生物信息学及其在病毒研究中的应用

肖 明

(上海师范大学生命与环境科学学院,上海 200234)

摘 要:介绍生物信息学的原理与方法及其在病毒研究中的应用. 保守序列承担着极其重要的功能,通过多重序列对齐搜寻保守序列,是生物信息学方法的基础;敏感位点是一种反映蛋白质或核酸功能的特定模式,因此,通过数量关系的优化推导敏感位点,可用于分离病毒蛋白质与核酸相互作用位点;核苷酸和氨基酸序列只有形成了三级或四级结构才能表现功能,通过同源建模预测蛋白质的高级结构,有助于疫苗的研制、抗病毒药物的筛选以及药物的分子设计. 预测RNA的三级结构大多从RNA折叠入手. 病毒蛋白质三级结构预测比较成功的是日本脑炎病毒包膜糖蛋白的三级结构.

关键词:生物信息学;保守序列;同源建模

中图分类号:Q73 **文献标识码**:A **文章编号**:1000-5137(2003)03-0096-07

至2003年3月止,收录在GenBank已测基因组全序列的病毒种类达1075种,而且正以每年20.1%的速度递增[1]. 依靠传统的研究思想和实验手段注释如此庞大的生物信息资源是不可能的. 生物信息学的出现为解决这个难题带来了希望,而且在短短的几年里,做出了惊人的成绩. 那么怎样将生物信息学的原理与方法应用于病毒研究中?作者综合了国内外的最新研究成果和本实验室的研究工作,对这个问题进行了探讨.

生物信息学的主要任务是以计算机为主要研究手段研究基因与蛋白质的功能. 如何将数字符号与序列符号联系起来?生物信息学的记分法能将氨基酸或核苷酸序列(生物信息学将它们称为"元素")通过某种手段转化为简单的、直观的、便于计算机处理的数值.

几种主要记分法如下:

(1)性质矩阵法 用能体现元素特征的理化性质(如疏水性、极性、带电性、芳香性、分子大小等)来描述序列中出现的特定元素. 具有某种性质的元素记为1,不具此性质的记为0[2].

(2)遗传密码矩阵(genetic code matrix, GCM) 该方法依据对比的两个氨基酸中的一个转变为另一个氨基酸所需要改变密码子碱基的个数. 适应于进化分析.

(3)结构-遗传矩阵(structure genetic matrix, SGM) 主要根据对比双方元素的相似性以及相应遗传密码的变化频率. 如对比的两个元素相同时记最高分,不相同记较低分,最低为0分.

(4)突变数值矩阵(mutation data matrix, MDM或pointed accepted mutation, PAM)是Dayhoff根据71组蛋白质家族进化过程中氨基酸替代的实际数目以及相对突变率所创立的半经验记分法[3].

(5)氨基酸替换矩阵(amino acid substitution matrix) 该方法主要基于2000个Blocks中对齐片段聚

收稿日期:2003-06-02

基金项目:国家自然科学基金项目(猪瘟病毒基因组3非编码区结构与功能关系的研究)资助(30170214);上海市高等学校科学技术发展基金项目(猪瘟病毒基因组启动RNA合成的必需位点的研究)资助(03DZ08)

作者简介:肖明(1961-),男,博士,上海师范大学生命与环境科学学院生物系副教授.

类百分比与氨基酸匹配的关系[4]. SGM 和 PAM 是目前应用较多的记分法,尤其是 PAM 在确定类似性较小的序列关系时很有效,所以 PAM 的应用比较普遍.

