Poster G-7

Learning Classifiers for Assigning Protein Sequences to Subcellular Localization Families

Authors:
Carson Andorf (*Iowa State University*)
Drena Dobbs (*Iowa State University*)
Vasant Honavar (*Iowa State University*)

Short Abstract: We explore machine learning approaches to build classifiers for protein subcellular localization prediction using a class conditional probabilistic representation of amino acid sequences. We combine these methods with a homology tool to develop HDTree. HDTree builds a classifier that outperforms current more computationally expensive methods used to predict subcellular localization.

Long Abstract:

Background

Assigning the subcellular localization to novel proteins is an important challenge in bioinformatics. We explore several machine learning approaches to data-driven construction of classifiers for assigning protein sequences to appropriate subcellular localization families using a class conditional probabilistic representation of amino acid sequences. Specifically, we represent protein sequences using class conditional probability distribution of amino acids (amino acid composition) or short (k-letter) subsequences (k-grams) of amino acids. We compare a model (NB k-grams), which ignores the statistical dependencies among overlapping k-grams with an alternative, NB(k), which uses an undirected probabilistic graphical model that captures the relevant dependencies. In addition, PSI-BLAST was used to query the subcellular localization data set to find the top matching hit. By combining the two complimentary methods we build a two-stage machine learning algorithm. We compare our methods with a support vector machine (SVM) based on the amino acid composition of a protein, SubLoc, and the state-of-the-art methods ESLpred and LOCSVMPSI which uses SVM classifiers trained on a hybrid of protein features.

Results

We report the performance of the resulting classifiers on two subcellular localization data sets. The first data set is Prokaryotic proteins with localization families of periplasmic, extracellular, and cytoplasmic. The second data set is based on Eukaryotic proteins with localization families of extracellular, cytoplasmic, mitochondria, and nuclear. Each of the proposed methods is effective in correctly assigning localization categories to protein sequences. When the optimal value of k is used, NB k-gram or NB(k) outperformed or matched SubLoc on 4 of the 7 localization classes, but was outperformed on all 7 classes by ESLpred and was outperformed on all 4 Eukaryotic classes by LOCSVMPSI. The performance of our methods can be improved by combining the outputs of our NB(k) and NB k-gram methods as input to a simple decision tree algorithm (DTree). Further improvements can be made by adding an additional input to the decision tree algorithm based on a homology search method such as PSI-BLAST (HDTree). HDTree was able to get accuracies of [94.8, 96.3, 97.0] on the three classes of Prokaryotic proteins with correlation coefficients

of [.84, .81, .93]. On the Eukaryotic data HDTree was able to get accuracies of [94.0, 99.5, 95.3, 96.3] on the four localization classes with correlation coefficients of [.85, .98, .80, .93]. This method outperformed both the SubLoc method and ESLpred method on 7 out of 7 (100%) of the classes and outperformed or matched LOCSVMPSI on 4 out of 4 (100%) of the Eukaryotic classes. In addition the proposed NB(k) and NB k-grams methods are based on undirected graphical models of overlapping k-grams that allow for incremental update of the learned protein function classifiers as additional training data become available. The decision tree method is based on 8 attributes and can be constructed with minimal computational effort. Using a SVM has an expense of increased computational demands if frequent updates to the classifier are required as new training data become available.

Conclusions

We have shown that amino acid k-gram compositions of protein sequences offer an inexpensive, yet highly effective source of information for predicting the subcellular localization annotations of proteins. Our experimental results demonstrate the feasibility of using machine learning approaches that require only the amino acid k-gram compositions to automatically and reliably generate subcellular localization annotations of protein sequences. These methods work very well when sequences within a class have similar k-gram compositions. Homology tools work well when sequences have relatively high sequence identity. HDTree combines the complimentary information found between these two approaches and builds a classifier that outperforms the current more computationally expensive methods used to predict subcellular localization.