The Tissue Microarray Object Model

A Data Model for Storage, Analysis, and Exchange of Tissue Microarray Experimental Data

Hye Won Lee, MS; Yu Rang Park, MS; Jaehyun Sim, MS; Rae Woong Park, MD, PhD; Woo Ho Kim, MD, PhD; Ju Han Kim, MD, PhD

• *Context.*—Tissue microarray (TMA) is an array-based technology allowing the examination of hundreds of tissue samples on a single slide. To handle, exchange, and disseminate TMA data, we need standard representations of the methods used, of the data generated, and of the clinical and histopathologic information related to TMA data analysis.

***Objective.*—To create a comprehensive data model with flexibility that supports diverse experimental designs and with expressivity and extensibility that enables an adequate and comprehensive description of new clinical and histopathologic data elements.**

***Design.*—We designed a tissue microarray object model (TMA-OM). Both the array information and the experimental procedure models are created by referring to the microarray gene expression object model, minimum information specification for in situ hybridization and immunohistochemistry experiments, and the TMA data exchange specifications. The clinical and histopathologic information model is created by using College of American Pathologists cancer protocols and National Cancer Institute common data elements. Microarray Gene Expression Data Ontology, the Unified Medical Language System, and the terms extracted from College of American Pathologists cancer protocols and NCI common data elements are used to create a controlled vocabulary for unambiguous annotation.**

***Result.*—The TMA-OM consists of 111 classes in 17 packages to represent clinical and histopathologic information as well as experimental data for any type of cancer. We implemented a Web-based application for TMA-OM, supporting data export in XML format conforming to the TMA data exchange specifications or the document type definition derived from TMA-OM.**

***Conclusions.*—The TMA-OM provides a comprehensive data model for storage, analysis, and exchange of TMA data and facilitates model-level integration of other biological models.**

(*Arch Pathol Lab Med.* 2006;130:1004–1013)

DNA microarray and proteomics surveys allow researchers to analyze expression levels of thousands of genes and proteins at once. The development of these high-throughput technologies has fundamentally affected biomedical research. Large-scale industrial efforts have been increased to apply genomics and proteomics for the identification of markers for new diagnostics and therapeutics.[1]

The concept of DNA microarrays was extended to pathologic research based on embedded tissue samples. Tissue microarray (TMA) technology is an array-based, high-throughput technology used to examine molecular alterations in a large number of tissues on a single slide in parallel.[2] Because TMA experiments are performed in parallel, TMA technologies have some distinct advantages over traditional research methods with the whole sections. Tissue microarry yields high-throughput data in a cost-effective manner and provides internally consistent staining conditions and allows many potential biomarkers to be assessed for the same case series. These advantages allow many researchers to examine marker genes in cancer studies.[3]

As the number of cancer studies using high-throughput technologies increases, TMA technology has been proven to be a high-throughput validation tool of the marker genes identified in DNA microarray experiments.[4] When validating candidate genes in other studies, such as serial analysis gene expression[5] and array comparative genomics hybridization,[6] TMA technology has been also used.[7,8]

Despite several advantages, many laboratories have difficulties in studying the results from TMA experiments. First, a single TMA experiment generates a vast quantity of data, resulting in difficulties in data collection, storage, and interpretation. Second, although TMA experiments with different cancer specimens may require clinical and histopathologic information, a generalized data model to

Accepted for publication January 25, 2006.

From the Seoul National University Biomedical Informatics (Ms Lee, Ms Y. R. Park, Mr Sim, and Dr J. H. Kim), Human Genome Research Institute (Dr J. H. Kim), and Department of Pathology, Seoul National University College of Medicine (Dr W. H. Kim), Seoul, Korea; and the Department of Medical Informatics, Ajou University School of Medicine, Suwon, Korea (Dr R. W. Park).

The authors have no relevant financial interest in the products or companies described in this article.

Presented in abstract form at the 10th Annual Meeting of Advancing Practice, Instruction and Innovation through Informatics, Lake Tahoe, Calif, August 2005.

Reprints: Ju Han Kim, MD, PhD, Seoul National University College of Medicine, 28 Yongon-dong, Chongno-gu, Seoul 110-799, Korea (e-mail: juhan@snu.ac.kr).

Comparison of Related Studies With This Study*

Source, y	Data Model	Generalized Clinical Information Model	Generalized Histopathologic Information Model	No. of Cancers Supported	Extensibility for Other Cancers	Integration With MAGE-OM	Controlled Vocabulary	Data Export in XML Format
Manley et al,[17] 2001	ER	Not included	Not included	1 (prostate cancer)	No	Not easy	Not considered	Not considered
Shaknovich et al,[18] 2003	ER	Not included	Not included	1 (large cell lymphoma)	No	Not easy	Not considered	Not considered
Demichelis et al,[19] 2004	ER	Not included	Not included	2 (breast and lung cancers)	No	Not easy	Not considered	Not considered
Berman et al,[20] 2004	XML Database	Not included	Not included	1 (prostate cancer)	No	Not easy	CPCTR CDEs and TMA CDEs	Enabled
Xperanto-TMA	Object	Model based on CAP cancer protocols and NCI CDEs	Model based on CAP cancer protocols and NCI CDEs	43 (and extensible)	Yes	Easy	MGED Ontology, UMLS, and terms extracted from CAP cancer protocols and NCI CDEs	Enabled

* MAGE-OM indicates microarray gene expression object model; ER, entity relationship; CPCTR, cooperative prostate cancer tissue resource; CDEs, common data elements; TMA, tissue microarray; CAP, College of American Pathologists; NCI, National Cancer Institute; MGED, microarray gene expression data; and UMLS, Unified Medical Language System.

manage the vastly different sets of clinical and histopathologic information is lacking. Third, different laboratories may use different experimental protocols and instruments and capture data using different data elements, formats, and structures. Thus, it is difficult to consistently combine the findings from different laboratories, even if they use the same TMA block. Fourth, it is not easy to integrate TMA experimental results with other biological data such as DNA microarray and array comparative genomics hybridization data. If it were possible to integrate findings from other studies, these laboratories could greatly increase the value of their experimental findings.[9]

These difficulties are not unique to TMA experiments. In DNA microarray, there have been efforts by the Microarray Gene Expression Data (MGED) group to solve the difficulties by developing three standards. (1) Minimum information about microarray experiment (MIAME) describes a specification for the minimum information that is needed to enable the interpretation of the experiment.[10] (2) The microarray gene expression object model (MAGE-OM) is the data model for gene expression data.[11] (3) Microarray gene expression markup language (MAGE-ML) defines an XML format for gene expression data exchange.[11]

The ArrayExpress database has been developed on the basis of MAGE-OM and fully supports MIAME.[12] The Gene Expression Omnibus project, which is based on a simpler data model, simple omnibus format in text, has also adopted MIAME.[13]

In the field of proteomics, a set of standards is being developed along MIAME lines. The Proteomics Standards Initiative (PSI) of the Human Proteome Organization has developed (1) the minimum information about a proteomics experiment (as a guide to minimum reporting requirements, (2) PSI-OM as proteomics data object model, and (3) PSI-ML as markup language. The proteomics experiment data repository model and markup language were created and implemented as a database.[14,15] Jones et al[16] proposed the functional genomics experiment object model in an effort to integrate MAGE-OM, the proteomics experiment data repository, and the Glasgow proposal for the PSI.

