

Toward Data-Driven Research on Biomedical Images

Statement of Research

Vebjorn Ljosa
ljosa@cs.ucsb.edu
www.ljosa.com

1 Introduction

The DNA sequences in Genbank now total more than 100 billion basepairs, ten times as much as just five years ago. More impressive than the sheer increase in size, however, is how the data, combined with novel techniques for searching and mining them, have led to a new, data-driven way of doing research: Large-scale analysis has become a rich source of new hypotheses, and has led to discoveries that would otherwise go unnoticed.

Whereas data-driven research is well established for genome and gene expression (microarray) data, and significant inroads have been made for protein interaction networks, it is still in its infancy for microscopy images. Robotic microscopes and motorized stages allow for high-throughput image acquisition, and databases of images are becoming available [12, 14, 15], but much work is needed for effective search, analysis, and mining. The work can be expected to yield a tremendous scientific return, however, for microscopy is at the very core of almost every field of biology. Micrographs capture the morphology of everything from subcellular structures to entire organisms. With appropriate labeling, they become quantitative measurements of the distribution of proteins in cells. Finally, live-cell imaging provides a window into the dynamic behavior of a biological system.

Many of the database and mining challenges posed by large image datasets stem from the gulf between image data and their biological meaning: An image is just a collection of pixels (or voxels), but the queries and the notions of similarity or distance are in the domain of biological objects. For instance, a neuroscientist may want to search for images that show Müller cells that extend into the outer nuclear layer (which normally contains photoreceptor cell bodies).

It may in some cases be possible to circumvent the gulf. For instance, my colleagues and I have worked with image similarity at the image level [7, 8]. We have also developed a method for constructing a vocabulary of descriptive terms directly from image data [1]. Such methods work best, however, as exploratory tools. In order to gain insight into biological mechanisms, they must be complemented by mining

and searching measurements of biological objects and processes—in other words, by bridging the semantic gap.

Through working with biological images, it has become clear that segmenting, classifying, and tracking cells and other objects is an inexact endeavor at best. Even experts cannot perform these analysis tasks reliably in many cases. Automated techniques may perform well on some cells and not so well on others. Consequently, the key to making use of their results is to have the analysis techniques give confidence values for their results and develop mining and database techniques that work with such probabilistic values [9, 10]. (Probabilistic values also have applications in sensor networks, moving object databases, and certain business databases.)

Storing, indexing, and searching probabilistic values poses many technical challenges from a database point of view. The nature of queries changes dramatically when probability is introduced. As an example, the result of a range query becomes a graded set, where each object is associated with an *appearance probability*, i.e., the probability that the object is in the query range. Answering a query requires new index structures that quickly prune objects with low appearance probability.

I have developed the *APLA-tree* [11], an index structure that can answer range and k-NN queries on arbitrary probability distributions efficiently, but challenges remain. For instance, many biological applications map naturally to spatial join queries, which are expensive to compute on probabilistic datasets. Additional difficulties arise when objects cannot be treated as points, but must be modeled as having (uncertain) extent. Finally, there are many open questions about how to mine uncertain data.

2 Research Highlights

2.1 Probabilistic Biomedical Data

Analysis of biomedical images yields probabilistic values, either because individual measurements are imprecise or because the value summarizes a population of measurements [11]. Queries on probabilistic values have been studied by several authors [2, 3, 4, 5], but under the assumption that the pdfs are uniform or Gaussian. Fitting a model distribution, such as a Gaussian, to the data is only appropriate when the data are well understood. In a scientific database, the most interesting data to query are precisely the ones that are *not* well understood. We are therefore concerned with queries on *arbitrary* pdfs.

Our index structure, the *APLA-tree* [11], represents each distribution by a sequence of linear approximations. By constraining where the lines intersect, the domain of each approximation becomes implicit. This gives the combination of accuracy and compactness necessary to answer queries twice as fast as existing methods. A new formulation of probabilistic k-NN queries allows the same index structure to answer them—so far the only efficient method for answering k-NN queries on arbitrary probability distributions.

Using the *APLA-tree*, a biologist can quickly find “images with outer nuclear layer thickness between 30 and 50 μm ” and “the ten images with the thinnest inner nuclear layer.” The thickness of layers of cells in the retina is related to cell death and injury

response, and queries such as these make it possible to find relevant images from past experiments.

Another part of my work is concerned with extracting a neuron’s morphology, which is crucial to understanding its function and behavior. What is the shape of the cell, and where is it located relative to other cells? In order to answer these questions, we need to *segment* the cell, i.e., find out which pixels (or voxels) belong to it. This, however, has turned out to be difficult, e.g., for horizontal cells in the retina. We have proposed the idea of *probabilistic segmentation* [9, 10], where each pixel is assigned a probability of belonging to the cell. The segmentation result (the *probabilistic mask*) is then a probabilistic value. An algorithm based on repeated random walks computes the probabilistic mask. The probabilistic mask provides additional information for cytometry and high-level mining methods. As a very simple example, in order to measure the thickness of a neurite, we can use the probabilistic mask to compute a probability density function for the thickness—in other words, compute the thickness as a probabilistic value.

