生物物理学报(2001) 17(3), 542-549

PDF文档

DNA序列高维空间-数字编码的运算法则

陈惟昌[1]、陈志义[2]、陈志华[3]、王自强[1]、邱红霞[1]
1 中日友好临床医学研究所生物物理研究室
2 中国科学院自动化研究所国家模式识别实验室
3 中日友好临床医学研究所生物化学与分子生物学研究室

DNA序列的高维空间二进制数字编码,除可以对DNA序列的碱基结构、功能基团、碱基互补、氢键强弱等性质进行编码之外,还可以方便地进行数学运算和逻辑运算。DNA序列高维空间数字编码的运算法则是:(1)根据DNA序列数码的奇偶性质,可以推导出其与末位碱基的对应关系。当DNA序列S的数值X(S)=4n,4n+1,4n+2,4n+3时,其末位碱基依次为C,T,A,G,(n=0,1,2,...)。(2)提出DNA序列高维空间的表观维数N_v,数值维数N_x及差异维数N_d的概念。当N_d=0时,首位碱基为A或G,当N_d=2n或2n+1(n=1,2,...)时,首位碱基为$(C)^n$或$(C)^nT$。(3)推导出DNA序列点突变(单核苷酸多态性SNP)的运算法则。(4)推导出DNA重复序列(Tandemrepeat)的运算法则。(5)提出DNA子序列(subsequence)的概念并定义DNA子序列的定值部X_I(digitalvalue)和定位部Q_I(locationvalue)及其计算公式。(6)推导出DNA序列的延长运算、删除运算、缺失运算、插入运算、转位运算、换位运算和置换运算等的运算法则。(7)通过按位加运算求得DNA序列的汉明距离d_h,碱基距离d_h',基团距离d_h''和共轭距离d_G以及这些距离的意义与联系。(8)分析结果表明DNA序列的数字编码比常规的字符编码在数学运算上具有明显的优越性。

OPERATIONAL RULES OF THE DIGITAL CODING OF DNA SEQUENCES IN HIGH DIMENSION SPACE

Digital coding of DNA sequence has great advantages of mathematical and logical operations. (1). According to the parity of DNA digital sequences, the last nucleotide bases can be determined. When the digital value of DNA sequence X(s)=4n, 4n+1, 4n+2, 4n+3, (n=0, 1, 2, …), the last nucleotide base is C, T, A, G respectively. (2). The difference between the visual dimension N_v and the digital dimension N_x is called the difference dimension N_d of DNA sequence. When N_d=0, the initial nucleotide is A or G, and when N_d=2n or 2n+1,(n=1, 2,…), then the initial nucleotide bases are $(C)^n$ or $(C)^nT$. (3). Operation rules for three kinds of point mutation of DNA sequences (transition, transversion and transformation) are derived. (4). The digital coding for a tandem repeat $(S_p)^n$ is, $X(S_p)^n=X(S_p)(2^{np}-1)/(2^p-1$

.(5). DNA sequence S_k with m subsequences, $X(S_k)=\Sigma^m_{i=1}X(S_I)Q_I$. $X(S_I)$ and Q_I are the digital value and location value of the DNA subsequence S_I repectively. (6). The formulae of truncation operation, the elongation operation, the deletion operation, the insertion operation, the translocation operation, the transformation operation and the substitution operation of DNA subsequences are also deduced. (7). The Hamming value of even bits V_h' in DNA sequence represents the number of purine base and the Hamming value of odd bits V_h'' is the number of keto base. (8). The relationship of the Hamming distance d_h, the base distance d_b, the functional group distance d_f and the conjugate distance d_G between two DNA sequences are also discussed.

关键词

DNA序列数字编码(Digital coding of DNA sequence); 奇偶性(Parity); 表观维数(Visual dimension); 单核苷酸多态性(Single nucleotide polymorphism); SNP; 表达序列标签(Expressed sequence tags); EST; 重复序列(Tandem repeat); DNA序列运算法则(Operation rules for DNA sequences)