MEDINFO 2007

Proceedings of the 12th World Congress on Health (Medical) Informatics

Building Sustainable Health Systems

Part 1

Edited by

Klaus A. Kuhn

University Medical Center, Technische Universität München, Germany

James R. Warren

Department of Computer Science, University of Auckland, New Zealand

and

Tze-Yun Leong

School of Computing, National University of Singapore, Singapore

IOS Press

Amsterdam • Berlin • Oxford • Tokyo • Washington, DC

MEDINFO 2007
K. Kuhn et al. (Eds)
IOS Press, 2007

Cancer Genomics Object Model: An Object Model for Multiple Functional Genomics Data for Cancer Research

Yu Rang Park[1], Hye Won Lee[1], Sung Bum Cho[1], Ju Han Kim[1,2*]

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

Abstract

The development of functional genomics including transcriptomics, proteomics and metabolomics allow us to monitor a large number of key cellular pathways simultaneously. Several technology-specific data models have been introduced for the representation of functional genomics experimental data, including the MicroArray Gene Expression-Object Model (MAGE-OM), the Proteomics Experiment Data Repository (PEDRo), and the Tissue MicroArray-Object Model (TMA-OM). Despite the increasing number of cancer studies using multiple functional genomics technologies, there is still no integrated data model for multiple functional genomics experimental and clinical data. We propose an object-oriented data model for cancer genomics research, Cancer Genomics Object Model (CaGe-OM). We reference four data models: Functional Genomic-Object Model, MAGE-OM, TMA-OM and PEDRo. The clinical and histopathological information models are created by analyzing cancer management workflow and referencing the College of American Pathology Cancer Protocols and National Cancer Institute Common Data Elements. The CaGe-OM provides a comprehensive data model for integrated storage and analysis of clinical and multiple functional genomics data.

Keywords:

cancer, genomics, data model, standards

Introduction

Functional genomics includes studies of the abundance of gene transcripts by microarrays (transcriptomics), the abundance, localization and interactions of the translated proteins (proteomics), the flux in related metabolites (metabolomics), and various others [1]. For managing and representing of these functional genomics data, several technology-specific data models have been proposed, including MAGE-OM for microarray [2], PEDRo for proteomics [3], SMAR [4], ArMet [5] and MIAMET [6] for metabolomics, and TMA-OM [7] for tissue microarray.

Current researches emphasize the need to integrate data from multiple functional genomics [8, 9]. Following these trends, many cancer researches have been conducted using multiple functional genomics technologies including DNA microarray, 2DE/MS and Tissue Microarray for the understanding of global biological characteristics [3-5].

As the number of cancer studies using multiple functional genomics technologies increases, there are increasing demands for flexible solutions for systematic management of these data. Several databases have been implemented for specific functional technologies or specific purposes [12]. Despite the necessity, there is a few number of integrated data models (Table 1). It only supports a few genomics technologies or document models for genomics and clinical data. Furthermore, most approaches are designed without consideration of extendibility for integration with new biological data models.

In the present study, we designed a new data model for cancer genomics research using multiple functional genomics data, Cancer Genomics Object Model (CaGe-OM).

Table 1 - Previous approaches for integrated model

	Method	Target data	Reference model	Implementation
FGE-OM (Jones A et al., 2004)	Integrated object model	Transcriptomics, and proteomics (2DE and MS)	MAGE-OM, PEDro, Gla-PSI	RAPAD (microarray, 2DE and MS data)
SysBio-OM (Xirasagar S, et al., 2004)	Integrated object model	Transcriptomics, proteomics and metabolomics	MAGE-OM, PEDro, and a model for protein-protein interaction and metabolite	CEBS (only for microarray data)
Genotype Shared Model (HL7 clinical genomics SIG)	Document (XML)	Transcriptiomic, proteomics, sequence and clinical data	HL7 CDA	Genetic testing : BRCA Tissue typing: BMT
IBM GMS (Robson B, et al., 2004)	Document (XML)	Clinical and genomics (protein structure) data	HL7 CDA	Genomic Messing System Language (GMSL)
caCORE (Covitz PA, et al., 2003)	Object oriented API (caBIO)	Clinical and genomics data	Object Model	caBIG, CGAP, MMHCC, caArray etc..
XDesc (Shifman MA, et al., 2004)	EAV and Relational model	Clinical and genomics (Transcriptomis) data	TrialDB	YMD

Materials and methods

We used class diagram of Unified Modeling Language (UML) to represent the concepts, objects and relationships in multiple functional genomics data for cancer research. We reference four experimental data models (FuGE-OM, MAGE-OM, PEDRo and TMA-OM) and two clinical and histopathological data models (College of American Pathologist Cancer Protocols and National Cancer Institute Common Data Element) to design a data model for cancer genomics research.

