6 由基因重复和外显子混匀造成的进化

最先注意到基因重复在进化中的重要性的是霍尔丹（Haldane,1932）和马勒(Muller,1935)。他们认为，一个基因的多余复本也许能发生引起岐化的突变，因而最终将会以一个新基因的形式出现。大野（Ohno,1970）以分子的、生物化学的和细胞学的证据为凭，把这种看法引向了极端主张基因重复是唯一能引起新基因产生的途径。虽然，现在已经知道还有一些别的产生新功能的方式（见第 91-92 页），但大野的观点在很大程度上还是成立的。

断裂基因的发现启发了吉尔伯特（Gilbert,1978），于是他提出，内含子间的重组为基因间外显子序列交换提供了一种机制。已经发现的许多这类外显子交换的例子表明，这种机制在真核生物的基因以出现新功能的形式进化中，起着十分显著的作用。

6.1 DNA 重复的类型

一个ＤＮＡ片段的拷贝数增加可由几种类型的ＤＮＡ重复（DNA duplication）所引起。这通常根据所涉及的基因组区域的幅度来分类。已经知道有以下几种类型的重复:（１）部分基因重复或基因内重复（partial or internal gene duplication）,(2)全基因重复(complete gene duplication),(3)部分染色体重复(partial chromosomal duplication), (4)非整数倍重复或染色体重复(aneuploidy or chromosomal duplication),和(5) 多倍性重复或基因组重复(polyploidy or genome duplication) 。前４种类型又称区域性重复（regional duplication）,因为它们影响的不是整个单倍的染色体组。大野(Ohno,1970) 曾极力主张,基因组重复一般要比区域性重复更为重要一些,因此在后一种情况下, 结构基因的调节系统可能只有部分发生了重复，而这种不平衡可能会破坏重复基因的正常功能。然而，正如以下所讨论的，区域性重复显然在进化中也起着非常重要的作用。

ＤＮＡ重复长期以来一直被认为是造成基因组大小进化的一个重要因素（见 Ohno,1970)。特别地，全基因组重复或它的某一主要部分，如一条染色体的重复，可能会造成基因组大小突然而极大的增长。基因组重复事件曾在各种不同的生物类群的进化中反复地被记录到，而在植物、真骨鱼类和两栖类中最为突出。造成基因组扩大的进化途径将在第八章中进行讨论。

6.2 域和外显子

一个蛋白质域（domain）是蛋白质中一个定义明确的区域，它区别于蛋白质中的其它部分，或者执行某一特殊功能，如与基质结合，或者构成该蛋白质内的一个稳定、紧密的结构单位，前者称为功能域（functional domain），后者则称结构域(structural domain)或组件(module)(Go¯和 Nosaka,1987)。定义一个功能域的边界常常是很困难的，因为在许多情况下功能是由散布在整个多肽里的氨基酸残基执行的。另一方面，一个结构组件则是由一段连续的氨基酸片段所构成的。

在考虑产生多重域蛋白质的可能进化机制时，以上区别是相当重要的。如果一个功能域相当于一个组件，那么，它的重复将会增加功能片段的数目。反之，如果功能是由散布在不同组件中的氨基酸残基执行的，则一个组件重复所造成的影响也许从功能上看是不成气候的。在许多蛋白质中看到的内部重复常常对应于结构组件或者单组件的功能域（Barker 等,1978)。

从理论上讲，结构域和外显子在基因中的排列间也许能设想出几种可能的关系（图 6 － 1)。乡(Go¯,1981)发现，在许多内部结构域划分已经确定的球状蛋白质中，基因的外显子和该域之间或多或少

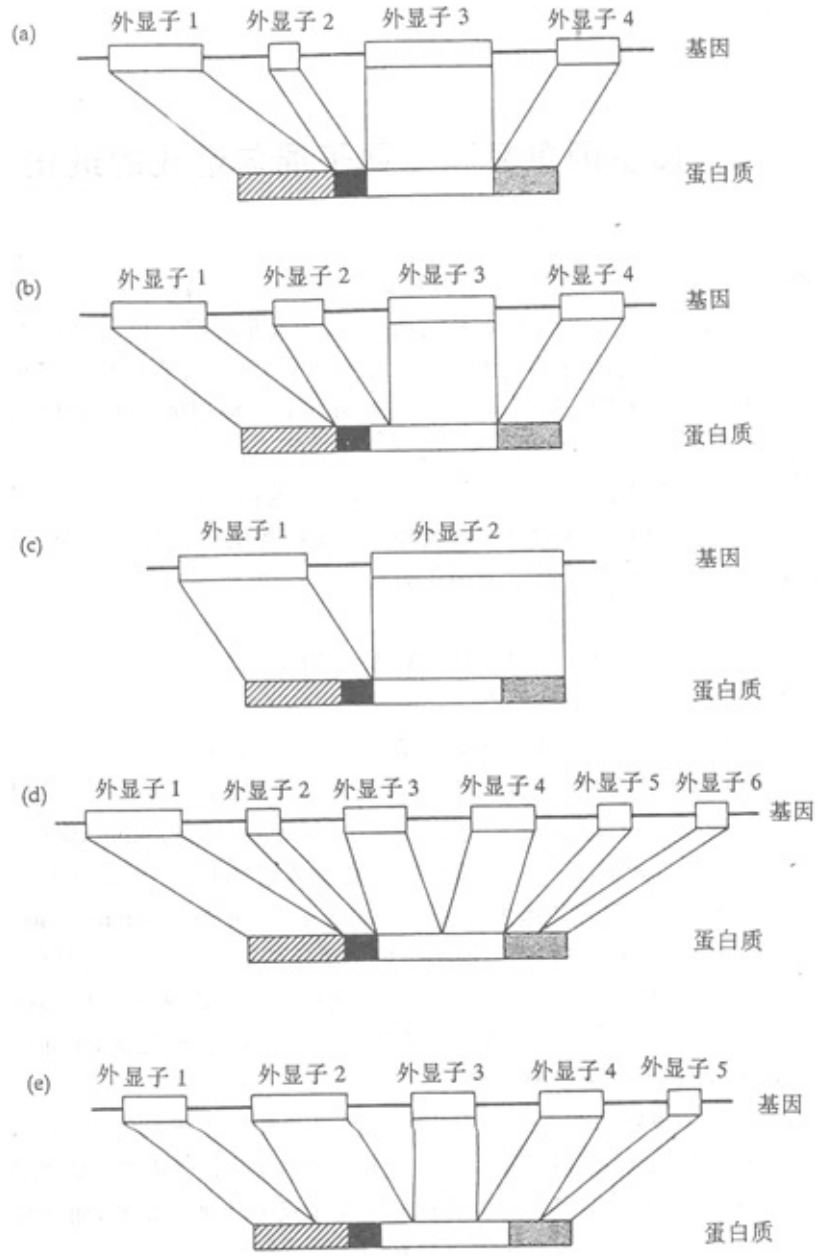


图 6 - 1 基因中外显子的排列与它所编码的蛋白质的结构域之间的可能关系：（ a ）每一外显子正好对应于一个结构域；（ b ）仅近似地对应；（ c ）一个外显子为两个或更多的域编码；（ d ）一个结构域由 2 个或多个外显子编码；和（ e ）外显子和域之间不对应。该蛋白质的 4 个结构域用不同的矩形块（画有斜条纹的，黑色的、白色的和打上点的）表示。

地存在着精确对应（图 6 — 1 a 、 b ）。在某几个例子中，可看到一个组件由一个以上的外显子编码的现象图（ 6 — 1 d ）。在她的研究中，没有发现一个蛋白质的组件结构与其基因的外显子划分间完全不一致的情况（图 6 — 1 e ）。然而，在为数不少的例子中，却可以看到几个邻近的域是由同一个外显子编码的（图 6 — 1 c ）。例如，血红蛋白 α 和 β 分别由 4 个域构成，而它们的基因却只分别由 3 个外显子所组成，其中第 2 个外显子则为 2 个邻近的域编码。乡认为，由于两个外显子间的内含子丢失，结果出现了两外显子的合并。事实上，存在于植物中的同源蛋白质——豆血红蛋白，其基因中就可看到在由珠蛋白的域结构预测的位置处（第 6 8 个氨基酸之后）正好含有一个额外的内含子。所以，珠蛋白基因家族进化期间，有的谱系失去了一些或者全部内含子（图 6 — 2 ）。

在大多数情况下，蛋白质水平上的域重复常指示出在 D N A 水平上出现了外显子重复。所以，它表明外显子重复是内部重复的最重要类型之一。真核生物的基因一般由许多外显子和内含子组成（第一章），而相邻的外显子常常是等同的或相互间非常相似的。这些事实表明，现代生物中许多复合基因

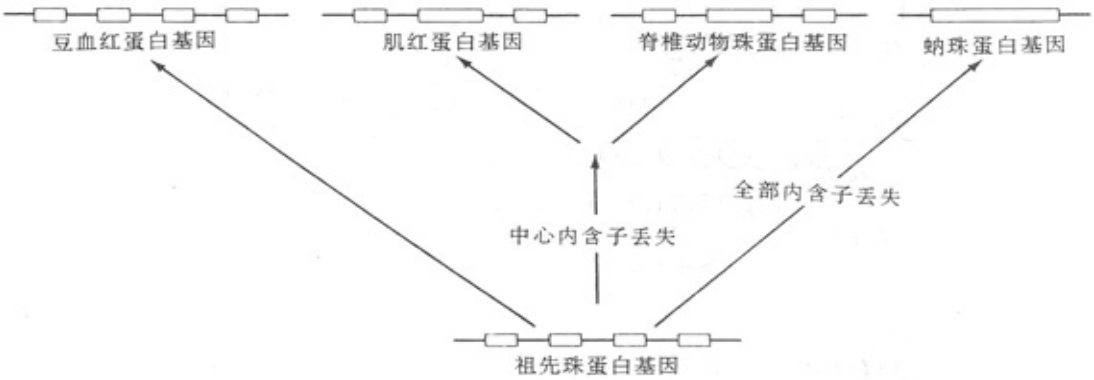


图 6 - 2 珠蛋白基因进化期间的内含子丢失。原始珠蛋白基因有 3 个内含子和 4 个外显子。豆血红蛋白基因保留着祖先结构，而其他谱系则至少丢失了一个内含子。注意，内含子并未按比例大小画。哺乳动物（牛，人，小鼠，猪和海豹）的肌红蛋白基因中的两个内含子长各为~4800bp 和~3400bp，，而珠蛋白和豆血红蛋白基因的同源内含子则分别只有 108 - 192 bp 和 103 - 904 bp 长。豆血红蛋白基因的中间内含子为 99 - 234 bp 长（Blanchetot 等,1983）。珠蛋白的资料来自许多两栖类，鸟类和哺乳类。豆血红蛋白的资料来自 3 种豆类（ Phaseolus vulgaris, Glycine max 和 Vicia faba ）。

是通过原始基因的内部重复和随后的修饰进化而来的。这类原始基因假定只含 1 个或少数几个外显子，且只能执行简单的生物学功能（Li,1983）。

6.3 域重复和基因的延长

对真核生物的现有基因的勘测表明，内部重复在进化中是频繁发生的。这种在基因大小上的增加，或基因的延长（gene elongation），是简单基因向复合基因进化中最重要的步骤之一。理论上基因的延长也可通过其他方式发生。例如，将一个终止密码子转变成一个有意义的密码子的突变也能使基因延长（第一章）。类似地，一个外来 DNA 片段插入某一外显子中，或出现删除拼接位点的突变，也能得到同样的结果。不过，这类分子变化大多数将破坏延长后的基因的功能，因为加进去的区域是由几乎随机排列的氨基酸所构成的。事实上，在绝大多数情况下，这类分子变化是与病理学表象一起而被发现的。例如，异常血红蛋白恒春（Constant Spring）和 Icaria 分别是由将终止密码子变成谷氨酰胺和赖氨酸的突变所引起的。由于这种突变，这些变异型的 α 链上增加了 30 个残基（Weatherall 和 Clegg,1979）。相比之下，一个结构域的重复倾向不会造成这类问题。事实上，这类重复有时甚至能加强新产生的蛋白质的功能，例如，增加活性位点的数目即可达到这一点。