2.2 Indexing Spatially Sensitive Distance Measures

Content-based search, clustering, and outlier detection require a notion of distance between images. For many classes of images, spatial location is important for whether two images should be considered similar. For instance, both macrophages and microglia express isolectin B4, so the location in the retina is crucial in order to tell them apart. The earth mover’s distance (EMD), a distance metric that takes spatial location into account, yields superior results, but its computational cost can be prohibitive, especially with high-dimensional image features and fine-grained spatial structures.

In order to make the EMD a feasible alternative, we developed a multi-resolution indexing approach [7, 8]. We derived practical lower bounds for the EMD, and incorporated multiple levels of lower bounds, one for each resolution of the index structure, into algorithms for answering range queries and k-NN queries, both by sequential scan and using an M-tree index structure. Experiments show that using the lower bounds reduces the running time of similarity queries by a factor of up to 36. Computing separately for each dimension of the feature vector yields a speedup of ~ 14 . By combining the two techniques, similarity queries can be answered more than 500 times faster.

2.3 Visual Vocabulary Construction

Given a large collection of medical images of several conditions and treatments, how can we succinctly describe the characteristics of each setting? For example, given a large collection of retinal images from several different experimental conditions (normal, detached, reattached, etc.), how can data mining help biologists focus on important regions in the images or on the differences between different experimental conditions?

If the images were text documents, we could find the main terms and concepts for each condition by existing information retrieval methods (e.g., *tf/idf* and LSI). My colleagues and I proposed something analogous, but for the much more challenging case of an image collection: We proposed to automatically construct a *visual vocabulary* by breaking images into tiles and deriving key tiles (“ViVos”) for each image and

condition [1]. We experimented with numerous *domain-independent* ways of extracting features from tiles (color histograms, textures, etc.), and several ways of choosing characteristic tiles (PCA, ICA).

We performed experiments on two disparate biomedical datasets. The quantitative measure of success is classification accuracy: Our “ViVos” achieve high classification accuracy (up to 83 % for a nine-class problem on feline retinal images). More importantly, our “ViVos” do an excellent job as “visual vocabulary terms”: they have biological meaning, as corroborated by domain experts; they help spot characteristic regions of images, exactly like text vocabulary terms do for documents; and they highlight the differences between pairs of images.

2.4 String Alignment by Indexing Frequency Vectors

Data-driven research is, of course, well-established in genomics. Even so, eukaryotic genomes can be billions of nucleotides long, and aligning whole genomes—or any pair of long DNA sequences—is computationally challenging. We tackle the problem of speeding up whole-genome alignment by translating the two sequences to paths in a four-dimensional space [6]. The problem of aligning the sequences then becomes the problem of finding places where the two paths are close to each other. The key to the speedup is to build an index structure on the points so points close together in 4-space are combined into a box. We then perform an index-based join on the boxes, thus making pruning and scheduling decisions on a higher level. By contrast, the existing, hash-table-based or suffix-tree-based, techniques are intrinsically bound to work at the level of single nucleotides.

Although there is little connection to biomedical images, this technique serves as a useful example of how attacking a well-studied problem with a different toolchest of ideas (in this case database techniques) can be profitable.

3 Research Methodology

I have worked closely with biologists for several years, and have learned that spending time to understand the application domain is a good investment. By studying the biological systems and the questions being asked about them, new computer science problems invariably emerge—problems that are interesting in their own right and also apply to other domains.

Focusing on specific problems in biology has many benefits. First, a problem that initially seems to be solvable by existing techniques may turn out to require new solutions when the particular constraints of the application are taken into account. (For instance, probabilistic spatial joins become necessary when cells cannot be reliably segmented.) Second, applying novel algorithms to a specific problem adds a feedback loop to the research process and leads to insight beyond analysis and simulation. Third, applying new techniques to the actual problem is a way of giving back to collaborators in biology, laying the foundation for further collaboration. Finally, deploying the algorithms outside computer science provides final validation of its effectiveness and usefulness. To increase impact, I plan to make implementations publicly available.

I plan to take on two PhD students the first year, then add one student to the group each of the next four years. As de facto advisor for summer students in an outreach program and as mentor for younger PhD students, I have gained some experience in judging when to assume a hands-on advisor role and when to leave a student alone with a problem. I prefer my students to start doing research already in their first year, before completing required coursework. My graduate classes will include research projects, and will be the primary means of recruiting students. Many students will find it easier to write their first papers working closely together with another student, then let their work evolve into an independent thesis topic.

