

蛋白质怎样识别 DNA ——几个 DNA 结合蛋白结构研究的启示

王家槐

(中国科学院生物物理研究所, 北京)

一、引言

蛋白质与核酸的相互作用是分子生物学的中心课题之一。核酸是生命信息的载体。在所有生命体系中, 遗传信息的表达又是在时间和空间上高度精确地得到调节和控制的。以人类为例, 体内各类细胞内 DNA 所包含的基因是一样的, 都是 10 万个左右。不同细胞或同一细胞在不同的时间、不同的环境下只有很少一部分有关的基因获得表达, 致使脑细胞、肝细胞、肌肉细胞等等得以各不相同, 执行各个不同的功能。因此基因表达的调控对于了解分化、发育、老化、包括癌变在内的许多病变及机体对环境因子的协调反应等机理是一个关键的问题。从目前研究较多的原核细胞的情况看来, 基因调节的主要环节就是一些特定的蛋白质与核酸的专一性相互作用, 尤其是那些调控蛋白如何识别 DNA 特定部位上的碱基排列结构。事实上蛋白质与核酸的相互作用在遗传过程的复制、转录、转译的几乎每一个环节以及其它许多生命过程中都极其重要, 本文主要从近两年来几个与 DNA 有专一性结合的蛋白质结构的研究成果来讨论蛋白质是怎么识别 DNA 的。这些蛋白是 λ -噬菌体中的一对阻遏蛋白 Cro 和 cI 以及大肠杆菌中的异化作用基因激活蛋白 (Catabolite gene activator protein, 简称 CAP)。重点则是讨论两个 λ -阻遏蛋白。

二、分子遗传学背景

λ -噬菌体属于温和型噬菌体。它以两种方

式感染大肠杆菌。一是溶菌方式。一个小时内, 噬菌体在大肠杆菌内繁殖, 大约产生 100 个后代, 随后大肠杆菌死亡。二是溶原方式。噬菌体的 DNA 通过基因重组, 插入细菌 DNA, 处于休止状态, 其包含的基因不被表达, 大肠杆菌得以存活。在细菌繁殖时, 噬菌体 DNA 也一起被复制遗传到下一代细菌的染色体里。这种带有噬菌体的细菌叫溶原菌。在适当条件下 (如紫外光照等致癌因子的作用), 半小时内, 插入细菌 DNA 内的 λ -DNA 即可以分离出来, 其基因也重新活跃起来, 噬菌体生长转入溶菌相。什么因素决定 λ -DNA 上的基因表达与否? 这是个很有趣的问题。早在十几年前,

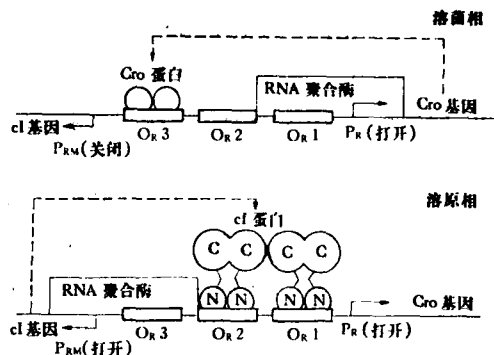


图 1 λ -噬菌体内基因调控机构示意图^[1]

在溶菌相, Cro 蛋白以二聚体形式优先结合操纵基因 O_R3 , 使启动子 P_{RM} 位上不能再结合 RNA 聚合酶, cI 基因不能表达, 而 RNA 聚合酶可结合到启动子 P_R 上, 右向基因得到表达, 噬菌体得以繁殖。在溶原相, cI 蛋白二聚体 (以其氨基端 "N") 优先结合到 O_R1 。由于协同作用 (羧基端 "C" 间的相互作用), 另一个二聚体结合到 O_R2 位, 从而使 RNA 聚合酶不能结合到 P_R , 右向转录关闭, 噬菌体处于休眠态。但由于 RNA 聚合酶与结合在 O_R2 上的 cI 蛋白相互作用, 可促进左向的 cI 基因的表达, 维持溶原相。

