Extracting meaning from biological imaging data

Andrew R. Cohen, Department of Electrical and Computer Engineering
Drexel University, Philadelphia, PA 19104, acohen@coe.drexel.edu http://bioimage.coe.drexel.edu

Abstract

Biological imaging continues to improve, capturing continually longer-term, richer and more complex data, penetrating deeper into live tissue. How do we gain insight into the dynamic processes of disease and development from terabytes of multidimensional image data? Here I describe a collaborative approach to extracting meaning from biological imaging data. The collaboration consists of teams of biologists and engineers working together. Custom computational tools are built to best exploit application-specific knowledge, to visualize and analyze large and complex datasets. The image data is *summarized*, extracting and modeling the features that capture the objects and relationships in the data. The summarization is *validated*, the results visualized and errors corrected as needed. Finally, the customized analysis and visualization tools together with the image data and the summarization results are *shared*. This perspective will provide a brief guide to the mathematical ideas that rigorously quantify the notion of extracting meaning from biological images, and to the practical approaches that have been used to apply these ideas to a wide range of applications in cell and tissue optical imaging.

Introduction

What is meaning? From a mathematical perspective, this question has been eloquently answered by a trilogy of papers on meaningful information (Gacs *et al.* 2001, Vereshchagin and Vitanyi 2004, Vitanyi 2006). This mathematical formulation of meaning will be described briefly below. From my perspective, as a computer engineer whose passion is analyzing biological images of live cell and tissue obtained by optical microscopy, the answer is more direct. Simply put, there are three steps to extracting meaningful information from imaging data. First, the data must be summarized concisely (Cohen *et al.* 2009). Next, this summary must be validated (Winter *et al.* 2011, Wait *et al.* 2014). Finally, the data needs to be shared. In theory, these steps apply to any type of data (Li and Vitanyi 1997). In practice they have been applied to a wide range of applications using time lapse phase and/or fluorescence microscopy. This perspective will describe these steps in detail.

In order to accomplish summarization, validation and sharing for biological imaging data, computational tools are required. *Visualizing the data together with summarization results is key*. Quite often this visualization is both the means and the end for making imaging data meaningful. Visualizing image data can be difficult, particularly as the *dimensionality* of the data grows. In imaging, dimensions are first spatial. A pixel, or picture element, at location (x, y) for 2-D images. A voxel, or volume element, at location (x, y, z) for 3-D images. In time-lapse imaging a voxel location is specified as a spatiotemporal (4-D) point, (x, y, z, t). In fluorescence microscopy we add a spectral channel λ to represent different imaging channels and the data is 5-D, (x, y, z, t, λ). As dimensionality continues to grow, visualization becomes even more important for extracting meaning and value from our imaging data. Visualizing complex image data together with summarization results is a challenge that requires sophisticated hardware and software solutions (Peng *et al.* 2014, Wait *et al.* 2014).

Another challenge is the size of the data. Current-generation time-lapse microscopes include integrated incubation and can typically acquire 100 movies or time-lapse image sequences in a single experiment. Each movie can consist of thousands of images. In our ongoing work analyzing stem cell image sequence data, a single dataset of 200 movies requires 350 gigabytes (GB) of image data or more. This is obviously too much data to analyze by hand or by eye – we must turn to computational analysis. There are many existing software packages for working with smaller and less complex image data sets (Eliceiri *et al.* 2012), but here I will focus on software solutions custom-written for the specific characteristics of the image data in order to best summarize the data in the context of a particular biological question. One of the key challenges in biological image analysis is the lack of computational tools for interactively and collaboratively summarizing, visualizing and validating image data.

Figure 1 shows an overview of the summarization, validation and sharing steps.