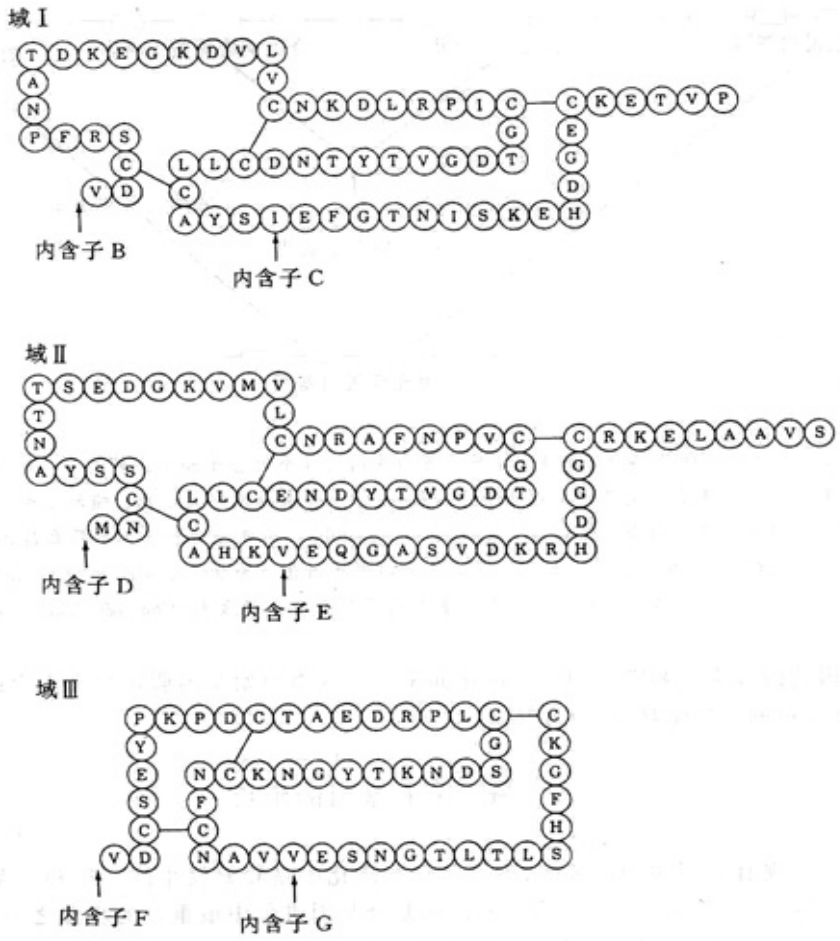
所以，进化期间基因的延长看来主要靠域的重复来实现。在下节中，我们将给出一个基因内重复的例子，以说明进化期间基因延长的后果。

卵类粘蛋白基因

卵类粘蛋白是一种存在于鸟类的卵白中的蛋白质，它能抑制一种催化蛋白质分解的胰蛋白酶的活性。卵类粘蛋白多肽可被划分成 3 个功能域（图 6 — 3 ）。每一个域都能和一个分子的胰蛋白酶或其他丝氨酸类蛋白酶结合。为这 3 个功能域编码的 DNA 区域明显地有着共同的进化起源，且相互间由内含子所隔开（Stein 等,1980）。这 3 个区域中，每一个都是由被一个内含子隔断的两个外显子所构成，且这两个外显子间不表现出相似。于是，卵类粘蛋白基因看来是由一个原始的单域基因经两次内部重复而得来的，其中每次重复都涉及两个邻近的外显子。由于域 I 和 II 相互间比它们中的任一个和域III之间都更为相似，所以它们可能是经第 2 次重复而得到的，而域III则是第 1 次重复的产物。

域重复的普遍性

表 6 — 1 列出了几个证据表明它们在其进史中发生过内部重复的基因名单。这些基因全



域	相似性百分数	
	氨基酸	核苷酸
Ⅰ vs. Ⅱ	46	66
Ⅱ vs. Ⅲ	30	42
Ⅰ vs. Ⅲ	33	50

图 6 - 3 分泌性卵类粘蛋白的 3 个功能域，和域间氨基酸水平与核苷酸水平的顺序相似程度。自 Stein 等，(1980)。

表 6 — 1 具有内部域重复的蛋白质

序　　列	蛋白质的长度[a]	重复的长度	重复的次数	重复的百分比[b]
免疫球蛋白∈-链 C 区(人)	423	108	4	100
免疫球蛋白 γ-链 C 区(人)	329	108	3	98
血清白蛋白(人)	584	195	3	100
小白蛋白(人)	108	39	2	72
蛋白酶抑制因子,Bowman-Birk 型(大豆)	71	28	2	79
蛋白酶抑制因子,颌下腺型(啮齿类)	115	54	2	94
铁氧还蛋白(*Clostridium pasteurianum*)	55	28	2	100
血纤蛋白溶酶原(人)	790	79	5	50
钙依赖性调节蛋白(人)	148	74	2	100
原肌球蛋白 α 链(人)	284	42	7	100

自 Barker 等(1978)　a　氨基酸残基数。　b　由重复顺序占据的部分占蛋白质总长的百分比。

都涉及一个或多个域重复，而其中有些序列则是由一个原始序列经多次重复而得到的，结果使这种重复性结构占据了整个蛋白质的长度。在这些例子中，每一例的重复事件都可从蛋白质或ＤＮＡ顺序的类似而轻易地推测出来。也许还有许多别的复合基因也是经基因内重复而进化的，但它们的重复区域相互间可能已分岐到这样一种程度，以致它们间的顺序同源性已经不能被辨认出来了。在某些情况下，如免疫球蛋白基因的恒定区和可变区，我们可以通过比较这些域的二级结构来推测其共同祖先，因为二级结构有比氨基酸顺序更强的保守性。所以，蛋白质中的内部重复极有可能比经验数据所指示的更多地普遍地存在着。

6.4 基因家族的形成与新功能的获得

一次全基因重复产生两个等同的拷贝。它们将如何进化则会因情况而异。例如，这些拷贝可能保留其原始功能，从而使该生物产生更多的某种ＲＮＡ或蛋白质。此外，其中一个拷贝可能会因一次有害突变而丧失能力，从而变成一个无功能的假基因（见第 85 页）。然而，更重要的是，基因重复可能会导致产生遗传新型或新基因的结果。如果重复中一个拷贝保留其原始功能，而另一个则累积分子变化，以致于最终变得能执行完全不同的功能，那么，产生遗传新型或新基因的情况就会出现了。

重复的基因可以分成两类：变异的重复和不变的重复。不变的重复（invariant repeats）相互间在顺序上是等同的或近似地等同的。在有些情况下表明，等同顺序的重复与某一基因产物的增量合成有关，该产物则是生物的正常功能所必需的。这样的重复称为剂量重复（dose repetitions）。无论何时出现需产生大量特别的ＲＮＡ或蛋白质产物的代谢需要，剂量重复就会十分普遍的出现（Ohno,1970）。代表性的例子有：执行翻译功能不可缺少的 ｒＲＮＡ的基因和 tRNA 的基因，以及染色体首要结构蛋白，组蛋白的基因，因此必须被大量地合成。

变异的重复（variant repeats）由一个基因的多拷贝所构成，虽然这些拷贝相互类似，但在其顺序方面却或多或少地有一定程度的差异。有趣的是，变异的重复有时能执行显然不同的功能。例如，在血液凝结过程中起裂解血纤蛋白原作用的凝血酶，和消化性酶胰蛋白酶，都是源自一个原始基因的重复。类似地，乳清蛋白，催化乳糖合成的酶的一个亚基，和通过裂解某些细菌细胞壁中的多糖成份来溶解它们的溶菌酶，在谱系上是相关的。功能上的分化通常需要大量的替换。不过，在某些情况下，一个新的功能有可能在数目相对较小的替换之后产生（例如，见 Betz 等,1974）。

在一个基因组中属于某一群重复顺序的所有基因，合起来被称为一个基因家族（gene family）或多基因家族（multigene family）。一个基因家族的成员通常位于同一染色体上相互间极靠近的地方。在某些情况下,一些功能性的或非功能性的家族成员可能会位于别的染色体上。

当重复基因在功能或顺序上变得相互间差异很大时，再把它们归成同一基因家族也许就不合适了。戴霍夫(Ｄａｙｈｏｆｆ，１９７８)造了一个词:超家族(ｓｕｐｅｒｆａｍｉｌｙ)来描绘关系密切和关系疏远的蛋白质间的联系。据此，在氨基酸水平上相互至少展示出５０％相似性的蛋白质，可以被看成是一个家族的成员；而当同源蛋白质展示出的相似性小于５０％时则被看成是一个超家族的成员。例如，α－珠蛋白和β－珠蛋白被分类在两个不同的家族中，而它们和肌红蛋白一起则构成了珠蛋白超家族（见第 59 页）。然而，这两术语并不总是能按戴霍夫的标准来严格地应用的。例如，人和鲤的α珠蛋白链仅展现出４６％的顺序相似性，这就低于为归于同一基因家族所划的界限。为此缘故，将蛋白质归类成家族和超家族，不仅要根据顺序的相似性，而且还要考虑关于功能类似性或组织特异性等方面的辅助证据才能决定。

基因家族内的基因数变化极大。有些基因仅在基因组内重复几次，它们被称为轻度重复（lowly repetitive）。另一些则可能在基因组中上百次地重复，因而被称为是高度重复的（highly repetitive）。在以下几节中，ｒＲＮＡ和ｔＲＮＡ基因将被作为例子，以说明高度重复的不变基因。轻度重复的基因将由同功酶和色敏感的色素蛋白基因作代表。

确定ＲＮＡ的基因

表６－２列出了几种有机体的ｒＲＮＡ和ｔＲＮＡ基因的数目。哺乳动物的线粒体基因组只含有一个拷贝的１２Ｓ ｒＲＮＡ基因和一个拷贝的１６Ｓ ｒＲＮＡ基因。这对线粒体翻译系统来说显然是足够了，

因为其基因组仅含１３个为蛋白质编码的基因（第四章）。枝原体是最小的自我复制的原核生物，它含有两组ｒＲＮＡ基因。大肠杆菌的基因组大小是它的４－５倍，含有７组ｒＲＮＡ基因。酵母中ｒＲＮＡ基因的数目大约是１４０，果蝇中和人中的数目则更大。爪蟾 Xenopus laevis 有比人更大的基因组和更多的ｒＲＮＡ基因。所以，ｒＲＮＡ基因的数目与基因组大袖存在着很强的正相关。这一规则对ｔＲＮＡ基因（表６－２）和其他确定ＲＮＡ的基因来说也是成立的。

表６－２ 各种有机体中每单倍体基因组的ｒＲＮＡ和ｔＲＮＡ基因数

基因组来源	rRNA 基因的数目[a]	tRNA 基因的数目	基因组的近似大小(bp)
人线粒体	1	22	1.7×10^4
枝原体 *Mycoplasma capricolum*	2	ND[b]	1×10^6
大肠杆菌 *E. coli*	7	～100	4×10^6
酵母 *Saccharomyces cerevisiae*	～140	320－400	5×10^7
果蝇 *D. melanogaster*	130－250	～750	2×10^8
人	～300	～1300	3×10^9
爪蟾 *Xenopus laevis*	400－600	～7800	8×10^9

自 Li(1983) a．对于ｒＲＮＡ基因，该值指整个ｒＲＮＡ基因组的数。 b．ＮＤ＝未定。

高度重复的基因，象ｒＲＮＡ基因，一般相互间是非常相似的。造成这种同源性的一个因素可能是纯洁化选择，因为这些基因应该遵守非常特殊的功能和结构要求。然而，同源性常常会延伸到没有任何功能意义的区域，而这类同源性的维持就需要求助于别的机制（见第 95 页）了。

同功酶

除了不变的重复以外，高等生物的基因组还含有许多其成员相互间已发生不同程度岐化的多基因家族。其中最能说明问题的例子是为同功酶编码的基因家族，象乳酸脱氢酶、醛缩酶、肌酸激酶和丙酮酸激酶等。同功酶（isozymes） 是催化同样的生化反应、但在组织特异性、生长调控、电泳移动性或生化特性等方面相互间可能有差别的一类酶。注意，同功酶是由不同的基因座位，通常是重复后的基因来编码的，这与异型酶（allozymes）不同，后者是由同一个基因座位上的不同等位基因编码的、某种酶的不同形式。

