KNOWLEDGE REPRESENTATION AND DATA MINING FOR BIOLOGICAL IMAGING

Wamiq M. Ahmed
Purdue University Cytometry Laboratories, Bindley Bioscience Center
1203 W. State Street, West Lafayette, IN 47907
wahmed@flowcyt.cyto.purdue.edu

Biological and pharmaceutical research relies heavily on microscopically imaging cell populations for understanding their structure and function. Much work has been done on automated analysis of biological images, but image analysis tools are generally focused only on extracting quantitative information for validating a particular hypothesis. Images contain much more information than is normally required for testing individual hypotheses. The lack of symbolic knowledge representation schemes for representing semantic image information and the absence of knowledge mining tools are the biggest obstacles in utilizing the full information content of these images. In this paper we first present a graph-based scheme for integrated representation of semantic biological knowledge contained in cellular images acquired in spatial, spectral, and temporal dimensions. We then present a spatio-temporal knowledge mining framework for extracting non-trivial and previously unknown association rules from image data sets. This mechanism can change the role of biological imaging from a tool used to validate hypotheses to one used for automatically generating new hypotheses. Results for an apoptosis screen are also presented.

1. INTRODUCTION

Microscopic imaging is extensively used to image cell samples in two or three spatial dimensions, a spectral dimension, and a temporal dimension.[1] This leads to five-dimensional image sets, and any combination of these dimensions can be acquired depending on the requirements of the application.[2] The general approach is that biologists first develop a hypothesis and then image biological samples to validate their hypotheses. However, images contain much more information than is needed for analyzing a particular hypothesis. For example, a drug-screening study may use an apoptosis assay and image analysis tools to find out which drug is best for inducing apoptosis in cancer cells. This type of analysis, while very useful for a particular application (apoptosis in cancer cells), is not able to extract all of the information from the imaging data. For example, let us assume there is a link between a cell undergoing apoptosis and induction of apoptosis in its neighboring cells after a certain time because of some underlying biological phenomenon. This information, while present in the images collected during the above mentioned drug screening study, will not be extracted. We believe a data mining approach for analyzing biological data can be extremely useful for harnessing the full potential of the information content of biological images. Such an approach can be used to generate new hypotheses and greatly facilitate biological research. Realization of this goal, however, requires the development of schemes for capturing the semantic content of biological images and development of data mining formalisms for extracting association rules. In order to achieve this goal, we present a graph-based knowledge representation scheme that captures the semantic knowledge contained in multi-dimensional biological images. Then we present a framework for extracting non-trivial, previously unknown association rules in such data. This approach can be used for analyzing large repositories of cellular images and can significantly help in biological discovery.

Association-rule mining for knowledge discovery in databases was proposed by Agrawal et al. and has since been extensively used for finding association rules.[3] Application of these tools to imaging data is hampered by the fact that image data require the extraction and representation of semantic information before data mining algorithms can be applied. Classification and clustering techniques have been previously applied to image data in different domains such as medical imaging and weather monitoring[4], but there has not been any work on association-rule mining on cellular images. The challenge lies in developing powerful knowledge representation schemes to capture the semantic information contained in multi-dimensional images and developing formalisms for mining association rules using these schemes. In this paper we propose a graph-based knowledge representation scheme and a data mining formalism for capturing the semantic image information and for extracting association rules. This approach has the potential to exploit the maximum information content of imaging data for automated biological discovery and can potentially change the role of biological imaging from merely a tool for hypothesis validation to a more powerful tool for generating new hypotheses as well.

