

基于贝叶斯网络的DNA序列剪接位点预测

李骞、王涛、冯焕清*、王明会
中国科学技术大学生物医学工程研究所

采用基于贝叶斯网络的建模方法, 预测真核生物DNA序列中的剪接位点。分别建立了供体位点和受体位点模型, 并根据两种位点的生物学特性, 对模型的拓扑结构和上下游节点的选择进行了优化。通过贝叶斯网络的最大似然学习算法求出模型参数后, 利用10分组交互验证方法对测试数据进行剪接位点预测。结果显示, 受体位点的平均预测准确率为92.5%, 伪受体位点的平均预测准确率为94.0%, 供体位点的平均预测准确率为92.3%, 伪供体位点的平均预测准确率为93.5%, 整体效果要好于基于使用独立和条件概率矩阵、以及隐Markov模型的预测方法。表明利用贝叶斯网络对剪接位点建模是预测剪接位点的一种有效手段。

PREDICTING SPLICE JUNCTION SITE IN DNA SEQUENCES WITH BAYESIAN NETWORK

Two new models for predicting the splice junction in eukaryotic DNA sequences were developed by exploiting Bayesian network, one for donor site and the other for acceptor site. The topology structures and the upstream (downstream) nodes of these two models were optimized in consideration of the biological characters of acceptor site and donor site. Both of the models were trained by a ML(maximum likelihood) algorithm for Bayesian network learning, then the testing DNA sequence data were feed into the model and a 10-fold cross validation method was used to evaluate the performance of prediction. The experimental results show that in average, the sensitivity of acceptor site detection was 92.5% and the specificity was 94.0%, the sensitivity of donor site detection was 92.3% and the specificity was 93.5%. These results proved that the models were better than the models based on independent matrix and conditional probability matrix, as well as the hidden Markov model for splice junction site detection in some ways. These conclusions indicate that the optimized Bayesian network models are powerful tools for splice junction detection in eukaryotic genes.

关键词

贝叶斯网络(Bayesian network); 剪接位点(Splice junction site); 供体位点(Acceptor site); 受体位点(Donor site)