让我们来考虑为脊椎动物中乳酸脱氢酶（ＬＤＨ）的Ａ和Ｂ亚基编码的两个基因。这两种亚基可形成５种四聚体同功酶：A_4,A_3B,A_2B_2,AB_3和B_4，所有这些酶都可在氧化态辅酶，尼克酰胺腺嘌呤二核苷酸（ＮＡＤ$^+$）的存在下催化将乳酸转变成丙酮酸的反应，或在还原态辅酶（ＮＡＤＨ）的存在下催化相反方向的反应。曾经有人提出，ＬＤＨ－B_4和另外一些富含Ｂ亚基的同功酶对ＮＡＤ$^+$有较高的亲和性，它们在一些行有氧代谢的组织如心脏中，行使着真正的使乳酸脱氢酶的功能，而ＬＤＨ－A_4和另一些富含Ａ亚基的同功酶则对ＮＡＤＨ有较高的亲和性，所以，它们在一些无氧代谢的组织如骨骼肌里，被特别地安排作为丙酮酸还原酶而起作用（Everse 和 Kaplan,1975; Nadal Ginard 和 Markert,1975）。图６－４展示了心脏中产生的ＬＤＨ的发育次序。我们看到，心脏存在的环境越是厌氧，特别是在怀孕的早期阶段，则富含Ａ亚基的ＬＤＨ同功酶所占的比例就越高。于是，这两个重复基因已对不同的组织和不同的发育阶段发生了特化。因为这两种亚基存在于几乎所有曾做过年代测定研究的脊椎动物中，所以，产生ＬＤＨ－Ａ和ＬＤＨ－Ｂ的基因重复可能发生在脊椎动物进化的早期阶段之前或期间。ＬＤＨ的一个有趣特性是，这两个亚基能形成异源多聚体，从而进一步增加了该酶的生理学功能多样性。

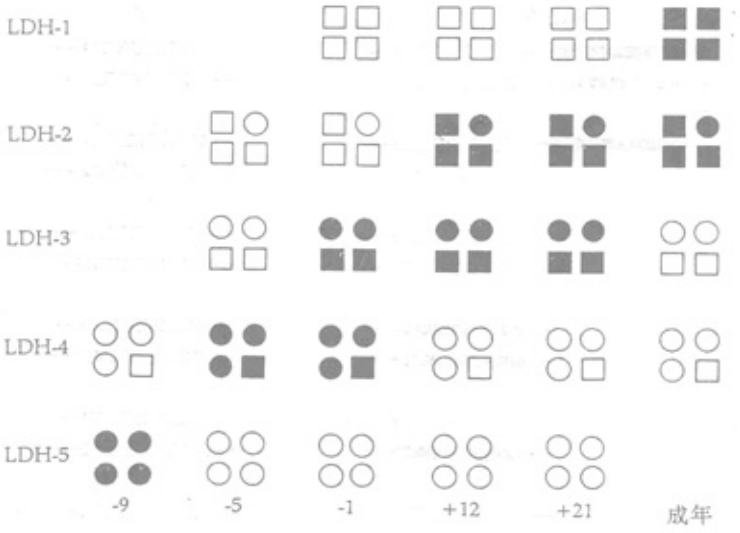


图 6 - 4 心脏中 5 种乳酸脱氢酶（ＬＤＨ）同功酶的发育次序。负数和正数分别表示出生前、后的天数。方块表示Ｂ亚基、圆圈表示Ａ亚基。被涂黑了的图形表示在数量上占优势的形式。注意在个体发育期间从Ａ亚基向Ｂ亚基的转移。资料取自 Markert 和 Vrsprung(1971)。

色敏感色素蛋白

人、猿和古世界猴具有 3 种色敏感色素蛋白。蓝色素由一个常染色体基因编码，而红色素和绿色素则各由一个 X-连锁基因编码（Nathans 等,1986)。红和绿色素的氨基酸顺序有９６％是等同的，但它们与蓝色素的相似性却只有４３％。蓝色素基因和绿、红色素基因的祖先在大约 5 亿年前发生分岐。相比之下，红和绿色素间的紧密连锁和高度的同源性指出，它们来源于非常近期的基因重复。因为新世界猴只有一个 X-连锁的色素基因，而古世界猴和人则有 2 个或多个色素基因，所以可以假定，重复发生在大约３５００－４０００万年以前、古世界猴与新世界猴分岐以后的祖先中。作为该重复的结果，人、猿和古世界猴能分辨 3 种颜色（即它们是三色性的)，而新世界猴，如松鼠猴，则只能对蓝与绿或蓝与红加以区别，但不能对绿与红加以区别（即它们是二色性的)。

有趣的是，两个 X-连锁的等位基因呈杂合状态的雌性松鼠猴是三色性的（Jacobs 和 Neitz,1986)。另一方面，只携带一条 X 染色体的雄体则从未获得过三色性视觉。于是，在人和古世界猴的场合，三色性视觉是通过类似于同功酶的机制（即，两个有区别的蛋白质由不同的基因座位编码）而获得的，而杂合体雌性松鼠猴达到同样目的，则是通过应用两种异型酶（即，同一基因座位上两个不同的等位基因形式）来实现的（图６－５)。如果三色性视觉能给予其携带者以选择优势，那么，新世界猴的一基因座位上的两个色敏感等位基因的长期维持，可能是通过一种超显性选择形式来实现的（第二章)。

6.5 重复基因的无功能化

多余的重复基因更可能是变成无功能基因、而不是进化成一个新基因，因为有害突变远比有利突变发生得频繁。一个重复基因的无功能化即产生一个假基因。这样产生的假基因称未加工的（unprocessed）假基因，这与将在第七章中讨论的经过加工的假基因相反。表６－３列出了在几种珠蛋白假基因中发现的结构缺陷。这些未加工的假基因大多数含有多重缺陷，象阅读框架移动、成熟终止前终止、和拼接位点或调控位点的删除等，以至很难看出哪种突变是使基因沉默的直接原因。在有几个例子里，也许能找到“元凶”。例如，人的ψζ只含有一个严重缺陷，无义突变，所以它可能是无功能化的直接原因。(符号ψ用来将假基因与其有功能的对应物加以区别。）有些假基因，象在山羊β－珠蛋白多基因家族中的ψβ^{x}和ψβ^{z}，则是由一个预先存在的假基因重复而得来的。

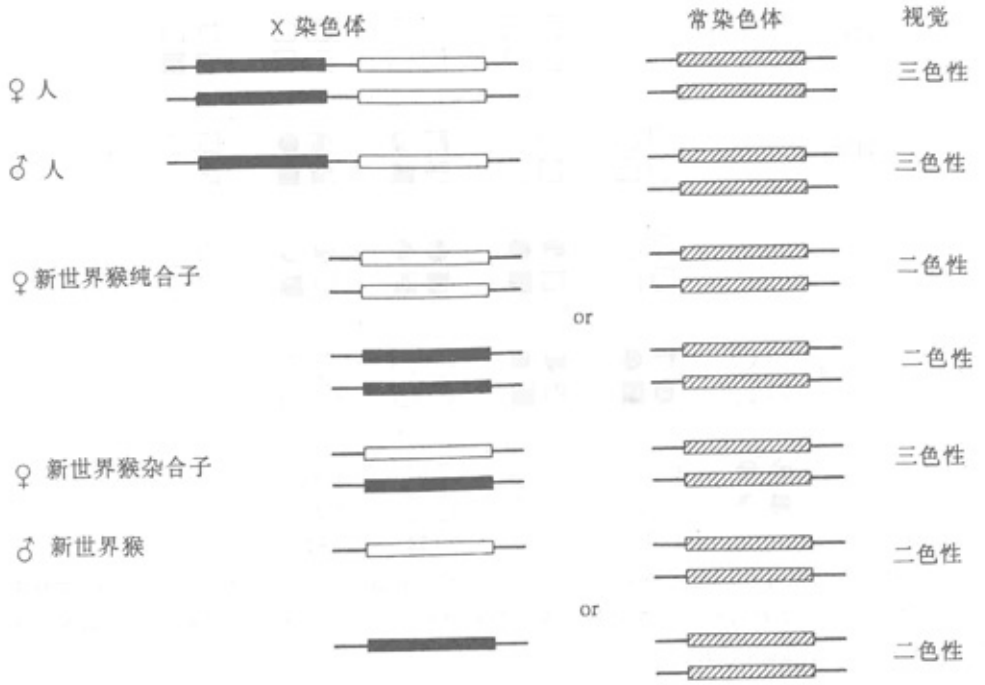


图 6 - 5 人和新世界猴（NWM）的雄性和雌性中，三色性视觉的分子基础。注意，雄性新世界猴不能获得三色性视觉。涂黑了的、空心的和画斜线的矩形分别表示绿、红和蓝色素基因。

表 6 — 3 珠蛋白假基因中的缺陷[a]

假基因	TATA 框	起始密码子	阅读框架移动	成熟终止前终止	缺乏必需氨基酸	拼接 GT/AG 规则	改变了的终止密码子	多聚腺苷化信号 AATAAA
人 ψα1		+	+	+	+	+	+	+
人 ψζ1				+				
小鼠 ψα3	+		+	+		+		
小鼠 ψα4			+		+			
小鼠 ψβh3	?	+	+	+	+	+	?	?
山羊 ψβ^a	+		+	+	+	+	+	+
山羊 ψβ^b	+		+	+	+	+	+	+
兔 ψβ2			+	+	+	+		

自 Li(1983) a、加号表示存在一种特别类型的缺陷；问号表示有存在该缺陷的可能性。

6.6 基因重复的年代测定

两个基因，如果它们是从一次重复事件中得来的则称为平行相关的（paralogous），如果它们是从一次物种形成事件中得来的则称为垂直相关的（orthologous）。例如在图 6 — 6 中，基因 α 和 β 是从一个祖先基因的重复中得到的，因而是平行相关的，而来自物种 1 的基因 α 和来自物种 2 的基因 α 则是垂直相关的，来自物种 1 的 β 基因和来自物种 2 的 β 基因间的关系也是如此。

如果我们知道基因 α 和基因 β 中的替换速率，则我们即可从序列资料中估出重复的年代，即 T_D。而替换速率则可根据垂直相关的基因间的替换数，结合有关物种 1 和物种 2 间分岐时间 T_S 的知识（图 6 — 6 ）来估出。下面我们将说明 T_D 的估值是怎样得到的。

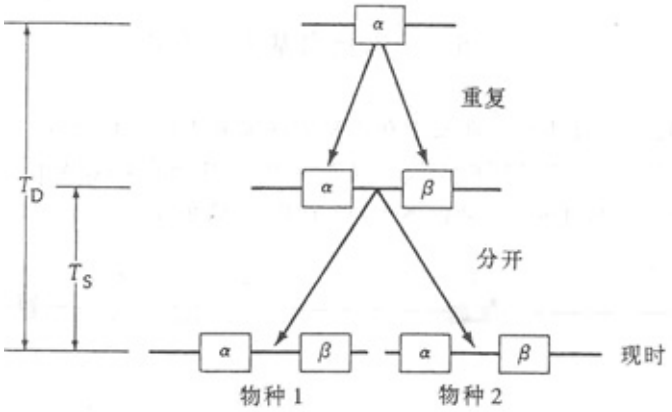


图 6 - 6 用来估计一个基因重复事件的时间（T_D）的模型。α和β这两个基因来自 T_D 单位时间以前一个祖先物种中发生的重复事件。后来该物种分裂成两个物种，1 和 2，这发生在 T_S 时间单位以前。在物种 1 和物种 2 中的两个α基因是垂直相关的，两个β基因也是如此，但α基因和β基因间则是平行相关的。

对于基因α，设 K_α 为这两个物种间的每位点替换数。那么，在基因α中的替换速率则由

$$r_\alpha = K_\alpha / 2T_S \qquad (6.\ 1)$$

来估计。在基因β中的替换速率，r β可用同样方式得到。这两个基因的平均替换速率则为：

$$r = r_\alpha + r_\beta / 2 \qquad (6\ .2)$$

为了估出 TD，我们需要知道基因α和β间的每位点替换数（K α β）。这个数可从以下 4 个对子的比较中得到：（1）来自物种 1 的基因α和来自物种 2 的基因β，（2）来自物种 2 的基因α和来自物种 1 的基因β，（3）来自物种 1 的两个基因，（4）来自物种 2 的两个基因。从这 4 个估值我们能算出 K α β 的平均值（K α β），进而我们能估出 TD 为：

$$T_D = K_{\alpha\beta} / 2r \qquad (6.\ 3)$$

注意，在蛋白质编码基因的情况下，分别用同义替换数和非同义替换数，我们能得到两个相互独立的 TD 估值。这两个估值的平均值也许能作为 TD 的最后估值来使用。然而，如果基因α和β间每同义位点的替换数太大，比如说大于 1，则同义替换数就不能被精确地估出，这样同义替换也许不能提供一个可靠的 T_D 估值。在这种情况下，将只有非同义替换数得到应用。反之，如果这种平行相关的基因间每非同义位点的替换数太小，那么，非同义替换数的估值将承受较大的取样误差，在这种情况下，就只应该用同义替换数了。

