建立蛋白质序列的同源性

王 槐 春

(军事医学科学院基础医学研究所,北京 100850)

提 要

蛋白质序列同源性比较是研究蛋白质结构、功能与进化的基础。分子生物学家依靠计算机程序进行序列对准比较。本文主要综述计算机介导的序列比较的基本原理和评价序列类似之显著性水准的方法,并介绍多序列同时比较的方法。

关键词 序列比较,同源性, Needleman-Wunsch 算法,多序列比较

近年来分子生物学家日益广泛地依靠计算机对核酸和蛋白质大分子序列进行比较和分析,其范围从先前的主要是不同种属生物的同一类分子的序列比较扩展到对不同来源、不同功能和类似性很小的序列的比较和结构分析,揭示出许多新的结构与功能关系[1]。目前国内大多数实验室所用的序列比较软件引自国外,有关序列比较的数学方法研究在我国几乎是空白,这大大影响了有关软件的设计和开发。蛋白质序列比较主要用于建立蛋白质序列的同源性,国外对其理论研究和算法的研究很多,并有很多很好的综述发表[2-4]。本文主要介绍计算机介导的序列比较的基本原理及建立序列同源性的方法。

序列比较的基本原理

一、四种计分方法

所有比较序列的方法都是将序列间残基的相似与不相似点转化为数值后进行比较,也就是给相同或相似的残基(氨基酸或核苷酸残基)的配对赋予正分,对不同的配对进行罚分,再计算总积分。在作序列对准比较时,有时需在序列中引入空位(gap)以获得最佳对准。然而,空位的引入意味着两序列间的距离增加,因而通常要对它扣分。目前已有四种预置数值的矩阵对相同、相似和不同的残基匹配进行记分,现分别介绍如下。

1. 酉矩阵(Unitary Matrix, UM),又称单位矩阵(identity matrix)。这是一种最简单的矩阵,也适用于比较核苷酸序列。UM 记分法是基于残基的相同或不同两种情况。当被比较的两序列在同一位置上的残基相同时,其配对积分设为 1.0,否则取 0.0。这种方法只对相同的残基记分,因而只适用于比较两个同源性很大的序列或在一个序列中寻找某一个非常短的残基片段。因此该矩阵的应用受到限制。

2. 遗传密码矩阵(genetic code matrix, GCM)不同氨基酸所对应的三联密码子的个数不同,有些只有一个,有些有多个,GCM 记分法是根据使一种氨基酸转化为另一种氨基酸所需要的密码子碱基的改变个数来确定[5]。如果变化一个碱基使某些氨基酸的密码子改变为另一些氨基酸的密码子,则这些氨基酸的配分积分为 1;如果需要 2 个碱基的改变,其积分为 2;如此类推。GCM 最常用于进化距离的计算,其优点是计算结果可以直接用于绘制进化树。但是它在蛋白质序列对准比较尤其是类似程度很低的序列比较中很少被使用。

3. 结构-遗传矩阵(structure-genetic matrix, SGM)众所周知,二十种氨基酸中,有些残

基在结构和化学性质上都很相似，例如氨基酸对 D/E，N/Q，R/K，S/T 和 V/I 等等. SGM 把残基之间的相似性与 GCM 中遗传密码子改变的频率结合起来加以考虑[6](表 1)，理论上应该比 UM 和 GCM 计分法更合理，实际使用效果也比它们好.

表 1　两组衡量20种氨基酸相似性的矩阵元素值

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	6	4	2	2	2	3	2	1	0	1	2	2	0	2	2	2	2	3	3	3	C
		6	5	4	5	5	5	3	3	3	3	3	3	1	2	2	2	3	3	2	S
C	12		6	4	5	2	4	2	3	3	2	3	4	3	3	2	3	1	2	1	T
S	0	2		6	5	3	2	2	3	3	3	3	2	2	2	3	3	2	2	2	P
T	-2	1	3		6	5	3	4	4	3	2	2	3	2	2	2	5	2	2	2	A
P	-3	1	0	6		6	3	4	4	2	1	3	2	1	2	2	4	1	2	3	G
A	-2	1	1	1	2		6	5	3	3	4	2	4	1	2	1	2	1	3	0	N
G	-3	1	0	-1	1	5		6	5	4	3	2	3	0	1	1	3	1	2	0	D
N	-4	1	0	-1	0	0	2		6	4	2	2	4	1	1	1	4	0	1	1	E
D	-5	0	0	-1	0	1	2	4		6	4	3	4	2	1	2	2	1	2	1	Q
E	-5	0	0	-1	0	0	1	3	4		6	4	3	1	1	3	1	2	3	1	H
Q	-5	-1	-1	0	0	-1	1	2	2	4		6	5	2	2	2	2	1	1	2	R
H	-3	-1	-1	0	-1	-2	2	1	1	3	6		6	2	2	2	3	0	1	1	K
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6		6	4	5	4	2	2	3	M
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5		6	5	5	4	3	2	I
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6		6	5	4	3	4	L
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5		6	4	3	3	V
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6		6	5	3	F
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		6	3	Y
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		6	W
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10		
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17	

