Biological Genome Evolution and Codon Usage

生物基因组进化密码子的使用

王志坚　王芳平　著

国防工业出版社
National Defense Industry Press

生物基因组进化
密码子的使用

王志坚　王芳平　著

国防工业出版社
·北京·

内 容 简 介

本书主要用生物信息学理论从基因组进化的角度研究了不同进化水平生物基因组中密码对使用的规律。全书共分8章，内容包括：密码子和密码对使用研究的发展历史及研究现状；密码对相对模式数随频数的分布和分布模型拟合理论；基于密码对使用偏好的基因组相似性分析方法；密码对使用在DNA双链上的不对称性分析理论；密码对使用与基因组进化的线性相关分析；依赖上下文关系的密码对使用偏好性理论。本书总结了最近几年该领域最新的理论研究成果，提供了大量的资料、图表和数据。

本书可供高等学校生物信息学及相关专业的师生以及从事生物基因组进化和密码子偏好性使用研究的科研人员参考。

图书在版编目(CIP)数据

生物基因组进化密码子的使用 / 王志坚，王芳平著. —北京：国防工业出版社，2014.4
ISBN 978-7-118-09366-7

Ⅰ. ①生… Ⅱ. ①王… ②王… Ⅲ. ①基因组—密码子—研究 Ⅳ. ①Q343.2②Q755

中国版本图书馆CIP数据核字(2014)第064219号

※

国防工业出版社出版发行
（北京市海淀区紫竹院南路23号 邮政编码100048）
北京嘉恒彩色印刷有限责任公司
新华书店经售
*
开本 710×960 1/16 印张 8 字数 185千字
2014年4月第1版第1次印刷 印数 1—2000册 定价 48.00元

（本书如有印装错误，我社负责调换）

国防书店：(010)88540777 发行邮购：(010)88540776
发行传真：(010)88540755 发行业务：(010)88540717

前　言

1953 年，沃森和克里克弄清 DNA 的双链双螺旋结构之后，科学家就围绕遗传密码的破译不断地推测探索，最终找到了答案，即遗传密码是指 DNA 或 mRNA 中的核苷酸序列与其所编码蛋白质中氨基酸序列之间的对应关系。这个发现是 20 世纪 60 年代分子生物学最辉煌的成就，也是后来蓬勃兴起的基因工程和人类基因组计划得以实现的基础。生物最本质的特征是进化，其中密码子的进化是当今基因组学研究的热点命题之一。研究发现，不论生物简单到只一个细胞，还是复杂到与人一样高等，它的遗传密码是一样的。虽然在几乎所有的生物中，遗传密码都是通用的，但是，在生物进化的过程中，遗传密码也是不断地进化的，例如线粒体的基因组已发现 27 个改变了的密码子，在 Candida albicans 这种生物中，亮氨酸密码子 CUG 被更多的解码为丝氨酸，终止密码子有时会被硒代半胱氨酸（selenocysteine）和吡咯赖氨酸（pyrrolysine）解读。揭示不同基因组中密码子的使用模式以及影响这种模式形成的内在因素，对于了解基因组特征和分子进化历史事件具有重要的启示。有关基因组进化的信息必须会在其 DNA 序列中反映出来，也就是说序列的组成和碱基的搭配必须包含基因组进化的信息，为了寻找这些信息，遗传密码包括 61 种编码氨基酸的有义密码子和 3 种通常不编码任何氨基酸的终止密码子（UAA、UAG 和 UGA）。一般每种氨基酸对应一种或多种密码子。编码同一种氨基酸的密码子称为同义密码子。在蛋白质合成过程中编码氨基酸的同义密码子并不被随机使用，这就是同义密码子使用的非随机性。大量研究表明，不同物种或同一物种的不同基因之间都在密码子使用上存在明显偏好。密码子使用偏性主要受到突变偏好、翻译选择、蛋白质二级结构、复制和转录选择、蛋白质疏亲水性以及外部环境等多种因素的影响，正如密码子的使用一样，两个紧邻的密码子，即密码对的使用也是高度偏好的，这种偏置现象在原核和真核生物中都广泛存在。