分子遗传学家就选中这个体系来研究基因的调控机理。

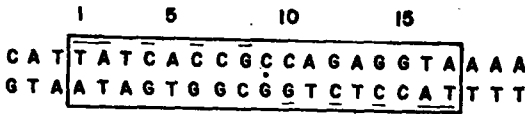


图2 操纵基因 O_{R1} 的碱基排列顺序

长方框内 17 个碱基是这个操纵基因。注意整个操纵基因以第九对碱基为中心,具有近似二重旋转对称性,横线划出部分为符合对称性的碱基对。

原来, λ -DNA 包含约 66 个基因,有 48502 个碱基对 [1A],其中只有百来个碱基对长的区域是起决定性控制作用的。图 1 是这个控制机构的关键部位的示意图。图中 O_{R1} , O_{R2} , O_{R3} 是三个操纵基因,都是 17 个碱基对长的 DNA 片段,其碱基序列大致相似,又不全相同。相邻操纵基因之间有几个碱基对的间隔。每个操纵基因有近似的二重旋转对称性(图 2)。图 1 中与操纵基因相邻的,是两个调节基因,右边一个叫 Cro , 左边一个叫 cI 。它们分别模制产生两个起调节作用的阻遏蛋白。这两个蛋白都能与三个操纵基因发生高度专一的结合,但结合的亲和力有所不同。 cI 阻遏蛋白的亲和力依次为 $O_{R1} \gg O_{R2} = O_{R3}$, 而 Cro 阻遏蛋白的亲和力则为 $O_{R3} \gg O_{R2} = O_{R1}$, 与 cI 恰好相反。如果细胞内 Cro 蛋白的量相对的多,它优先结合到 O_{R3} , 其后果是, RNA 聚合酶可结合到位于操纵基因 O_{R1} 右边的启动子 P_R , DNA 的右向转录得以进行,繁殖噬菌体所必需的大量基因得以表达,其中包括产生 Cro 蛋白本身的基因。这时噬菌体就处于溶菌相。在溶菌相,只有一个位于操纵基因 O_{R3} 左边的 cI 基因被关闭。因为 Cro 蛋白在 O_{R3} 上的存在,妨碍了 RNA 聚合酶结合到启动子 P_{RM} 上,所以 DNA 的左向转录无法进行。相反,如果由于某种原因, cI 蛋白的量相对的多,它们将优先结合到 O_{R1} 。由于协同作用,在 O_{R1} 上的结合,也促进了 cI 蛋白对 O_{R2} 的结合。 cI 蛋白在 O_{R1} 和 O_{R2} 上的结合阻止了 RNA 聚合酶结合到 P_R , 其后果是,噬菌体 DNA 的绝大部分基因被关闭,处于休止态。同时,由于 cI 蛋白与 RNA 聚

合酶之间的相互作用, cI 对 O_{R1} 和 O_{R2} 的结合非但不妨碍,反而促进 RNA 聚合酶在 P_{RM} 位上的结合,导致 cI 基因的活跃,反过来产生 cI 蛋白,来维持这种溶原相 [1B]。

总之, cI 与 Cro 两个阻遏蛋白形成了一对分子开关^[2]。 cI 多时,绝大部分基因,包括 Cro 被关闭,只有 cI 本身的基因是活跃的,噬菌体处于溶原相; Cro 多时则相反,绝大部分基因包括 Cro 本身得到表达,只有 cI 基因被关闭,噬菌体处于溶菌相。而整个调控机构的决定性环节,在于这两个蛋白质与三个相似又不相同的操纵基因之间的竞争性的蛋白-DNA 相互作用。

