EMORY UNIVERSITY

Center for Comprehensive Informatics

Technical Report

A Data Model and Database for High-resolution Pathology Analytical Image Informatics

Fusheng Wang
Jun Kong
Lee Cooper
Tony Pan
Tahsin Kurc
Wenjin Chen
Ashish Sharma
Cristobal Niedermayr
Tae W. Oh
Daniel Brat
Alton B. Farris
David Foran
Joel Saltz

CCI-TR-2011-2
July 26, 2011

A Data Model and Database for High-resolution Pathology Analytical Image Informatics

Fusheng Wang[+], Jun Kong[+], Lee Cooper[+], Tony Pan[+], Tahsin Kurc[+], Wenjin Chen[*], Ashish Sharma[+], Cristobal Niedermayr[+], Tae W. Oh[#], Daniel Brat[$], Alton B. Farris[$], David Foran[*], Joel Saltz[+]

[+]Center for Comprehensive Informatics, Emory University
[*]The Cancer Institute of New Jersey, UMDNJ-Robert Wood Johnson Medical School
[#]Department of Computer Information Systems, Georgia State University
[$]Department of Pathology and Laboratory Medicine, School of Medicine, Emory University

Abstract

The systematic analysis of imaged pathology specimens often results in a vast amount of morphological information at both the cellular and sub-cellular scales. The information generated by this process has tremendous potential for providing insight regarding the underlying mechanisms of disease onset and progression. While microscopy scanners and computerized analysis are capable of capturing and analyzing data rapidly, microscopy image data remains underutilized in research and clinical settings. One major obstacle which tends to reduce wider adoption of these new technologies throughout the clinical and scientific communities is the challenge of managing, querying, and integrating the vast amounts of data resulting from the analysis of large digital pathology datasets. This paper presents a data model, which addresses these challenges, and demonstrates its implementation in a relational database system.

Context:

This paper describes a data model, referred to as Pathology Analytic Imaging Standards (PAIS), and a database implementation, which are designed to support the data management and query requirements of detailed characterization of micro-anatomic morphology through many interrelated analysis pipelines on whole slide images and tissue microarrays.

Aims:

Aim 1: Development of a data model capable of efficiently representing and storing virtual slide related image, annotation, markup, and feature information.

Aim 2: Development of a database, based on the data model, capable of supporting queries for data retrieval based on analysis and image metadata, queries for comparison of results from different analyses, and spatial queries to assess relative prevalence of features and classified objects and to retrieve collections of segmented regions and features.

Settings and Design:

The work described in this paper originated from the challenges associated with characterization of micro-scale features for comparative and correlative analyses involving whole slides tissue images and tissue microarrays. Technologies for digitizing tissues have advanced significantly in the past decade. Slide scanners are capable of producing high-magnification, high-resolution images from whole slides and tissue microarrays within several minutes. Hence, it is becoming increasingly feasible for basic, clinical, and translational research studies to produce thousands of whole slide images. Systematic analysis of these large datasets require efficient data management support for representing and indexing results from hundreds of interrelated analyses generating very large volumes of quantifications such as shape and texture and of classifications of the quantified features.

Methods and Material:

We have designed a data model and a database to address the data management requirements of detailed characterization of micro-anatomic morphology through many interrelated analysis pipelines. The data model represents virtual slide related image, annotation, markup and feature information. This set of information includes a) context relating to specimen preparation, special stains etc, b) human observations involving pathology classification and characteristics, c) algorithm and human-described segmentations, features and classifications, and d) a description of the computation being carried out and identification of input and output datasets. The database supports a

wide range of queries. Typical queries include: a) queries needed to obtain combinations of image and metadata required for certain analytic procedures, b) queries needed to compare results obtained from different algorithms and to compare algorithm results with human annotations and markups, c) spatial queries, such as those used to assess relative prevalence of features or classified objects in various portions of slides or to assess spatial coincidence of combinations of features or objects, d) queries needed to support selection of collections of segmented regions, features, objects within and across virtual slides used to carry out machine learning or content based information retrieval algorithms.

Results:

We currently have three databases running on a Dell PowerEdge T410 server with CentOS 5.5 Linux operating system. The database server is IBM DB2 Enterprise Edition 9.7.2. The set of databases consists of 1) a tissue microarray (TMA) database containing image analysis results from 4740 cases of breast cancer, with 641MB storage size; 2) an algorithm validation database, which stores markups and annotations from two segmentation algorithms and two parameter sets on 18 selected slides, with 66GB storage size; and 3) an in silico brain tumor study database comprising results from 307 TCGA slides, with 365GB storage size. The latter two databases also contain human generated annotations and markups for regions and nuclei.

The data model and the database infrastructure are being employed in applications that 1) implement a systematic approach for validating image segmentation algorithms; 2) employ the database to investigate whether glioma morphology correlates with gene expression data; and 3) investigate relationships between microscopic and macroscopic features.

Conclusions:

Our experience with the in silico study of brain tumors has shown that data sets resulting from analyses of digitized slides can be extremely large. Modeling and managing pathology image analysis results in databases provides immediate benefits on the value and usability of data through standardized data representation, data normalization, and

semantic annotation. The database provides powerful query capabilities, which are otherwise difficult or cumbersome to support by other approaches such as programming languages. Standardized, semantic annotated data representation and interfaces also make it possible to more efficiently share image data and analysis results.

Key-words:

Digitized slides, data models, databases, image analysis.

Key Messages:

Effective use of large microscopy image datasets in research requires the application of many interrelated analyses for the detailed characterization of morphological characteristics. Modeling and managing image analysis results in databases provides powerful capabilities to store and index analysis results efficiently and to perform complex queries for data exploration, analysis comparison, and analysis validation.

Introduction:

High-resolution digitized pathology images contain a wealth of spectral and morphologic features related to the microanatomy of the tissues under study. Examination of the subtle differences exhibited by diseased tissue at the cellular and sub-cellular levels has potential to improve characterization of the histologic type, stage, prognosis, and likely treatment response. For example, the morphologies of cell nuclei, their infiltrative patterns, the development and extent of new blood vessels, and degree of necrosis, are all measurable features of significant interest in the study of diffuse gliomas. The classifications of brain tumor nuclei based on morphology can be studied to look for genetic correlations, create image-based computational biomarkers, and assess patient survival.

Technologies for digitizing microscopy have advanced significantly in the past decade. Slide scanners are capable of producing high-magnification, high-resolution images from whole slides and tissue microarrays within several minutes. It is rapidly becoming feasible for even medium-scale studies to routinely generate thousands of whole slide images. At this scale, the subjective process of manually capturing and classifying histopathological features is both time consuming and likely to increase observer variability and errors [1].

