A Novel Ensemble Approach for the Subcellular Localization of Proteins.

A. Rozza , G. Lombardi, M. Re, E. Casiraghi, G. Valentini, P. Campadelli
Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano,, Italy

Motivation: The subcellular localization of a protein is highly correlated with its function. Despite most proteins are synthesized in the cytoplasm they need to be transported to their final location in order to fulfill their biological role. This is the reason why the knowledge of the subcellular localization of a protein is pivotal in understanding its function. The knowledge of the localization of the proteins in cell has many applications in the research fields of proteomics, drug target discovery and systems biology. Despite the commonly accepted usefulness of information about the subcellular localization of proteins, the biomolecular processes governing the sorting of proteins in living cells is not yet completely understood and thus the use of computational biology approaches has become commonplace in order to reduce the high costs required by large scale experiments aimed at characterize the subcellular localization of whole proteomes. The automated prediction of protein subcellular localization often starts with the comparison of the fasta sequence of the protein with the content of many public databases such as PFam, SMART, INTERPRO (among the others). This result into feature vectors composed by hundreds, if not thousands, of elements each describing the presence/absence of a certain feature (i.e. a particular protein domain or the presence/absence of a known sorting signal) in the investigated protein. Several machine-learning methods have been proposed for the automated prediction of subcellular localization of proteins but most of them are not well suited to deal with the high dimensionality characterizing this prediction problem. It is known that the robustness of any predictor working on high dimensional data can be increased by means of techniques able to reduce the dimensionality of the input vectors while preserving the information that are relevant for the prediction problem at hand. Moreover, in protein subcellular localization prediction, the reduction of the input space should be considered desirable because this allows the usage of the entire amount of information about proteins contained in public databases.

Methods: Here we present a novel ensemble of classifiers that performs multiclass prediction of proteins subcellular localization by combining several kernel based classifiers. Each component classifier produces a score that is combined by means of a decision directed acyclic graph (DDAG). The reduction of the input space is performed by the component classifiers, called K-TIPCAC, and is based on the projection of given points on the Fisher subspace estimated on the training data by means of a novel technique.

Results: Our approach is both sensitive and accurate with respect to comparable methods and may be applicable in the prediction of the localization of whole proteomes. Moreover, its modularity allows the complete separation of the prediction step from the initial interrogation of the database(s) enabling investigators to select different combination of data sources.