Content-Based Tissue Image Mining

Abhi Gholap, Gauri Naik, Aparna Joshi, CVK Rao
BioImagene, Inc.
1510 Fashion Island Boulevard, San Mateo, California 94404, USA
abhijeet@bioimagene.com

Abstract

Biological data management and mining are critical areas of modern-day biology research. High throughput and high information content are two important aspects of any Tissue Microarray Analysis (TMA) system. Tissue image mining is efficient and faster if the tissue images are indexed, stored and mined on content.

A four-level system to harness the knowledge of a pathologist with image analysis, pattern recognition, and artificial intelligence is proposed in this article. At Image Processing and Information Level, information such as contrast or color is used. At Object Level, pathological objects, including cell components, are identified. At Semantic Level, layout and formation of individual cells into sheets in a tissue image are analyzed. At the highest level, Knowledge Level, inference of the expert is indicated.

A pilot system that uses two levels of harnessing involving the first two levels' features of tissue images with immunohistochemical markers is implemented.

1. Introduction

Biological image data management, which addresses the inherent problems in acquisition, analysis, storage, management, and integration of ever-evolving heterogeneous biological data, is the key to the critical areas of modern-day, data-intensive biology research [1]. Over the past several years, the focus has been on the development of methods and technologies supporting high-throughput generation of biological data, such as DNA sequence and gene expression data [2-5].

As pharmaceutical companies head into the target validation mode, validation studies at tissue level become more prevalent. Techniques, such as Tissue Microarray (TMA), are preferred as compared to DNA sequence studies. Since the amount of data available with TMA is much larger than the DNA sequence and gene expression data, a sophisticated software solution becomes imperative in mining and managing such data. With such solution, collection, interpretation, and validation of TMA data are comparatively easier, and the information generated can easily be integrated with other diagnostic methods. One can derive meaningful information from TMA images, which is useful for diagnostic as well as for research purposes. High-throughput is important but high-information content is even more important. Tissue image mining will be efficient and faster only if the tissue images are indexed, stored and mined on content. What is needed is an efficient tissue image-mining tool.

Scope of content used in tissue image mining varies by large degree, depending on the associated image analysis algorithms. There are image management systems, which rest on standard databases for efficient data storage and management, such systems might consider image parameters like width, height, spatial resolution, magnification and textual information provided as a means for indexing, sorting and searching images. In some of these image management systems, parameters are given an uniform weight irrespective of the significance of a parameter from domain perspective. Generally, a life science researcher is interested in the images of same morphology and score, and not in their size.

Clustering is another commonly used approach by some image mining systems. However, this approach suffers from some distinct disadvantages. The success of clustering depends on the parameters and methodology used for clustering. Most of the parameters used by life science researchers do not have precise values, and, therefore, are not suitable for clustering. For example, in the case of breast carcinoma membrane-grading using Her2/neu, clustering on the basis of 0+,1+,2+ and 3+ is possible only if the boundaries between these classes can be defined precisely. Furthermore, there may be no

correlation between the way a life science researcher groups tissue images into classes like 0+,1+,2+ and 3+, and the features used for clustering. One consequence of this lack of correlation is difference of opinion between the automated and manual system, especially in transition or border-line cases. Technology developers find it extremely difficult to meet the expected performance.

2. Knowledge-driven image informatics

Content-based mining, or harnessing, the domain knowledge of human pathologists with image analysis, pattern recognition and artificial intelligence methods is essential in providing efficient content-based tissue mining facility [6]. Figure 1 below shows proposed architecture for harnessing domain specific knowledge with voluminous image data.