The notions of concise and meaningful as used here are not qualitative measures. *Algorithmic Information Theory* (AIT) is a theoretical framework for image understanding that provides mathematical and computational techniques to quantify how concise a representation is possible and how well a model captures the meaningful information from a given digital object. The foundation of AIT is Kolmogorov Complexity (Li and Vitanyi 1997). The Kolmogorov Complexity of a digital object, a movie or experiment or dataset, gives a measure of the most concise possible description of the object. Think of it as the file size in bytes that the perfect file compression algorithm would achieve on the given data. AIT also characterizes the relationship between data and models. Randomness deficiency measures how much meaningful information a model has extracted from the data (Cohen *et al.* 2009). AIT gives the capability to quantify how well our summary represents the image data.

A concise and meaningful summary of the image data

The first step is to summarize the data, to find a more concise representation compared to the hundreds of gigabytes of images. The tasks associated with summarization include denoising, segmentation, tracking, lineaging and modeling. The best approaches to summarize image data will exploit information in both directions among these tasks. Not every application will require all of these tasks. For example, image compression algorithms have been used to quantify the developmental potential of stem cells from single image frames (Zhang *et al.* 2012).

Denoising algorithms need to be matched to the specific characteristics of the images. One of the simplest and most effective denoising approaches is the median filter. This robust estimator is particularly good at removing the ubiquitous “salt and pepper” noise. More complicated approaches model the imaging noise and the background structure separately, and use a combination of filters (Michel *et al.* 2007). As the first task in the summarization step, denoising is highly specific to the imaging conditions. In a recent paper on visualization and analysis of 5-D images (Wait *et al.* 2014), different denoising algorithms were used for different image channels, each with carefully chosen parameters to match the imaging and noise characteristics. Following denoising, the images are segmented to identify the individual objects.

Segmentation, or delineation of individual objects, is a two-step task. First, a *threshold* divides the image into foreground and background regions. The foreground contains the important objects. Thresholding picks an intensity value to separate the two regions. Picking this value automatically, called adaptive thresholding, is one of the very few easy tasks in image analysis (Otsu 1979). If the objects in your image are not touching each other, then congratulations! You are done with segmentation, your results should be near perfect. More likely is that your objects will be in contact, at least occasionally, and more sophisticated segmentation is needed.

Following thresholding, separating touching objects is the second segmentation step. Separating touching objects is far and away the hardest task you face. If you are using 2-D imaging to look at 3-D objects, they can overlap partially or completely. This overlap is called occlusion. Occlusion can make it impossible for even a human domain expert (that’s you) to manually segment the objects. If you have time sequence data, incorporating temporal context to improve the low-level image processing tasks has been widely used with good success (Cohen *et al.* 2010, Winter *et al.* 2011). As a rule of thumb, if you can see the correct segmentation by eye, an algorithm will often, although not always, be able to extract the correct answer. Likewise, if you are unable to determine the correct segmentation by eye, the algorithm will rarely, although not never, extract the correct answer. After segmentation, if you have time sequence data, tracking is next.

Tracking establishes temporal correspondences between segmentation results. Simpler tracking algorithms establish these correspondences between pairs of image frames (Clark *et al.* 2011). More sophisticated algorithms solve the correspondence over multiple image frames simultaneously, often achieving significantly better accuracy. For biological applications, our Multitemporal Association Tracking is a multiframe tracking

