Use Cases and System Requirements

for AIM Pathology

*Summary.* 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 modeling, managing, querying, and integrating the vast amounts of data resulting from the analysis of large digital pathology datasets.

# 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. Whole-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.

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. 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.

# Use Case: Whole Slide Imaging

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) and REMBRANDT 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 5x104 by 5x104 pixels at 20X objective magnification. Brain tumor image analysis algorithms segment and classify 105 to 107 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.

# Use Case: Tissue MicroArray Image Analysis

As part of a separate project, several of the key computational and imaging tools that our team has developed have already been migrated to core research facilities for use in ongoing investigative studies at The Cancer Institute of New Jersey. They have been used to analyze microarrays consisting of cancers of the breast, head & neck, and prostate. Similar to the ISBTRC use case, the Histopathology and Imaging Shared Resource is steadily imaging hundreds of glass specimens a year with the majority of the specimens being Tissue Microarrays(TMA). The support for TMA specimens presents particular challenges to pathology data management – instead of having only one tissue specimen on the slide, each TMA specimen can easily incorporate tissue specimen originated from hundreds of patients; researchers routinely stain multiple slices of same TMA block with different markers to explore their relations or the combined diagnostic/prognostic power; TMA studies are often designed with build-in redundancies for which all analyss/evaluation results need to be properly consolidated. PAIS model is designed to help users handle the above complications when studying their TMA data set. The algorithms developed at the Cancer Institute of New Jersey successfully isolate and quantify designated marker staining from counter-stained immunohistochemically stained TMA specimen. The tissue structures were captured using filter-banks and texton libraries build from gold-standard databases.

As part of a recent study the automated software was used to quantify Beclin1 expression which was shown to be predictive of autophagy (DiPaola, et al., 2008). Our team has since conducted a series of man-machine performance studies. In the first experiments we utilized the TMA analysis tools to evaluate IHC staining intensity on imaged breast cancer TMA specimens comprised of 1407 tissue cores. The results showed that the computer software algorithms achieved similar interpretations to those provided by a panel of 3 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 (Goodell A, et al., 2010). As an extension of those studies we examined the expression patterns of a cohort array of several hundred tissue samples originating from Head and neck squamous cell carcinoma (HNSCC) patients. ROC curve analysis showed that the automated and manual scoring were generally consistent with area under the curve (AUC) values of 0.9677 for Smad2 and 0.885 for Smad3 (Xie W, et al., 2009).

Utilizing the image-based and correlated clinico-pathologic features gathered during the course of those studies our team has established a database and an image archive to enable us to conduct iterative prototyping and performance analysis of the proposed PAIS data model.

To date, the following query use-case scenarios have been identified and several are already supported:

**TMA use case 1:** Locate and retrieve all those imaged tissue microarray cores which originated from patients exhibiting similar histological types, diagnoses and/or stages of disease progression.

**TMA use case 2:** Locate and retrieve all imaged TMA cores a specific cancer type that responded poorly to specific treatment regimens complete with correlated survival data.

**TMA use case 3:** Perform digital noise reduction and color decomposition of imaged specimens and identify all of those pixels exhibiting specific DAB and HER2 staining intensities and distributions

**TMA use case 4:** Detect and delineate all tumor regions of tissue within a given cohort of imaged TMA cores and classify each as a specific stage or subtype of disease progression (e.g. dysplastic, ductal carcinoma in situ, lobular carcinoma in situ, etc)

**TMA use case 5:** Detect, delineate and classify all nuclei within a given cohort of discs as cancerous or benign based upon fusion of shape and texture descriptors

**TMA use case 6:** Systematically identify all regions within the imaged specimens exhibiting the statistically most similar staining and expression signatures while drilling down at varying levels of granularity (i.e. disc, tissue, cellular, sub-cellular levels)

**The following use-case scenarios are longer-term projects for which feasibility studies will be conducted:**

**TMA use case 7:** Develop the means to support query and retrieval of specimens which have been acquired using multispectral imaging

**TMA use case 8:** Develop the means to support query and retrieval of imaged specimens which have been prepared using multiple quantum dot antigen-antibody conjugates

**TMA use case 9:** Develop the means for conducting systematic experiments to conduct iterative prototyping of new image analysis algorithms leading to the optimal approach for interrogating federated databases and image archives

**TMA use case 10**: Query for existing image feature datasets or gold-standard databases, based on specimen domain and image type, and provide support for content-based image retrieval

[Note: many of these use-cases, especially 5-10, apply to a spectrum of pathology imaging applications which are not limited to the TMA imaging domain.]

In the next phase of the project, we plan to expand the archive of imaged specimens and correlated clinical data to include a wider range of tissues, cancer types and biomarkers by increasing the number of breast, prostate and glioblastoma specimens under study with a long-term plan to include a new, representative set of head & neck, lymphoma and melanoma cases.

# System Requirements

## Information Objects

The following information is essential to represent metadata generated from pathology image analysis or human annotations.

*Image Reference.* Meta provides metadata that describes an image or a group of images that are used as the base for making markup and annotation, and can be used to identify and retrieve them from an image archive. The image reference should also include the resolution of the image in microns/pixel, the z-axis resolution and coordinate, if available.