1　通过多重序列对齐搜寻保守序列

通过多重序列对齐搜寻保守序列是生物信息学方法的基础,几乎所有的注释序列的意义、研究序列的结构的方法都是建立在此基础上的. 多重序列对齐是指多条相关序列的对齐分析. 保守序列是指病毒在进化过程中基因组序列保持不变或变异很小的序列. 在进化过程中,变化很小,或者不变的序列往往承担着极其重要的功能,一旦出现变化,功能就会受影响或者被破坏,物种就有被淘汰的危险. 因此,保持不变或变化很小的一组序列可能具有相同的功能,尤其当其中有已被实验证明了的更是这样. 从严格意义上讲,相同的序列,只有证明了它们是同源才能判定具有相同功能. 确定保守序列的基本步骤是,首先从一大堆无规则的序列中聚积相似序列,通过多重序列对齐从相似序列集合中建立起反映序列特性的对齐片段,然后在对齐片段中确定反映功能的保守序列. 国际上已有专门的数据库(如 Blocks[5],PROSITE[6],IDENTIFY[7] 和分析软件(如 BLAST, DNAsis, FASTA, GCG, MOST[8], Emotif[7], Tool[9])用于保守序列的分析. 动态程序算法(dynamic program algorithm)[9],点矩阵作图法(dot matrix)[3]等也是常见的方法.

病毒自身携带或者利用宿主细胞与复制有关的酶,以复制方式进行增殖[11]. 几乎所有复制酶作用底物是核苷三磷酸或脱氧核苷三磷酸. 按照酶与底物相互作用机制可以推测这些复制酶一定具有某些共同的特征. 猪瘟病毒(classical swine fever virus, CSFV)的 NS5B 是一个比较重要的基因,它编码的蛋白质与病毒基因组的复制有关. 将 CSFV 的石门株(shimen)和兔化弱毒株(HCLV)的序列与两个黄病毒属的 West Nile Virus(WNV),Yollow Fever Virus(YFV)和瘟病毒属的 BVDV 相应序列对齐进行分析. 发现在相应序列内有一保守性很强的序列对齐即 Block. 在此 Block 中不难发现隐含在序列相应位置的保守三肽,这就是保守序列, 见图 1。

WNV	ERLSRMAVSGDDCVVKPLD
YFV	DRLKRMAVSGDDCVVRPID
BVDV	NRLVRIHVCGDDGFLITEK
HCLV	DRVAKIHVCGDDGFLITER
Shimen	DRVAKIHVCGDDGLLIFER

组成一个 Block, 保守序列 GDD 位于其中. WNV 即 West Nile Viurs,YFV 即 Yellow fever virus 同属黄病毒科黄病毒属. BVDV, HCLV, Shimen 同属黄病毒科瘟病毒属. 各序列起始氨基酸的位置没有标明.

图 1　几个病毒的具有 RdRp 功能的蛋白质相应序列

现已发现所有依赖 RNA 的 RNA 聚合酶(RNA-dependent RNA polymerase, RdRp)中均含有保守的三肽结构 GDD. BVDV 的 NS5B 为 BVDV RNA 复制所必需,与 CSFV 同属黄病毒科的人丙型肝炎病毒(HCV)的 NS5B 基因编码的产物已被证明具有 RdRp 活性,该蛋白也具有保守的 GDD 三肽结构[12]. 黄病毒科的其他病毒相应 NS5 基因产物也具有 GDD 三肽结构. 据此可以推测 CSFV 的 NS5B 具有 RdRp 的功能.

2　通过数量关系的优化推导敏感位点

敏感位点是一种反映蛋白质或核酸功能的特定模式. 一个具有显著意义的敏感位点能较真实地反

映被检核酸或蛋白质序列的结构和功能特征. 因此,敏感位点的推导是功能预测、相互作用位点的搜索、基因组图谱的绘制、分子模拟设计的基础,特别对定点突变具指导作用. 通过数量关系的优化推导敏感位点的方法有:

(1) 最大期望值算法(expectation maximization algorithm) 简称 EM 法. EM 法主要原理是计算能描述数据的期望参数值,然后反复优化这些参数值,直到能从多重序列对齐中找出多序列共有的分子结构和生物学特性的局部模式(local patterns)[13].

