Queries for Author

Journal: **Journal of the American Medical Informatics Association**
Paper: **amiajnl-2011-000170**
Title: **ImageMiner: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology**

The proof of your manuscript appears on the following page(s).
Please read the manuscript carefully, checking for accuracy, verifying the reference order and double-checking figures and tables. When reviewing your page proof please keep in mind that a professional copyeditor edited your manuscript to comply with the style requirements of the journal.
This is not an opportunity to alter, amend or revise your paper; it is intended to be for correction purposes only.

During the preparation of your manuscript for publication, the questions listed below have arisen (the query number can also be found in the gutter close to the text it refers to). Please attend to these matters and return the answers to these questions when you return your corrections.

Please note, we will not be able to proceed with your article and publish it in print if these queries have not been addressed.

Query Reference	Query
	Please ensure all author names are correct because we are close to publishing your paper online - these data will be recorded on PubMed and CrossRef
	Please check that the "Provenance and peer review" statement is correct about your article
1	Please provide a structured abstract
2	NB the journal does not divide articles into Sections. OK to usa 'described above' here?
3	Ref items 6 & 37 and ref items 63 & 70, were identical. The duplicate items were deleted and the reference list was renumbered accordingly. Please check if this is appropriate.
4	Please provide complete details for reference 31.
5	Please provide publisher location for references 34, 52 and 56.
6	Please provide publisher name and location for reference 45, 46, 55 and 57.
7	Please provide volume number and page ranges for references 48, 58.
8	Please provide publisher name for references 59, 63.

If you are happy with the proof as it stands, please email to confirm this. Changes that do not require a copy of the proof can be sent by email (please be as specific as possible).
Email: production.jamia@bmjgroup.com

If you have any changes that cannot be described easily in an email, please mark them clearly on the proof using the annotation tools and email this by reply to the eProof email.

PLEASE RESPOND WITHIN 48 HOURS

ImageMiner: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology

David J Foran,[1] Lin Yang,[1] Wenjin Chen,[1] Jun Hu,[1] Lauri A Goodell,[1] Michael Reiss,[1] Fusheng Wang,[2] Tahsin Kurc,[2,3] Tony Pan,[2] Ashish Sharma,[2,3] Joel H Saltz[2]

► An additional material is published online only. To view this file please visit the journal online (www.jamia.org).

[1]Center for Biomedical Imaging & Informatics, The Cancer Institute of New Jersey, UMDNJ-Robert Wood Johnson Medical School, New Brunswick, New Jersey, USA
[2]Center for Comprehensive Informatics, Emory University School of Medicine, Atlanta, Georgia, USA
[3]Department of Biomedical Engineering, Emory University, Atlanta, Georgia, USA

Correspondence to
Wenjin Chen, Center for Biomedical Imaging & Informatics, The Cancer Institute of New Jersey, UMDNJ-Robert Wood Johnson Medical School, 195 Little Albany St, Room 3521, New Brunswick, NJ 08901, USA; chenwe@umdnj.edu

Received 7 February 2011
Accepted 9 April 2011

ABSTRACT

The design and implementation of ImageMiner, a software platform for performing comparative analysis of expression patterns in imaged microscopy specimens such as tissue microarrays (TMAs), is described. ImageMiner is a federated system of services that provides a reliable set of analytical and data management capabilities for investigative research applications in pathology. It provides a library of image processing methods, including automated registration, segmentation, feature extraction, and classification, all of which have been tailored, in these studies, to support TMA analysis. The system is designed to leverage high-performance computing machines so that investigators can rapidly analyze large ensembles of imaged TMA specimens. To support deployment in collaborative, multi-institutional projects, ImageMiner features grid-enabled, service-based components so that multiple instances of ImageMiner can be accessed remotely and federated. The experimental evaluation shows that: (1) ImageMiner is able to support reliable detection and feature extraction of tumor regions within imaged tissues; (2) images and analysis results managed in ImageMiner can be searched for and retrieved on the basis of image-based features, classification information, and any correlated clinical data, including any metadata that have been generated to describe the specified tissue and TMA; and (3) the system is able to reduce computation time of analyses by exploiting computing clusters, which facilitates analysis of larger sets of tissue samples.

INTRODUCTION

Tissue microarray (TMA) technology preserves limited tissue resources and reagents by providing the means of producing large numbers of individual core biopsy samples (histospots) rather than a limited number of standard-sized histology sections. A carefully planned array can be constructed using TMAs such that a 20-year survival analysis can be performed on a cohort of 600 or more patients using only 100–200 μl of antibody.[1] TMAs can provide insight into the underlying mechanisms of disease progression and can be used as a tool to improve prognostic accuracy and personalize treatment regimens for specific subpopulations of patients.[1–3]

The dense matrix configuration of histospots utilized in the TMA technology lends itself to high-throughput quantitative analysis. However, unlike DNA microarrays, wherein each tiny spot within a given array is homogeneous and represents a unique cloned complementary DNA or oligonucleotide, individual spots within a TMA often consist of a complex, heterogeneous mix of tissues. These factors complicate quantitative analysis of TMA specimens tremendously. Currently, the primary methods used for such evaluations involve manual and interactive review of TMA samples, while they are subjectively analyzed and scored. An alternate, but less utilized, strategy is to sequentially digitize specimens for subsequent semi-quantitative assessment.[4 5] Both procedures ultimately entail the interactive evaluation of specimens, which makes for a slow and tedious process which is prone to both inter- and intra-observer variability. Together, these factors significantly reduce the reliability and reproducibility of the assessment.

Since about 2001, the idea of developing more effective methods and protocols to conduct the quantitative analysis of TMA specimens has become an extremely active area of research.[6–12] For example, the automated quantitative analysis (AQUA) system was developed to help automate the process of characterizing the staining intensities of tissue samples. The AQUA system is a molecular based approach for quantitatively assessing protein expression with the intent to reduce the degree of variability associated with pathologist-based evaluation of samples.[13] Several other groups have also undertaken projects to read immunohistochemistry TMA specimens using commercial complementary DNA microarray readers.[14 15] Although significant progress has been made in the development of automated methods for assessing TMAs, most of the existing efforts are limited by the fact that they are closed and proprietary, do not exploit the potential of advanced computer vision techniques, do not integrate well with a TMA data management system, and/or do not conform with emerging data standards. Future advances in histopathology imaging will rely on the availability of reliable, portable algorithms and computational tools that can provide automated data management operations, perform high-throughput quantitative analysis, and support query and retrieval of image-based analysis results in collaborative environments.

In this paper we present the design and implementation of ImageMiner, an open source and

novel software platform, designed to address many of the stated challenges and requirements. ImageMiner is designed to provide scientists and physicians with a set of portable, reliable investigative tools for performing high-throughput comparative analysis and reproducible characterization of expression profiles in imaged TMAs. The distinguishing characteristics of the system are the capacity to support queries and perform comparisons across large datasets originating from both standard robotic and virtual microscopes and the capacity to automatically locate and retrieve those imaged tissue discs from within distributed, 'gold standard' archives that exhibit expression patterns that are most similar to those of a given query disc. The ImageMiner system encapsulates the following three main functions to support quantitative and comparative TMA analysis.

Objective, computer-based analysis of TMA images

The system implements a library of image processing operations, including automated registration, segmentation, feature extraction, and classification. This functionality is designed and implemented for use in both stand-alone mode and in conjunction with high-performance computing platforms. The ImageMiner image analysis methods have been successfully executed on parallel machines and in distributed environments, making it practical to reliably assess multiple imaged specimens concurrently, thereby facilitating rapid high-throughput analysis.

Management of TMA image datasets and analysis results

ImageMiner provides support for investigators to manage and reliably search through TMA image data and corresponding analytical and experimental results. The ImageMiner data model is designed to capture imaged specimen information, correlated clinical data, and image markups and annotations.[16] The design of this model is based on input from a panel of consulting medical oncologists and pathologists.[17] The image archival subsystem and data management software modules are developed in keeping with emerging guidelines from the cancer Biomedical Informatics Grid (caBIG) program and the Association for Pathology Informatics.

Remote access to and federation of ImageMiner instances

Our implementation uses a service-based architecture and grid-computing technologies. The data analysis and data management components of ImageMiner can be accessed from remote clients via service interfaces, and multiple ImageMiner deployments can be federated in a distributed setting. This function allows collaborating investigators to manage local analytical and database resources and share these resources with other project participants across institutions.

To evaluate the system in real-case scenarios and conditions, a multi-institutional, grid-enabled consortium has been established among investigators located at strategic sites at the Cancer Institute of New Jersey, Emory University School of Medicine, the Ohio State University, Rutgers University, University of Texas, and the University of Pennsylvania School of Medicine.

BACKGROUND AND RELATED WORK

Microscopy image analysis

Automatic quantification or computerized processing of TMA specimens remains an extremely active and challenging area of research because: (1) high-throughput microscopy imaging easily generates thousands of cores for each TMA study; (2) there is a relatively limited number of pathologists who are available and experienced with the process of evaluating TMAs; (3) it is generally accepted that the development of reliable algorithms could lead to objective and reproducible measures while reducing or eliminating the tedium and fatigue associated with manual assessment; and (4) since all the cores on a TMA slide are stained using the same, identical protocol as well as the same processing times and temperature, large-scale, computerized analysis of these specimens is facilitated.

