基于二次判别的果蝇启动子识别

林昊、李前忠*
内蒙古大学

通过对果蝇polⅡ启动子和非启动子的序列特征分析，计算了序列每个位点单碱基保守性M1(1) 值和六联体保守性M6(1)值。从而分别选取两个区域的六联体频数作为离散源参数，利用离散增量结合二次判别函数（IDQD）对启动子进行了预测。对于从编码区和内含子中选取的非启动子数据集，启动子的预测成功率分别达到93%和89%。比较结果显示IDQD模型能够有效地提高启动子预测成功率。

A Model for Drosophila melanogaster Promoter Prediction

Based on the statistical analysis of D. melanogaster promoter characteristics, the M1(1) and M6 (1) were calculated. By utilizing intrinsic features, take the hexamers frequency of polⅡ promoter sequences as parameters of diversity source, Increment of Diversity with Quadratic Discriminant (IDQD) model was used to predict promoters. The non-promoter sets were selected from introns and coding regions. The predicted results of 10-fold cross-validation exhibited that the sensitivity was 93% for promoter vs CDS, and 89% for promoter vs intron. It was showed that IDQD could improve predictive capability.

关键词

polⅡ启动子；(polⅡ promoter)；离散增量(Increment of Diversity)；二次判别函数(Quadratic discriminant)；六联体(Hexamer)