文章编号:1000 - 582X(2003)07 - 0063 - 03

核酸与分子建模及结构表达启动子强度预测*

李 波[1],仇亮加[1],李劲为[1],叶 楠[1,2],李志良[1,2]
(1. 重庆大学 化学化工学院,重庆 400044; 2. 重庆大学 生物工程教育部与重庆市重点实验室,重庆 400044)

摘 要:从核酸分子的一级结构出发,基于分子中原子间距离及各原子电负性,构建了能描述核酸分子结构的系列参数:分子电性边数矢量简称分子电边矢量。据此对 38 个脱氧核糖核酸(DNA)启动子序列的强度进行定量结构活性相关(QSAR)及定量序列活性模型(QSAM)研究,取得良好结果。与其它方法相比,分子电边矢量具结构分辨率高、活性相关性好、计算简便等特点,可望应用于生物大分子的结构表征及活性预测。

关键词:定量构效相关;定量序效模型(QSAM);分子电边矢量;脱氧核糖核酸;启动子序列
中图分类号:R945.1; R927.2 **文献标识码**:A

举世瞩目的人类基因组计划[1]启动和实施,取得了惊人成绩。一个崭新的称为后基因组或功能基因组和蛋白质组的时代已到来。新近以不同常规于碱基对的氢键作用模式的新型碱基对已有开发[2]。现在一种定量构效关系(QSAR)新模型称为定量序列活性模型(简称定量序效模型 QSAM)[2-3]业已提出,其原理是核酸(DNA, RNA)或多肽与蛋白质序列可以表征为一个行矢量即一个数字序列描述子,并构成所谓校正集的独立变量矩阵,对应生物活性称为相依变量或函数,建立定量构效关系模型可预报其它称之为校验集的生物活性[4-6]。笔者从核酸一级结构出发基于 DNA 链中原子间距及各非氢原子电负性构建了能描述 DNA 链结构的分子电性边数矢量(MEEV),继借多元线性回归(MLR)分别建立 DNA 定量结构活性相关模型,结果良好。

1 原理与方法

对于生物大分子,其生物活性将决定于化学结构。故结构表达是定量结构活性相关的关键前提。分子中各原子好比物理学中点电荷,其相互作用可按点电荷相互作用即库仑定律(1)定义为:

$$E = \sum_{i=1\,to\,all}(q_i * q_j)/d_{ij} \propto (q_i * q_j)/d_{ij} \quad (1)$$

其中 q_i、q_j 分别表示第 i、第 j 个原子的相对电负性,d_{ij} 表示 i、j 两原子间的相对距离。一般来说,原子间的距离相隔越远,其相互作用就越弱,因此距离一般取到相隔 30 键联距离就足够了。现以一种脱氧核糖核酸启动子序列片段为例(其碱基 A 腺嘌呤 C 胞嘧啶 G 鸟嘌呤 T 胸腺嘧啶 U 尿嘧啶与有关 ACGT 片段结构隐氢图见图 1)来说明计算法:首先据隐氢图原子编号写出脱氧核糖核酸启动子序列距离矩阵,从而知相隔一定键距的原子相互作用项数。在确定变量时,按原子间相隔最短键距分别将所有作用加和,就得到一组能表征结构的描述变量,称为分子电距矢量(MEDV,简记为 ν)。ν_k 表示第 $k(k = 1,2,\cdots m)$ 个描述变量,即原子相距 k 键的所有作用。如启动子序列中间相邻原子间的相互作用有:多组 C - C 作用即 1 - 2、2 - 4、7 - 8,…;多组 C - N 相互作用即 2 - 3、4 - 6、6 - 7,…;多组 C - O 相互作用即 4 - 5、8 - 9、8 - 10,…;共多组相互作用。所有相加和得:$\Sigma_{i=1\,to\,all}(q_i * q_j)/d_{ij} = 17.185$。同法可计算其它元素 $\nu_k(k = 2,3,\ldots 30)$,获得矢量为 ν = (17. 185, 6. 568, 3. 169, 1. 398, 445. 265, 237. 503, 58. 309,

* 收稿日期:2003 - 03 - 11
基金项目:重庆市应用基础项目(01 - 3 - 6);国家春晖计划教育部启动基金(99 - 38/04);霍英东基金(98)及国家新药基金(96 - 101 - 01 - 08)资助
作者简介:李波(1980 -),男,广西桂林人,回族,重庆大学化学工程与工艺本科生。