以上我们是在速率恒定的假定下进行的。此假定可用以上提及的 4 个对子的比较来加以检验。如果这 4 个对子的比较中近似相等性不成立，则该假定也不能成立。如后面（见第 95 页）将要讨论的那样，起因于具体的进化事件的一些问题也可能会产生，并使 TD 的估计复杂化。

测定基因重复事件年代的另一种方法是，结合有关被研究物种分岐年代的古生物学资料，来考虑基因在系统发育中的分布。例如，除无颌鱼（Agnatha）外所有脊椎动物都编码α和β珠蛋白链。对此观察到的事件有两种可能的解释。一种是：产生α和β珠蛋白的重复事件发生在无颌类与其他脊椎动物的共同祖先中，但后来所有无颌类都丢失了其中的一个重复。这是有可能的但可能性不大，因为这样一种设想需要在许多进化谱系中发生的这种丢失是非独立的(即与物种有关)。另一种解释是：重复事件发生在无颌类与其他所有脊椎动物的祖先分岐之后，但在其他脊椎动物幅射演化之前（4.5-5 亿年前）。后一种解释想来要更合理一些，所以重复的年代普遍取 4.5-5 亿年前（Dayhoff，1972;Dickerson 和 Geis,1983）。

显然，以上方法只能粗略给我们提供重复年代的估值，所以，对所有估值都应该小心采用。

6.7 珠蛋白基因超家族

珠蛋白超家族曾经历过所有可能发生在重复顺序家族中的进化路线：（1）原始功能的保留，（2）新功能的获得，和（3）某些重复中功能的丧失。对人类而言，珠蛋白超家族由3个家族构成：肌红蛋白家族，它的唯一的一个成员位于第22染色体上，位于第16染色体上的α—珠蛋白家族，和位于第

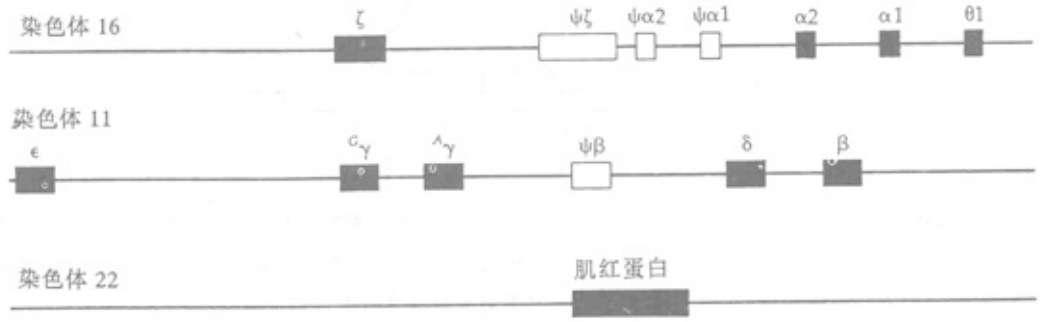


图6－7 人的珠蛋白基因超家族的3个基因家族的染色体排列：α－珠蛋白家族在第16染色体上，β－珠蛋白家族在第11染色体上，而肌红蛋白则在第22染色体上。涂黑了的矩形块表示有功能的基因，空心矩形块则表示假基因。

11染色体上的β—珠蛋白家族（图6—7）。这3个家族合在一起产生两种功能蛋白质：肌红蛋白和血红蛋白。这两种蛋白质在约6—8亿年前发生分岐（见图6—8；Dayhoff,1972;Doolittle,1987），并已经在某些方面发生了特化。从组织特异性角度看，肌红蛋白变成了肌肉中的储氧蛋白质，而血红蛋白则变成了血液中氧的运输员。就四级结构而言，肌红蛋白保留着单体性结构，而血红蛋白则变成了四聚体。从功能上看，肌红蛋白已进化为有比血红蛋白更高的氧亲合力，而血红蛋白的功能则变得更加精密而可调节（见 Stryer,1988）。例如，哺乳类的血红蛋白，有根据血液中有机磷酸盐的水平来调节其对氧的亲合性的能力。显然，异源多聚体结构有助于血红蛋白功能的精密化。

人和绝大多数脊椎动物的血红蛋白是由两种链所构成，一种由α家族的成员编码，另一种则由β家族的成员编码。如以上所讨论的，α家族和β家族是在约4.5-5亿年前发生分岐的(图 6-8)。由于无颌鱼只具有一种单体的血红蛋白，所以，脊椎动物血红蛋白的多聚体化一定是在靠近α—β分岐的时刻出现的。

在人类中，α家族由4个功能基因，即ζ，α_1、α_2和最近发现的θ1所构成（图6—7）。它还含有3个假基因：ψζ，ψα1和ψα2。β家族由5个功能基因：即∈，$^{G}\gamma$，$^{A}\gamma$，β和δ，以及一个假基因ψβ所组成。这两上家族已经在生理学特性和个体发育调节等两个方面都发生了分岐。事实上，在不同的发育阶段上有不同的珠蛋白出现；胚胎期为$\zeta_2\in_2$和$\alpha_2\in_2$,胎儿期为$\alpha_2\gamma_2$，而在成年期则为$\alpha_2\beta_2$和$\alpha_2\delta_2$；θ_1在何时表达尚且不知。而且，与氧结合的亲合性方面的差别，也在这些珠蛋白中发生了进化。例如，胎儿血红蛋白$\alpha_2\gamma_2$有比任何一种成人血红蛋白（$\alpha_2\beta_2$和$\alpha_2\delta_2$）都高的氧亲合力，因而能在处于相对缺氧（低氧）环境的胎儿中更好地行使功能（Wood 等,1977）。这一现象再次为这样一个事实作出了例证，即基因重复能导至生理系统精细化的结果。

在α家族的成员中，胚胎型ζ是分岐程度最高的，在3亿多年以前就已经分枝了（图6—8）。θ1珠蛋白在大约2.6亿年前发生分枝。由于两个α基因间的分岐时间还未确定，所以该图中只画出了α_1基因。α_1基因和α_2基因有几乎等同的ＤＮＡ顺序，且产生同样的多肽。看起来这好象表示它们的分岐时间离现在非常近。然而，这种类似性也可能是趋同进化的结果（Zimmer 等,1980），这一现象我们放在后面讨论（见第 101-102 页）。这两个基因存在于人类和所有猿类中，所以可能是在２０００多万年以前产生的。

在β家族的成员中，成年型（β和δ）与非成年型（γ和∈）大约在1.55-2亿年前分岐(Efstratiadis 等,1980)。两个γ基因的祖先大约在1—1.4亿年以前与∈基因分岐。产生$^{G}\gamma$和$^{A}\gamma$的重复在大约3500万年前，人谱系与新世界猴谱系分开之后出现（Shen 等,1981）。δ基因和β基因间的分岐以前估计是在

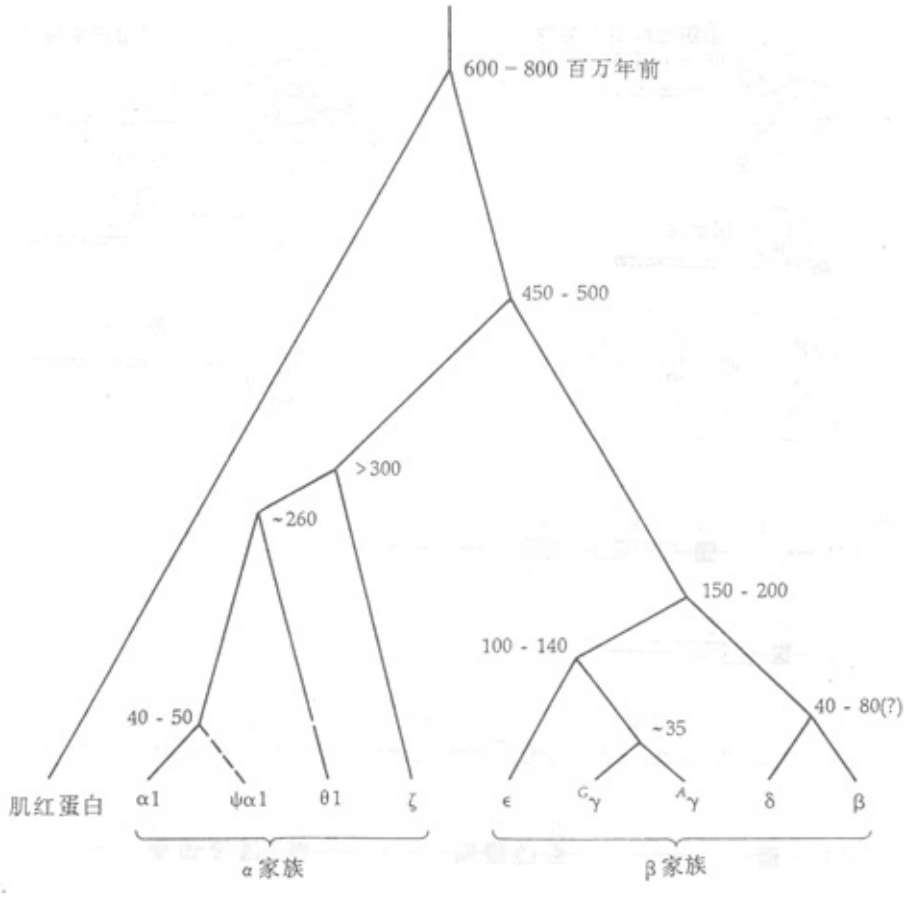


图 6 - 8 人的珠蛋白基因的进化史。虚线表示一个假基因谱系。图中只标出了两个α－珠蛋白基因中的一个，因为它们相互间发生分歧的年代还未确定。

4000 万年前（Dayhoff,1972;Efstratiadis 等,1980），但最近ＤＮＡ顺序资料表明，它可能早于真兽类的辐射，即在约 8000 万年前出现（Hardison 和 Margot,1984;Goodman 等,1984）。从以上讨论中我们注意到，在两个家族中，基因间分岐时间与基因间功能或调节方面的分岐程度之间，存在着明显的相关。

6.8 外显子混匀

有两类外显子混匀（exon shuffling）：外显子重复和外显子插入。外显子重复指一个基因中的一个或多个外显子的重复，所以它是一种内部重复，这已在基因的延伸一节中讨论过（见第 81 页）。外显子插入是这样一种过程，通过该过程结构域或功能域在蛋白质之间发生交换，或者被插入一个蛋白质之中。这两类混匀都曾在产生新基因的进化过程中被采用。这里，我们将讨论一个外显子从一个基因插入另一个基因，结果产生镶嵌或嵌合蛋白质的情况（Doolittle,1985;Patthy,1985）。

镶嵌蛋白质

第一个被发现的镶嵌蛋白质是组织血纤蛋白溶酶原活化因子（ＴＰＡ）（图６—９）。血纤蛋白溶酶原经ＴＰＡ作用转化成它的活化形式：血纤蛋白溶酶，后者则将血纤蛋白、血块中的一种可溶性的纤维状蛋白质溶解。在底物血纤蛋白的存在下，血纤蛋白溶酶原转化成血纤蛋白溶酶的过程将被大大加速。血纤蛋白能与血纤蛋白溶酶原和ＴＰＡ两者结合，从而将它们联系起来而起催化作用。这种分子排列方式允许血纤蛋白溶酶仅以非常接近血纤蛋白的形式产生，从而给予血纤蛋白溶酶原以对血纤蛋白的特异性。相比之下，尿激酶（ＵＫ），一种尿液中的血纤蛋白溶酶原活化因子，则缺乏血纤蛋白特异性。对ＴＰＡ和ＵＫ的前体尿激酶原的氨基酸顺序比较表明，ＴＰＡ在其氨基末端含有４３个残基序列，而ＵＫ中却没有相应

的对应物（Banyai 等,1983)。这一片段能形成一种手指样结构（图６－９ａ），而它和另一种蛋白质的与血纤蛋白亲合性有关的指状域是同源的。后一种蛋白质是一种存在于血浆中或细胞表

