

# 1

## 生命的蓝图

### 基因和基因组

生命是存储并加工信息的复杂系统。生命信息不仅在细胞之间世代相传，同时在生物个体发生的过程中，信息也能够在细胞内得以表达。在人类历史发展的早期，人类对遗传现象，也就是父代与子代之间的信息传送，一直不明其理。直到 19 世纪中期，Gregor Mendel 的豌豆实验才打开了探索遗传机理的大门；遗憾的是，Mendel 实验的重要意义直到 1900 年才被人们重新发现。遗传单位，后来被称为基因，它决定了生物个体在某一方面的特性。Mendelian 遗传定律描述了两个基本的遗传规律：首先，作为信息流的纵向传递，基因的遗传是相对独立的；其次，作为信息的横向表达，由于存在多个等位基因的相互竞争，生物个体的外在表现型并不一定能反映它的内在基因型。

生物学是研究生命现象的科学，其目的是为了了解生物体的结构与功能。19 世纪下半叶，不仅遗传学（分析遗传功能的生物学分支）取得了巨大的成就，而且由于显微镜的发明，细胞生物学的发展也突飞猛进。最重要的

2 后基因组信息学

是，人们发现遗传物质就在细胞核内的染色体上。染色体上的全套基因，或者更准确地说是单倍体染色体上的全套基因，被称为基因组。从生命信息流的角度看，基因组是信息传递的结构和功能单位，而基因则是信息表达的结构和功能单位（表 1.1）。

表 1.1 基因组和基因

| 物 质 | 定 义    | 分子机理               |
|-----|--------|--------------------|
| 基因组 | 信息传递单位 | DNA 复制             |
| 基因  | 信息表达单位 | 转录为 RNA；<br>翻译成蛋白质 |

19 世纪发展起来的遗传学和细胞生物学规律是现代生物学尤其是分子生物学的基础。20 世纪后半叶则是分子生物学发展的黄金时期。在分子生物学领域，实验技术的确立和迅猛发展为基因和基因组的研究建立了分子基础，揭示了各种细胞过程中生物大分子之间的结构与功能的关系，并最终提出了人类基因组计划。人类基因组计划的目的就是为了解读生命的蓝图。接下来让我们快速浏览一下这些发展。

DNA 和蛋白质

染色体是由脱氧核糖核酸（DNAs）和蛋白质组成的大分子复合物。最初，人们猜想蛋白质可能是遗传物质；但是到了 20 世纪中期，研究表明 DNA 包含遗传信息，而蛋白质则是在细胞内合成的。1953 年 James Watson 和 Francis Crick 提出了 DNA 结构的双螺旋模型。这一结构模型是从 DNA 纤维的 X-ray 衍射数据中推导出来的，而这些衍射数据由 Rosalind Franklin 和 Maurice Wilkins 得到。同时实验发现，在任何 DNA 组成中，腺嘌呤（A）和胸腺嘧啶（T）的总和都等于鸟嘌呤（G）和胞嘧啶（C）的总和<sup>\*</sup>。实际上，每个 DNA 分子包含两条链，每条链都是由 A、T、G 和 C 4 种核苷酸重复组成的线

译者注：原文可能有误，应为嘌呤（A+G）的总和都等于嘧啶（T+C）的总和。

性多聚体(表 1.2 和表 1.3)。A 与 T, 以及 G 与 C 结构互补, 而且中间有氢键连接, 因而两条链缠绕形成的双螺旋结构很稳定(图 1.1)。遗传信息由核苷酸序列编码, 一旦给定某一条链上的核苷酸序列, 那么根据互补规则, 另一条序列上的核苷酸序列也会自动确定因此, DNA 双螺旋模型对遗传的分子机理产生了深远的影响。DNA 分子的复制(表 1.1)揭示了遗传信息的传递, 而互补的双链则确保了信息复制的正确性。

表 1.2 核苷酸和蛋白质

| 大分子 | 主 链 |       | 重复单元                                                     | 长 度                                  | 角 色        |
|-----|-----|-------|----------------------------------------------------------|--------------------------------------|------------|
| 核 酸 | DNA | 磷酸二酯键 | 脱氧核糖核苷酸<br>(A,C,G,T)                                     | $10^3 \sim 10^8$                     | 基因组        |
|     | RNA | 磷酸二酯键 | 核糖核苷酸<br>(A,C,G,U)                                       | $10^3 \sim 10^5$                     | 基因组        |
|     |     |       |                                                          | $10^3 \sim 10^4$<br>$10^2 \sim 10^3$ | 信使<br>基因产物 |
| 蛋白质 |     | 肽键    | 氨基酸<br>(A,C,D,E,F,G,H,I,<br>K,L,M,N,P,Q,R,S,<br>T,V,W,Y) | $10^2 \sim 10^3$                     | 基因产物       |

表 1.3 核苷酸代码

|   |           |   |                  |
|---|-----------|---|------------------|
| A | 腺嘌呤       | W | 弱键(A 或 T)        |
| G | 鸟嘌呤       | S | 强键(G 或 C)        |
| C | 胞嘧啶       | M | 氨基(A 或 C)        |
| T | 胸腺嘧啶      | K | 酮类(G 或 T)        |
| U | 尿嘧啶       | B | 非腺嘌呤(G 或 C 或 T)  |
| R | 嘌呤(A 或 G) | H | 非尿嘌呤(A 或 C 或 T)  |
| Y | 嘧啶(C 或 T) | D | 非胞嘧啶(A 或 G 或 T)  |
| N | 任意核苷      | V | 非胸腺嘧啶(A 或 G 或 C) |

