

分子进化基础

原作：李文雄（美国得克萨斯大学）

D. 戈劳尔（以色列特拉维夫大学）

翻译：陈建华

Fundamentals of Molecular Evolution

Wen-Hsiung Li (The University of Texas, Houston)

Dan Graur (Tel Aviv University)

Sinauer Associates, Inc. , Publishers

Sunderland, Massachusetts, 1991

序

1869年, Johann Friedrich Miescher 发现了DNA。他曾经“大胆”地认为, DNA可能和遗传有关。但过了一段时间, 他还是放弃了这种“荒谬”的观点, 并花费30多年的时间去研究鱼精蛋白—精子细胞中非常基本的蛋白质组分。按照Miescher的观点, 这种蛋白质是遗传的关键所在。后来发现, DNA分子不仅仅是遗传的关键, 正如一位分子进化学先驱Emile Zuckerkandl所说的那样, 它们还是“进化史的文献资料”。实际上, 每一种活组织的DNA都是历史过程的积累。然而, 包容在这些过程里的信息处在一种杂乱无章的无序状态中, 是分散的或片段的。有一些信息是隐藏的或是“伪装”的, 很难辨认出来; 有一些已丢失, 毫无踪迹可寻。分子进化的目的就是要澄清这些历史过程, 整理信息使之有序, 读出并解译信息。

由于进化过程难以用基因材料研究清楚, 用分子资料不仅可以重建进化年代, 而且也可以弄清进化过程的驱动力。分子生物学的重大突破, 如基因克隆技术、DNA序列分析、限制性核酸内切酶片段分析等, 在某种意义上, 已使科学家们处于一种新的十分有利的位置。我们能够洞察一个从未见过的世界, 在这里, 基因是通过复制、DNA混匀、核苷酸替代、转移及基因转化进化的。在这个世界里, 基因组或静止, 或流动, 有时在过了很长时间后会有微小的变化, 有时又会在眨眼间发生戏剧性的地质学尺度的变化。

通过研究遗传材料, 我们也能试着将生物界按系统发育事实来划分, 并建立分类系统。与传统的研究进化方法相反, 例如比较解剖学、形态学和古生物学, 都毫无必要地将自己限制在极端相似组织的进化关系研究中。我们现在已经能够建立巨大的家系树, 它连接脊椎动物、昆虫、植物、真菌和细菌, 并能够追溯它们的共同祖先, 一直到无法追溯的时代。

本书的宗旨是从分子水平上描述进化的原动力, 进化过程的驱动力及其基因组, 基因和它们的产物长期进化的各种分子机制所产生的结果。另外, 本书从进化观点为分子材料的比较和系统发育分析提供了基本的方法论。尽管个体受自然选择和其它过程所影响, 我们依然强调: 是群体和基因随着进化时间而变。为了在这些看起来完全不同的水平间构造联系, 我们借用了群体遗传学的基本概念。

我们着手为分子进化“初学者”写了这本书。同时, 也努力保持了方法的科学和规范, 并包含了一些定量方法。因此, 可以用数学和直觉解释, 从分子水平上描述进化现象和机制。这两种解释都不会削弱彼此的说服力, 还会相互补充, 以帮助读者更好地领会观点。我们没有试图达到泛泛的完满性, 而是提供了大量实例, 以支持和澄清许多理论观点和有争议的问题。

许多年来, 生物化学家和分子生物学家们把进化研究看作是大胆推测, 无根据的假设和无学科方法论的统一体, 而这种评价从未正确过。分子方法的引进不容置疑地将进化论变成一种“硬”科学。在这里, 能够从经验数据测度、计数或用计算机处理相关的参数, 也能用事实去检验理论的正确性。在今天的进化研究中, 推测学也服务于同物理学同样的目的。它们是定量性假说, 以刺激实验性工作。由此, 理论能够得以校正、精炼或被驳斥。本书的主要目的之一是通过加强本领域的事实基础, 表明进化研究已经达到了William Herschel先生于1831年所说的所有自然科学的共同目的, 即可以这样陈述其主张: “不是模糊和泛泛的, 而是在位置、重量和量度上尽可能精确”。

我们感谢许多同事、学生和朋友, 是他们帮助我们编辑了本书。他们的评论、建议、指正和讨论大大地提高了本书的质量, 也避免了许多。我们荣幸地感谢Sara Barton, Adina Breiman, David Cutler, David Hewett-Emmett, Winston Hide, Austin Hughes, Li Jin, Margaret Kidwell, Amanda Ko, Giddy Landan, William S. Lewis, Volker Loeschcke, Ora Manheim, David Mindell, Tatsuya Ota, Lori Sadler, Paul Sharp, Yuval Shuali, Jurgen Tomiuk, Danid Wool 和 Chung-I Wu. 特别地尴尬, 我们感谢C. William Birky Jr. 和 Bruce Walsh, 他们两位从头到尾地校阅了文稿并提出了很多有益的建

议。我们也向 Masatoshi Nei 博士表达感激之情，他对我们的调查研究工作给予了极大的鼓励和指导，来自世界卫生组织和美国—以色列两国科学基金会的支持也是本项合作成果得以面世的原因之一。

李文雄

D·戈劳尔

前 言

什么是分子进化

分子进化包括两个研究领域：(1) 大分子进化；(2) 基因和有机体进化史的重建。关于“大分子进化”，我们指的是在进化年代中，出现在遗传材料（如 DNA 序列）中的变化速率、模式及产物（如蛋白质）以及与此变化相关的机制，也被称做分子种系发生史。旨在探讨有机体和大分子进化史（正如从分子材料所推知的那样）。

这两个研究领域似乎由不相关的研究项目组成，因为第一个研究领域的目的是要阐明分子进化的原因和结果；而第二个研究领域是仅仅把分子用作工具，以重建有机体及其遗传组分的演化史。然而在实践中，这两个方面是紧密相关的，一个领域的进步也会促进另一领域的发展，例如：种系发生知识对决定分子特征的变化顺序是很必要的；相反，给定分子的变化模式和速率知识对试图重建类群的进化史是很重要的。

传统上，第三个研究领域，前生物（Prebiotic）进化或称“生命起源”也包含在分子进化框架中。然而，它涉及到很多猜想，也极少用到定量方法，并且，在前生物系统（即缺乏可复制基因的系统）信息传递过程的规律，目前还不清楚。因此，本书并不侧重生命起源，有兴趣的读者可以参阅 Oparin (1957), Cairns-Smith (1982), Dyson (1985) 和 Loomis (1988) 的著作。

分子进化研究起源于两个不同的学科：群体遗传学和分子生物学。群体遗传学为研究进化过程提供了理论基础，而分子生物学则为其提供了经验数据。因此，掌握群体遗传学和分子生物学的基本知识对理解分子进化是很有裨益的。

1 基因结构与突变

本章提供为研究DNA水平上的进化过程所必需的、关于分子生物学方面的基础知识。最为基本的部分是一种典型的真核生物基因的基因组结构和各种突变类型。更进一步的背景知识将在以后的有关章节中讲述出。某些术语请参看词汇表。

1.1 DNA序列

除某些病毒外，所有生物的遗传信息都由脱氧核糖核酸DNA（deoxyribonucleic acid）分子所携带。DNA通常是由两条相互缠绕的互补链所组成，形成一种右手螺旋状。每条链都是一条由四种核苷酸组成的线性多聚核苷酸。有两种嘌呤（purines）：腺嘌呤A(adenine)和鸟嘌呤G(guanine)，两种嘧啶(pyrimidines)：胸腺嘧啶T(thymine)和胞嘧啶C(cytosine)。两条链靠核苷酸之间的氢键联系在一起。腺嘌呤与胸腺嘧啶通过两个氢键，又称弱键（weakbond）配对；而鸟嘌呤与胞嘧啶通过三个氢键，又称强键（strong bond）配对(图1-1)。

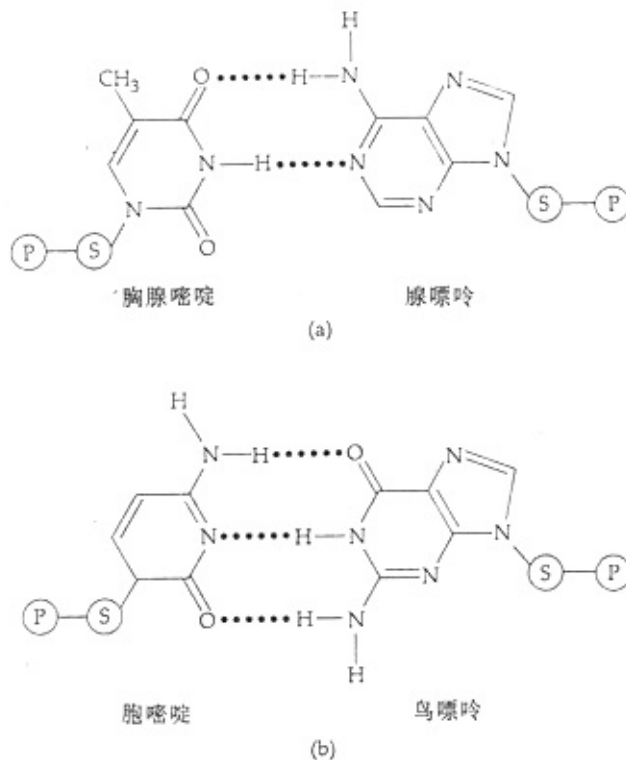


图1-1 互补碱基靠氢键（由圆点构成的线）配对，（a）在胸腺嘧啶和腺嘌呤之间（弱键），和（b）在胞嘧啶和鸟嘌呤之间（强键）。P：磷酸根；S：糖。

DNA序列中的每一个核苷酸都包含一个戊糖（脱氧核糖）、一个磷酸基团和一个嘌呤或嘧啶碱基。DNA分子的主干由糖和磷酸等部分构成，它们通过不对称的5' -3' 磷酸二酯键共价连结在一起。所以，DNA分子是一种有极性的分子，它的一端，在末端核苷酸的5' 碳原子上有一个磷酸根（-P），另一端，其末端核苷酸的3' 碳原子上有一个游离羟基（-OH）。磷酸二酯键的方向决定着分子的特性；

例如，序列 5' -G-C-A-A-T-3' 与序列 3' -G-C-A-A-T-5' 是两种不同的序列。为了方便起见，DNA 序列按其转录的顺序，即从 5' 端到 3' 端书写，它们又分别被称为上游 (upstream) 和下游 (downstream) 方向。DNA 的双螺旋形式有两条链，按反向平行方向排列着 (图 1-2)。

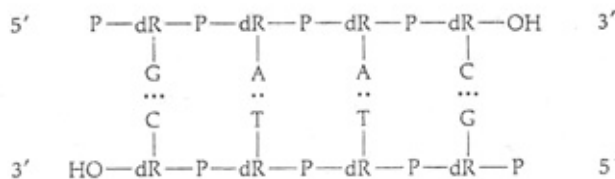


图 1-2 双链 DNA 的反向平行结构的图示

P, 磷酸根; dR, 脱氧核糖; CH, 羟基; A, 腺嘌呤; G, 鸟嘌呤; C, 胞嘧啶; T, 胸腺嘧啶; —, 共价键; .., 弱键; ..., 强键