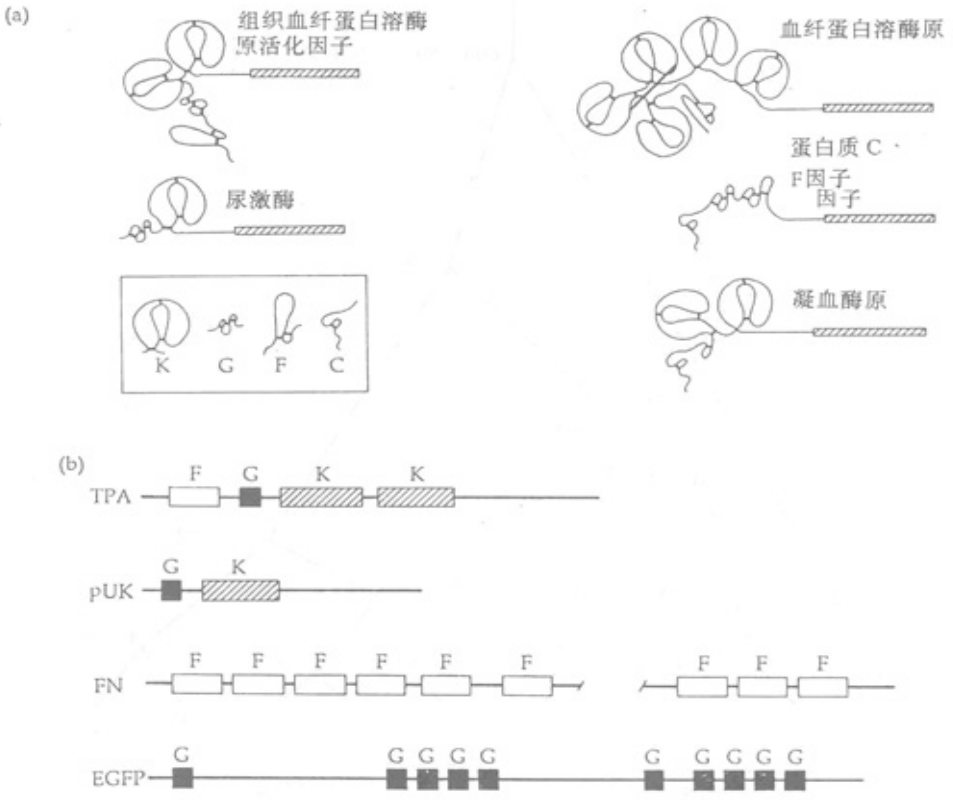


图６－９ （ａ）组织血纤蛋白溶酶原活化因子（ＴＰＡ）与其他涉及血液凝结和血纤蛋白溶解的蛋白质中观察到的各种结构组件。方框插图显示了非蛋白酶区域中组件的结构：Ｋ，纽结杆状组件；Ｇ，生长因子组件；Ｆ，指状组件；Ｃ，依赖维生素Ｋ的钙结合组件。画有斜线的杆状图形表示与胰蛋白酶同源的蛋白酶区域。自 Patthy(1985)修改而成。（ｂ）通过外显子插入组织血纤蛋白溶酶原活化因子（ＴＰＡ）蛋白质而获得的组件的起源。pUK，尿激酶原，EGFP，表皮生长因子前体；ＦＮ，纤粘连蛋白（fibronectin)。

面上的、起促进细胞粘着作用的大分子糖蛋白称纤粘连蛋白（ＦＮ）（图６－９ｂ）。若指状域片段缺失会导至ＴＰＡ的血纤蛋白亲合性丧失。ＴＰＡ与ＦＮ的同源性限制在这种指状域内。所以，外显子混匀必定是与ＴＰＡ从ＦＮ或某一类似蛋白质那里获得该域的事件有关的。

ＴＰＡ还含有一个与表皮生长因子（ＥＧＦ）和另一些蛋白质的类生长因子区域同源的片段。另一些蛋白质是指象因子Ⅸ和因子Ⅹ之类的蛋白质，它们是在血液凝结过程中促使血液结块的酶。此外，ＴＰＡ的羧基端区域与胰蛋白酶和其他胰蛋白酶样的丝氨酸蛋白酶的酶蛋白部分同源，这些酶都起着将蛋白质水解成多肽片段的作用。最后，ＴＰＡ的非酶蛋白部分含有两个类似于血纤蛋白溶酶原的纽结杆(Kringle)的结构。(一个纽结杆是一个富含半胱氨酸的序列,序列中含有３个内部二硫桥并形成一种有点象丹麦蛋糕的纽结杆状结构,这种丹麦蛋糕的名字是(Kringle)。所以，在ＴＰＡ的进化期间，它至少从其他３个基因：血纤蛋白溶酶原，表皮生长因子和纤粘连蛋白那里，夺取了至少４个ＤＮＡ片段（图６－９ｂ）。而且，这些得来的单位的连结处和外显子与内含子间的边界精确地一致，于是，就使得关于外显子事实上曾从一个基因转移到另一个基因的想法更加可信。有关外显子混匀的更多例子，可见 Doolittle(1985)和 Patthy(1985)。

外显子混匀上的相位限制

为了使一个插入某一基因的内含子中的外显子不至于引起阅读框架上的框架移动，接受基因时的相位限制就必须遵守。为了把这种限制搞清楚，让我们根据内含子相对于编码区的可能位置来将内含子归于不同的类型。位于两编码区之间的内含子，根据编码区被打断的方式可被分成３种类型。如果内含子处在两

个密码子之间，则它是具有相位 0 的内含子；如果它位于一个密码子的第 1 和第 2 个核苷酸之间，则它具有相位 1；如果它位于一个密码子的第 2 和第 3 个核苷酸之间，则它具有相位 2（图 6－10）。

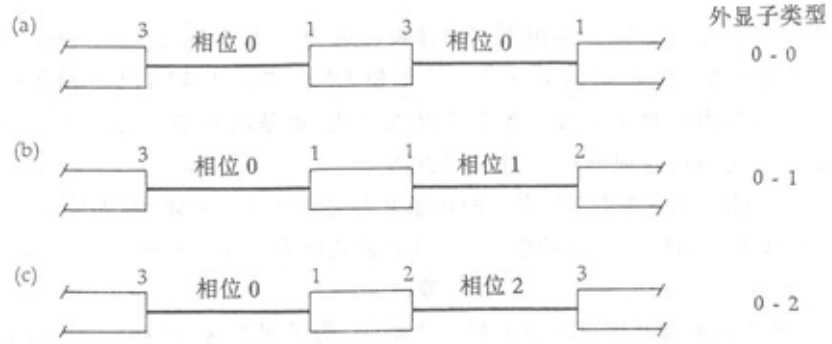


图 6－10 内含子的相位和外显子的类型。外显子用矩形块表示。外显子—内含子连结处上的数字指示外显子的最后一个核苷酸的密码子位置，内含子—外显子连结处上的数字则指示外显子的第一个核苷酸的密码子位置。9 种可能的外显子类型中只有 3 种在图中表示出来了。

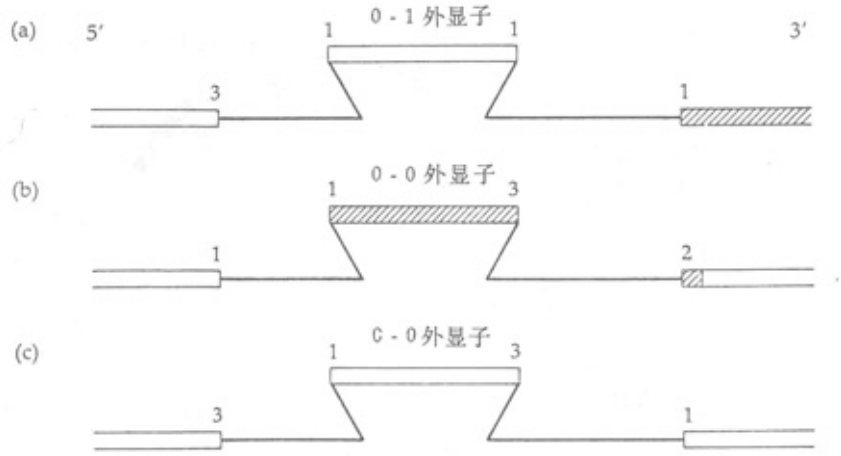


图 6－11 外显子插入内含子的后果。画有斜条纹的矩形块指示阅读框架移动。（a）一个 0－1 不对称外显子插入一个相位－0 内含子；（b）一个 0－0 对称外显子插入一个相位－1 内含子；（c）一个 0－0 对称外显子插入一个相位－0 内含子。（a）和（b）中的插入为不成功的插入。

外显子则根据其两侧的内含子而组合归类。例如，图 6－10 b 中处在中间的外显子，其 5′ 端与相位－0 内含子相邻，其 3′ 端与相位－1 内含子相邻，因而说它是 0－1 类型。两端由同样相位的内含子包围的外显子称为对称的外显子（symmetrical exon），否则即为非对称的（asymmetrical）。例如，图 6－10 a 中处在中间的外显子是对称的。在 9 种可能的外显子类型中，3 种是对称的（0－0，1－1 和 2－2），6 种是非对称的。

只有对称的外显子才能被插入内含子中。例如，图 6－11 a 中，一个 0－1 外显子插入一个相位－0 内含子，结果造成后面所有外显子的阅读框架移动。而且，对称外显子的插入也是有限制的；一个 0－0 外显子只能插入相位为 0 的内含子，类似地，一个 1－1 外显子只能插入相位为 1 的内含子，一个 2－2 外显子也只能插入相位为 2 的内含子，以图 6—11b 的情况为例，一个 0—0 外显子插入了一个相位为 1 的内含子，结果造成被插入外显子和在它 3′ 端侧所有外显子的阅读框架移动。图 6－11 c 则显示，一个 0－0 外显子插入一个相位为 0 的内含子之中将不会引起阅读框架移动。

6.9 产生新功能的变通途径

除了基因重复和外显子混匀以外，还有许多别的产生新基因或新多肽的机制。以下将考虑 3 种这样的机制。

重叠基因

已经发现，一个ＤＮＡ片段能通过用不同阅读框架来为一个以上的基因编码。这一现象在病毒、细胞器和细菌中普遍存在。图６－１２ａ展示了一个单链ＤＮＡ噬菌体ΦＸ１７４的遗传图。其中已观察到几个重叠基因。例如，基因Ｂ整个地被包含在基因Ａ之内，而基因Ｋ在5′端与基因Ａ重叠，在3′端与基因Ｃ重叠。后一情况的更详细分析于图６－１２ｂ给出。

重叠基因也可通过应用一个ＤＮＡ序列的两条互补链而产生。例如，人线粒体基因组中，确定 $tRNA^{Ile}$ 和 $tRNA^{Gln}$ 的基因分别位于不同的链上，并且它们之间有一个３－核苷酸重叠，前者中读成 5′-CTA-3′，后者中读成5′-TAG-3′(Anderson 等,1981)。

问题是，在进化期间重叠基因可能是怎样产生的呢？为了回答这一问题，我们注意到，开放阅读框架在整个基因组中大量存在。因此，相当长度的潜在编码区存在于已有基因的不同阅读框架中或互补链上，这是完全可能的。因为６４种可能的密码子中只有３种是终止密码子，所以，即使一个随机的

