Current Challenges in Bioimage Database Design

Ambuj K. Singh Arnab Bhattacharya Vebjorn Ljosa
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA 93106-5110
{ambuj, arnab, ljosa}@cs.ucsb.edu

Abstract

Information technology research has played a significant role in the high-throughput acquisition and analysis of biological information. The tremendous amount of information gathered from genomics in the past decade is being complemented by knowledge from comprehensive, systematic studies of the properties and behaviors of all proteins and other biomolecules. Understanding complex systems such as the nervous system requires the high-resolution imaging of molecules and cells and the analysis of these images in order to understand how distribution patterns (e.g., the localization of specific neuron types within a region of the central nervous system, or the localization of molecules at the subcellular level) change in response to stress, injury, aging, and disease. We discuss two kinds of bioimage data: retinal images and microtubule images. We argue that supporting effective access to them requires new database techniques for description of probabilistic and interpreted data, and analysis of spatial and temporal information. The developed techniques are being implemented in a publicly available bioimage database.

1. Introduction

The understanding of complex cellular behavior and systems is critically enhanced by the capture and analysis of biological images. Significant progress in our understanding of biological events can be made by applying advances in information technologies, such as image processing, pattern recognition, and databases, to the enormous volume of such images that are being generated.

Bioimage databases obtain their semantics and utility through tools that interpret them. At a primitive level, we have tools that detect edges, segments and contours. Then, we have tools that assign likely labels (from a domain-specific ontology) to these extracted objects. At the next level, we have tools that infer group-specific characteristics (e.g., a retinal layer consisting of cells of a specific type, or the pattern of neurons entering a collection of cells). Finally, we have tools that gather spatio-temporal attributes from a collection of images and relate them to higher-level models of change (e.g., a disease or response to stimuli). The multitude of such tools that interpret the observed data as well as other interpretations results in a complex collection of values and objects. Supporting the above set of tasks calls for a database design that is flexible, can support multiple layers of inherently probabilistic interpreted information, and supports a collection of analysis tools.

Typical queries on bioimages are posed using a combination of attribute-based descriptors (querying of metadata) and content-based descriptors (querying of raw and processed data). Typically, the raw images are analyzed, and visual descriptions extracted (manually or automatically). This analysis can lead to a multifold increase in the amount of storage and complexity. Clearly, the amount of information to be maintained and accessed in such databases is complex and enormous. Based on the degree of semantics and interpretation, queries in image databases can be divided into a number of types. At the basic level, we have queries that use only the experimental conditions and other tags easily associated with the images. Next, we have queries on the spatial features extracted from the images. For example, images with a specified subcellular localization pattern can be found by extracting texture features and searching using a suitable distance metric. Then, we have queries based on high-level semantic interpretations, such as cell types, that are extracted from the images manually or automatically. Finally, we have queries that are based on spatio-temporal changes of features and high-level objects such as protein localization.

The rest of the paper is organized as follows. Section 2 discusses the biological background of retinal and microtubule images, the two main components of our bioimage database. Section 3 discusses distance metrics and index structures useful for spatial analysis of bioimages. Section 4 discusses probabilistic data and queries. Section 5 dis-

cusses the temporal analysis of bioimages. We end with some concluding remarks in Section 6.

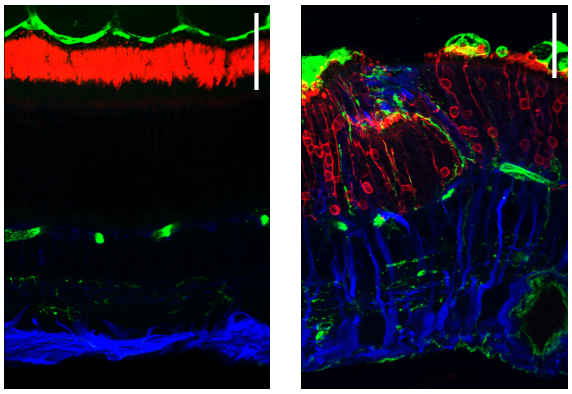
2. Retinal and microtubule image datasets