核糖核酸 RNA (ribonucleic acid) 既有双链分子又有单链分子。RNA 与 DNA 不同之处在于，其主干糖的部分是核糖而不是脱氧核糖，由尿嘧啶核苷酸代替了胸腺嘧啶核苷酸。腺嘌呤，胞嘧啶，鸟嘌呤，以及胸腺嘧啶/尿嘧啶，都是标准核苷酸。某些功能 RNA 分子，特别是 tRNA 分子，常含有非标准核苷酸，即在转录之后导入 RNA 分子中的、将标准核苷酸加以化学修饰的核苷酸。

1.2 基因结构

过去，基因被定义成为一个多肽链编码或确定一个功能 RNA 分子的一个 DNA 片段。然而，最近的分子研究却从根本上改变了我们对基因的认识，所以，我们将采用一个更有点生气的定义，即基因是对某一特殊功能至关重要的基因组 DNA 或 RNA 的一个序列。执行这一功能可能并不需要该基因被翻译出来，甚至无需被转录出来。

现在，已有三类基因被认识：(1) 为蛋白质编码的基因 (protein coding genes)，它们在被转录成 RNA 之后再被翻译成蛋白质；(2) RNA 基因 (RNA specifying genes)，它们仅被转录；(3) 调节基因 (regulatory genes)。按照狭义的定义，第三类只包括那些不转录的序列。被转录的调节基因本质上属于前两类之一。蛋白质编码基因和 RNA 基因又被看成是结构基因 (structural genes)。注意，某些作者将结构基因的定义，限制在只包括蛋白质编码基因的范围。

在细菌中，结构基因的转录仅由一种依赖 DNA 的 RNA 多聚酶执行。与之不同的是，在真核生物里要用到三种 RNA 多聚酶。核糖体 RNA (rRNA) 基因由 RNA 多聚酶 I 转录，蛋白质编码基因由 RNA 多聚酶 II 转录，小分子细胞质 RNA (scRNA) 基因，象转移 RNA (tRNA) 的基因，则由 RNA 多聚酶 III 转录。某些小分子细胞核 RNA (snRNA) 基因由多聚酶 II 转录，另一些则由多聚酶 III 转录。一种 snRNA 基因 U6，可能既被多聚酶 II 又被多聚酶 III 转录。

蛋白质编码基因

真核生物的一个标准的蛋白质编码基因由转录部分和不转录部分构成 (图 1-3a)。不转录部分根据它们相对于蛋白质编码基因的位置，被命名为 5' 和 3' 侧序列 (flanking sequences)。5' 侧序列含有几种信号 (特异序列)，它们决定转录过程的起始、缓和和终止。因为这些调节序列起始转录过程，所以，它们也被称为启动子 (promoters)，而它们所处的部位则叫启动子区 (promoter region)。启动子区由以下信号构成：TATA 块 (TATA box)，位于转录起始点上游 19-27 碱基对 (bp) 处，更上游一些的 CAAT 块 (CAAT box)，以及一个或多个 GC 块 (GC box) 拷贝，后者由序列 GGGCGG 或其变异体组成，位于 CAAT 块的周围 (图 3a)。可能在定向方面起作用的 CAAT 块和 GC 块，控制 RNA 多聚酶的起始联结，而 TATA 块则控制转录起始点的选取。不过，我们注意到，以上信号对启动子的功能而言都不是必不可少的。某些基因不具有 TATA 块，所以没有一种独一无二的转录起始点。另一些基因既没有 CAAT 块也没有 GC 块，它们的转录起始是受 5' 侧区域中的其他因素所控制的。3' 侧序列含有关于转录过程终止的信号和多聚腺苷酸尾部的信号。目前，还不能精确地标明一个基因开始和终止的位点。

真核生物中为蛋白质编码的基因的转录起始于转录起始部位 (transcription initiation site) (RNA 转录物中的帽子部位 (cap site)), 结束于终止部位 (termination site), 后者与成熟信使 RNA (mRNA) 分子的多聚腺苷化作用 (polyadenylation) 或多聚腺苷酸尾部位 (poly(A) addition site) 可能等同, 也可能不等同。换言之, 转录的终止可能发生在比多聚腺苷酸尾部更下游的地方。转录而成的 RNA, 又称前信使 RNA (pre messenger RNA) (pre mRNA) 包含 5' 和 3' 不翻译区, (untranslated regions), 外显子 (exons) 和内含子 (introns)。内含子, 或插入序列, 是一些虽被转录但在前 mRNA 分子的加工中被剪除的序列。所有保留在经拼接而达成熟的 mRNA 中的基因组序列称外显子。被翻译出来的外显子或外显子的某些部分称为蛋白质编码的外显子或编码区 (coding regions)。

有几种类型的内含子, 按它们从前 mRNA 中剪除的特异机制来界定其性质 (见 Lewin, 1990)。这里我们只关心那些受 RNA 多聚酶 II 转录的细胞核基因中的内含子。这些内含子在分子成熟期间, 在酶促作用下从前 mRNA 中剪除。这些内含子的拼接位点 (splicing sites or junctions) 可能由每一内含子的 5' 和 3' 端上的核苷酸来决定, 分别称给体 (donor) 和受体 (acceptor) 部位。例如, 所有真核生物的细胞核内含子都以 GT 开始, 以 AG 结束 (GT-AG 规则 (GT-AG rule)), 而已表明这些序列对于正确的剪切和拼接是至关重要的 (图 1-3 a)。邻近内含子的外显子序列可能对拼接部位的决定也有贡献。此外, 每一内含子都含有一段特别的序列, TACTAAC 框 (TACTAAC box), 位于内含子 3' 端上游

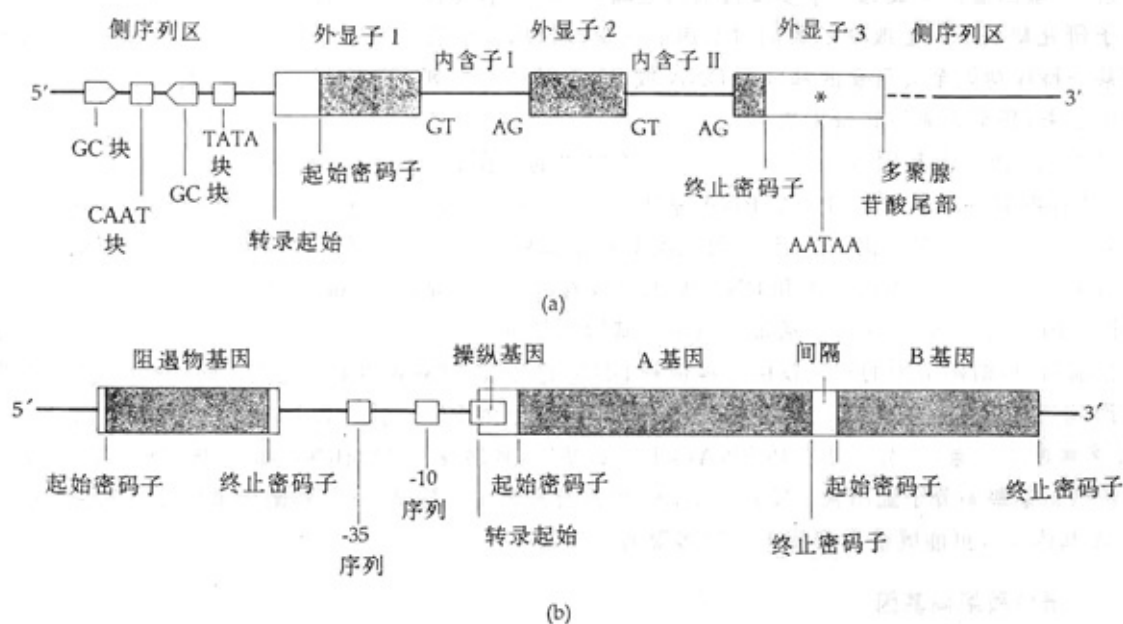


图 1-3 (a) 典型真核生物的蛋白质编码基因的图示结构。注意, 按照习惯 5' 端放在左边。(b) 受诱导的原核生物的操纵子的图示结构。基因 A 和 B 是为蛋白质编码的基因, 并被转录形成一个信使 RNA。阻遏物基因为一个阻遏蛋白编码, 后者结合在操纵基因上, 并通过遏制 RNA 多聚酶的移动从而阻碍结构基因的转录。操纵基因是一段至少有 10 个碱基长的 DNA 区域, 它可能与操纵子中基因的被转录区域重叠。经与一种诱导物 (一种小分子) 结合后, 阻遏物转化成一种不能与操纵基因结合的形式。于是 RNA 多聚酶就能起动操纵子中基因 A 和 B 的转录 (见 Lewin 1990)。(a) 和 (b) 两图中的各区都不是按其尺度来画的,

约 30 个碱基的地方。这一序列在酵母细胞核基因中是相当保守的而它在更高等的真核生物的基因中则要变化得多样一些。拼接过程包括 5' 拼接部位的裂开, 以及内含子 5' 端的 G 和 TACTAAC 框中第 6 位上的 A 之间形成磷酸二酯键。随后, 3' 拼接部位裂开, 而两个外显子则被连在一起。

内含子的数目因基因而异。某些基因具有十几个内含子, 其中有些内含子可能几千个核苷酸长。另一些 (例如, 大多数组蛋白基因) 则完全没有内含子。外显子在整个基因中并不是均匀分布的。

有些外显子密集在一起，另外一些则位于与邻近外显子相距很远的地方。图 1—4 给出了说明这一现象的一个例子。注意，细胞核蛋白质编码基因的绝大多数具有内含子。并非所有内含子都起分裂编码区的作用。有些内含子出现在不翻译区域，主要在转录起始位点和起始密码子之间的区域中。

真细菌类的蛋白质编码基因与真核生物中的这类基因有几方面不同，最重要的是，它们不含内含子(图 1—3 b)。真细菌中的启动子含有一个—10 序列(—10 sequence) 和一个—35 序列(—35 sequence)，分别位于转录起始位点上游 10 bp 和 35bp 的地方。前者也称普里伯劳块(Pribnow box)具有序列 T A T A A T 或其变异形式，而后者则具有序列 T T G A C A 或其变异形式。原核生物的启动子在离—35 序列更上游的地方，可能还含有其他一些特别的序列。

原核生物中几个结构基因可能呈连续排列以形成一个基因表达单位，它被转录成一个 mRNA 分子，接着再翻译成不同的蛋白质。这种单位通常都含有控制属于该单位的基因协调地表达的遗传因素。这一由基因排列成的整体称为一个操纵子(operon)(图 1-3b)。

RNA 基因

原核生物和真核生物中的 RNA 基因，其结构通常是相似的。这些基因一般不含内含子。然而，在有些生物中，象纤毛虫、粘菌和细菌中，确定 RNA 的基因就可能含有在 RNA 变成有功能分子前必须被剪除的内含子。涉及某些 RNA 基因的转录调节的序列因素，有时被包括在决定功能终产物的序列的范畴内。特别地，所有确定 tRNA 的基因都含有一个受 RNA 多聚酶 III 识别的内部转录起始点。