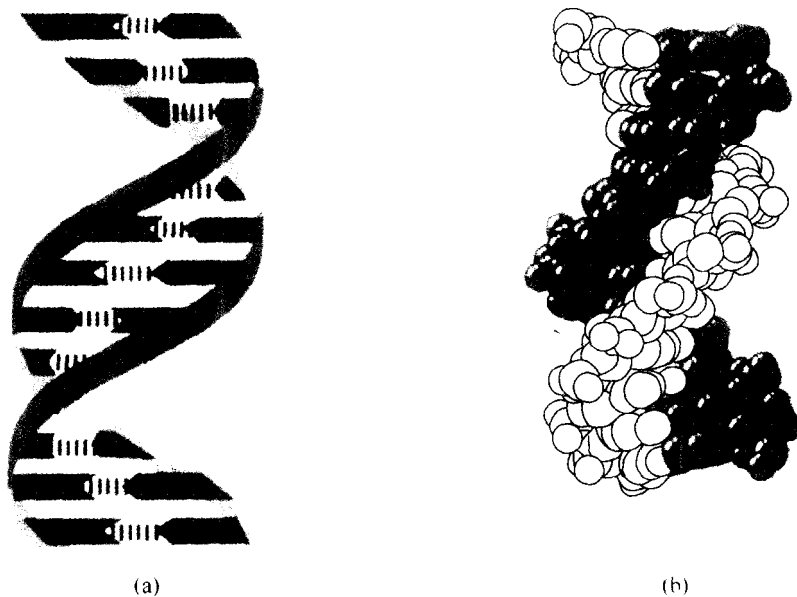


图 1.1 DNA 双螺旋  
(a) Watson-Crick 模型示意图；(b) 一段合成 DNA 的三维结构，  
序列为“CGCGAATTCGCG”(PDB:7BNA)

蛋白质是由 20 种氨基酸组成的线性多聚体，氨基酸之间由肽键相连（表 1.2 和表 1.4）。由多肽链折叠形成的蛋白质三维结构远比 DNA 双螺旋结构复杂。氨基酸序列又称为蛋白质的一级结构；1953 年，Frederick Sanger 第一次测定了胰岛素的氨基酸序列。蛋白质三维结构又称为三级结构；1960 年，John Kendrew 用蛋白质晶体 X-ray 衍射方法，第一次测定了肌血球素的三维结构。Max Perutz 改善了 X-ray 衍射方法，创造出著名的同晶置换法。如图 1.2 所示，肌血球素由 6 个  $\alpha$  螺旋组合而成。 $\alpha$  螺旋最初由 Linus Pauling 于 1951 年提出，本来是用来解释单个氨基酸的同聚物的；Linus Pauling 还预言了蛋白质中  $\alpha$  螺旋的存在。越来越多的蛋白质三维结构被解析出来，同时也发现了一个蛋白质功能的基本原理：就像 DNA 双螺旋的互补结构一样，正是由于蛋白酶和底物结构上的互补使得蛋白酶能够成功地识别底物并对其发挥作用。蛋白酶与底物结构的互补关系就像锁与钥匙的配对关系，不过两者之间的相似性并不严格，因为分子结构相当灵活

并且可以被诱导。正是两个分子之间结构骨架的互补最终决定了两者的相互作用。

表 1.4 氨基酸代码

|     |   |            |
|-----|---|------------|
| Ala | A | 丙氨酸        |
| Arg | R | 精氨酸        |
| Asn | N | 天冬酰胺酸      |
| Asp | D | 天冬氨酸       |
| Cys | C | 半胱氨酸       |
| Gln | Q | 谷氨酰胺酸      |
| Glu | E | 谷氨酸        |
| Gly | G | 甘氨酸        |
| His | H | 组氨酸        |
| Ile | I | 异亮氨酸       |
| Leu | L | 亮氨酸        |
| Lys | K | 赖氨酸        |
| Met | M | 甲硫氨酸       |
| Phe | F | 苯丙氨酸       |
| Pro | P | 脯氨酸        |
| Ser | S | 丝氨酸        |
| Thr | T | 苏氨酸        |
| Trp | W | 色氨酸        |
| Tyr | Y | 酪氨酸        |
| Val | V | 缬氨酸        |
| Asx | B | 天冬氨酸或天冬酰胺酸 |
| Glx | Z | 谷氨酸或谷氨酰胺酸  |
| Sec | U | 硒代半胱氨酸     |
| Unk | X | 未知氨基酸      |

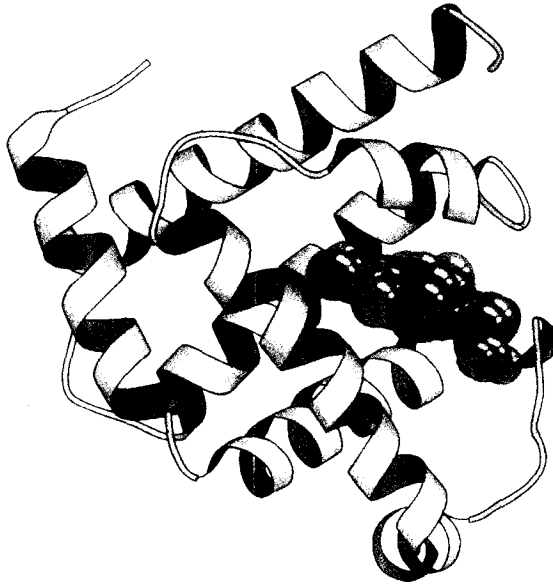


图 1.2 抹香鲸肌血球素的三维结构(PDB:1MBN)

