（真核）3′端核苷酸序列中部分碱基相互作用的识别方式显然不足以解释人线粒体内 mt-mRNA与核蛋白体的相互作用。这一问题有待今后的研究加以阐明。

参 考 文 献

[1] Anderson, S. et al.: *Nature*, 290, 457, 1981.

[2] Borst, P. & et al.: *Nature*, **290**, 443, 1981.

[3] Borst, P. & et al.: *Nature*, 289, 439, 1981.

[4] Montoya, J., et al.: *Nature*, **290**, 465, 1981.

[5] Ojala, D., et al.: G.: *Nature*, **290**, 470, 1981.

[本文于 1981 年 7 月 13 日收到]

核酸序列分析与蛋白质结构功能研究的关系

王 培 之

（中国科学技术大学 生物系）

自从生物学的研究发展到分子水平以后，科学家们竞相从新的高度在生物大分子结构、遗传学和生物膜等各个研究领域中开展研究，并且很快都取得了十分重要的成果。但在开始时，他们相互间的联系和渗透却不多，正如 J. C. Kendrew[1] 所指出的那样：当时，实际上存在着两类分子生物学家——结构学家（structurists）和信息学家（informationists）。有一个时期，这两者几乎是完全分立的两大学派。以 M. Delbrück 和 S. E. Luria 为代表的“噬菌体分子生物学家”，他们的主要兴趣是用一维的、亚显微的信息实体来分析、解释病毒和细菌的遗传学，这个信息实体最终归之于 DNA 分子；另一个学派是以 W. T. Astbury、J. D. Bernal 和他们的学生为代表的结构学派，他们的兴趣在于发展测定生物大分子，尤其是蛋白质的三维结构的技术，但对遗传学等却不关心，或者很不感兴趣。

现在这两个学派之间已有了较大的渗透，因为无论如何，遗传现象及其信息实体——核酸，与蛋白质的结构功能总是不可分割地相互联系的。正如 Crick 在 1958 年就提出的那样：生物信息的传递方向是 DNA → RNA → 蛋白质，以后又进一步发展成为分子生物学的所谓“中心法则”。到 1966 年提出了遗传密码之后，则从根本上解决了信息联系的具体方式和确切内容，亦即基因和蛋白质的同一线性关系，或者说解决了核酸一级结构和蛋白质一级结构的专一对应关系。

1954 年 Sanger 发表了胰岛素的全部氨基酸排列顺序之后，至今已有近千个蛋白质的一级结构被测出，其中最大的蛋白质包含了一千多个氨基酸残基，成百种蛋白质的空间结构已得到确定或正在研究，并据此解释这些蛋白质是怎样利用自身的结构来行使其功能的。

但是，核酸的顺序测定工作长期以来落后于蛋白质的顺序测定，直到 1975 年，还是由 Sanger 等建立了测定 DNA 顺序的“加”、“减”法，接着 Maxam 和 Gilbert 又建立了“化学法”之后，核酸的顺序测定才变得非常快速而简便，只要几天功夫就可以测定一个含有一、二百个核苷酸的 DNA 片段，测定四、五千个核苷酸长度的 DNA 分子也只要花半年到一年的时间，以至于许多研究人员认为，这些技术的建立预告了分子生物学的新纪元的到来[2]。而从近几年所取得的成果来看，这样的说法确实是一点也不过份。这方面已有不少专文综述，本文想就核酸序列分析与蛋白质结构功能研究的关系方面作一些介绍。

第一、可以利用测定核酸的顺序来校对和推测蛋白质的顺序。

早在 1972 年，W. M. Jon[3] 等就测定了一种称为 MS_2 的 RNA 噬菌体中决定其外壳蛋白结构的 RNA 片段的一级结构，并且和它的外壳蛋白中的氨基酸排列顺序进行对照，其结果完全一致。

最近(1979 年) Dennis Piszkiewicz[4] 等测定了 ATP 磷酸核糖基转移酶的一级结构，又拿它和另一个实验室的 W. M. Barnes 等在 1978 年测定的并为之编码的 his G 基因一级结构作比较，结果也完全匹配，而且确定氨基末端的残基是甲硫氨酸。把氨基端和 his G 基因结构比较后指出，在蛋白质合成后，N-甲酰基已被除去。