上三角为 SGM 矩阵，下三角为 Dayhoff 的 MDM 矩阵
本文中氨基酸用标准的单字母代码表示，A：丙氨酸 C：半胱氨酸，D：天冬氨酸 E：谷氨酸 F：苯丙氨酸 G：甘氨酸 H：组氨酸 I：异亮氨酸 K：赖氨酸 L：亮氨酸M：蛋氨酸 N：天冬酰胺 P：脯氨酸 Q：谷氨酰胺 R：精氨酸 S：丝氨酸 T：苏氨酸 V：缬氨酸 W：色氨酸 Y：酪氨酸

4. Dayhoff 突变数值矩阵 (mutation data matrix, MDM) 或称 PAM (pointed accepted mutation) 矩阵. Dayhoff 等(1978 年)选择71组密切相关的蛋白质序列进行对准比较，绘制出这些蛋白质家族的进化树，并据此推导出它们的祖先序列. 根据推导出的祖先序列，她计算了在进化过程中，一种氨基酸(如 A_{ai}) 被另一种氨基酸 (A_{aj}) 所替代的实际数目 (A_{ij})，又称可接受的点突变数 (PAMs). 根据这些改变，她确定了 20 种氨基酸中每一种氨基酸的相对突变率 m_i. 结合 A_{ij} 和 m_i 就形成了 MDM 中的元素值(表 1)[7]. 该矩阵是上述四种矩阵中唯一的根据蛋白质进化过程中所观察到的改变而设计的矩阵，为半经验计分方法. Feng 等 (1985)[6] 分别用上述四种矩阵记分法使用 Needleman-Wunsch 对准方法详细比较了酪氨酸激酶家族和球蛋白家族的序列，发现当两个序列之相同率大于 30% 时，四种记分法所得的对准结果几乎一样；当两个序列的相同率小于 20% 时，不同记分法产生结果差异很大，其中数 Dayhoff 记分法在确定序列的微弱同源关系时最有效，其次是 SGM 记分法. 但是 Dayhoff 计分法也有其缺陷[8]. 例如，该矩阵赋予色氨酸的自身配对以最高分数 (表 1)，这将导致有关的序列比较程序偏向于使两序列的色氨酸对准而忽略了寻找有重要生物学意义的对准类型，诚然，色氨酸在进化上确实是一个高度保守并只有一个密码子的残基，它还有较大的侧链基团，使之可能在二级结构确定中发挥重要作用. 但是给色氨酸的高计分会干扰获得其它明显的对准片段. 半胱氨酸可能比色氨酸更重要，因为它在许多蛋白质的活性部位起重要作用，而且由两个半胱氨酸所形成的二硫键是蛋白质三维结构的重要特征.

考虑到上述各种记分方法的各自的合理性和不足之处，Argos (1988) 提出把 Dayhoff 计分法、残基的 5 个主要物理特征(疏水性、侧链基团的大小及形状、形成二级结构转角的趋势、形成反平行 β 折叠的趋势和氨基酸的折射率)和用于序列比较的多个序列窗口 (window) 大小的方法综合加以考虑，既增加了对同源序列识别的敏感性，又大大降低了无关序列的局部类似性的干扰[4].

二、Needleman-Wunsch 算法

Needleman 和 Wunsch (1970) 提出的求两个序列最佳对准的数学方法至今仍是编制对准程序的依据[9]. 该方法的关键是在计算机上设计一个二维阵列，它们的两个轴就是要比较的二序列 (图 1)[10]. 首先根据上述的残基匹配记分法在该阵列的每一个位置上赋予一定数值，得到一个原始矩阵；然后从该矩阵的右下角(对应于序列的羧基末端)开始，把矩阵中的数值按如下规定逐行相加得到一个“转化”的矩

(1)原始矩阵(用SGM矩阵计分)