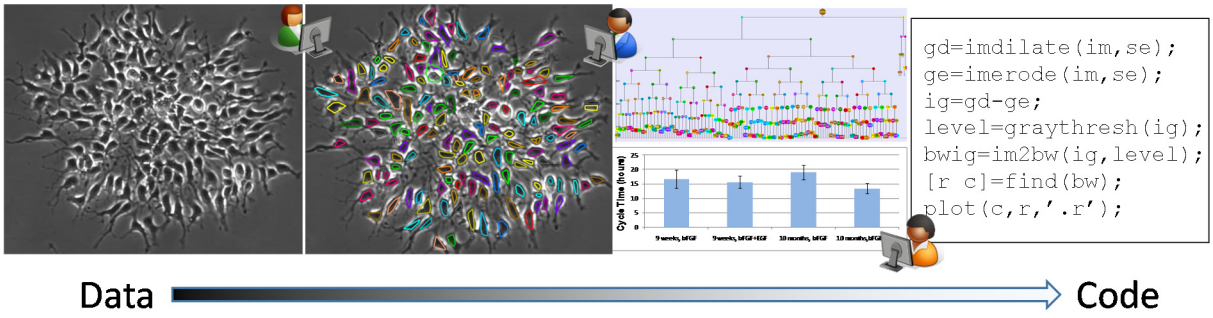

Figure 1. Interactive and collaborative use of the image data, together with the summarization results and the visualization and analysis code. From left to right, single image from a 2000 frame sequence, same image with segmentation and tracking results overlaid, lineage tree with time as the vertical axis (top) and statistical model (bottom), and open source code fragment. The arrow on the bottom shows the progression from image data to source code, with summarization results falling in the gray area.

solution that has proven widely effective for tracking organelles and stem cells (Winter *et al.* 2011, Winter *et al.* 2012, Chenouard *et al.* 2014, Mankowski 2014). If you capture images frequently enough, so that *e.g.* objects overlap by 50% or more between frames, and the segmentation is reliable, tracking should be straightforward. For image sequences with proliferating cells, one additional task is to establish the parent-daughter relationship or *lineage* (Al-Kofahi *et al.* 2006, Winter *et al.* 2011, An-An *et al.* 2012). The results of the segmentation, tracking and lineaging algorithms, taken together, are referred to as the *dynamic phenotype* (Cohen *et al.* 2009). The dynamic phenotype is a far more concise representation compared to the image data, but it is not the ultimate goal. For the example described above, starting with 350GB of image data, the dynamic phenotype still contains a considerable 2GB of data. In order to extract meaning from the image data, we must fit a model to this data.

The final step of summarization is to model the dynamic phenotype. A model, in both the mathematical (AIT) and biological sense is a concise representation of our data. The simplest models are statistical. In unsupervised approaches to AIT, models are based on clustering, or partitioning the data based on *e.g.* meaningful differences in behavior. More complex models span the fields of physics, pattern recognition, machine learning, *etc.* and can typically include domain or application specific knowledge. For example, generative models learn simulation parameters from the image data and are scored by how well they recreate object behaviors (Peng and Murphy 2011). The current state of the art in AIT gives a theoretical basis for analyzing distinct classes of models, including finite sets, recursive functions, and probability distributions (Vitanyi 2006), and a practical set of tools for unsupervised (Cohen *et al.* 2009), or semi-supervised (Cohen *et al.* 2010) analyses based on AIT principles. Importantly, these practical applications of AIT for summarization and modeling have consistently found that the algorithmically meaningful characteristics of the image data were also biologically meaningful. Integrating new types of models into the AIT framework will be another very active research area moving forward. Although AIT provides rigorous tools to characterize the relationships between data and models, ultimately the judgment of the biologists and engineers most familiar with the application must be brought to bear.

Validating the Summary

Validation is the next step after summarization. There is no completely computational approach to extracting meaningful information from image data. Summarization algorithms for complex data will always require human assistance, at the very least to provide domain knowledge on the imaging and application characteristics. There is also often the need to correct any errors in some parts of the automatically generated summarization. This is the validation step.

AIT is robust to segmentation and denoising errors but for some applications any tracking errors can render the summary invalid (Cohen *et al.* 2009). Tools like LEVER (Winter *et al.* 2011) have been developed to allow users to correct any errors in the automated segmentation, tracking and lineaging. The guiding principle behind such

approaches is to minimize the amount of human effort required to correct any errors. In LEVER, this is accomplished by learning from user provided corrections to automatically correct related mistakes. The validation incorporates the ability to correct errors, automatically utilizing the information provided by the human observer to update the summary. One significant challenge is how to handle the visual ambiguity inherent in biological images. There are two ways to handle the situation where human observers are unable to determine, or to agree on, a ground truth. Either the data must be discarded, or it must be marked as ambiguous in the summarization so that subsequent analysis can determine how best to handle the ambiguity. The question of how to best integrate human expertise into the data summarization process and to manage ambiguity in the summarization results is another very active research area.