Relational database implementations for specific cancers (including prostate cancer,[17] large cell lymphoma,[18] and breast and lung cancers[19]) have been reported without necessarily considering the standardization effort (Table). The Association of Pathology Informatics proposed an open access TMA data exchange specification (TMA DES), which is a well-formed XML document with 4 required sections, 80 common data elements (TMA CDEs), and 6 semantic rules.[9] A TMA database was created by transforming the Microsoft (Microsoft Corporation, Redmond, Calif) Excel-based central database of Cooperative Prostate Cancer Tissue Resource into XML files conforming to TMA DES.[20] Nohle and Ayers[21] evaluated the Association of Pathology Informatics specification of TMA DES by developing a document type definition (DTD) defining the 80 CDEs and validating the exported XML files from AIDS (Acquired Immunodeficiency Syndrome) and Cancer Specimen Resource TMA data.

The greatest value of all the -omics data must be in providing us with integrated views of microarray, proteomics, metabolomics, and any other data. Given that the data models for microarray and proteomics have been largely based on object modeling technology (ie, MAGE-OM and PSI-OM), and the utility of TMA technology has been a high-throughput clinical validation tool of the marker genes identified in DNA microarray experiments,[4] development of an object model for TMA data may facilitate tight integration of all -omics data models and consequently improve combined interpretation.

Object model is a conceptual representation of objects with attributes and functions, and associations between objects. Object model provides high expressivity and flexibility with easy maintenance. Most importantly, an object model for TMA may enable the integration with other biological data models such as MAGE-OM and PSI-OM.

The MAGE-ML is created with reference to the automatically generated DTD from the unified modeling language description of MAGE-OM, as is PSI-ML to the DTD from that of PSI-OM. Although TMA DES provides a well-formed XML document with meta data tags,[9] and a DTD development scheme for the CDEs has been provided,[21] an object model for TMA is not yet available. We have developed the TMA-OM. Model-level integration with other data models may provide a common frame of reference for ''omics'' studies. Following the wisdom of MAGE lines, the DTD that is used for the creation of TMA markup language can be automatically generated from TMA-OM. Consistent development and integration of data models for the clinical and histopathologic information as well as controlled vocabularies and ontologies may benefit systematic investigations of the fundamental clinicopathologic processes we are studying.

We propose that a data model for TMA should have the following characteristics. First, it is essential for a TMA data model to have sufficient expressivity to describe the diverse information and data concerning TMA experiments, including clinical and histopathologic information. Second, it needs to have flexibility to support diverse designs of TMA experiments. Third, it should use controlled vocabularies and conform to standards and standard protocols for common understanding among users. Fourth, integration with other biological data models (eg, MAGE-OM and PSI-OM) should be considered. Finally, the data model should have extensibility that permits us to adequately describe new clinical and histopathologic data elements that are not yet predefined at the time of use. In other words, it must be possible to describe newly proposed CDEs and cancer protocols without necessarily changing the data model or database implementations. By eliminating the need for predefining all data elements that may be used in the future, the extensibility can make it possible to incrementally develop and maintain a TMA information system.

We developed a data-centric model, called TMA-OM, using unified modeling language. The TMA-OM consists of three models for (1) array information, (2) experimental procedure, and (3) clinical and histopathologic information. Both the array information model and the experimental procedure model are created by referring to three external resources (ie, MAGE-OM, minimum information specification for in situ hybridization and immunohistochemistry experiments [MISFISHIE], and TMA CDEs). The clinical and histopathologic information model is created by referring to the 43 College of American Pathologists (CAP) cancer protocols and the National Cancer Institute Common Data Elements (NCI CDEs) (Figure 1). The use of controlled vocabularies is essential for unambiguous representation of TMA experiments. We developed a controlled vocabulary using MGED Ontology, Unified Medical Language System (National Library of Medicine, National Institutes of Health, Besthesda, Md), and the data elements in CAP cancer protocols and NCI CDEs.

Although the experimental procedure for TMA is similar to that of DNA microarray, there are limits to describe TMA data by MAGE-OM. DNA microarray contains thousands of probes to measure the gene expression levels in a specimen having the clinical and histopathologic information for the *single* specimen, and TMA contains hundreds of tissues having the clinical and histopathologic information for the *population* (Figure 2). The fundamental difference in the array fabrication methods makes it impossible to reuse MAGE data model for describing TMA experiments. Moreover, MAGE-OM does not support clinical and histopathologic information, for which we had to design data models (namely the ClinInfo and HistoPathol packages, respectively) for the development of TMA-OM.

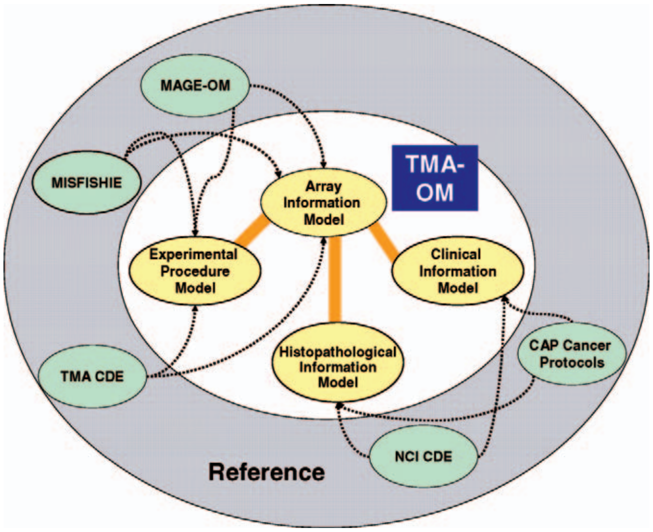
MAGE-OM
MISFISHIE
Array Information Model
TMA-OM
Experimental Procedure Model
Clinical Information Model
TMA CDE
Histopathological Information Model
CAP Cancer Protocols
NCI CDE
Reference

Figure 1. *Architecture of tissue microarray object model (TMA-OM). Array information and experimental procedure models are created by referring to microarray gene expression object model (MAGE-OM), TMA common data elements (CDEs), and minimum information specification for in situ hybridization and immunohistochemistry experiments (MISFISHIE). Clinical and histopathologic information models are created using College of American Pathologists (CAP) cancer protocols and National Cancer Institute (NCI) CDEs.*

A web-based database application called Xperanto-TMA was built to implement the TMA-OM. Exporting data both into an XML format that conforms to the TMA DES and into another format that conforms to the DTD automatically generated from the TMA-OM is a function supported by Xperanto-TMA. We designed a relational schema according to object-to-relational mapping rules in contrast to the previous studies that used simple relational modeling technology (Table). Applying systematic object-to-relational mapping technology, although not without effort, supports the plausible characteristics of the TMA data model described here, including the support for XML interfaces.