许多 RNA 分子在转录之后都受到修饰。这类修饰包括标准和非标准核苷酸的掺入，标准核苷酸转变成非标准核苷酸的修饰，向 5' 或 3' 末端的核糖核苷酸的终端序列的酶促添加等。

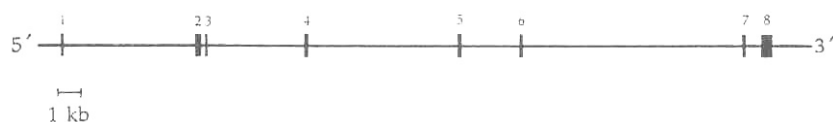


图 1—4 人的 IX 因子基因中 8 个外显

子位置。图中竖线代表 8 个外显子。本图只标出了被转录的区域，外显子和内含子均按其尺度绘出。注意，5' 不翻译区比 3' 不翻译区短。资料取自 Yoshitake 等(1985)

调节基因

我们对调节基因的认识比其他类型基因的认识要晚一些。暂时只有几种这类基因或基因家族被鉴别出来。它们是：(1)复制子基因(replicator genes)，作为 DNA 复制的起始和终止的特异位点而起作用，(2)重组子基因(recombinator genes)，提供受与重组有关的酶识别的特异位点，(3)分离子基因(segregator genes)，在有丝分裂和减数分裂期间，提供将染色体附着到分离机构上的特异位点，以及(4)附着位点(attachment sites)，为蛋白质、激素，或其他分子所附着。可能还存在更多的调节基因，它们中有些可能在功能和位置两方面都是独立于结构基因的，并且可能与更复杂的调节功能，象多细胞生物的个体发育有关。

1.3 遗传密码

蛋白质合成涉及到一个解码过程，藉此，一个由 mRNA 分子携带的遗传信息通过转移 RNA(tRNA)介体的使用，而被翻译成氨基酸。20 种基本氨基酸及其缩写形式列于表 1—1。翻译从翻译起始位点开始，进行到一个终止信号时为止。翻译过程涉及相互邻近的不重叠核苷酸三联体，称为密码子(codons)的连续识别。一个将被翻译的序列的相位由起始密码子决定，称为阅读框架(reading frame)。在处于核糖体和 mRNA 分子间的中间相位的翻译机构中，每一个密码子被翻译成一个特异氨基酸，它被连续地添加到正在延长的多肽链上。密码子与氨基酸之间的对应由一组称为遗传密码(genetic

code)的规则来决定。除了几个例外情况(见后面)，与细胞核为蛋白质编码基因有关的遗传密码是“通

用的”(universal), 即几乎所有真核生物的细胞核基因和原核生物的基因的翻译, 都是由同一组规则所决定的。

通用密码子由表 1 — 2 给出。由于 1 个密码子由 3 个核苷酸组成, 且存在四种不同类型的核苷酸, 所以, 应有 $4^3=64$ 种可能的密码子。其中 61 个为特异的氨基酸编码, 并被称为有意义密码子(sense codons), 其余 3 个则为无意义密码子或终止密码子 (nonsense or stop codons), 其作用是充当使翻译过程终止的信号。因为有意义密码子有 61 个, 而蛋白质中的氨基酸却只有 20 种, 所以, 大多数氨基酸由 1 个以上的密码子编码。这样一类密码被称为是简并密码 (degenerate code)。编码同一氨基酸的不同密码子称同义密码子 (synonymous codons)。相互之间仅在第 3 位上不同的同义密码子被指定为一个密码子族 (codon family)。例如, 为缬氨酸编码的 4 个密码子组成一个四密码子族。与之对比, 为丝氨酸编码的 6 个密码子被分成一个四密码子族和一个二密码子族。

表 1 — 1 基本氨基酸及其三字母和一字母缩写

名 子	三字母缩写	一字母缩写
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

表 1 — 2 通用密码子

密码子	氨基酸	密码子	氨基酸	密码子	氨基酸	密码子	氨基酸
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

大多数真核生物的蛋白质中第 1 个氨基酸是甲硫氨酸, 由起始密码子 (initiation codon)AUG 编码。

这一氨基酸在成熟蛋白质中通常被移除了。大多数原核生物的基因也用AUG密码子作为起始，起始翻译过程的氨基酸则是甲硫氨酸的衍生物，称甲酰甲硫氨酸。

表 1－3 哺乳动物线粒体的密码子

密码子	氨基酸	密码子	氨基酸	密码子	氨基酸	密码子	氨基酸
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Trp
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Met	ACA	Thr	AAA	Lys	AGA	Stop
AUG	Met	ACG	Thr	AAG	Lys	AGG	Stop
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

*与通用密码子的差别用黑体字表示。

通用密码子在质体—象维管植物的叶绿体—基因组所采用的独立的翻译过程中也被使用着。比较起来，大多数线粒体基因组和几种细胞核基因组(例如，枝原体属 *Mycoplasma* 和四膜虫属 *Tetrahymena*)则使用与通用密码子不同的密码子。不过，一般说来这些密码子与通用密码子之间只有较小的差别。作为一例，哺乳动物线粒体的密码子列于表 1－3。注意，通用遗传密码中编码丝氨酸的密码子有两个被用作终止密码子，而色氨酸和甲硫氨酸都是被两个密码子编码，而不是一个。

1.4 突变

DNA序列在染色体复制过程中，正常情况下被精确地拷贝。然而，极偶然地也会出现错误，从而产生新的序列。这些错误称之为突变(mutations)。突变既可发生在体细胞内又可发生在种系细胞内。由于体细胞突变是不遗传的，所以，在只讨论进化问题时我们将不考虑它们。本书通篇提及的“突变”这个词，将只指在种系细胞中的突变。

突变可以根据受突变事件影响的DNA序列的长短来分类。例如，突变可能影响到一个核苷酸(点突变(point mutations))或几个相互邻近的核苷酸。我们也可按突变事件造成的变化类型，将突变分成：(1)替换(substitutions)，即一个核苷酸被另一个所取代，(2)缺失(deletions)，即从DNA中移去一个或多个核苷酸，(3)插入(insertions)，向序列中添加一个或多个核苷酸，(4)倒位(inversions)，即含有2个或多个碱基对的双链DNA片段转动180°(图1－5)。

核苷酸替换

核苷酸替换分成转换(transition)和颠换(transversions)两类。转换是A和G(嘌呤)之间或T和C(嘧啶)之间的替换。颠换是一个嘌呤和一个嘧啶之间的替换。

发生在蛋白质编码区的核苷酸替换也可通过它们对翻译产物蛋白质的影响来定性。如果替换不引起氨基酸改变(图1－6a)，则被称为是同义的(synonymous)或沉默的(silent)。非同义的和引起氨基酸改变的突变，又可进一步被划分成误义的(missense)和无义的(nonsense)突变。误义突变将受到

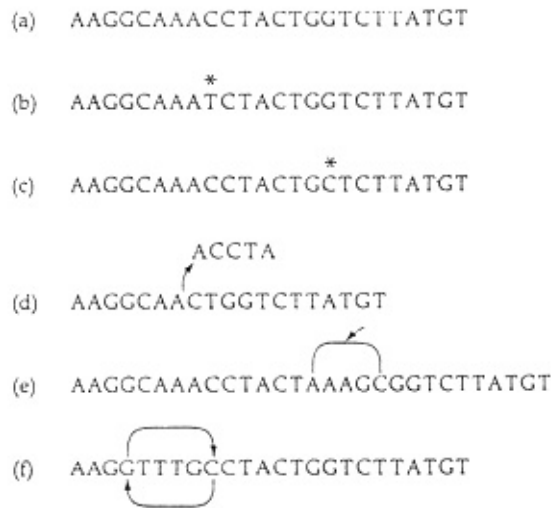


图 1-5 突变的类型

(a) 原序列; (b) 从 C 到 T 的转换; (c) 从 G 到 C 的颠换; (d) 缺失序列 A C C T A; (e) 插入序列 A A A G C; (f) 5' -GCAAAC- 3' 倒位成 5' -CAAACG- 3'

影响的密码子变成另一种密码子, 后者确定的氨基酸与前者编码的氨基酸互不相同 (图 1-6 b)。无义突变则将密码子变成一个终止密码子, 这样, 由于翻译过程在成熟前就结束了, 所以最终结果是产生一个截掉尾巴的蛋白质 (图 1-6 c)。

(a)	Ile	Cys	Ile	Lys	Ala	Leu	Val	Leu	Leu	Thr
	ATA	TGT	ATA	AAG	GCA	CTG	GTC	CTG	TTA	ACA
	ATA	TGT	ATA	AAG	GCA	CTG	GTA	CTG	TTA	ACA
	Ile	Cys	Ile	Lys	Ala	Leu	Val	Leu	Leu	Thr

(b)	Ile	Cys	Ile	Lys	Ala	Asn	Val	Leu	Leu	Thr
	ATA	TGT	ATA	AAG	GCA	AAC	GTC	CTG	TTA	ACA
	ATA	TGT	ATA	AAG	GCA	AAC	TTC	CTG	TTA	ACA
	Ile	Cys	Ile	Lys	Ala	Asn	Phe	Leu	Leu	Thr

(c)	Ile	Cys	Ile	Lys	Ala	Asn	Val	Leu	Leu	Thr
	ATA	TGT	ATA	AAG	GCA	AAC	GTC	CTG	TTA	ACA
	ATA	TGT	ATA	TAG	GCAAACGTCCTGTTAACA					
	Ile	Cys	Ile	Ter						

图 1-6 发生在编码区中的替换类型

(a) 同义的, (b) 误义的, 和 (c) 无义的

每一个有意义密码子通过一次核苷酸替换可以突变成 9 种其他密码子。例如, C C U (p r o) 可以经历 6 种非同义的替换, 变成 U C U (Ser), A C U (Thr), G C U (Ala), C U U (Leu), C A U (His) 或 C G C (Arg), 以及 3 种同义的替换, 变成 C C C, C C A 或 C C G。因为通用遗传密码由 61 个有意义密码子组成, 所以有 $61 \times 9 = 549$ 种可能的核苷酸替换。如果我们假定核苷酸替换随机地进行且编码区中所有密码子具相同的频率, 那么, 我们就能根据遗传密码算出不同类型核苷酸替换的期望比例。这些数据列于表 1-4。由于遗传密码的结构, 同义替换主要发生在密码子的第 3 位上。事实上, 第 3 位上所有可

能发生的改变中，大约 70 %是同义的。相比之下，密码子第 2 位上的所有替换都是非同义的，而第 1 位上发生的核苷酸变化，绝大多数也是如此（96 %）。

表 1—4 随机蛋白质编码序列中不同类型突变替换的相对频率

替换	数目	百分比
全部密码子总数	549	100
同义的	134	25
非同义的	415	75
误义	392	71
无意义	23	4
第一位总数	183	100
同义的	8	4
非同义的	175	96
误义	166	91
无意义	9	5
第二位总数	183	100
同义的	0	0
非同义的	183	100
误义	176	96
无意义	7	4
第三位总数	183	100
同义的	126	69
非同义的	57	31
误义	50	27
无意义	7	4