Sharing the Results

Like everything else in science, the real value of our images and summarization methods and results comes when others can use them. Many software tools, including most of the ones mentioned in this perspective are provided open source. Generally "open source" means that you are allowed to download, use and modify the code as you like. Redistribution is generally allowed, but with varying restrictions. One limitation is that if image data is not available, there is no practical way to visualize the results. Without the ability to visualize summarization results together with the image data, it becomes impossible to reuse the results with any confidence. It is also difficult to evaluate the quality of the code. Data and code need to go together.

New hardware and software infrastructure designed for viewing and interacting with complex data over the internet continue to improve. Recently, HTML5 and WebGL standards have been developed. These standards provide a widely available framework for high-performance interactive and distributed applications, exactly what we need to make our imaging data and summarization results ubiquitously available. Even for large imaging experiments an "open data" approach is feasible. For the stem cell image data example described above, the 350GB of image data can be lossy compressed (*e.g.* JPEG) down to less than 10GB. This is a reasonable amount of data to download. Although lossy compression should never be used in segmentation or denoising, it is perfectly acceptable for visualizing and validating the summary. There is no technical hurdle to providing open source code together with all of the image data and the summarization results.

Conclusion

The best way to show the importance of the biology, the beauty of the imaging and the truth of the summarization is to show all of the data together with the summarization results. The size and complexity of our imaging data will continue to grow, incorporating new imaging modalities and additional data types. Collaboration between teams of biologists and engineers will be needed to design experiments and to guide the analyses of this rich data. In order to realize the full potential of our imaging data to provide insight into fundamental questions in biology and medicine, we must enable all of the data to be visualized together with the validated summarization. Extracting meaningful information from biological imaging data will require leveraging the best capabilities of interdisciplinary teams of human domain experts working together with sophisticated computational hardware and software. The success of our methods will be measured by how easy it is to reuse our techniques and build on our results. The future of biological imaging has never looked brighter!

Acknowledgements

Image data from Figure 1 courtesy Sally Temple, Susan Goderie, Mo Liu and Maria Apostolopoulou from the Neural Stem Cell Institute, Rensselaer, NY. Thanks to Walt Mankowski, Mark Winter, Eric Wait and Sally Temple for feedback and suggestions on the manuscript. Portions of the research described were supported by Drexel University, by the National Institute of Neurological Disorders and Stroke (R01NS076709), by the National Institute of Aging (R01AG040080) and by a Human Frontier Science Program grant (RGP0060/2012).

References

Al-Kofahi, O., R. J. Radke, S. K. Goderie, Q. Shen, S. Temple and B. Roysam (2006). "Automated cell lineage tracing: a high-throughput method to analyze cell proliferative behavior developed using mouse neural stem cells." Cell Cycle **5**(3): 327 - 335.

An-An, L., L. Kang and T. Kanade (2012). "A Semi-Markov Model for Mitosis Segmentation in Time-Lapse Phase Contrast Microscopy Image Sequences of Stem Cell Populations." Medical Imaging, IEEE Transactions on **31**(2): 359-369.

Chenouard, N., I. Smal, F. de Chaumont, M. Maska, I. F. Sbalzarini, Y. Gong, J. Cardinale, C. Carthel, S. Coraluppi, M. Winter, A. R. Cohen, W. J. Godinez, K. Rohr, Y. Kalaidzidis, L. Liang, J. Duncan, H. Shen, Y. Xu, K. E. Magnusson, J. Jalden, H. M. Blau, P. Paul-Gilloteaux, P. Roudot, C. Kervrann, F. Waharte, J. Y. Tinevez, S. L. Shorte, J. Willemse, K. Celler, G. P. van Wezel, H. W. Dan, Y. S. Tsai, C. O. de Solorzano, J. C. Olivo-Marin and E. Meijering (2014). "Objective comparison of particle tracking methods." Nat Methods **11**(3): 281-289.