CAP 蛋白又叫 cAMP 受体蛋白。它在大肠杆菌中可促进几个与异化作用有关的操纵子的基因表达。以著名的乳糖操纵子为例,当细胞内作为第二信使的 cAMP 浓度足够高时,它就与 CAP 蛋白结合。cAMP 会引起 CAP 蛋白构象发生显著变化。然后, cAMP-CAP 复合物(而不是单独的 CAP)能与乳糖操纵子的启动子前的一段 DNA 发生专一性结合,促进 RNA 聚合酶对该启动子的结合,导致相应结构基因的表达^[3]。在这里, CAP 与大约 21 个碱基对那么长的 DNA 片段结合。

三、调控蛋白的三维结构及其与 DNA 的结合

上述三个蛋白质的三维结构已用 X-射线晶体学方法获得测定。 Cro 是由 66 个残基组成的碱性蛋白,整个结构只有一个结构域,大体上由三段 α -螺旋和三段 β -链组成(图 3)^[4]。 cI 蛋白由 236 个残基组成,有两个结构域。氮端域包括残基 1—92,是与 DNA 相结合的结构域;羧端域包括残基 132—236,是主要负责寡聚化的结构域,中间是联接肽^[5]。目前已测得三维结构的仅是氮端域部分。它包括一个伸展的前臂部,五段 α -螺旋及一些无规的联接片段(图 4)^[6]。CAP 蛋白由 209 个残基组成,也有两个结构域。氮端域较大,结合变构因子 cAMP,其主体是两层 β -折叠层围成的一个桶状结构和

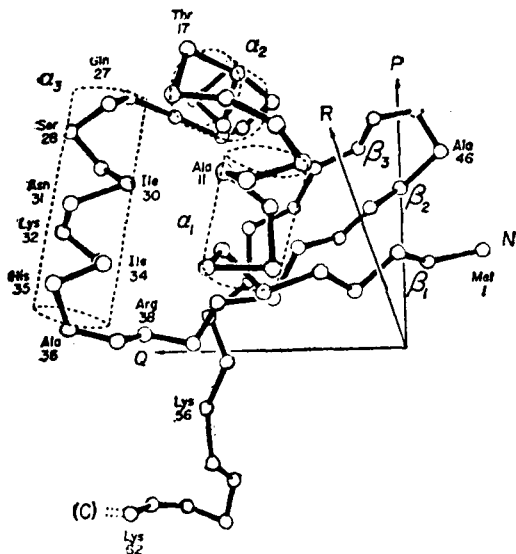


图3 Cro蛋白的三维结构^[4]

图中示出分子的主链骨架结构。分子羧端从63—66位残基在晶体内处于无序状态,未能确定其位置。

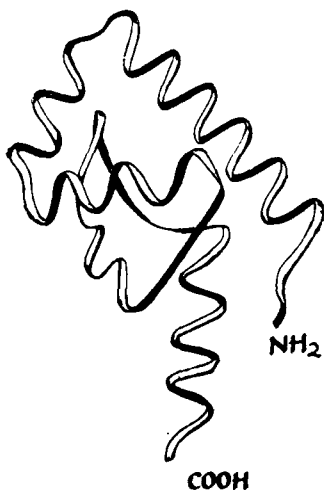


图4 cI蛋白氨端域的三维结构^[4]