Functional genomics experiment data modeling

For designing a framework to represent results from multiple functional genomics investigation, we reference four data models; the FuGE-OM for common aspects of experiments, the MAGE-OM for microarray, the PEDRo for proteomics, and the TMA-OM for tissue microarray. The FuGE-OM focused on modeling the common artifacts of functional genomics, such as sample preparation, protocols, instruments, and contact details [1]. Following the wisdom of the FuGE-OM, we reference packages associated with common aspects of functional genomics in three data models (MAGE-OM, PEDRo and TMA-OM) and modify the existing packages and classes within FuGE-OM for describing common biological experimental data which belongs to CommonBioData namespace. We extract technology-specific packages for each three data model. These extracted packages comprised in TechnologySpecific namespace.

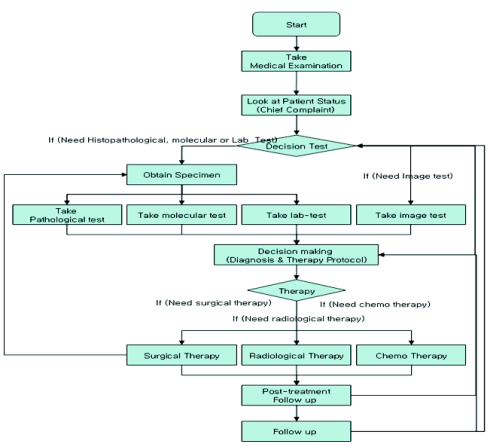

Figure 1 - Workflow diagram of clinical management of cancer. Diamonds indicate events and rectangles are physical entities.

Clinical and histopathological data modeling

For designing the clinical and histopathological data modeling, we analyzed cancer management and referenced document models of clinical and histopathological information, such as College of American Pathologists (CAP) Cancer Protocols (CPs) [13] and National Cancer Institute (NCI) Common Data Element (CDE)[14]. Figure 1 shows the workflow diagram of clinical management for cancer. To obtain comprehensive and extensible data models, we have created 6 packages (i.e., MedicalExamination, Histopathol, Specimen, DecisionTest, Therapy, and FollowUp) by systematically capturing the event and process from the workflow diagram (Figure 1) and category and value from the 43 CAP CPs and NCI CDE.

Results

Workflow of clinical management of cancer

For structured modeling of clinical data for cancer, it is required to analyze workflow of clinical management for cancer like diagnosis and therapy (Figure 1). When a patient arrives at a hospital, she/he takes a medical examination (captured by MedicalExamination class). Medical examination is an event to look at a patients status by a doctor based on physical examination (i.e. inspection, auscultation, and palpation) (PhysicalExam). Then the doctor writes down chief complaints of the patient. Then the patient takes a decision test such as clinical laboratory, images, histopathological and molecular tests (DecisionTest). The doctor makes a decision about the diagnosis and therapy protocols based on the results from various decision tests (Diagnosis & Plan). There are three types of therapy: surgical, radiological and chemotherapy (Therapy). In solid tumor, tumor specimen is obtained after surgical therapy. And then, histopathological test is taken on specimen (HistoPathol). After the therapy, post-treatment follow up is taken on the patient (FollowUp). After primary treatment, the patient is observed according to the follow up schedule (FollowUp).