MATERIALS AND METHODS

Development of MIAME and MAGE-OM

DNA microarray data must be presented and exchanged to promote the sharing of well-annotated data within the life sciences community. To meet that need, the MGED group developed standards for microarray data annotation and exchange: MIAME, guidelines for describing a microarray experiment, MAGE-OM as a data model, and MAGE-ML as an XML-based markup language directly derived from MAGE-OM.

The MGED group has developed MAGE-OM as an object model for describing experiments performed on all types of DNA microarrays, including spotted and synthesized arrays, and oligonucleotide and cDNA arrays. The MAGE-OM is a standard data model representing microarray gene expression data by 132 classes contained in 17 packages.[11,22] Packages of MAGE-OM are generic process templates that can involve all types of experiments with DNA microarrays. Generic templates make the model reusable in other technologies such as proteomics.[15] The MAGE-

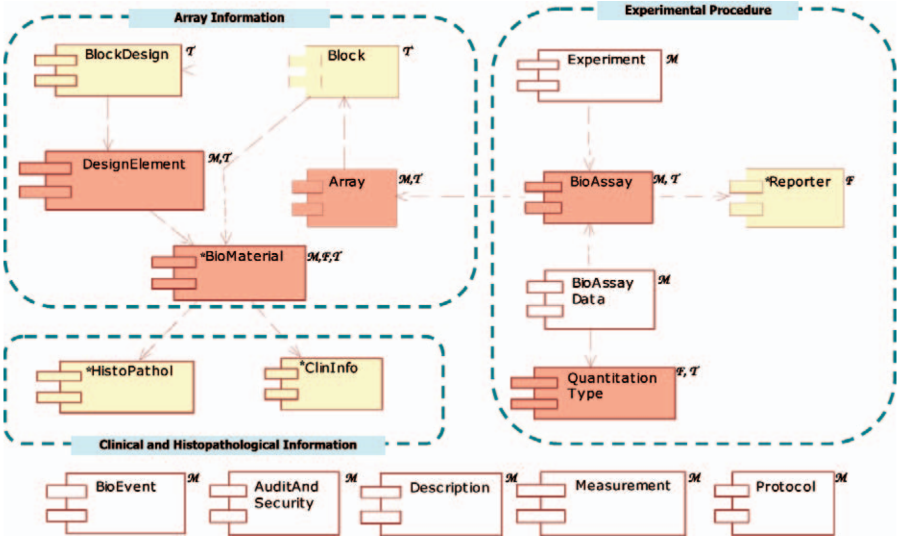

Figure 2. *The relationships of 17 packages in tissue microarray object model (TMA-OM). Twelve packages are categorized to 3 groups: array information, experimental procedure, and clinical and histopathologic information. Dashed arrows depict the reference relationships. The remaining 5 packages have relationships with most other packages such that the arrows are omitted. White-colored packages are created by reusing the corresponding packages from microarray gene expression object model (MAGE-OM) and red-colored ones by modifying those of MAGE-OM. Yellow-colored packages are newly created to represent TMA-specific components including clinical and histopathologic information (see "Materials and Methods"). Labels are according to reference resources (***M,** MAGE-OM; **F,** MISFISHIE [minimum information specification for in situ hybridization and immunohistochemistry experiments]; and **T,** TMA DES CDEs [TMA data exchange specifications common data elements]). *See unified modeling language class diagrams in Figures 4 and 5.*

OM has a modular structure robustly designed for the evolution of array technology.

The OntologyEntry Class of MAGE-OM and MGED Ontology

The OntologyEntry class of MAGE-OM provides the model with flexibility by eliminating the need for predefining all possible data elements and allowing one to refer MGED Ontology terms or to contain user-defined terms.

The MGED Ontology defines sets of common terms and annotation rules for microarray experiments, enabling unambiguous annotation and efficient queries, data analysis, and data exchange without loss of meaning.[23] The MGED Ontology contains the OntologyEntry class itself to refer to an external source.[24]

Association of Pathology Informatics TMA Data Exchange Specification

The Association of Pathology Informatics proposed an open access TMA DES, which is a well-formed XML document with 4 required sections (header, block, slide, and core) and 80 TMA CDEs, which are well-defined XML meta data tags that can be used to consistently describe data in different XML files. A set of 6 semantic rules describes the complete data exchange specification.[9] Users are allowed to add their own tags. It was demonstrated that a TMA database of XML files conforming to TMA DES can be successfully created for prostate cancer TMA data.[20] A DTD defining the 80 TMA DES CDEs was implemented as an external file that can be supplemented by internal DTD extensions for locally defined data elements.[21]

MISFISHIE Standard Working Group

Minimum information specification for in situ hybridization and immunohistochemistry experiments has been developed in the MISFISHIE Standard Working Group at MGED. The MISFISHIE specification describes what information should be provided for the representation of immunohistochemistry and in situ hybridization data. Because the MISFISHIE specification covers the most important types of TMA experiments yielding core results, it may serve as a valuable resource for the development of a data model for TMA experiments. We used the MISFISHIE version that was revised on December 19, 2004.[25]

CAP Cancer Protocols

The CAP has designed protocols for the communication of pathologic information from cancerous specimens. The CAP cancer protocols are the standard of surgical pathology reporting and have been revised annually.[26] These protocols specify the information the referring physician needs to select primary or adjuvant treatment, assess prognosis, and analyze outcome. The protocols provide common reporting formats to help tumor registrars and others collect pathologic data in a uniform manner. All protocols are divided into clinical information, macroscopic examination, and microscopic evaluation. Each protocol is stratified according to the procedure used to obtain specimens.[27]

The 43 CAP cancer protocols that cover clinical and histopathologic information for most major cancers provide rich resources for developing comprehensive data models. Because they provide a standardized way of describing clinical and histopathologic information for major cancers, the protocols can be used as a common framework to develop an extensible data model, enabling one to represent new (ie, previously unseen) clinical and histopathologic data elements without modifying the data model or database implementation.