	V	E	D	Q	K	L	S	R	C	N
V	6	4	3	2	3	5	2	3	2	2
E	4	6	5	4	4	1	3	4	0	3
N	2	3	5	3	4	1	5	4	2	6
K	3	4	3	4	6	2	3	6	0	4
L	5	1	1	2	2	6	2	2	2	1
T	3	3	2	3	4	2	5	4	2	4
R	2	2	2	3	5	2	3	5	2	2
P	3	3	2	3	2	3	4	2	2	2
K	3	4	3	4	6	2	3	6	0	4
C	2	0	1	1	0	2	4	0	6	2
D	3	5	6	4	3	1	3	3	1	5

(2)转化的矩阵

	V	E	D	Q	K	L	S	K	C	N
V	(50)	46	40	35	30	26	19	14	8	2
E	42	(44)	42	37	31	22	20	15	6	6
N	35	36	(38)	36	31	22	20	15	7	6
K	33	34	32	31	(33)	23	20	17	5	4
L	34	30	30	29	25	(27)	19	13	7	1
T	29	29	28	29	27	23	(22)	15	7	4
R	23	23	23	24	26	23	20	16	7	2
P	20	20	19	20	19	20	21	13	7	2
K	14	15	14	15	17	13	14	(17)	5	4
C	8	6	6	6	5	7	9	5	(11)	2
D	3	5	6	4	3	1	3	3	1	(5)

序列1 V E D Q K L S K C N

序列2 V E N K L T R P K C N

图1 Needleman-Wunsch 对准方法的工作原理示例

阵：把下一行各盒子（box）中的最大允许值加到该盒子左边的上一行及上一列的所有盒子中．转化完成后，再追踪出最大积分路线，它代表最佳对准方式．其执行情况是：从左上角（相当于序列的氨基末端）开始向下向右方向追踪（垂直向下和向下向左的方向均不允许），得到最高积分的脊状路径．该路径自动离开矩阵对角线位置就意味着在某一个序列中要引入一个“空位”．空位的引入自然能产生较好的对准结果，但一般不能代表真正的进化事件，因而要用对总积分罚分的方法来调整路径．当脊状路径离开对角线越远时，对总积分的罚分就越多．

Needleman 和 Wunsch 算法以及在此基础上发展起来的 Sellers 和 Waterman 等人的方法都使用动态编程方法进行全序列的对准，这些方法的成功主要取决于两序列间的相似程度，随着所选择的对引入空位的罚分参数的不同，对准结果也不同，甚至对紧密相关的序列[11]．由于残基的插入或删除事件大多发生在蛋白质二级结构的无规则卷曲部位，因此在形成 α-螺旋或 β-折叠等部位引入空位就应该罚分，这样可以增强序列对准与结构的相关性．对于相关性极远的序列，同源性可能局限于一些关键性残基或序列片段，它们在不同蛋白质中的分散程度变化很大．目前已经发展了几种算法鉴别这种局部同源的序列[12-14]．

三、序列类似的显著性水平的评价

用序列比较程序所得到的序列最佳对准结果并不等于提供了两个序列的某种关系．那么应该如何对序列类似性进行评价呢？这一问题的答案并非简单直接．序列对准通常出现如下几种情况[4]：（1）两个较长的微弱相似的序列（通常指相同率在 15—25% 之间）需要作类似性可信限的概率估计．（2）两个序列的某一部分对准得很好，而其余部分很难配对．（3）空位的罚分问题，不同的严紧程度产生不同的对准结果，严紧的罚分会不允许空位插入，而松弛的罚分甚至使任何两个无关序列达到 100% 的配对．（4）两个序列中一串四个或更多的残基完全相同，而其余区域不能很好地对准，或者至多只是少数对准．（5）一段较短序列（20 到 50 个残基）对准得较好（相同率在 20—40%），但其余部分的相同率很低．

序列类似之显著性检验的典型方法是使两个序列的每一个序列随机打乱（jumbling），再在同样条件下（即使用相同的程序与变量参数）进行对准比较，计算这些随机序列的相同率或其它对准积分，这一过程重复一定次数（通常 50—100 次），得到随机序列对准积分的正态分布曲线，其平均值与标准差分别用 μ 和 δ 表示，该两个真序列比较的对准积分为 x，利用公式

$$z = \frac{x - \mu}{\delta}$$

计算大于或等于 x 的对准积分的概率（z 值用标准差（SD）为单位）[7]。根据图 2，当 z 值为 3.1、4.3 和 5.2SD 单位时，对准积分为 x 的随机