The vertebrate retina has a layered structure in which each layer consists of well-defined cell types. There are five kinds of neurons: photoreceptors (rods and cones), horizontal cells, bipolar cells, amacrine cells, and ganglion cells. The primary flow of the electrical signals generated by the photoreceptors in response to light stimuli is from the photoreceptors to the bipolar cells to the ganglion cells, and finally through the optic nerve into the rest of the brain. Horizontal and amacrine cells provide lateral connections across the retina. Non-neurons such as Müller cells, astrocytes, and microglia are also present in the retina. The retina can be divided into a number of layers: retinal pigment epithelium (RPE), outer segment (OS), inner segment (IS), outer nuclear layer (ONL), outer plexiform layer (OPL), inner nuclear layer (INL), inner plexiform layer (IPL), and ganglion cell layer (GCL).

Under injury, many proteins show remarkable variability in their spatio-temporal distribution because of changes both to the morphology of the retina and to the distribution of proteins in each cell. These distributions are observed by coupling fluorophores to protein-binding antibodies and other labels. Sections of the retina are then observed under a confocal, laser-scanning microscope. In Figure 1, we see that in response to detachment, the protein rhodopsin redistributes so as to label not only the outer segments of the rods, as in Figure 1(a), but also the rod cell bodies (in the outer nuclear layer), as in Figure 1(b) [5]. We also see that the Müller cells, which express glial fibrillary acidic protein, hypertrophy [4]. These images form the raw image data for the ensuing spatio-temporal analysis.

Microtubules are long, hollow, unbranched protein structures found within cells. They are about 25 nm in diameter and up to several μm long. They are polymers of tubulin and formed by assembly of α and β subunits. The microtubules are part of the cytoskeleton, but in addition to structural support they are used in many other processes. They are capable of growing and shrinking in order to generate force, and there are also motor proteins that move along the microtubule. Microtubules play a major role in cell division, where they attach to the chromosomes in order to segregate them correctly.

Gain or loss of tubulin subunits from the (+) end can change the length of a microtubule. Growth occurs by the assembly of GTP-bound tubulins, and shrinking occurs by the hydrolysis of GTP to GDP. The switch from growth to shrinking is called a *catastrophe* event and the switch from shrinking to growth is called a *rescue* event. The dynamic equilibrium between these events changes under different conditions such as drug concentrations. For example, the drug taxol makes the microtubules more stable (reduces the likelihood of shrinking), while Colchicine has the opposite effect [2]. Microtubule dynamics is observed through a stack of images taken at different time points. These form the raw image data for microtubule behavior analysis.



(a) Normal (b) Detached for 28 days

Figure 1. Normal cat retinas labeled with anti-rhodopsin (red), anti-glial fibrillary acidic protein (anti-GFAP, green), and isolectin B4 (blue). Scale bars: 50 μm

3. Spatial analysis of bioimages

Supporting efficient access to the complex bioimage data and metadata requires the design of appropriate distance metrics and index structures. For many classes of images, including retinal images, the spatial location of objects (e.g., layers, cells or parts of cells) in the image is important in determining whether two images should be considered similar. The earth mover's distance (EMD), first proposed by Werman et al. [11] and later applied to computer vision problems by Peleg et al. [8], captures the spatial aspect of the different features extracted from the images. The distance between two images measures both the distance in the feature space and the spatial distance.

Though the EMD is theoretically attractive and is a better distance function for images than L_p norms [9], it involves solving a linear programming problem, which may take a long time. For example, for 256-dimensional features extracted from images that are partitioned into 8×12 tiles, each EMD computation takes 27 s, so, a similarity search on a database of 4,000 images can take almost 30 h to complete!

We have proposed the *LB-index*, a multi-resolution approach to computing the EMD [6]. We formulate the EMD in a new way that works directly with feature vectors of any dimensionality without requiring the images to have the same total "mass." The formulation allows us to condense

Rows	Columns	Number of bins	Time [s]	Ratio of distance to actual distance
2	3	6	0.13	86%
4	6	24	0.58	93%
8	12	96	26.68	100%

Table 1. Average time to compute and average ratio to the actual distance between two histograms as a function of histogram size for 256-dimensional CSD feature vectors.

the representation of an image in feature space into progressively coarser summaries. We have developed lower bounds for the EMD. The lower bounds can be computed from the summaries at various resolutions, and then applied to the problem of similarity search (k-nearest-neighbor search and range search) in an image database. Higher-level lower bounds are less tight, but less expensive to compute.