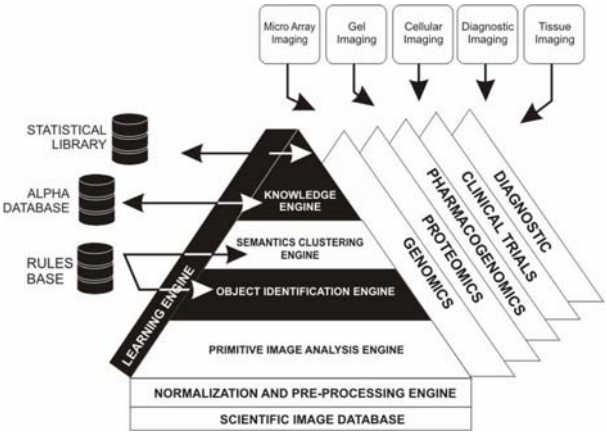

Figure 1. Architecture for harnessing knowledge with image data

Figure 2 shows a sample image, which an experienced life science researcher would perhaps interpret as follows:

This is a TMA image of a breast tissue with IHC stain showing the invasive ductal carcinoma (IDC). There are several hundred epithelial cells with 98% cytoplasm positivity. The epithelial cells are arranged in six sheets, six cords and seventeen tubules, which are infiltrating into stroma. Stroma is composed of large amount of collagen, and has more than 1,000 lymphocytes arranged in small clusters. Epithelial cells are showing pleomorphism. Nuclei of epithelial cells are also showing pleomorphism and prominent nucleoli, where nuclei and membranes are not positive.

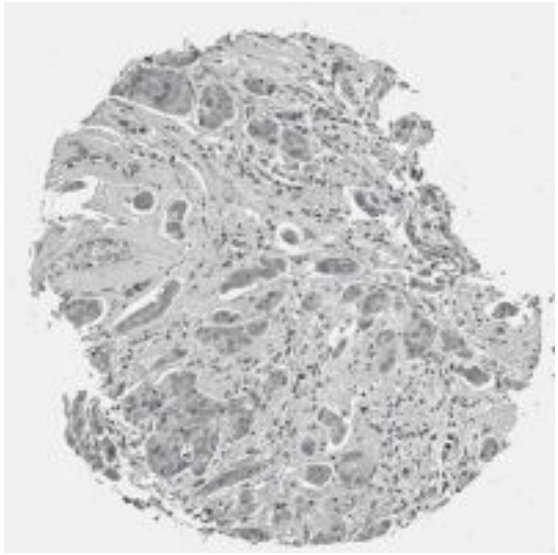

Figure 2. Sample Tissue Core Image

Alternatively, the contents of the sample tissue could be broken down into five levels of information, where different types of parameters are extracted at each level. These levels correspond to the harnessing model shown in Figure 1. The parameters provide us information about content, which can be used to index large pool of images. Following is a brief description of each of these levels of the prospoed framework.

2.1. Knowledge level

At this level, one gets the description of the image from domain perspective. For example, this is a TMA image with IHC stain. It is a breast tissue, showing IDC. There are several hundred epithelial cells with 98% cytoplasm positivity.

2.2. Semantic level

The sematic level provides detailed description of the image from domain objects point of view. For example, the epithelial cells are arranged in six sheets, six cords and seventeen tubules, which are infiltrating into stroma. Stroma is composed of large amount of collagen, and more than 1,000 lymphocytes arranged in small clusters. Epithelial cells are showing pleomorphism. Nuclei of epithelial cells are also showing pleomorphism and prominent nucleoli, where nuclei and membranes are not positive.

2.3. Object level

This block consists of object level measurements. Nature and number of parameters measured at this level could determine the scope of semantics that could be realized at next level. More the number of parameters, better the set of semantic rules. Nature of the parameters decides the consistency of derived semantic rules. All parameters at this level are mostly domain-specific or pathology terms, and none is pure image processing or statistical in nature. In the pilot system, which is being experimented, more than 40 different parameters are being used. Some of these parameters are percentage positivity over:

- All Nuclear grades
- All cytoplasm grades
- All membrane grades
- Percentage cells stained
- Percentage cells with nuclear stained
- Percentage cells with cytoplasm stained
- Percentage cells with membrane stained
- Percentage cells with both nuclear and cytoplasm stained
- Percentage cells with both cytoplasm and membrane stained
- Percentage cells with nuclear and membrane stained
- Percentage cells with all three components, nuclear cytoplasm and membrane stained
- Total number of cells
- Background stain intensity
- Mask properties
- Percentage background stained
- Percentage stromal area
- Number of vesicular cells
- Mean area of vesicular nuclei
- Standard deviation of vesicular nuclei area