Computerized image analysis offers a means of rapidly carrying out quantitative, reproducible measurements of micro-anatomical features in high-resolution pathology images and large image datasets. Nevertheless, image data is often an underutilized resource in biomedical research, since reliably analyzing even moderate numbers of virtual slides leads to a formidable information synthesis and management problem. As we shall describe in the next section, systematic analysis of large-scale image data can involve many interrelated analyses on hundreds or thousands of images, generating billions of quantifications such as shape and texture, as well as classifications of the quantified features.

In this paper, we describe a data model, referred to as Pathology Analytic Imaging Standards (PAIS), and a database implementation, which are designed to support the data management and query requirements of detailed characterization of micro-anatomic morphology through many interrelated analysis pipelines on whole slide images and tissue microarrays.

The data model represents virtual slide related image, annotation, markup and feature information. This set of information includes a) context relating to patient data, specimen preparation, special stains, etc.; b) human observations involving pathology characteristics; and c) algorithm and human-described segmentations, features, and classifications. Moreover, it supports the provenance of the markups and annotations through a description of the computation being carried out and an identification of input and output datasets.

The database supports a wide range of queries. Typical queries include: a) those needed to obtain combinations of image and metadata required for certain analytic procedures, b) those needed to compare results obtained from different algorithms and to compare algorithm results with human annotations and markups, c) spatial queries, such as those used to assess relative prevalence of features or classified objects in various portions of slides or to assess spatial coincidence of combinations of features or objects, d) queries needed to support selection of collections of segmented regions, features, objects within and across virtual slides used to carry out machine learning or content based information retrieval algorithms.

The data model and the database have been successfully used for algorithm validation and integrative in silico study of brain tumors as shall be presented later in the paper. The PAIS data model is capable of capturing detailed markups and annotations, while allowing for an efficient implementation using the relational database technology. We have shown that the PAIS database enables more expressivity and efficiency in retrieving, comparing, and mining vast amounts of results than those achieved using a programmatic approach (i.e., MATLAB scripts) only.

Background:

We will use a research project underway at the In Silico Brain Tumor Research Center (ISBTRC) as an example to illustrate data management challenges that arise from analyzing large numbers of high-resolution microscopy images. The ISBTRC is a cancer Biomedical Informatics Grid (caBIG®) In Silico Research Center of Excellence established as a collaboration of four institutions: Emory University, Thomas Jefferson University, Henry Ford Hospital, and Stanford University. It conducts integrative in silico study of diffuse glioma brain

tumors using Pathology image data, omics data, Radiology image data, and clinical outcome data obtained from The Cancer Genome Atlas (TCGA)[2] and REMBRANDT [3] and from the partner institutions. The center develops techniques that extract and correlate information from these complementary data types in order to improve disease classification and better understand biology of disease progression.

The example project is the characterization of micro-anatomic elements, such as cells and nuclei, in whole slide tissue images. The morphology of these elements varies in shape and texture across different classes and grades of gliomas. For example, nuclei appear to be round shaped with smooth regular texture in oligodendrogliomas, whereas they are generally more elongated with rough and irregular texture in astrocytomas. However, there are also many nuclei that appear to be transitions and are difficult to classify. The goal of the project is to use image analysis algorithms in whole-slide scans linked to patient outcome and genomic data to better define such structures in order to improve the classification and grading of these diseases.

The project has already gathered over 700 whole slide images of diffuse gliomas (219 images at 20X objective magnification and 517 at 40X), derived from the TCGA repository, Henry Ford Hospital, and Emory University, with a long term goal of expanding the studies to approximately 3500 slides from about 700 patients in the course of the project. With this many slides, it is not feasible to manually examine each slide image, mark microscopic objects, and annotate them. Computerized analysis of the images is necessary to extract, quantify, and classify micro-anatomic features. The effectiveness of a computer analysis pipeline, however, depends on many factors including the nature of the histological structures being segmented, the classifications being performed, and sample preparation and staining. Thus, detailed computer-aided characterization of brain tumor morphology requires coordinated use of many interrelated analysis pipelines on a large number of images. Results produced from multiple runs by varying the algorithms and input parameters of the analysis pipelines can help determine priority pipelines for a particular set of images and study objectives. The priority pipelines are executed on the image dataset and are further refined by comparing and correlating the results in order to increase the accuracy of output. This strategy leads to a very challenging data management problem.

Whole slide brain images are roughly $5x10^4$ by $5x10^4$ pixels at 20X objective magnification. Brain tumor image analysis algorithms segment and classify 10^5 to 10^7 cells in each virtual slide. Classification categories include a variety of classes of brain tumor cells, several categories of normal brain cells (astrocytes, oligodendrocytes, microglia and neurons), endothelial cells, red blood cells, and macrophages. Brain tumor tissue analyses can encompass discrimination from normal tissue, analysis of tumor cell density, classification of nuclei, quantification of mitotic figures, identification and classification of angiogenesis, and identification of differing types of necrosis, including the pseudopalisades that are often seen around necrosis in glioblastoma. Reliable identification of subcellular structures, such as mitotic figures in brain tumor cells, is done through additional processing in cells or regions identified as being brain tumor. Identification and classification of angiogenesis and pseudopalisades requires a synthesis of regional texture analysis, cell segmentation, and classification along with ability to recognize and characterize larger scale histological structures. A systematic analysis of datasets consisting of thousands of images, therefore, can result in classification of roughly tens of billions to trillions micro-anatomic structures. The process of classifying a given cell involves roughly 10-100 features describing morphometry, texture, and stain quantification. An in-depth analysis even if limited to classifying the constituent cells of the specimens can easily encompass a very large amount of features. These data sets need to be stored and indexed so that investigators can query and interrogate the results to search for patterns and correlations as well as validate and refine computer analysis algorithms.

Software systems and data models have been developed for managing and accessing digitized microscopy images and large image datasets. The virtual microscope system[4, 5] is designed to support the storage, retrieval, and processing of very large microscopy images on high-performance systems. The Open Microscopy Environment (OME) project [6] has developed a data model and a database system that can be used to represent, exchange, and manage image data and metadata. The OME provides a data model of common specification for storing details of microscope setup and image acquisition. Cell-Centered Database (CCDB) [7, 8] is a system and data model developed to capture image analysis output, image data, and information on the specimen preparation and imaging conditions that generated the image data.