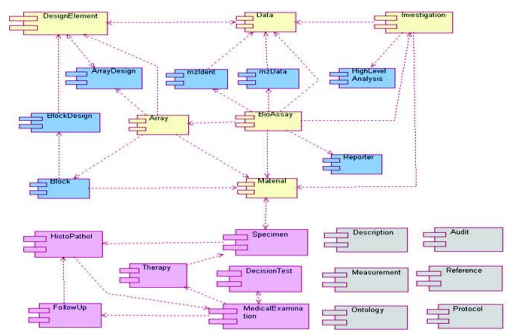

Figure 2 The relationships of 26 packages in Cancer Geno-mics Object Model (CaGe-OM). Most packages in this model are catego-rized into three namespaces; the CommonBioData (yellow-colored), ClinicalData (pink-colored) and Technolo-gySpecificData (blue-colored) namespae. Six packages (gray-colored) are remaining for general purpose.

Overview of cancer genomics object model

CaGe-OM is a data model that contains 183 classes grouped into 25 packages. Figure 2 shows the relationship of CaGe-OM packages, which grouped in three namespaces: the CommonBioData, TechnologySpecific, and ClinicalData models. The CommonBioData contains a

set of packages that describe common aspects of functional genomics from microarrays, proteomics, tissue microarray or potentially other functional genomics techniques. The packages belong to ClinicalData namespace represents clinical and histopathology data of cancer. The TechnologySpecific namespace contains the packages for describing technology-specific data. The remaining 6 packages for common annotation: Description, Measurement, Ontology, Audit, Reference, and Protocol. This model has three abstract classes at the top-level, Extendable, Describable and Identifiable that are unchanged from MAGE-OM. Most classes inherit their attributes. CommonBioData and TechnologySpecificData can refer to ClinicalData through Material packages in CommonBioData namespace.

CommonBioData namespace

The CommonBioData namespace represents common aspects of functional genomics experiments. Experimental design (captured by Investigation package), biological sample preparation (Material) and biological molecules such as DNA or protein sequences (ConceptualMolecule) are common components of all functional genomics investigation. The CommonBioData namespace consists of six packages; Investigation, Material, Array, DesignElement, Data and BioAssay.

To represent data from experiments using any type of technology, packages contained in this namespace have a generic structure. In Data packages, for instance, the Data object represents a container for a set of multidimensional data matrices and the coordinate set found in each of the dimensions.

The Material is a package for all biological and physical materials involved in an experimental workflow. For integrating genomics and clinical data, this package has a relationship with Specimen package that belongs to ClinicalData namespace. As a result, CommonBioData namespace has a reference to ClinicalData via Material and Specimen classes, representing the clinical and histopathological information of the specimen used in functional genomics experiments.

The Array, ArrayDesign, and DesignElement in the MAGE-OM contain information regarding the design, manufacturer and contents of microarrays (http://www.omg.org/docs/formal/ 03-02-03.pdf). In these packages, the ArrayDesign package is a microarray specific package. However, Array and DesignElement is commonly used in the functional genomics experiments such as microarray, tissue microarray and proteomics. Thus we are adding these two packages into CommonBioData namespace.

Clinical Data namespace

The Clinical Data namespace includes package with classes covering clinical and histopathological data 43 cancer types considered by CAP CPs, and clinical contexts from the workflow analysis of cancer management and NCI CDEs. The ClinicalData namespace is composed of six packages; MedicalExamination, Histopatho, Specimen, DecisonTest, Therapy and FollowUp.

The MedicalExamination package, the core package in ClinicalData namespace, contains classes for Demography, PhysicalExam, History, Diagnosis and Plan. Through MedicalExamination classe, all the other packages contained in ClinicalData namespace have associations with Medical-Examination package (Figure 3(a)).

The HistoPathol package provides classes describing histopathological information of specimens (Figure 3(b)). The BasicHistoPathol class stores elements that should be included regardless of the organ and tissue. The OrganSpecific class store elements for specific organs. The BasicHistoPathol class is an abstract class, the subclasses of which are the TumorInfo and Histology classes.

Classes in DecisonTest package are describing several medical tests such as image, laboratory, molecular and histopathological test. Therapy package contain classes to store data from medical therapy; surgical, radiological, and chemotherapy. Specimen package provide classes describing information of tissue obtained by surgical therapy or biopsy. The FollowUp package defines the classes for follow up data like recurrence and vital sign of patient.