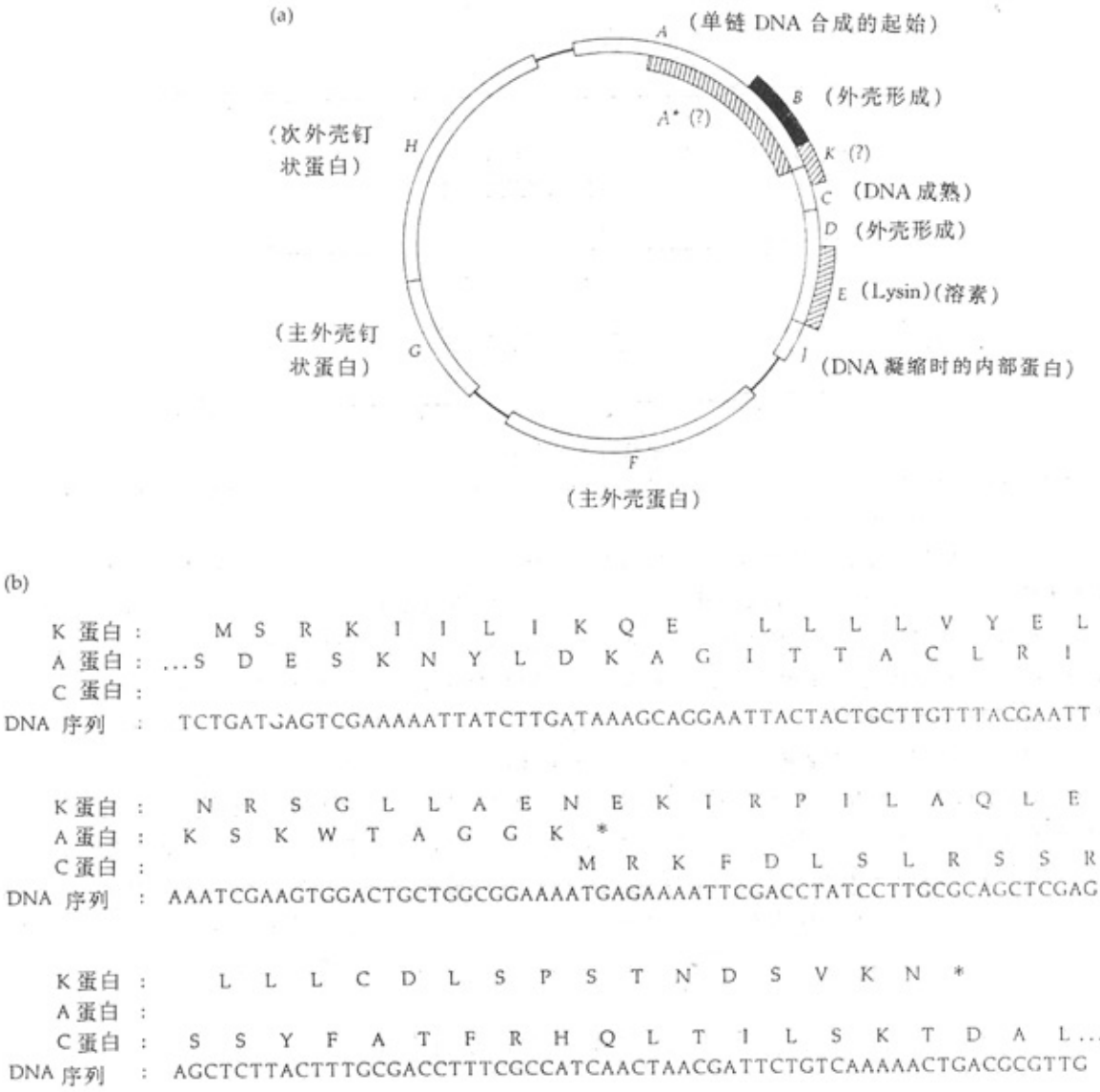


图６－１２ ΦＸ１７４噬菌体单链环状ＤＮＡ图。注意，为Ｂ蛋白质编码的基因（黑色）被完全包含在为Ａ蛋白质编码的基因内，而基因Ｋ则与两个基因Ａ和Ｃ重叠。自 Kornberg(1982)修改而成。（ｂ）显示出与基因Ａ的5′部分和基因Ｃ

的 3′ 部分重叠的K基因的顺序。星号表示终止密码子。(关于氨基酸的单字母缩写见表 1 - 1)。

ＤＮＡ顺序也可能含有成百个核苷酸长度的开放阅读框架。如果碰巧这样一个阅读框架中含有一个起始密码子和一个转录起始位点，或者通过突变产生了这些位点，那么，一条额外的 mRNA 就将会被转录出来，并随后被翻译成一个新蛋白质。这一新产物是否有有利功能那就是另一回事了，但如果它确有，则这种性状就可能会在群体中固定。

我们也注意到，为重叠基因编码的ＤＮＡ区段上的进化速率，预期要低于只用一种阅读框架的类似ＤＮＡ序列。其原因是，在重叠基因中非简并位点的比例要高于非重叠基因中的同类比例，这就大大降低了同义突变在总突变中的比例（Miyata 和 Yasunaga,1978)。

变通性的拼接

原始ＲＮＡ转录产物的变通性拼接，可能会造成从同一个ＤＮＡ片段产生不同多肽产物的结果。在这种情况下，外显子和内含子间的界限就不再是绝对的了，而是有赖于所涉及的 mRNA。许多ＲＮＡ变通性加工的例子已在多细胞生物中被发现。

变通性拼接常被用作生长调节的手段。在涉及果蝇 D. melanogaster 的性别决定过程的几个基因中，曾看到一种非常有趣的情形。至少有 3 个基因：性致死基因(Sxl)、转化基因(tra)和倍性基因(dsx)，在雄性和雌性中是以不同的方式进行拼接的（图６－１３)。在 dsx 的情况中，该基因有 6 个外显子；外显子 1、２、３和 4 用于雌性，而外显子 1、２、３、５和 6 则用于雄性。在 Sxl 和 tra 的情况下，雄性中变通性拼接的产物含有成熟前终止密码子，因而是无功能的。例如， Sxl 的外显子 3 中含有一个框架内的终止密码子，但雌性的 mRNA 却不含这种外显子。

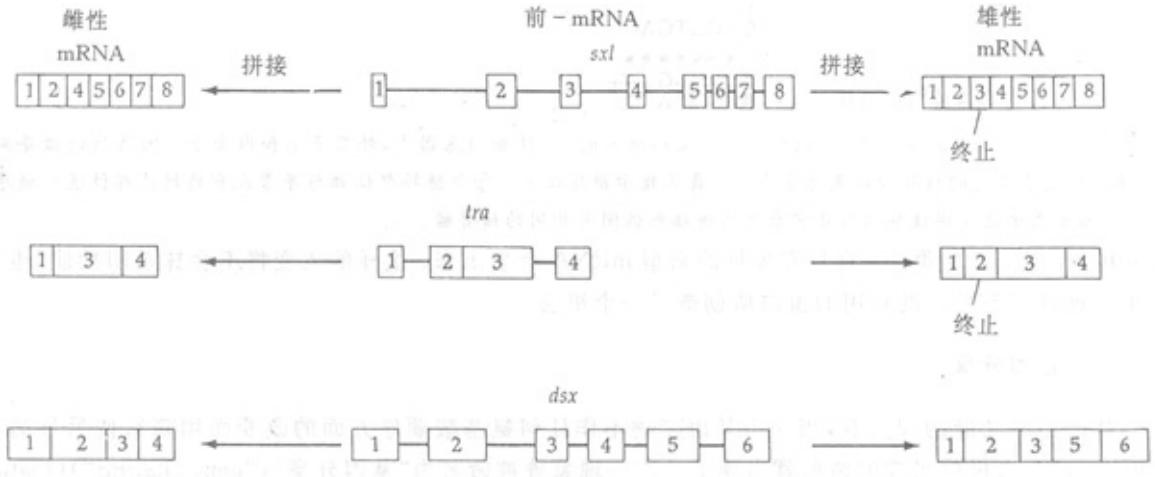


图 6 - 1 3 果蝇 D.melanogaster 雌性(左)和雄性(右)中,性致死基因(Sxl), 转化基因(tra) 和倍性基因(dsx)的拼接模式。“停止”指示一个终止密码子，它截断成熟 mRNA 的编码区，从而造成产物无功能。自 Baker(1989)。

变通性拼接的一个特例可用由内含子编码的蛋白质的例子来加以说明（Perlman 和 Butow,1989)。在这类例子中，该内含子含有一个开放阅读框架，它为功能完全不同于其两侧外显子所编码的蛋白质的某种蛋白质全部或其一部分编码。在有些情况下，这类开放阅读框架是其上游外显子的延伸物，例如，酵母线粒体基因 cox Ⅰ中的内含子ａ１４α（图６－１４ａ)。在另一些情况下，内含子不仅包括一个游离态的编码蛋白质的基因，而且还含有关于转录起始和终止的必要信号（图６－１４ｂ)。内含子ａ１４α十分有趣，因为它为一种被称为成熟酶的酶蛋白编码。成熟酶对这个内含子从它的前体 mRNA 上准确地自我拼接去除是必要的。这种成熟酶在ＤＮＡ重组中还起着内切核酸酶的作用。

要出现变通性拼接的进化，就需要重新产生一个变通性拼接的联结位点。因为拼接信号通常长为５－１０个核苷酸，所以通过突变这类位点以一种可以查觉的频率产生是有可能的。事实上，从文献中知道已有许多这样的例子。例如，图６－１５所示的例子中，甘氨酸密码子中的一次同义替换即把某一编码区变成了拼接部位。在图６－１５关于β^{+}-地中海贫血症的病理学查证的例子中，新的拼接位点通常比老的拼

接位点更强（即，该突变发生后合成的 mRNA 大多数为已发生改变的那种类型）。这样的突变显然具有有害的后果，预期绝不会在群体中固定。然而，如果新产生的拼接位点要弱得多，则大多数

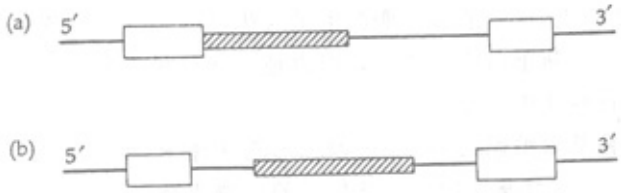


图 6 - 1 4 内含子为蛋白质编码的例子:（ a ）一个开放阅读框架（画有斜条纹），它是上游外显子（空心矩形块）的一个延伸物（例如，酶母线粒体基因 cox Ⅰ中的内含子 a 1 4 α）;（ b ）一个游离存在的开放阅读框架，其转录起始和终止信号位于内含子之中（例如，噬菌体 T4 中 sum γ基因的内含子）。资料取自 Perlman 和 Butow(1989)。

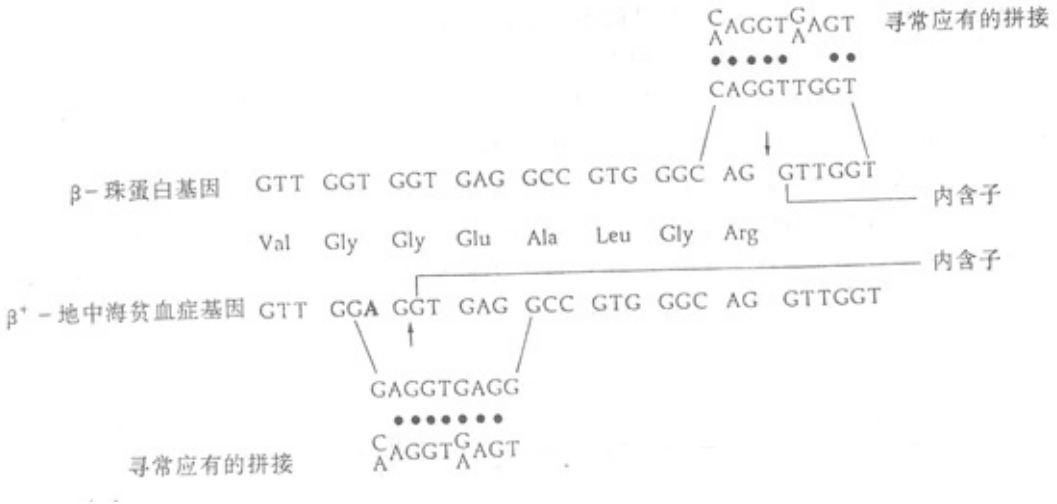


图 6 - 1 5 来自正常个体和患β + 地中海贫血症的病人的β-珠蛋白基因中，外显子 1 和内含子Ⅰ间区域的核苷酸顺序。发生了突变的核苷酸以黑体字表示。箭头指示拼接位点。每个拼接部位都与寻常应有的拼接部位进行顺序比较，园点表示这些拼接部位与寻常应有的拼接部位间有相同的核苷酸。

mRNA 将是原始类型，且只有少量的新型 mRNA 产生出来。这样的改变将不会抹杀旧功能，也为产生一种具有新功能的有用的蛋白质创造了一个机会。

基因分享

从产生新功能的观点看，当一个基因产物不作任何氨基酸顺序方面的改变而用来行使另外的功能时，一种令人极感兴趣的情形就出现了。这一现象曾被命名为“基因分享”(“gene sharing”) (Piatigorsky 等,1988)。基因分享的意思是，一个基因在没有重复也没有失去原始功能的情况下获得了并保持着第二种功能。不过，基因分享可能会要求在组织特异性或发育时序性的调节系统方面发生一点变化。