又如大肠杆菌 i 基因的全结构测定以后，发现乳糖操纵子的阻遏蛋白氨基酸顺序的测定结果有错误，后来证实，在其氨基酸顺序测定过程中，果然漏掉了一个肽段，从而得到了纠正。由此可见，这样的校对颇有好处。

正因为核酸的顺序测定发展到了比蛋白质顺序测定要方便得多的地步，因此，在比较容易取得要测蛋白质的基因片段时，或者在要测蛋白质产物比较难以得到时，可以考虑把蛋白质的顺序测定转化为 DNA 的顺序测定。

例如，SV 40(多型瘤病毒)的全结构的测定完成之后，只要分离到 SV40 所编码的蛋白质，测定其 N—端开始几个氨基酸，就可以从对应的核苷酸顺序推测出该蛋白的全结构了。

再比如，在测定了线粒体 DNA 的顺序以后，不仅可以弄清其基因是根据原核还是真核方式组织的，而且还可以较方便地确定线粒体基因翻译产物的一级结构，而用一般方法这是较难做到的[5]。最近，Sanger 实验室在进行人的线粒体 DNA 顺序分析的过程中，发现有一段核苷酸所对应的氨基酸顺序与牛的细胞色素氧化酶亚基 II 的氨基酸顺序非常相似。而且这种相似性在整个分子中都可以看到。因此，他们得出结论，此段 DNA 顺序相当于细胞色素氧化酶亚基 II 基因的顺序。当然，关于这个基因翻译产物的一级结构也就自然能够确定了。

此外，在寻找激素等一类多肽或蛋白的前体时，如果其 mRNA 比较容易分离或提取的话，就可以利用麦胚的核糖核蛋白体系，将纯化的 mRNA 翻译成对应的前体蛋白质或多肽，再进行氨基酸排列顺序的测定。如胰岛素原和甲状旁腺激素原等的前体都是以这种方法发现的。当将甲状旁腺激素的 mRNA 加入上述体系中时，合成出了比甲状旁腺激素原还要大的蛋白质，经一级结构测定，确定它是甲状旁腺激素原的前体，即前甲状旁腺激素原[6]。

但是，如果某些多肽或蛋白其对应的 mRNA 在组织里面所有具有 PolyA 顺序的 RNA 中，仅以十分低的比例存在，比如低于 1% 时，就很难使用上述方法了。1979 年 B. E. Noyes[7] 等在分离和鉴定胃泌素（又名促胃液激素）的 mRNA 和它的前体时，就采用了一个新的可供普遍使用的方法。他们取胃泌素中一段独特的氨基酸顺序 Trp–Met–Glu–Glu 为模板，合成一段对应的脱氧十二核苷酸 d(C–T–C–C–T–C–C–A–T–C–C–A)。再用这个脱氧十二核苷酸作引物，加到猪胃窦的 RNA 中，就能专一地使猪胃泌素的 mRNA 被反转录成胃泌素的 cDNA (即互补 DNA 链)；以凝胶电泳分离 cDNA，再对它进行顺序分析，结果找出了对应于 G34(胃泌素原，34 肽) 的 mRNA 顺序的 cDNA 片段。再将这个经过鉴定的 cDNA 片段进行末端标记作为探针，在存在甲汞化氢氧化物的条件下，对从猪胃窦提取的富 Poly A 的 RNA 混合物进行琼脂糖凝胶电泳，使经过标记的 cDNA 与之杂交，就能检测出胃泌素的 mRNA 了。经过定位以后，成功地鉴定了一个对 G 34 编码的 620 个核苷酸大小的 mRNA。同时指出，在 G34 编码区的两边还存在着附加的氨基酸顺序，这个实验结果可以预言胃泌素的前体是一个 110—140 个氨基酸残基长度的多肽，由此提供了纯化和最后克隆(扩增)这个至今未弄清楚的前体分子 mRNA 的必要方法。这个对于存在量较少的 mRNA 种类的核苷酸水平的研究，进而找到对应的蛋白或多肽的前体的方法所达到

