Chapter 52
Classification of Biological Sequences

Pratibha Rani
International Institute of Information Technology Hyderabad, India

Vikram Pudi
International Institute of Information Technology Hyderabad, India

ABSTRACT

The rapid progress of computational biology, biotechnology, and bioinformatics in the last two decades has led to the accumulation of tremendous amounts of biological data that demands in-depth analysis. Data mining methods have been applied successfully for analyzing this data. An important problem in biological data analysis is to classify a newly discovered sequence like a protein or DNA sequence based on their important features and functions, using the collection of available sequences. In this chapter, we study this problem and present two Bayesian classifiers RBNBC (Rani & Pudi, 2008a) and REBMEC (Rani & Pudi, 2008c). The algorithms used in these classifiers incorporate repeated occurrences of subsequences within each sequence (Rani, 2008). Specifically, Repeat Based Naive Bayes Classifier (RBNBC) uses a novel formulation of Naive Bayes, and the second classifier, Repeat Based Maximum Entropy Classifier (REBMEC) uses a novel framework based on the classical Generalized Iterative Scaling (GIS) algorithm.

INTRODUCTION

With the development of biology, biotechnology, bioinformatics and biomedical research, more and more biological data is getting collected and is available for analysis (Wang et al., 2005). Data mining methods have been applied successfully for analyzing this data and many sophisticated mining tools such as *GeneSpring*, *Spot Fire* and *VectorNTI* have also been developed (Wang et al., 2005). The trend of developing data mining based solutions for biological data analysis is rapidly evolving. Details can be found in (Wang et al., 2005, chapter 2).

A critical problem in biological data analysis is to classify biological sequences based on their important features and functions. This problem is important due to the exponential growth of newly

DOI: 10.4018/978-1-4666-2455-9.ch052

generated sequence data during recent years, which demands for automatic methods for sequence classification. The advantage of automatic sequence classifier is that, prediction of class of an unclassified sequence reduces the time and cost required for performing experiments on the new sequence in laboratory to find its functions and properties. Since the sequences belonging to the same class have similar characteristics, the predicted class will give idea about the function and properties of the new sequence. For example, (1) a protein's structure and functions depend on its amino acid sequence, so if we can predict the class of a new protein sequence on the basis of its amino acid sequence, then we can predict its structure and functions; (2) frequently, it is unknown for which proteins a new DNA sequence codes or if it codes for any protein at all. If we can predict the class of a new coding sequence on the basis of known coding sequences then there is a high probability to predict the proteins it will code for; and (3) prediction of the type of disease can be done by predicting the class of a sample sequence using a set of known sample sequences divided in different classes according to the type of diseases.

The known state-of-the-art solutions for classification problem are mainly based on Sequence Alignment (Altschul et al., 1990, 1997; Pearson & Lipman, 1988), Hidden Markov Model (HMM) (Krogh et al., 1994; Durbin et al., 1998; Eddy, 1998), Probabilistic Suffix Trees (PST) (Bejerano & Yona, 1999; Eskin et al., 2003) and Support Vector Machines (SVM) (Leslie et al., 2002; Ben-Hur & Brutlag, 2003a, 2003b; Weston et al., 2005). Recent approaches (Melvin et al., 2007; Marsolo & Parthasarathy, 2006a, 2006b) have been trying to improve SVM by incorporating domain knowledge, using complex features based on structures and combining it with other classifiers.

In this chapter we discuss two totally data mining based, simple but effective *Bayesian* classifiers for the biological sequences. These classifiers are called Repeat Based Naive Bayes Classifier (RBNBC) and Repeat Based Maximum Entropy Classifier (REBMEC). These classifiers use generic domain independent feature extraction method which requires comparatively less memory and time with the advantage of no need of domain expertise. Also these classifiers incorporate repeated occurrences of subsequences within each sequence known as *repeats* of the subsequences. Note that the existing domain based feature extraction methods are highly memory intensive and time consuming and they need extensive domain knowledge (Ferreira & Azevedo, 2005b, 2006; Lesh et al., 1999, 2000; Huang & Brutlag, 2001).

Naive Bayes is well known as a surprisingly successful classification method that has outperformed much more complicated methods in many application domains (Domingos & Pazzani, 1996; Kotsiantis & Pintelas, 2004; Zhang, 2004). However a direct implementation of *Naïve Bayes* does not work well for biological sequences. In RBNBC it is adapted to work for biological sequences.

On the other hand REBMEC uses a novel framework based on the classical *Generalized Iterative Scaling* (GIS) (Darroch & Ratcliff, 1972) algorithm to find the maximum entropy model for the given collection of biological sequences. The maximum entropy principle has been widely used for various tasks including discretization of numeric values of features (Kotsiantis & Pintelas, 2004), feature selection (Li et al., 2003; Tatti, 2007; Ratnaparkhi, 1998), and various text related tasks like translation (Berger et al., 1996), document classification (Nigam et al., 1999), and part-of-speech tagging (Ratnaparkhi, 1998). REBMEC's approach is inspired by these works because comparison between biological sequence data and natural languages are commonplace (Buehler & Ungar, 2001). Unlike other *Bayesian* classifiers like *Naive Bayes*, maximum entropy based classifiers do not assume independence among features. These classifiers build the model of the dataset using an iterative approach to find the parameter values that satisfy the constraints generated by the features and the training data (Thonangi & Pudi,