图 1—7 不等价交换，结果是子链之一中某一 DNA 序列缺失，而另一子链中则出现同一序列的重复。图中矩形表示某一特定长度的 DNA

缺失和插入

缺失和插入可能由几种机制造成。机制之一是不等价交换 (unequal crossing over)。图 1—7 绘出了一个简单模型，按此方式，两条染色体间不等价交换，结果在一条染色体上某一 DNA 片段缺失，而在另一条上则出现相应的添加。另一个机制是复制滑脱 (replication slippage) 或滑脱链误配 (slipped strand mispairing)。这类事件发生在含有邻接短重复序列的 DNA 区域中。图 1—8 a 表示，在 DNA 复制期间，滑脱可因邻接重复间的误配而引起，而滑脱又可造成某一 DNA 片段缺失或重复的后果，至于究竟是缺失还是重复，则要看滑脱发生在 5' → 3' 方向还是与之相反的方向。图 1—8 b 表明，滑脱误配也可能发生在非复制 DNA 中。造成 DNA 序列插入或缺失的第三种机制是 DNA 转座，这将在第 7 章中加以探讨。

缺失和插入统称裂缝 (gaps)，因为当一个带有缺失或插入的序列与原序列比较时，两序列中就将有一个会出现“裂缝” (第 3 章)。裂缝事件中涉及的核苷酸数目，范围从一个或几个到包括成千个核苷

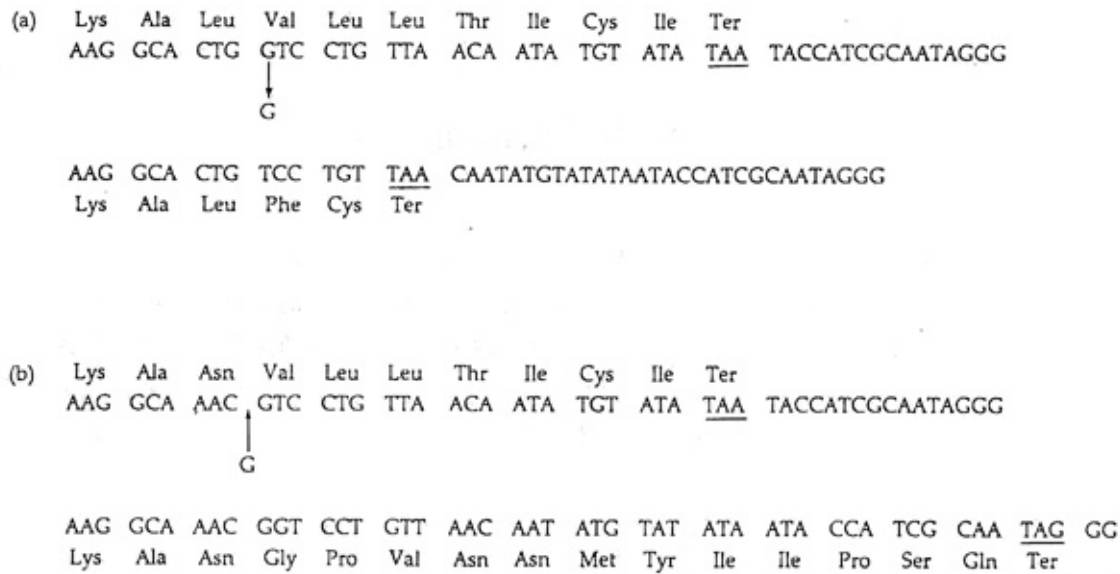


图 1 - 9 由缺失或插入造成的阅读框架发生框架移动的例子。(a) 缺失一个 G 造成成熟前终止。(b) 插入一个 G, 结果抹掉了一个终止密码子。终止密码子下划有横线。

习题

- 1 从某种哺乳动物中找出两个完全定序了的为蛋白质编码的基因 (你可以应用数据库或一本杂志) 。确定它们每一个的内含子和编码外显子间的长度比。用取自果蝇的两个为蛋白质编码的基因再做上述处理。哪些基因的内含子与外显子长度比更大?
- 2 从某种哺乳动物中找出一个完整的 m R N A 序列。删去你在编码区中最先碰到的两个核苷酸 C T 。判断是否有一个成熟前终止密码子进入阅读框架, 或者是否翻译会超越原终止密码子而延长。
- 3 用习题 2 中的序列, 在编码区中第 6 个 C 后插入一个 A 。插入将会怎样影响翻译?
- 4 用习题 2 中的序列, 判断有多少密码子能经一次核苷酸替换而突变成终止密码子。
- 5 用习题 2 中的序列, 判断密码子第 3 位中有多少转换是同义的, 而又有多少颠换是同义的。

后继阅读文献

- Darnell, J. E., H. F. Lodish and D. Baltimore. 1990. *Molecular Cell Biology*, 2nd Ed. Scientific American Books, New York.
- Lewin, B. 1990. *Genes* IV. Oxford University Press, New York.
- Stryer, L. 1988. *Biochemistry*, 3rd Ed. Freeman, New York.
- Suzuki, D. T., A. J. F. Griffiths, J. H. Miller and R. C. Lewontin. 1989. *An Introduction to Genetic Analysis*, 4th Ed. Freeman, New York.
- Watson, J. D, N. H. Hopkins, J. W. Roberts, J. A. Steitz and A. M. Weiner. 1987. *Molecular Biology of the Gene*, 4th Ed. Benjamin/Cummings, Menlo Park, CA.

2 群体中基因的动力学

群体遗传学探讨发生在群体内的遗传变化。本章将介绍某些对理解分子进化来说至关重要的、群体遗传学的基本原理。群体遗传学的一个基本问题，是确定一个突变基因的频率在各种进化力量的影响下，将如何随时间而变化。此外，从长期的观点看，重要的是决定一个新的突变变异型完全替代群体中的老变异型的概率，以及估计替代过程将会进行得多快。与形态上的改变不同，许多分子变化好象对生物的表型只有很小的影响，所以，分子变异型的频率受机遇的影响强烈。因此，在处理分子进化时，机遇因素应该被考虑进去。

2.1 等位基因频率方面的改变

一个基因在染色体或基因组上的位置称基因座位 (locus)，某一给定基因座位上所选取的基因形式称等位基因 (alleles)。在一个群体中，某一基因座位上可能有一个以上的等位基因存在，它们的相对比例称等位基因频率 (allele frequencies) 或基因频率 (gene frequencies)。例如，我们假定，一个大小为 N 的单倍体群体中，某一基因座位上有各具 n_1 和 n_2 个拷贝的两个等位基因。那么，它们的等位基因频率就分别等于 n_1/N 和 n_2/N 。注意，其中 $n_1 + n_2 = N$ ，而 $n_1/N + n_2/N = 1$ 。

进化是在群体的遗传组成方面发生变化的过程。因此，进化过程中最基本的部分是等位基因频率随时间的变化。事实上，从进化的观点看，一个新突变要变得有意义就必须增加它自己的频率，并最终在群体中被固定 (fixed) (即在随后的世代中所有个体将共有同一种突变型等位基因)。如果不增加自己的频率，那么这个突变将对该物种的进化史几乎没有影响。对于一个要增加频率的突变型等位基因来说，必须是某些因子而不是突变来掺入作用。这些因子包括自然选择，随机遗传漂变，重组和迁徙。

为了认识进化的过程，我们必须研究以上因子是如何影响等位基因频率的变化的。本书里，我们只讨论自然选择和随机遗传漂变。在涉及形态学性状的经典进化研究中，自然选择被看成是进化的主要驱动力量。与此成鲜明对比的是，在分子水平上随机遗传漂变被认为在进化中起重要作用。

研究群体中的遗传变化有两种数学途径：决定性的和随机性的。决定性模型 (deterministic model) 较简单。它假定，群体中等位基因频率从这一代到下一代的改变是以唯一的方式发生的，并且能根据关于初始条件的知识毫不含糊地预测出来。严格说来，此途径仅当两个条件满足时才能应用：(1) 群体在大小上是无限的，和 (2) 环境或者不随时间而变或者按决定性规则而变。这些条件在自然界显然绝不会得到满足，因此，纯决定性途径对于描述群体中等位基因频率随时间的变化可能是不够的。随机的或不可预料的等位基因频率方面的波动也必须被考虑进去。

处理随机波动需要一种不同的数学途径。随机性模型 (stochastic models) 假定等位基因频率的变化按或然性方式发生，即，从这一世代中关于条件的知识我们不能含糊地预料下一世代里的等位基因频率，而只能决定可能出现的某些等位基因频率的出现概率。显然，随机性模型比决定性模型更可取，因为它们建立在更为现实的假定之上。不过，决定性模型在数学处理上要容易得多，而且在某些条件下它们也能得出足够精确的近似结果。以下，我们将按决定性方式处理自然选择。

2.2 自然选择

自然选择 (natural selection) 被定义成：一个群体内遗传上不同的个体或基因型的有差别的增殖。有差别的增殖是由个体间如死亡率、能育性、生殖力、交配成功和后代生活力等因子方面的差异所造成的。

自然选择是个体间在与增殖有关的性状方面存在遗传变异的必然结果。如果一个群体由在这类性状方面相互无差别的个体所组成，则它将不会受到自然选择。选择导致等位基因频率随时间而变。然而，仅仅是等位基因频率从一代到另一代发生变化，并不一定表示自然选择在起作用。别的过程，例如随机遗传漂变，也能导致等位基因频率随时间的改变（见后面的内容）。

基因型的适合度 (fitness), 通常用 w 表示, 是一个关于该个体的生存和增殖能力的尺度。不过, 由于一个群体的大小通常受其所处环境的负载容量限制, 所以, 某一个体的进化成功不是由其绝对适合度 (absolute fitness) 而是由其与群体中其他基因型相比的相对适合度 (relative fitness) 所决定的。在自然界, 不能预期一个基因型的适合度在所有世代和所有环境条件下保持不变。然而, 对每种基因型指定一个恒定的适合度值, 我们就可以得出一些简单的理论公式, 它们对理解由自然选择导致的群体遗传结构变化的动力学是很有用的。在某些最简单的模型里, 我们假定生物的适合度仅由其遗传组成所决定。我们还假定, 所有基因座位对个体的适合度的贡献相互独立, 所以每一基因座位都能被隔离开来处理。

群体中产生的大多数新突变型都会降低其携带者的适合度。这类突变将受到淘汰性选择并且最终将会被从群体中去除。这种类型的选择称负选择或纯洁化选择 (negative or purifying selection)。偶尔, 某一突变可能与群体中最好的等位基因一样合适, 这样的突变即为选择中性的 (neutral), 且其命运将不由选择所决定。极罕见地, 可能会产生一个能给其携带者以选择优势的突变型。这类突变将受到正选择或有利选择 (positive or advantageous selection)。如果这一新突变型仅在杂合子情况下有利, 在纯合子情况下则无优势, 那么结果选择体制将为超显性选择 (overdominant selection)。

