生物物理学报(2006) 22(1), 39-48

PDF文档

# 一种基于特征筛选的原核生物启动子判别分析方法

杜耀华
国防科技大学机电工程与自动化学院

启动子识别是研究基因转录调控的重要环节，但目前方法的识别正确率偏低。在深入分析原核启动子特征的基础上，提出了一种基于特征筛选的原核启动子判别分析方法，首先在启动子序列的组成特征、信号特征和结构特征中选取备选特征，为每个特征建立适当的描述模型，并对主要的保守模式采用复合模式模型；再通过模型计算对备选特征进行逐步筛选，优化特征集，将序列表示为组合特征向量；最终利用二次判别分析实现识别。对大肠杆菌和枯草杆菌实际启动子数据进行的刀切法测试验证了方法的有效性和通用性。对于大肠杆菌非编码区sigma70启动子，识别的平均正确率达到了85.8%，优于其它几种典型识别方法；对于大肠杆菌编码区内部sigma70启动子和其它几种原核启动子，平均正确率也都超过了80%。方法框架还具有良好的可扩展性，能够方便的容纳新特征，使识别性能不断提高。

# A Discriminant Analysis Method for Prokaryotic Promoter Regions Based on Feature Selection

Promoter identification is an essential task in the research of transcription regulation, but computational prediction of promoters has been one of the most elusive problems despite considerable effort devoted to the study. In this paper, a discriminant analysis method based on feature selection for prokaryotic promoter regions is proposed. In each region, the candidate features of primary sequence, including content features, signal features and structure features, are calculated by selecting proper models. Especially for the main conserved signal features, a composite motif model is adpoted. Through a stepwise selection process, the optimal features are determined from candidate feature sets and combined as a multidimensional vector, then the vector of combined features is further used by quadratic discriminant analysis to predict the potential promoter regions. The algorithm has been trained and tested on E. coli and B. subtilis promoter datasets by the jackknife method. For E. coli ‘non-coding’ sigma70 promoters, the average prediction accuracy is 85.8%, and for E. coli ‘coding’ sigma70 promoters and several other kinds of prokaryotic promoters, their prediction accuracies are also higher than 80%. The results indicate that our method is a universal algorithm that outperforms most of the existing approaches based on several performance measurements. Furthermore, the framework of the method is extendable, which can accept more new features to improve the prediction results efficiently.

## 关键词

原核生物(Prokaryote); 启动子识别(Promoter identification); 复合模式(Composite motif); 特征筛选(feature selection); 二次判别分析(Quadratic discriminant analysis); 刀切法(Jackknife method)