分子氨端前三个残基处于无序状态

一长两短共三段 α -螺旋。羧端域较小,是与DNA相结合的部位,包括三段 α -螺旋和由四段较短的 β -链组成的折叠层(图5)^[7]。

有了这三个蛋白(或其片段)的三维结构,我们能不能获得一点关于它们与DNA作用的信息呢?这是人们普遍关心的一个问题。其答案是肯定的。

1. 调节蛋白与DNA的复合物具有二重对称性 已知这三个蛋白都以二聚体形式与DNA相结合;而由图2,我们也已看到,操纵基

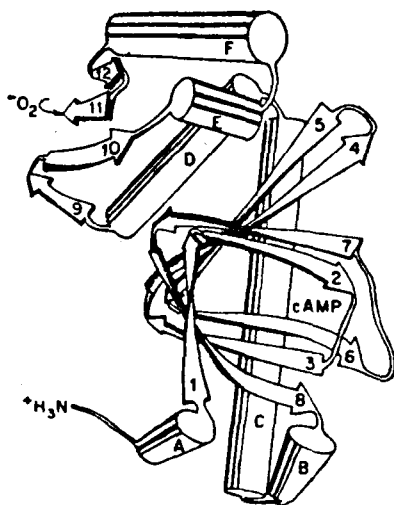


图5 CAP蛋白分子的三维结构

氨端域较大,是cAMP结合域;羧端域较小,是DNA结合域。 α -螺旋E和F可能参与DNA的结合^[7]。

因上的碱基排列具有近似二重对称。这启示我们,蛋白质二聚体两个分子应该分别与操纵基因对称相关的两半相结合,形成的复合物从而也具有近似的二重对称性。这种结合特性很可能在其它许多蛋白-DNA复合物中普遍存在。这可扩大蛋白与DNA的结合面,从而提高专一性。不仅如此,至少对于cI和Cro,分子的电荷分布是不均匀的。蛋白质分子的偶极向量在形成二聚体时,其合成向量将躺在二重对称轴上,这对于蛋白质与带负电场的DNA的相互作用是有利的^[8]。关于这一点,后面还要进一步讨论(图6)。

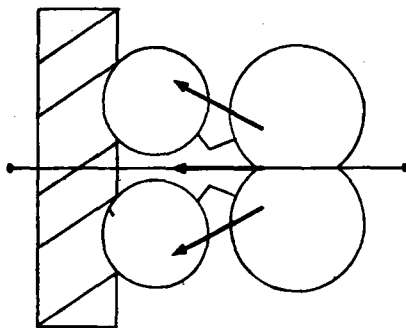


图6 调节蛋白与DNA复合物的对称性示意图

蛋白分子的偶极向量在形成蛋白质二聚体时的合向量躺在二重轴上,正指DNA浅沟^[7]。

2. 螺旋-转折-螺旋超二级结构是蛋白质与DNA结合的一种折叠单位 生物化学和遗传

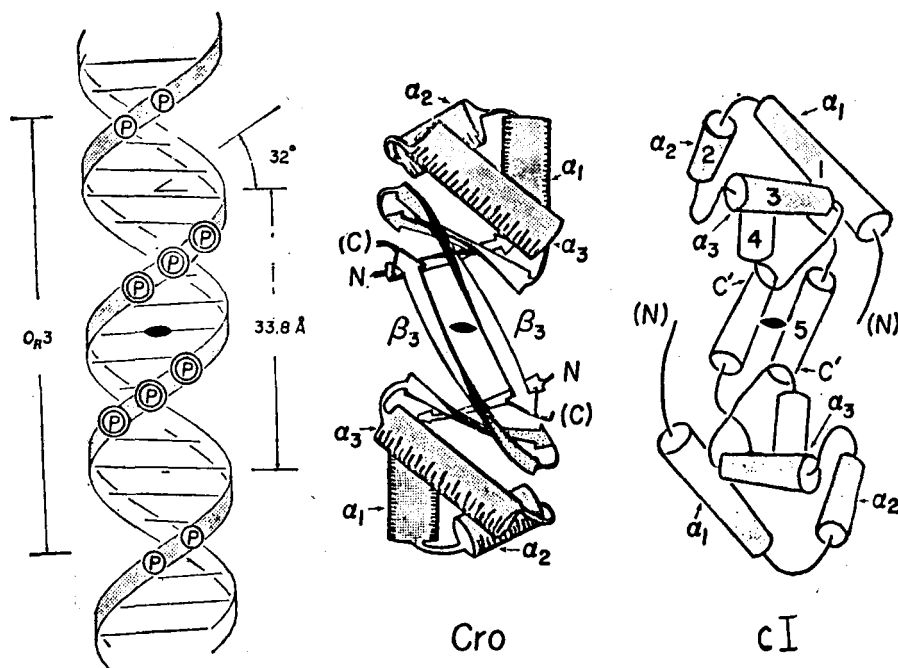


图 7 (A) Cro 蛋白和 cI 蛋白与 DNA 的结合^[10]