TechnologySpecificData namespace

For storing technology-specific data of the experiment, the TechnologySpecificData namespace contains the packages derived from MAGE-OM, PEDRo, and TMA-OM. The TechnologySpecificData namespace is composed of eight packages; ArrayDesign, HighLevelAnalysis, Assay, mzData, mzIdent, Block, BlockDesign and Reporter.

The ArrayDesign and HighLevelAnalysis packages are microarray-specific packages. These packages are reused from corresponding MAGE-OM packages. ArrayDesign includes the manufacturing protocols, contacts, and details of the exact materials used for each feature in Array. The HighLevel-Analysis is a package for the analysis results.

The Assay, mzData and mzIdent packages, which come from the PEDRo, are proteomics-specific packages. The Assay package provides classes and attributes that contain information and annotation on the event of proteomics experiment using 2D or MS and the acquisition of images. The mzData package stores the output from mass spectrometry (MS). The mzIdent package contains the output (and input parameters) from database searches with mass spectrometry data to identify proteins or to quantify protein abundance.

The Block, BlockDesign and Reporter are tissue microarray-specific packages. These packages are identical to packages of the same name in TMA-OM. The BlockDesign package stores the intended pattern of individual block elements. The block with large number of tissues is constructed according to the BlockDesign and the block is sliced to arrays. The Block package records information on the actual events manufacturing blocks. The Reporter package contains classes for reporters used in TMA experiments. The reporter represents materials to identify a particular molecule like gene, protein, or DNA sequence.