The CCDB implements an ontology link to support semantic queries and data sources federation. The ImageMiner system[9] implements capabilities for content-based image retrieval for tissue microarray datasets. The Bio-Imaging Semantic Query User Environment (BISQUE) [10] and associated tools like the Digital Notebook allow a biologist to capture image datasets and associated experiment metadata and manage them in a BISQUE database, which is built on a relational database system. Content-based image retrieval approaches and systems have also been implemented to support rich queries on image data[11-16]. One of the early systems with application in biomedicine employed methods to express the global characteristics of images as a measure of the Gleason grade of prostate tumors[17, 18]. Another system, developed by Wang et al.[19], indexes image block segments at different scales by dividing the original image into smaller overlapping regions. It employs integrated region matching distances to characterize images and allows users to browse the regions of a matched image at different scales. The use of parallel and distributed computing for analysis has increased over the years, enabling researchers to process image datasets quickly and generate large volumes of analysis results. Yang et al. has demonstrated a distributed system for computer-aided analysis of digitized breast tissue specimens[20]. Gurcan et al. employed parallel and distributed computing to efficiently support automated characterization of Neuroblastoma using a multi-resolution grid based framework[21].

Most of the previous work in microscopy image data management is targeted at remote access to and sharing of microscopy images and annotations, and is not primarily designed to handle large volumes of analysis results and large images for correlative studies and algorithm validation. Our work, on the other hand, targets the following closely interrelated tasks: 1) to systematically manage, query and analyze results produced by data analyses composed from large numbers of interrelated algorithms, 2) to compare results produced by workflows consisting of cascades of multiple algorithms, 3) to efficiently manage resulting datasets that in aggregate can contain billions of imaging derived features, and 4) to support histological feature query and analysis patterns.

DICOM Working Group 26[1] is developing a DICOM based standard for storing microscopy images. The metadata in this model captures information such as patient, study and equipment information. Image tiles are managed as series and the mapping relationship is represented in an XML format. However, the metadata is limited and not easy to extend to efficiently represent and manage image analysis results. Similarly, DICOM Structured Report standard[22] has been used to model and store image annotations and markups in DICOM. The standard does not provide an approach for managing and querying data. Annotation and Image Markup (AIM) [23] is a data model developed in the caBIG® program. It is designed to facilitate standardization for image annotation and markup for radiology images. AIM is motivated by the characteristics and requirements of Radiology imaging applications. Pathology images have characteristics that are not taken into account by the current AIM model. For example, pathology image annotations are done at microscopic levels in multiple granularities. AIM, on the other hand, takes a "flat" structure on annotations and markups, where fine-grained annotations and markups and their relationships are difficult to represent.

Subjects and Methods:

Data Model

The Pathology Analytical and Imaging Standards (PAIS) model is designed to provide a flexible, efficient, and semantically enabled data model for pathology image analysis and characterization. The logical model of PAIS is defined in Unified Modeling Language (UML), and consists of 62 classes and associations between them. The major components of the model (main classes and relationships, not including attributes) are shown in Figure 1.

The *ImageReference* class provides metadata that describes an image or a group of images, which have been used as the base for markups and annotations. This class can be used to identify and retrieve the relevant images from an image archive. The metadata includes the resolution of the image in microns/pixel, the z-axis resolution and coordinate, if available. The subclasses derived from the ImageReference class reference specific types of images such as

[1] http://medical.nema.org/DICOM/minutes/WG-26/

whole slide images (WholeSlideImageReference) and tissue microarray (TMA) images (TMAImageReference). The DICOMImageReference class is used to maintain image metadata in the case images are stored as DICOM images. Note that there can be multiple ImageReference instances for multi-sliced images or multi-modality registrations. The ImageReference class is associated with Subject, Specimen, AnatomicEntity, and Equipment classes, which collectively capture metadata about how the corresponding image has been acquired.

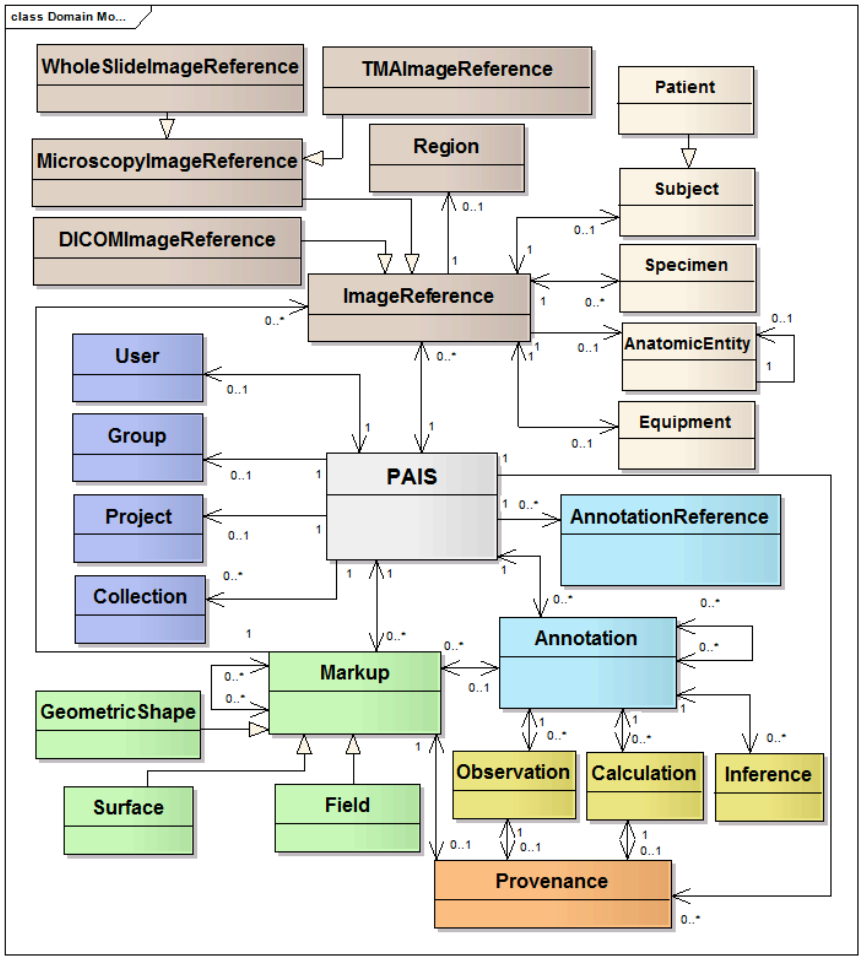

Figure 1. PAIS object model.

The *Region* class is used to identify the area of interest from an image (e.g., a specific tile from a whole slide image, or an area that contains a disc image in TMA image) for the purpose of

markup and annotation. It also captures the relative zoom resolution of the region over the original image. The coordinate reference of the markups on the image can be either local – relative to the region, or global – relative to the original image. The units used for markups or measurements are mostly based on pixels: The values of coordinate, width, and length are represented in pixels; square pixels are the units of measurement for area, while resolution uses microns/pixel.