2. GRAPH-BASED REPRESENTATION OF SEMANTIC CONTENT

Attribute relational graphs (ARGs) have been used for representing image content for content-based retrieval.[5] The nodes of the ARG represent the objects and the edges represent the relations. In order to develop an integrated representation for multi-dimensional biological images that include two or three spatial dimensions, a spectral, and a temporal dimension we extend the concept of an ARG to a colored attribute relational graph (CARG). A CARG is a special ARG where each node of the ARG contains a color attribute that specifies the spectral band (fluorescence channel) in which this image was acquired. Formally a CARG is a four tuple $G = (V, E, A_v, A_e)$ where V is a set of vertices, E is a set of edges between vertices, A_v is a set of attributes of vertices, and A_e is a set of attributes of edges. The vertices represent the objects in the images and vertex attributes contain attributes of objects such as area, perimeter, texture, and shape descriptors. Edge attributes represent the spatial relations between objects. In our experiments we use four spatial relations that include 'overlap' (o), 'contain' (c), 'near neighbor' (n), and 'far neighbor' (f), and four object attributes that include area, major axis length, minor axis length, and perimeter. CARG captures the image information in three spatial dimensions and the spectral dimension. Information in the temporal dimension is captured using a temporal sequence of CARGs each representing the spatial and spectral information at a time instant. An example of a series of CARGs showing 3 cells imaged in three different fluorescence channels for an apoptosis screen is shown in Figure 1 (a-c). Here white, light gray, and dark gray nodes represent Hoechst, Annexin V fluorescein isothiocyanate (FITC), and propedium iodide (PI) respectively. At time instant 1, Cell 1 is in an early apoptotic state (overlap of white and light gray nodes) whereas it is in a late apoptotic state at time instants 2 and 3. Similarly, Cell 3 is in the live state (white node disjoint from other nodes) at time instant 1 and 2 and in an early apoptotic state at time instant 3.

Most biological events involve spatio-temporal changes in the attributes of biological objects (cells, intracellular compartments) or changes in the spatial relations between different objects.[6,7] The ARG model can be used for representing the information about spatio-temporal events and the spatial relations among them. Each node of the ARG represents an event. Attributes of the nodes include the type of event, participating objects along with their attributes, start time, duration, and the decomposition into simpler events for composite events. For example an apoptosis event may be considered to be made up of sub-events such as 'normal' when cell is alive and 'apoptotic' when the cell undergoes apoptosis. Figure 1(d) shows the representation of four apoptosis events as an ARG.

3. DATA MINING FRAMEWORK

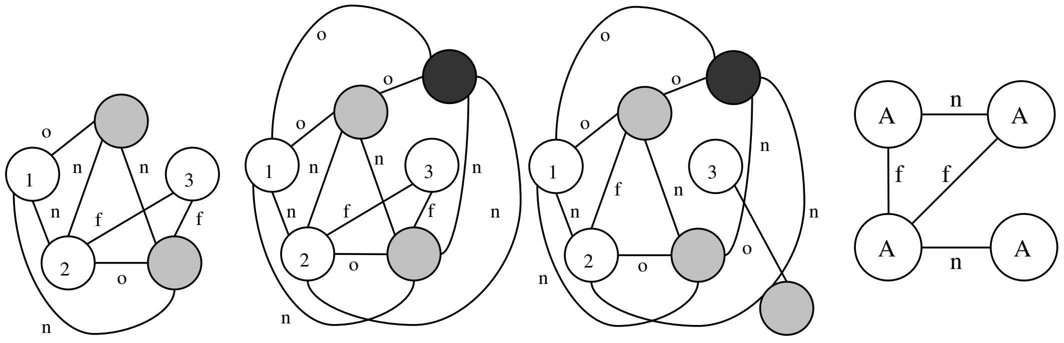
The graph-based knowledge representation scheme proposed in Section 2 provides a data structure for storing image information in terms of the objects and the events happening in the images. Data mining algorithms can then be applied on this symbolic representation. This can help discover interesting patterns in imaging data. Such patterns could be in the form of association rules between different features of biological objects (between roundness and size) or between features of biological objects and different semantic classes of objects (between roundness and mitotic state). These association rules may also have a temporal dependence (between roundness and size after a time interval, or between roundness and cell division after a time interval). In order to capture these patterns we mine six different types of rules as shown in Figure 2. Association rules are generally represented as (X→Y) where X is the antecedent and Y is the consequent. The support and confidence values for mined rules are defined as follows.

Support = Number of transactions where both X and Y appear / Number of transactions in the database.

Confidence = Number of transactions where both X and Y appear / Number of transactions where only X appears.