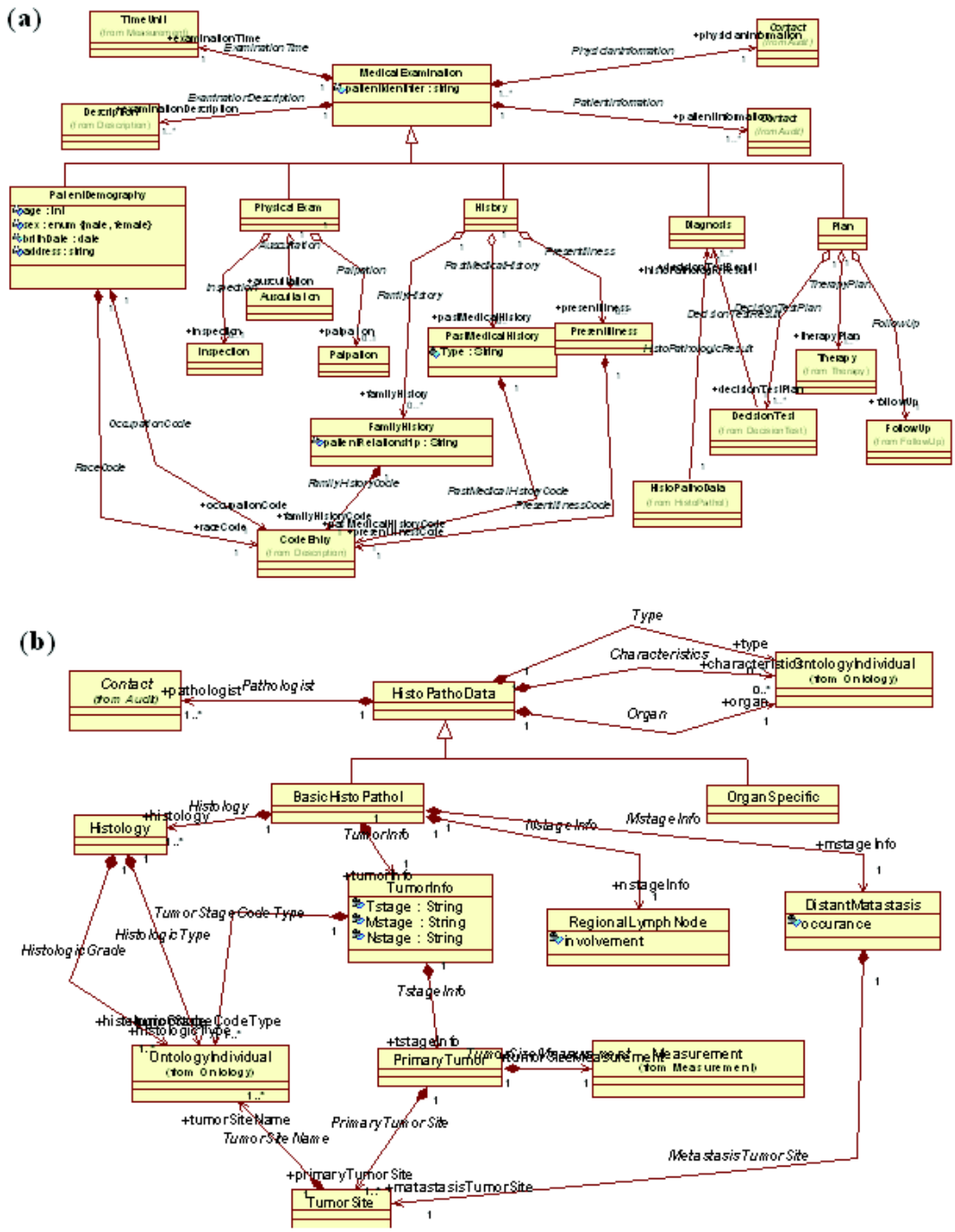

Figure 3 - Class diagram of (a) MedicalExamination and (b) Histopathol packages in ClinicalData namespace.

Discussion

We developed a data model, CaGe-OM, to store and integrate data generated from microarray, proteomics and tissue microarray experiments performed on the same biological samples. The CaGe-OM can represent clinical and histopathological information as multiple functional genomics data for any type of cancer. This integrated data model allows the combined analysis of multiple functional genomics data for understanding of the underlying biological nature in a systematic fashion.

The CaGe-OM can integrate easily a new biological data model without significant difficulty by representing common aspects of the new models as CommonBioData and technology-specific parts as TechnologySpecificData separately, while it is hard to modify the models in previous studies to condider and integrate a new model (Table 1). Because the CaGe-OM is independent of implementation, several applications based on this model such as relational database schema, web application and XML document can be constructed.

The development of an integrated data model for cancer genomics researches may facilitate tight integration of technology-specific data models and clinical data models. As functional genomics are increasingly used in cancer research, the CaGe-OM will be useful for the structured data management of clinical data and for the analysis of functional genomics data combined with clinical data.

Acknowledgments

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea (0412-MI01-0416-0002).

References

[1] Jones AR, Pizarro A, Spellman P, et al. FuGE: Functional Genomics Experiment Object Model. OMICS. Jun 2006; 10(2): 179-84

[2] Spellman PT, Miller M, Stewart J, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol. Aug 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. Sep 17 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. Jun 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. Dec 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. Sep 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. Jan 25 2006; 130(7):1004-13

[8] Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet. 2001;2:343-72.

[9] Waters M, Boorman G, Bushel P, et al. Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. EHP Toxicogenomics. Jan 2003;111(1T):15-28.

[10] Callagy G, Pharoah P, Chin SF, et al. Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays. J Pathol. Feb 2005;205(3):388-396.

[11] Ippolito JE, Xu J, Jain S, et al. An integrated functional genomics and metabolomics approach for defining poor prognosis in human neuroendocrine cancers. Proc Natl Acad Sci U S A. Jul 12 2005;102(28):9901-6.

[12] Elfilali A, Lair S, Verbeke C, et al. ITTACA : a new database for integrated tumor transcriptome array and clinical data analysis. Nucleic Acids Res. Jan 1 2006;34(Database issue):D613-6.

[13] Warzel DB, Andonaydis C, McCurry B, Chilukuri R, Ishmukhamedov S, Covitz P. Common data element (CDE) management and deployment in clinical trials. AMIA Annu Symp Proc. 2003:1048.

[14] The College of American Pathologists. Cancer protocols in January 2006 revision version. Available at: http://www.cap.org/apps/docs/cancerprotocols/protocolsindex.html. Accessed November 8, 2006.

Address for correspondence

Ju Han Kim, MD, PhD,
Seoul National University College of Medicine,
28 Yongon-dong, Chongno-gu, Seoul 110-799, Korea.
Tel: +82-2-740-8320; Fax:+82-2-742-5947
e-mail: juhan@snu.ac.kr