本书总结了近几年密码子及密码对使用与基因组进化研究的最新成就，提供了系统的理论分析、数据计算结果和大量的资料及实验结果。本书内容分为

8章，第一章、二、七、八章由王志坚著，第三、四、五、六章由王芳平著。本书主要用生物信息学理论从基因组进化的角度研究了不同进化水平生物基因组中密码对使用的规律。

由于出书时间比较仓促，加上作者学术水平有限，书中不足之处在所难免，希望读者予以谅解，也欢迎读者不吝赐教，提出宝贵的意见，作者将不胜感激！

2013年12月

目　录

第1章 绪 论

1.1 引 言

生物信息的序列化，是生命科学进入21世纪的划时代里程碑，基因组测序技术的发展，使得基因组数据呈指数增长，而大规模生物基因组全序列的测定对于从整个基因组规模深刻认识、研究物种，阐明基因的结构与功能关系，利用某些基因组图谱和测序获得的信息推测其他生物基因组的基因数目、位置、功能、表达机制和物种进化等发挥着巨大作用。基因组序列是关于生命细胞的语言，仅仅通过4个字母来代表DNA化学亚基的字母表，出现了生命过程的语法，其最复杂形式就是人类。阐明和使用这些字母来组成新的“单词和短语”是分子生物学领域的中心焦点。其中遗传密码的破译是分子生物学和分子遗传学发展中的一个重大里程碑，也是后来蓬勃兴起的基因工程和人类基因组计划得以实现的基础。

生物最本质的特征是进化，其中密码子的进化是当今基因组学研究的热点命题之一。揭示不同基因组中密码子的使用模式以及影响这种模式形成的内在因素，对于了解基因组特征和分子进化历史事件具有重要的启示。有关基因组进化的信息必须会在其DNA序列中反映出来，也就是说序列的组成和碱基的搭配必须包含基因组进化的信息，为了寻找这些信息，本书将分析编码序列中密码对（密码对指编码序列中紧邻密码子）的搭配和非编码序列中三联体对搭配与基因组进化的相关性，密码对中二核苷酸偏好与基因组进化的关系等相关问题。如同密码子的非随机使用一样，密码对的使用也是高度偏好的。密码对组合模式（61×64）的高维性，使分析难度大大增加，一些重要信息往往被湮灭在复杂的背景噪声之中。因此从核苷酸序列出发来研究密码对的使用在生物信息学研究领域是一个比较棘手的问题。但是近年来，随着基因组学研究技术的快速发展，全基因组测序不断大规模地进行，庞大的基因组数据信息源源不断地从一系列新技术中产生，使得通过对不同物种的基因组数据进行比较分析，揭示不同物种间进化上的差异成为可能。

1.2 密码子研究的发展

遗传密码(表1.1)是指 DNA 或 mRNA 的碱基序列与其编码的蛋白质的氨基酸序列间的相互关系。

表 1.1 64 种密码子以及氨基酸的标准配对

第一个字母	第二个字母				第三个字母
	U	C	A	G	
U	苯丙氨酸 苯丙氨酸 亮氨酸 亮氨酸	丝氨酸 丝氨酸 丝氨酸 丝氨酸	酪氨酸 酪氨酸 终止 终止	半胱氨酸 半胱氨酸 终止 色氨酸	U C A G
C	亮氨酸 亮氨酸 亮氨酸 亮氨酸	脯氨酸 脯氨酸 脯氨酸 脯氨酸	组氨酸 组氨酸 谷氨酰胺 谷氨酰胺	精氨酸 精氨酸 精氨酸 精氨酸	U C A G
A	异亮氨酸 异亮氨酸 异亮氨酸 甲硫氨酸 (起始)	苏氨酸 苏氨酸 苏氨酸 苏氨酸	天门冬酰胺 天门冬酰胺 赖氨酸 赖氨酸	丝氨酸 丝氨酸 精氨酸 精氨酸	U C A G
G	缬氨酸 缬氨酸 缬氨酸 缬氨酸 (起始)	丙氨酸 丙氨酸 丙氨酸 丙氨酸	天门冬氨酸 天门冬氨酸 谷氨酸 谷氨酸	甘氨酸 甘氨酸 甘氨酸 甘氨酸	U C A G

遗传密码具有以下特点:

(1) 密码子的基本单位,即三个核苷酸(Triplet)组成一个密码子 (Codon),每个密码子由三个前后相连的核苷酸组成,一个密码子只为一种氨基酸编码。共有 64 个密码子。

