PSLID, the Protein Subcellular Location Image Database: Subcellular location assignments, annotated image collections, image analysis tools, and generative models of protein distributions

Estelle Glory[123], Justin Newberg[123], Tao Peng[12], Ivan Cao-Berg[14], Robert F. Murphy[12345]

[1]Center for Bioimage Informatics, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[2]Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[3]Molecular Biosensor and Imaging Center, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[4]Department of Biological Sciences, Department of Machine Learning, Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
[5]External Fellow, Freiburg Institute for Advanced Studies, Freiburg, Germany

email: Robert F. Murphy (murphy@cmu.edu)
URL for the overall project web site: http://pslid.cbi.cmu.edu
URL for accessing the code: http://murphylab.web.cmu.edu/software
The particular Open Source License being used: GPL

Subcellular location is a fundamental characteristic of proteins, and knowledge of subcellular location on a proteome-wide basis will be critical to building accurate models of complex cellular behaviors. Subcellular patterns are often quite complex and difficult, if not impossible, to capture using text-based approaches. Since in many cases the primary data for determining subcellular location is in the form of images of fluorescently-tagged proteins, widespread availability of annotated protein images is essential. The Protein Subcellular Location Image Database (http://pslid.cbi.cmu.edu), which is currently at Release 4, is the earliest publicly available image database of subcellular patterns, with the first release occurring in 2001. It contains image collections from a number of research groups, in particular, large collections of 2D and 3D fluorescence microscope images that have been widely used to train and test systems for automatically analyzing subcellular patterns. Over the past decade, these systems have been demonstrated to be more sensitive at distinguishing subcellular patterns than visual analysis. PSLID provides access to either entire collections or to user-specified subsets of the collection via either an interactive search interface or a programmatic search interface. In addition to providing images, PSLID provides public access to a large suite of tools for automated analysis of those images. These include preparatory functions (such as segmentation into individual cells and extraction of numerical features to capture subcellular distributions), and core analysis functions (including the statistical comparison of location between image sets, training and using of machine classifiers, and clustering of images or proteins by their location patterns). Most critically, PSLID includes state-of-the-art methods for inference of patterns across multi-cell images using graphical models and tools for building and distributing generative models of subcellular patterns (models that allow synthesis of new images drawn from the same underlying distribution as a set of training images). The database and analysis software are not only accessible via web service but can be downloaded for local use. A number of the tools are also available separately so that they can be used without requiring installation of the PSLID database engine. These are available both as Matlab toolboxes and as standalone applications.