4 Future plans

In ongoing work, I plan to examine the effect of probabilistic data on spatial joins and on mining for patterns of spatial colocation. How does working with probabilistic datasets change the nature of spatial joins? First of all, not all result pairs are equally desirable: A pair of high-confidence points is more important than a pair of low-confidence points. Add to this that the datasets generally have many low-confidence points and few high-confidence points, and it becomes clear that the spatial join is changing into a top- k query: Given two sets of probabilistic points, find the k top-scoring pairs according to a ranking function that takes into account the confidence values of the points and the distance between them. Answering such queries in reasonable time requires an effective heuristic for finding good pairs early in the search so other pairs can be pruned. Preliminary results show that order statistics provide a promising basis for such heuristics.

Another aspect of probabilistic image data is the fact that objects have non-zero extent (they are not points), yet cannot be segmented reliably. We can, however, compute segmentations as probabilistic values (see Section 2.1). I plan to develop efficient index structures for such objects.

Together with biologists from Steven Fisher's retinal laboratory, I plan to investigate the response of horizontal cells to retinal detachment and to the genetic disorder retinoschisis. It is already known that horizontal cells extend away from their usual position in the outer plexiform layer, toward either the inner or the outer retina, under both these conditions. A quantitative study is needed, however, to determine whether (and how) the response is different between the two conditions and between detachments of different lengths.

I plan to address the many open questions about how to perform clustering, outlier detection, and spatio-temporal correlation detection on probabilistic data. The only existing clustering technique for uncertain data can only handle uniform distributions [13].

Content-based browsing and retrieval is crucial for a large database of biomedical images to be more than an expensive file cabinet, but has so far seen limited use because the perceptual distance between two images varies with the interest of the user and because the lack of ground truth makes feature selection difficult. I would like to explore relevance feedback on such image datasets, using the ideas behind our visual vocabulary [1].

References

- [1] A. Bhattacharya, V. Ljosa, J.-Y. Pan, M. R. Verardo, H. Yang, C. Faloutsos, and A. K. Singh. ViVo: Visual vocabulary construction for mining biomedical images. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM)*, pages 50–57, 2005.
- [2] C. Böhm, A. Pryakhin, and M. Schubert. The Gauss-tree: Efficient object identification in databases of probabilistic feature vectors. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, 2006.
- [3] R. Cheng, D. V. Kalashnikov, and S. Prabhakar. Querying imprecise data in moving object environments. *IEEE Transactions on Knowledge Engineering*, 16(9):1112–1127, Sept. 2004.
- [4] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. S. Vitter. Efficient indexing methods for probabilistic threshold queries over uncertain data. In *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, 2004.
- [5] A. Faradjian, J. Gehrke, and P. Bonnet. GADT: A probability space ADT for representing and querying the physical world. In *Proc. ICDE*, pages 201–211, 2002.
- [6] T. Kahveci, V. Ljosa, and A. K. Singh. Speeding up whole-genome alignment by indexing frequency vectors. *Bioinformatics*, 20(13):2122–2134, 2004.
- [7] V. Ljosa, A. Bhattacharya, and A. K. Singh. Indexing spatially sensitive distance measures using multi-resolution lower bounds. In *Proceedings of the 10th International Conference on Extending Database Technology (EDBT)*, pages 865–883, 2006.
- [8] V. Ljosa, A. Bhattacharya, and A. K. Singh. LB-index: A multi-resolution index structure for images. In *Proceedings of the 22nd International Conference on Data Engineering (ICDE)*, 2006.
- [9] V. Ljosa and A. K. Singh. Probabilistic segmentation of horizontal cells. In *Proceedings of the 2006 workshop on multiscale biological imaging, data mining & informatics (Bioimage Informatics)*, pages 39–40, 2006.
- [10] V. Ljosa and A. K. Singh. Probabilistic segmentation and analysis of horizontal cells. In *Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM)*, Dec. 2006.
- [11] V. Ljosa and A. K. Singh. APLA: Indexing arbitrary probability distributions. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE)*, 2007. To appear.
- [12] M. Martone, S. Zhang, A. Gupta, X. Qian, H. He, D. Price, M. W. M, S. Santini, and M. Ellisman. The Cell-Centered Database: A database for multiscale structural and protein localization data from light and electron microscopy. *Neuroinformatics*, 1(3):379–396, 2003.
- [13] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, and K. Y. Yip. Efficient clustering of uncertain data. In *Proceedings of the Sixth IEEE International Conference on Data Mining (ICDM)*, pages 436–445, 2006.
- [14] A. K. Singh, B. Manjunath, and R. F. Murphy. A distributed database for bio-molecular images. *SIGMOD Record*, 33(2):65–71, 2004.
- [15] J. R. Swedlow, I. Goldberg, E. Brauner, and P. K. Sorger. Informatics and quantitative analysis in biological imaging. *Science*, 300:100–102, 2003.

All papers and application materials are available from www.ljosa.com.