2.4. Image processing level

At the preliminary stage of this level, parameters, such as image quality and image characteristic, are measured. These parameters give an indication on the variations in generic aspects of tissue image, such as staining process, staining marker, and image capturing device settings. One could use standard image processing and statistical methods to measure these parameters. Some of the parameters included in the pilot system are:

- Gray scale mean of input image
- Gray scale standard deviation of input image
- Mean value of stained pixels intensity in input image
- Standard deviation of stained pixels in input image
- Stained pixels percentage
- CS mean value of counter stained pixels in input image
- Standard deviation of counterstained pixels in input image

As a subsequent step in this level, input is provided to the system externally by pathologist or end-user. These input parameters are dependent on equipment used, scanning device used, and staining process followed. These parameters - type of tissue, type of stain, type of marker, antibody, cell localization, magnification - could be same for several tissues in a batch. Content preparation, content update, and content mining are three important aspects of any content management system. Efficiency of a given content management system is decided by the outcome of mining content. Let us look at some typical instances in which the content could be mined by harnessing domain-specific knowledge with standard image informatics.

3. Examples of content-based mining and ranking of search results

There are several situations that a researcher may encounter while experimenting with a new marker on a specific tissue.

- Case 1: Researcher wants to know the effect of the same marker on other tissues. For example, researcher on conclusive study of cytokeratin markers in breast tissue wants to study the effect the cytokeratin on liver tissue.
- Case 2: Researcher wants to know the effect of same marker on different cell localization of the same tissue.
- Case 3: Researcher wants to know all other markers that gave similar response on the same tissue. For example, while studying the effect of cytokeratin expression in liver tissue, researcher might be interested in knowing all other markers that have resulted in cytoplasm positive in breast tissues. This will help in validation of the new experiments with repository.
- Case 4: Researcher wants to know all other markers that gave similar response on all other tissues. For example, while studying the effect of cytokeratin, researcher wants to

know all other markers that gave cytoplasm positive on different tissues. These tissues could be from different animals.

- Case 5: Researcher is experimenting with a known marker on a specific tissue for percentage positivity and wants to know all other tissues with similar percentage positivity to the same marker. For example, a nuclear marker, Ki-67, being expressed in many cancers.

Ranking search results in the situations described above is one of the most powerful features that can be provided by search engines. Harnessing domain-specific knowledge (pathological descriptors) with image informatics provides us with similar facility. Here, the emphasis is given to each of the measured parameters, and the difference between query tissue image and tissue image in database is based on the researcher's knowledge and expertise.

For example, on mining "breast carcinoma, TMA, IHC, ER with 70% positivity", the result would be:

- Top ranks: All tissue images of breast carcinoma, TMA, IHC, ER with 70% positivity
- Next ranks: All tissue images of breast carcinoma, TMA, IHC, ER with 69% positivity
- Next ranks: All tissue images of breast carcinoma, TMA, IHC, ER with 71% positivity
- Bottom ranks: All tissue images of breast carcinoma, TMA, IHC, ER with 1% positivity

In another example, on mining "Breast carcinoma/TMA/IHC/VR/50%epithelialarea/ER/70%positivity", the result would be:

- Top ranks: All tissue images of breast carcinoma, TMA, IHC, VR, 50% epithelial area, ER with 70% positivity
- Next ranks: All tissue images of breast carcinoma, TMA, IHC, VR, 50%epithelial area ER with 69% positivity
- Next ranks: All tissue images of breast carcinoma, TMA, IHC, VR, 50% epithelial area ER with 71% positivity
- Middle ranks: All tissue images of breast carcinoma, TMA, IHC, DAB, 50%epithelial area ER with 70% positivity

4. Pilot studies

Proposed harnessing concept is being experimented with a four-level feature for indexing and searching tissue images. At the highest level, the features include inference drawn like "High-grade infiltrating ductal carcinoma of Breast: 3+ membrane (Her2/neu) positivity", rich in domain knowledge but difficult to extract automatically. At the lowest level, the features include the "percentage positivity range", indicating an aggregate assessment, which could be automated with reasonable accuracy.

