Pavan Kumar
MB200510

Optimization of HMM Parameters to Increase Efficiency in Predicting Protein Secondary Structure.

ABSTRACT

The present day probabilistic models like HMM are performing the vital task of structure prediction. These models are required to improve its working to increase the prediction accuracy. The third problem of HMM is to maximize the model parameters in order to achieve a better prediction.

The Hidden Markov Model (HMM) is a natural and highly robust statistical methodology for automatic speech recognition and secondary structure prediction. It is also being tested and proved considerably important in a wide range of applications. The model parameters of HMM are essential in describing the behavior of the utterance of most of the observation sequences. Also it is well established that HMM efficiently interprets the predictable properties in biological sequences because of having transition probability parameter.

Many successful heuristic algorithms are developed to optimize the model parameters in order to best describe the trained observation sequences. However, most of the methodologies explore for only the local maxima in practice. None of the methodology can recover the global maxima or other more optimized local maxima. The stochastic search method, Genetic Algorithm (GA) is presented for HMM parameter training without using iterative training method. GA applies the principle of natural evolution and performs global searching within the defined search space. Experiments showed that using GA for HMM training result in a better performance than using other heuristic algorithms for optimization purpose.

In our work the problem of optimization of parameters is considered to obtain this goal. It seems to have a better outcome as we are utilizing a global technique to deal with the prediction problem. It is quite interesting to maximize the model parameters utilizing a computational approach. The parameters of HMM (viz. π, A, B) i.e. initial, transition, emission probability are calculated from a taken database comprising of a pair of protein secondary structure and primary sequence. Here the global search is initialized with multiple inputs generated for each parameter variables. Fitness function has been used to operate selection procedure. Random multipoint crossover is chosen to allow an unbiased exchange of bits among the strings. Also random mutation with very little probability is allowed to alter the strings. Evaluating the fitness optimization is done. The prediction efficiency of the model has been found to enhance remarkably.