下面我们将考虑一个有两个等位基因, A_1 和 A_2 的基因座位的例子。每一等位基因可被指定一个内在适合度值; 它可能是优势的、劣势的, 或中性的。然而, 这种指定只适用于单倍体生物。在二倍体生物中, 适合度最后由该基因座位上的两个等位基因间的相互作用所决定。两个等位基因, 将有三种可能的二倍体基因型: A_1A_1 , A_1A_2 和 A_2A_2 , 它们的适合度则可分别用 w_{11} , w_{12} 和 w_{22} 来表示。

在随机交配下 A_1A_1 , A_1A_2 和 A_2A_2 的频率将分别为 p^2 , $2pq$ 和 q^2 。一个维持着这种基因型比的群体, 被认为是处于哈迪-温伯格平衡 (Hardy-Weinberg equilibrium)。注意, $p = p^2 + \frac{1}{2}(2pq)$ 以及 $q = \frac{1}{2}(2pq) + q^2$ 。

在一般情况下, 三种基因型被指定为以下适合度值和初始频率:

基因型	A_1A_1	A_1A_2	A_2A_2
适合度	w_{11}	w_{12}	w_{22}
频率	p^2	$2pq$	q^2

现在让我们考虑一下等位基因频率随选择而改变的动力学。假定三种基因型的频率及其适合度如上, 则三种基因型 A_1A_1 , A_1A_2 和 A_2A_2 对下一世代的相对贡献, 将分别为 p^2w_{11} , $2pqw_{12}$ 和 q^2w_{22} 。因此, 等位基因 A_2 的频率在下一世代将变成

$$q' = \frac{pqw_{12} + q^2w_{22}}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} \quad (2.1)$$

每世代等位基因 A_2 的频率改变量被表示成 $\Delta q = q' - q$ 。我们可以证明

$$\Delta q = \frac{pq[p(w_{12} - w_{11}) + q(w_{22} - w_{12})]}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} \quad (2.2)$$

下面, 我们假定 A_1 是该群体中的原等位基因或“老”等位基因。然后, 我们来考虑一个新突变型等位基因 A_2 出现后, 等位基因频率变化的动力学。为了数学上的便利, 我们给予 A_1A_1 基因型的适合度值为 1。新产生的基因型 A_1A_2 和 A_2A_2 的适合度将有赖于 A_1 和 A_2 间相互作用的模式。例如, 如果 A_2 对 A_1 是完全显性的, 那么 w_{11} , w_{12} 和 w_{22} 可分别写成 1, $1+s$, 和 $1+s$, 如果 A_2 完全隐性, 则适合度分别为 1, 1, $1+s$ 。这里 s 是带有 A_2 的基因型与 A_1A_1 间的适合度差异。 s 的值为正表示与 A_1A_1 比较

适合度增高,而为负时则表示适合度降低

共显性

在选择共显性模式(codominant mode of selection)或基因选择(genic selection)中,两种纯合子有着不同的适合度值,而杂合子的适合度则为两种纯合基因型的适合度的平均值。三种基因型的相对适合度值可写成:

基因型	A_1A_1	A_1A_2	A_2A_2
适合度	1	$1+s$	$1+2s$

根据等式 2.2,在共显性下我们得等位基因 A_2 的每世代频率改变如下:

$$\Delta q = \frac{spq}{1 + 2spq + 2sq^2} \tag{2.3}$$

图 2-1 显示具 $s=0.01$ 的等位基因 A_2 的频率增加情况。我们看到,共显性选择总是在消耗别的等位基因频率的前提下增加其中一种等位基因的频率,而与等位基因在群体中的相对频率无关。因此,基因选择是一种定向选择(directional selection)。注意,在低频率下对一个共显性等位基因的选择

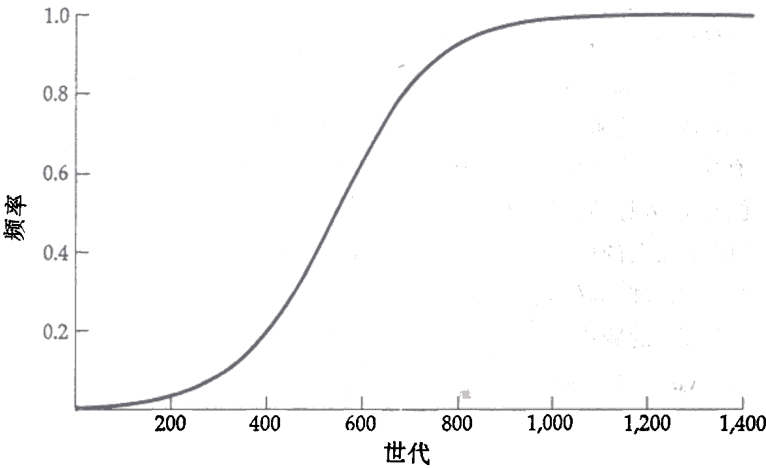


图 2-1 具 $S=0.01$ 的共显性优势等位基因,在其于第 0 代作为突变的结果而出现后的频率。不是十分有效(即,等位基因频率的变化缓慢)。原因在于,当 A_2 频率低时,等位基因 A_2 处于杂合子中的比例较大。例如,当 A_2 的频率为 0.5 时,50%的 A_2 等位基因由杂合子携带,而若 A_2 的频率为 0.01,则 99%的这种等位基因都处于杂合子中。因为杂合子带有两种等位基因,受到的选择压力比纯合子 A_2A_2 的弱(即 s 和 $2s$ 的关系),所以在低 q 值下等位基因频率的总改变量就小。

超显性

在选择的超显性模式(overdominant mode of selection),杂合子有最高的适合度。于是:

基因型	A_1A_1	A_1A_2	A_2A_2
适合度	1	$1+s$	$1+t$

这里 $s>0$ 且 $s>t$ 。根据 A_2A_2 的适合度是高于、等于或低于 A_1A_1 的适合度, t 可相应地取正值、零、或负值。等位基因频率的变化表示成

$$\Delta q = \frac{pq(2sq - tq - s)}{1 + 2spq + tq^2} \tag{2.4}$$

图 2-2 表示一个受超显性选择的等位基因的频率变化。与造成一个等位基因最终从群体中消失的共显性选择体制不同,在超显性选择下,群体迟早将达到一种两等位基因共存的平衡状态。平衡达到后将看不到等位基因频率的进一步变化(即 $\Delta q=0$)。所以,超显性选择属于一种称为平衡选择或稳定化选择(balancing or stabilizing selection)的选择体制。

处于平衡时的等位基因 A₂ 的频率可通过令等式 2.4 的 $\Delta q=0$ 而解出：

$$\hat{q} = \frac{s}{2s-t} \quad (2.5)$$

当 $t=0$ （即两种纯合子有相同的适合度值）时，两种等位基因的平衡频率将都是 50%。

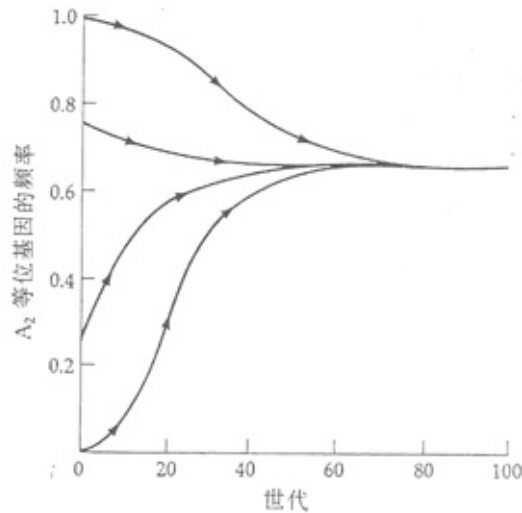


图 2-2 一个受超显性选择的等位基因的频率变化。起始频率从上到下各为：0.99, 0.75, 0.25 和 0.01; $s=0.250$ 及 $t=0.125$ 。因为 s 和 t 的值都特别大，所以等位基因频率的变化迅速。注意，在 $q=0.667$ 处有一个稳定平衡。自 Hartl 和 Clark (1989) 修改而成。

2.3 随机遗传漂变

正如上面提到的，自然选择不是引起等位基因改变的唯一因子。等位基因频率变化也可因机遇而产生，虽然在这种情况下变化不是有方向性的而是随机的。产生等位基因频率的随机波动的一个重要因子是生殖过程中配子的随机取样（图 2-3）。之所以出现取样问题，是因为在自然界中绝大多情况下，任一个世代中可用配子的数目都远大于下一世代所产生的成年个体数。换句话说，只有一小部分配子成功地发育到成体。在一个经受孟德尔式分离的图 2-3 配子的随机取样。假定每一世代的配子库（大长方形）中的等位基因频率精确地反映了亲本世代（小长方形）的成体中的等位基因频率。因为群体大小是有限的，所以等位基因频率上下波动。自 Bodmer 和 Cavalli Sforza (1976) 修改而成两倍体群体中，即使没有配子过剩，即每一个体对下一世代正好贡献两个配子，取样问题也会发生。原因是，杂合子能产生两种配子，而传给下一世代的两个配子由于机遇却可能是同一类型的。

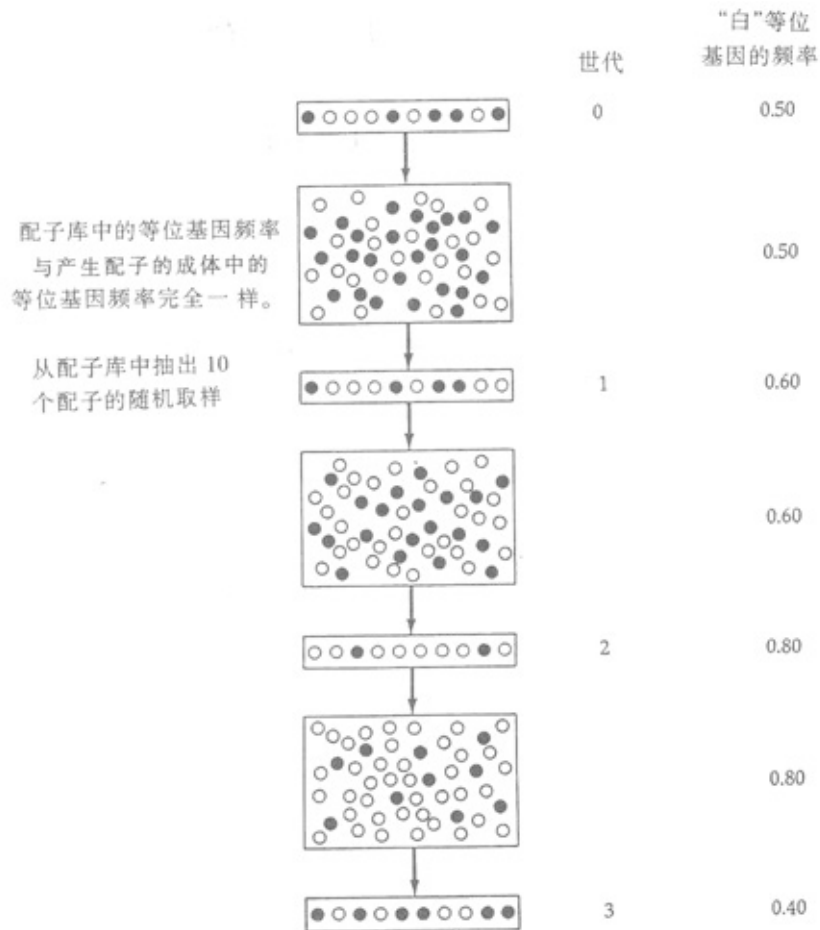
为了看取样造成的后果，让我们考虑一个群体中所有个体有同样的适合度以至于选择不起作用的理想情形。我们再把问题进一步简化，考虑一个没有重叠世代的群体（即一群同时繁殖的个体），这样，任一给定世代都能与前一世代和后一世代毫不含糊地区分开来。该被考虑的群体为两倍体，由 N 个个体所组成，所以，对任一给定的基因座位来说该群体含有 $2N$ 个基因。接下来我们处理具有两个等位基因 A_1 和 A_2 、频率分别为 p 和 $q=1-p$ 的一个基因座位的简单例子。当从无限的配子库中抽取 $2N$ 个配子时，样本中正好含有 i 个 A_1 型基因的概率 p_i 由二项式概率函数给出

$$p_i = \frac{(2N)!}{i!(2N-i)!} p^i q^{2N-i} \quad (2.6)$$

对于两种等位基因共存（即 $0 < p < 1$ ）的群体而言， p_i 总是大于零的，所以，等位基因频率可以无需选择的帮助而一代一代地改变。

只是由于机遇而造成等位基因频率变化的过程称随机遗传漂变（random genetic drift）。不过要注意，随机遗传漂变也可由不是配子取样的过程所引起。例如，选择强度上的随机变化也能导致等位基因频率方面的随机变化。

