Chapter 8
Data Mining Methods for a Systematics of Protein Subcellular Location

Kai Huang and Robert F. Murphy

Summary

Proteomics, the comprehensive and systematic study of the properties of all expressed proteins, has become a major research area in computational biology and bioinformatics. Among these properties, knowledge of the specific subcellular structures in which a protein is located is perhaps the most critical to a complete understanding of the protein's roles and functions. Subcellular location is most commonly determined via fluorescence microscopy, an optical method relying on target-specific fluorescent probes. The images that result are routinely analyzed by visual inspection. However, visual inspection may lead to ambiguous, inconsistent, and even inaccurate conclusions about subcellular location. We describe in this chapter an automatic and accurate system that can distinguish all major protein subcellular location patterns. This system employs numerous informative features extracted from the fluorescence microscope images. By selecting the most discriminative features from the entire feature set and recruiting various state-of-the-art classifiers, the system is able to outperform human experts in distinguishing protein patterns. The discriminative features can also be used for routine statistical analyses, such as selecting the most typical image from an image set and objectively comparing two image sets. The system can also be applied to cluster images from randomly tagged genes into statistically indistinguishable groups. These approaches coupled with high-throughput imaging instruments represent a promising approach for the new discipline of location proteomics.

8.1 Introduction

8.1.1 Protein Subcellular Location

The life sciences have entered the post-genome era where the focus of biological research has shifted from genome sequences to protein functionality. With whole-genome drafts of mouse and human in hand, scientists are putting more and more effort into obtaining information about the entire proteome in a given cell type. The properties of a protein include its amino acid sequences, its expression levels under various developmental stages and in different tissues, its 3D structure and active sites, its functional and structural binding partners, and its subcellular location. Protein subcellular location is important for understanding protein function inside the cell. For example, the observation that the product of a gene is localized in mitochondria will support the hypothesis that this protein or gene is involved in energy metabolism. Proteins localized in the cytoskeleton are probably involved in intracellular trafficking and support. The context of protein functionality is well represented by protein subcellular location.

Proteins have various subcellular location patterns [250]. One major category of proteins is synthesized on free ribosomes in the cytoplasm. Soluble proteins remain in the cytoplasm after their synthesis and function as small factories catalyzing cellular metabolites. Other proteins that have a target signal in their sequences are directed to their target organelle (such as mitochondria) via posttranslational transport through the organelle membrane. Nuclear proteins are transferred through pores on the nuclear envelope to the nucleus and mostly function as regulators. The second major category of proteins is synthesized on endoplasmic reticulum(ER)-associated ribosomes and passes through the reticuloendothelial system, consisting of the ER and the Golgi apparatus. Some stay in either the ER or the Golgi apparatus, and the others are further directed by targeting sequences to other organelles such as endosomes or lysosomes. Protein subcellular location patterns often result from steady states or limit cycles and can also change under specific conditions. Proteins are continuously being synthesized, localized, and finally degraded. This process forms the distribution of a protein inside the cell. In addition, intracellular signal transduction pathways often involve translocation of either specific signal or cargo proteins among compartments and intercellular signal transduction employs endocytosis and exocytosis of certain signal proteins. The static and dynamic properties of protein subcellular location patterns provide a significant challenge for machine learning and data mining tools.

8.1.2 Experimental Methods to Determine Protein Subcellular Location

Several experimental methods have been developed to determine protein subcellular location, such as electron microscopy, subcellular fractionation,