As of this writing, only a few TMA management systems have been reported that provide algorithmic support for applying automatic methods to quantify and assess TMAs,[18 19] and, because several commercial projects are proprietary with development taking place in isolation, they cannot be easily adopted by the general research community. At the same time, there are few efforts that have been shown to successfully support high-performance computing applications in TMA analysis.[20]

Recently, because of the advances that have been made with regard to computational capacities and pattern recognition technologies, there has been renewed interest throughout the research community in applying content-based image retrieval (CBIR) techniques to the analysis and mining of image data in biomedical applications.[21–24] Individual strategies and approaches used in these systems differ as to the degree of generality (general vs domain specific), level of feature abstraction (primitive vs logical), overall dissimilarity measure used in retrieval ranking, level of user intervention (with or without relevance feedback), and the methods used to evaluate the system's performance. One of the early systems reported by the Pittsburgh Supercomputing Center used global characteristics of images to provide a measure of the Gleason grade of prostate tumors.[25 26] Results obtained using this system exhibited a strong correspondence between the image distance generated by the computer algorithm and the pathological significance as judged by certified anatomical pathologists. Wang *et al* from Pennsylvania State University introduced the use of wavelet technology and integrated region matching distances for characterizing pathology images.[27] Over the years, a rich set of techniques have been used to classify pathology images. An excellent recent review of this active area of research has been produced by Gurcan *et al*.[28] While our work draws from some of these earlier efforts, the algorithms that our team are developing have been directed towards, and optimized for, performing automated analysis of TMA specimens. In an earlier study, we reported using a mixed set of 3744 breast tissue samples, including normal tissue, ductal hyperplasia, fibroadenoma, atypical ductal hyperplasia, ductal carcinoma in situ, and invasive lobular carcinoma, to carry-out experiments to determine the efficacy of those algorithms for their capacity to systematically classify imaged tissue discs.[29] During the course of those experiments, the system provided an average correct classification performance rate of 89% when used to distinguish between primary breast carcinomas and non-cancerous breast tissue including normal tissue, ductal hyperplasia, fibroadenoma, and atypical ductal hyperplasia. During subsequent experiments, the algorithms provided an average accuracy of 80% when the prototype was used to discriminate between two subgroups of breast cancer and non-cancerous breast tissue samples.[29] As an extension of this work, ImageMiner now features a quick, reliable segmentation module which has been integrated with the CBIR module, making it possible to perform classifications based on the texton signatures of specific sub-regions—that is, tissue, cell, or subcellular level within a given histospot rather than on the signatures of an entire disc. This new feature improves the performance significantly and expands the range of applications

for which the system can be used. ImageMiner now also features options that enable investigators to automatically locate, retrieve, characterize, and display high-resolution versions of imaged discs individually or lower-resolution ensembles of ranked retrievals based on their similarity to a gold standard image archive of previously classified cases.

In an attempt to address the challenges of high-throughput analysis, several investigators have begun to exploit distributed computing technologies. For example, our own team recently demonstrated the use of a high-performance computing system for automatic analysis of imaged histopathology breast tissue specimens.[30] Gurcan *et al* reported the successful application of distributed computing in a pilot project to support automated characterization of neuroblastoma using the Shimada classification system.[31] The ImageMiner system that we are developing is a logical extension of our early successful efforts developing network-based clinical decision support systems[32–36] and large-scale, feasibility studies that we conducted on IBM's World Community Grid in July 2006, using more than 100 000 imaged tissue samples.[29 30]

TMA data model and data management

Because of the high density of histospots and the intrinsic difficulties that arise during their evaluation, TMA analysis requires tight communication and coordination among the modules that are used to conduct computerized data management and those responsible for digital imaging. To address the increasing need for interoperable ways of exchanging TMA slides and related data within the clinical and research communities, the tissue microarray data exchange specification (TMA DES) was published in 2004 as a guideline for standardizing TMA data exchange.[6] Several systems that were subsequently developed adopted these specifications and can export compatible XML datasets. However, the fact that this specification was designed to be sufficiently general and flexible to accommodate any user-defined tags, parsing, and sharing of TMA information among institutions and/or systems was still highly problematic. The updated TMA DES,[37] published in 2005, provided a well-defined XML document type definition (DTD) for validating documents to improve compliance with emerging standards. It also allowed extensions to the core DES DTD by permitting local data element definitions.

Several object-oriented TMA data models and data management systems have been developed to work with a backend, relational database. Aside from the 'donor—core location—image—evaluation' relations, which are common across each, these designs vary significantly with regard to the following features and capabilities: (1) the system developed by Barsky *et al*[38] was integrated with a custom constructed arrayer to allow it to directly manage the construction process, whereas commercially available arrayers lacked such capabilities; (2) varying levels of flexibility are supported to enable investigators to manage clinical information and other salient data which are gathered throughout the course of a research study. For instance, most systems do not provide adequate support for multiple simultaneous users to evaluate the same TMAs, nor do they provide sufficient flexibility for creating user-defined data tables; (3) only a few models and systems provide computational support for image segmentation and classification; and (4) the crucial role of semantic interoperability has not yet been addressed in the research community. Recent TMA data models and data management systems have adopted one or more standards including the NCI Thesaurus.[18 39–41]

The ImageMiner data model that we are developing is based on input from a panel of consulting oncologists and pathologists and designed to capture imaged specimen information, correlated clinical data, and image markups and annotations. It has been implemented to support quantification and classification of imaged tissue specimens and it provides extensions to accommodate new data types resulting from human and computer analyses. The ImageMiner system is implemented with the goal of interoperability (by leveraging caBIG standards and interoperability guidelines) and enables investigators to export XML data compliant with TMA DES standards.

Grid services for resource federation

Group efforts are critical to the study of scientific problems that require complementary sets of expertise and resources. During the course of many modern collaborative projects, it is necessary to provide collaborating teams with access to datasets and analysis methods that may be distributed across multiple institutions. We have designed the ImageMiner system to allow secure federation of local ImageMiner instances across distributed networks. To implement this functionality, we have used a service-oriented architecture design and we have leveraged grid-computing technologies.

Grid computing has been successfully used in an increasing number of large-scale, biomedical research efforts. The Biomedical Informatics Research Network (BIRN) project,[42 43] for example, is funded by the National Institutes of Health (NIH) to provide collaborative access to, and analysis of, distributed datasets generated from neuroscience studies. The MammoGrid and eDiamond projects[44–46] build and federate medical image databases for mammography datasets and to facilitate collaboration among researchers and clinicians throughout the European Union. The Cardio Vascular Research Grid provides applications and software tools in a service-oriented grid framework for cardiovascular research groups (http://cvrgrid.org). Another large-scale informatics effort is the caBIG (https://cabig.nci.nih.gov) program, which has its primary focus on advancing cancer research.[47–49] The overarching goal of caBIG is to develop the requisite grid infrastructure, standards, processes, and applications to allow more effective sharing of data and analytical resources across institutions, while providing support to facilitate collaborative, multi-institutional projects. Our work draws from the principles, architectures, and tools being developed in these efforts. In particular, we leverage the service-based tool design guidelines of caBIG and the grid infrastructure, called caGrid,[49 50] developed in that program in order to create services that are customized to support collaborative, multi-site TMA studies.

DESIGN OBJECTIVES

The primary design criteria for ImageMiner is to develop a software platform that enables investigators to perform rapid, large-scale, and reproducible comparative analysis of expression patterns in digitized TMAs. Using ImageMiner, investigators can apply cascades of computerized image-processing methods on multiple arrays, each of which contains hundreds or even thousands of imaged tissue discs. These methods segment each disc image into spatial structures, compute a set of features for each segmented structure, and classify the disc images on the basis of the resulting feature signature.

Another central objective of this work is to support mining of imaged specimens and experimental data. Toward that end, metadata about TMAs, analysis results (ie, image segmentations, features, and classification results), and metadata about analyses (eg, which methods and input parameters were used for a given set of results) are stored in the system in a format that

can be queried by users. The types of queries supported by the system include: (i) retrieval of TMAs and imaged discs based on TMA metadata; (ii) queries to search for image discs that exhibit expression and staining patterns which are most similar to those of a given image disc; (iii) spatial queries on assessing the relative prevalence of features or classified objects, or assessing spatial coincidence of combinations of features or objects; and (iv) queries to support selection of collections of segmented regions, features, objects for further machine learning or content-based retrieval applications.

The third objective of the design is to facilitate more efficient use of the TMA technology multi-site studies. The ImageMiner system is designed such that multiple deployments of the system can be federated in multi-institutional settings. The members of a collaborative project can remotely access each ImageMiner instance, execute TMA analyses, and query analysis results across those instances. The ability to choose various analysis algorithms can assist algorithm developers in algorithm evaluation and validation as well as researchers in comparative analyses. For example, an algorithm researcher could select similar algorithms implemented and hosted as services by other researchers and compare the results from those algorithms with his/her own results. Similarly, biomedical researchers could use results from multiple algorithms, which might extract different types of information (eg, texture vs anatomic structures), to create different views of the specimens under study and compare results from these different views.

SYSTEM DESCRIPTION

The design objectives described above are realized by three components. The first component implements data analysis functionality and supports image segmentation and feature extraction operations. The second component provides support for data management and content-based search and retrieval of imaged specimens. The third component enables grid-based concurrent execution of image analysis operations on cluster machines and standards-based mechanisms for data and algorithm sharing. Figure 1 illustrates the architecture and main components of the ImageMiner system. Each of these components is described in detail in the sections that follow. The dependencies between these components are shown in table 1.

Data analysis component

This component implements a library of operations to perform automatic analysis of TMA arrays and can leverage parallel computing platforms to speed up analysis of multiple TMA images.