54.332, 53.738, 44.602, 37.417, 33.066, 27.256, 29.992, 27.257, 24.303, 20.148, 16.299, 13.381, 13.090, 13.606, 12.763, 10.695, 8.933, 7.530, 8.275, 8.079, 7.813, 6.578, 5.595)。其它脱氧核糖核酸启动子序列的 ν 矢量计算方法与此相同。

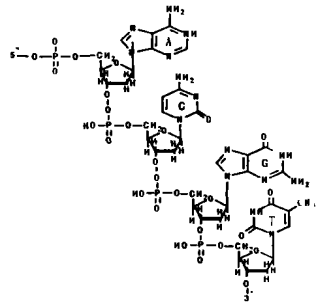

图1 DNA 序列 ACGT 片段隐氢结构图

采用多元线性回归分析进行分子模拟，对不同序列化合物 i ($i = 1,2,...,n$)，将生物活性即启动子强度 $y(i)$ 可表示为若干描述子变量 $x(i,k)$ ($k = 1, 2,...,m$) 的线性贡献。

2 结果与讨论

采用脱氧核糖核酸启动子序列中每个侧链碱基对的立体、几何或电性参数来描述序列链结构，取得一些令人满意进展。但弱点是随着序列链增长、碱基对侧链增加，结构描述参数将成倍地增加。如描述每个侧链碱基对用立体、几何或电性参数为 m，而组成脱氧核糖核酸启动子序列链的碱基对残基数为 n，则结构描述子应为 $n \times m$。对由较少碱基对组成的序列链来说，描述子数目不太多，相对来说结构表征较易；但链较长、碱基对残基较多，那么描述子数目将过多，况并非每碱基对的立体、几何或电性参数都较易得到，因此启动子结构表征就显得较难。另这些方法都未考虑到主序列键影响，这从某种程度上不合理，因对多肽链来说，在形成高级结构时，由于肽键氧原子要参与氢键形成维系高级结构，故在肽链结构表征时只考虑侧链而完全不考虑主链是不妥的。本文提出的分子电边矢量 ν 是基于序列链的基本结构(一级结构) 出发，借助鲍林电负性根据点电荷间作用原理而提出的，有一定物理基础。只需知道启动子一级结构和各原子电负性，毋需要碱基对残基的其它任何性质，故分子电边矢量计算十分简易。

表1 部分脱氧核糖核酸序列启动子活性的观察值与计算值

编号	序列名称	观测活性	计算活性	相对	绝对
01	D/E20	1.748	1.829	-0.081	-4.633
02	N25	1.477	1.365	0.112	7.583
03	G25	1.278	1.353	-0.075	-5.868
04	A1	1.881	1.833	0.048	2.551
05	A2	1.301	1.158	0.143	10.99
06	CON	0.602	0.794	-0.192	-31.89
07	LAC/UV5	0.518	0.683	-0.165	-31.85
08	N25/03	0.903	0.858	0.045	4.983
09	CON/03	0.903	0.772	0.131	14.51
10	CON/PEX	1.204	1.070	0.134	11.12
11	CON/ANTI	0.255	0.343	-0.088	-34.51
12	L-8A	1.672	1.483	0.189	11.30
13	L/CON	1.146	1.437	-0.291	-25.39
14	N25/04	1.246	1.433	-0.187	-15.00
15	N25/05	1.173	1.355	-0.182	-15.51
16	CON/05	1.173	1.011	0.162	13.81
17	L/N25USR	1.763	1.523	0.24	13.61
18	LS2	2.217	2.195	0.022	0.992