Experiments are carried out to validate the harnessing concept. Some of the points validated are:

- Pathological descriptors are more appropriate for searching similar images.
- Weights given to each of the pathological descriptors need not be the same.
- The pathological descriptors, sensitive to the size of available tissue image, should be given lesser weight.
- Searching or sorting on generic image parameters, such as mean value of color, mean value of gray scale, and standard deviation of gray value, gives a large list.
- Harnessing pathological descriptors with image parameters leads to proper ranking of search results.

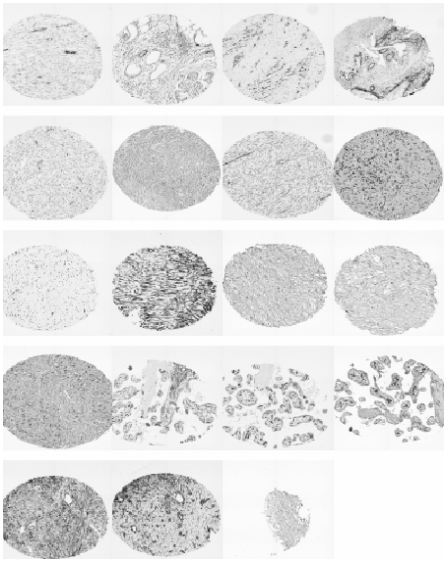

Figure 3. A thumbnail of tissue microarray cores used for experimentation

Figure 3 shows thumbnails of a sample set of 19 tissue micro array cores. Figure 4 shows two sections

of a core from this sample set which are used for search. Searching on only nuclear percentage positivity parameter gave 7 out of 19 images in the range 90-100% nuclear positivity. Searching on minimum gray value or maximum gray value is found to be sensitive to the pre-processor used.

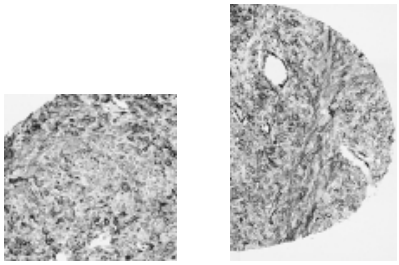

Figure 4. Sections of a core used for searching

Figure 4 shows sections of a core used for searching. Searching the sample set using sections based on nuclear percent positive together with stained mean gave the respective core on the top of the search list. Nuclear percent positivity is measured at object level and stained mean is measured at pixel level. Nuclear identification at object level uses a host of algorithms that are based on information available at and around the pixel as well as contextual information. Stained mean is a statistical metric based on pixel level measurement.

A pilot system is being implemented using tissue images with IHC markers to extract features at the knowledge level and the semantic level of harnessing concept. Investigation into the other two levels is underway.

5. References

[1] M. A. Branca,, N. Goodman,, and T.V. Venkatesh, , "*Bioinformatics: Getting Results in the Era of High-Throughput Genomics*", Cambridge Healthtech Institute Report 9, May 2001.

[2] LION Discovery Center, http://www.lionbioscience.com/solutions/discoverycenter.

[3] Oracle, Solutions for Life Sciences, http://www.oracle.com/industries/life_sciences/index.html?content.html.

[4] IBM Life Sciences, http://www-3.ibm.com/solutions/lifesciences/.

[5] EMC, Life Sciences Infrastructure Solutions, http://www.emc.com/vertical/pdfs/life_sciences/interstitial_data_warehouse.jsp.

[6] BioImagene, METHOD AND SYSTEM FOR QUANTITATIVELY ANALYZING BIOLOGICAL SAMPLES, patent application WO2005027015.