Image analysis library

The current library includes operations to correct for artifacts and compensate for mechanical distortions and other artifacts within an imaged specimen and to perform automatic segmentation, feature extraction, and classification of tissue samples.

Figure 2A shows a representative TMA exhibiting bowing of rows and columns. In order to develop a reliable means to compensate for the mechanical distortion of arrays, it is necessary to devise an algorithm that could accurately extract the exact grid location of each disc throughout the specimen. To achieve this objective, our algorithms were designed to operate on a low-resolution image map of the array. The registration algorithm uses a combination of template matching and Hough transformation to effectively identify tissue cores and accurately model the rows and columns of the matrix structure.[51 52]

The feature extraction operation is subsequently performed by automatically generating texton measurements for each tissue disc. Textons were first defined by Bela Julesz, the late cognitive scientist, as conspicuous repetitive local features that humans perceive as being discriminative among textures.[53] A computational model for textons was later introduced by Leung and Malik, using cluster centers in a feature space which are generated in response to a fixed set of filter banks.[54] Textons have been successfully used to perform texture classification by a host of investigators.[55–57] In the ImageMiner system, the feature extraction process computes multi-scale texton histograms for each imaged disc using the Schmidt[58] and LM[54] filter banks. We demonstrated that trained strong classifiers can be successfully used to automatically delineate the tumor and non-tumor regions within a breast cancer specimen.[59 60] Thus we have developed an automated means of performing classification of the tissue discs using an Adaboost classifier.[61] Our library also includes the soft margin support vector machine (SVM) and the boosting techniques to improve classification accuracy. The SVM is a set of supervised learning methods for classification

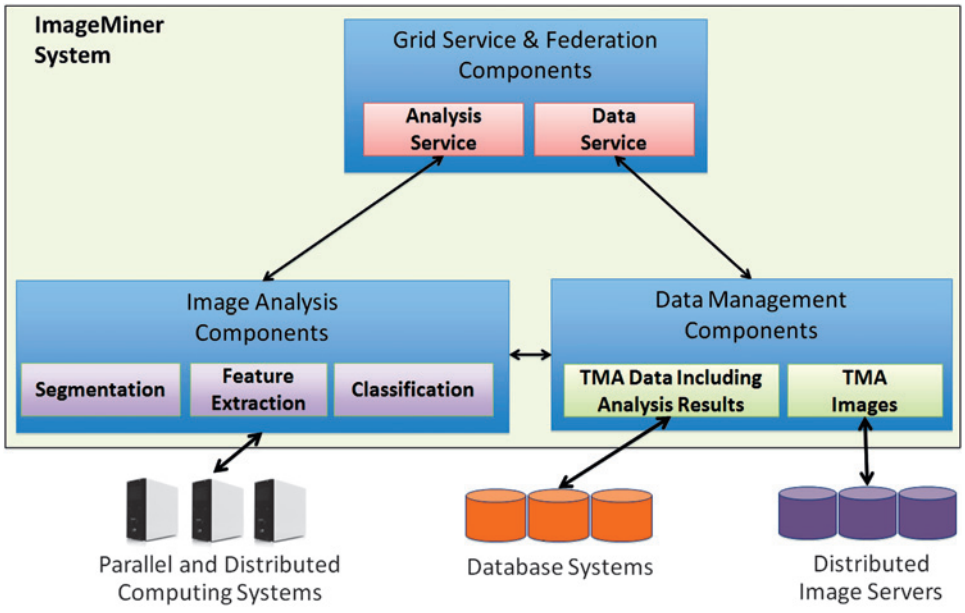

Figure 1 Architecture of ImageMiner. The system consists of three components. The image analysis component encapsulates a library of image processing operations. These operations can be run on parallel and distributed computing systems in order for the system to concurrently process multiple arrays and image discs. The data management component stores and manages metadata about TMA images and analysis results in the form of image markups and annotations. The Grid service and federation component enables remote federated access to ImageMiner instances across institutions.

Foran DJ, Yang L, Chen W, *et al*. *J Am Med Inform Assoc* (2011). doi:10.1136/amiajnl-2011-000170

Table 1 Dependencies between the various ImageMiner core components

ImageMiner component	Description	Dependencies
ImageMiner system	Integrated ImageMiner software system	ImageMiner analysis methods, ImageMiner analytical service, graphical user interface, PAIS database, ImageMiner database
ImageMiner database	Database on pathology imaging-related data and TMA image data	ImageMiner data model, relational database system
PAIS database	Database on pathology and microscopy image analysis results (markups and annotations), provenance information	PAIS data model, relational database with spatial query capabilities (eg, IBM DB2)
ImageMiner analytical service	Analytical service with parallel computing backend	caGrid service infrastructure, parallel computing backend system (eg, Message Passing Interface program, DataCutter, and job scheduler)

TMA, tissue microarray.

and regression. Boosting is one of the most important recent developments in machine learning. It works by sequentially applying a classification algorithm on a reweighted version of the training data while the final label is determined by weighted voting.

Over the course of the past 12 months, the key computational and imaging tools have been migrated to the histopathology and imaging shared resources at the Cancer Institute of New Jersey for use in ongoing investigative studies. During the course of the deployment of these Java-based software tools, it was shown that they pose minimal requirements in terms of client computers (Windows XP or Windows 7, 2 GHz, 2 GB RAM). They have been successfully used to analyze microarrays consisting of cancers of the breast, head and neck, and prostate. As part of a fairly recent study, the automated software was used to quantify Beclin1 expression, which was shown to be predictive of autophagy.[62] Our team has since conducted a series of man–machine performance studies. In the first experiments, we used the TMA analysis tools to evaluate immunohistochemical staining intensity on imaged breast cancer TMA specimens comprising 1407 tissue cores. The results showed that the computer software algorithms achieved similar interpretations to those provided by a panel of three board-certified pathologists and was consistent with inter-pathologist concordance. These results were presented at the 2010 Annual Conference of the United States and Canadian Academy of Pathology.[63]

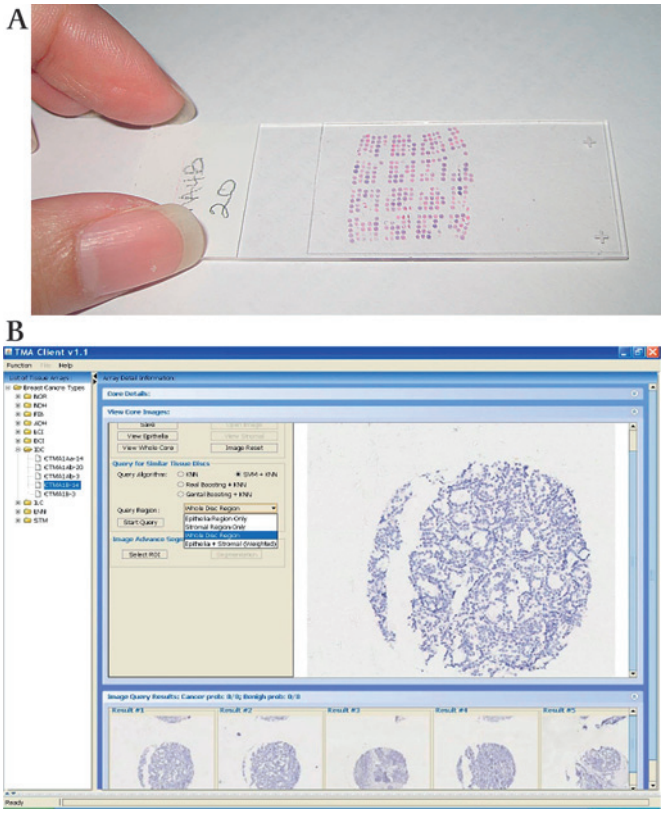

Figure 2 (A) Tissue microarray (TMA) mechanical distortion. (B) ImageMiner Prototype Client Query Interface.

High-performance computing

Processing of large micro-tissue arrays takes excessive amounts of time using a client's workstation. Although each image disc is relatively small, an array may contain hundreds of tissue discs, and a large-scale study may use hundreds to thousands of arrays. To address the computational requirements of analyses, we have designed the data analysis component to take advantage of computing clusters. The high-performance computing support draws from the DataCutter framework.[64–66] DataCutter is designed and implemented as a stream-filter framework in which a data processing pipeline can be composed as a network of interacting components, referred to as filters. The filters interact with each other by sending and receiving data through communication channels referred to as streams. In our current implementation, we use the bag-of-tasks execution model using DataCutter. A single image processing operation or a group of interacting operations is treated as a single task. Multiple instances of these tasks are instantiated on different computation nodes of a cluster. Images received by ImageMiner for analysis are distributed to these instances using a demand-driven strategy (to balance computational load among the task instances).

Data management and image search component

The ImageMiner system is designed to support both indexing and querying of imaged tissue samples and correlated clinical data based on the staining characteristics and expression signatures of a given specimen. In addition to providing CBIR capability, it also provides support for queries using standard, text-based criteria, such as the diagnosis of record, histological type, tumor grade, and biomarker used in a given study and queries based on the measurable parameters of a given disc, such as effective staining area and staining intensity. Figure 2B shows the client interface of the prototype image search and retrieval module. The client interface allows users to interactively select any region or object of interest within a given disc and initiate a query based on the texton histogram of that particular subregion, tissue, or cell. Users can subsequently refine queries by clicking on any one of the ranked retrievals, in which case the selected ranked retrieval serves as the new query input image. This feature makes it possible to iteratively modify the search until the desired image ensemble has been obtained. The ImageMiner system is compliant with TMA DES.[6 37] It is able to export valid TMA data for exchange in accordance with these specifications.