The set of *Project*, *Group*, *User*, and *Collection* classes stores information related to the study and analysis experiments. The Project class represents the study being conducted. A Group is the collection of scientist and/or clinician users conducting the study. A User is the person who has marked and annotated the images in the study – a user may be associated with multiple annotated images and objects, and multiple users may annotate an image or image region. A Collection is a group of items of the same type. For example, when an experiment is done to validate algorithms against human assessment, all the PAIS instances to be compared are in the same collection. Similarly, results from the same algorithm but obtained using different input parameters applied on the same image can be grouped into the same collection. One PAIS instance can have multiple collections.

The Markup class is used to delineate a spatial region in images and represents a set of values derived from pixels. Markup symbols are associated with one or more images and can be in the form of geometric shapes, surfaces, and fields. Geometric shapes can be points, lines, polylines, polygons, rectangles, circles, and ellipses. We employ the representation format of Scalable Vector Graphics (SVG)[2] for markups. For example, we use closed points to represent polygons. They result in compact representations and save disk space. Since major Web browsers natively support SVG, SVG based markups can easily be displayed as overlays on top of images. 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.

---

[2] http://www.w3.org/Graphics/SVG/

The Annotation class associates semantic meaning to markup entities through coded or free text terms that provide explanatory or descriptive information. There are three types of annotations: Observation, Calculation, and Inference. Observation holds information about interpretation of a markup or another annotation entity. Observations can be quantified based on different measure scales such as ordinal and nominal scales. Calculation stores information about the quantitative results from mathematical or computational calculations, such as Scalar, Array, Histogram, and Matrix. Inference is used to maintain information about disease diagnosis derived by observing imaging studies and/or medical history. PAIS only captures image based annotations. It uses the AnnotationReference class to link to external annotations such as molecular or genetic annotations. This provides endpoints for queries integrating data from different data types.

The Provenance class captures the derivation history of a markup or annotation, including algorithm information, parameters, and inputs. Such information is critical for validating analysis approaches and comparing algorithms.

To identify objects and relationships within a PAIS instance, an id attribute (unique in the PAIS instance scope) is associated with each PAIS object. Globally unique ids (UIDs) are also associated with objects that could potentially be shared across multiple PAIS object instances, such as Specimen, Project, Group, Collection, Annotation and Markup. The UIDs are based on the UUID standard [8].

Database Implementation

We have implemented a database infrastructure (Figure 2) to manage microscopy analysis results expressed in the PAIS model. The PAIS Data Repository component encapsulates the database and the data loading and query subcomponents. The database is designed to support queries on both metadata and spatial features for data retrieval, comparative data analysis, and algorithm validation. The types of queries include:

- Queries involving combinations of image and algorithm metadata to retrieve analysis results. An example of this type of queries is: Find all markups with area between 200 and 500 square pixels, and eccentricity between 0 and 0.5 on image ”astroII”.

- Queries to compare results obtained from different algorithms and to compare computer-generated results with human annotations and markups. Examples are: Find the average glioma grades of nuclei segmented and calculated by algorithm ”NSMORPH” for each human segmented region on image ”OligoIII” grouped by human classification. Compare the average “Sum Canny Pixels” feature calculated from algorithms between the “Proneural” and “Mesenchymal” tumor subtypes.

- Spatial queries, such as those used to assess relative prevalence of features or classified objects in various portions of slides or to assess spatial coincidence of combinations of features or objects. Examples of spatial queries are: Find all segmented nuclei from algorithm “NSMORPH” with parameter set 1 in the region segmented by human as ”Astro grade II” on image “gbm0”. Find nuclei in region [100,100:1000,1000] that are detected by Algorithm “NSMORPH-1” and that intersect with those detected by Algorithm “NSMORPH-2” on image ”OligoIII”.

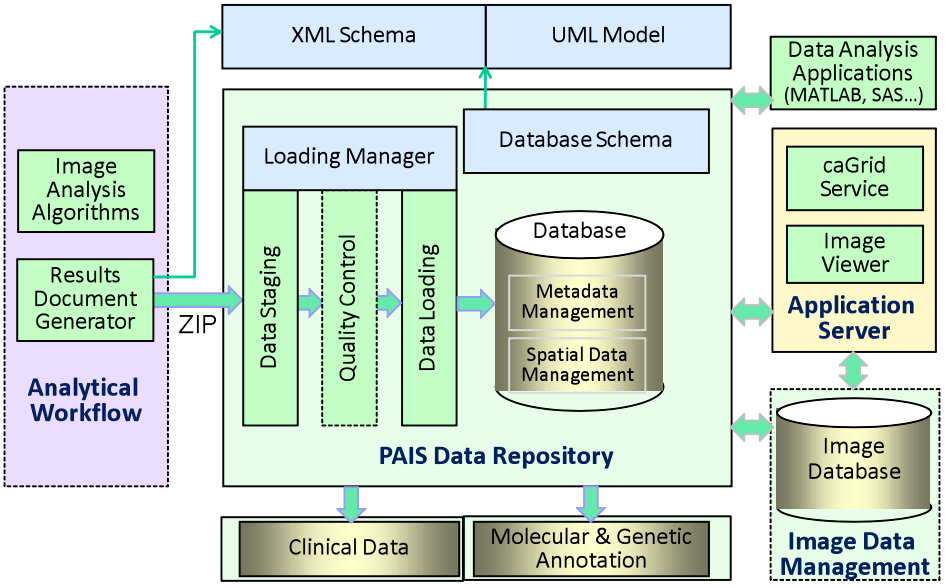

Figure 2. PAIS database implementation. The architecture includes analytical workflow, PAIS data repository, application server, image database, and data analysis applications.

We have used a relational database backend in our implementation, although results expressed in the PAIS model are exchanged using XML documents. Our performance evaluation has showed that the relational database approach is more efficient than a native XML based approach in our case for a wide range of queries. The PAIS database is comprised of a set of tables mapped from the PAIS logical model. The database schema has 1) a data staging table for storing compressed PAIS documents submitted from clients and tracking jobs of data mapping; 2) metadata tables for storing metadata on images, subjects, projects, and experiments; 3) spatial tables for storing markup shape objects; 4) calculation tables for computed image features – multiple calculation tables provided for different feature sets; 5) observation tables (nominal or ordinal) for annotations; 6) vocabulary tables to define the common data elements used for calculations, observations and anatomic entities; 7) provenance tables for storing algorithm information and analysis parameters; and 8) application tables such as validation tables for storing pre-computed markup intersection information

between different methods. The database also provides a set of extended functions and stored procedures for manipulating data.