我们用图 2-4 来说明随机取样对不同大小群体的等位基因频率的影响。等位基因频率一代一代地改变，但变化的方向在任一时间点上都是随机的。随机漂变的最明显特征是，等位基因频率的波动



在小群体中比在大群体中要突出得多。

让我们跟踪由相继世代中随机遗传漂变过程造成的等位基因频率改变的动力学。等位基因 A_1 的频率被写成 $p_0, p_1, p_2, \dots, p_t$, 其中下标指世代数。等位基因 A_1 的起始频率为 p_0 。类似地, 等位基因 A_2 的频率为 $q_0, q_1, q_2, \dots, q_t$ 。在没有选择时, 我们期望 p_1 与 p_0 相等, q_1 与 q_0 相等, 且对以后世代都是如此。然而, 群体是有限的这一事实, 则意味着 p_1 只是在平均上(即重复取样无限次时)才等于 p_0 。而实际情况是, 取样在每一世代里只发生一次, 所以 p_1 通常不同于 p_0 。在第 2 代里, A_1 的频率 (p_2) 将不再依赖于 p_0 而依赖于 p_1 。类似地, 在第 3 代 A_1 的频率 (p_3) 将既不依赖于 p_1 也不依赖于 p_0 , 而是依赖于 p_2 。所以, 随机遗传漂变的最重要的性质是其累积行为 (cumulative behavior), 即随着一代代地传下去, 某一等位基因的频率将倾向于越来越偏离其起始频率。

为了看一看随机遗传漂变的累积效应, 让我们考虑一下下面的数字例子。某一群体由 5 个两倍体个体所组成, 在某一基因座位上的两个等位基因 A_1 和 A_2 , 频率均为 50%。现在我们要问, “在下一世代得到同样的等位基因频率的概率是多少?” 应用等式 2.6, 我们得到值为 0.25 的概率。换句话说, 在 75% 的情况下第 2 代的等位基因频率将不同于起始频率。而且, 在以后的世代里保留着起始等位基因频率的概率也不再是 0.25, 而将会越来越小。例如, 在第 3 代群体中有 50% A_1 的概率约为 18%, 而 10 代以后该概率就只有约 5% (图 2-5)。相应地, A_1 或者 A_2 丢失的概率则随着时间的推移而增加, 因为, 所有被选取的配子正好带有同样的等位基因的事件, 在每一世代中都具有一个有限的概率。一旦某一种等位基因的频率达到 0 或者 1, 它的频率在以后的世代里就不再变化了。第一种情形被称为丢失 (lose) 或灭绝 (extinction), 第二种则称之为固定 (fixation)。如果取样过程持续一个相当长的时期, 则这类可能情形将必然会达到。

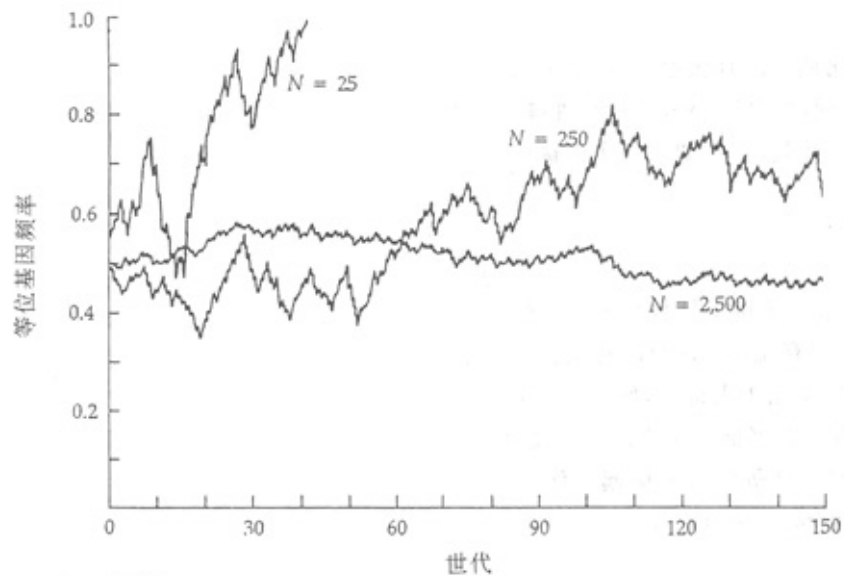


图 2 - 4 不同大小的群体经受随机遗传漂变时等位基因频率的变化。最小的群体 4 2 代后达到固定。另两个群体 1 5 0 代以后还是多态性的，但如果实验继续得足够长则终将会达到固定（等位基因频率 = 0 % 或 1 0 0 %）。自 Bodmer 和 Cavalli Sforza (1976)。

随机遗传漂变的最终结果是一个等位基因的固定以及所有别的基因的丢失。要不发生这样的结果，除非通过象突变或迁移这样的过程使等位基因不断地向群体中输入，或者由某种平衡选择积极地维持着多态现象。

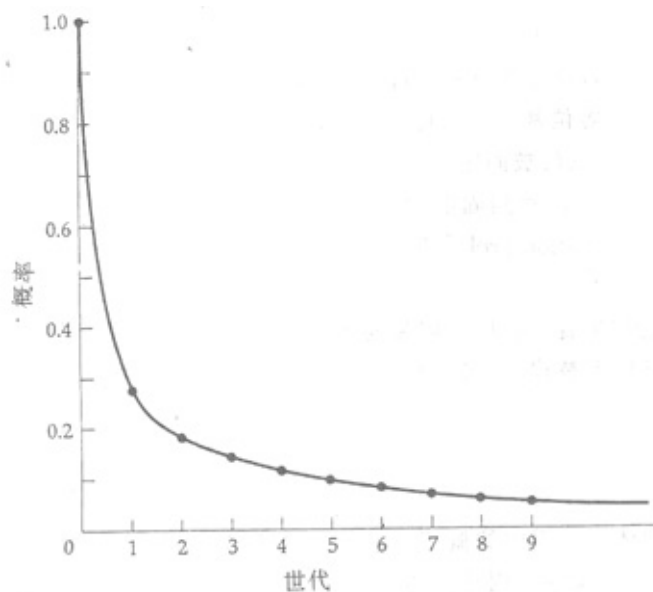


图 2 - 5 两个选择中性的等位基因，有 $N = 5$ 和 $p = 0.5$ ，其维持与起始等位基因频率相同的概率随时间而变的情形。

2.4 有效群体大小

群体生物学的一个基本参数是群体大小 (population size) N ，它被定义成：一个群体中的总个体数。然而，从群体遗传学和进化的观点看，将被考虑的有关数仅由那些积极参与繁殖过程的个体所构成。由于并非所有个体都参与繁殖，所以，影响进化过程的群体大小与调查统计的大小是不同的。群体的这部分称有效群体大小 (effective population size)，并用 N_e 来表示。

一般来说， N_e 小于 N ，有时甚至比 N 要小得多。有各种各样的因素能造成这种差异。例如，在一个具有重叠世代的群体中，在任何时间里，群体的部分将由处于生殖前或生殖后阶段的个体所组成。由于发育阶段重叠，有效大小可能比调查统计大小要小很多。例如，据根井和今泉 (Nei 和 Imaizumi, 1966) 报导，人类的 N_e 仅略大于 $N/3$ 。

有效群体大小比调查统计大小要小，也可能发生在涉及生殖的雄性个体数不等于雌性个体数的时候。这种情况在多配偶制的物种，象社会性哺乳类和领土性鸟类，或者存在无生殖阶层的物种（例如蜜蜂、蚂蚁和白蚁）中尤其明显。如果一个群体由 N_m 个雄体和 N_f 个雌体所组成，则 N_e 将由以下公式给出：

$$N_e = \frac{4N_m N_f}{N_m + N_f} \quad (2.7)$$

注意，除非雌性个体数与雄性个体数相等，否则 N_e 总是要比 N 小。举一个极端的例子，我们假定一个大小为 N 的群体中，所有雌体 ($N/2$) 以及仅有一个雄体参与生殖过程。应用等式 2.7，我们看 $N_e = 2N/(1+N/2)$ 。如果 N 比 1 大得多，则 N_e 将等于 4，而与调查统计到的群体大小无关。

有效群体大小也可能因群体大小的长期变动而大为降低，后者可由象环境的灾变，生殖的循环模式地区性灭绝和重建事件等因素所引起。例如，在一个经历 n 个世代的物种中，长期有效群体大小 (long term effective population size) 由：

$$N_e = \frac{n}{\frac{1}{N_1} + \frac{1}{N_2} + \dots + \frac{1}{N_n}} \quad (2.8)$$

给出，其中 N_i 是第 i 世代的群体大小。换言之， N_e 等于 N_i 值的调和平均，结果它更接近于 N_i 中的最小值而不是最大值。类似地，如果一个群体经历一次瓶颈，则有效群体大小将大为降低。

2.5 基因替换

基因替换 (gene substitution) 被定义成一个过程，经此，一个突变型等位基因将完全代替群体中占优势的或“野生型” (wild type) 等位基因。在此过程中，一个突变型等位基因在群体中产生，通常以一个拷贝出现，并且经一定世代数以后被固定。一个新的等位基因固定所花的时间称固定时间 (fixation time)。不过，并非所有新突变型都能达到固定。事实上，它们中的大多数在几个世代后就会被丢失。所以，我们还需要谈谈固定概率 (fixation probability) 问题，以及讨论影响着一个新突变型等位基因在群体中固定的机会的因素。