图中●代表二重旋转轴。可以看出,两个蛋白的螺旋 2 和螺旋 3 以及螺旋 1 的一部分具有相似的结构。二聚体两个分子的两段螺旋 3 相间 34 Å,可以嵌入 DNA 相邻的深沟中。DNA 分子上的⊙代表当蛋白质与 DNA 结合时,与蛋白质接触的磷酸根。把蛋白质分子翻转 180°,覆盖到 DNA 分子上,使两者的二重轴吻合,就可得到复合物。

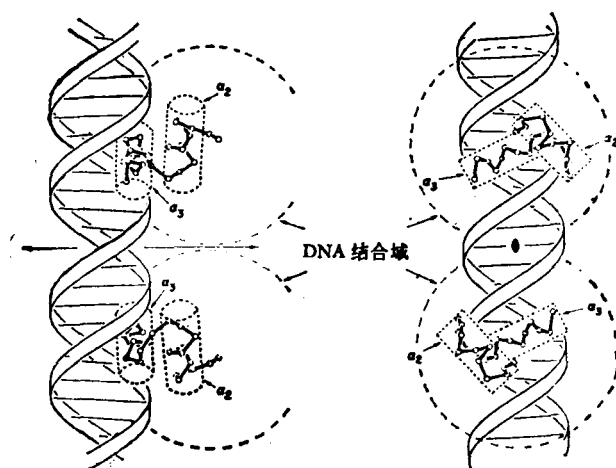


图 7 (B) 螺旋-转折-螺旋作为一个与 DNA 结合的超二级结构单位与 DNA 的结合模型^[11]

学证据指出^[1,9],这三个蛋白与 DNA 的结合,主要在 DNA 分子的同一侧,作用部位在相邻的两个深沟。DNA 分子并不因此发生重大变化,而保持右手 B-DNA 构象。结构比较表明^[10],Cro 和 cI 氨端域的螺旋 2-转折-螺旋 3 的主链构象(包括 24 个连续的 α-碳原子)基本

相同。在二聚体中,分属两个分子的两段 α-螺旋 3 从分子表面突出,间隔 34 Å,正好与 DNA 相邻两个深沟的距离一致。模型研究告诉我们,当蛋白二聚体与 DNA 结合时,这两段螺旋 3 可嵌入 DNA 的相邻深沟中,使氨基酸侧链与核苷酸碱基对的外露部分发生专一性相互作用。

用；而两段螺旋 2 则可与 DNA 的磷酸根发生非特异性相互作用(图 7 A、B)^[4,5]。这种螺旋-转折-螺旋超二级结构单位也在 CAP 的 DNA 结合域中存在^[11]。只是对于 CAP，目前还没有一个令人满意的与 DNA 结合的模式。更广泛的氨基酸序列比较研究以及立体化学研究表明^[12]，这种结构单位至少可能在除 λ 外另两种温和型噬菌体的四个基因调控蛋白中存在。也就是说，在普遍意义上，这种特定的螺旋-转折-螺旋结构是蛋白质在进化过程中获得的一种适于与 DNA 相结合的折叠单位。这是一个相当重要的发现。

3. 蛋白分子的一些柔性部位参与了与 DNA 的结合 Cro 蛋白的羧端和 cI 蛋白的氮端都有几个氨基酸残基在晶体结构中处于无序状态，观察不到(图 4、5)。生物化学和模型研究表明，cI 与操纵基因形成复合物时，除了螺旋-转折-螺旋折叠单位与 DNA 正面接触外，它的氮端像一个伸展的臂部，可顺深沟抱向 DNA 分子的背面(图 8)。如果把氮端头部三个残基除去，这种作用不复存在^[13]。这个独特的柔性“臂部”的参与，使整个 cI 蛋白与操纵基因的结合更牢，专一性更强。Cro 蛋白的羧端也很可能有同样的作用。