基因分享最初在晶状体中发现，这里晶状体是构成眼睛的水晶体赖以维持透明和适当的光线折射的物质。最初的发现是，来自鸟类和鳄类的∈晶状体在氨基酸顺序上与乳酸盐脱氢酶 B （LDH-B4,见第 84 页）等同，且具有同样的ＬＤＨ活性（Wistow 等,1987)。后来的工作表明，这“两种”蛋白质事实上是同一种蛋白质，而且是由同样的基因来编码的（Hendriks 等,1988)。第二种晶状体δ存在于所有鸟类和爬行类之中，也已被证明在顺序方面与另一种酶等同。这种酶即精氨琥珀酸裂解酶，它催化将精氨琥珀酸转化成精氨酸的反应。这两个蛋白质好象也是由同样的基因编码的（Piatigorsky 等,1988)。类似地，七鳃鳗，真骨鱼类，爬行类和鸟类中的τ－晶状体，已被证明与α－烯醇酶等同并由同样的基因编码。α－烯醇酶是糖酵解中的一种酶，将 2 －磷酸甘油酸转化成磷酸烯醇式丙酮酸（Piatigorsky 和 Wistow，1989)。于是，δ－、 ∈－和τ－晶状体的例子说明，一个未经重复的基因能通过基因分享而获得额外的功能。另一方面，α、β和γ晶状体则是另一类蛋白质的经典例子，这类蛋白质通过基因重复，以及其后由祖先基

因分化成为不同蛋白质编码的基因而进化(例如,热震惊基因,为暴露于过热环境后才表达的蛋白质编码)。

基因分享可能是相当普通的现象。事实上，在以上例子中，那些酶和晶状体自身就可能有两种以上的功能。例如，τ—晶状体/α—烯醇酶也能象一个热震惊蛋白质那样起作用。显然基因分享增添了基因组的简洁性，即使在真核生物中简洁性看来并没有很高的优越性（第八章)。还要注意，在晶状体基因分享的例子中，同一个多肽既起着酶的作用又有结构蛋白质的功能，这就搅乱了酶和非酶、或结构蛋白质之间的传统界限。

6.10 多基因家族的协同进化

从十九世纪七十年代中期到十九世纪八十年代中期，关于ＤＮＡ重退火和ＤＮＡ杂交的研究大量涌现，目的在于探明真核生物基因组的结构和组织。这些研究揭示，较高等生物的基因组是由高度重复顺序、中度重复顺序和单拷贝顺序等３类序列所构成的（第八章)。它们还揭示出一个有趣的进化现象，即：一个重复顺序家族的成员在一个物种内相互间一般是非常相似的，而来自不同物种的该家族的成员，即使这些物种间亲缘关系很近，相互间也可能是很不一样的。这一现象最先被布郎等(Brown 等,1972) 查觉，他们是在比较来自非洲蟾蜍 Xenopus laevis 和X.borealis 的核糖体DNA时发现的,后一种蟾蜍那时曾被误认为是X. mulleri。

在 Xenopus (爪蟾属)和大多数别的脊椎动物中，确定１８Ｓ和２８Ｓ核糖体ＲＮＡ的基因以成百的拷贝数存在着，且以一列或几个串联的列的形式排列着。每一重复单位由一个转录片段和一个不转录片段构成（图６—１６)。转录片段产生一个４５Ｓ ＲＮＡ前体，该前体经酶切割而被划分成有功能的１８Ｓ和２８Ｓ核糖体ＲＮＡ。这种转录重复片段通过不转录间隔片段（ＮＴＳ）而相互隔开。

在对 X.laevis 和X.borealis 间的核糖体ＲＮＡ基因的比较中，布朗等(Brown 等,1972)发现，虽然这两个物种的１８Ｓ和２８Ｓ基因极为相似，但这两个物种间的ＮＴＳ区域却大不相同。相比之下，在每一个体内以及一个物种的不同个体间，该ＮＴＳ区域则是非常相似的。于是，看起来好象是这样一种情况：虽然ＮＴＳ区域在不同物种间迅速分岐，但它们在每一物种中却是一起进化的。布朗等（Brown 等,1972）的结论是，一定有一种“校正”机制在起作用，以使某一突变从一个间隔顺序传向邻近的间隔顺序，其速度快于在这些顺序中出现新的变化。他们称这种在一个个体内显现的现象为水平进化（horizontal evolution)，以用来与垂直进化相对照。垂直进化是指某一突变在一个繁殖群体中的传播。后来，有人提出了并发进化（coincidental evolution）或协同进化（concerted evolution） 等等术语。后一个术语由齐默尔等(Zimmer 等,1980)提出，是目前文献中最通用的术语。

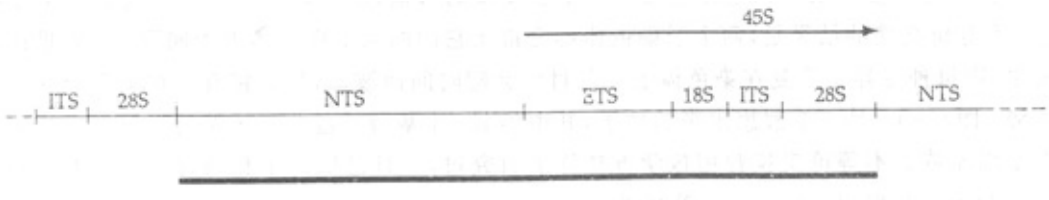


图６－１６ 脊椎动物ｒＲＮＡ基因的一个典型重复单位的图解表示。黑色杆线表示重复单位，箭头指示转录单位。ＥＴＳ，外转录间隔; ＩＴＳ，内转录间隔; ＮＴＳ，不转录间隔。自 Arnheim(1983)。

随着限制酶分析和ＤＮＡ顺序测定等技术的出现，已有大量资料证明多基因家族中协同进化的普遍性（见 Ohta，1980;Dover,1982;Arnheim,1983 等综述)。图６—１７展示了一个来自人和黑猩猩核糖体基因的限制酶分析的例子。人类中，每一重复单位在２８Ｓ基因3′端的ＮＴＳ区中有一个 *Hpa* Ⅰ位点，而在黑猩猩和其他大型猿类中则缺少这一位点。该 *Hpa* Ⅰ位点很有可能是在人—猿分枝之后而在人谱系中产生的,并且最终在每一个人的重复中固定了。NTS 区中其他限制性位点也同样展示出物种特异的同源性。

协同进化本质上意味着，一个基因家族的某个成员并不是与该家族的其他成员毫不相干地进化着的。通过其成员间的遗传相互作用，一个多基因家族是以协同的方式，象一个整体一样地一起进化着。

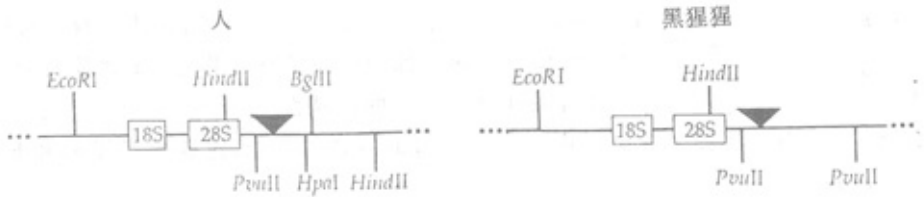


图 6 - 1 7 人和黑猩猩 1 8 S和 2 8 S核糖体基因中的限制性位点。所用限制酶为 EcoR Ⅰ, Hind Ⅱ, Pvu Ⅱ, Ⅲ, 和 *Hpa* Ⅰ 。基因上面标出的限制位点在物种中是多态的。基因下面的那些位点则是单态的。倒三角形表示N T S中长度上的多态性。自 Arnheim(1983)修改而成。

协同进化的机制

不等价交换（unequal crossing over）和基因转变（gene conversion）（图 6 — 1 8 ）近来认

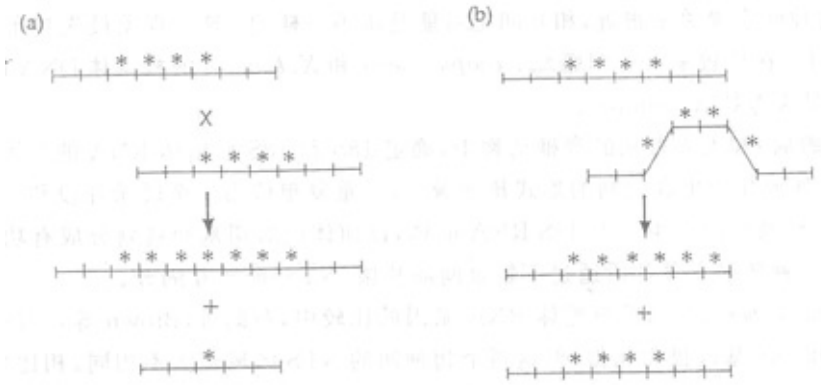


图 6-18 (a)不等价交换模型和(b)基因转变模型。不等价交换的结果是，两条子染色体都出现了重复数上的改变和两种重复类型（其中一种用星号标出）的频率上的改变，后者是与亲本频率（ 5 0 % ）相比较而言的。另一方面，基因转变则仅在其中一条子染色体中改变两类重复的频率,而且对两条染色体都不改变它们的总重复数。自 Arnheim(1983)修改而成。

为是造成协同进化的两个最重要的机制。不等价交换可以发生在生殖细胞减数分裂时某一染色体的两条姐妹染色单位体间，也可以发生在有丝分裂时的两同源染色体间。一个交互重组的过程就是在一条染色单体或染色体中产生某一顺序的重复，而在另一条中则造成相应的缺失。图 6 — 1 8 a展示出的例子中，一次不等价交换事件导至一条子染色体上出现 3 个重复段的增幅，而在另一条上则出现 3 个重复段的缺失。这种不等价交换的结果是，两个子染色体都变得比它们的亲本染色体更为同源化。如果这种过程反复发生，则每种变异型重复在染色体上的数目将会随时间而波动，最后将有一种类型会在该家族中处于优势。图 6 — 1 9 是一个假想出来的例子，其中类型— 4 基因通过反复多轮的不等价交换而传遍了某一个基因家族。不等价交换曾用数学方法详细研究过，并且已取得了相当程度的实验支持（见 Ohta,1980;Dover,1982;Li 等,1985a 等综述）。

基因转变是一种非交互重组的过程，在此过程中两个序列相互作用的方式为，其中一个被另一个转化（见 Lewin，1990）。从协同进化过程的观点看，基因转变中最重要的类型是非等位基因转变（即，位于不同基因座位的基因间的转变，而不是不同的等位基因形式间的转变）。图 6 — 1 8 b是一个非等位基因转变的例子，其中野生型重复中有两个转化成了突变型。结果，第一条子染色体变得比亲本染色体更为同源些，而在第二条子染色体中却未发生变化。理论研究表明，和不等价交换一样，基因转变也能产生协同进化(Ohta,1984; Nagylaki, 1984) 。基因转变曾经作为γ-珠蛋白基因(Jeffreys,1979;Scott 等, 1984)和另外许多基因中(见 Dover,1982)出现同源化的机制而提出来过。

作为一种协同进化的机制，基因转变看来有几个胜过不等价交换的优点。首先，不等价交换使一个家族中的重复基因的数目发生改变，故而有时可能会造成严重的份量不平衡。基因转变则相反，并不造成基因数目改变。其次，基因转变不仅能对串联的重复、而且能对分散的重复起着校正机制的作用。相比之下，不等价交换在所涉及的重复散布于非同源染色体上时就受到了限制。如果这些重复基因

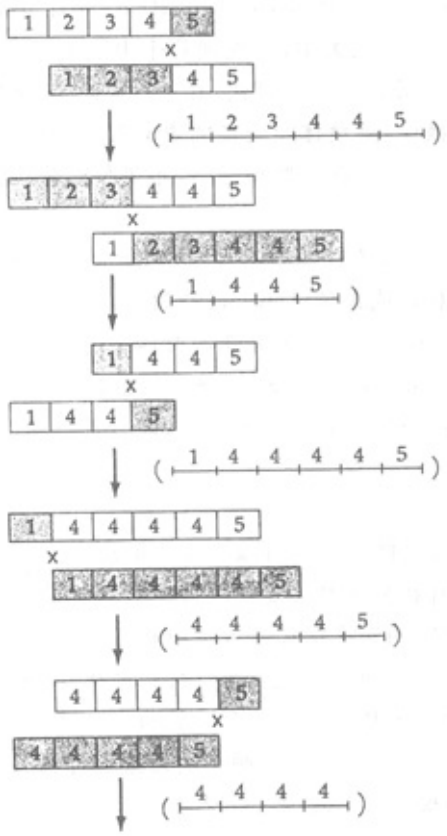


图 6 - 1 9 由不等价交换导至的协同进化。不等价交换事件反复循环使得每一染色体上重复基因变得越来越同源化。该过程还会影响每一染色体上重复顺序的数目。自 Ohta，(1980)。