(2) 密码子之间不重叠使用核苷酸,也无核苷酸间隔。从起点至终止信号之间所阅读的碱基对,称为一个读码框架 (Reading Frame)。插入或去掉一个碱基,就会使以后的读码发生错误,称为移码,由于移码引起的突变称为移码突变。

（3）密码子中第三位碱基具有较小的专一性，称为“摆动性”或“变偶性”。

（4）一种氨基酸可有多个密码子，这个特点称为密码子的简并性或多态性。可以编码相同氨基酸的密码子称为同义密码子。除 Trp 和 Met 只有 1 个密码子外，其他 18 种氨基酸均有 1 个以上的密码子，Phe、Tyr、His、Gln、Glu、Asn、Asp、Lys、Cys 各有 2 个密码子；Ile 有 3 个密码子；Val、Pro、Thr、Ala、Gly 各有 4 个密码子；Leu、Arg、Ser 各有 6 个密码子。

（5）64 个密码子中，有 3 组不编码任何氨基酸，而是多肽合成终止密码子：UAA；UAG；UGA。此外，AUG 既是甲硫氨酸的密码子，又是肽链合成的起始密码子。

（6）密码子的通用性。所有生物从最低等的病毒直至人类，蛋白质合成都使用同一套密码子表。

（7）变异性。虽然遗传密码在不同生命之间有很强的一致性，但亦存在非标准的遗传密码，称为变异性。在线粒体中，便有和标准遗传密码相异之处，甚至不同生物的线粒体有不同的遗传密码。如支原体会把 UGA 转译为色氨酸。纤毛虫则把 UAG（有时候还有 UAA）转译为谷氨酰胺（一些绿藻也有同样的现象），或把 UGA 转译为半胱氨酸。一些酵母会把 GUG 转译为丝氨酸。在一些罕见的情况下，一些蛋白质会有 AUG 以外的起始密码子。

1.2.1 同义密码子使用偏好性及其生物学基础

从 20 世纪 60 年代中期开始，人们就对密码子的特性进行了深入研究，发现同义密码子的使用概率并不相等，不同的生物，甚至同种生物不同的蛋白质编码基因，对简并密码子使用频率并不相同，具有一定的偏爱性这种现象称为同义密码子使用的偏好性。同义密码子偏好性的研究一直是进化领域中一个热门的话题。多年的研究结果表明从原核生物到真核生物，其基因组中同义密码子使用偏好性的现象广泛存在[1-8]，这一现象的产生有诸多生物学基础，如翻译机制[9-18]、基因的碱基组分[19-23]、基因长度[24]、tRNA 丰度等[25-30]，下面对这些因素进行简要概述。

（1）基因序列碱基组成的偏好性。在不存在弱的自然选择压力的情况下，一定方向的突变压力会影响序列本身的碱基组成，而这一效应同时也会反映在同义密码子的第三位上，如细菌基因组中核苷酸含量变化范围较广。在某些细菌中 GC ~ AT 突变压力高，使得密码子第三位 AT 含量很高，而另一些细菌中，AT ~ GC 突变压力高，因而密码子第三位 AT 含量很高。这样的偏好性仅仅是反映了序列组成的特征，而与蛋白功能或表达水平无关[31-41]。

（2）tRNA 丰度。Ikemura[42-46] 的实验证明在大肠杆菌（Escherichia coli）、

鼠伤寒沙门氏菌(Salmonella typhimurium)和酿酒酵母(Saccharomyces cerevisiae)中密码子的偏好性与同源 tRNA 的丰度有关。其中蛋白质复制数和那些同源 tRNA 含量最高的密码子关联最强。由于密码子在蛋白翻译过程中需要和携带对应反密码子的 tRNA 相互识别作用,才能把游离的氨基酸残基转移到多肽链上,因此这些对应 tRNA 的丰度决定了蛋白质合成的资源。在高表达基因中那些偏好使用的密码子对应 tRNA 含量也较高,这些密码子被称为最优密码子,它们靠减少与对应的 tRNA 匹配时间而加快翻译速度。