The database implementation uses the IBM DB2 Universal Database server with the DB2 Spatial Extender as the underlying database system. We have chosen IBM DB2 since it is available free of charge for research and education, and provides integrated support for spatial data types and queries through the spatial extender component. To support efficient management and query of spatial information, we model and manage markup objects as spatial objects as supported by the spatial extension of DB2. We also employ in queries dozens of spatial functions implemented in DB2 such as spatial relationship functions and functions that return information about properties and dimensions of geometries. Many of our spatial queries are different from traditional GIS queries. We have implemented additional optimizations to reduce query execution times. Data are clustered by image and tiles. With such clustering, queries at the level of tiles can be efficiently supported with minimal I/O request.

To enable convenient data exchanging between analysis programs and the PAIS database, we use XML based representation for the PAIS model, based on an XML schema derived from the logical model. PAIS XML documents are generated via the PAIS document generator by each client application (image analysis applications and human markup and annotation applications). To reduce the size for processing, PAIS documents are often generated on partitioned regions such as tiles, and different PAIS document instances from different regions of the same image will share the same document unique identifier. For efficient data transportation, PAIS XML documents are further compressed into zip files.

When a PAIS XML document zip file is received by the system, an entry is created in the staging table for storing the zip file. The data loading manager on the database server parses each document in the staging table with an efficient event based XML parser, and maps the contents of the document to the database tables by generating SQL batch insertion requests. The data uploads are optimized in a batch and resource-efficient manner, and populate relational and spatial tables. To provide smooth workflow, we also track the loading status of each document, and log any exceptions in the workflow. This could guarantee continuous workflow under error conditions.

Results:

We currently have three PAIS databases running on a Dell PowerEdge T410 server with CentOS 5.5 Linux operating system. The database server is IBM DB2 Enterprise Edition 9.7.2. The set of databases consists of 1) a tissue microarray (TMA) database containing image analysis results from 4740 cases of breast cancer, with 641MB storage size; 2) an algorithm validation database, which stores markups and annotations from two segmentation algorithms and two parameter sets on 18 selected slides, with 66GB storage size; and 3) an in silico brain tumor study database comprising results from 307 TCGA slides, with 365GB storage size. The latter two databases also contain human generated annotations and markups for regions and nuclei.

Applications

We present three applications that demonstrate the use of PAIS for algorithm validation, in silico research, and correlative analysis. The first application implements a systematic approach for validating image segmentation algorithms. Sampling and result comparison queries in this application are supported by the PAIS database. The second application employs the PAIS database to investigate whether glioma morphology correlates with gene expression data. The third application is a new project investigating relationships between microscopic and macroscopic features.

Application 1: Algorithm Validation

The evaluation and validation of image analysis algorithms is an important component in imaging studies, because the efficacy of an analysis pipeline will generally be dependent on the characteristics of specimens and images used in the study, the types of algorithms employed, and the study's objectives. We have developed a systematic framework and workflow (Figure 3)

for evaluating results generated by computer algorithms employed in studies carried out in ISBTRC.

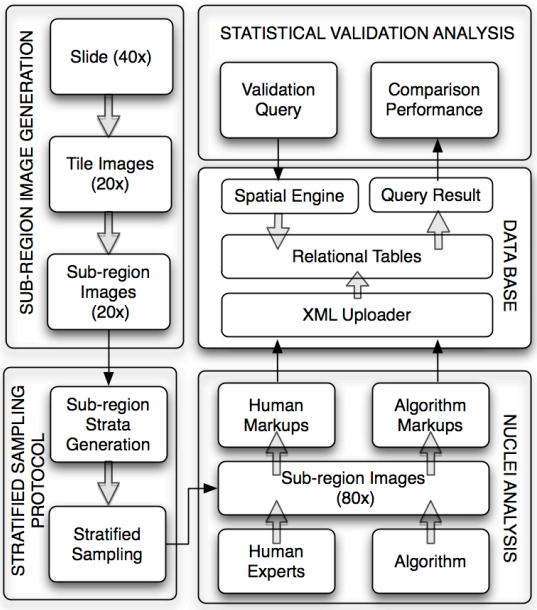

Figure 3. Algorithm validation workflow.

In this framework, nuclear boundaries marked and annotated by pathologists and pathology residents are considered ground truth against which algorithm-generated results are compared. Image datasets used in validation are organized by three hierarchical spatial concepts in increasing order of granularity: slide, tile, and subregion. To make computer analysis tractable on large whole slide images, each slide is partitioned into a series of tiles with 4096x4096-pixels in size under 20X objective magnification. Even at this resolution, a tile can contain tens of thousands of nuclei, making it infeasible for human experts to mark boundaries for all nuclei in the tile. Hence, the tiles are further divided into 8x8 sub-regions for human processing – our experiments with different subregion sizes showed that human reviewers could analyze one 8x8 sub-region within 10 minutes on average.

The overall validation workflow includes five major steps as shown in Figure 3: 1) image partitioning of whole slide images into small regions for processing and annotation, 2) stratified sampling to generate a sound set of regions, 3) nuclei analysis with algorithms and human annotations, 4) result representation and loading into the database, 5) and statistical validation

analysis. The PAIS database is employed in multiple steps. In the stratified sampling step, the database is first used to create a set of subregions for each tile with a stored procedure. An initial algorithm result set is loaded into the database. When a tile is partitioned into sub-regions, the subregion and markup containment relationship is computed through a spatial containment query and persisted in a table. The number of nuclei in each subregion can then be obtained by searching the table with a count and group-by query. All of the subregions for each tile are grouped into sets of “low”, “average”, “high”, and “very high” counts and stored in another table. A set of sub-regions are then randomly selected from each group for comparisons. Sub-regions from stratified sampling are reviewed by human experts with a graphical user interface. The markups from each review are captured into an XML document, which is then validated and parsed into the PAIS database.

The PAIS database is also used to compute statistics for differences and similarities between human generated results and computer generated ones. For example, the database is queried to retrieve only nuclei that have one-to-one match between algorithm- and human-generated results. If a human marked nucleus contains multiple machine-segmented nuclei, it is not included in the final evaluation study. For each of one-to-one nuclear pairs, three measures - overlapping to union ratio, centroid distance, and Hausdorff distance - are computed with either database built-in functions or user-defined functions.