The National Cancer Institute Common Data Elements

The NCI CDEs are data elements that are collected and stored uniformly across institutions and studies funded by NCI.[28] The NCI CDEs are defined in a data dictionary that contains at a minimum the item name, the way the item is collected, valid values, coding, and data type.[29] The NCI CDEs provide rich resources with detailed descriptions of clinical and histopathologic information for cancer research and clinical trials. The NCI CDEs for cancer clinical trials are defined, and CDEs for other fields are in progress. We used clinical-related and histopathologic-related data elements from NCI CDES, and excluded clinical trials-related ones. By providing detailed specification and a broader coverage of data elements, NCI CDEs complement the 80 TMA DES CDEs[9] in developing clinical and histopathologic data models for TMA-OM.

Development of TMA Object Model

We used class diagrams of unified modeling language to represent the concepts, objects, and relationships in TMA experiments. Unified modeling language is a standard notation to represent the design and visualization of the architecture of a system during development. Class diagrams give an overview of a system by showing its classes and the relationships between them. Figure 2 demonstrates the relationship of the 17 TMA-OM packages, which are grouped in 3 categories (the array information model, the experimental procedure model, the clinical and histopathologic information model), and the remaining 5 basic packages: BioEvent, Protocol, Description, Measurement, and AuditAndTrail.

Array Information Model and Experimental Procedure Model.—Because the TMA is an array-based technology, array manufacturing and experimental procedures for the TMA are similar to those of DNA microarray. To develop Experiment and BioAssayData packages, the corresponding packages of MAGE-OM are reused. By referring to TMA DES CDEs and MISFISHIE, we modified the corresponding packages from MAGE-OM to develop Array, DesignElement, BioMaterial, BioAssay, and QuantitationType packages for the TMA-OM. Block, BlockDesign, and Reporter packages are created by referring to TMA DES CDEs and MISFISHIE.

The experimental procedure model refers to the array information model through the reference from BioAssay to Array packages, representing the array platform and the applied bioassay. Array information model refers to the clinical and histopathologic information model via the Block and BioMaterial packages, representing the clinical and histopathologic information of the tissues on the array.

We referred to TMA DES CDEs to obtain experimental and array manufacturing processes, and MISFISHIE to obtain results from a bioassay such as immunohistochemistry and in situ hybridization.

In TMA DES CDEs, the CDEs in the block section are referred to for creating Array, Block, BlockDesign, and DesignElement packages, and the CDEs in the slide section are used to create BioAssay package. BioMaterial and QuantitationType packages are created by referring to both MISFISHIE and the CDEs in the core section of TMA DES, and Reporter packages by referring to MISFISHIE (Figure 2).

Clinical and Histopathologic Information Model.—Clinical and histopathologic information is not supported by MAGE-OM. The CAP cancer protocols and NCI CDEs provide rich resources of clinical and histopathologic information. To obtain comprehensive and extensible data models, we created ClinInfo and HistoPathol packages by systematically capturing the categories and valid values of the common and organ-specific data elements from the 43 CAP cancer protocols, which are stratified according to the procedures used to obtain specimens, and the NCI CDEs (Figure 1). The data elements were structured with abstraction and hierarchy to create object models.

Those data elements of CAP cancer protocols and NCI CDEs that are common to all cancer types are incorporated in ClinInfo and HistoPathol packages as classes or attributes. Although ClinInfo and HistoPathol packages provide essential data elements for clinical and histopathologic information, it is impractical to predefine all the clinical and histopathologic data elements. The remaining cancer-type specific elements of CAP cancer protocols and NCI CDEs are represented as the corresponding entities in the OntologyEntry class. The OntologyEntry class of TMA-OM references MGED Ontology terms for those concepts already defined in MGED Ontology.[24] Furthermore, for those concepts that cannot be readily defined, TMA-OM allows users to create user-defined terms. The attributes (ie, category, value, and description) of user-defined terms are registered in the OntologyEntry class of TMA-OM, by which one can represent a new concept by referring the OntologyEntry class.

Development of Controlled Vocabulary for TMA Object Model

We need to develop a controlled vocabulary to specify the terms that describe TMA experiments and clinical and histopathologic information. For the description of experimental procedure and array information, MGED Ontology terms are included in the controlled vocabulary. The terms extracted from TMA DES CDEs and MISFISHIE are used. For describing clinical and histopathologic information, terms are extracted from CAP cancer protocols, NCI CDEs, and anatomy-related terms in the Unified Medical Language System.[30]

RESULTS

Workflow Analysis of TMA Experiments

TMA-OM Models TMA Data.—To develop TMA-OM, we first analyzed the TMA experimental procedure to obtain a workflow diagram, a conceptualized model of the biological workflow (Figure 3). This figure demonstrates how TMA-OM captures TMA data through TMA manufacturing (Figure 3, a) and the creation and manipulation of BioAssay (Figure 3, b).

The specimen (captured by the Specimen class) in a tissue block is treated to create small core biopsies (Core) by a treatment (Treatment) according to a Treatment Protocol *method*. The core is transferred into defined array coordinates in a recipient block (Block). The transferring process is repeated. A completed block is sliced to arrays (Array). Tens of identical arrays can be made from a block. A bioassay such as immunohistochemistry and in situ hybridization (PhysicalBioAssay) is created by the hybridization event (Hybridization) of the array (Array) and antibodies or probes (Reporter). After hybridization, a series of treatments are applied. The final form of PhysicalBioAssay is obtained by an ImageAquisition event. The images are analyzed by pathologists with macroscopic or microscopic methods, resulting in a MeasuredBioAssay. The image analysis procedure is captured by the FeatureExtraction class. The actual image is stored in the Image class and the extracted features in the MeasuredBioAssayData class. DerivedBioAssayData can be obtained by mapping and transforming MeasuredBioAssayData (Figure 3.).

The Object Model: TMA-OM

The TMA-OM contains 111 classes in 17 packages. Most packages are categorized into 3 models: the array information, experimental procedure, clinical and histopathologic information models (Figure 2). Because the remaining 5 packages are reused from corresponding MAGE-OM packages, specifications for them have been already described.[31] The TMA experimental information can refer to clinical and histopathologic information through array information. A bioassay such as immunohistochemistry or fluorescent in situ hybridization, for example, is performed by joining reporters (eg, antibodies, probes) with an array, each core of which has links to clinical and histopathologic information.

Figures 4 and 5 depict 4 among 17 packages and their elements to illustrate the representation of concepts and relationships in biological materials (BioMaterial), reporters (Reporter), histopathologic information (HistoPathol), and clinical information (ClinInfo).