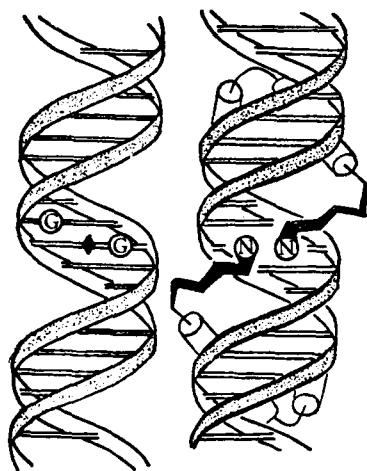


图 8 cI 蛋白与 DNA 结合时，其氮末端部沿 DNA 深沟抱向背部，使结合更牢，专一性更强^[13]

4. DNA 的深沟是更有利的识别部位
任何特异性识别当然取决于 DNA 的碱基排

列。正如下文将要指出的，侧链碱基在相互配对之后余下的一些极性基团可与蛋白质相应基团形成氢键，这是特异性识别的主要基础。这样的极性基团在深沟和浅沟都有，然而上述调控蛋白的研究表明，相互作用发生在深沟。可以认为作为蛋白质识别的位点，深沟优于浅沟，很可能有一定的普遍意义。这不仅因为深沟比浅沟宽一倍，有利于蛋白质与之相互作用，更重要的是，深沟比浅沟有更特异的潜在的形成氢键的能力。A. Rich 等人曾从立体化学角度初步探讨过这个问题^[14]。图 9 给出四种不同类型的突变替换： $G \cdot C \rightarrow A \cdot T$ ， $G \cdot C \rightarrow C \cdot G$ ， $G \cdot C \rightarrow T \cdot A$ 和 $A \cdot T \rightarrow T \cdot A$ （注意，从结构化学角度看， $T \cdot A \rightarrow C \cdot G$ 只是 $G \cdot C \rightarrow A \cdot T$ 的反转，属于同一类型，如此等等，所以一共才四种类型）。显而易见，在浅沟，三个潜在的氢键的位点中有两个对所有四种类型的突变替换都是保守的，它们都是氢键的受体；而在深沟，任何一类突变替换所引起的氢键形成方式的变化却要大得多。换句话说，从立体化学角度看，深沟可能提供更好的特异性识别，是与蛋白质发生专一性作用的更好部位。

四、蛋白质与 DNA 专一性识别的本质

专一性识别的问题对分子生物学家来讲，是一个极为重要的问题。在核酸与核酸的相互作用中，碱基之间按 Watson-Crick 方式正确配对是决定性的，这导致了遗传密码的发现。曾经有一种观点，认为在蛋白质与核酸的专一性作用中，有“第二密码”。事实上，对于 cI 和 Cro 的研究，使我们相信，那种“第二密码”是不大可能存在的。的确，模型研究显示，无论是 cI^[15] 或 Cro^[16]，在蛋白质二聚体与操纵基因之间，至少可形成十对左右氢键。它们是这些蛋白专一地识别 DNA 的分子基础。但是，三个操纵基因的碱基组成并不完全一样；更重要的是，两个蛋白质与 DNA 的结合面的氨基酸侧链分布相当不同。在这种情况下，一个固定的“第二密码”这样一种模式是不可想像的。

