

DNA 序列中模式发现的一种快速算法

李冬冬*、王正志、杜耀华、晏春
国防科技大学自动控制系

模式发现是生物信息学一个重要的研究方向，目前的大部分算法不能保证获得最优的模式。本文提出了一种新的穷举搜索算法：判据搜索算法（Criterion Search Algorithm, CRISA），把模式发现问题转化成了一个层次图的路径搜索问题，推导了适用于三个节点的判据，在对局部路径的逐层扩展过程中，使用这一判据淘汰不合格的路径，获得了较快的计算速度。采用深度优先策略实现了搜索算法，使得算法仅仅需要大约4兆的内存。对仿真的和实际的生物序列数据的测试表明了算法的有效性。

A fast motif finding algorithm for DNA sequence

Motif finding is an important research field in bioinformatics, and many algorithms have been developed. But among these algorithms, only few of them can find the correct motif surely, such as MITRA. In this paper, we proved a new exhaust search algorithm named CRISA (criterion search algorithm), which can find the correct motif surely and fast. We first translate motif finding problem to a path searching problem in a level graph. Then, we expand a local path from the first level to the last one, one level per step. Using the criterion deduced in this paper, we can wash out many illegal paths in each step and consequently reduce the calculation time. We realize the algorithm by depth first search

so as to save memory, and the total memory used in the algorithm is about 4M. We demonstrate that

CRISA performs well for both simulative and biological data, and it achieves favorable results relative to other motif finding algorithms.

关键词