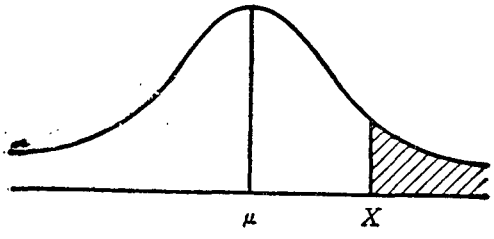

图 2　随机序列对准积分的正态分布的概率
对准积分大于或等于 x 的概率用 z 值表示，即 x 距离均值（μ）有多少个标准差（δ）单位，$z = (x - \mu)/\delta$

对准积分≥X 的概率	Z (SD单位)	Z (SD单位)	对准积分≥ X 的概率
10^{-1}	1.28	0.0	0.500
10^{-2}	2.33	1.0	0.159
10^{-3}	3.09	2.0	0.227×10^{-1}
10^{-4}	3.72	3.0	0.135×10^{-2}
10^{-5}	4.26	4.0	0.317×10^{-4}
10^{-6}	4.75	5.0	0.287×10^{-6}
10^{-7}	5.20	6.0	0.987×10^{-9}
10^{-8}	5.61	7.0	0.128×10^{-11}
10^{-9}	6.00	8.0	0.622×10^{-15}
10^{-10}	6.36	9.0	0.113×10^{-18}
10^{-15}	7.94	10.0	0.762×10^{-23}
10^{-20}	9.26		

出现概率分别是 10^{-3}、10^{-5} 和 10^{-7}。一般假定 z 值大于 5SD 时，表示两个被比较的蛋白质在进化上相关；z 值在 3—5SD 之间时，如果两者有其它方面相类似的证据（如功能相似）也表明同源，z 值小于 3SD 时表示不同源[15]。国外设计的许多序列比较软件包都带有计算 z 值的程序，可直接用于评价序列对准的显著性水平，例如 NBRF 蛋白库的 ALIGN 程序给出序列对准的结果及其 z 值[7]，Kanehisa 的 IDEAS 程序包和 Lipman-Pearson 的 FASTP 程序包分别含有 SEQDP 和 RDF 程序计算 z 值[16,17]。然而，并非每个人都有机会接触计算机及有关软件，因此有时需要有一些经验规则来判断序列对准的显著性。

经验法则[10]：（1）如果两个序列都长于 100 个残基，在适当地加入空位之后，它们配对的相同率达到 25% 以上，则两个序列确实相关（即同源）。（2）如果配对的相同率小于 15%，则不管两个序列的长度如何，它们都不可能相关。（3）如果两序列的相同率在 15—25% 之间，它们很可能相关，可用下面公式计算标准化的对准积分（normalized alignment score, NAS），并参考图 3 来确定类似性的显著水平。

NAS = [相同残基(半胱氨酸除外)配对数 ×10＋半胱氨酸配对数×20－空位数 ×25]÷两序列之较短序列的残基数 ×100

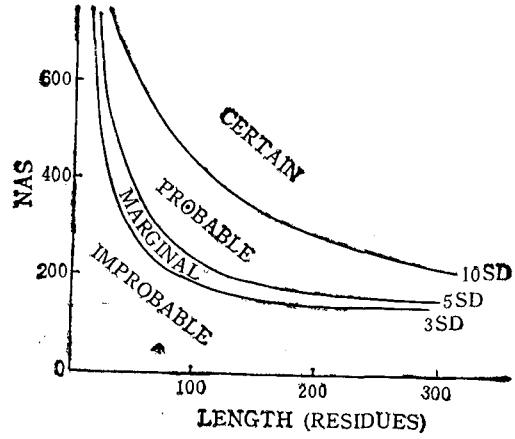

图 3　确定序列类似的显著水平的粗略参考曲线

短片段（如 20 到 40 个残基）的对准提出另一个评价问题。通常它们都包含一个或多个完全相同的四、五甚至六聚体片段，有时这些寡聚体由重复的残基组成，如精-精-精-精-精或丙-脯-丙-脯-丙-脯等模式。Argos 和 McCaldon (1988) 给出一个概率值表[4]，包括了不同长度不相关的蛋白质序列中出现一个或多个残基n-聚体的概率，由此可以确定被比较序列的相关性，此外还应从生物学和功能角度判断有关序列的同源性。

多序列对准比较

如果说两个序列对准比较主要用于建立两个序列的同源关系和推测它们的结构与功能联系，那么同时比较一组序列对于分子进化和结

构功能研究更为有用. 例如，某些在生物学上有重要意义的相似性只能通过把一组序列对准后才能识别，同样地，只有在多序列比较之后功能相关蛋白质的序列类型或结构成分才可能被发现. 此外，多个类似性较小的同源蛋白质的对准比较可能揭示，两个或三个从来不被替换的对功能有重要影响的残基，可以作为蛋白质工程的定点诱变部位.