注：obs. 为活性观察值；cal. 为活性计算值；* 为预测值。

运用多元线性回归技术建立启动子序列分子电边矢量 ν 与相应启动活性间定量结构活性相关模型，根据模型标准回归系数可以看出 ν 中第7～16、18、20～22、24、26 等16个(或20个加23、28、29、25) 元素对活性的贡献较大，而这几个元素中第7、10、14 元素的贡献尚没第12、15、16、21、18 等几个元素的贡献大；而后几个元素表示键距分别相隔12、15、16、21、18 等。为什么会出现键距相隔12、15、16、21、18 的相互作用比键距相隔1、2、3、4～11 的影响还大这种奇特现象呢?这可能是因形成高级结构后，键距分别相隔12、15、16、21、18 等的原子间距由于形成氢键或其它作用维系高级结构而拉近，其相互作用变大。该结果从另一方面说明：对于启动子序列，其生物活性不仅与一级结构有关，且与其高级结构有关；一级结构是基础，高级结构是影响活性主要因素。于是选取第12、15、16、21、18 七个元素建立 QSAR 模型，其结果与选用30个元素建模结果很相近。表1列出了相应观察与计算值。然后运用分子电边矢量(MEEV) 对所有38个启动子序列作训练集采用逐步多元回归方法 SMR 建模，据标准回归系数从中选取7或4个元素作描述变量建模，结果良好，其计算值亦列入表1。用 ν 计算的结果与文献结果相近或更好。但计算简便，只需知道序列一级结构便可计算出分子电边矢量 ν，不需其它任何理化参数或理论参数。

基于脱氧核糖核酸启动子序列链的一级结构，借助鲍林电负性提出了一组结构表征新参数即分子电边矢量 ν，并分别以38个启动子序列借多元线性回归建

立 QSAR 模型。根据提取出来的主要影响因子可找到某些高级结构信息。且分子电边矢量 ν 只需序列链一级结构信息，毋需其它任何有关侧链碱基对的立体、几何或电性参数即可得到。具有计算简便、活性相关性好的优点，可望在多肽、核酸等生物大分子的结构表达及定量结构活性相关(QSAR) 研究中得到更广应用。

致谢　绿色化学与药物设计研究室及化学生物与分子药学研究室李声时教授、刘树深教授、周原副教授、杨胜喜工程师、廖春阳工程师、张梦军讲师，应用化学 98－2 班王远强、覃仁辉、邹竹惠、吴世容学士等提供有关帮助或协助。谨致谢忱。

参考文献：

[1] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome[J]. Nature, 2001, 409:860－921.

[2] PICCIRILLI J A, KRAUCH T, MORONEY S E, et al. Quantitative sequence－activity models of nuclei acids by a multivariate approach[J]. Nature, 1990, 343: 33－35.

[3] CHARTON M. The quantitative description of amino acid, peptide, and protein properties and bioactivities[J]. Prop Phys Org Chem, 1990, 18:163－284.

[4] COCCHI M, JOHANSSON E. Amino acids characterization by GRID and multivariate data analysis[J]. Quant Struct Act Relat, 1993, 12:1－8.

[5] HELLBERG S, SJOSTROM M, SKAGERBERG B, et al. Peptide quantitative structure－activity relationships, a multivariate approach[J]. J Med Chem, 1987, 30: 1 126－1 135.

[6] LIU S, LI Z. Approach to estimation and prediction for normal boilins points of alkanes based on a molecular distance－edge rector(MDE)[J]. J Chem Inf Compt Sci, 1998, 38(3):387－394.

Molecular Electronegativity Distance－edge Vector for Structural Expression and Activity Prediction of Deoxyribonucleic Acid Promoter Sequences Quantitative Sequence－Activity Models (QSAMs)

LI Bo[1], *QIU Liang-jia*[1], *LI Jin-wei*[1], *YE Nan*[1,2], *LI Zhi-liang*[1,2]

(1. College of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, China;
2. Ministry of Education and Chognqing City key Laboratory of Biomedical, Chongqing University, Chongqing 400044, China)

Abstract: Based on the distance between atoms and firstly electronegativity of each atom, a new set of descriptors called molecular electronegativity edge-distance vector (VMED) applied to describe molecular structure of *E. coli* transcriptional DNA－promoter sequences, is firstly proposed, because it is easy to calculate, only from primary structure of DNAs. Here a new type of quantitative structure-activity relationships (QSARs) called quantitative sequence-property models (QSAM) is developed with good forecasting ability by multiple linear regression (MLR) method. The results show that VMED has both excellent structural selectivity and good activity estimation. Besides, this novel vector will be useful to structure characterization and activity prediction of biological molecules. The resulting structural descriptors can be used to investigate requirments for new nucleic acids(NAS) in order to obtain sequences with altered activities.

Key words: molecular electronegativity edge-distance vector (VMED); structural parameterization; theoretical descriptors; quantitative sequence-activity models (QSAM); *E. coli* transcriptional DNA-promoter sequences; Nucleic acid bases; multiple linear regression (MLR); Quantitative structure activity relationship (QSAR)

(编辑　张小强)