(2) 权值矩阵(Weight matrix)法 上述方法给出了反映序列特征的最优模式. 模式中元素对反映序列特征的贡献是平均化的. 事实上,蛋白质、酶以及核酸的活性部位中元素的作用是有差异的,因此,包含在敏感位点中的各元素除了出现的频率外还应有能反映贡献差异的数学模式. 权值矩阵法在这方面有所侧重. Gribskov 的权值矩阵法[14]是在对齐序列与最佳对齐序列之间引入一个 Profile,即权值矩阵. 这个 Profile 是通过元素频率矩阵转化而来的. 而最佳对齐序列是使用动态程序算法在权值矩阵的基础上得到的. 权值矩阵法的本质是定位罚分以及氨基酸偏性二者的位置依赖性,前者引入了结构信息,后者引入了在每个位置允许侧链特性的信息.

(3) 信息量(Information Content)法 用于分析蛋白质与核酸相互作用的信息理论首先是由 Schneider 等提出的[15],后经 Berg[16], Stormo[17,18]等进行了补充完善. 该理论的核心是,具有相同亲和性的序列在进化过程中有相同的概率被与之作用的蛋白质或酶选为结合位点,而结合的自由能与亲和性直接相关,即两序列有同样的自由能,就有同样的结合蛋白质的亲和性. 而结合的自由能直接与碱基利用率相关. 这样,我们就可将某序列被选为结合位点的概率通过亲和性、自由能与碱基利用率联系起来,也就是说碱基利用率的大小与序列被选为结合位点的概率有关. 而信息量实质上就是表示一个序列被选为结合位点的机会的多少,是通过碱基利用率计算出来的. 该方法主要用于从大量的核酸序列中分析蛋白质的结合位点. 至今已对结合阻遏蛋白的操纵区[16,19]、结合核糖体的翻译起始区[15]、cAMP 结合蛋白在转录起始区的结合区[17]、DNA 转录的启动子[15,16]以及 Archaea 的转录调节位点[20]进行了信息量的分析. 2002 年,我们用信息熵的原理对信息量法进行了补充,使其更接近实际情况(见图 2)[21].

$$\langle p(b_l)\rangle = \int p(b_l)\rho^{(l)}(p(b_l),nb)\delta(1-\sum p(b_l))\prod \mathrm{d}p(b_l) = \frac{n(b_l)+1}{N+4}, \tag{1}$$

$$I_l = -\sum_{b_l=A,C,G,T} p(b_l)\ln[p(b_l)/p^0(b_l)] \tag{2}$$

$$I_{(seq)} = \sum_{l=1}^{s} I_l. \tag{3}$$

$I_{(seq)}$ 为核酸与蛋白质发生相互作用位点(site)的信息量,而 Il 则是位点中某个位置(position)的信息量,N 为某位置包含的总的碱基数,Il 则是位点中的位置. $p(b_l)$ 是碱基出现的概率, $p^0(b_l)$ 是碱基在基因组出现的比率.

图 2 对猪瘟病毒基因组 3′非翻译区进行信息量分析的公式

我们利用补充了的信息量法对猪瘟病毒基因组 3′非翻译区(3′UTR)进行了信息量的分析[21]. 猪瘟病毒基因组 3′UTR 是病毒复制酶与 RNA 相互作用的起始位点,为了确定 3′UTR 中与复制酶相互作用的位点,我们按照图 2 的公式,设计了一个程序对 20 条 3′UTR 序列进行了分析. 我们以此为例来说明如何利用信息理论分析病毒的核酸与蛋白质的相互作用. 表 1 是从 20 个 CSFV 毒株的 3′UTR 序列中分离出的每个毒株在 3′UTR 的复制酶起始识别位点. 也就是说,复制酶在这里结合,开始复制各毒株的基因组. 这个序列矩阵是从大量的序列矩阵中经反复抽提离析出来的,因此被认为是最好的. 图 3 为 3′UTR 信息量分析结果,图 3-1 为碱基频率矩阵,其数据来自表 1. 纵列 A,C,G,T 对应的数字分别为位置(1)1-21 的碱基 A,C,G,T 出现的次数,图 3-2 为位置 1-21 的 Il 值以及 21 个 Il 值的和, I(seq)值,是通过公式 1 和 2,按图 3-1 的数据计算出来的. 选择最好序列矩阵的标准就是 Il 值最大. 从结果可以看出位置 9,12,8 和 13 的 Il 值较大,其中位置 9 的 Il 值最大,表明这些位置可能对复制酶在这里起始结合所发挥的作用较大;13.9102 是被选择出来的最大的 I(seq)值,其数据来源是表 1 中的序列矩阵,也就

是说,该序列矩阵组成的位点在各个毒株3′UTR中含有最大的信息量,表明该序列矩阵中各个序列最有可能是不同毒株的复制酶在3′UTR的起始结合位点.