Array Information Model

Information regarding the design, manufacture, and contents of an array is contained in 5 packages in TMA-OM: DesignElement, BlockDesign, Block, Array, and BioMaterial (Figure 3). The DesignElement package allows users to specify information about the biological materials (in the BioMaterial package) deposited on an array. The BlockDesign package stores the intended pattern of individual block elements. The Block and the Array packages record information on the actual events manufacturing blocks and arrays. According to BlockDesign, a block with a large number of tissues is constructed and the block is sliced into arrays. All features of a block are equal to those of arrays that were made from the block. An array has the

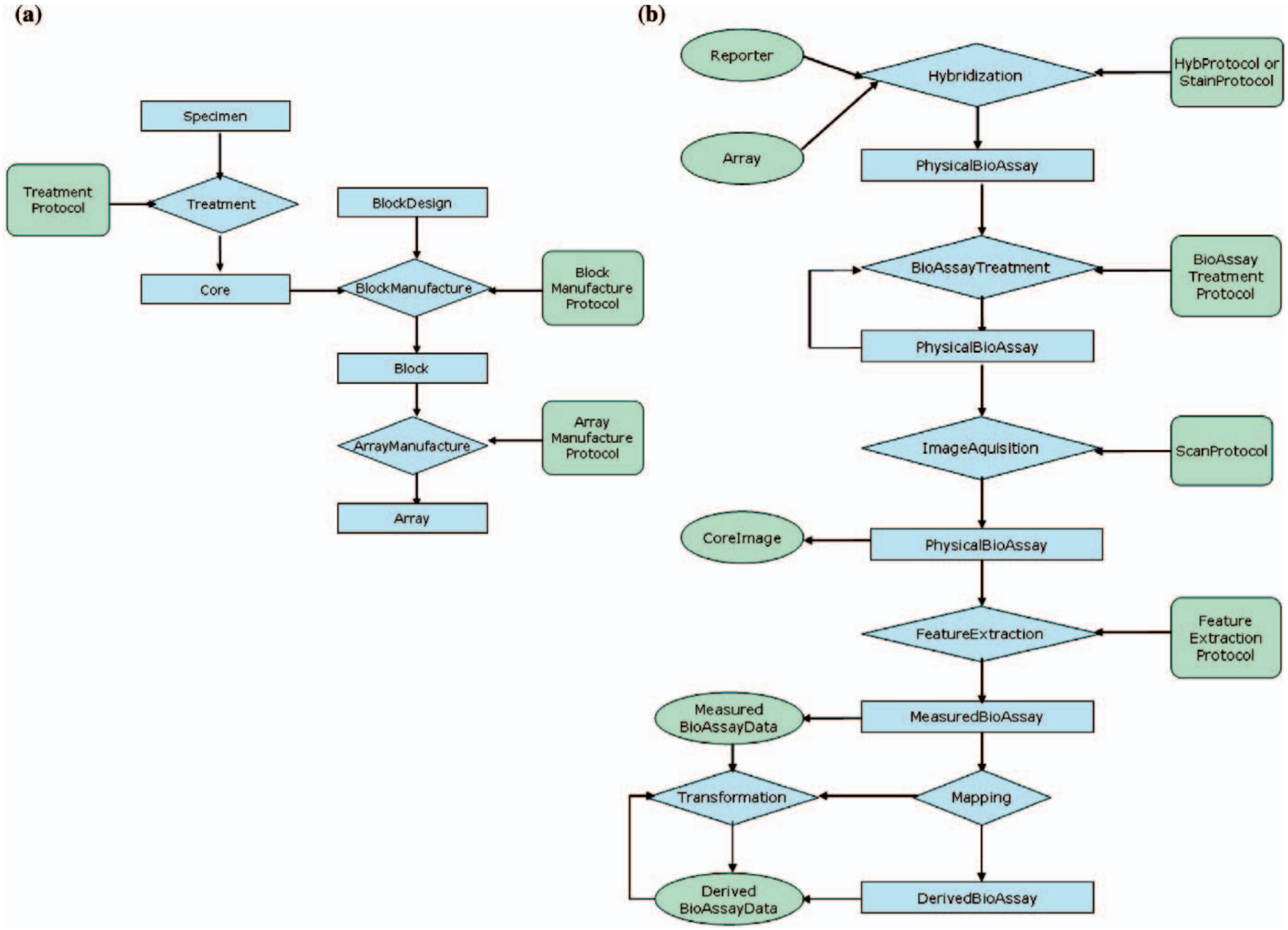

Figure 3. *Workflow diagram showing the procedure for tissue microarray (TMA) manufacturing, experiments, and data analysis. a, Workflow diagram for TMA manufacturing. b, Workflow diagram for the creation and manipulation of BioAssay. An array and one or more reporter(s) are joined by hybridization to create a PhysicalBioAssay. Rectangles indicate physical things; diamonds, events; ovals, data; and rounded rectangles, methods.*

same features as the corresponding block used to manufacture the arrays.

The classes of the DesignElement package describe what is intended to be at each location in the Block through the Feature class. The Feature class describes an intended location in the Block. The features have associations to the BioMaterial class.

Classes in the BlockDesign packages describe a design for blocks that are constructed with a set of cores. A BlockDesign consists of several features in which cores are placed. BlockDesign allows a user to specify the layout of features and the protocols used.

The Block package stores information about blocks that were constructed on the basis of a BlockDesign. This includes the manufacturing protocols, contacts, numbers of arrays from the block, and details of the biological material (ie, a core) used for each feature. Information for positional changes and other feature defects can be recorded for each block.

Classes in the Array package contain information and annotation on arrays that are created from a block. This includes the manufacturing protocols and relevant contacts.

Biological samples in an experiment are termed *biomaterials* and are a part of the BioMaterial package (Figure 4, a) that describe the process of treatment of a specimen to obtain a core. The Specimen class has association to the ClinInfo and HistoPathol classes. Each feature of DNA microarray references the BioSequence class because each feature has probe sequence information, but that of the TMA references the BioMaterial class, which in turn references the ClinInfo and HistoPathol classes through the Specimen class (Figure 4, a).

Experimental Procedure Model

The experimental procedure model is composed of 5 packages: Experiment, Reporter, BioAssay, BioAssayData, and QuantitationType.

Experiment represents a collection of results from one or more BioAssay(s). Experiment, through its association with ExperimentDesign, contains records of information including replicate, the type of experiment, and a set of the parameters of the experiment. Experimental factors, through the association of FactorValues, are represented by the ExperimentFactors. Further information is available through the supplementary Web address.[32]

Materials used to identify specific molecules are termed *reporters* and are part of the Reporter package (Figure 4,