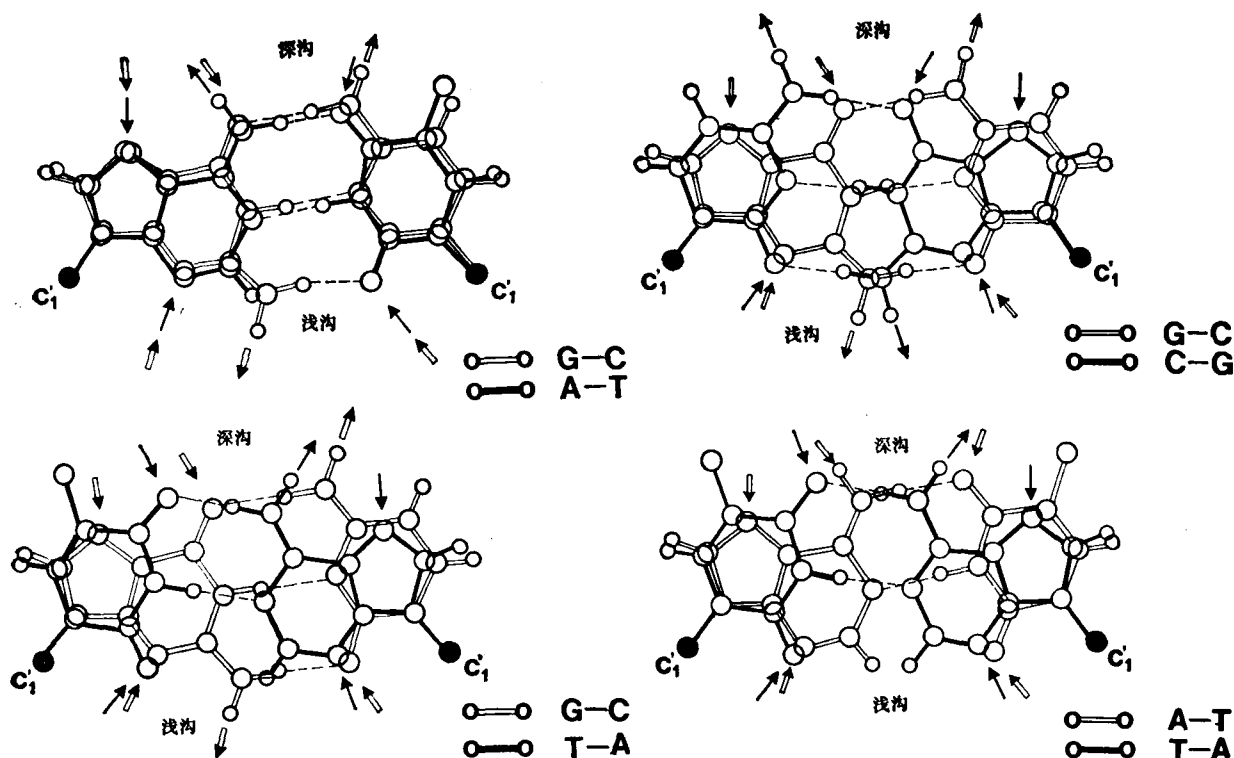


图9 四种不同类型的突变替换

最小圆代表氢原子,其余圆代表氧、氮、甲基和碳原子。虚线代表碱基配对的氢键。指向原子的箭头代表氢键受体,背离原子的箭头代表氢键供体。对所有这四种突变替换,浅沟的可能的氢键形成方式变化不大,而深沟的却变化较大。也就是说,作为蛋白质识别位点,深沟比浅沟的特异性更强。

我们知道,DNA 是极性很强的大分子。它的骨架是磷酸-戊糖链,布满带负电的和亲水性基团。它的侧链碱基的芳香环在形成双螺旋分子时,其疏水平面相互堆积覆盖,真正暴露在外并能与蛋白质相作用的,多数是环外侧一些极性基团,只有胸腺嘧啶有一个甲基外露在深沟,是唯一突出的疏水基团。不难推论,DNA 与蛋白质的相互作用必然以极性相互作用为主体。这是与蛋白-蛋白相互作用大相迥异的。有许多证据指出,蛋白-蛋白相互作用中主要的是疏水作用。经过仔细研究的例子之一,是胰蛋白酶与其抑制剂之间的相互作用^[17],诚然,氢键对于蛋白质相互间专一性识别也是很重要的。但是,一个立体分布的特定氢键体系对于蛋白质专一性识别 DNA 显然起着占主导地位的作用,这一点是 Cro 和 cI 结构研究所提供的一个重要信息(图 10)^[15,16]。前面提到,这两个蛋白