4. EXPERIMENTAL RESULTS AND DISCUSSION

In this section we report the results of applying the association-rule mining algorithms to the images generated by an apoptosis screen. A fluorescent marker (Hoechst) was used for labeling the nuclei whereas Annexin-V-FITC was used to label cells as apoptotic or non-apoptotic. Nuclear features, including area, major axis length, minor axis length, and perimeter, were extracted. Nuclei neighbors were determined using the distance between the centroids of different nuclei. The extracted features were discretized by dividing the range of each feature into 4 ranges as shown in Table 1.



(a) t = 1 (b) t = 2 (c) t = 3 (d)

Fig. 1. (a-c) A sequence of CARGs as an integrated representation of spatial, spectral, and temporal information for three cells. White nodes represent nuclear stain which is used to identify the cells. Light gray nodes represent Annexin-V-FITC and dark gray nodes represent PI (d) Representation of 4 apoptosis events.

Same object spatial rules

1. $Attr_X(A) \to Attr_Y(A),\ X \neq Y$
2. $Attr_X(A) \to SemClass_Y(A),\ X \neq Y$
3. $SemClass_X(A) \to SemClass_Y(A),\ X \neq Y$
4. $SemClass_X(A) \to Attr_Y(A),\ X \neq Y$

Same object temporal rules

1. $Attr_X(A) \xrightarrow{TempRel} Attr_Y(A),\ X \neq Y$
2. $Attr_X(A) \xrightarrow{TempRel} SemClass_Y(A),\ X \neq Y$
3. $SemClass_X(A) \xrightarrow{TempRel} SemClass_Y(A),\ X \neq Y$
4. $SemClass_X(A) \xrightarrow{TempRel} Attr_Y(A),\ X \neq Y$

Object neighborhood spatial rules

1. $Attr_X(A) \to Attr_Y(B),\ X \neq Y, B \in Neighborhood(A)$
2. $Attr_X(A) \to SemClass_Y(B),\ X \neq Y, B \in Neighborhood(A)$
3. $SemClass_X(A) \to SemClass_Y(B),\ X \neq Y, B \in Neighborhood(A)$
4. $SemClass_X(A) \to Attr_Y(A),\ X \neq Y, B \in Neighborhood(A)$

Object neighborhood temporal rules

1. $Attr_X(A) \xrightarrow{TempRel} Attr_Y(B),\ X \neq Y, B \in Neighborhood(A)$
2. $Attr_X(A) \xrightarrow{TempRel} SemClass_Y(B),\ X \neq Y, B \in Neighborhood(A)$
3. $SemClass_X(A) \xrightarrow{TempRel} SemClass_Y(B),\ X \neq Y, B \in Neighborhood(A)$
4. $SemClass_X(A) \xrightarrow{TempRel} Attr_Y(A),\ X \neq Y, B \in Neighborhood(A)$

Spatial event rules

1. $Event_X(A) \to Attr_Y(A)$
2. $Event_X(A) \to Event_Y(A)$
3. $Event_X(A) \to Event_Y(B),\ A,B \in ObjList(X)$
4. $Event_X(A) \to Event_Y(B),\ B \in Neighborhood(A)$

Temporal event rules

1. $Event_X(A) \xrightarrow{TempRel} Attr_Y(A), TempRel \in \{D,M,O,C,S,E,CO\}$
2. $Event_X(A) \xrightarrow{TempRel} Event_Y(A), TempRel \in \{D,M,O,C,S,E,CO\}$
3. $Event_X(A) \xrightarrow{TempRel} Event_Y(B),\ A,B \in ObjList(X), TempRel \in \{D,M,O,C,S,E,CO\}$
4. $Event_X(A) \xrightarrow{TempRel} Event_Y(B),\ B \in Neighborhood(A), TempRel \in \{D,M,O,C,S,E,CO\}$

Fig. 2. Different types of spatial and temporal rules. $Attr_X(A)$ means attribute X of object A, $SemClass_X(A)$ means semantic class X involving object A, and $Event_X(A)$ means event X involving object A. *TempRel* refers to the temporal relations between different events.[6]

We also use two other features, 'state' and 'nbr,' where state can be either 'live' or 'apoptotic' and 'nbr' can be either 'none,' implying none of the cell's neighbors is in apoptotic state or 'oneplus,' implying one or more of a cell's neighbors are apoptotic. Association-rule mining formalism was then applied to the semantic information extracted from a set of 200 images (100 fields of view x 2 fluorescent channels). Using a minimum support of 0.2 and confidence of 0.6, a total of 90 rules were found. A subset of these rules is shown in Table 2. Some of these rules are obvious such as the dependence of area on the major and minor axis lengths. However the relationship between the apoptotic state of a cell and its features, or that between the apoptotic state of a cell and the state of its neighbors, can be a useful one.