The data model underlying the ImageMiner database[16] is developed on the basis of input from a panel of consulting

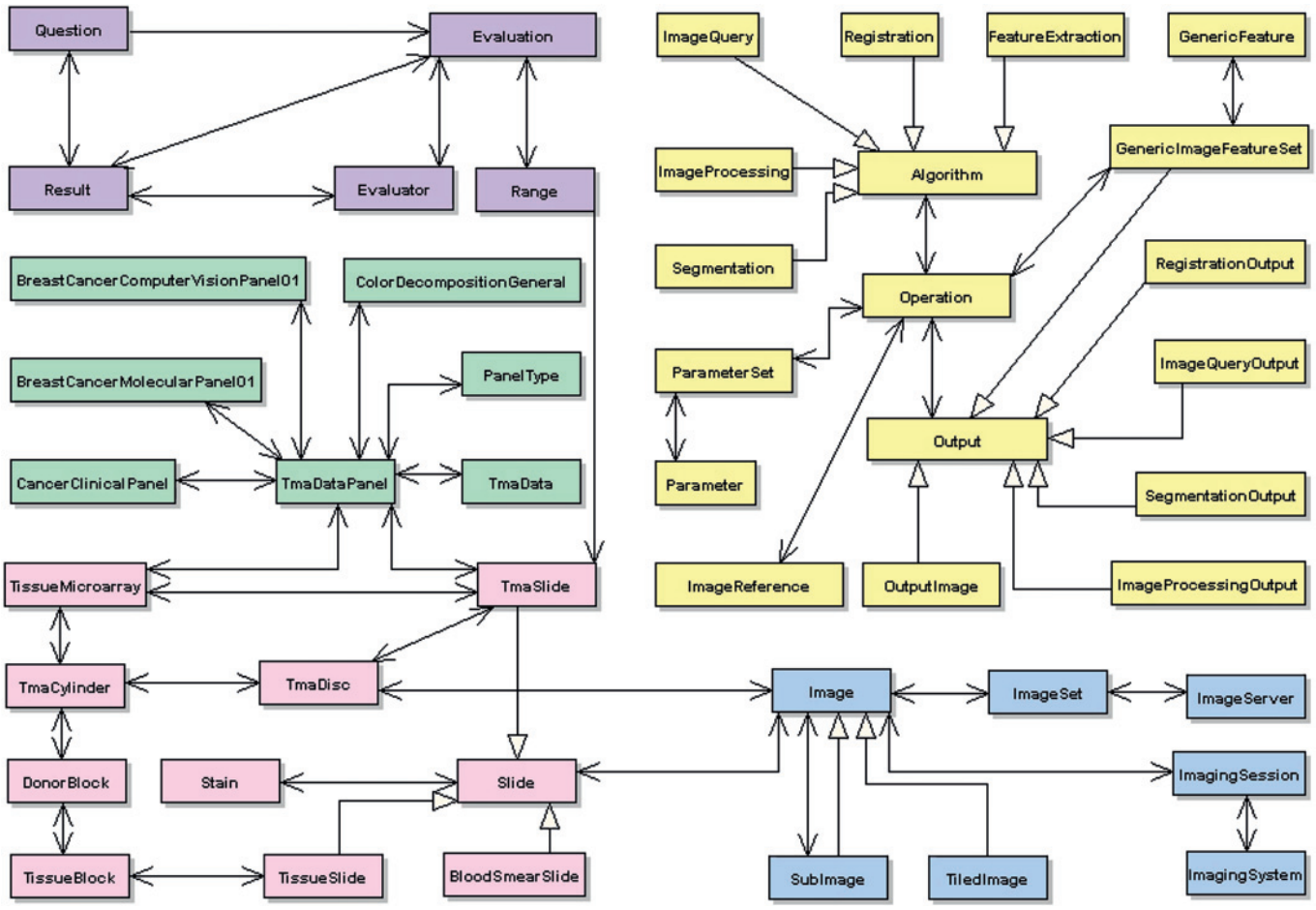

Figure 3 Overview of ImageMiner data model. Only the main classes and class relationships are illustrated; class attributes are not shown. Tma, tissue microarray.

oncologists and pathologists.[36] The main classes of the data model are shown in figure 3. The data model is designed to host quantitative and qualitative information derived from the physical and digital specimens, including clinical data and research data. The implementation of a generic table solution, 'TmaDataPanel', provides flexibility in managing polymorphous datasets that are derived from TMA studies (classes rendered in green). In the design phase, researchers are free to design and reuse different data tables, which are then stored in the generic TmaDataPanel table. In later phases of the research, the data can be reorganized into specialized classes that are optimized to facilitate search and interoperability. The 'Evaluation' classes (rendered in purple) of the model are designed specifically to support TMA evaluation studies using registered TMA images. The model supports evaluation studies while providing flexibility in terms of the configuration of tissue discs within the specimens, which evaluators are to participate in the studies, and which question sets are used. Once a study has been completed, the organizer is free to finalize and transform the data into a TmaDataPanel for sharing and searching the study results. The data model version 1.0 consists of 58 classes and 262 data elements (attributes). It has undergone review by the NCI Enterprise Vocabulary Services program to ensure compliance with caBIG standards and has been loaded into the Cancer Data Standards Repository (caDSR). The model can be viewed and retrieved via caBIG CDE Browser (https://cdebrowser.nci.nih.gov/CDEBrowser/).

The classes associated with image analysis results (classes rendered in yellow in figure 3) are used to manage image analysis results and metadata about the image analysis methods and parameters. Our team is currently working to extend and harmonize the analysis results component of the ImageMiner data model with the PAIS (Pathology Analytical Imaging Standards) model[67] to support markup and annotations in TMA, pathology, and microscopy imaging applications, while maintaining interoperability with corresponding standards in the radiology domain. The PAIS model was motivated by the requirements of our ongoing TMA project as well as the general problem of analyzing whole-slide microscopy images. PAIS is being developed in keeping with the Annotation and Image Markup (AIM) model,[68 69] which is under development in the caBIG In-Vivo Imaging Workspace to support radiological image annotation and markup in healthcare and clinical trial environments. The PAIS model can take advantage of the AIM model to represent observations and markups (geometric shapes) for image segmentations. However, PAIS has been optimized for representing fine-grained markups and annotations and provides additional information for data provenance, such as algorithms and parameters used for image segmentation.

Figure 4 shows the major components of PAIS. Please note that the figure illustrates the main classes and class relationships in the model, whereas the attributes of each individual class is not shown. (1) 'ImageReference' provides metadata that describe an image or a group of images (eg, DICOM images, TMA images, and whole-slide microscopy images), which are used as the base for markup and annotation and can be used to identify and retrieve the images from an image archive. The

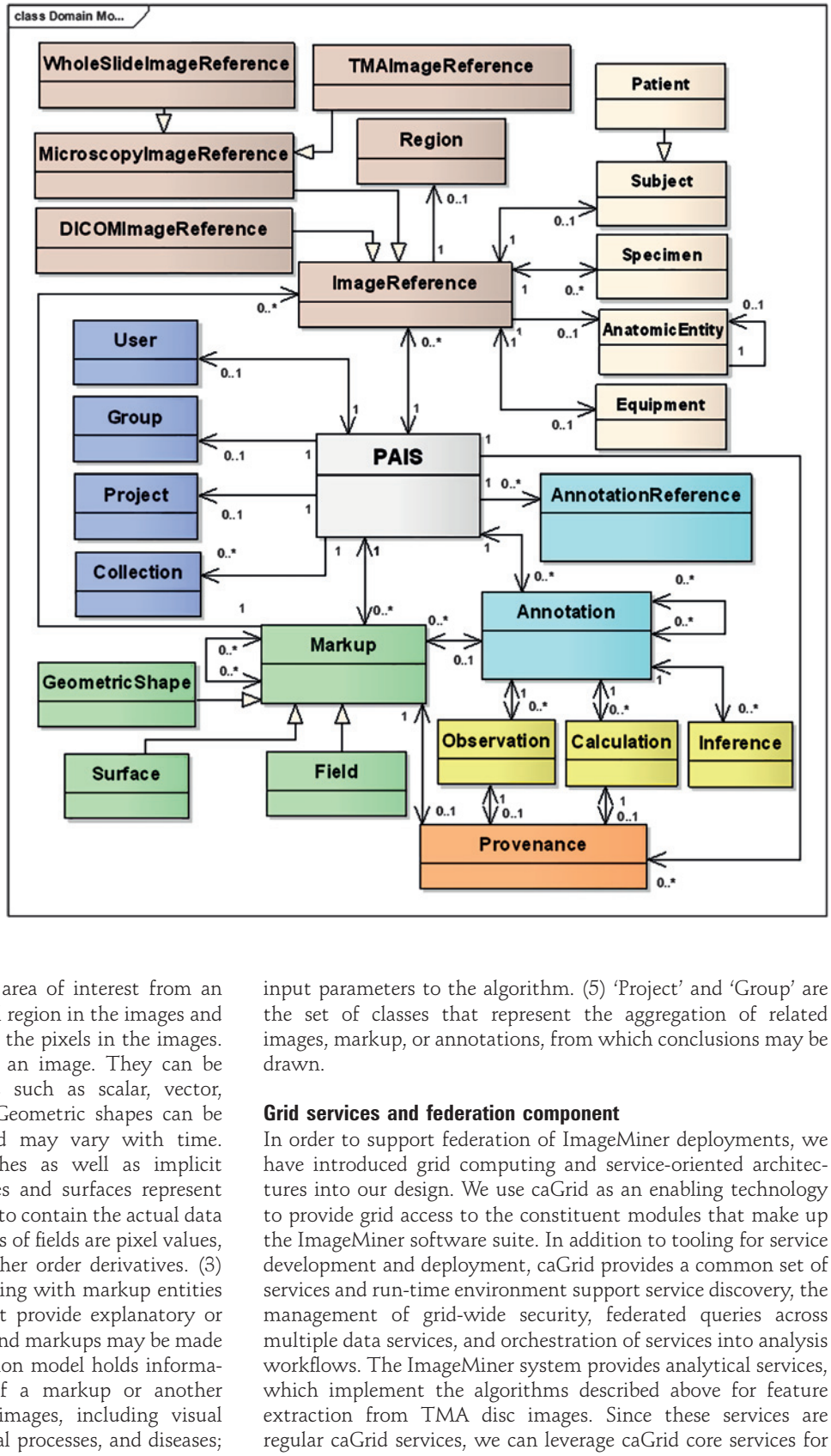

Figure 4 PAIS data model. Only the main classes and class relationships are illustrated; class attributes are not shown.