## 中心法则

DNA 和蛋白质是活细胞中发挥关键作用的生物大分子；它们都是由重复单元组成的线性多聚体（表 1.2）。4 种核苷酸组成的 DNA 序列编码并存储了遗传信息，而 DNA 复制则是遗传信息传递的分子机理。与此相比较，在细胞中真正执行功能的是蛋白质因此，需要有一种机制，使得 DNA 序列能够将 4 种核苷酸的编码信息转化为 20 种氨基酸编码的蛋白质序列。这是遗传信息得以表达的至关重要的一步。这一步转化的分子机理就是翻译。在翻译过程中，每 3 个核苷酸编码（称为一个密码子）被翻译成为一个氨基酸。表 1.5 列出了 64 种密码子翻译成为 20 种氨基酸的遗传密码表。最初，人们以为所有的物种都通用这张遗传密码表，但是现在已经发现了许多不同的地方，表 1.6 列出了这些不同之处。

DNA 信息并不是直接转化为蛋白质信息的。首先，DNA 被转录为 RNA（核糖核酸）；然后，RNA 被翻译成蛋白质。为了同其他类型的 RNA

(如转运 RNA 即 tRNA 和核糖体 RNA 即 rRNA) 相区别, 这一类 RNA 被称为信使 RNA 即 mRNA。RNA 也是一种线性大分子, 同 DNA 十分相像; 二者的区别仅在于 RNA 的骨架糖链由核糖而不是脱氧核糖组成, 另外, 尿嘧啶(U) 取代了胸腺嘧啶(T) 表 1.2 和表 1.3)。从序列符号上看, 转录只是把 DNA 序列上的 T 转化成 U 就得到了 RNA 序列。因此, 从 DNA 到 RNA 然后再到蛋白质的信息表达过程是一个单向信息流。这个单向信息流同 DNA 之间的信息传递一起构成了分子生物学的中心法则 (图 1.3)。1958 年, Francis Crick 首先提出了中心法则。

表 1.5 标准遗传密码表

| 第一位 | 第二位 |            |     |     |     |     |     |     | 第三位 |
|-----|-----|------------|-----|-----|-----|-----|-----|-----|-----|
|     | U   |            | C   |     | A   |     | G   |     |     |
| U   | UUU | Phe        | UCU | Ser | UAU | Tyr | UGU | Cys | U   |
|     | UUC | Phe        | UCC | Ser | UAC | Tyr | UGC | Cys | C   |
|     | UUA | Leu        | UCA | Ser | UAA | 终止  | UGA | 终止  | A   |
|     | UUG | Leu        | UCG | Ser | UAG | 终止  | UGG | Trp | G   |
| C   | CUU | Leu        | CCU | Pro | CAU | His | CGU | Arg | U   |
|     | CUC | Leu        | CCC | Pro | CAC | His | CGC | Arg | C   |
|     | CUA | Leu        | CCA | Pro | CAA | Gln | CGA | Arg | A   |
|     | CUG | Leu        | CCG | Pro | CAG | Gln | CGG | Arg | G   |
| A   | AUU | Ile        | ACU | Thr | AAU | Asn | AGU | Ser | U   |
|     | AUC | Ile        | ACC | Thr | AAC | Asn | AGC | Ser | C   |
|     | AUA | Ile        | ACA | Thr | AAA | Lys | AGA | Arg | A   |
|     | AUG | Met,<br>起始 | ACG | Thr | AAG | Lys | AGG | Arg | G   |
| G   | GUU | Val        | GCU | Ala | GAU | Asp | GGU | Gly | U   |
|     | GUC | Val        | GCC | Ala | GAC | Asp | GGC | Gly | C   |
|     | GUA | Val        | GCA | Ala | GAA | Glu | GGA | Gly | A   |
|     | GUG | Val        | GCG | Ala | GAG | Glu | GGG | Gly | G   |

表 1.6 其他的遗传密码

|     | T1  | T2  | T3  | T4  | T5  | T6  | T9  | T10 | T12 | T13 | T14 | T15 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| CUU | Leu | —   | Thr | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| CUC | Leu | —   | Thr | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| CUA | Leu | —   | Thr | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| CUG | Leu | —   | Thr | —   | —   | —   | —   | —   | Ser | —   | —   | —   |
| AUU | Ile | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| AUC | Ile | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| AUA | Ile | Met | Met | —   | Met | —   | —   | —   | —   | Met | —   | —   |
| AUG | Met | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| UAU | Tyr | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| UAC | Tyr | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| UAA | 终止  | —   | —   | —   | —   | Gln | —   | —   | —   | —   | Tyr | —   |
| UAG | 终止  | —   | —   | —   | —   | Gln | —   | —   | —   | —   | —   | Gln |
| AAU | Asn | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| AAC | Asn | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| AAA | Lys | —   | —   | —   | —   | —   | Asn | —   | —   | —   | Asn | —   |
| AAG | Lys | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| UGU | Cys | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| UGC | Cys | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| UGA | 终止  | Trp | Trp | Trp | Trp | —   | Trp | Cys | —   | Trp | Trp | —   |
| UGG | Trp | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| AGU | Ser | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| AGC | Ser | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   | —   |
| AGA | Arg | 终止  | —   | —   | Ser | —   | Ser | —   | —   | Gly | Ser | —   |
| AGG | Arg | 终止  | —   | —   | Ser | —   | Ser | —   | —   | Gly | Ser | —   |