Clark, B., M. Winter, A. R. Cohen and B. Link (2011). "Generation of Rab-based transgenic lines for in vivo studies of endosome biology in zebrafish." Developmental Dynamics **240**(11): 2452-2465.

Cohen, A. R., C. Bjornsson, S. Temple, G. Banker and B. Roysam (2009). "Automatic Summarization of Changes in Biological Image Sequences using Algorithmic Information Theory." IEEE Trans Pattern Anal Mach Intell **31**(8): 1386-1403.

Cohen, A. R., F. Gomes, B. Roysam and M. Cayouette (2010). "Computational prediction of neural progenitor cell fates." Nat Methods **7**(3): 213 - 218.

Eliceiri, K. W., M. R. Berthold, I. G. Goldberg, L. Ibanez, B. S. Manjunath, M. E. Martone, R. F. Murphy, H. Peng, A. L. Plant, B. Roysam, N. Stuurman, J. R. Swedlow, P. Tomancak and A. E. Carpenter (2012). "Biological imaging software tools." Nat Methods **9**(7): 697-710.

Gacs, P., J. Tromp and P. Vitanyi (2001). "Algorithmic statistics." Information Theory, IEEE Transactions on **47**(6): 2443-2463.

Li, M. and P. M. B. Vitanyi (1997). An Introduction to Kolmogorov Complexity and Its Applications. New York, Springer Verlag.

Mankowski, W., M. Winter, E. Wait, S. Naik, M. Lodder, T. Shumacher and A. R. Cohen (2014). Segmentation of Occluded Hematopoietic Stem Cells from Tracking. 36th annual IEEE Conference on Engineering in Medicine and Biology. Chicago IL.

Michel, R., R. Steinmeyer, M. Falk and G. S. Harms (2007). "A new detection algorithm for image analysis of single, fluorescence-labeled proteins in living cells." Microsc Res Tech **70**(9): 763-770.

Otsu, N. (1979). "A Threshold Selection Method from Gray-Level Histograms." IEEE Transactions on Systems, Man, and Cybernetics **9**(1): 62-66.

Peng, H., A. Bria, Z. Zhou, G. Iannello and F. Long (2014). "Extensible visualization and analysis for multidimensional images using Vaa3D." Nat. Protocols **9**(1): 193-208.

Peng, T. and R. F. Murphy (2011). "Image-derived, three-dimensional generative models of cellular organization." Cytometry A **79**(5): 383-391.

Vereshchagin, N. K. and P. M. B. Vitanyi (2004). "Kolmogorov's structure functions and model selection." Information Theory, IEEE Transactions on **50**(12): 3265-3290.

Vitanyi, P. (2006). "Meaningful information." Information Theory, IEEE Transactions on **52**(10): 4617 - 4626.

Wait, E., M. Winter, C. Bjornsson, E. Kokovay, Y. Wang, S. Goderie, S. Temple and A. R. Cohen (2014). "Visualization and Correction of Automated Segmentation, Tracking and Lineaging from 5-D Stem Cell Image Sequences." BMC Bioinformatics **15**(1): 328.

Winter, M., E. Wait, B. Roysam, S. Goderie, E. Kokovay, S. Temple and A. R. Cohen (2011). "Vertebrate Neural Stem Cell Segmentation, Tracking and Lineaging with Validation and Editing." Nature Protocols **6**(12): 1942-1952.

Winter, M. R., C. Fang, G. Banker, B. Roysam and A. R. Cohen (2012). "Axonal transport analysis using Multitemporal Association Tracking." Int J Comput Biol Drug Des **5**(1): 35-48.

Zhang, X., H. Wang, T. Collins, Z. Luo and M. Li (2012). Classifying Stem Cell Differentiation Images by Information Distance. Machine Learning and Knowledge Discovery in Databases. P. Flach, T. De Bie and N. Cristianini, Springer Berlin Heidelberg. **7523:** 269-282.