都能与三个操纵基因发生专一性结合,但亲和性又不相同。事实上,它们的确可能有共同的氢键对,如 Cro 的 Gln 27 及 cI 的 Gln 44 都可与操纵基因第二位以及对称相关的第十六位上的腺嘌呤形成两对氢键,而这两个位置上的腺嘌呤在所有操纵基因上都存在,等等。同时,它们又有不同的相互作用。正是这种竞争性相互作用构成了它们成为分子开关的基础。

当我们指出,一个特定的立体分布的氢键体系是蛋白质专一性识别 DNA 特殊位点的立体化学基础时,还应该明确,它并不是蛋白质与 DNA 形成复合物时结合能的唯一来源。事实上,无论从热力学或动力学角度看,真正促进并稳定 DNA 与这些调节蛋白之间形成复合物的,是 DNA 上的负电基团与蛋白质结合面上碱性氨基酸侧链的正电基团间的静电力。这是一种非特异性相互作用。一个可能的模式是,主

构,直接观察其间的立体化学关系。现在, **CrO** 和 **cI** 与有特定碱基序列的 DNA 片段的共晶已经获得^[15,9]。**cI** 与操纵基因片段共晶的晶体学研究初步证实了,已经提出的复合物模型是正确的^[20]。DNA 与蛋白质相互作用的直观图景将很快展示在分子生物学家的前面。

- [1] Sanger, F., et al.: *J. Mol. Biol.*, **162**, 162, 1982.
(B) Ptashne, M. et al.: *Cell*, **19**, 1, 1980.
- [2] Johnson, A. D., et al.: *Nature*, **294**, 217, 1981.
- [3] Zubay, G. et al.: *Proc. Natl. Acad. Sci. USA*, **66** 104, 1970.
- [4] Anderson, W. F. et al: *Nature*. **290**, 654, 1981.
- [5] Pabo, C. O. et al.: *Proc. Natl. Acad. Sci. USA*, **76**, 1608, 1979.
- [6] Pabo, C. O. et al.: *Nature* **298**, 443, 1982.
- [7] Mekey, D. B. et al.: *Nature*, **290**, 744, 1981.
- [8] 王家槐, Lewis, M., «中国科学»待发表(1984)。
- [9] Simpson, R. B., *Nucleic Acids Res.*, **8** 759, 1980.
- [10] Anderson, W. F., et al., *J. Mol. Biol.*, (1983) 在印刷中。
- [11] Steitz, T. A. et al., *Proc. Natl. Acad. Sci. USA*, **79**, 3097, 1982.
- [12] Ohlendorf, D. H. et al.: *J. Mol. Evol.*, **19**, 109, 1983.
- [13] Pabo, C. O. et al.: *Nature*, **298**, 441, 1982.
- [14] Sceman, N. C. et al.: *Proc. Natl. Acad. Sci. USA*, **73**, 804, 1979.
- [15] Lewis, M., et al.: *Cold Spring Harbor Symp. Quant. Biol.*, **43**, 435, 1983.
- [16] Ohlendorf, D. H., et al. *Nature*, **298**, 718, 1982.
- [17] Huber, R. *Protein-Protein Interaction* Edited by Jaenicke d Helmreich, 111. 1972.
- [18] Berg, O. G. et al.: *Trends in Biochem. Scien.* **7**, 52, 1982.
- [19] Matthews, B. W., *Cold Spring Harbor Symp. Quant. Biol.*, 1983, 在印刷中。
- [20] Wang J-H. et al.: "The Crystallographic Studies on the Co-crystal Containing the N-terminal Domain of λ -Repressor and the 11 Base-pair Synthetic DNA Fragment", The Report in the Eighth Course of International School of Crystallography June 1982, in Erice, Italy.

著优于非特异性结合，一对氢键就可提高结合常数达七倍。这时，识别过程终于完成。至于 λ -噬菌体一对阻遏蛋白与三个操纵基因之间的竞争性结合，也必定是这种特异性契合在程度上的差别所致。要彻底弄清这一点，最好的途径就是研究阻遏蛋白与操纵基因共生晶体的结

• 8 •