Region information is needed to identify the area of interest from an image (e.g., a 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 should also include attributes for the left corner coordinate (x, y) and the width and height of the region on the original image. It should also include 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 coordinate resolution should be defined.

*Subject* information is needed for a person or animal investigated in a study, and specimen information is needed for a single unit of tissue or blood or urine that is taken from the subject and used for the study. Anatomic entity is needed to locate the place in the body where the image is taken and the finding is located. Equipment is the imaging device that acquires images from the subject or specimen.

*Analysis information*. We need to track analysis related information such as study, experiment, the person who performs the experiment. Collection information is also needed for representing a group of items of the same type, gathered for display or study. For example, when we perform an experiment to validate algorithms against human assessment, all the instances to be compared are of the same collection.

*Markup*. 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 one or multiple images, and can be in form of vector based (0-2D shapes), Surface*, or* Field*.*  Markup representation should also be extremely compact for saving storage, as there could be millions of markups per slide. 2D shapes could use standard based representation like SVG (Scalable vector graphics), or WKT (spatial well known text representation).

*Annotation.* Annotation associates semantic meaning to markup entities through coded or free text terms that provide explanatory or descriptive information. Annotations could include features (calculations), information about the quantitative results from mathematical or computational calculations. Annotation also includes observations, information about interpretation of a markup or another annotation entity in an image or images, including visual features, morphologic or physiologic processes, and diseases. Due to the complexity of image analysis methods, Calculation can contain a wide range of feature types, such as Scalar, Array, Histogram, and Matrix. Complex calculation results (e.g., files) may also be binary encoded as text based representation, such as Base-64 or HEX based encoding. Similarly, observations can be quantified based on different measure scales: i) Nominal scale, which offer names or labels for certain characteristics; ii) Ordinal scale, which uses numbers to represent rank order of the entities assessed (e.g., glioma classification with grade 0-10); iii) interval scale, and iv) Ratio scale, where measurement is the estimation of the ratio between a magnitude of a continuous quantity and a unit magnitude of the same kind.

Provenance. We need to keep track provenance to determine the derivation history of a markup or annotation, including algorithm information, parameters, and inputs. Such information is critical for validating approaches and comparing algorithms.

## Major Relationships

An AIM Pathology document should contain image references (one or many), markup (one or many), and annotation (one or many). The model should capture major relations between markup and annotation. One or multiple markups could be associated with annotation, and annotation could include multiple features and observations. Image reference should be associated with subject, specimen, anatomic entity and equipment information. A region is also needed to be associated with each image reference. Provenance could be associated with the document root, or markups or annotations.

Nested annotations should be supported, and similar for nested markups. Document root should also be associated with user and project information.

# Related Standards

## DICOM Supplement 145

The Digital Imaging and Communications in Medicine (DICOM) standard was originally developed for handling, storing, printing, and transmitting information in radiology images among different hardware and systems. The DICOM standard includes file format definition as well as network communication protocols. It enables the integration of scanners, servers, workstations, printers, and network hardware from multiple manufacturers into a Picture Archiving and Communication System (PACS). The National Electrical Manufacturers Association (NEMA) holds the copyright to this standard. It was developed by the DICOM Standards Committee, whose members are also partly members of NEMA.

In recognize of the fact that digital imaging, especially whole slide imaging, was gaining recognition in pathology and that the existing DICOM standard lacked the capacity to handle large two-dimensional images such as WSI digital slide, the DICOM committee formed working group 26 to form supplements so as to explore the subject of including pathology images in DICOM. The resulting DICOM Supplement 145: Whole Slide Microscopic Image IOD and SOP classes were generated by working groups 26 in 2009. The Supplement detailed on a description of the tiling operation that allowed large pathology images to be dissected into tiles that hence stored in a fashion of image series that were similar to radiology image series and the retrieval of specific image area in designated levels.

In general DICOM stores image annotations as separate objects from the image. There are several types of annotation objects serving different purposes: Presentation States refer to parameter of a display rendering the image; Segmentation provides a categorization of areas of image; and Structured Reporting captures measurements, clinical observations, analyses as well as findings with robust reference to image evidence.

The following IOD (Information Object Definitions) are extracted from Supplement 145 that were identified to be potentially useful for our model design.

C.7.9

|  |  |  |  |
| --- | --- | --- | --- |
| Lossy Image Compression | (0028,2110) | 3 | Specifies whether an Image has  undergone lossy compression.  Enumerated Values:  00 = Image has NOT been  subjected to lossy  compression.  01 = Image has been subjected to  lossy compression.  See C.7.6.1.1.5  Lossy Image |
| Compression Ratio | (0028,2112) | 3 | Describes the approximate lossy  PS 3.3 - 2009  Page 370  - Standard -  compression ratio(s) that have been  applied to this image.  See C.7.6.1.1.5 for further explanation.  May be multivalued if successive lossy  compression steps have been applied.  Notes: 1. For example, a compression  ratio of 30:1 would be described  in this Attribute with a single value  of 30.  2. For historical reasons, the  lossy compression ratio may also  be described in Derivation |
| Lossy Image Compression Method | (0028,2114) | 3 | A label for the lossy compression  method(s) that have been applied to this  image.  See C.7.6.1.1.5 for further explanation.  May be multivalued if successive lossy  compression steps have been applied;  the value order shall correspond to the  values of Lossy Image Compression  Ratio (0028,2112).  Note: For historical reasons, the lossy  compression method may also be  described in Derivation  Description (0008,2111). |