新突变在群体内不断地产生着。结果，基因替换不断地发生，用一个等位基因替代另一个，而有时它自己又被一个新的等位基因所替代。所以，我们才能谈到基因替换的速率，即每单位时间的替换或固定数。

固定概率

某一特定等位基因在群体中被固定的概率有赖于：(1) 起始频率，(2) 选择优势度或劣势度 s ，(3) 有效群体大小 N_e 。下面，我们将考虑基因选择的情形，并假定三种基因型 $A_1 A_1$ ， $A_1 A_2$ 和 $A_2 A_2$ 的相对适合度分别为 1， $1+s$ 和 $1+2s$ 。

木村 (Kimura, 1962) 曾证明, A_2 的固定概率由

$$p = \frac{1 - e^{-4N_e q}}{1 - e^{-4N_e s}} \quad (2.9)$$

给出，其中 q 为等位基因 A_2 的起始频率。由于当 x 很小时 $e^{-x} \approx 1-x$ ，所以等式 2.9 在 x 趋近于 0 时简化成 $p \approx q$ 。所以，对于一个中性等位基因而言，固定概率等于它在群体中的频率。例如，一个频率为 40% 的中性等位基因，它将在 40% 的事例中固定而在 60% 的事例中丢失。这可以从直观上认识，因为在中性等位基因的场合，固定是由不偏袒任何等位基因的随机遗传漂变所造成的。

我们注意到，一个新突变型在大小为 N 的两倍体群体中以一个拷贝出现，它将有 $1/(2N)$ 的起始频率。一个突变型等位基因的固定概率 p ，可在等式 2.9 中用 $1/(2N)$ 代替 q 来得出。当 $s \neq 0$ 时，

$$p = \frac{1 - e^{-\frac{2Ns}{N}}}{1 - e^{-4Ns}} \quad (2.10)$$

对于一个中性突变，等式 2.10 变为

$$p = \frac{1}{2N} \quad (2.11)$$

如果群体大小与有效群体大小相同，则等式 2.10 简化成

$$p = \frac{1 - e^{-2s}}{1 - e^{-4Ns}} \quad (2.12)$$

如果 s 的绝对值较小，则我们得

$$p \approx \frac{2s}{1 - e^{-4Ns}} \quad (2.13)$$

对于 s 取正值而 N 的值较大时，等式 2.13 简化为

$$p \approx 2s \quad (2.14)$$

所以，如果一个有利突变产生于一个大群体中且其超越其他等位基因的选择优势度较小，比如说不到 5%，那么其固定的概率将近似等于其选择优势度的两倍。例如，若一个具 $s=0.01$ 的新突变在群体中产生，则其最终固定的概率即为 2%。

现在我们考虑一个数学例子。一个新的突变型在一个有 1000 个个体的群体中产生。如果 (1) 它是中性的，(2) 它给出 0.01 的选择优势度，或 (3) 它具有 0.001 的选择劣势度，那么该等位基因将在群体中固定的概率各为多少？为了使问题简化，我们假定 $N=Ne$ 。对中性场合来说，固定的概率是 $1/(2N)=0.05\%$ 。由等式 2.12，我们得到优势和劣势情况下固定的概率各为 2% 和 0.004%。这些结果是很值得注意的，因为它们从本质表示一个有利突变并不总是会在群体中固定。事实上，具有 $s=0.01$ 的选择优势度的所有突变中，98% 将会因机遇而丢失。此理论结果极为重要，因为它表明，把适应性进化当成有利突变从群体中产生并总是会在以后的世代中兴旺的过程是幼稚的观点。而且，即使是轻微有害的突变也具有一个在群体中固定的有限概率，尽管这是一个小概率。一个有害的等位基因也可能在群体中固定，仅仅这一事实就能有力地说明机遇作用在进化期间对决定突变的命运方面的重要性。]

固定时间

一个等位基因固定或丢失所需要的时间与该等位基因的频率和群体的大小有关。达到固定或丢失的平均时间随着等位基因的频率分别趋近于 1 或 0 而越来越短。

在处理新突变时，对固定和丢失分开处理要方便一些。以下，我们对那些最终将在群体中固定的突变的平均固定时间进行处理。这一变量称为条件固定时间 (conditional fixation time)。在一个两倍体群体中一个新突变的起始频率，据定义为 $q=1/(2N)$ ，此例的平均条件固定时间 $[AKt-]$ ，已由木村和太田 (Kimura 和 Ohta, 1969) 算出。对于一个中性突变而言，它由下式所近似给出

$$\bar{t} = 4N \text{ 代} \quad (2.15)$$

而对于一个具选择优势度 s 的突变来说，它近似为

$$\bar{t} = (2/s) \ln(2N) \text{ 代} \quad (2.16)$$

这里 $\ln$ 表示自然对数函数。

为了说明不同突变类型间的差别，我们假定，某种哺乳动物具有约 10^6 的有效群体大小和 2 年的世代 (间隔) 时间。在这些条件下，一个中性突变在群体中固定平均要花 8 百万年。相比之下，一个具 1% 的选择优势度的突变在同样的群体中固定则只花约 5800 年。有趣的是，一个具选择劣势度 s 的有害等位基因其条件固定时间与具选择优势度 s 的有利等位基因的正好相等 (Maruyama 和 Kimura, 1974)。如果一个有害等位基因被给予高丢失概率，则这一点就可从直观上加以理解。即，对于一个注定要在群体中固定的有害等位基因而言，固定必须发生得非常快。

图 2-6 中我们给出了关于有利的和中性的突变的基因替换动力学过程。我们注意到，有利突变或者迅速丢失或者迅速在群体中固定。与之不同的是，中性等位基因的频率变化则是缓慢的，且固定时间比有利突变型的要长得多。

基因替换速率

我们现在考虑一下替换速率 (rate of substitution), 它被定义为: 每单位时间里达到固定的突变型

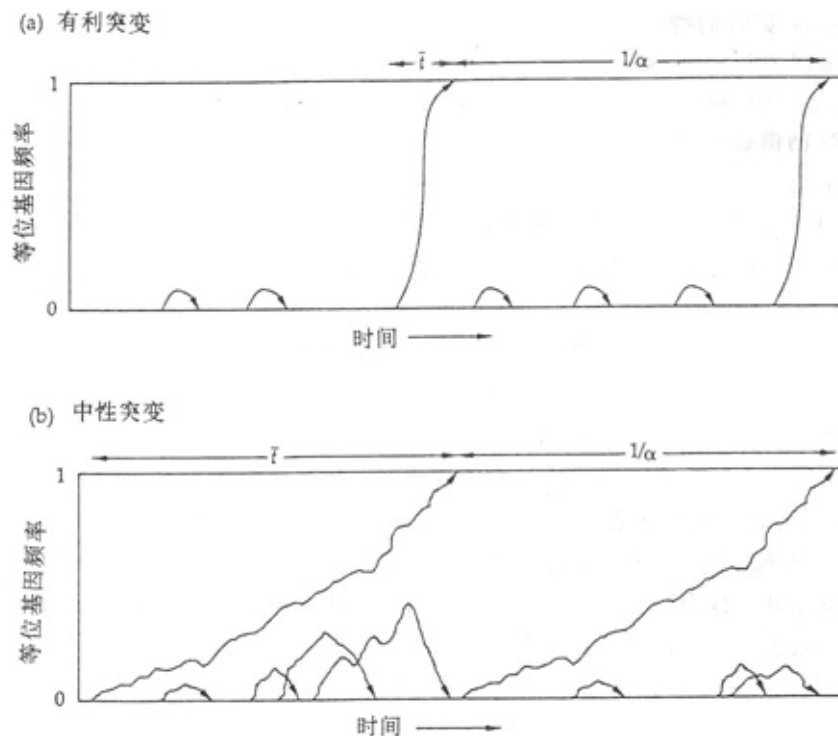


图 2-6 基因替换的动力学过程 (a) 有利的和 (b) 中性的突变。有利突变或者很快从群体中丢失或者迅速被固定, 所以它们对遗传多态性的贡献小。另一方面, 中性等位基因的频率相比之下变化非常缓慢, 所以产生大量的瞬时多态。为条件固定时间, 而 $1/\alpha$ 是接连的两固定事件间的平均时间。自 Nei (1987)。

的数目。我们将先考虑中性突变。如果突变以每世代每基因 u 的速率发生, 那么在一个大小为 N 的两倍体群体中产生的突变型数即为每世代 $2Nu$ 。因为这些突变中每一个的固定概率都是 $1/(2N)$, 所以, 我们用总突变数乘以它们的固定概率即可得到中性等位基因的替换速率:

$$K=u \quad (2.17)$$

于是, 对中性突变来说, 替换速率等于突变率, 这是一个惊人的简单的结果 (Kimura, 1968a)。

如果我们注意到, 在一个大群体中, 每代产生的突变数虽高但每个突变的固定概率却低, 则该结果就可以直观地加以理解。相比之下, 在一个小群体中每代产生的突变数虽低, 但每一突变的固定概率却高。结果是, 中性突变的替换速率与群体大小无关。

对有利突变而言, 替换速率可用突变率乘以如等式 2.14 给出的每基因的固定概率来求得。对于具 $s > 0$ 的基因选择, 我们得:

$$K=4Nsu \quad (2.18)$$

换言之, 对基因选择的情况, 替换速率有赖于群体大小 (N)、选择优势度 (s)、以及突变率 (u)。

2.6 遗传多态性

如果一个基因座位在群体中有两个或更多的等位基因存在, 则它就被说成是多态性的 (polymorphic)。然而, 如果其中一个等位基因有非常高的频率, 比如说 99% 或更高, 那么, 其他等位基因看来都不能在样本中被观察到, 除非样本特别大。所以, 从实际出发, 如果那种最常见的等位基因的频率小于 99%, 则该基因座位通常被定义为多态性的。该定义显然有点武断, 而且在文献中还可发现有人用别的标准。

测度某一群体的多态性程度的最简单方法之一, 是通过将多态性基因座位数用所取样本的基因座位总

数来除，以算出多态性基因座位的平均比例。然而，该测度与被研究的个体数有关。群体内遗传变异性的更为合适的尺度是平均期望杂合度（expected heterozygosity）或基因多样性（gene diversity）。该尺度不依赖多态性的武断描绘，可以从有关基因频率的知识中直接算出，并且受取样作用的影响较小。一个基因座位上的基因多样性定义为

$$h = 1 - \sum_{i=1}^m x_i^2 \quad (2.19)$$

这里 x_i 是等位基因 i 的频率而 m 为该基因座位上等位基因的总数。对任一给定的基因座位而言， h 是随机地从该群体中选取的两个等位基因互不相同的概率。所有被研究过的基因座位的 h 值的平均值， H ，可用作群体内遗传变异性程度的一种估计。

基因多样性尺度 h 和 H ，曾被广泛用于电泳数据和限制酶数据。不过，它们可能不适用于 DNA 顺序数据，因为自然界中 DNA 水平上的遗传变异程度是相当高的。特别地，当所考虑的序列很长时，样本中的每一序列看来都与别的序列有一个或多个核苷酸的差别，在大多数情况下 h 和 H 都将接近于 1。所以，这些基因多样性尺度将无法区别不同的基因座位或群体，也不再是多态性的信息尺度。