For 256-dimensional Color Structure Descriptor (CSD) [7], Table 1 gives the average running time for different resolutions, as well as the average ratio of the lower bound distance computed from a coarser resolution to the actual distance. Using different levels of lower bounds for sequential scans can speed up range queries and k-NN queries by factors of 30–60 on a database of 3,932 cat retinal images. We also applied the lower bounds to a variant of the M-tree [1] algorithms. This can lead to a speedup of factors of 5–25 for range and k-NN queries compared to the original M-tree algorithms.

4. Probabilistic data and queries

Probabilistic data are abundant in scientific databases: Measurements have varying accuracy (numerical values, spatial, or temporal extents), and computational methods produce results of varying confidence. For example, a program that identifies cells in fluorescent confocal retinal micrographs may output a region of the image, and state that the region is 45 % likely to be a horizontal cell body and 30 % likely to be an amacrine cell body. Managing and querying probabilistic data is challenging. Relevant scientific questions include schema design for modeling, storing and accessing probabilistic information, as well as support for different kinds of queries (top-k, range, similarity, and join) and discovery (outlier detection, classification) on such data.

If we have a model for the imprecision in images, an imprecise value can be represented as a probability distribution. Probability distributions can also summarize a population of exact measurements. As an example of the latter, consider retinal micrographs. Much can be learned about the effect on the retina of diseases and treatments by measuring the thickness of these layers. In Figure 1(b), retinal detachment has caused some of the photoreceptors to die. As a result, the outer nuclear layer (ONL), which contains photoreceptor cell bodies, has become thinner. The thickness of the layer varies, however, so measuring it in just one place is useless. Keeping only the mean of several measurements is also insufficient, as the spread of the measurements is important: Highly variable thickness is indicative of degenerate retinas.

Whatever the source of the probability distribution, it can be represented in different ways. Fitting the distribution to a parametric distribution such as a Gaussian makes sense when the source of uncertainty is well understood. In contrast, histograms have the advantage of being able to represent arbitrary distributions without an a priori assumption of what the distribution should look like. This is important because it supports data-driven research, in which new hypotheses are generated by analysis of the wealth of poorly understood measurements (raw data) in the database. For instance, the distribution of a certain measurement may appear Gaussian with a small amount of noise, but careful analysis of measurements from a large number of images may reveal that some of the supposed noise is really a systematic contribution from a previously unknown mechanism. If the data were just fitted to a Gaussian, the mechanism would elude discovery.

Groups of bins in a histogram can be merged hierarchically to yield a multi-tier histogram that can accelerate database queries. Another possible way of summarizing distributions is to take spatial locality into account. In this case, one obtains a space-varying measure of a feature (e.g., layer thickness). These can again be obtained at varying spatial resolutions.

When more than one feature is of interest, there are two alternatives, depending on whether spatial variations in an image need to be coordinated. When there is no need to coordinate the feature distributions spatially, one can extract independent histograms from an image and combine them using range queries or NN-queries. When the variations in space need to be coordinated, the distributions become multi-variate, and storing and querying them becomes more difficult.

Next, we present some examples of queries that can be posed using the above probability distributions. Consider the example of querying retinal images for understanding layer thicknesses. A range query $\mathrm{R}(t, p)$ asks for all images for which a particular layer is thicker than t units with a probability at least p. This query can be answered using a hierarchical index structure of histograms: Raw image histograms are stored at the leaves, histograms that are similar are aggregated, and a subtree of leaf nodes is represented by a min–max histogram storing the lower and the upper bound values for each bin. A nearest-neighbor query $\mathrm{NN}(t, k)$ asks

for the k images that have the highest probability of being the closest to a thickness of t units. This query is much harder to answer: The thickness distributions of all images have to be considered before we can decide the k nearest images. An image with a distribution B *dominates* another image with a distribution C if the expected distance of B to t is less than the expected distance of C to t. (Alternative formulations are also possible). This relationship defines a partial order and the top k images in the partial order are returned by the nearest neighbor query. Speedups are obtained if approximate moments are computed using multi-level histograms, and the approximate values are used to prune the solution space. The most interesting query gives an input distribution (or, an input image) and asks for the best matches. To answer this query, we need to compare the distance between two distributions. The EMD discussed earlier is useful for such comparisons. It defines the metric space for returning the best matches.