表1　CSFV 3′UTR中可能的复制酶起始识别位点

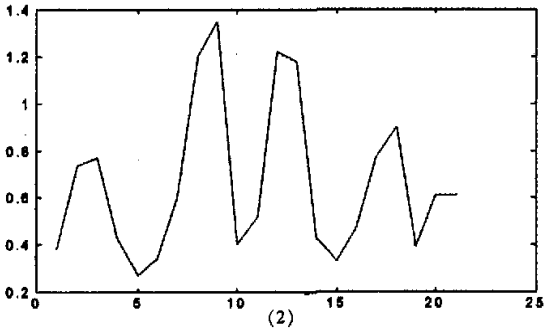
A	B	C	D
SEQ 1(	CAP)	207	AAGGTAATTTCCTAACGGCCC
SEQ 2(	CON-TREM. UK)	162	CAGCACTTTAGCTGGACAGAA
SEQ 3(	RIEMS,C)	208	AAGGTAATTTCCTAACGGCCC
SEQ 4(	CHINESE)	221	TAAGGTAATTTCTAACGGCCC
SEQ 5(	ALFORT-187)	208	AAGGTAATTTCCTAACGGCCC
SEQ 6(	P97)	162	CAGCACTTTAGCTGGAGGGAA
SEQ 7(	D4990. I)	162	CAGCACTTTAGCTGGAGGAAA
SEQ 8(	GLENTORF)	162	AGCACTTTAGCTGGAAGGAAA
SEQ 9(	ALFORT-A19)	208	AAGGTAATTTCCTAACGGCCC
SEQ10(	GPE-)	208	AAGGTAATTTCCTAACGGCCC
SEQ11(	19. SK)	161	CAGCACTTTAGCTGGAGGAAA
SEQ12(	SHIMEN)	208	AAGGTAATTTCCTAACGGCCC
SEQ13(	HCLV)	218	TAAGGTAATTTCTAACGGCCC
SEQ14(	MOORE. UK)	162	CAGCACTTTAGCTGGAAGGAA
SEQ15(	ALD)	208	AAGGTAATTTCCTAACGGCCC
SEQ16(	OLD-LEDER. US)	162	CAGCACTTTAGCTGGAAGGAA
SEQ17(	BRESCIA)	207	AAGGTAATTTCCTAACGGCCC
SEQ18(	ALFORT)	207	AGGTAATTTCCTAACGGCCCC
SEQ19(	U744. D)	162	CAGCACTTTAGCTGGAGGAAA
SEQ20(	A. SK)	161	CAGCACTTTAGCTGCACGAAA

A:序列编号; B:CSFV的不同毒株; C:位点序列第一个碱基在3′UTR的位置; D:可能的复制酶起始结合位点序列.

	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	10	18	2	1	9	9	10	2	1	8	0	0	1	11	11	9	2	1	5	9	9
C	8	0	1	8	1	8	0	0	0	1	10	18	0	0	1	10	1	1	11	11	11
G	0	2	17	10	2	0	0	0	0	1	8	0	1	9	8	1	17	18	4	0	0
T	2	0	0	1	8	3	10	18	19	10	2	2	18	0	0	0	0	0	0	0	0

I_l 0.38 0.74 0.77 0.43 0.27 0.34 0.60 1.20 1.35 0.40 0.52 1.22 1.18 0.43 0.33 0.47 0.77 0.90 0.39 0.61 0.611$_{(seq)}$ = 13.9102

(1)



(2)

图3(1)为3′UTR碱基频率矩阵,数据分别来自表1中的序列. 图3(2)为位点1-21的I_l值,$I_{(seq)}$值以及相应的直观图,I_l值通过公式(1,2)根据碱基频率矩阵的数据计算出来,$I_{(seq)}$值通过公式(3)计算出来. 图3(2)中的横坐标表示l,纵坐标表示I_l值,从图中可观察各位置重要程度.