T1, 标准密码; T2, 脊椎动物线粒体密码; T3, 酵母线粒体密码; T4, 霉菌、原生动物和腔肠动物线粒体密码以及支原体和螺原体密码; T5, 无脊椎动物线粒体密码; T6, 纤毛虫、粗枝藻和六鞭虫的核密码; T9, 棘皮动物线粒体密码; T10, 整倍体核密码; T12, 可选的酵母核密码; T13, 海鞘线粒体密码; T14, 扁形虫线粒体密码; T15, 昆虫核密码。

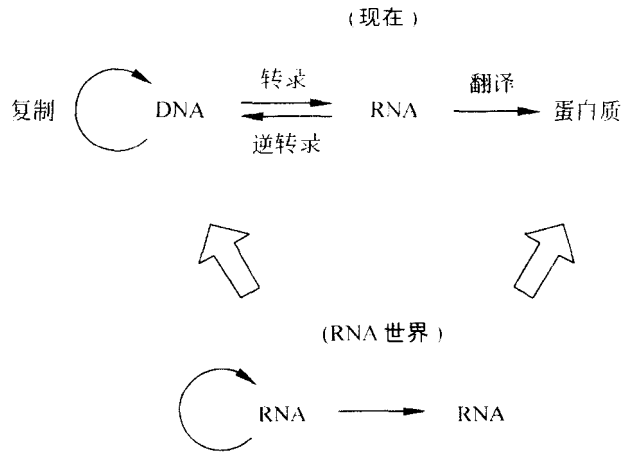


图 1.3 中心法则及其演变

但是，人们很快发现，在一种很特别的逆转录病毒里面存在着反向信息流——RNA 也可以转录为 DNA。Howard Temin 预言了这种反转录现象，而执行反转录功能的反转录酶也在 1970 年得到鉴定。逆转录病毒令人很感兴趣，这一类病毒包括了许多致癌病毒和 AIDS（艾滋病）病毒，它们的遗传信息存储在 RNA 而不是通常的 DNA 里。因此，遗传信息的传递需要首先从 RNA 反转录为 DNA 然后才能进行 DNA 复制。

到了 1977 年，又出现了一项惊人的发现。转录后的 mRNA 并不全部参与下一步的翻译。人们发现，编码蛋白质的基因被分为若干个片断，这就是所谓的外显子 / 内含子结构。转录后的 mRNA 必须经过加工过程，内含子被剪切掉，剩余的外显子拼接成为成熟的 mRNA。mRNA 的拼接在高等生物中十分普遍，一般在转录发生后就在核内进行。然后，成熟的 mRNA 被运送到核外，在核糖体上被翻译成蛋白质。某些物种的 tRNA 和 rRNA 基因也会发生剪接。另外，在某些物种中，对于 RNA 的后加工并不局限于仅仅剪切掉一些核苷酸（RNA 拼接），还包括一些更为复杂的加工过程，比如插入或者取代一些特定的核苷酸，这些操作被称为 RNA 编辑。从信息流方向的角度看，这些 RNA 转录后加工过程同中心法则并不矛盾，但是毕竟改变了 RNA 信息加工的复杂程度。

## RNA 世界

在我们这个星球上，几乎所有物种的基因组信息都存储在 DNA 上，并且通过 DNA 进行信息传递，而基因信息都通过蛋白质这样的功能分子得以表达。那么是不是 RNA 的主要作用就像中心法则所描述的那样，只是仅仅为更重要的 DNA 和蛋白质做信使呢？结论恰恰相反。人们普遍相信，生命始于 RNA。在还没有 DNA 和蛋白质的时候，存在着一个 RNA 的世界。此时，RNA 必然发挥着双重作用，既充当信息传递的基因组，又充当信息表达的功能分子。实际上，RNA 可以行使基因组的功能，而现在还存在一些病毒是以 RNA 作为基因组的。RNA 也可以发挥功能分子的作用。首先，tRNA 和 rRNA 都是基因表达得到的功能分子，而更重要的是，Thomas Cech 在 1981 年发现了一类新的 RNA 分子具有催化活性。以往，人们一直认为只有蛋白酶才能催化化学反应，而核糖酶的发现推翻了这一观点。从那时起，越来越多的具有催化作用的核糖酶被发现，其中一部分是天然存在的核糖酶，但是绝大部分都是通过体外分子进化实验设计并筛选得到的人工核糖酶。