In all the above queries, questions of scale are an important factor when comparing images. Two images should be normalized to the same scale (resolution) before we can compare them. This information is available in the meta-data associated with each image. Query algorithms must also take into account that images are generally of different size after normalization.

Other important queries, like "what is the typical thickness distribution of the INL after the retina has been detached for 28 days?" or "does a given treatment have any effect on detached retinas?", require images to be clustered and classified by their distributions. Distance metrics such as the EMD can be used along with existing data mining algorithms. Another possibility is to represent a distribution as a mixture of primitive distributions (such as Gaussian), and to represent a class of distributions by a set of appropriate parameters.

Because of probabilistic data, developing a schema for our image database has proven to be more difficult than expected. The parts of the schema that store the raw images, the details for the acquisition process, and simple image features build on previous systems, such as the Open Microscopy Environment (OME) [10]. Our contributions pertain to the part of the schema that stores information interpreted from the images by image analysis methods. This information differs from simple image features by (1) having semantic (biological) meaning and (2) being probabilistic in nature. Examples include a count of bipolar cell bodies, and the thickness of the inner nuclear layer of the retina.

The semantic information also differs from simple features in another, crucial way: It is necessary to track its lineage. Whereas simple image features are well-defined and can only be computed correctly in one way, semantic information can have many different origins, including manual entry and analysis methods of varying reliability.

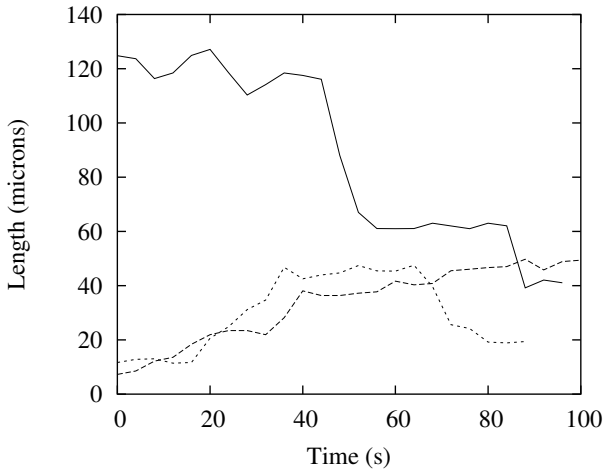


Figure 2. Example life-history plots of microtubules. The x-axis shows time in seconds and the y-axis shows length of the microtubules in μm. Three individual microtubules are shown.

Some semantic information, such as the count of bipolar cell bodies in the outer nuclear layer, are simply modeled as univariate probability distributions. Whereas there are many open questions about how to index and query such distributions, they are fairly simple from the perspective of designing a schema. This is not the case when it comes to modeling the spatial locations and extents of objects identified in the image. It is not clear how to model, for instance, an axon, a cell body, or a layer of cells.

5. Understanding temporal behavior

Bioimages can be captured at different time instants to understand the temporal processes behind a disease or behavior. Two such examples are retinal images recorded after different periods of retinal reattachment and microtubule behavior under different conditions. In this section, we summarize a temporal study of microtubule behavior.

Microtubules in a cell change their behavior of growth and shortening under different drug injections and different drug concentrations. The dynamic behavior of microtubules is an extremely important biological process. For example, the dysfunction of the microtubule-associated protein *tau* has been correlated with a variety of neurodegenerative diseases, including Alzheimer disease, fronto-temporal dementia with Parkinsonism associated with chromosome 17 (FTDP-17), Pick disease, and progressive supranuclear palsy [3].

To ascertain the growth and shortening behaviors under different conditions, biologists record videos of the microtubules both in vivo and in vitro and track individual micro-

Cluster	Conditions		
	Buffer	3R tau	4R tau
Cluster 1	17	10	-
Cluster 2	-	12	4
Cluster 3	-	-	18

Table 2. The distribution of microtubules in the three clusters.