图3　3′UTR信息量分析结果

3 通过同源建模预测序列的高级结构

核苷酸和氨基酸序列只有形成了三级或四级结构才能表现功能. 了解病毒蛋白质和核酸高级结构是非常重要的,它有助于疫苗的研制、抗病毒药物的筛选以及药物的分子设计. 在科学实践中,我们经常会遇到一级结构相似,而空间结构相差较远,以及一级结构差异很大而高级结构、功能非常相似的现象. 目前对大分子空间结构测定的方法一般是用 X 光衍射以及核磁共振(NMR). 这些方法能较精确地测定大分子的高级结构. 著名的蛋白质和核酸三维结构数据库 PDB(http://www.pdb.bnl.gov/),专门收集通过 X 光衍射和 NMR 确定了结构的蛋白质和核酸. 然而仅靠 X 光衍射和 NMR 远远跟不上序列测定的速度,X 光衍射需要高纯度的结晶,周期要求长;NMR 也只能测定较小的蛋白质分子的结构. 不了解空间结构,就很难确定大分子的功能,更谈不上作用机理的阐明. 因此,随着计算机科学的发展人们开始着手高级结构预测的研究. 常见的预测蛋白质核酸高级结构的方法是同源建模(homology modeling). 所谓同源建模就是选择行使同一功能、同源性较高的且空间结构已被 X 光衍射或 NMR 确定了的蛋白质或核酸作为参考模板,从而构建序列三级结构模型的方法. 一般分如下几个步骤:(1) 选定参考模板. (2) 一级结构、二级结构对比分析. (3) 三维结构模型构建. (4) 模型精炼. (5) 模型评估.

现在有一种新的预测蛋白质三级结构的方法,称之为穿线法(threading),从蛋白质折叠入手,将各种蛋白质分成不同的折叠类型,将折叠类型与结构类型对应起来,也是用于病毒蛋白质结构功能研究的好方法.

与蛋白质三级结构研究相比,RNA 三级结构研究更慢. 人们大多从 RNA 折叠入手预测 RNA 的三级结构[22]. 对 RNA 折叠的研究不仅将最终确定 RNA 的三级结构,也将阐明 RNA 从一级结构向三级结构转变的过程,弄清 RNA 在行使功能时,其空间结构以及各种结构成分构象的变化,这样我们预测模拟出的 RNA 分子是一个"会动的小虫". 探求 RNA 折叠的奥秘无疑对阐明空间结构以及对研究生命起源有重要意义.

最近有人提出所有 RNA 的结构只不过是各种结构 Motifs 的连结. 这些结构 Motifs 有独立的三级结构,而且大都已被弄清. 这些结构 Motifs 是 RNA 分子的基本结构成分,它们的有机结合则是 RNA 空间结构的实质[23]. 于是将庞大的 RNA 分子的空间构象转变成几个容易对付的结构 Motifs 的研究. 下面是几类已确定晶体结构的结构 Motifs:

(1) 出现在 RNA 转折处的终端回环 Motifs(Terminal loop Motifs),如 U-转折(U-torn),四核苷酸环(Tetraloops).

(2) 不形成互补碱基对的螺旋片段的内环 Motifs(internal loop Motifs),如链交叉嘌呤集结团(Cross-Strand Purine Stacks),鸟嘌呤突出 Motifs(Bulged-G Motifs),腺嘌呤平台(A-Platforms),突出-螺旋-突出 Motif(Bulge-Helix-Rulge Motif),结合金属的 Motifs.

(3) 序列上相距较远,通过堆积折叠聚积一团的三级 Motifs(Tertiary Motifs)如核糖拉链(Ribose Zippers),四核苷酸环-螺旋相互作用 Motif(Tetraloop-Helix Interactions).