从理论上讲，用于两个序列比较的Needleman-Wunsch 算法可以扩展到三个或多个序列的比较，即在计算机上建立一个三维或多维矩阵. 但是除掉多个很短序列的对准外，三维以上矩阵所需要的内存空间和程序动态优化所需要的时间太长，使得这种直接延伸 Needleman-Wunsch 算法的方法变得不现实. 因此，直到几年以前，多数的多序列对准都是在一系列的两两对准之基础上，通过用肉眼对一些关键性残基进行对准后再调整其他残基的方法获得，Doolittle 称之为"肉眼对准"(eyeball alignment)[18].

Murata 等 (1985 年) 在 Needleman-Wunsch 算法的基础上，在 VAX 计算机上建立了两个三维矩阵来比较三个序列[19]. 虽然这种方法减少了运算时间，但是它所占的空间太大，只能用于比较长度在 100 个残基左右的三个序列. Doolittle 等[20,18]提出"渐近对准"(Progressive alignment) 方法，它首先用 Needleman-Wunsch 算法把所有序列做两两对准，然后进行全面对准. 它遵循的一条简单定律是"一旦有了一个空位，总有一个空位" (once a gap, always a gap). 当第一对最类似序列对准后，第三个与它们最相似的序列加入进去对准，按照上述定律，一旦在某一部位引入了空位，随后加入的序列在相同部位都引入空位. 接着第 4 个最类似的序列加入进去，不断重复上述过程，直至达到使所有的序列对准. Doolittle 等用这种方法比较了多达 11 种球蛋白序列，说明了其有效性. 但是，该方法未使用动态编程方法，因而缺少一种对序列对准质量进行精确的和全面的衡量——什么使一种对准比另一种对准更佳. Lipman 等[21]提出多序列对准的一种新算法，既把动态编程方法引入其中从而提高了对准的质量，又大大减少了该算法对计算时间和空间的需要，这种方法在多个类似性很小的、全序列的对准比较中最有效.

结 束 语

以上主要介绍了序列比较的基本原理和建立序列同源性的方法. 目前国际上已发展了几十个序列分析软件包，这种软件包能进行序列比较、核酸或蛋白质序列数据库的快速同源性检索以及蛋白质结构预测等功能[22]. 有了分子序列数据库及其应用程序，我们就能很方便地对一个新测定序列或大量报道的序列进行分析研究，并可能揭示出新的结构与功能的关系.

本所姚志建研究员审阅了全文，特此致谢.

参 考 文 献

1 Doolittle R F. *Trends Biochem Sci*, 1985; **10**: 233
2 Doolittle R F. *Science*, 1981; **214**: 149
3 Dayhoff M O, Barker W C, Hunt L T. *Methods Enzymol*, 1983; **91**: 524
4 Argos P, McCaldon P. In: Setlow J K ed., *Genetic engineering-Principles and methods*, New York: Plenum Press 1988: Vol. **10**, p. 21
5 Sellers P H. *SIAM J. Appl Meth*, 1974; **26**: 787
6 Feng D F, Johnson M S & Doolittle R F. *J Mol Evol*, 1985; **21**: 112
7 Dayhoff M O, Schwartz R M, Orcutt B C. In: Dayhoff M O ed., *Atlas of protein sequence and structure*, Natn Biomed Res Found, Washington, D C, 1978: Vol. 5, Suppl. 3, p 345—352
8 Chtterjee D, Maizel J V Jr. *Gene Anal Techn*, 1987; **4**: 27
9 Needleman S B, Wunsch C D. *J Mol Biol*, 1970; **48**: 443
10 Doolittle R F. In: Doolittle R F ed. *Of URFS and ORFS——a primer on how to analyze derived amino acid sequences*. California, University Science Books, 1987: 27—29
11 Blundell T L et al. *Nature*, 1987; **326**: 347
12 Goad W B, Kanehisa M I. *Nucleic Acids Res*, 1982; **10**: 247
13 Taylor W R. *J Mol Biol*, 1986; **188**: 233
14 Patthy L. *J Mol Biol*, 1987; **198**: 567
15 Barker W C, Dayhoff M O. *Proc Natn Acad Sci USA*, 1982; **79**: 2836
16 Kanehisa M I. In: Kanehisa M I ed. *IDEAS-Integreated*