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0.00	40.70	50.00
Cluster 2		0.00	37.15
Cluster 3			0.00

Table 3. The inter-cluster distances. "Buffer" is closer to "3R tau" than to "4R tau."

tubules. The length of a microtubule versus the time is then plotted as a "life-history" plot. Figure 2 shows three such life-history plots of microtubules.

Each life-history plot is a time series, so there is a collection of such time series for each condition. We need to understand the patterns within a collection, and also how different collections relate to one another. Such analysis can be improved by data transformation (such as Fourier, wavelet, or discrete cosine transform). The resulting summaries can be clustered or classified to provide new biological insights, e.g., how periodicities change under a treatment, or whether a drug leads to similar growth dynamics as another drug.

A particularly useful analysis is the comparison of microtubule behavior in different concentrations of wild-type three-repeat tau (3R tau) and wild-type four-repeat tau (4R tau) [3]. We computed the Fourier transformation of the time series of microtubules of three different conditions, then clustered them. The conditions were "Buffer", "3R tau", and "4R tau". Table 2 shows the results of the clustering. Clusters corresponding to each condition can be identified. Table 3 shows the distances among the three clusters. We see that microtubules treated with different wild-type tau proteins behave more similarly than when not treated with tau. We also observe that "Buffer" is closer to "3R tau" than to "4R tau." This indicates that the effect of 4R wild-type tau is stronger than the effect of 3R wild-type tau.

6 Conclusion

Many areas of biology depend heavily on the acquisition and analysis of images. Advances in database field are necessary in order to efficiently store, query, and analyze high volumes of such images.

In this paper, we have discussed requirements for bioimage databases. Motivated by real datasets, we have described challenges for bioimage databases and outlined possible solutions. Bioimages must provide distance measures that are sensitive to spatial locality, support probabilistic data and queries, and support mining of interpreted spatio-temporal data.

Acknowledgements

We would like to thank Geoffrey P. Lewis from the laboratory of Steven K. Fisher at UCSB for providing the retinal images. We would also like to thank Stuart C. Feinstein and Leslie Wilson for providing microtubule data collected in their laboratories. This work is supported by the National Science Foundation under ITR-0331697. The project URL is http://www.bioimage.ucsb.edu.

References

[1] P. Ciaccia, M. Patella, and P. Zezula. M-tree: An Efficient Access Method for Similarity Search in Metric Spaces. In *Proc. VLDB*, pages 426–435, 1997.

[2] J. S. Hyams and C. W. Lloyd. *Microtubules*. Wiley-Liss, Inc., 1994.

[3] S. F. Levy, A. C. LeBoeuf, M. R. Massie, M. A. Jordan, L. Wilson, and S. C. Feinstein. Three- and Four-Repeat Tau Regulate the Dynamic Instability of Two Distinct Microtubule Subpopulations in Qualitatively Different Manners: Implications for Neurodegeneration. *Journal of Biological Chemistry*, 280(14):13520–13528, 2005.

[4] G. P. Lewis and S. K. Fisher. Müller cell outgrowth after retinal detachment: Association with cone photoreceptors. *Investigative Ophthalmology & Visual Science*, 41(6):1542–1545, 2000.

[5] G. P. Lewis, C. S. Sethi, K. A. Linberg, D. G. Charteris, and S. K. Fisher. Experimental retinal detachment: A new perspective. *Molecular Neurobiology*, 28(2):159–175, Oct. 2003.

[6] V. Ljosa, A. Bhattacharya, and A. K. Singh. LB-index: A multi-resolution index structure for images. In preparation.

[7] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley, 2002.

[8] S. Peleg, M. Werman, and H. Rom. A Unified Approach to the Change of Resolution: Space and Gray-Level. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:739–742, 1989.

[9] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[10] J. R. Swedlow, I. Goldberg, E. Brauner, and P. K. Sorger. Informatics and quantitative analysis in biological imaging. *Science*, 300:100–102, 2003.

[11] M. Werman, S. Peleg, and A. Rosenfeld. A Distance Metric for Multi-Dimensional Histograms. *Computer, Vision, Graphics, and Image Processing*, 32(3):328–336, 1985.