Chapter 60
JFeature: A Java Package for Extracting Global Sequence Features from Proteins for Functional Classification

Xin Chen
Zhejiang University, China

Hangyang Xu
Zhejiang University, China

ABSTRACT

Prediction of various functional properties of proteins has long been a central theme of bioinformatics in the post-genomic era. Statistical learning, in addition to analysis based on sequence similarity, was proven successful to detect complex sequence-function associations in many applications. JFeature is an integrated Java tool to facilitate extraction of global sequence features and preparation of example sets, in statistical learning studies of sequence-function relationships. With a user-friendly graphical interface, it computes the composition, distribution, transition and auto-correlation features from sequence. It also helps to assemble a negative example set based on the most-dissimilar principle. The Java package and supplementary documentations are available at http://www.cls.zju.edu.cn/rlibs/software/jfeature.html.

INTRODUCTION

In the last decade, genome projects have been generating a huge amount of sequences at an exponential rate. By the year 2009, more than 500 species have been sequenced, whereas the characterization of newly discovered genes is lagging far behind. Computational approaches are usually sought to predict function or functional class of many new genes yet to be studied experimentally. Sequence similarity based approaches, in the forms of global alignment, local alignment or hidden Markov model, are extensively used to assign putative functions for new genes. Recent studies show that statistical learning approaches could be an effective alternative and/or supplement,

DOI: 10.4018/978-1-4666-3604-0.ch060

especially in the case of predicting function of distantly-related proteins or homologous proteins with different functions (Cai et al., 2003). Their successful applications have also been reported in predicting protein functional classes (Karchin et al., 2002), protein-protein interactions (Bock & Gough, 2001), sub-cellular localizations (Zhang, 2006) and many other functional aspects of proteins from sequences.

Statistical learning approaches construct statistical models, such as neural network models or support vector machine (SVM) models, trained from known examples of the target sequence-function relationship. The models are then used to predict the functions of unknown sequences. The learning algorithms, in most cases, only deal with encoded representations of the examples sequences (known as the feature vectors), instead of the raw sequences themselves. Moreover, they usually require both positive and negative examples (sequences with and without a certain function) for training.

The feature vectors representing example sequences are usually required to have a fixed length for most learning algorithms. Therefore, researchers often compute global sequence features to build a feature vector with the same number of components. In this case, each component (a global feature) describes sequence characteristics on a whole protein level. A global feature is typically a real-valued function of some physicochemical property and/or sequential order of the residues in the sequence. Four types of functions have been demonstrated to be useful in extracting informative features for functional classifications. They are the composition, distribution, transition and auto-correlation functions. The four different function types, in combination with the over 500 different amino acid properties (Kawashima & Kanehisa, 2000), result in thousands of possible global features. A set of the most discriminative features is often chosen empirically to form the feature vector. Moreover, many rounds of optimizations are usually necessary.

In many applications, we want to predict proteins with certain functions, for instance, the prediction of potential allergen proteins. This requires both allergen (positive) and non-allergen (negative) protein examples to train. While positive examples are usually reported in literatures and can be collected, there is hardly any primary data source where negative examples can be found. Without much knowledge on the potential distributions of negative examples in a particular objective, many researchers resort to the intuitively solution of constructing a negative example dataset that is most unlike the positive one. Such an approach has produced satisfactory results in many applications, including the prediction of drug-target likeness (Xu et al., 2007), protein-protein interactions (Ben-Hur and Noble, 2005; Gomez, et al., 2003; Zhang, et al., 2004), and proteins functional classes (Lin, et al., 2006). In addition, similar to positive examples, negative examples are better to be sampled from all potential negative examples uniformly without any bias. Empirical evidence suggests that positive examples are usually clustered in association with their protein family classifications. Assuming the same for negative examples, protein family based sampling of negative examples would help to get an unbiased sampling of negative examples. This is the idea implemented for negative example preparations in several recent studies (Cai, et al., 2003; Lin, et al., 2006; Xue, et al., 2004). In the above example, satisfactory results classifying allergen and non-allergen proteins were achieved by assembling negative examples from randomly picked PFam families that do not show sequence similarities to any of the allergic effect-causing proteins (Cui, 2007).

Till 2009, many software tools and packages are available to facilitate statistical learning analysis. However, there is still no convenient software package assisting the extraction of sequence features and preparation of negative examples. Therefore, the *JFeature* toolbox was developed as a bridging