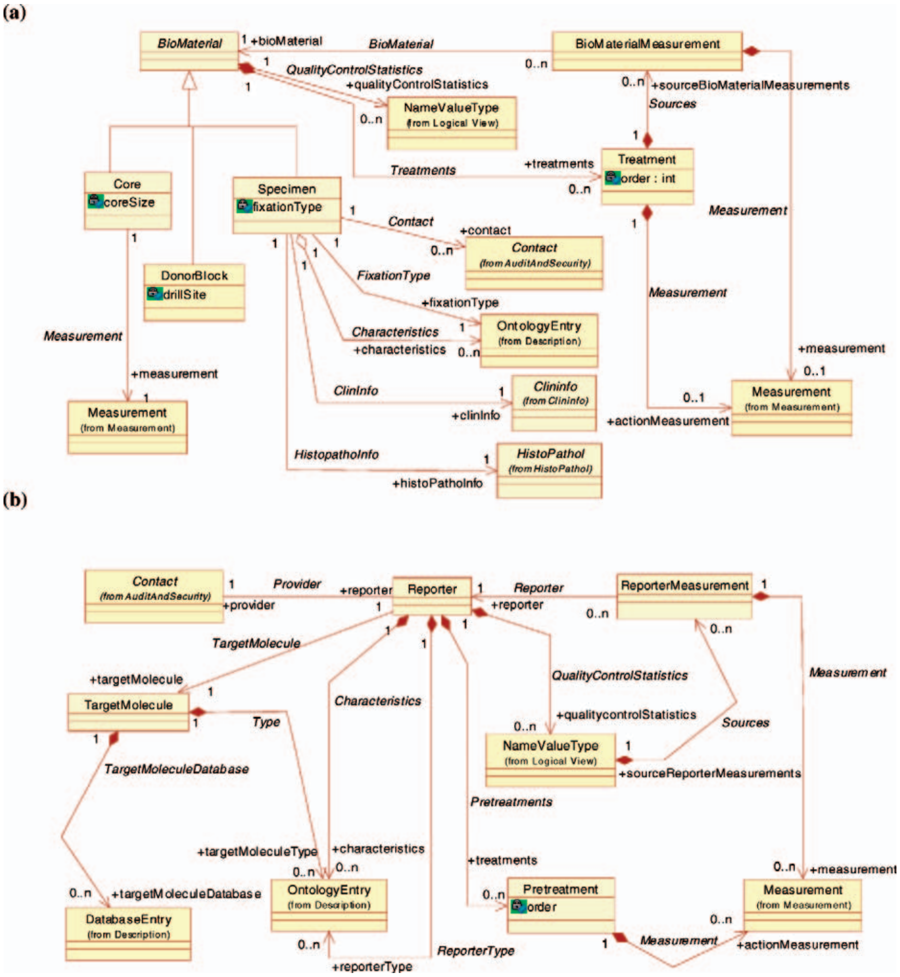

Figure 4. *Unified modeling language models of the BioMaterial and Reporter packages. a, The classes in the BioMaterial package describe the characteristics and the biological material (ie, specimen) treatment processes. The Specimen class references the ClinInfo and HistoPathol classes. b, The classes in the Reporter packages represent experimental-staining materials, termed* reporters, *identifying specific molecules (TargetMolecule), such as proteins.*

b). This package describes the characteristics and the treatment procedure of Reporters. A reporter identifies a particular molecule, such as a gene, a protein, or a DNA sequence, called the TargetMolecule. The TargetMolecule class has an association with the DatabaseEntry class to reference an individual record in a database such as GenBank and SwissProt.

The BioAssay package provides classes that contain information and annotation on the event of joining an array with one or more reporter(s), the acquisition of images, and the extraction of data for an image per feature. BioAssay is an abstract class and has 3 derived classes: the PhysicalBioAssay, which leads to the production of Images, the MeasuredBioAssay, which is associated with the set of quantitation produced by FeatureExtraction, and the DerivedBioAssay, which groups BioAssays that have been analyzed together to produce further refinement of the quantitations.

The classes defined in the BioAssayData package represent bioassay data and the information and annotation on the data derivation.

The QuantitationType package defines the classes for quantitation, such as measured and derived signals of the reporters. The StandardQuantitationType class is designed to store structure and meaning of the molecule-expression data. The StandardQuantitationTypes consist of MeasuredSignal (eg, probe intensity), DerivedSignal (eg, binarized data [positive/negative]), and PresentAbsent (eg, present/absent). The MeasuredSignal consists of TissueIntensity, PercentTissueStaining, and NumOfNucleiCounted. By allowing control values to be included in the molecular-expression data, the accuracy and reliability of the measurements provided can be evaluated.

Clinical and Histopathologic Information Model

Classes in the ClinInfo package contain comprehensive clinical data on patients (Figure 5, a). The ClinInfo package contains classes for Demography, Diagnosis, Resection, MolecularAnalysis, Followup, and RelevantHistory, which is composed of PreviousTherapy and OtherHistory. Through the MolecularAnalysis class, TMA-OM can store the description and results of other types of experiments, including DNA microarray experiments using tissue from the same patient. The ClinInfo object has information on physicians who are responsible for the care of the patient. Additional types of clinical information can be described through *type* and *characteristics* associations with the OntologyEntry class.

The HistoPathol package provides classes describing histopathologic information of specimens (Figure 5, b). The BasicHistoPathol class stores elements that should be included regardless of the organ or tissue. The OrganSpecific class stores elements for specific organs. The BasicHistoPathol class is an abstract class, subclasses of which are the TumorInfo and Histology classes. TumorInfo contains PrimaryTumor, RegionalLymphNode, and

