Research Issues in Protein Location Image Databases

Robert F. Murphy
Carnegie Mellon University
murphy@cmu.edu

Christos Faloutsos
Carnegie Mellon University
christos@cs.cmu.edu

1. DESCRIPTION

Which proteins have similar locations within cells? How many distinct location patters do cells display? How do we answer these questions quickly, from a large collection of microscope images such as in on-line journals?

This tutorial focuses exactly on questions like the above, with the goal to highlight the database and data mining research questions in "location proteomics." The field of proteomics tries to understand the organization of cells by studying all proteins in parallel, using high-throughput methods. Of particular importance is determining which proteins associate with each other and the locations within the cell where these interactions take place.

The Human Genome Project and other genome projects have led to the population of large sequence databases as well as methods such as BLAST for retrieving sequences from them. The focus of most biomedical research is now shifting from simply identifying gene sequences to determining the properties and functions of the proteins encoded by those genes. Since mammalian cells have a number of distinct sub-compartments (such as the nucleus, mitochondria, and lysosomes) that play different roles, knowledge of the subcellular locations of all proteins is critical to understanding their functions and the mechanisms by which those functions are accomplished. Using fluorescence microscopy, we can illuminate and track the molecules of a given protein and take a photograph of them inside a human cell; this way we can figure out the places in the cell where a protein is found.

Current approaches to using fluorescence microscopy rely on visual interpretation of images and description of proteins with a limited vocabulary. Recent work provides an automated and systematic alternative that involves creation of databases of fluorescence microscope images for many different proteins and tools for automated analysis and retrieval of images from those databases.

One fundamental problem is how to compare two such images. We present such solutions, using low-level features from the protein images. We show how these features can be used to

- Train classifiers that can recognize the patterns of all major organelles with high accuracy in previously unseen images,
- Rank images in a set by how typical they are of that set (e.g., to find the most representative or typical image),
- Provide sensitive, statistically sound comparisons of sets of images collected for different proteins or for the same protein under different conditions (e.g., with and without a drug) to determine whether the patterns are the same,
- Group proteins by their location patterns, forming a tree hierarchy,
- Index image databases to permit content-based retrieval Retrieve and analyze microscope images from on-line journal articles.

The combination of these automated methods with methods for large-scale fluorescent tagging of proteins (such as CD-tagging) and automated fluorescence microscopes enables the new field of Location Proteomics, providing for the first time sensitive, objective and systematic determination of the subcellular location for all proteins in a proteome. This approach also permits determination of the ways in which these locations change in response to a variety of phenomena, including growth and development, onset of disease, or exposure to drugs or pathogens.

2. OUTLINE

The material consists of the following topics.

- Problem definition and introduction to relevant cell and molecular biology (Murphy)
- Results on supervised and un-supervised learning on microscope images (Murphy)
- Current approaches to creating databases of fluorescence microscope images (Murphy)
- Automated retrieval of images from on-line journal articles (Murphy)
- Automated and interactive feature selection methods (Faloutsos)

SIGMOD 2005 June 14-16, 2005, Baltimore, Maryland, USA.

- Fast and flexible feature-based image retrieval methods (Faloutsos)
- Clustering and Singular Value Decomposition (Faloutsos)
- Future directions (Faloutsos)

3. AUDIENCE

Computer scientists interested in learning about the challenges and opportunities in this new area of computational biology research, as well as current methods for interpretation and retrieval of complex images that differ dramatically from natural scenes.

4. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. EF-0331657Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

5. ABOUT THE INSTRUCTORS

Robert F. Murphy earned an A. B. in Biochemistry from Columbia College in 1974 and a Ph.D. in Biochemistry from the California Institute of Technology in 1980. He was a Damon Runyon-Walter Winchell Cancer Foundation postdoctoral fellow at Columbia University from 1979 through 1983, after which he became an Assistant Professor of Biological Sciences at Carnegie Mellon University. He received a Presidential Young Investigator Award from the National Science Foundation shortly after joining the faculty at Carnegie Mellon in 1983 and has received research grants from the National Institutes of Health, the National Science Foundation, the American Cancer Society, the American Heart Association, the Arthritis Foundation, and the Rockefeller Brothers Fund. He has co-edited two books and published over 90 research papers. His research group at Carnegie Mellon focuses primarily on the application of fluorescence methods to problems in cell biology, with particular emphasis on automated interpretation of fluorescence microscope images. He has a long-standing interest in computer applications in biology, and developed the first formal undergraduate degree program in computational biology in 1987. He also founded and directs the Merck Computational Biology and Chemistry Program at Carnegie Mellon. In 1984, he co-developed the Flow Cytometry Standard data file format used throughout the cytometry industry and he is Chair of the Cytometry Development Workshop held each year in Asilomar, California. He is currently Professor of Biological Sciences and Biomedical Engineering and Voting Member in the Center for Automated Learning and Discovery in the School of Computer Science at Carnegie Mellon.

Christos Faloutsos is a Professor of Computer Science at Carnegie Mellon University. He has received the Presidential Young Investigator Award from the National Science Foundation (1989), four “best paper” awards, and four teaching awards. He is a member of the executive committee of SIGKDD; he has published over 120 refereed articles, one monograph, and holds four patents. His research interests include data mining for streams and networks, fractals, bioinformatics, spatial and multimedia bases, and database performance.