Application 2: In Silico Correlative Morphometric Study

A previous study of glioblastoma has defined four clinically-relevant tumor subtypes by differences in gene expression and characteristic genomic alterations[24]. We have utilized the PAIS database in an effort to examine the morphological correlates of these tumor subtypes [25]. Computer algorithms were used to analyze diffuse glioma brain tumor images in a large-scale dataset consisting of 307 slides corresponding to 77 distinct patients. Each analysis computed 74 features for each segmented nucleus. The segmentation results and features were stored in the PAIS database. In order to correlate micro-anatomic morphometry with molecular profiles and clinical outcome, summary statistics on image features were computed for each

image. This process involved calculating the mean feature vectors and the feature covariance values of all possible feature pairs over all nuclei in the image. The PAIS database was queried to search for feature pairs and retrieve corresponding feature values. The summary statistics for each image were then combined in a separate program to create a single-feature vector for the image. This allowed us to represent each image as a point in the summary statistics feature space – in our case, it was a 2849-dimensional space, since a nucleus had 74 features.

Queries for mean, standard deviation, and covariance of feature calculations are supported through IBM DB2 SQL queries with DB2's built-in aggregation functions: the AVG, STDDEV, and COVARIANCE functions, respectively. With the PAIS database query support on morphological signature computation for whole slide images, we were able to correlate nuclear morphometry with clinically relevant molecular characterizations and to produce preliminary result suggesting a possible relationship between nuclear morphometry and the established clinically relevant molecular glioblastoma (GBM) tumor subtypes.

Application 3: Correlative Study on Liver Biopsy

We are currently carrying out a study to quantify the relationship between the area of liver steatosis regions, clinical parameters such as liver functional studies, and radiology quantization measurements. This study involves a large set of liver biopsies with both microscopy and radiology images. The properties of the liver organs reflected in the radiology images, such as measurements of steatosis (i.e., fat content) and fibrosis (i.e., scarring), will be measured by experienced radiologists. The microscopy images will be analyzed by machine algorithms. Due to vast number of steatosis regions in each image, manual segmentation and annotation of images becomes very difficult. We are developing machine algorithms to identify all steatosis regions with certain constraints (e.g., constraints on size and shape). This information derived from microscopy imaging will then be integrated with the radiology readouts from the associated magnetic resonance imaging (MRI) images.

All numerical features derived from steatosis as well as the locations of the steatosis regions will be captured in the PAIS database. The radiology readouts will be stored in a database built

on the AIM data model[23]. These two databases will be used to investigate how the measure of correlation between structures at different scales (e.g., microvesicular versus macrovesicular steatosis) is varied as the cut-off values of properties used in machine algorithms are changed. This is particularly important because certain features such as the type of steatosis can be very crucial in predicting the functional status of the liver[26]. This will be done by generating queries on the PAIS database to search for and retrieve only the steatosis regions that satisfy a set of user-defined criteria on the properties and spatial locations of the regions, and by comparing the query results with the radiology readouts from the same images.

Database Performance

The PAIS database is designed to be fast for metadata and spatial queries and queries involving comparisons of results from different analyses. To undertake a performance evaluation of the database, we selected 18 slides, and loaded image analysis results from two different algorithm parameter sets and human annotated results. The total volume of data amounts to about 18 million markups and 400 million features. We selected different types of queries that are typical in our use cases and ran them against the PAIS database and as MATLAB programs. We chose MATLAB for comparison instead of a C/C++ implementation, because MATLAB is a platform more commonly used for algorithm development and analysis by imaging researchers, although an implementation of the same operations in C/C++ could achieve lower execution times. There are additional performance improvements that can be implemented on these implementations[27-31]. In an earlier work we performed a detailed comparison of different database configurations for spatial joins[29]. Our results showed we could take advantage of combined processing power and memory capacity of multiple machines by carefully distributing database contents and modifying queries. We plan to carry out a similar performance study for the PAIS database in a future work.

The queries selected for performance evaluation are: (Query 1) Count nuclei on each slide processed by a specific algorithm; (Query 2) Compute intersection ratio and distance between nuclei segmented by two different algorithms on the same slide. This query is important for

algorithm validation studies, in which results obtained from different algorithms are compared to look for similarities and differences in the analysis outcome; and (Query 3) Retrieve the mean nuclear feature vector and covariance of features on nuclei segmented by an algorithm on a slide. This query is used to examine the relationship between nuclear morphology and tumor subtypes defined by molecular analyses.

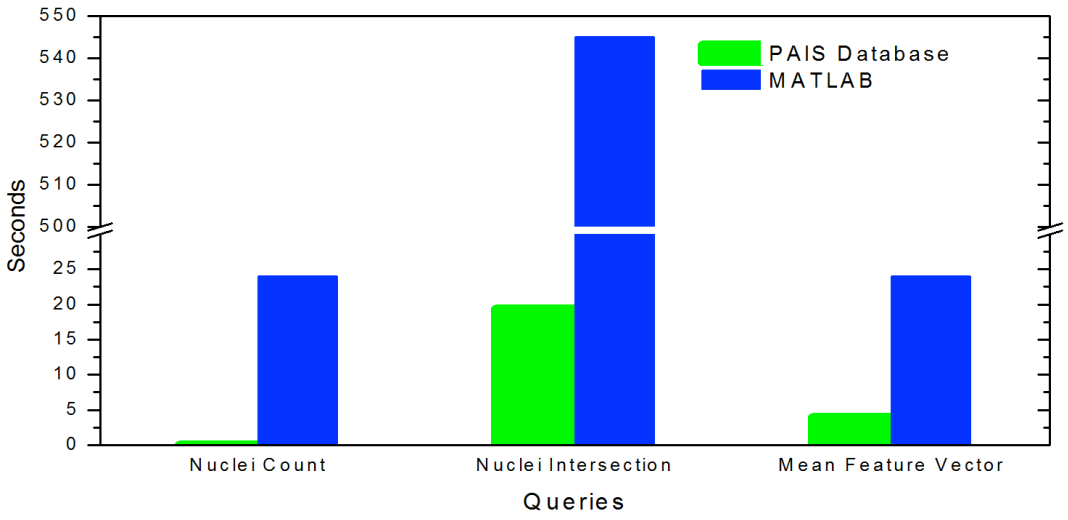

Figure 4. Comparison of Performance of PAIS vs MATLAB. PAIS database has significant performance advantage over programmatic approach (270, 28 and 5 times faster). Highly expressive query language: 5 lines vs 60 lines.

The execution times for these queries are shown in Figure 4. The first query takes 18.4 seconds to execute for a single slide using MATLAB, and only 0.068 second with the PAIS database. The execution times for the second and third queries are 545 and 24 seconds using MATLAB, where the same queries take 19.5 and 4 seconds with the PAIS database, respectively. Our results show that the PAIS implementation achieves significant speedup over the MATLAB based implementation.

