewart J, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. 2002;3(9):Research0046.1-9
3. Garwood K, McLaughlin T, Garwood C, *et al.* PEDRo: a database for storing, searching and disseminating experimental proteomics data. BMC Genomics. 2004;5(1):68.
4. Castle AL, Fiehn O, Kaddurah-Daouk R, *et al.* Metabolomics Standard Workshop and the development of international standards for reporting metabolomics experimental results. Brief Bioinform. 2006;7(2):159-65.
5. Jenkins H, Hardy N, Beckmann M, *et al.* A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol. 2006;22(12):1601-6.
6. Bino RJ, Hall RD, Fiehn O, *et al.* Potential of metabolomics as a functional genomics tool. Trends Plat Sci. 2004;9(9):418-25.
7. Lee HW, Park YR, Sim J, *et al.* The tissue microarray object model: a data model for storage, analysis and exchange of tissue microarray experimental data. Arch Pathol Lab Med. 2006; 130(7):1004-13
8. Weng C, Gennari JH, Fridsma DB. User-centered semantic harmonization: a case study. J Biomed Inform. 2007;40(3):353-64.
9. Saltz J, Oster S, Hastings S, *et al.* caGRID: design and implementation of the core architecture of the cancer biomedical informatics grid. Bioinformatics. 2006;22(15)1910-6.
10. Xirasagar S, Gustafson SF, Huang CC, *et al.* Chemical efforts in biological systems (CEBS) object model for toxicology data, SysTox-OM: design and application. Bioinformatics. 2006;22(7):874-82.
11. Park YR, Lee HW, Cho SB, *et al.* Cancer Genomics Object Model: An object model for multiple functional genomics data for cancer research. Proc MEDINFO'07, K. Kuhn *et al.* (Eds), Amsterdam, IOS Press 2007;:1235-1239.
12. Tuck D, O'Connell R, Gershkovich P, *et al.* An approach to object-relational mapping in bioscience domains. Proc AMIA Symp. 2002;:820-4.
13. Bleeker SE, Derksen-Lubsen G, van der Lei J, *et al.* Structured data entry for narrative data in a broad specialty: patient history and physical examination in pediatrics. BMC Med Inform Decis Mak. 2006;6:29.
14. Oliveira AG, Salgado NC. Design espects of a distributed clinical trials information system. Clin Trials. 2006;3(4):385-396.
15. Park JH, Park YR, Park CH, *et al.* Xperanto: a web-based integrated system for DNA microarray data management and analysis. Genomics and Informatics. 2005;3(1):39-42.
16. Genome Research Informatics Pipeline (GRIP). Available at http://grip.snubi.org/. Accessed September 18, 2007.