如果 RNA 世界确实存在，那么中心法则的进化过程就应该像图 1.3 中所显示的那样。就信息存储的稳定性和复制的准确性而言，DNA 确实比 RNA 更有优势；就催化功能的多样性和特异性而言，蛋白质也确实比 RNA 更有优势。这也可能是为什么 DNA 和蛋白质会取代 RNA，分别独自执行信息存储（模板）和信息加工（催化）的功能，而在此之前，RNA 同时承担了这两项功能。像 RNA 剪接和 RNA 编辑这样奇怪的现象很可能就是当初 RNA 世界里 RNA 加工的痕迹，而 RNA 加工的本来目的是要从 RNA 基因组模板中生产出催化分子的。在像 ATP, cAMP, NAD 和辅酶 A 这样的具有生物活性的化学分子中，核糖比脱氧核糖更加常见，这也可以看做是 RNA 世界的痕迹。遗传密码的多样性（表 1.6）也说明了中心法则是在不同的物种中独立进化得到的。

那么，RNA 世界是如何变迁到 DNA / 蛋白质世界的呢？从 RNA 到

DNA 的转化是可以理解的，因为这两种分子十分相像，而且它们存储信息的方式是一样的。与此相比较，RNA 和蛋白质的差异就非常大了。然而值得注意的是，在从 RNA 向蛋白质的转换中，需要保存的是加工信息的能力，而不是存储信息的能力。因此，只要催化活性能够得以保存，RNA 向蛋白质的转换就能够发生在生命的进化之中；这就好比在计算机发展的早期，晶体管取代了阴极射线管一样。因为催化活性依赖于发生相互作用的分子之间的结构与化学互补性，因此二者之间接触面的“锁-钥匙”关系必须在两个世界的转化过程中保留下来。

## 细胞

正如我们所看见的，生命的本质在于信息的存储与加工。信息存储在分子之中，并且也在分子中进行加工。对存储的信息进行复制，这样的信息加工就是信息传递；从存储的信息模板中生产出催化剂，这样的信息加工就是信息表达。虽然中心法则巧妙地总结了参与这两种信息加工过程的分子，但是并没有说明信息加工发生的时间和地点。实际上，生命不能脱离细胞而单独存在，细胞为生命提供了信息加工的时间和空间。

早在 19 世纪早期的时候，人们就建立起这样一个概念：细胞是所有生物体的结构与功能单位。细胞内含有更小的亚细胞单元，比如存储遗传信息的细胞核。传统上，根据细胞的组织结构，生物体被划分为两大类：真核生物，其细胞内含有细胞核；原核生物，其细胞内不含细胞核。但是值得注意的是，根据分子层面的组织结构，生物体也可以被划分为三大类，因为原核生物又可以被分为细菌和古细菌。图 1.4 是一个真核细胞的结构示意图，其中包含细胞质、细胞核、内质网、高尔基体、线粒体以及其他一些细胞器。整个细胞以及每个细胞器都被特异性的膜包被着，这些膜只允许特异性的分子穿膜运输。细胞的组织结构对于分子执行信息加工功能提供了关键性的空间环境。

除了空间环境，细胞还提供了时间环境。根据细胞内部时钟或者对外部刺激发生反应，一个细胞会分裂为两个细胞。细胞分裂有两类：减数分裂

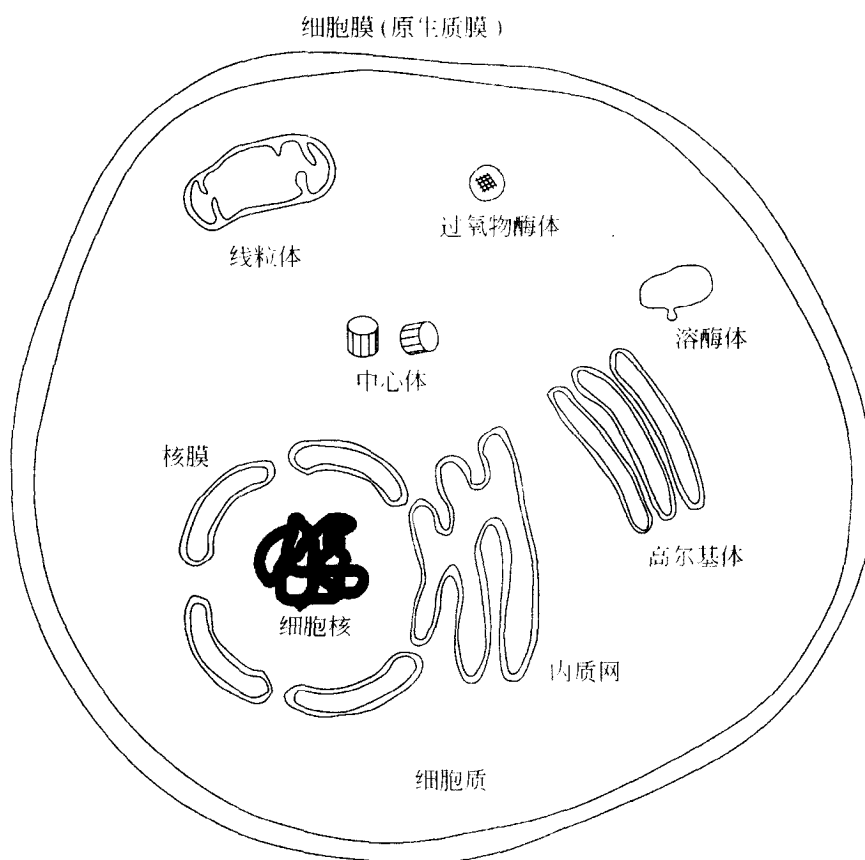


图 1.4 真核细胞示意图

和有丝分裂，分别对应于纵向信息传递和横向信息表达图 1.5 描述了两类分裂。减数分裂产生了生殖细胞或者配子，然后配子结合成受精卵，受精卵通过有丝分裂发育成为一个新的生物个体。因此，对于我们这个星球上存在的或者曾经存在过的生物而言，都必然存在着引发生命起源的细胞连续性，这就是所谓的生殖细胞系。