C.7-10

|  |  |  |  |
| --- | --- | --- | --- |
| Image Orientation (Patient) | (0020,0037) | 1 | The direction cosines of the first row and  the first column with respect to the patient.  See C.7.6.2.1.1 for further explanation. |
| Image Position (Patient) | (0020,0032) | 1 | The x, y, and z coordinates of the upper  left hand corner (center of the first voxel  transmitted) of the image, in mm. See  C.7.6.2.1.1 for further explanation. |
| Slice Thickness | (0018,0050) | 2 | Nominal slice thickness, in mm. |
| Slice Location | (0020,1041) | 3 | Relative position of the image plane  expressed in mm. C.7.6.2.1.2 for further  explanation. |

C.8-73

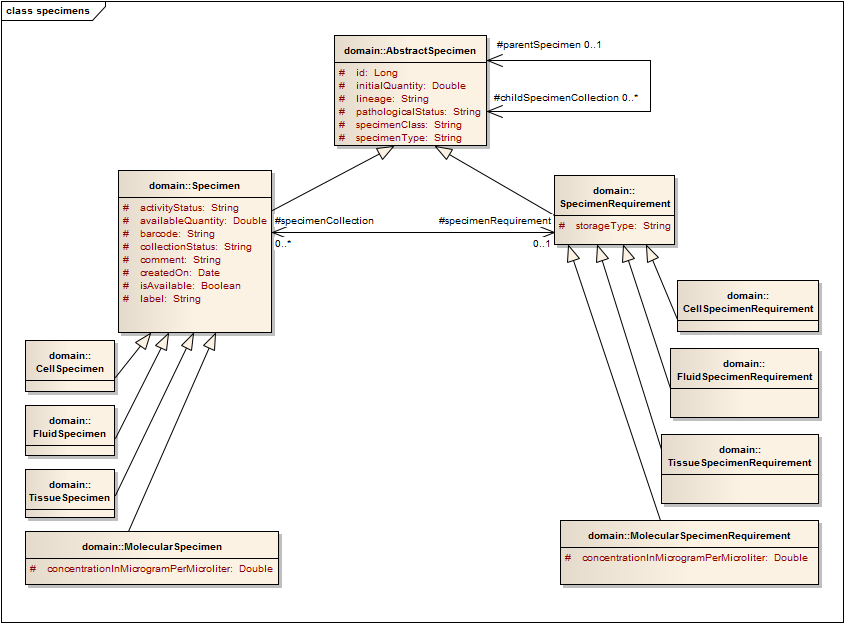
|  |  |  |  |
| --- | --- | --- | --- |
| Modality | (0008,0060) | 1 | Type of equipment that originally acquired  the data used to create the images in this  Series.  Enumerated Value:  MG  See section C.7.3.1.1.1 for further  explanation. |

## caTissue

caTissue Suite is caBIG's biorepository tool for biospecimen inventory management, tracking, and annotation. This tool permits users to enter and retrieve data concerning the collection, storage, quality assurance, and distribution of biospecimens. The caTissue data model is the underlying data structure that supports the above functionalities.

The caTissue is relevant to our model development in that we consider the pathology specimens upstream information from the pathology images. The information related to the specimens are important to PAIS users as it helps understanding and interpreting the images. Therefore we paid specific attention to data related to pathology specimen generation and the related caTissue data structure is extracted as the figure below.

Specimen gathering info:



Pathology Annotations\_SCG

Pathology Annotations\_Specimen

These two diagrams in caTissue model also have many elements defined in disease-specific annotation format.

# LS DAM

The caBIG Life Sciences Domain Analysis Model (LS DAM) is a shared view of the semantics for Life Sciences which includes hypothesis driven basic and pre-clinical research as well as discovery sciences. It is aligned, where appropriate, with the Clinical Sciences Biomedical Research Integrated Domain Group (BRIDG) model, which supports protocol driven clinical and pre-clinical research. The LS DAM is a foundational component for achieving semantic interoperability among the various applications across caBIG.

Some interesting classes to our model design include:

Experiment: A coordinated set of actions and observations designed to generate data, with the ultimate goal of discovery or hypothesis testing.

ExperimentalStudy: A detailed examination or analysis designed to discover facts about a system under investigation. Systems may include intact organisms, biologic specimens, and natural or synthetic materials.

## Open Microscopy Environment (OME)

OME (Open Microscopy Environment) develops open-source software and data format standards for the storage and manipulation of biological light microscopy data. One focus of OME model is on modeling experiment, microscope, images, acquisition, analyses and results. While OME provides a comprehensive acquisition model, it has a limited “flat” structure on analytical results.

Thus, instead of using OME’s approach for presenting data analysis results or human annotations, we could provide a reference to OME document to complement our model.