'Region' class is used to identify the area of interest from an image. (2) 'Markup' delineates a spatial region in the images and represents a set of values derived from the pixels in the images. Markup symbols are associated with an image. They can be geometric objects, surfaces, or fields such as scalar, vector, matrix, and more generally tensors. Geometric shapes can be one-, two-, or three-dimensional and may vary with time. Surfaces include finite element meshes as well as implicit surfaces. While both geometric shapes and surfaces represent boundaries in space, a field can be used to contain the actual data values within a spatial region. Examples of fields are pixel values, binary masks, gradient fields, and higher order derivatives. (3) 'Annotation' associates semantic meaning with markup entities through coded or free text terms that provide explanatory or descriptive information. Annotations and markups may be made by humans or machines. The annotation model holds information about: (*a*) the interpretation of a markup or another annotation entity in one or more images, including visual features, morphological or physiological processes, and diseases; (*b*) the quantitative results from mathematical or computational calculations; and (*c*) the disease diagnosis derived by observing imaging studies and/or medical history. (4) 'Provenance' is information that helps to determine the derivation history of a markup or annotation. This information includes the algorithm name, specification of input datasets, and the values of the input parameters to the algorithm. (5) 'Project' and 'Group' are the set of classes that represent the aggregation of related images, markup, or annotations, from which conclusions may be drawn.

Grid services and federation component

In order to support federation of ImageMiner deployments, we have introduced grid computing and service-oriented architectures into our design. We use caGrid as an enabling technology to provide grid access to the constituent modules that make up the ImageMiner software suite. In addition to tooling for service development and deployment, caGrid provides a common set of services and run-time environment support service discovery, the management of grid-wide security, federated queries across multiple data services, and orchestration of services into analysis workflows. The ImageMiner system provides analytical services, which implement the algorithms described above for feature extraction from TMA disc images. Since these services are regular caGrid services, we can leverage caGrid core services for support, such as service discovery, workflows, and security as shown in Table 2.

The analytical resources in ImageMiner are wrapped as caGrid-compatible services with well-defined interfaces. The backend of the analysis service can be a single machine or a cluster. An ImageMiner client can communicate to these

Table 2 ImageMiner provides analytical services for feature extraction from tissue microarray disc images. Various caGrid core services can be leveraged to support service discovery, multi-service workflows, and secure access to ImageMiner services

Service	Description
ImageMiner analytical service	Regular caGrid application service. Encapsulates algorithms for feature extraction
caGrid index service	An ImageMiner service can register to the caGrid index service. Clients could query the index service to discover the ImageMiner services
caGrid workflow service	A client can use the workflow service to compose and execute workflows that may involve multiple analytical and data services. In our implementation, our client program does not provide support for composing workflows using multiple services. The caGrid workflow service could be leveraged for that purpose
caGrid security services (Dorian, GridGrouper, GTS, CDS)	These services can be used to limit access to an ImageMiner service. The ImageMiner service can be deployed as a secure service requesting a client have a grid identity (using Dorian) and limit access to service and service operations based on the groups (GridGrouper) a client belongs to

services remotely using standard method invocation mechanisms and information exchange protocols. A depiction of the ImageMiner grid-enabled architecture is included in online supplemental material. In our current implementation, an image feature extraction module has been implemented as a caGrid service with a cluster backend and has been successfully deployed at the Cancer Institute of New Jersey, Emory, and Ohio State University. These sites serve as a test-bed for performance analysis. A client program can use the client application programming interface (currently bound to Java) of the service to compose and submit requests to the service and retrieve the results of the analysis. As part of the ongoing project, our team regularly implements additional analytical services, which are subsequently optimized for deployment.

SYSTEM PERFORMANCE

In this section, we present an experimental evaluation of the ImageMiner system, describing performance results from automated segmentation and content-based retrieval of image data. We also gauge performance from a high-performance, grid-enabled technology perspective.

Automated segmentation and CBIR

Results from analysis of TMA images using our library are given in figure 5. In figure 5A, we show the delineation of the outer boundary of individual tissue discs. After the whole disc masks have been obtained, a refined region-of-interest segmentation is executed using texton histograms to delineate the boundaries of stromal and epithelial sub-regions throughout the specimens.

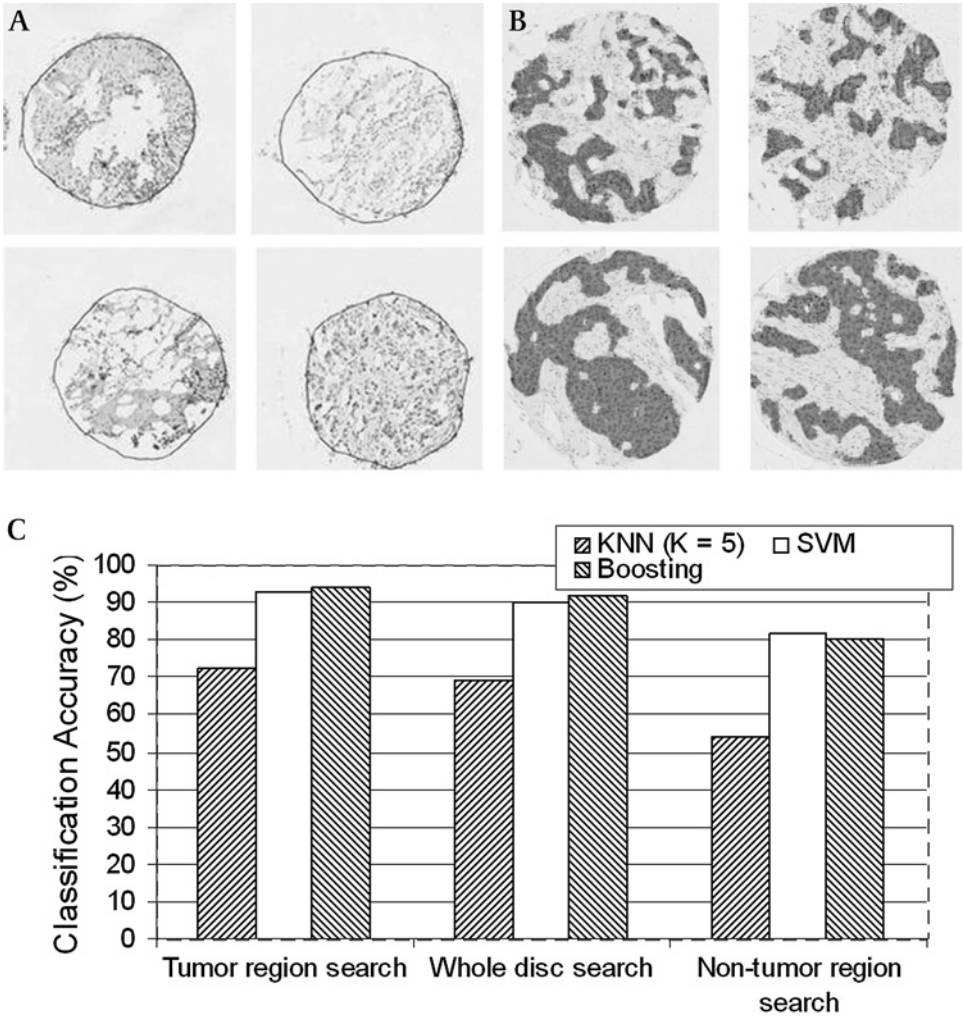

Figure 5 Analysis results from the image analysis library. (A) The delineation of the outer contours of the breast tissue disc. (B) The segmentation results for four representative breast cancer tissue discs showing delineation of tumor versus non-tumor regions using texton-based descriptor and an integral histogram approach. (C) The classification accuracy using different regions for searching. KNN, K nearest neighbors; SVM, support vector machine.

The Adaboost method is chosen as the classifier for segmentation. In the current implementation, the strong classifier is assembled from all of the weak classifiers to minimize the cost function representing the classification accuracy.

Some representative segmentation results are shown in figure 5B. When the segmentation results were compared with those that had been hand-drawn by a board-certified anatomical pathologist for 100 cancer tissue discs, the average pixel wise false positive rate was 6.62% and the average pixel wise false negative rate was 3.15%.

