生物物理学报(2006) 22(3), 177-196

PDF文档

k-长DNA子序列频数分布研究

王树林[1,2]、王戟[1]、陈火旺[1]、张鼎兴[1]
1　国防科技大学
2　湖南大学计算机与通信学院

在详细阐述了生成DNA序列分形图像的Hao方法后，提出一种能够直观显示k-长DNA子序列频数分布差异性的三维频数分布图生成方法。把3D频数分布图转化为1D对数频谱图，突出显示了频数分布的局部特征，提出k-长DNA子序列频数区划分准则，并详细研究了甚高频数区的n阶零间隔现象，指出n阶零间隔分布就是基因组进化过程所留痕迹的假设，并给出对数频谱图特征的生物学解释。实验发现许多DNA序列频数概率分布近似服从非中心F分布，对于分布呈多峰现象的基因组序列，可采用多个非中心F分布的叠加来拟合。在比较非中心F分布与Gamma分布后，提出一种结合二者在拟合方面具有互补优势的新分布，实验证明这种新分布能够更好地吻合实际DNA序列的频数分布。最后研究了两种特异出现频数（最高出现频数与出现频数为1的k-长子序列个数）与k值的关系，发现不同物种的这两种关系具有良好的一致性。

The research of the occurrence frequency distribution of k-mer in whole dna sequence

The research of the k-mer distribution in genome is helpful for understanding the relationship between the structure of genome and its function, and it plays an important role in the recognition of repetitive subsequences, the partition into intron and exon and the investigation of genome evolution. After introducing Hao method which allows the depiction of frequency of k-mer in the form of fractal image, a novel method that can generate 3D frequency distribution map of k-mer in genome is proposed, and the advantage of the 3D frequency distribution map is that the difference of the k-mer occurrence frequency is exhibited obviously for biologist. Then the criterion of the partition of occurrence frequency segment is proposed on the basis of the 1D histogram which is transformed from 3D occurrence frequency distribution. 1D histogram can show the local feature of the occurrence frequency distribution of k-mer, i.e. the occurrence frequency of k-mer in ultrahigh frequency segment appears discontinuous in integer. The palindromes in forbidden k-mer are roughly studied in forbidden segment. Phenomena of n-order zero interval in ultrahigh frequency is deeply investigated. Moreover, it is proposed that the distribution of n-order zero interval is the mark of the process of genome evolving and many features of the logarithm histogram of occurrence frequency are successfully explained from the view of biology. On the basis of many experiments, it is discovered and validated that the occurrence frequency distribution of k-mer is subjected to non-central F distribution. Applying several non-central F distributions can fit the density distribution of the occurrence frequency of k-mer in genome which has the same number peaks. On the basis of experiments, the comparison between non-central F distribution and Gamma distribution which was proposed to fit genome distribution by Hsieh and Luo is studied through experiments. Due to the complement of the two distributions in fitting genome density distribution, a new distribution which combines non-central F distribution with Gamma distribution is presented, and experiments show that the new distribution is better than any single of the two distributions in fitting genome density distribution. After the relationship between the maximal frequency of k-mer in genome and the length of k-mer and the relationship between the number of different k-mer which occur only once in genome and the length of k-mer are deeply investigated, and it is discovered that the two relationships among many species are consistent, which are the evidences of neutral evolution theory of genome.

关键词

DNA序列(DNA Sequences)；三维频数分布图(3D Frequency Distribution Map)；分形(Fractal)；非中心Ｆ分

布(Non-central F Distribution); k-长DNA子序列(k-mer)