Queries on multiple slides are generally linearly scalable in the PAIS database. For example, computing the covariance of features on one slide takes on average about 219 times less than it does on 213 slides. Such scalability is ensured through data clustering on images and tiles at the data loading stage. Since most query operations are tile or slide based, such clustering will minimize the number of disk reads during query execution.

The PAIS database loading tool is also optimized for efficiency. We use an efficient event based XML document parsing approach to process XML documents. The approach only needs a single scan of a document and requires minimal resource. We also optimize insertions through batch transactions. We are able to load results from a singe whole slide image (~0.5 million objects) into the PAIS database within 10 minutes.

Another significant advantage of the PAIS database, compared to accessing and retrieving data through MATLAB, is the highly expressive power of queries. SQL query language, with the spatial capabilities through the use of spatial data types and spatial functions, makes it very easy to express such queries. For example, Query 2 is expressed as a single SQL statement with five lines of SQL code. With MATLAB, the same query is written as 30 lines of computation code plus another 30 lines of code for handling disk read/write operations.

Spatial queries are common in our use cases. They are used for retrieving contained markup objects, the density of markup objects, comparisons between different algorithm results, or between human generated and algorithmic results. Spatial queries are also used for constraining analysis on certain regions, such as human annotated regions (e.g., tumor regions) or regions classified by an analysis algorithm. The spatial database engine in DB2 Spatial Extender provides automatic query optimization on spatial predicate-based queries through its grid based spatial index. We take advantage of this index to speed up queries. Each spatial region is divided into multi-level grids and indexed. These grids can be used to efficiently identify markups (segmented regions) that are in two different datasets and that intersect each other. Instead of comparing a markup with all markups across the whole image from the other dataset, the grid-based index can be invoked to retrieve markups that intersect the same grid as the query markup. In this way, the set of comparisons is reduced significantly, and linear scalability can be achieved.

Conclusions and Discussion:

Effective use of large microscopy image datasets in basic, clinical, and translational research requires the application of many interrelated analyses for the detection and detailed classification of morphological characteristics. Our experience with the in silico study of brain tumors has shown that data sets resulting from these analyses can be extremely large. Modeling and managing pathology image analysis results in databases provides immediate benefits on the value and usability of data through standardized data representation, data normalization, and semantic annotation. The database provides powerful query capabilities, which are otherwise difficult or cumbersome to support by other approaches such as programming languages. Advanced database access methods can be employed to make queries efficient. Besides, all query interfaces are through standardized SQL query language, which is highly expressive and natural for data retrieval and comparison operations. Standardized, semantic annotated data representation and expressive interfaces also make it possible to more efficiently share image data and analysis results.

However, moving data from unstructured representation to a structured one is challenging. The first major challenge is the big gap between researchers who work on imaging algorithms and database researchers and developers. Image algorithms researchers focus on algorithms and programming languages, whereas database developers tend to look at the problem from the point of view of data models, query languages, and query optimization methods. This project has been made possible through extensive collaboration as a team of multi-disciplinary people to map imaging questions into database questions. The second major challenge is to generate valid structural data. Databases have rigid requirements on data validity, such as integrity constraints and data types, especially complex spatial data types. Human generated annotations such as freehand drawing boundaries are often invalid polygons (e.g., unclosed or self-crossing polygons). We have detected in our studies more than a dozen scenarios of invalid polygons. We have developed a set of computational geometry algorithms and validation tools to fix such scenarios and transform them into valid spatial data types acceptable by the spatial database engine. The third challenge is to provide a generic and user-friendly document

generator that can be used in diverse applications. Applications oftentimes have their own proprietary data representations and naming conventions, e.g., format on encoding patient id in file names. To let application users develop their own document generation tool is difficult. Instead, we have developed a customizable PAIS document generator framework, which takes a simple plain file based representation of results and annotate the data with additional metadata conventions defined in an XML based customization file. This approach significantly simplifies users' effort for document generation. They only need to convert their algorithms' results into this simple plain file format, and documents are then automatically generated.

In our current implementation we have chosen a relation database implementation of the PAIS model. XML native databases are becoming mature technologies and have many advantages on managing XML documents. In this approach, XML documents are managed as they are, and no mapping between data models and query languages is needed. Besides, XML databases provide XML query languages such as XQuery and SQL/XML to express powerful queries. One immediate benefit of XML-based approach is that data exchanging is made easy. XML databases are also much tolerant to schema evolutions. Another significant benefit is that application development is simplified because no mapping from the XML schema to the relational schema is needed. However, native XML based approach is more suitable for managing small sized XML documents such as those generated from tissue microarray image analyses. The relational database based implementation, on the other hand, is highly efficient on both storage and query performance for managing and querying large-scale result data. The side effect is major effort needed on developing efficient tools for mapping XML documents into relational and spatial tables.

We plan to extend our current work in several ways: The current PAIS database server runs on a single node machine where no parallel I/O is provided. We have observed that the database performance is mainly bounded by the I/O bottleneck. We are working on scaling up database with parallel I/O capabilities and data partitioning through parallel database infrastructure. We are putting together multiple physical RAID-5 disk arrays on a single node, and testing data partitioning on multiple physical nodes as well. We expect this will boost the performance by an order of magnitude.

Another ongoing work is the investigation of MapReduce[32] based query processing capability. In one ongoing project on algorithm sensitivity, we perform on demand algorithm result comparison and identify patterns of quality of results versus change of parameters. In this case, intermediate testing results are not persisted in the database, but the process requires rapid generation and processing of the intermediate results. We are investigating the use of the MapReduce approach for scalable query execution and data processing on commodity clusters.

Image analysis results often need to be queried together with images to, for example, visualize images with markups and annotations, retrieve image regions based on characteristics, etc. This needs a database of managing images used in the studies, thus integrated querying capabilities are possible through the image database and the PAIS database. We are designing the database and implementing a set of tools to support such combined queries leveraging the work done in virtual microscope software systems.