(3) 基因长度[24,29]。基因越长,能够容纳的密码子越多,在没有其他压力的情况下,同义密码子被选择的概率不会受样本容量限制而出现统计上的误差;反之,基因长度越短,可以编码的密码子数量和种类越少,甚至有的密码子根本不会出现。这种使用偏好性和其他进化压力无关。

(4) mRNA 二级结构[17]。有报道对具有不同超二级结构的蛋白编码基因按照同义密码子使用模式进行聚类分析,结果表明这两种蛋白的密码子使用模式有明显差异,暗示着密码子使用与蛋白质结构具有一定相关性。

(5) 蛋白质的亲疏水性以及氨基酸保守性[26,28]。不同的基因编码序列其氨基酸含量有可能不同,一方面稀有氨基酸由于本身出现机率小,一旦使用某种密码子,则其他密码子出现概率更小;另一方面,对于比较保守的氨基酸,不容易发生突变,则其密码子使用模式固定为序列本身组成。

(6) 基因水平转移和重组。如果某些基因是由其他基因组中水平转移而来,则基因序列上的一些特征,如 GC 含量会出现与原基因组中不一样的模式;如果某基因是重组而来,则在基因内部会出现 GC 异质区,这些不同寻常的变异同时也会反映在密码子使用模式的差异上。

(7) 密码子碱基组成的上下文关系 (Codon Context)。密码子的上下文关系是影响 mRNA 译码精确性的基因一级结构的主要特征。其研究内容主要包括两个方面:密码子所处环境对碱基偏好[47-62]和对密码子偏好[63,64]。如果密码子第一、二位是 G、C,那么密码子第三位倾向于使用 A 或 U,反之亦然。由于碱基互补配对原则的关系,A 与 U 配对,G 与 C 配对,如果密码子三个位点都是 A、U 或 G、C,则密码子和反密码子配对容易出现位置差错,或是影响配对速度,这样会影响到基因表达的速度。

1.2.2 遗传密码子的集中研究

遗传密码子的研究集中在以下几个方向:

(1) 遗传密码子的通用性与变异性的研究。无细胞提取液体外实验和大量事实表明从低等的原核生物到高等的真核生物,大多都使用相同的遗传密码,即

遗传密码子具有通用性，这已成为现代分子生物学的理论基础。但 1979 年 Barrell 等[65]首次发现在人的线粒体中，通用密码子 AUA（异亮氨酸）和 UGA（终止密码）分别编码甲硫氨酸和色氨酸。随后，人们又在牛、酵母和链孢霉等的线粒体及支原体、腺病毒和几种原生动物中发现了变异密码子的存在，如草履虫将 2 个终止密码子 UAA 和 UAG 作为编码谷氨酸或谷胺酰胺的密码子。变异密码子在一定程度上反映了遗传密码的进化过程。生物界丰富的多样性要求遗传信息也存在相应的多样性，而密码子的多样性是遗传信息多样性的重要组成部分。由于存在这些变异密码子，大大地增加了遗传信息的含量，这样可能提高进化的速度，促进生物的进化和生物多样性的发展。

（2）遗传密码子的起源和进化。虽然遗传密码的发现已经有近 50 年的历史，但是有关其起源和进化仍然是悬而未决的问题。从 20 世纪 60 年代中期遗传密码宣告全部破译之后，就出现了两个对立的理论。1967 年，Woese[66]提出立体化学相互作用论，认为遗传密码子起源于三聚体与氨基酸的直接配对；1968 年，Crick[67]提出偶然事件冻结论，认为密码子与氨基酸得到对应关系的出现，纯粹是一种偶然现象，而后在进化过程中被固定下来。半个世纪以来，两条研究路线各有进展，后来又陆续提出一些新的理论，影响较大的有：氨基酸和密码子共进化论、离体选择理论、解码机理起源理论、第二密码观点、信息理论和流体静力学压力假说等。这些理论和假说各执一词，莫衷一是。虽然在几乎所有的生物中，遗传密码都是通用的，但是在生物进化的过程中，遗传密码也是不断进化的。现在关于遗传密码的进化主要有两种假说，分别是渐进进化与随机进化。关于遗传密码的起源和进化理论至今仍未得到令人满意的诠释，各种生物基因组测序任务的完成，为研究遗传密码的起源和检验上述理论的真伪提供了新的素材；现代生物技术的组合运用也为研究密码子的进化提供了新的思路。