从传统的角度看，基因组是生命的蓝图：如果给定一个固定的环境，那么基因组就决定了生物发育的方方面面这种观点暗中假定了基因组是生命的指令系统，而其他的一切都归为环境因素。但是，必须强调指出，在生命的纵向信息传递中，被传递的不仅仅是基因组或者细胞核，而是整个细

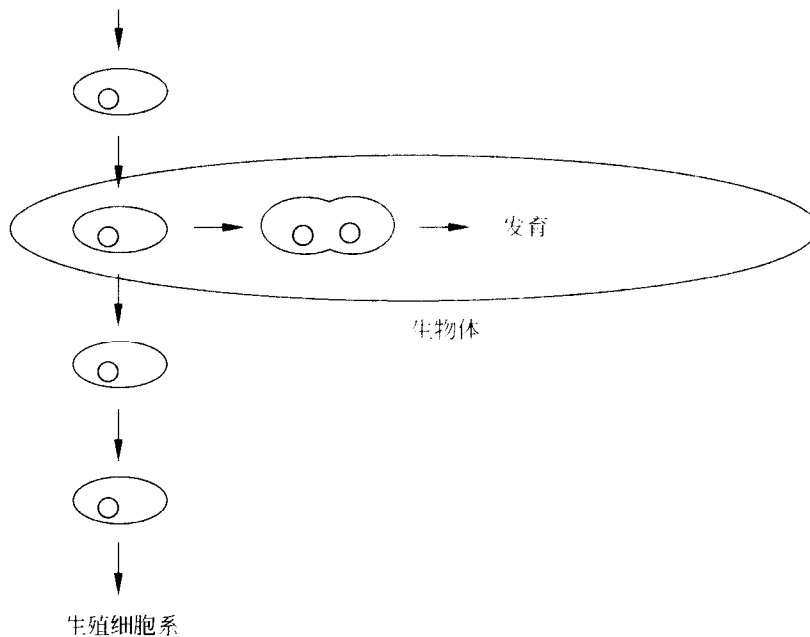


图 1.5 生命的双向信息流：纵向的遗传信息传递和横向的遗传信息表达

胞。如果指令系统存在于细胞的其他部位，那么一切又会如何呢？我们将在本章的后面讨论这一点。

## 分子生物学技术的发展

一部分子生物学发展的历史就是一部技术发展的历史，尤其是测定生物大分子（DNA, RNA 和蛋白质）序列与结构的技术发展。新的测定结构信息的实验技术极大地促进了生命科学各个方面的发展，因为分子结构信息为了解分子功能提供了线索。这是结构-功能关系的基本原理。图 1.6 列出了一系列分子生物学的历史中测定分子结构的里程碑事件，其中一些我们已经提及过。

Watson-Crick 的 DNA 双螺旋结构模型的基础就是低解析度的 X-ray 衍射数据。直到 20 世纪 70 年代末期，合成 DNA 短片段的单结晶才能得到高解析度的 X-ray 衍射数据。只有在得到了原子级的 DNA 结构解析图之

