Protein Subcellular Localization Prediction Using Artificial Intelligence Technology

Rajesh Nair and Burkhard Rost

Summary

Proteins perform many important tasks in living organisms, such as catalysis of biochemical reactions, transport of nutrients, and recognition and transmission of signals. The plethora of aspects of the role of any particular protein is referred to as its “function.” One aspect of protein function that has been the target of intensive research by computational biologists is its subcellular localization. Proteins must be localized in the same subcellular compartment to cooperate toward a common physiological function. Aberrant subcellular localization of proteins can result in several diseases, including kidney stones, cancer, and Alzheimer’s disease. To date, sequence homology remains the most widely used method for inferring the function of a protein. However, the application of advanced artificial intelligence (AI)-based techniques in recent years has resulted in significant improvements in our ability to predict the subcellular localization of a protein. The prediction accuracy has risen steadily over the years, in large part due to the application of AI-based methods such as hidden Markov models (HMMs), neural networks (NNs), and support vector machines (SVMs), although the availability of larger experimental datasets has also played a role. Automatic methods that mine textual information from the biological literature and molecular biology databases have considerably sped up the process of annotation for proteins for which some information regarding function is available in the literature. State-of-the-art methods based on NNs and HMMs can predict the presence of N-terminal-sorting signals extremely accurately. Ab initio methods that predict subcellular localization for any protein sequence using only the native amino acid sequence and features predicted from the native sequence have shown the most remarkable improvements. The prediction accuracy of these methods has increased by over 30% in the past decade. The accuracy of these methods is now on par with high-throughput methods for predicting localization, and they are beginning to play an important role in directing experimental research. In this chapter, we review some of the most important methods for the prediction of subcellular localization.
Key Words: Protein subcellular localization prediction; sorting signals; neural networks; support vector machines; hidden Markov models; amino acid composition; text analysis.

1. Introduction

1.1. Proteins Are the Machinery of Life

Proteins are the workhorses that are responsible for transforming the genetic information for life, stored in the nucleic acids (DNA), into physical reality. A protein molecule consists of a long unbranched chain of 20 amino acid residues; each amino acid is linked to its neighbor through a covalent peptide bond. The most distinguishing characteristic of proteins is that they have well-defined three-dimensional (3D) structures. A stretched-out polypeptide chain has no biological activity. Protein function arises from the “conformation” of the protein, which is the 3D arrangement, or shape, of the molecules in the protein. Proteins are the most structurally complex and functionally sophisticated macromolecules known and they perform a wide array of tasks in organisms, such as the catalysis of biochemical reactions, transport of nutrients, and recognition and transmission of signals. The coordinated effort of many different types of proteins is required to realize the genetic program that is encoded in DNA. All the multiple aspects of the role of any particular protein is referred to as its “function.”

1.2. Predicting Protein Function: A Major Challenge for Modern Biology

To date, the genome (DNA) sequences of over 400 organisms, including the draft sequence of the human genome (1), has been completed. The number of entirely sequenced genomes has been growing exponentially for many years, and this growth is expected to continue for at least the next several years. With the availability of genome sequences of entire organisms, we are, for the first time, in a position to understand the expression, function, and regulation of the entire set of proteins encoded by an organism. This information will be invaluable for understanding how complex biological processes occur at a molecular level, how they differ in various cell types, and how they are altered in disease states. Identifying protein function is a large step toward understanding diseases and identifying novel drug targets (2). However, experimentally determining protein function continues to be a laborious and painfully slow task requiring enormous resources. For example, more than a decade after its discovery, we still do not know the precise and entire functional role of the prion protein (3). To compound this problem, the rate at which expert annotators add experimental information...