A comparative performance analysis of classification accuracy was carried out using three different algorithms: the K nearest neighbors (K=5), the soft margin SVM, and the boosting. The K nearest neighbor classifier is a simple classifier based on the Euclidean distance among feature vectors. Because the training data are not linearly separable, we chose the soft margin SVM, which allows training vectors to be on the wrong side of the support vector classifier with certain penalty. The key parameters, which affect the accuracy of soft margin SVM, are the penalty and the kernel. The penalty parameter was selected according to cross-validation (CV) errors in our case. For the kernel selection, we tested linear, polynomial, and Gaussian kernel, where the last one outperformed all the others. For the boosting method, instead of using Adaboost with a simple linear classifier as the weak classifier, we apply boosting using an eight-node classification and regression decision tree (CART) as the weak learner, which empirically provided higher accuracy. The number of nodes of CART can be selected using CV. The number of iterations was chosen as 40 to achieve satisfactory accuracy while avoiding over-fitting.

These algorithms were applied to the region masks of the tissue image. It is clear that the maximal margin classifiers, such as boosting and SVM, provides significantly better performance than simple classifiers such as K nearest neighbors. Figure 5C shows that using the tumor region mask provided appreciable improvements in classification and CBIR accuracy. Figure 5C also shows the results when queries are formulated using the texton histograms corresponding to three different sub-regions within the breast cancer specimens (tumor region alone, whole tissue sample, and non-tumor region alone). The tumor and non-tumor regions were automatically delineated using the segmentation algorithms.

As new therapy options become available, it has become increasingly important to distinguish among subclasses of pathology to determine which medications are appropriate and what level of risk is justified for a given patient population. The subtle visible differences exhibited by the digitized TMA specimens can sometimes give rise to inconsistent scoring and interpretation of results. Passing specimens through a reliable high-throughput, computer-based system, however, could potentially improve the accuracy with which patient populations are assigned to specific treatment regimens, improve the accuracy of prognosis, and reduce the costs of drug discovery.

High-performance computing

The current implementation of the ImageMiner analysis component supports the processing of multiple disc images on a computation cluster using a master–slave parallelization scheme. One of the computation nodes (or the head node of the cluster) is designated as the master node; the other nodes become slave nodes (processing nodes). When a slave node is idle, it sends a request for work to the master node. The master node fetches the next disc image from the disc and sends it along with processing parameters to the slave node.

We have experimented with the implementation in order to evaluate its performance. The evaluation was performed on a cluster system at the Ohio State University. Each node of the cluster has AMD Dual 250 Opteron CPUs, 8 GB DDR400 RAM, and 250 GB SATA hard drive. The computing nodes are connected through dual GigE Ethernet. The performance numbers obtained on 2 to 32 nodes are shown in figure 6. The graph in figure 6A displays the execution time when the number of processors is increased with the number of disc images fixed at 64. The processing time decreases almost linearly as more processors are added. The master–slave scheme with the demand-driven assignment of tasks to the processing nodes results in good computational load balance among the processing nodes. This is a feasible approach, since each imaged disc can be processed independently of other disc images. Figure 6B shows the execution time when both the number of disc images and the number of processing nodes are scaled simultaneously. The processing times remain almost the same across all data points, indicating good scalability of the system. As an extension of these experiments, we have also processed a dataset containing 624 imaged tissue discs, wherein each image originated from a different patient. These studies were carried out on 16, 32, and 48 nodes, respectively, resulting in execution times of 6, 3, and 2.1 h. Using this large dataset as a test ensemble showed that the use of parallel computing enabled us to execute in several hours what would otherwise require several days using a standard workstation.

The experimental evaluations reported in this article show that parallelization provides a viable mechanism for researchers to manipulate large datasets consisting of many imaged tissue samples. As the resolution and speed of digitized microscopy instruments improve, we can anticipate that modern

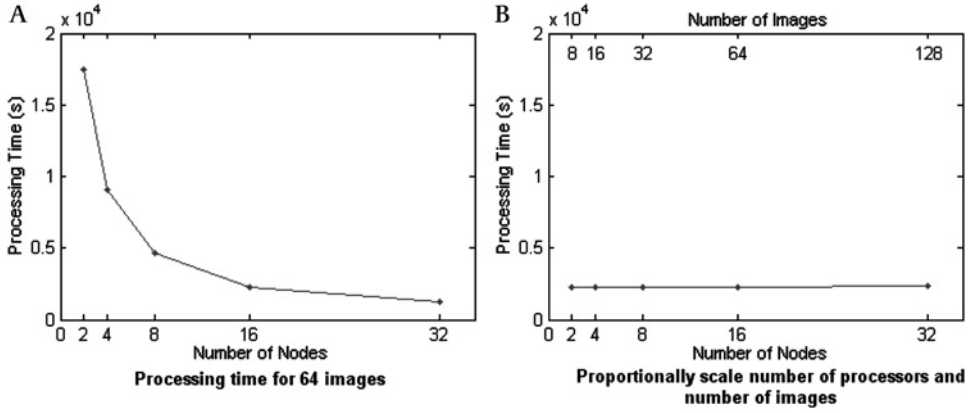

Figure 6 Performance assessment of the grid-enabled implementation. (A) Performance of processing 64 images while number of processors used varied from 2 to 32. (B) The number of processors and the number of images are scaled proportionately—that is, the number of images is doubled when the number of processors is doubled. Please notice that the y axis is at the same scale as in (A).

collaborative research studies will continue to generate increasingly large amounts of data for analysis. In addition, it is likely that, with advances in machine-learning and pattern recognition techniques, it will be desirable to analyze a given dataset or image ensemble many times, using a range of different algorithms in order to systematically identify which provides optimal results for a given application. In fact, such multi-analysis strategies may be invaluable for enabling investigators to objectively evaluate the efficacy of using different combinations and permutations of algorithmic modules and identify additional feature measurements that may be salient to a given classification. The processing power of a desktop machine is not sufficient for performing rapid, high-throughput analyses in such scenarios.

Grid-enabled feature extraction service

We have implemented a caGrid analytical service to provide remote access to the ImageMiner analysis component as well as to support federation of multiple instances of the component. The backend implementation of the analysis component can be a sequential program, which can be deployed on a single machine, or a parallel program (as described above), which can be deployed on a cluster machine. A client can submit a batch of input images and analysis parameters to the service remotely. After the processing of all the images is completed, the service returns the results (eg, histograms) to the client.

The service implementation facilitates access to remote resources without a client having to know the details of the resource or whether the analysis component is deployed on a cluster machine or not. The default client–service interaction protocols used in web services, and hence in caGrid, are SOAP and XML. The caGrid infrastructure provides a transfer service to address performance issues that may arise because of encoding large volumes of data in XML and using SOAP as a communication protocol. We used the caGrid transfer service in our ImageMiner services. Our experimental evaluation of the caGrid transfer service has shown that it achieves much lower transfer times than the XML and SOAP mechanisms for large datasets.

DISCUSSION

The primary focus of this research is to design, develop, and evaluate a software system for scoring and performing comparative analysis of expression patterns in TMAs. To summarize this work, we have identified several core requirements that are needed to facilitate multi-site, collaborative studies involving TMAs:

1. A suite of analysis methods that investigators can use for reproducible analysis of TMA images.
2. Both image data and analysis results (from expert reviews as well as computerized methods) should be managed in a coordinated way along with rich metadata about the images and results. This is an important requirement so that investigators can carry out additional mining of data, perform validation of analysis methods, and share information with collaborators.
3. Secure and remote access to data and tools should be implemented in collaborative projects.

These requirements have driven our design objectives. Towards that end, we have established a framework of image analysis, data management, and grid modules to support these functions and have evaluated their utility for performing comparative analysis of expression patterns in histology specimens.

Our system can detect and delineate tumor regions within imaged tissue discs and generate the texton histogram corresponding to the specified region of interest. This information can be used to formulate queries into a 'gold standard' database of cases and identify those tissue discs that contain lesions exhibiting staining signatures most similar to that of the query (ie, the test image). This capability would allow a researcher to retrieve data from multiple samples based on image characteristics. In addition, the PAIS model allows for queries that compare and contrast results obtained from one algorithm with results from other algorithms for algorithm evaluation—such queries may look for features and classifications obtained by multiple algorithms, or features that are different across algorithms. The annotation component of the combined ImageMiner and PAIS data model provides limited support for storing and managing semantic information. We plan to extend this component and enable more comprehensive semantic query capabilities in our future work to facilitate exploration of analysis results using semantic queries as well as better interoperability and sharing of the data.

By incorporating distributed execution capabilities, ImageMiner enables investigators to carry out large-scale analytical studies, which may not be feasible on a desktop machine. The current implementation achieves good performance for processing of TMA discs by taking advantage of the fact that TMA discs can be processed independently. In the next phase of development, this implementation will be extended to support parallel extraction and processing of disc images originating from a single TMA slide. It will require a more sophisticated parallelization of the registration to compensate for mechanical distortions during image acquisition and disc segmentation algorithms.

We have designed and implemented the ImageMiner system as a client–service system, using caGrid as the enabling middleware infrastructure for services. Having a standards-based and service architecture as the underlying foundation has a few advantages. First, the architecture facilitates access to remote resources, and the backend systems of these remote resources can be changed or upgraded with minimal impact to the existing ImageMiner clients, as long as the service interfaces remain the same. Second, the client program does not need to know whether the service has a single-machine backend or is deployed on a cluster system. This simplifies the implementation of the client. Third, ImageMiner resources can be federated and accessed along with other types of services by client applications. This enables investigators to leverage aggregate processing power on a high-performance system and scale to much larger volumes of data than can be managed on a workstation. Lastly, in our current implementation, the core services of the caGrid infrastructure can be leveraged (see table 1) for service discovery, security, federated queries, and workflows.