这些方法和 Motifs 无疑对病毒特别是对 RNA 病毒的三级结构的预测以及功能的相互关系的确定奠定了坚实的基础.

病毒蛋白质三级结构预测比较成功的是日本脑炎病毒(Japanese encepalitis virus, JEV)包膜糖蛋白(envelope glycoprotein, Egp)的三级结构[24]. 日本脑炎是严重的地方性流行疾病,其病原体 JEV 表面的 Egp 是抗体中和反应的主要目标抗原. 对 Egp 三级结构的了解是研制高效疫苗的首要条件. 1999 年 Kolaskar 对 Egp 的三级结构进行了预测,建立了它的空间结构模型,并获得了 Egp 在水溶液中的构象. 4 个独立的评估方法对其预测的结果进行了论证,认为在立体化学和几何学上都是可以接受的. Kolaskar 等使用的参考模板为已定晶体结构的与 JEV 同属黄病毒科黄病毒属的蜱传媒脑炎病毒(tick borne encephalitis)的 Egp. RNA 三级结构被弄清的为数较少,其中包括 Hepatitis delta virus 的核酶(ri-

bozyme)[25],为预测病毒 RNA 折叠与三级结构提供了参考模板.

利用生物信息学的方法注释病毒核苷酸、氨基酸序列的意义存在的主要问题是需要提高准确度,得到的结论需要通过实验手段进一步完善. 我们认为,一门学科发展的初期过分强调准确会扼杀这门学科. 事实上,这些方法都在逐步地改进. 在啤酒酵母完整基因组确定的 5932 个基因中,大约 60% 是通过信息分析得来的,其中 2950 个编码基因的功能也是通过相似性搜索确认的,占编码基因的 48%[26]. 最近在研究恶性传染性疾病非典型肺炎(severe acute respiratory syndrome, SARS)的病原体的过程中,就是利用生物信息学的方法对其基因组进行分析,为确定该病原体为一新的冠状病毒(coronavirus)起到了关键作用[27,28]. 可见,生物信息学的作用是很大的,因此生物信息学在病毒研究的作用会越来越明显.

参考文献:

[1] http://www.ncbi.nlm.nih.gov.

[2] ZHANG B H, Ding D F. Search and Databank establishment [J]. Acta of Biochemistry and Biophysics, 1995, 27(4): 367-373.

[3] 王槐春. 蛋白质与核酸序列分析的基础[M]. 北京:人民军队出版社, 1994.

[4] HENIKOFFS, HENIKOFF J G. Amino acid substitution matrices from protein blocks [J]. Proc Natl Acad Sci USA, 1992, 89:10915-10919.

[5] HENIKOFF J G, PIETROKOVSKI S, HERIKOFFS. Recent enhancements to the Blocks Database servers [J]. Nucleic Acids Research, 1997, 25(1):222-225.

[6] BARROCH A, BUCHER P, HAFMAMK. The PROSITE databases, its Etatus in 1997 [J]. Nucleic Acids Research, 1997, 25(1):217-221.

[7] NEVILL-MANNING C G, WU T D, BRUTLAG D L. Highly specific protein sequence motifs for genomes analysis [J]. Proc. Natl. Acad. Sci. USA, 1998, 95:5865-5871.

[8] TATUSOV R L, ALTSCHUL S F, KOONIN E V. Detection of conserved segments in proteins: Iterative scanning of sequence databases with alignment blocks [J]. Proc. Natl. Acad. Sci. USA, 1994, 91:12091-12095.

[9] ITHIMURE D. DNA analysis: New kids on the Block [J]. Science, 1999, 285:355-356.

[10] NEEDLEMAN S B, WUNSCH C D. A general method applicable to the search for similarities in the amino acid sequence of two proteins [J]. J Mol Biol, 1970, 48:443-453.

[11] STRAUSS J H, STRAUSS E G. Viral RNA replication: with a little from the host [J]. Science, 1999, 283:802-804.

[12] MEYERS G, RUMENAPF T, THIEL H J. Molecular cloning and nucleotide sequence of the genome of Hog Chol-era Virus [J]. Virology, 1989, 171:555-567.