Table 1. Feature ranges used for discretization of features.

Feature	Range1	Range2	Range3	Range4
Area (A)	$0<A_{R1}<=200$	$200<A_{R2}<=400$	$400<A_{R3}<=600$	$600<A_{R4}$
Major axis (Mj)	$0<Maj_{R1}<=20$	$20<Maj_{R2}<=40$	$40<Maj_{R3}<=60$	$60<Maj_{R4}$
Minor axis (Mi)	$0<Min_{R1}<=10$	$10<Min_{R2}<=20$	$20<Min_{R3}<=30$	$30<Min_{R4}$
Perimeter (P)	$0<P_{R1}<=50$	$50<P_{R2}<=100$	$100<P_{R3}<=150$	$150<P_{R4}$

Table 2. Mined rules with support and confidence values.

No.	Antecedent	Consequent	Support	Confidence
1	Mj = Maj_{R1}	Mi = Min_{R2}	0.447	0.981
2	Mi = Min_{R2}	A = A_{R2}	0.585	0.778
3	State = live	Nbr = none	0.244	0.854
4	State = apoptotic, Mj = Maj_{R1}	Mi = Min_{R2}	0.274	0.992
5	State = apoptotic, A = A_{R2}	P = P_{R2}, Mi = M_{R2}	0.432	0.836
6	Nbr = oneplus	State = apoptotic	0.221	0.841
7	P = P_{R2}, Mi = Min_{R2}	A = A_{R2}	0.582	0.819
8	Mi = Min_{R2}, Nbr = none, State = apoptotic	A = A_{R2}	0.287	0.831
9	P = P_{R2}, A = A_{R2}, Mj = Maj_{R2}	Mi = Min_{R2}	0.301	1
10	A = A_{R2}, Mj = Maj_{R1}	P = P_{R2}, Mi = Min_{R2}	0.301	1

5. CONCLUSION

Mining association rules from cellular images can be a powerful tool for discovering new biological knowledge in an automated manner. In this paper we have presented a graph-based model for representation of the semantic content of cellular images and a formalism for mining association rules at the level of object features and at the level of biological events. Our experiments did not involve temporal data mining, although our formalism provides for mining temporal association rules. In the future we plan to generate significantly large data sets for applying our data mining formalism on spatial as well as temporal data.

References

1. Ahmed WM, Leavesley SJ, Rajwa B, Ayyaz MN, Ghafoor A, and Robinson, JP. State of the art in information extraction and quantitative analysis for multi-modality biomolecular imaging *Proceedings of the IEEE* 2008; **96, 3**: 512-531.
2. Swedlow JR, Goldberg I, Brauner E, Sorger P K. Informatics and Quantitative Analysis in Biological Imaging. *Science* 2003; **300, 5616**: 100-102.
3. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases *Proc.of the ACM SIGMOD Conference on Management of Data* 1993; 207-216.
4. Antonie M, Zaiane OR, Coman A. Application of data mining techniques for medical image classification *Second International ACM SIGKDD Workshop on Multimedia Data Mining* 2001; 94-101.
5. Petrakis EGM, Faloutsos C, Lin K. ImageMap: An image indexing method based on spatial similarity *IEEE Transactions on Knowledge and Data Engineering* 2002; **14, 5**; 979-987.
6. Ahmed WM, Ghafoor A, Robinson JP. Knowledge extraction for biological imaging *IEEE Multimedia* 2007; **14, 4**: 52-62.
7. Ahmed WM, Lenz D, Liu J, Robinson JP. XML-based data model and architecutre for a knowledge-based grid-enabled problem-solving environment for high-throughput biological imaging *IEEE Transactions on Information Technology in Biomedicine* 2008; **12, 2**: 226-240.