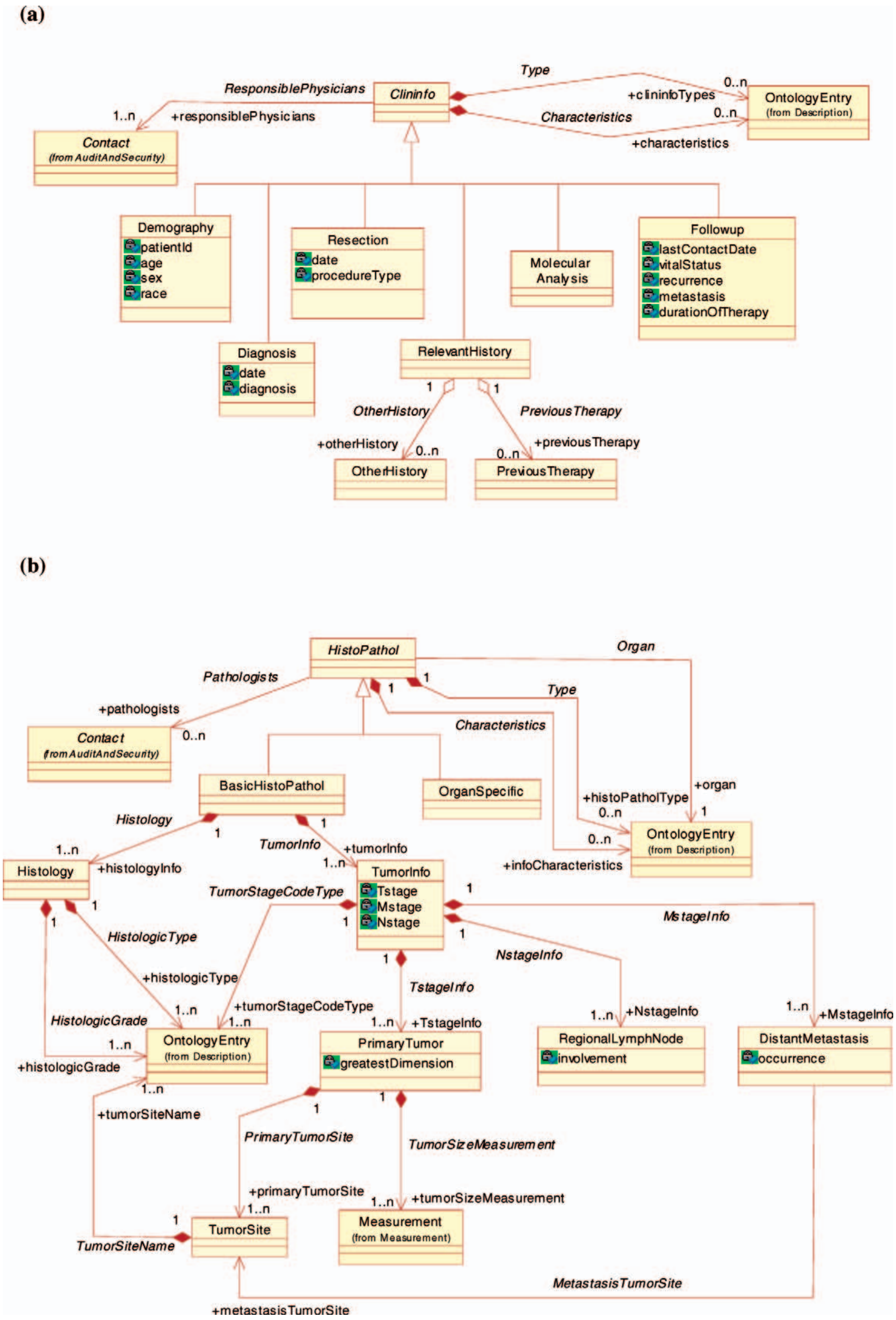

Figure 5. *Unified modeling language models of the ClinInfo and HistoPathol packages. a, The classes defined in the ClinInfo package describe clinical information of a patient from whom a cancerous specimen was extracted. b, The classes in the HistoPathol package describe the histopathologic information of the cancerous specimen.*

DistantMetastasis to describe detailed data under Tumor-Node-Metastasis staging information. HistologicType and HistologicGrade are modeled in the Histology class. The OrganSpecific class has no subclasses for flexibility; instead, elements of each cancer can be described through *organ*, *type*, and *characteristics* associations with the OntologyEntry class. The HistoPathol object contains information about pathologists who analyzed the specimen.

Implementation: Database and Web Interface

The TMA-OM has been implemented as a Web-based database application, Xperanto-TMA,[33] powered by MySQL 4.1 (MySQL Inc, Cupertino, Calif) database management system.

The relational schema is derived from TMA-OM by a formal object-relational mapping strategy. The mapping rule is as follows. Each class, except for abstract classes, is mapped onto a table. Associations between classes are stored in a single table (AssnList table) making the relational schema simpler. Each tuple in the AssnList table corresponds to one association in the class diagram and contains the identifiers of the 2 classes connected by the association. Abstract classes are not captured. The associations of abstract classes are passed on to those of the subclasses. The mapping produced 107 relational tables.

Although database management systems supporting object-oriented models and/or XML are available, some are proprietary, and available technologies are not yet mature enough to form a global standard. In the development of Xperanto-TMA, we tried to adhere to nonproprietary and open-source efforts and resources. Applying a formal object-relational mapping rule allows us to make use of the advantages of the relational database, including high performance, without losing the structure and constraints of the object model. Further information is available through the supplementary Web address.[32]

COMMENT

We have developed TMA-OM with a Web application to represent clinical and histopathologic information as well as experimental data for any type of cancer. Although previous studies considered only 1 or 2 cancers (Table), TMA-OM is designed to support the 43 cancer types considered by CAP cancer protocols, enriched by NCI CDEs, and allow users to incrementally extend data elements to other cancer types. We are developing templates that support guided entry for the individual cancer protocols and meta data registry that supports ISO11179 for detailed description of data elements (including CDEs). Creation of an abstract class for templates, which are composed of data elements in OntologyEntry class and the instances of which represent CAP cancer protocols and other protocols and their variations, will provide guiding structure.

There is a recent trend in bioinformatics toward the integration of microarray gene expression and other biological data by extending MAGE-OM. Xirasagar et al[34] proposed a data model called SysBio-OM for systems biology and Jones et al[16] proposed a data model called functional genomics experiment object model for functional genomics. The purpose of both studies was to represent gene expression data integrated with other biological data. The TMA-OM is designed with consideration of extensibility for integration with MAGE-OM.

The TMA-OM can be integrated with MAGE-OM by 2 methods. First, records in the TargetMolecule class in TMA-OM are designed to allow a link with the BioSequence class of MAGE-OM. Both the Reporter classes in TMA-OM (containing the TargetMolecule class) and in MAGE-OM (containing the BioSequence class) equally represent reporters to identify a specific molecule such as a gene. Second, the Specimen class in TMA-OM corresponds to the BioSource class in MAGE-OM. Both classes equally represent specific tissue samples.

Although the Reporter class of MAGE-OM can represent only the sequence information of DNA probes, the Reporter class of TMA-OM can represent a variety of probes, including DNA, RNA, and proteins as well as treatment information applied to them. Although the Specimen class of TMA-OM is primarily designed to represent only tissue specimens, it is possible to represent other forms of biological samples that are used in DNA microarray experiments.

Integrating object models may allow the combined analysis of TMA and DNA microarray data in a systematic fashion such as TMA-based, high-throughput validation of gene or protein expression patterns hypothesized by DNA microarray experiments. Array comparative genomics hybridization data, a high-throughput genomic technology measuring DNA copy-number alteration, can be represented in MAGE-OM, and hence in TMA-OM. It is likely that SysBio-OM and functional genomics experimet object model can be integrated with TMA-OM.

Because TMA-OM is independent of implementation, several applications can be constructed based on its use in different settings. We implemented the Xperanto-TMA, a Web-based database based on TMA-OM, by applying a formal object-relational mapping rule. Data in Xperanto-TMA can be converted into XML documents conforming to either TMA DES format or the native DTD that is automatically derived from TMA-OM. The TMA-OM is in support of TMA DES for efficient data exchange and it also provides a comprehensive data model for storage, analysis, and data integration with expressivity, flexibility, and extensibility. The TMA-OM supports MISFISHIE, TMA DES, and a wide range of clinical and histopathologic information extracted from CAP cancer protocols and NCI CDEs.