[13] CARDON L R, STORMO G D. Expectation maximization algorithm for identifying protern-binding sites with variable longths from unaligned DNA fragments [J]. J. Mol. Biol., 1992, 223:159-170.

[14] GRIBSKOW M, MCLACHLAN A D. EISERBERG D. Profile analysis: detection of distantly related proteins [J]. Proc Natl Acad Sci USA, 1987, 84:4355-4358.

[15] SCHNEIDER T D, STORMO G D, GOLD L, et al. Information content of binding sites or nucleotide sequences[J]. J. Mol. Biol., 1986, 188:415-431.

[16] BERG D G, VON HIPPEL P. H, Selection of DNA binding sites by regulatory proteins: statistical-mechanical theory and application to operators and promoters [J]. J. Mol. Biol., 1987, 193:723-750.

[17] STORMO G D, HARTZELL Ⅱ G W. Identifying protein-binding sites from unaligned DNA fagments [J]. Proc Natl Acad Sci USA, 1989, 86:1183-1187.

[18] STORMO G D, FIELDS D S. Specificity, free energy and information content in protein DNA interactions [J]. Trends in

Biochemical Sciences, 1998, 23:109-113.

[19] FIELDS D S, HE YI-YUAN, AL-UZRI A Y, et al. Quantitative specificity of the Mnt repressor [J]. J Mol Biol, 1997, 271:178-194.

[20] GELFAND M S, KOONIN E V, MIRONOV A A. Predication of transcription regulatory sites in Archaea by a comparative genomic approach [J]. Nucleic Acids Research, 2000,28(3):695-705.

[21] XIAO M, ZHU Z Z ,LIU J P, ZHANG C Y. Prediction of recognition sites for classical swine fever virus genomic replica - tion with information analysis [J]. Molecule biology, 2002, 36(1):48-57.

[22] FERRE-D'AMARE A R, DOUDNA J A. RNA folds: insights from rescent crystal structural motifs in RNA[J]. Annu Rev Biophys Biomol struct,1999,28:57-73.

[23] MOORE P B. Structural motifs in RNA. Annu. Rev. Biochem [J]. 1999,67:287-300.

[24] KOLASKAR A S KULKORNI-KALE U. Prediction of Three-Dimensional structure and mapping of conformational epitopes of envelope glycoprotein of Japanese encephalitis virus [J]. Virology, 1999, 261:31-42.

[25] FERRE-D'AMARE A R,ZHOU K, DOUDNA J A. Crystal of a hepatitis delta virus ribozyme [J]. Nature,1998, 395: 567-574.

[26] MEWES H W, ALBERMANNK, BAHR M, et al. Overview of the yeast genome [J]. Nature, 1997, 387(6632 suppl): 7-65.

[27] ROTA P A, OBERSTE M S, MONROE S S, et al. Characteraction of a novel coronavirus associated with severeacute respiratory syndrome [J]. Science, 2003, 300: 1394-1398.

[28] MARRA M A, JONES S J M, ASTELL C R et al. the genome sequence of the sars-associated coronavirus [J]. Science, 2003, 300: 1399-1404.

Bioinformatics and its Application in Virology

XIAO Ming

(Colloge of Life and Environment Sciences, Shanghai Normal University, Shanghai 200234, China)

Abstract: Application of theory and methodology of bioinformatics in virology was introduced. Conserved sequences played very important role. It was the basis for bioinformatics that the conserved sequences are searched for through multi-sequence alignment. Sensitive positions were the motifs by which the function of proteins and nucleic acids were assumed. The sensitive positions were induced through optimizing quantity relation, which was applied to isolation of interactive positions between proteins and nucleic acids. The sequences of nucleotide and amino acids were functional when the space structures were formed. Prediction of high structure of protein through homology modeling was helpful to research of vaccine, selection of the antiviral drug, and designation of molecule drug. Most predictions on RNA space structure began from RNA fold. The prediction of space structure of envelope glycoprotein of Japanese encepalitis virus was a successful sample.

Key words: Bioinformatics; conserved sequences; homology modeling