(下转第 274 页)

其机制仍与 P_{21}^{ras} 激活 PI 转换，继而激活 PKC 有关，未转移的乳腺癌细胞 PKC 水平低于转移细胞，应用 PKC 激活剂 PMA，能诱导细胞的转移表型[23]。

肿瘤转移趋化因子可以来自瘤细胞自身分泌即自泌性移动因子（autocrine motility factor），或来自周围组织。它们通过肿瘤细胞相应受体使细胞发生迁移（migration），促进其浸润与转移，而 PKC 可以提高这类受体对趋化因子的亲和力[24]。

此外，PKC 能降解肿瘤坏死因子受体，使该受体所介导的细胞杀伤作用减弱，提示 PKC 可促进肿瘤细胞生长[25]。一些抗癌剂如三苯氧胺和黄源酸盐等已被证实是特异性 PKC 抑制剂[21]，从另一角度证实了 PKC 在促进肿瘤生长中的作用。因而近来提出 PKC 抑制剂可能会是有效的抗癌剂。

综上所述，PKC 在许多环节上参与了细胞分化与增殖以及恶性转变的调控，但距彻底阐明这些作用相差甚远。由于 PKC 的调节作用广泛存在着双向性，细胞内对 PKC 活性的调节的多样性，以及生物信息传递过程的复杂性，使得深入研究 PKC 的作用机制难度较大。但是，进一步探讨这些机制无疑对于阐明肿瘤发生发展机理具有重大意义。

本文承赵明伦教授审校，在此致谢。

参 考 文 献

1 Housey G M *et al. Cell*, 1988; 52(3): 343
2 Nishizuka Y *Cancer*, 1989; 63(10): 1892
3 Strulovici B *et al. Biochemistry*, 1989; 28(8): 3569
4 Whitfield J F *et al. Cancer Metastasis Rev*, 1987; 5(3): 205
5 Rozengurt E. *Science*, 1986; 234(4773): 161.
6 Frick K K *et al. J Biol Chem*, 1988; 263(6): 2948
7 Susa M *et al. Cell*, 1989; 57(5): 817
8 Schroder H C *et al. Biochem J* ,1988; 252(3): 777.
9 Kaji H *et al. J Biol Chem*, 1988; 263(27): 13588
10 Martelly I *et al. Exp Cell Res*, 1989; 183: 92
11 Persons D A *et al. Cell*, 1988; 52(3): 447
12 Hsiao W L W *et al. Mol Cell Biol*, 1989; 9(6): 2641
13 Girard P R *et al. Cancer Res*, 1987; 47(11): 2892
14 Blackshear P J *et al. FASEB J*, 1988; 2(14): 2957
15 Angel P *et al. Cell*, 1988; 55(5): 875
16 Kammer G M *et al. Immunol Today*, 1988; 9(7—8): 222
17 Kolesnick R N *et al. J Biol Chem*, 1988; 263(14): 6534
18 Liss A R. 生物科学动态，1989；6：18
19 Hennings H *et al. J Invest Dermatol*, 1987; 88: 60
20 Lacal J C *et al. Nature*, 1987; 330(6145): 269
21 Lopez C A *et al. Cancer Res*, 1989; 49(4): 895
22 Halazonetis T D et al. Cell, 1988; 55(5): 917
23 Korczak B *et al. Cancer Res*, 1989; 49(10): 2597
24 Blood C H *et al. J Biol Chem*, 1989; 264(18): 10614
25 Johnson S E *et al. J Biol Chem*, 1988; 263(12): 5686

[本文于1990年4月23日收到，9月3日修回]

（上接第 268 页）

databases and extended analysis system for nucleic acids and proteins (users manual). Natn Cancer Institute, Frederick, Maryland, 1987
17 Lipman D J, Pearson W R. *Science*, 1985; 227: 1435
18 Feng D F, Doolittle R F. *J Mol Evol*, 1987; 25: 351
19 Murata M et al. *Proc Natn Acad Sci USA*, 1985; 82: 3073
20 Johnson M S, Doolittle R F. *J Mol Evol*, 1986; 23: 267
21 Lipman D J et al. *Proc Natn Acad Sci USA*, 1989; 86: 4412
22 Lewitter F I, Rindone W P. *Methods Enzymol*, 1987; 155: 582

[本文于1990年5月14日收到，7月13日修回]