位于染色体的端粒部位（染色体臂的末端），象在人和猿的ｒＲＮＡ基因中的情况那样，或许它能对非同源染色体起有效作用；而若分散的重复位于染色体的中间，象在小鼠的ｒＲＮＡ基因中的情况那样，则它将受到极大限制（Arnheim,1983）。如果这些重复散布在一条染色体上，则不等价交换可能会导致重复段间基因的缺失或重复这样的结果。例如，图６－２０表示两重复簇间不等价交换的一个假想的例

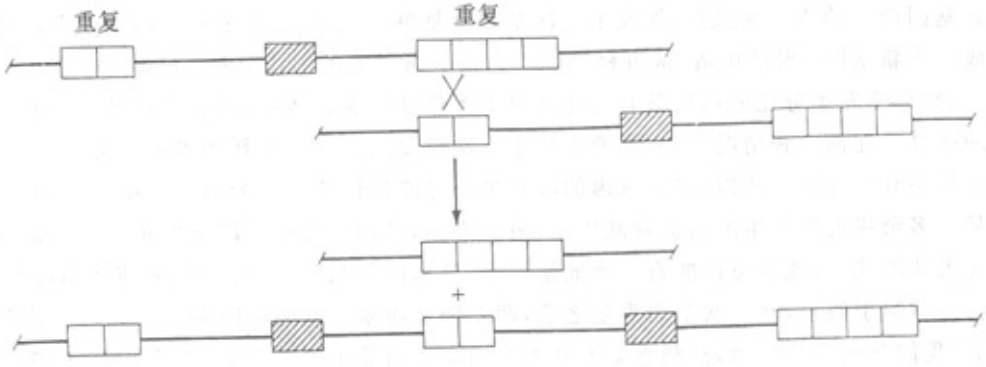


图 6 - 2 0 涉及散布重复（空心矩形块）的交换。画有斜条纹的矩形块表示一个唯一的基因。该交换事件后，这一基因在一条染色体上缺失，而在另一条染色体上则重复。

子，结果是一条染色体中某一唯一基因的缺失，而在另一条染色体中则相应地重复。其中一条染色体或者全部两条染色体，可能会对携带它们的生物带来有害影响。第三，基因转变可以有偏向性的方向。来自真菌的实验资料已经表明，基因转变方向上的倾斜是普遍的、而且常常是强烈的（Lamb 和 Helmi,1982），而理论研究业已证明，轻微的倾斜就可能对重复突变的固定概率有很大影响（Nagylaki 和 Petes,1982;Walsh,1985）。

在串联重复顺序的大家族中，不等价交换可能象基因转变一样是一个可接受的过程。首先，在这

类家族中，重复的数目显然能极大地波动而不至引起显著的逆效应。这是从以下观察中得出的结论：果蝇中确定ＲＮＡ的基因的数目在同一物种的不同个体间，以及不同物种间变化幅度大（Ritossa等,1966;Brown和Sugimoto,1973)。在人类中，已发现几个串联重复的家族，它们在拷贝数方面展现出异乎寻常的变化程度（Nakamura 等,1987)。其次，在一次基因转变事件中，通常只有一个小区域（异源双链区）涉及，而在不等价交换中，染色体间交换的重复数则可能非常大。显然，交换的重复数越大，协同进化的速率就越高（Ohta,1983)。在有些情形下，不等价交换的这一优点可大到足以与基因转变的优点相抗衡。

除不等价交换和基因转变之外，还有一些其他机制，象复制滑脱和转座（第一章和第七章)，也能造成某一家族中变异型基因的获得或丢失（Dover,1982)。最后，应注意到，协同进化不仅要求突变在该家族成员间的水平转移（同源化)，而且要求突变向群体中的所有个体传播（固定)。所以，我们还需要考虑随机遗传漂变的效应。多费（Dover,1982,1986）对在DNA 转移和随机遗传漂变等各种机制联合作用下,多基因家族的协同进化过程,起了一个名称,即分子驱动(molecular drive)。

协同进化的进化论含意

协同进化使得某一变异型重复能传向所有基因家族的成员。这种水平地传播的能力有着深远的进化后果，因为这样一来，一个有利突变型重复就能替代所有其他重复而在该家族中固定。我们注意到，单单一个变异型所能给予生物的选择优势通常是很有限的。然而,如果该突变传给了许多甚至所有成员的话，则这种优势就会大大地加强。于是，通过协同进化，一个较小的选择优势可以变成较大的选择优势。在这方面，协同进化优于基因家族各个成员的独立进化（见Arnheim,1983;Walsh,1985)。

阿恩海姆（Arnheim,1983） 曾对ＲＮＡ多聚酶Ⅰ的转录调控信号和ＲＮＡ多聚酶Ⅱ的转录调控信号的进化进行过比较。ＲＮＡ多聚酶Ⅰ只转录ｒRNA基因，而RNA多聚酶Ⅱ则转录所有为蛋白质编码的基因（第一章)。ＲＮＡ多聚酶Ⅰ的转录调控信号的进化看来比ＲＮＡ多聚酶Ⅱ的该信号的进化要快得多。例如，在无细胞转录系统中，一种小鼠 rRNA 克隆不能在人的细胞提取物中转录，但来自差异极明显的物种的蛋白质编码基因的克隆却能够在异源系统中转录（例如，家蚕的基因在人的细胞提取物中，和哺乳类的基因在酵母的提取物中)。阿恩海姆（Arnheim,1983）认为，在关于ＲＮＡ多聚酶Ⅰ的转录单位的例子中，倾向于影响转录起始的那些有利突变，因协同进化所造成的后果可能已传播到整个 rRNA 多基因家族。与之不同的是，在关于ＲＮＡ多聚酶Ⅱ的转录单位的例子中，在任何一个基因中发生的影响转录起始的有利突变，预期将不会传遍所有基因，因为它们属于许多不同的家族。

关于新基因产生的传统观点是，先发生一次基因重复事件，然后该重复产生的两个基因之一逐渐分化而变成一个新基因。现已搞清，该过程可能不象以前所假定的那样简单。只要两个基因分岐的程度不是很大，则那个发生分化的拷贝就有可能通过不等价交换而被清除，或者通过基因转变而转化成保持原样的拷贝。在前一种情况下，它需要再发生一次重复以产生一个新的多余拷贝；而在后一种情况下则必须从头开始分化。所以，重复基因的岐化进程可能比传统上认为的要慢得多，为此缘故一个新基因从某一多余拷贝中产生的机会就减少了。另一方面,基因转变也可能会阻止一个多余的拷贝长时期地成为无功能状态，或者也许能有选择地使一个“死基因”（假基因）复活过来（Walsh,1987)。

我们已经习惯于假定，在一次基因重复之后，两个随之而来的基因将随时间单调地岐化着。在这样的假定下，我们前面已经证明，推测重复事件的时间是相当简单的。例如，人β－和δ－珠蛋白的蛋白质顺序相互间相似的程度比与兔β 1或与小鼠β的主要和次要顺序的要高（Dayhoff,1972)。因此曾有过这样的推测：人的这两个基因是从约４０００万年以前的一次重复事件中衍生而来的，这个时间远在哺乳类辐射（约８０００万年以前）之后。考虑到重复基因能相互较正这样一个事实，则这一结论就可能是错误的。事实上，近来已经有人提出，β和δ基因起源于发生在哺乳类辐射之前的一次重复（Hardison 和Margot,1984)。该提议是根据这样的观察事实而作出的，兔假基因ψβ2的大内含子和3′ 不翻译区与人的δ的相似程度比与兔β 1的高，小鼠的假基因βｈ3相似于其3′末端上的δ。如果这一假说结果是正确的，则以上例子极好地说明了，基因－校正事件将会怎样部分或全部地抹擦掉重复基因间分岐进化的历史。在大的多基因家族中，基因－校正事件预期是频频发生的，在这种情况下，追踪家族成员间的进化关

系将会更为困难。

从进化论的观点看，多基因家族的进化和分群体的进化之间存在着某种类比。我们可以把多基因家族中的每一种重复看成是分群体中的一种同类群。那么重复间信息的传递就等价于同类群间基因或个体的迁移。众所周知，迁移将减少两同类群间遗传差异的量，但会增加一个同类群中的遗传变异的量（例如等位基因的数目）。类似地，重复间的信息传递将会减少重复间的遗传差异但将增加某一基因座位上的遗传变异的量（Ohta,1983,1984;Nagylaki,1984）。小鼠主组织相容复合体中某些基因座位是高度多态的，事实上，已观察到某一基因座位上的等位基因数多达５０个。所以，曾有人提出，这种程度较高的多态性是由基因转变所造成的（例如Weiss等,1983;但可参阅Hughes和Nei,1989）。

习题

1 与某一随机选出的ＤＮＡ片段的重复相比，外显子重复有什么优点？

2 在重叠基因中，简并位点的数目能大大地被削减。（ａ）如果下面的顺序仅按第1个阅读框架翻译，那么将有多少非简并位点？多少四重简并的位点？（ｂ）如果除了第一个阅读框架外，该顺序也按第2个阅读框架翻译，那么将有多少非简并位点和四重简并位点？（ｃ）如果3个阅读框架都进行翻译，那么该顺序中的非简并位点和四重简并位点将是多少？（3个阅读框架的起始点各用箭头标出）

3 许多多聚体蛋白质是由重复基因编码的亚基所构成。有两种可能情况：（ａ）这些亚基可能全部来自一个基因座位，也可能来自不同的基因座位。在前一种情况下该蛋白质被称为是“同质型的”，而后一种情况下则是“异质型的”。假定该蛋白质象乳酸脱氢酶（ＬＤＨ，见第84页原版）那样是一种四聚体酶，且其所有亚基由两个基因座位编码。那么，能产生多少种不同的同功酶？（ｂ）这样的蛋白质常常是异质型的（即，常常是由不同基因座位产生的亚基所构成的）。例如，哺乳类成体的血红蛋白是一个由两条α链和两条β链构成的四聚体。假定在某一种哺乳动物中有3个α样基因座位和2个β样基因座位。那么，如果每一个四聚体是由两个α样基因座位产生的亚基和两个β样基因座位产生的亚基所构成的话，则能产生多少不同的异质型四聚体？

4 在大鼠和小鼠的基因组中有两个为胰岛素编码的基因（前胰岛素原Ⅰ和Ⅱ），而在除啮齿类以外的哺乳动物中则只有一个胰岛素基因。前胰岛素原Ⅰ基因被认为是通过一个所谓“反录转座”（第七章）的过程而产生的。前胰岛素原Ⅰ和Ⅱ这两个基因都在5′不翻译区中含有一个小内含子（长为１１８个核苷酸）。大鼠和小鼠中的内含子对子间，核苷酸差异数如下：

	内含子		
内含子	小鼠Ⅰ	小鼠Ⅱ	大鼠Ⅰ
小鼠Ⅱ	21		
大鼠Ⅰ	15	25	
大鼠Ⅱ	16	24	18

（ａ）该矩阵中的数字指示，小鼠前胰岛素原Ⅱ（即小鼠Ⅱ）基因中的内含子比其他基因中的相应内含子进化快，为什么？（ｂ）假定核苷酸替换的速率恒定，并且假定小鼠和大鼠在１５００万年前发生分岐，请用小鼠Ⅰ和大鼠Ⅱ但排除小鼠Ⅱ，来估计前胰岛素原Ⅰ和Ⅱ间的分岐时间。

５、在以下顺序的内含子（虚线）中插入一个０－０对称外显子。那么对阅读框架而言将会发生什么？如果插入一个２－２对称外显子会发生什么呢？如果插入一个不对称外显子又会发生什么？

5′——CAT TCG TCT TTA TTC GAA ATC GCG---TGG ACA GCG GTG AAT CTC TTT GAC GCT GTG——3′

６、为什么一个串联重复的家族能比一个散布重复的家族更容易经历协同进化？ 请解释。

后继阅读文献

Cold Spring Harbor Symposium on Quantitative Biology.1987.Evolution of Catalytic Function, Vol. 52. Cold Spring Harbor Laboratory.Cold Spring Habor,NY.

Dayhff,M.O.1972.Atlas of Protein Sequence and Structure, Vol. 52. National Biomedical Research Foundation,Silver Spring,MD.

Dover,G.A.and R. B. Flavell. (eds) . 1982. Genome Evolution, Academic Press,New York.

Li,W,- H. 1983. Evolution of duplicate gene and pseudogene,pp 14-37.In M.Nei and R.K.Koehn (eds.) ,Evolution of Genes and Protein. Sinauer Associates,Sunderland MA.

Ohno ,S.1970.Evolution by Gene Puplication,Springer Verlag,Berlin.

Ohta,T.1980.Evolution and Variation of Multigene Families.Springer-Verlage,Berlin.