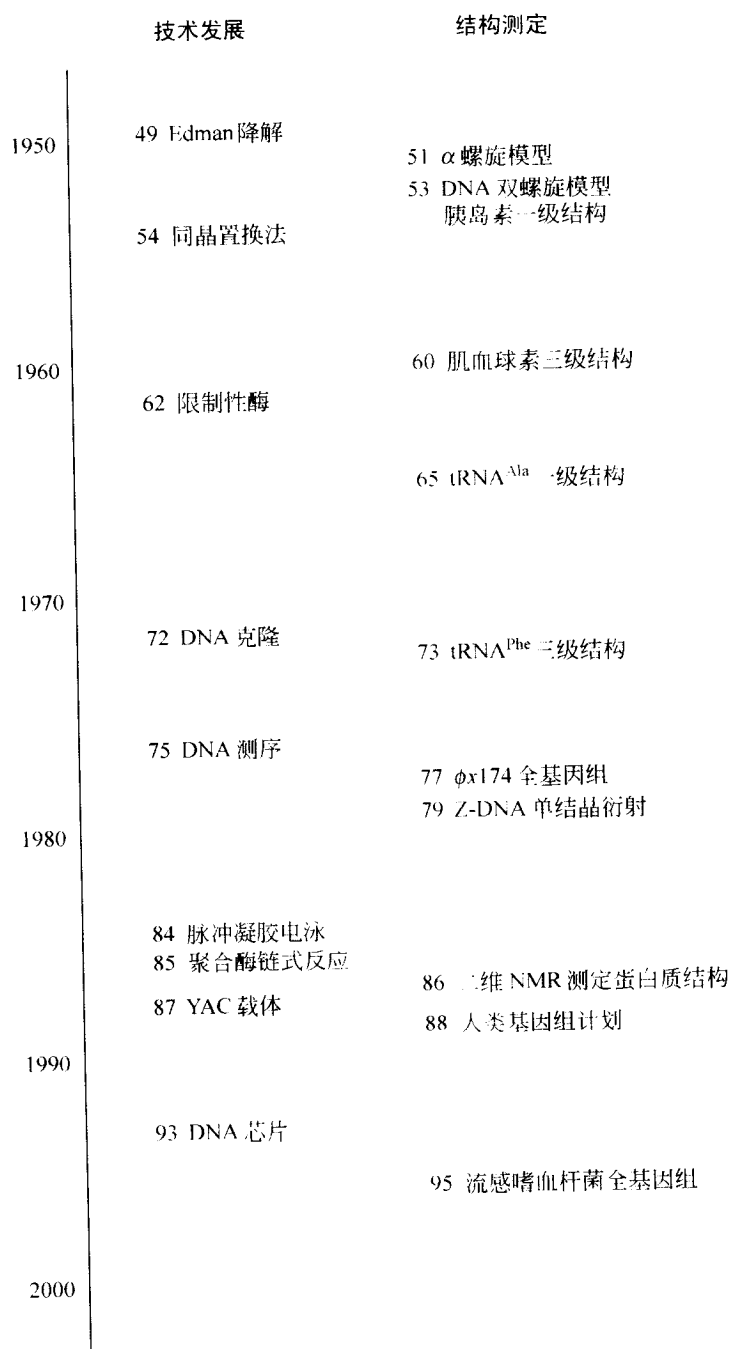


图 1.6 测定核酸与蛋白质结构信息的历史

后，人们才发现 DNA 结构的多态性远比原先设想的要多，图 1.7 显示了这一点。Watson-Crick 的 DNA 结构被称为 B 型 DNA；B 型 DNA 是右手螺旋结构，每个螺旋包含 10 个碱基对，还形成了一个深沟和一个小沟；碱基对的方向基本上垂直于螺旋的轴方向。在含水量较低的情况下，DNA 采取构象上比较倾斜的 A 型结构；另外，在高盐环境下，某些特定的序列会形成左手螺旋结构，被称为 Z 型结构。而且，即使是 B 型双螺旋结构，局部结构也同平均的 Watson-Crick 结构有所不同，呈现出显著的序列依赖性。

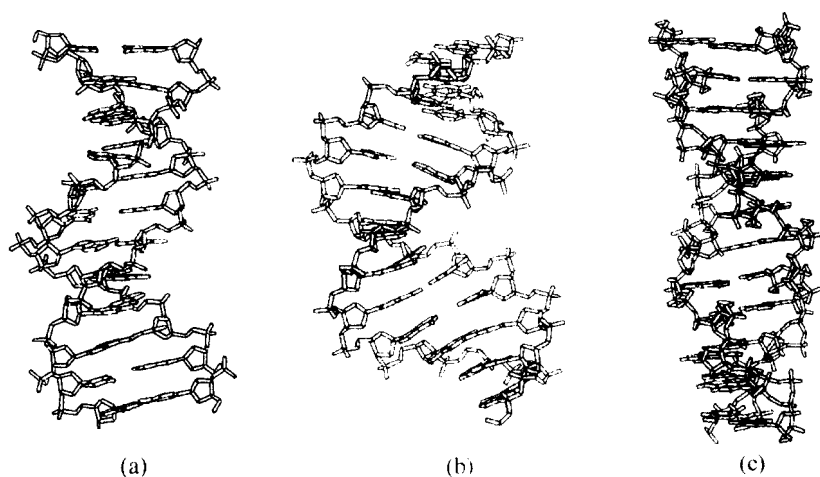


图 1.7 DNA 三级结构的多样性

(a) B 型 DNA(PDB:7BNA) ; (b) A 型 DNA(PDB:140D) ; (c) Z 型 DNA(PDB:2ZNA)

第一个测定 RNA 分子一级结构的是 Robert Holley 在 1965 年测定的丙氨酰转运 RNA 这比 DNA 测序方法早了 10 年。tRNA 具有自身互补的部分序列，能够形成自身的 Watson-Crick 碱基配对结构；正如图 1.8 所显示的那样，tRNA 形成了三叶草形状的二级结构。而后，在 1973 年，通过 X-ray 衍射得到的苯丙氨酰转运 RNA 的三维结构证实了这一结构模型；此次三维结构的测定还揭示了三维结构上的碱基配对稳定了 L 型结构。人们已经了解到 RNA 双螺旋结构采取的是 A 型结构。RNA 同蛋白质一样，特异性的三维结构决定了特异性的功能。