References:

1. Furness, P.N., N. Taub, K.J. Assmann, G. Banfi, J.P. Cosyns, A.M. Dorman, et al., *International variation in histologic grading is large, and persistent feedback does not improve reproducibility.* The American journal of surgical pathology, 2003. **27**(6): p. 805-10.
2. *The Cancer Genome Atlas (TCGA) Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
3. Ayala, G., D. Wang, G. Wulf, A. Frolov, R. Li, J. Sowadski, et al., *The prolyl isomerase Pin1 is a novel prognostic marker in human prostate cancer.* Cancer Res, 2003. **63**(19): p. 6244-51.
4. Catalyurek, U.V., M.D. Beynon, C. Chang, T.M. Kurc, A. Sussman, and J.H. Saltz, *The virtual microscope.* IEEE Transactions on Information Technology in Biomedicine, 2003. **7**(4): p. 230-248.
5. Afework, A., M.D. Beynon, F. Bustamante, S.H. Cho, A. Demarzo, R. Ferreira, et al., *Digital dynamic telepathology - the Virtual Microscope.* Journal of the American Medical Informatics Association, 1998: p. 912-916.
6. Goldberg, I., C. Allan, J.-M. Burel, D. Creager, A. Falconi, H. Hochheiser, et al., *The Open Microscopy Environment (OME) Data Model and XML File: Open Tools for Informatics and Quantitative Analysis in Biological Imaging.* Genome Biol. , 2005. **6**(R47).
7. Martone, M.E., J. Tran, W.W. Wong, J. Sargis, L. Fong, S. Larson, et al., *The Cell Centered Database project: An update on building community resources for managing and sharing 3D imaging data.* Journal of Structural Biology, 2008. **161**(3): p. 220 - 231.
8. Martone, M.E., S. Zhang, A. Gupta, X. Qian, H. He, D.L. Price, et al., *The cell-centered database: a database for multiscale structural and protein localization data from light and electron microscopy* Neuroinformatics, 2003. **1**(4): p. 379-395.

9. Chen, W., V. Chu, J. Hu, L. Yang, F. Wang, T. Kurc, et al., *ImageMiner: a medical image analysis and image management UML data model.*, in *APIII: Advancing Practice, Instruction & Innovation Through Informatics*. 2009: Pittsburgh, PA.
10. Kvilekval, K., D. Fedorov, B. Obara, A. Singh, and B.S. Manjunath, *Bisque: A Platform for Bioimage Analysis and Management.* Bioinformatics, 2010. **26**(4): p. 544-552.
11. Wang, J., G. Wiederhold, O. Firschein, and S. Wei, *Content-based image indexing and searching using Daubechies' wavelets.* Int. J. Digital Librar., 1998. **1**(4): p. 311-28.
12. Carson, C., S. Thomas, S. Belongie, J. Hellerstein, and J. Makik. *Blobworld: a system for region-based image indexing and retrieval*. in *Third Int. Conf. Vis. Inf. Sys.* 1999.
13. Schnorrenberg, F., C. Pattichis, C. Schizas, and K. Kyriacou, *Content-based retrieval of breast cancer biopsy slides.* Technology & Health Care, 2000. **8**(5): p. 291-7.
14. Guld, M.O., C. Thies, B. Fischer, and T.M. Lehmann, *A generic concept for the implementation of medical image retrieval systems.* Stud Health Technol Inform, 2005. **116**: p. 459-64.
15. Chen, W., D.J. Foran, and M. Reiss, *Unsupervised imaging, registration and archiving of tissue microarrays.* Proc AMIA Symp, 2002: p. 136-9.
16. Hadida-Hassan, M., S.J. Young, S.T. Peltier, M. Wong, S. Lamont, and M.H. Ellisman, *Web-based telemicroscopy.* J Struct Biol, 1999. **125**(2-3): p. 235-45.
17. Wetzel, A., P. Andrews, M. Becich, and J. Gilbertson, *Computational aspects of pathology image classification and retrieval.* J Supercomputing, 1997. **11**: p. 279-93.
18. Zheng, L., A.W. Wetzel, J. Gilbertson, and M.J. Becich, *Design and analysis of a content-based pathology image retrieval system.* IEEE Trans Inf Technol Biomed, 2003. **7**(4): p. 249-55.
19. Wang, J.Z., J. Nguyen, K.K. Lo, C. Law, and D. Regula, *Multiresolution browsing of pathology images using wavelets.* Proc AMIA Symp, 1999: p. 430-4.
20. Yang, L., W. Chen, P. Meer, G. Salaru, M.D. Feldman, and D.J. Foran, *High throughput analysis of breast cancer specimens on the grid.* Med Image Comput Comput Assist Interv Int Conf Med Image Comput Comput Assist Interv, 2007. **10**(Pt 1): p. 617-25.

21. Gurcan, M.N., J. Kong, O. Sertel, B.B. Cambazoglu, J. Saltz, and U. Catalyurek, *Computerized pathological image analysis for neuroblastoma prognosis.* AMIA Annu Symp Proc, 2007: p. 304-8.
22. Clunie, D.A., *DICOM structured reporting and cancer clinical trials results.* Cancer Inform, 2007. **4**: p. 33-56.
23. Channin, D., P. Mongkolwat, V. Kleper, K. Sepukar, and D. Rubin, *The caBIG Annotation and Image Markup Project.* Journal of Digital Imaging, 2009.
24. Verhaak, R.G., K.A. Hoadley, E. Purdom, V. Wang, Y. Qi, M.D. Wilkerson, et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
25. Cooper, L.A., J. Kong, D.A. Gutman, F. Wang, S.R. Cholleti, T.C. Pan, et al., *An integrative approach for in silico glioma research.* IEEE Trans Biomed Eng, 2010. **57**(10): p. 2617-21.
26. Marsman, H., T. Matsushita, R. Dierkhising, W. Kremers, C. Rosen, L. Burgart, et al., *Assessment of donor liver steatosis: pathologist or automated software?* Human pathology, 2004. **35**(4): p. 430-5.
27. Chung, W., S.-Y. Park, and H.-Y. Bae, *Efficient parallel spatial join processing method in a shared-nothing database cluster system*, in *International Conference on Embedded Software and Systems (ICESS)*. 2004. p. 81-87.
28. Gray, J., M.A. Nieto-Santisteban, and A.S. Szalay, *The zones algorithm for finding points-near-a-point or cross-matching spatial datasets.* Microsoft Technical Report, MSR-TR-2006-52., 2006.
29. Kumar, V.S., T. Kurc, J. Saltz, G. Abdulla, S. Kohn, and C. Matarazzo, *Architectural Implications for Spatial Object Association Algorithms*, in *The 23rd IEEE International Parallel and Distributed Processing Symposium (IPDPS 09)*. 2009: Rome, Italy.
30. Gaede, V. and O. Gunther, *Multidimensional access methods.* ACM Computing Surveys, 1998. **30**(2): p. 170-213.

31. Becla, J., K.-T. Lim, S. Monkewitz, M. Nieto-Santisteban, and A. Thakar, *Organizing the extremely large LSST database for real-time astronomical processing*, in *17th Annual Astronomical Data Analysis Software and Systems Conference (ADASS 2007), London, England.* 2007.
32. Dean, J. and S. Ghemawat, *Mapreduce: Simplified data processing on large clusters.* Communications of the Acm, 2008. **51**(1): p. 107-113.