对于 DNA 顺序数据，关于群体中多态性的更合适的尺度，是任意两个随机选取的序列间每位点的平均核苷酸差异数。该尺度称核苷酸多样性 (nucleotide diversity) (Nei 和 Li, 1979), 用 π 表示：

$$\pi = \sum_{i,j} x_i x_j \pi_{ij} \quad (2.20)$$

这里 x_i 和 x_j 分别为 DNA 序列的第 i 种和第 j 种类型的频率，而 π_{ij} 是 DNA 序列的第 i 种和第 j 种类型间不同核苷酸的比值。

目前，关于 DNA 序列水平上的核苷酸多样性，仅有为数不多的一些研究。这类研究之一涉及果蝇 *D. melanogaster* 的乙醇脱氢酶 (Adh) 基因座位。跨越 Adh 区域的 11 个序列已由克赖特曼 (Kreitman, 1983) 定序。这些线性排列的序列长达 2379 个核苷酸。如果不管缺失和插入，则有 9 个不同的等位基因，其中一个由 3 个序列表示，而其余的则各由一个序列表示 (图 2-7)。所以，频率 x_1 — x_9 各为 1/11，而频率 x_9 为 3/11。43 个核苷酸位点是多态性的。首先我们计算每对等位基因的不同核苷酸所占的比值。例如，等位基因 1-S 和 2-S 在 2379 个核苷酸中相互有 3 个不同，则 $\pi_{12} = 0.13\%$ 。样本中所有等位基因对的 π_{ij} 值列于表 2-1。由等式 2.20，估出核苷酸多样性为 $\pi = 0.007$ 。被研究的等位基因中 6 个是缓慢移动的电泳变异型 (S)，5 个为快速移动的电泳变异型 (F)。S 和 F 间的区别是由一个氨基酸替换，结果赋予蛋白质以不同的电泳运动性所造成的。对这些电泳类型的每一种，也曾分别地计算了它们的核苷酸多样性。我们得到：关于 S， $\pi = 0.006$ ；关于 F， $\pi = 0.003$ 。这些结果指出，S 等位基因的多态性是 F 等位基因的两倍。

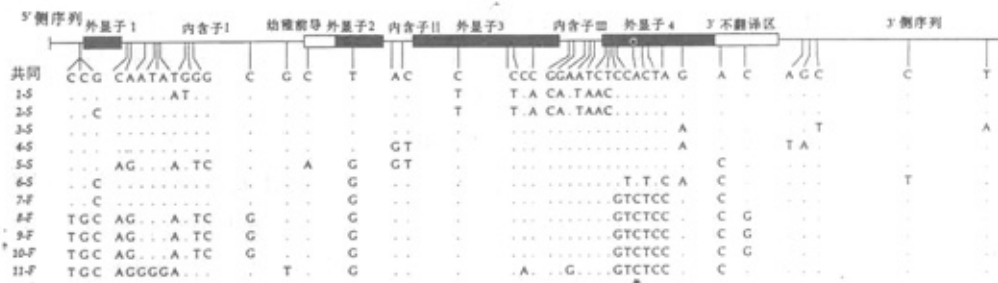


图 2-7 果蝇 *D. melanogaster* 的乙醇脱氢酶基因，其 11 个序列中的多态性核苷酸位点。只有那些与共同顺序有差异的地方被显示出来了。圆点表示与共同顺序相同。外显子 4 中的星号指示赖氨酸对苏氨酸替代的位点，该替代是两电泳型等位基因间出现快/慢运动性的原因。自 Hartl 和 Clark (1989) 修改而成。

表 2-1 果蝇 *D. melanogaster*^a 的醇脱氢酶基因座位的 11 个等位基因中，

成对地比较的百分比核苷酸差异

等位基因	等位基因									
	1-S	2-S	3-S	4-S	5-S	6-S	7-F	8-F	9-F	10-F
1-S										
2-S	0.13									
3-S	0.59	0.55								
4-S	0.67	0.63	0.25							
5-S	0.80	0.84	0.55	0.46						
6-S	0.80	0.67	0.38	0.46	0.59					
7-F	0.84	0.71	0.50	0.59	0.63	0.21				
8-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38			
9-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38	0.00		
10-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38	0.00	0.00	
11-F	1.22	1.18	0.97	1.05	0.84	0.67	0.46	0.42	0.42	0.42

自 Nei(1987)。数据自 Kreitman(1983)

a 加以比较的位点的总数为 2 3 7 9。S 和 F 分别表示缓慢和快速移动的电泳型等位基因

2.7 新达尔文学说与中性突变假说

达尔文提出他的进化学说根据的是自然选择，而没有关于群体中变异的来源的知识。孟德尔定律被重新发现以及遗传变异被证明是由突变产生的之后，达尔文主义和孟德尔主义被用作后来被称为进化的综合学说 (synthetic theory)，或新达尔文主义(neo-Darwinism)的那种理论的框架。根据该学说，虽然突变被看成是遗传变异的根本源，但自然选择却在决定群体的遗传构成，并在基因替换的过程中，被认为起着决定性的作用。

一时，新达尔文主义成了进化生物学中的法则，而选择终于被看成是能驱动进化过程的唯一力量。其他因素，象突变和随机漂变，被认为最多只有次要作用。新达尔文主义的这一特殊标记称为选择主义 (selectionism)。

根据选择论者或进化过程的新达尔文学派的观点，基因替换是以选择对有利突变作用的结果而出现的。另一方面多态性则是由平衡选择来维持的。于是，新达尔文论者把替换和多态性看成是由不同的进化力量驱动的、两种孤立的现象。基因替换是正的适应性过程的最后结果，如果且只有当一个新的等位基因能改善该生物的适合度时，它才会经此过程而占据群体的以后世代；而多态性则是在某一基因座位上两个或更多个等位基因共存对该生物或该群体有利时，才被维持的。新达尔文学派的理论主张自然界中大多数遗传多态性是稳定的。

十九世纪七十年代后期群体遗传学中出现了一次革命。蛋白质顺序数据的应用打破了群体遗传学研究中的物种界线，并且首次为检验与基因替换的过程有关的学说提供了足够的经验数据。1968年，木村提出，进化中大多数的分子变化是由于中性或近中性突变的随机固定所造成的 (Kimura, 1968a 又见;King 和 Jukes, 1969)。这一假说，现在以分子进化的中性学说 (neutral theory of molecular evolution) 而著称，极力主张：分子水平上的大多数进化变化以及物种内的大多数变异性，既不是由有利等位基因的正选择也不是由平衡选择所造成的，而是由选择上呈中性或近中性的突变型等位基因的随机遗传漂变所造成的。从理论上讲，中性并不意味着所有等位基因的适合度完全相等。它的意思只是说，这些等位基因的命运很大程度上是由随机遗传漂变所决定的。换句话说，选择可能会起作用，但它的强度太弱以至于不能抵消机遇作用的影响。要使这种情况成为事实，一个等位基因的选择优势度或劣势度的绝对值必须小于 $1/(2N_e)$ ，其中 N_e 为有效群体大小。

照中性学说的说法，等位基因的频率纯粹由随机规则所决定，而且我们在任一给定时刻得到的画面都只是一种瞬态，它代表取自进行着的动力学过程中的一种暂时构造。因此，多态性基因座位是由或者在走向固定的途中、或者将要灭绝的那些等位基因所组成。从这一场景来看，与进化过程有关的所有分子的表现形式，都应被看成是突变输入和与之相伴的等位基因的随机灭绝或固定的某一连续过程的结果。故而，中性学说把替换和多态性看成是同一现象的两个侧面。替换是一个长期而渐进的过程，

藉此突变型等位基因的频率随机地增加或减少,直到这些等位基因最终因机遇而固定或丢失。在任何给定时间里,某些基因座位所具有的等位基因,其频率将既不是 0%也不是 100%。这些就是多态性的基因座位。中性学说认为,群体中的大多数遗传多态性在自然界中都是瞬时的。

中性论者和选择论者间争论的本质,涉及突变型等位基因的适合度值的分布问题。两种学说都认为,大多数新突变是有害的并且它们很快被从群体中清除,所以,它们对替换速率和群体内的多态性的量都没有什么贡献。不同的是,关于非有害突变中中性突变的相对比例问题。选择论者主张很少的突变是选择中性的,中性论者却认为大多数非有害突变都是有效中性的。

前 20 年时间里关于中性突变假说的激烈争论给分子进化带来了很大影响。首先,它导致了在考虑分子变化的进化动力学时,随机漂变的作用不容忽视,这一点得到普遍承认。第二,分子生物学和群体遗传学间综合,通过分子进化和遗传多态性只是同一现象的两个侧面这一概念的导入,而大大加强(Kimura 和 Ohta, 1971)。虽然争论仍在继续,但任何令人满意的进化学说必须与分子水平上进化过程的这两个方面一致,这一点现在已得到承认。

在一系列研究中,根井等(Nei 等, 1978)曾从这一观点检验过中性突变假说。最近,赫德森等(Hudson 等, 1987)曾提出过一种方法,测试由种间 DNA 顺序比较揭示的高速进化着的 DNA 区域,是否象中性突变假说所预测的那样,在物种内也表现出高水平的多态性。

习题

- 1 从等式 2.2 导出等式 2.3。
2. 如果 A_2 对 A_1 是完全显性,那么等位基因 A_2 的每世代频率改变将是多少?
3. 从等式 2.4 导出等式 2.5 中的平衡频率。
4. 给定一个由 5 个两倍体个体组成的群体,其中 A_1 的频率为 0.5,且 A_1 和 A_2 有同样的适合度,那么,在下一世代 A_1 的频率为 (a) 0.0, (b) 0.5, 或 (c) 1.0 的概率各为多少?
5. 在一个雌体以 2:1 超过雄性的群体中,有效群体大小与调查统计的群体大小的比是多少?
6. 一个经历过瓶颈的群体,比如在连续 6 个世代中其群体大小为: 10^4 、 10^4 、 10^4 、10、 10^4 和 10^4 。其长期群体大小是多少?
- 7 在一个有效群体大小为 100, 且 $N_e=N$ 的群体中,一个具 0.01 的选择劣势度的新突变,其固定概率是多少?
8. 应用图 2-7 中的序列,计算等位基因 (a) 1-S、2-S 或 3-S 间, (b) 1-S 和 7-F 间,编码区(黑长方块)中的核苷酸多样性。该编码区长 771 个核苷酸。

后继阅读文献

Christiansen, F.B. and M.W. Feldman. 1986. Population Genetics. Blackwell Scientific Publications, Cambridge, MA

Crow, J.F. and M. Kimura. 1970. An Introduction to Population Genetics Theory. Harper & Row, New York.

Hartl, D.L. and A.G. Clark. 1989. Principles of Population Genetics. 2nd Ed. Sinauer Associates, Sunderland, MA.

Hedrick, P.W. 1983. Genetics of Populations. Science Books International, Portola Valley, CA.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Nei, M. and D. Graur. 1984. Extent of protein polymorphism and the neutral mutation theory. *Evol. Biol.* 17:73-118.