The potential for TMA technology to assist high-throughput validation of the clinical relevance of tumor markers is clear, and will greatly aid the rapid assessment of new therapeutic and prognostic markers in preclinical trials. Some applications in translational research can be suggested, such as progression-model TMA and outcome TMA. The TMA technology is applicable in clinical trials or in animal and experimental models.[35]. As TMA technologies have been developing, automated processes in experimental procedures have been growing. When various designs of TMA and new protocols are introduced, TMA-OM can sufficiently support them. When it is essential to modify the object model, it is easy to add or change the objects and associations because of its modular nature.

As the number of studies using TMA increases, there is a growing need for a data model to represent and exchange TMA data. It is meaningful to integrate TMA data with other biological data such as gene expression data and proteomics data to promote an understanding of the underlying biological nature. We hope that TMA-OM becomes more helpful as a data model for TMA experiments to meet these needs.

We have already developed and managed the Web database called Xperanto (Expressionist's Esperanto in XML),[36] which supports MGED standards. Xperanto supports most microarray platforms and the import and export of MAGE-ML. Tight coupling of Xperanto-TMA with Xperanto is being tested.

This study was supported by grant 2005-00162 from the Ministry of Science and Technology, Republic of Korea. Ms Lee and Ms Park's training was supported by a grant, 0405-BC02-0604-0004, Ministry of Health and Welfare, Republic of Korea. The authors thank all Seoul National University Biomedical Informatics members and external testers for the completion of this work.

References

1. Kallioniemi OP, Wagner U, Kononen J, Sauter G. Tissue microarray technology for high-throughput molecular profiling of cancer. *Hum Mol Genet.* 2001; 10:657–662.

2. Kononen J, Bubendorf L, Kallioniemi A, et al. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med.* 1998;4:844–847.

3. Au NH, Cheang M, Huntsman DG, et al. Evaluation of immunohistochemical markers in non-small cell lung cancer by unsupervised hierarchical clustering analysis: a tissue microarray study of 284 cases and 18 markers. *J Pathol.* 2004; 204:101–109.

4. Han H, Bearss DJ, Browne LW, Calaluce R, Nagle RB, Von Hoff DD. Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray. *Cancer Res.* 2002;62:2890–2896.

5. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science.* 1995;270:484–487.

6. Pinkel D, Segraves R, Sudar D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet.* 1998;20:207–211.

7. 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.* 2005;205:388–396.

8. Hustinx SR, Cao D, Maitra A, et al. Differentially expressed genes in pancreatic ductal adenocarcinomas identified through serial analysis of gene expression. *Cancer Biol Ther.* 2004;3:1254–1261.

9. Berman JJ, Edgerton ME, Friedman BA. The tissue microarray data exchange specification: a community-based, open source tool for sharing tissue microarray data. *BMC Med Inform Decis Mak.* 2003;3:5.

10. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME): toward standards for microarray data. *Nat Genet.* 2001;29:365–371.

11. Spellman PT, Miller M, Stewart J, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* 2002; 3:RESEARCH0046.

12. Brazma A, Parkinson H, Sarkans U, et al. ArrayExpress: a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 2003;31:68–71.

13. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002;30: 207–210.

14. Garwood K, McLaughlin T, Garwood C, et al. PEDRo: a database for storing, searching and disseminating experimental proteomics data. *BMC Genomics.* 2004;5:68.

15. Taylor CF, Paton NW, Garwood KL, et al. A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat Biotechnol.* 2003;21:247–254.

16. Jones A, Hunt E, Wastling JM, Pizarro A, Stoeckert CJ Jr. An object model and database for functional genomics. *Bioinformatics.* 2004;20:1583–1590.

17. Manley S, Mucci NR, De Marzo AM, Rubin MA. Relational database structure to manage high-density tissue microarray data and images for pathology studies focusing on clinical outcome: the prostate specialized program of research excellence model. *Am J Pathol.* 2001;159:837–843.

18. Shaknovich R, Celestine A, Yang L, Cattoretti G. Novel relational database for tissue microarray analysis. *Arch Pathol Lab Med.* 2003;127:492–494.

19. Demichelis F, Sboner A, Barbareschi M, Dell'Anna R. TMABoost: an integrated system for comprehensive management of Tissue Microarray data. IEEE *Trans Inf Technol Biomed.* 2006;10(1):19–27.

20. Berman JJ, Datta M, Kajdacsy-Balla A, et al. The tissue microarray data exchange specification: implementation by the Cooperative Prostate Cancer Tissue Resource. *BMC Bioinformatics.* 2004;5:19.

21. Nohle DG, Ayers LW. The tissue microarray data exchange specification: a document type definition to validate and enhance XML data. *BMC Med Inform Decis Mak.* 2005;5:12.

22. Microarray and Gene Expression-MAGE. MAGE standards resource web site. Available at: http://www.mged.org/Workgroups/MAGE/mage-om.html. Accessed March 5, 2005.

23. Microarray Gene Expression Data (MGED) Ontology. The ontology working group for the MGED project. Available at: http://mged.sourceforge.net/ontologies/. Accessed January 5, 2005.

24. Stoeckert C, Parkinson H. The MGED ontology: a framework for describing functional genomics experiments. *Comp Funct Genomics.* 2003;4:127–132.

25. MISHFISHIE Standing Working Group website. Available at: http://mged.sourceforge.net/misfishie/. Accessed January 18, 2005.

26. The College of American Pathologists. Cancer protocols in January 2005 revision version. Available at: http://www.cap.org/apps/docs/cancer_protocols/protocols_index.html. Accessed March 28, 2005.

27. Leslie KO, Rosai J. Standardization of the surgical pathology report: formats, templates, and synoptic reports. *Semin Diagn Pathol.* 1994;11:253–257.

28. National Cancer Institute Common Data Elements browser to search for data elements. Available at: http://cdebrowser.nci.nih.gov/CDEBrowser/. Accessed December 12, 2004.

29. 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.

30. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. *J Am Med Inform Assoc.* 1998;5:1–11.

31. Object Management Group documents. Available at: http://www.omg.org/cgi-bin/apps/doc?dtc/02-09-06.zip.

32. Web supplementary information is available at: http://www.snubi.org/software/xperanto-tma/suppl/. Accessed June 10, 2005.

33. Xperanto-TMA database. Available at: http://xperanto.snubi.org/TMA/. Accessed June 1, 2005.

34. Xirasagar S, Gustafson S, Merrick BA, et al. CEBS object model for systems biology data, SysBio-OM. *Bioinformatics.* 2004;20:2004–2015.

35. Henshall S. Tissue microarrays. *J Mammary Gland Biol Neoplasia.* 2003; 8:347–358.

36. Park JY, Park YR, Park CH, Kim JH. Xperanto: a web-based integrated system for dna microarray data management and analysis. *Genome Inform.* 2005; 3:39–42.