During the course of developing ImageMiner, the computational and imaging tools have been migrated to core research facilities at the Cancer Institute of New Jersey for use in ongoing investigative studies. They have been used to analyze microarrays consisting of cancers of the breast, head and neck, and prostate. As part of a recent study the automated software was used to quantify Beclin1 expression, which was shown to be predictive of autophagy.[62] Our team has since conducted a series of man–machine performance studies. In the first experiments, we used the TMA analysis tools to evaluate immunohistochemical staining intensity on imaged breast cancer TMA specimens comprising 1407 tissue cores. The results showed that the computer software algorithms achieved similar interpretations to those provided by a panel of three board-certified pathologists and was consistent with inter-pathologist

concordance. These results were presented at the 2010 Annual Conference of the United States and Canadian Academy of Pathology.[63] As an extension of those studies, we examined the expression patterns of a cohort array of several hundred tissue samples originating from patients with head and neck squamous cell carcinoma. Receiver operating characteristics curve analysis showed that the automated and manual scoring were generally consistent with area under the curve values of 0.9677 for Smad2 and 0.885 for Smad3.[70]

Limitations and future directions

As a result of the experience gained during the course of the project, we have learned the importance of sustaining a conscious effort to avoid the pitfalls of developing these tools in isolation and then later testing them in a clinical setting. Accordingly, our team is working closely with medical oncologists and surgical pathologists to enable us to remain focused on clinically relevant applications.

After completing these studies, we recognize the fact that the primary limitation of the image analysis algorithms currently used in ImageMiner is that a large labeled training dataset is required. To address this issue, our team is developing semi-supervised and online learning techniques to reduce the cited dependence.

While the parallel computing approaches described in this paper offer a solution for addressing the computational requirements of high-throughput analyses, such approaches also present limitations. For example, the high-level organization of the imaged TMAs lends itself to a bag-of-tasks type of parallelism, in which multiple imaged tissue samples can be processed concurrently on a parallel machine. This approach provides an efficient, yet relatively easy to implement, strategy for use of parallel computers to process images. In order to decrease the processing time for each image, our team has already begun to investigate the feasibility of automatically splitting each whole-slide image into image patches before processing. On the basis of experience gained during the course of these experiments, our team has begun to improve the efficiency of the data storage and management infrastructure used in ImageMiner. Specifically, we are exploring the use of a parallel database setup and compare the performance with a setup consisting of multiple database instances controlled by a frontend system in order to address the issue of conducting multi-analysis studies involving thousands of images and hundreds of algorithm variations.

In the current design of the system, multiple services can be federated at the client level—that is, the same client application can submit requests to multiple services, allowing a researcher to aggregate analysis resources from multiple sites in a collaborative study. However, the current system does not support load-balancing across services nor does it take into account computational power and load of individual services. An alternative approach would be to implement an aggregator service, which could control multiple analytical services using a demand-driven strategy, similar to the one implemented in the high-performance computing component, to distribute requests across multiple services.

Another key lesson learned during the course of these studies relates to the value of mitigating risks while trying to meet the primary design and development objectives of a project. To help mitigate risk in the ImageMiner project, the repository of imaged tissue samples and database has been designed and developed so that, in addition to supporting multi-modal indexing, querying, and retrieval of imaged tissue and correlated clinical data based on visual content, which was the primary goal of our efforts, the system also enables users to submit standard queries using the diagnosis of record, histological type, tumor grade, and image metrics for immunohistochemical staining intensities in order to retrieve the corresponding digitized arrays exhibiting those profiles.

To further mitigate risk in this project, the CBIR engine and interface was developed to achieve a user-friendly approach for conducting routine browsing and navigation through the datasets. For example, the interface enables users to interactively select any region or object of interest within a disc and initiate a query based on the texton signature of that particular subregion, tissue, or cell. The ImageMiner interface also enables users to refine queries by clicking on any one of the ranked retrievals in order to initiate subsequent (refined) queries using the selected ranked retrieval as the new query input image. By integrating these capabilities into the design to accompany the fully automated modules, we believe that the ImageMiner toolset and system is poised to achieve an added level of usability, versatility, and acceptance throughout a broader range of microscopy imaging applications.

We are currently expanding the scope of cancer types, tissues, and biomarkers under study and investigating the use of the system in performing sub-classifications in terms of the differential diagnosis, histological type, and tumor stage. These next steps will undoubtedly present new challenges and require the use of much more sophisticated statistical approaches for combining texton signatures and carrying out queries. Having recently established a multi-site consortium (Cancer Institute of New Jersey, Emory, Ohio State University, University of Pennsylvania, University of Texas) for performing iterative prototyping of the system throughout the course of its development, our team will begin to conduct retrospective and prospective performance analysis of real case scenarios and conditions.

CONCLUSIONS

Advances in imaging technologies have opened the door for investigators to employ high-resolution and high-throughput image data in their projects. While such data offer tremendous amounts of biomedical information, the size of datasets and labor-intensive nature of analyses create obstacles to more effective extraction and application of this information. Future advances in digital pathology will rely on the availability of reliable, portable algorithms and computational tools that can provide automated data management operations, perform high-throughput quantitative analysis, and support query and retrieval of image-based analysis results in collaborative environments. Our work aims to allow communities of end users to use standard, 'off-the-shelf', client-end software that can seamlessly access large image analysis libraries and grid-computing tools to reduce the obstacles of conducting collaborative research projects while supporting investigators through the efficient use of available resources.

Acknowledgments UMDNJ also thanks and acknowledges IBM for providing free computational power and technical support for this research through World Community Grid. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government.

Funding This research is funded, in part, by a grant from the NIH through contracts 5R01LM009239-04 and 3R01LM009239-03S2 from the National Library of Medicine and contracts SAIC/NCI#29XS154 and 9R01CA156386-05A1 from the National Cancer Institute. This work also received supported, in part, from NCI caBIG grants 79077CBS10, 94995NBS23, and 85983CBS43, the NHLBI R24 HL085343 grant, the NIH U54 CA113001, the NSF grants CNS-0403342 and CNS-0615155, by the NCI and

NIH under Contract No N01-CO-12400, by PHS Grant UL1RR025008 from the Clinical and Translational Science Award Program and by the US Department of Energy under Contract DE-AC02-06CH11357.