20 世纪 60 年代，人们发现了限制性酶。这一发现直接导致了 DNA 重组技术的发展，其中包括了 1975 年左右建立起来的 DNA 测序方法。II 型

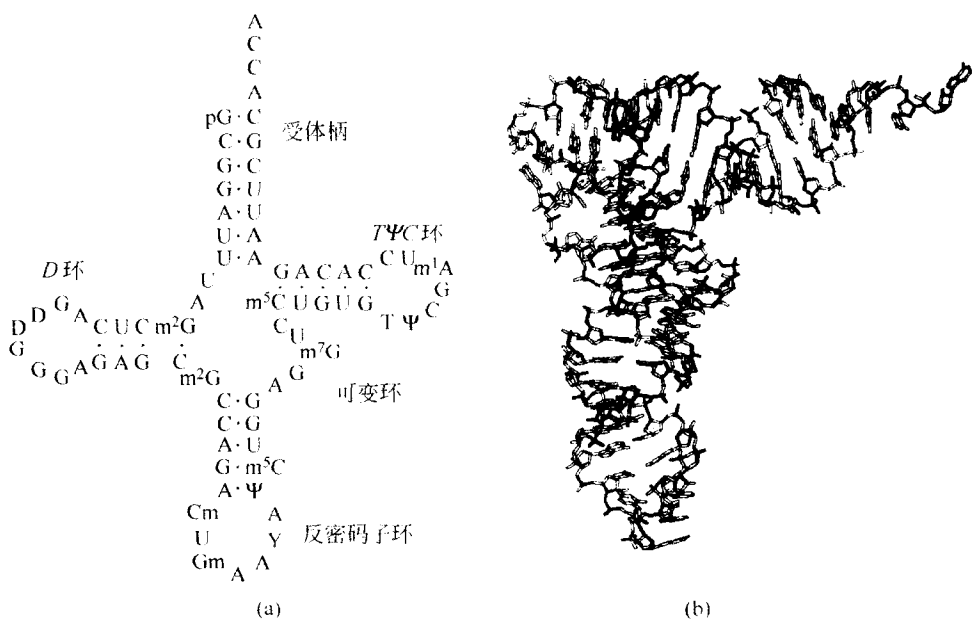


图 1.8 转运 RNA

(a) 酵母丙氨酰转运 RNA 的一级结构和二级结构；

(b) 酵母苯丙氨酰转运 RNA 的三级结构 (PDB: 1TRA)

限制性酶能够识别 DNA 双螺旋中的特异性模式，然后在识别位点附近剪切两条链。例如，从大肠杆菌中分离得到的 *EcoR* I 能够识别序列模式“GAATTC”然后在“G”和“A”之间发生剪切。因为 DNA 双链是以反方向平行的方式聚合在一起，而且彼此之间通过自身互补的方式进行识别，也就是说，DNA 双链具有二元对称性，因此，DNA 经过限制性酶剪切过后，会在每条链的剪切末端形成一段单链的末端，这就是所谓的粘性末端；粘性末端可以被用来同其他的 DNA 序列相连接。因此，限制性酶是一种选择性很好的剪切酶，经由限制性酶剪切过的序列易于识别，能够进行特异性的序列连接。

另一项针对 DNA 序列的重要手段是能够制作某条序列的多个拷贝。分子克隆技术是指把一段特异性的 DNA 片段引入载体，然后在细菌或者酵母细胞中得以扩增。PCR 技术（聚合酶链式反应）则不需要细胞就可以完成序列的扩增，只需要特定的聚合酶。Sanger 方法和 Maxam-Gilbert 方法都是 DNA 测序方法，这两种方法都是根据 DNA 序列片段长度的不同用电泳

的方法将它们分离开来。随着 DNA 测序方法的不断发展，新的克隆载体的不断发明，以及其他各种技术的不断进步，比如分离大的 DNA 片段的技术，最终导致了人类染色体解码成为可能。

## 人类基因组计划

1977 年，在 DNA 测序方法刚建立不久的情况下（图 1.6），一个很小的病毒（ $\phi_{x174}$ ）基因组被完全测序，它包含大约 5000 个核苷酸，11 个基因。而到了 1995 年，在 DNA 测序技术发展了 20 年后，第一个自由生物体被完全测序，它是流感嗜血杆菌（*H. influenzae*）。流感嗜血杆菌的全基因组包含了 180 万个核苷酸，大约 1700 个基因；在此之后，掀起了一股全基因组测序的浪潮，其部分原因是科学家们有计划地合作测定一系列生物体的全基因组，范围从细菌一直到高等真核生物。人类基因组计划的最终目的在于测定人类染色体上的 30 亿个核苷酸和大约 10000 个基因，并且利用这些信息为生物医学服务，以期获得更多的发现。这项计划启动于 20 世纪 80 年代末期，而测序本身可望在 21 世纪头几年完成。

图 1.9 比较了几个基因组的大小以及不同物种的全基因组测序历史。一般来说，病毒的基因组大小大约在  $10^3$  到  $10^5$  个核苷酸范围内，而自由生物的基因组大小则在  $10^6$  到  $10^9$  个核苷酸范围内。大体上，越高等的生物，其基因组也越大，但是人类基因组并不是最大的基因组；也许，用基因组中能够编码并转录为功能 RNA 和蛋白质的序列总和更能表明生物是否高等。细菌和古细菌的基因组大小大约在  $10^6$  个核苷酸左右，平均每 1000 个核苷酸大约就有一个基因。单细胞真核生物，如酿酒酵母（*S. cerevisiae*）的全基因组约为  $10^7$  个核苷酸，大约每 2000 个核苷酸就有一个基因。简单多细胞生物，如线虫（*Arabidopsis*）的全基因组大约包含  $10^8$  个核苷酸，大约每 5000 个核苷酸就有一个基因。因此，从中可以发现一个很明显的趋势：基因组越大，非编码序列区域所占的比重也越大。

由于人类基因组计划的启动，计算生物学，或者叫做生物信息学已经成为分子生物学的一个组成部分。生物信息学的首要功能就是要有效地组织