的新成就，是十分鼓舞人心的。

第二、可以通过对基因的核苷酸序列分析预言未知蛋白或多肽的存在，从而进一步找出这种新的蛋白或多肽。这方面的例子也是不少的，而且也是非常吸引人的。

比如在已知 $\phi \times 174$ 噬菌体的基因顺序后，通过对基因的核苷酸序列的仔细观察，发现有未被使用的起始和终止信号存在，就根据这一发现预言必定存在着它所编码的新的蛋白质。结果，果然找到了称之为K蛋白的新蛋白质[8]。

再比如，对于牛促肾上腺皮质激素-β-脂肪酸释放激素前体的克隆cDNA的顺序测定和分析以后，就推测有部分顺序是为其它尚未发现的激素编码的[9]，而且，从发现的为新的多肽编码的核苷酸序列分析来看，在它的两侧也存在着一对碱性氨基酸，这就推测可以由前体蛋白的水解、加工形成象 α-MSH（促黑激素）或 β-MSH 那样的新的多肽。根据 α-MSH 类推，这个多肽在加工时，还可能在C-端形成苯丙酰胺。同时，根据新的多肽片段和 α-MSH 或 β-MSH 在结构上的值得注意的相似，建议将它命名为 γ-MSH。

第三、最近几年有人提出由“外显子”（exon）或“小基因”（mini-gene）来推测蛋白质的“功能区域”[10]的设想，也是很有意思的。

有不少蛋白质在结构上有几个分立的部分，分别执行着一种专门的功能。这些执行专门功能的部分称之为“功能区域”。在弄清了一个蛋白质的氨基酸排列顺序（一级结构）和在自然状态时折叠的图式（空间结构）之后，如要进一步弄清在这个蛋白质分子上有多少个分立的“功能区域”，却是十分不容易的，有时简直是很难实现的。

但是，由于核酸顺序测定技术的进展，1977年夏季，在冷泉港的专题论文集里首次刊登了关于在SV40和 Ad_2 等几种病毒中的某些基因是具有外来插入顺序的报道。由于这些病毒是利用它们的宿主的酶系统的，因此包含了一个很清楚的意思，说明高等生物的基因可能也具有插入顺序，或者说也是隔裂的。过去几年里，美国和欧洲的不少研究人员发现，许多基因是被隔裂成片段的，它们的中间是由和基因编码蛋白无关的DNA片段（或者叫做插入顺序）所连接的，而不像细菌的基因那样，是独个的连续单元。现在，被发现的隔裂基因的名单日增月长，以至对于高等生物基因是否都是隔裂的这一点已不成问题了，相反，成问题的倒是这类基因中是否有不隔裂的。甚至认为在高等生物中，所有为蛋白编码的基因上，至少存在一个插入顺序。而从目前已发现的隔裂基因中看，大部分有2—4个插入顺序，有的甚至多达数十个。插入基因顺序是和结构基因顺序一起被转录到信使RNA的前体中去的，然后，经过酶的作用，把插入顺序“裁剪”出来，再把结构基因片段“拼接”（splice）起来，成为信使RNA，然后被翻译成蛋白质。MIT的 Walter Gilbert 建议用名词“intron”（内含子）来表示“非编码部分”或“插入顺序”；用“exon”（外显子）来表示“编码区”或“基因片段”。Roger Lewin 则还把“基因片段”称之为“mini-gene”（小基因）。关于插入顺序的生物功能已有很多讨论，不少人认为隔裂基因的这种组织方式，在进化和调控上是有特殊意义的。在本文中，我们感兴趣的是，这些被隔裂的各个“基因片段”或“小基因”，往往分别和整个基因编码的蛋白质的各个功能区域相对应，这在免疫球蛋白中是最明显的。免疫球蛋白分子的主要部分（重链恒定区）是由四个“部分”或“功能区域”所组成的。第一个是和细胞膜相作用的；第二个是和补体分子相结合的；第三个是起枢纽作用的；第四个是结合到免疫球蛋白分子的可变区的。而在免疫球蛋白的基因上正好有三个插入顺序把为四个功能区域编码的基因片段分开。或者说，它的重链恒定区的基因是有效地由四个“小基因”所构成的。还有，血红蛋白分子的蛋白部分的编码基因，是被两个插入顺序隔裂成三个片段的，它的中心片段，或者说，中间的那个“小基因”是为带有含铁的血红素基团的蛋白分子部分编码的。如果随着时间的推移，在这方面能得到更多的例证，从而证明这