Ethics approval IRB exemption approved by UMDNJ IRB office.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Kononen J,** Bubendorf L, Kallioniemi A, *et al*. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 1998;**4**:844—7.
2. **Moch H,** Schraml P, Bubendorf L, *et al*. High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *Am J Pathol* 1999;**154**:981—6.
3. **Torhorst J,** Bucher C, Kononen J, *et al*. Tissue microarrays for rapid linking of molecular changes to clinical endpoints. *Am J Pathol* 2001;**159**:2249—56.
4. **Mucci NR,** Akdas G, Manely S, *et al*. Neuroendocrine expression in metastatic prostate cancer: evaluation of high throughput tissue microarrays to detect heterogeneous protein expression. *Hum Pathol* 2000;**31**:406—14.
5. **Matysiak BE,** Brodzeller T, Buck S, *et al*. Simple, inexpensive method for automating tissue microarray production provides enhanced microarray reproducibility. *Appl Immunohistochem Mol Morphol* 2003;**11**:269—73.
6. **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.
7. **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.
8. **Ayala G,** Wang D, Wulf G, *et al*. The prolyl isomerase Pin1 is a novel prognostic marker in human prostate cancer. *Cancer Res* 2003;**63**:6244—51.
9. **Camp RL,** Chung GG, Rimm DL. Automated subcellular localization and quantification of protein expression in tissue microarrays. *Nat Med* 2002;**8**:1323—7.
10. **Camp RL,** Dolled-Filhart M, King BL, *et al*. Quantitative analysis of breast cancer tissue microarrays shows that both high and normal levels of HER2 expression are associated with poor outcome. *Cancer Res* 2003;**63**:1445—8.
11. **Camp RL,** Neumeister V, Rimm DL. A decade of tissue microarrays: progress in the discovery and validation of cancer biomarkers. *J Clin Oncol* 2008;**26**:5630—7.
12. **Sanders TH,** Stokes TH, Moffitt RA, *et al*. Development of an automatic quantification method for cancer tissue microarray study. *Conf Proc IEEE Eng Med Biol Soc* 2009;**1**:3665—8.
13. **Rubin MA,** Zerkowski MP, Camp RL, *et al*. Quantitative determination of expression of the prostate cancer protein alpha-methylacyl-CoA racemase using automated quantitative analysis (AQUA): a novel paradigm for automated and continuous biomarker measurements. *Am J Pathol* 2004;**164**:831—40.
14. **Haedicke W,** Popper HH, Buck CR, *et al*. Automated evaluation and normalization of immunohistochemistry on tissue microarrays with a DNA microarray scanner. *Biotechniques* 2003;**35**:164—8.
15. **Rao J,** Seligson D, Hemstreet GP. Protein expression analysis using quantitative fluorescence image analysis on tissue microarray slides. *Biotechniques* 2002;**32**:924—6, 8—30, 32.
16. **Chen W,** Chu V, Hu J, *et al*. *ImageMiner: A Medical Image Analysis and Image Management UML Data Model. APIII: Advancing Practice*. Pittsburgh, PA: Instruction & Innovation Through Informatics, 2009.
17. **Chen W,** Meer P, Georgescu B, *et al*. Image mining for investigative pathology using optimized feature extraction and data fusion. *Comput Methods Programs Biomed* 2005;**79**:59—72.
18. **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**:1004—13.
19. **Thallinger GG,** Baumgartner K, Pirklbauer M, *et al*. TAMEE: data management and analysis for tissue microarrays. *BMC Bioinformatics* 2007;**8**:81.
20. **Viti F,** Merelli I, Galizia A, *et al*. Tissue MicroArray: a distributed Grid approach for image analysis. *Stud Health Technol Inform* 2007;**126**:291—8.
21. **Schnorrenberg F,** Pattichis C, Schizas C, *et al*. Content-based retrieval of breast cancer biopsy slides. *Technol Health Care* 2000;**8**:291—7.
22. **Guld MO,** Thies C, Fischer B, *et al*. A generic concept for the implementation of medical image retrieval systems. *Stud Health Technol Inform* 2005;**116**:459—64.
23. **Chen W,** Foran DJ, Reiss M. Unsupervised imaging, registration and archiving of tissue microarrays. *Proc AMIA Symp* 2002:136—9.
24. **Hadida-Hassan M,** Young SJ, Peltier ST, *et al*. Web-based telemicroscopy. *J Struct Biol* 1999;**125**:235—45.
25. **Wetzel A,** Andrews P, Becich M, *et al*. Computational aspects of pathology image classification and retrieval. *J Supercomputing* 1997;**11**:279—93.
26. **Zheng L,** Wetzel AW, Gilbertson J, *et al*. Design and analysis of a content-based pathology image retrieval system. *IEEE Trans Inf Technol Biomed* 2003;**7**:249—55.
27. **Wang JZ,** Nguyen J, Lo KK, *et al*. Multiresolution browsing of pathology images using wavelets. *Proc AMIA Symp* 1999:430—4.
28. **Gurcan MN,** Boucheron LE, Can A, *et al*. Histopathological Image Analysis: a Review. *IEEE Rev Biomed Eng* 2009;**2**:147—71.
29. **Yang L,** Chen W, Meer P, *et al*. Virtual microscopy and grid-enabled decision support for large-scale analysis of imaged pathology specimens. *IEEE Trans Inf Technol Biomed* 2009;**13**:636—44.
30. **Yang L,** Chen W, Meer P, *et al*. High throughput analysis of breast cancer specimens on the grid. *Med Image Comput Comput Assist Interv* 2007;**10**:617—25.
31. **Gurcan MN,** Kong J, Sertel O, *et al*. *Computerized pathological image analysis for neuroblastoma prognosis. AMIA Annual Symposium Proceedings/AMIA Symposium*. 2007.
32. **Foran DJ,** Winkelmann DA, Goodell LA, *et al*. A network-based prototype for interactive telemedicine & automated management of distributed, clinical databases. *J Clin Eng* 1996;**21**:383—91.
33. **Comaniciu D,** Meer P, Foran DJ. Image-guided decision support system for pathology. *Mach Vis Appl* 1999;**11**:213—24.
34. **Comaniciu D,** Meer P. *Cell image segmentation in diagnostic pathology*. Springer, 2001.
35. **Chen W,** Foran DJ. Advances in cancer tissue microarray technology: towards improved understanding and diagnostics. *Anal Chim Acta* 2006;**564**:74—81.
36. **Chen W,** Reiss M, Foran DJ. A prototype for unsupervised analysis of tissue microarrays for cancer research and diagnostics. *IEEE Trans Inf Technol Biomed* 2004;**8**:89—96.
37. **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.
38. **Barsky S,** Gentchev L, Basu A, *et al*. Use and validation of epithelial recognition and fields of view algorithms on virtual slides to guide TMA construction. *Biotechniques* 2009;**47**:927—38.
39. **Shah NH,** Rubin DL, Espinosa I, *et al*. Annotation and query of tissue microarray data using the NCI Thesaurus. *BMC Bioinformatics* 2007;**8**:296.
40. **Viti F,** Merelli I, Caprera A, *et al*. Ontology-based, Tissue MicroArray oriented, image centered tissue bank. *BMC Bioinformatics* 2008;**9**(Suppl 4):S4.
41. **Sharma-Oates A,** Quirke P, Westhead DR. TmaDB: a repository for tissue microarray data. *BMC Bioinformatics* 2005;**6**:218.
42. **Ellisman M,** Peltier S. Medical data federation: the biomedical Informatics research network. In: Foster I, Kesselman C, eds. *The Grid 2: Blueprint for a New Computing Infrastrcture*: Elsevier, 2003:109—20.
43. **Grethe JS,** Baru C, Gupta A, *et al*. Biomedical informatics research network: building a national collaboratory to hasten the derivation of new understanding and treatment of disease. *Stud Health Technol Inform* 2005;**112**:100—9.
44. **Amendolia SR,** Brady M, McClatchey R, *et al*. MammoGrid: large-scale distributed mammogram analysis. *Stud Health Technol Inform* 2003;**95**:194—9.
45. **Rogulin D,** Estrella F, Hauer T, *et al*. *A Grid Information Infrastructure for Medical Image Analysis. Distributed Databases and processing in Medical Image Computing Workshop (DiDaMIC-2004)*. 2004.
46. **Solomonides A,** McClatchey R, Odeh M, *et al*. MammoGrid and eDiamond: Grids Applications in Mammogram Analysis. *Proceedings of the IADIS International Conference: e-Society 2003*. 2003:1032—3.
47. **Kakazu KK,** Cheung LW, Lynne W. The Cancer Biomedical Informatics Grid (caBIG): pioneering an expansive network of information and tools for collaborative cancer research. *Hawaii Med J* 2004;**63**:273—5.
48. **Fenstermacher D,** Street C, McSherry T, *et al*. *The Cancer Biomedical Informatics Grid (caBIG™)*. 2005.
49. **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**:1910—16.
50. **Oster S,** Langella S, Hastings S, *et al*. caGrid 1.0: an enterprise Grid infrastructure for biomedical research. *J Am Med Inform Assoc* 2008;**15**:138—49.
51. **Yang L,** Meer P, Foran DJ. Unsupervised segmentation based on robust estimation and color active contour models. *IEEE Trans Inf Technol Biomed* 2005;**9**:475—86.
52. **Trucco AV.** *Introductory Techniques for 3-D Computer Vision*. 1st edn. Prentice Hall, 1998.
53. **Julesz B.** A theory of preattentive texture discrimination based on first-order statistics of textons. *Biol Cybern* 1981;**41**:131—8.
54. **Leung T,** Malik J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int J Comp Vis* 2001;**43**:29—44.
55. **Cula OG,** Dana KJ. *Compact Representation of Bidirectional Texture Functions*. 2001:1041.
56. **Heeger D,** Bergen J. *Pyramid-Based Texture Analysis/Synthesis*. ACM, 1995:238.
57. **Leung T,** Malik J. *Recognizing Surfaces Using Three-Dimensional Textons*. 1999:1010.
58. **Schmid C.** Constructing models for content-based image retrieval. *IEEE Conf Comput Vis Pattern Recogn*: **2**; 2001.
59. **Georgescu B,** Shimshoni I, Meer P. *Mean shift based clustering in high dimensions: a texture classification example. 9th International Conference on Computer Vision*. Nice, France, 2003.
60. **Foran DJ,** Yang L, Tuzel O, *et al*. A caGRID-enabled, learning based image segmentation method for hisopathology specimens. *Proc IEEE Int Symp Biomed Imaging* 2009;**6**:1306—9.
61. **Freund Y,** Schapire RE. A decision-theoretic generalization of online learning and an application to boosting. *J Comp Sys Sci* 1997;**55**:119—39.
62. **DiPaola RS,** Dvorzhinski D, Thalasila A, *et al*. Therapeutic starvation and autophagy in prostate cancer: a new paradigm for targeting metabolism in cancer therapy. *Prostate* 2008;**68**:1743—52.
63. **Goodell LA,** Chen W, Javidian P, *et al*. *Use of Computer Assisted Analysis To Facilitate Tissue Microarray Interpretation. United States and Canadian Academy of Pathology 2010 Annual Meeting*. Washington, DC, 2010.

64. **Kumar VS,** Rutt B, Kurc T, *et al*. *Large Image Correction and Warping in a Cluster Environment. Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*. Tampa, Florida: ACM, 2006.
65. **Beynon M,** Chang CL, Catalyurek U, *et al*. Processing large-scale mufti-dimensional data in parallel and distributed environments. *Parallel Comput* 2002;**28**:827—59.
66. **Beynon MD,** Kurc T, Catalyurek U, *et al*. Distributed processing of very large datasets with DataCutter. *Parallel Comput* 2001;**27**:1457—78.
67. **Wang F,** Pan T, Kurc T, *et al*. *Unified Modeling of Image Annotation and Markup. APIII: Advancing Practice, Instruction & Innovation Through Informatics*. Pittsburgh, PA, 2009.
68. **Channin DS,** Mongkolwat P, Kleper V, *et al*. The caBIG annotation and image markup project. *J Digit Imaging* 2010;**23**:217—25.
69. **Channin DS,** Mongkolwat P, Kleper V, *et al*. The annotation and image mark-up project. *Radiology* 2009;**253**:590—2.
70. **Xie W,** Chen W, Foran DJ, *et al*. *Alterations of TGFβ/Smad Signaling in Human Head and Neck Squamous Cell Carcinomas*, 2010. In review.