个关于"小基因"和"功能区域"相对应的诱人的设想是普遍有效的话，那就意味着，我们将会在蛋白质的结构功能研究方面出现一个新的突破。即使不能达到这种程度，至少到目前为止，在你看到一个高等生物的隔裂基因时，只要数一下它所具有的"小基因"的数目，就有可能在有关这个基因所编码的蛋白中存在多少个"功能区域"方面，得到一个直接的"线索"，这个"线索"无疑是十分宝贵的。当然，就目前来说，关于珠蛋白的基因中，两个较外面的"小基因"是为什么"功能区域"编码的问题，还是一个"谜"；而卵清蛋白有8个"小基因"，它们又分别对应于什么"功能区域"呢？则是一个更大的"谜"。但它为我们提供了对蛋白质结构功能研究方面提出问题的一个崭新的方法。

第四、在核酸顺序测定的方法本身有可为蛋白质一级结构测定所借鉴的地方。

在DNA顺序测定方面一个十分重要的突破，就是一种能识别专一核苷酸顺序的限制性内切酶的发现和广泛应用，因而可以很容易地完成含几千个核苷酸以上的大片段的重迭，亦即建立起DNA顺序的物理图谱。同样，在测定蛋白质的一级结构中，如果也能找到许多只对特定的氨基酸所对应的肽链进行专一切断的肽链内切酶的话，那么，也就可以高效率地得到个数较少，而链却较长的几个多肽片段了。如果能使用多种专一性切断的肽链内切酶，则由于可以取得重叠多肽片段，即建立起蛋白的"物理图谱"，因而也就可以比较容易地排出大的蛋白质分子的一级结构了。为此，人们现在也开始象寻找DNA的限制性内切酶那样，努力寻找能识别氨基酸残基的蛋白质肽链内切酶[11]。除了原有的少数几个，如胰蛋白酶等以外，前些年又发现了几个能严格识别氨基酸残基的酶。第一个是Drapean等发现的*Staphylococcus aureus* V8的菌体外酶，它能专一地切断谷氨酸和天冬酰胺的羧基侧的肽链；第二个是由Doonan等发现的从担子菌*Armillaria mellea*中提取的对赖氨酸专一的肽酶。在含有半胱氨酸的蛋白质中，用三氟乙酰化学方法修饰赖氨酸的ε-氨基之后，只要使半胱氨酸胺乙基化，就可以用同样的酶，选择性地只切断半胱氨酸残基的键；第三个是由Walter等的小组发现的能专一地切断脯氨酸的羧基侧的键的酶。还有人提出，是否可以用人工合成的肽链作为细菌的营养物，有意识地去诱导产生出具有专一性的肽链内切酶。可以预见，随着被发现的对氨基酸残基专一的肽链内切酶的数目的增加，测定蛋白质一级结构的效率肯定也会得到明显的提高。

我们相信，随着分子生物学的进一步深入发展，各个学科之间，各个研究领域之间，各种研究技术和方法之间的互相渗透，互相促进和互相启发必然会日益增多，其结果将会促成新的、更大的进展和突破。

本文得到上海生化所李其梁和我系王贤舜、徐询老师的指导，特此致谢。

参考文献

[1] Price: F. W.: *Basic Molecular Biology*, **5**, 1979.
[2] GINA BARI KOLATA.: *Science*, **192**, 645, 1976.
[3] Jon, W. M. et al.: *Nature*, **237**, 82, 1972.
[4] Piszkiewicz, D. et al.: *Proc. Nat. Acad. Sci. USA.*, **76**, 1589, 1979.
[5] Tzagoloff, A. et al.: *Ann. Rev. Biochem.*, **48**, 419, 1979.
[6] 冯佑民、鲁子贤等：《生物化学与生物物理进展》，1979年，第5期，第1页。
[7] Noyes, B. E. et al.: *Proc. Nat. Acad. Sci. USA.*, **76**, 1770, 1979.
[8] Barrell, B. et al.: *New Scientist*, **78**, 21, 1978.
[9] Shigetada Nakanishi: *Nature*, **278**, 423, 1979.
[10] Roger Lewin.: *New Scientist*, **10**, 452, 1979.
[11] 東野一弥：《蛋白质·核酸·酵素》，**23**，197，1978.

[本文于1980年10月30日收到]