（3）遗传密码的扩张。遗传密码的“扩张”是指通过化学的、离体生物合成的或活体蛋白突变等方法，促使“渗漏”型终止密码子 UAG 或 UGA 编码 20 种标准氨基酸以外的任何非自然氨基酸[68]。这不同于硒代半胱氨酸和吡咯赖氨酸的发现[69,70]。后者是生物体的自主行为，而前者则是人为的结果。目前，对应无义密码子 UAG，至少已经有 30 种以上的非自然编码氨基酸被稳定和高效率地整合进了相应的蛋白质，获得了具有新或强化属性的蛋白质[71]。随着密码子研究手段、技术和策略的完善和发展，研究者发现遗传密码并非必须为三联体，在某些情况下 4 碱基密码子和 5 碱基密码子也可编码新的氨基酸[72-74]。

（4）遗传密码子简并性（多态性）的研究[75-79]。密码子简并主要取决于第

三位碱基。这样,两种生物的基因中碱基的组成可以不相同,但其编码的蛋白质中氨基酸的组成和功能可能基本相同。如果没有简并性的存在,变异的影响会非常大,这是不利于生物生存的。正因为有了简并由变异引起密码子一个核苷酸的改变,其结构只是变成了一个氨基酸的另一个密码子,而合成出与原来没有区别的蛋白质,这样密码子的简并性起到了防错的作用,从而维持了物种的稳定性。

1.2.3 衡量同义密码子偏好性的指标

为了进行同义密码子用法模式分析,20 世纪 80 年代以来,一些与密码子偏好性计算有关的统计量陆续被提出,在此作简要介绍:

(1) 相对同义密码子使用频率 (Relative Synonymous Codon Usage, RSCU)。该指标由 Sharp 等人于 1986 年提出,用于标准化那些来自不同氨基酸组成的密码子用法。由于该指标比较直观地反映了密码子使用的偏好性,因此应用最为广泛。RSCU 的定义是以某一同义密码子使用次数的观察值为分子,以该密码子出现次数的预期值为分母。其中,密码子预期出现的次数为当该密码子所编码的氨基酸的所有密码子平均使用时的次数。对于一个给定的氨基酸 i,其第 j 个密码子的 RSCU 值计算公式为

$$\mathrm{RSCU}_{ij} = \frac{X_{ij}}{\overline{X}} = \frac{X_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}}$$

式中:X_{ij}是该密码子出现次数的实际观察值;n_i 是编码此给定氨基酸的密码子简并数,其数值范围从 1 ~ 6。

如果密码子无偏好性,则 RSCU 值为 1;如果该密码子比其他同义密码子出现更为频繁,则 RSCU 值大于 1,反之亦然。

(2) 密码子适应指数 (Codon Adaptation Index, CAI)。由 Sharp 和 Li 在 1987 年提出,对于某个基因编码序列, CAI 是指实际编码该蛋白的所有密码子对完全使用最优密码子编码该蛋白的情况下的适应性指数,通过计算实际使用的密码子与其对应的最优密码子的 RSCU 值的几何平均数比值所得。用公式表示,对特定基因的第 k 个密码子:

$$\mathrm{CAI} = \frac{\mathrm{CAI}_{\mathrm{obs}}}{\mathrm{CAI}_{\mathrm{max}}} = \frac{\sqrt[L]{\sum_{k=1}^{L} \mathrm{RSCU}_k}}{\sqrt[L]{\sum_{k=1}^{L} \mathrm{RSCU}_{k\mathrm{max}}}} = \sqrt[L]{\sum_{k=1}^{L} \frac{\mathrm{RSCU}_k}{\mathrm{RSCU}_{k\mathrm{max}}}}$$

式中：$RSCU_{kmax}$指与第 k 个密码子编码同一氨基酸的最优密码子的 RSCU 值；L 指基因编码序列所有密码子数目。

当基因中密码子偏好性达到最大程度，即所有氨基酸都是由最优密码子编码，则 CAI 值为 1；CAI 值越小，则密码子偏好程度越低，反之亦然。该指标的计算和基因表达水平呈正相关，但是前提是需要有在高表达基因中的最优密码子表作为参照，因此应用上有一定局限。

（3）密码子不同位置上的 GC 含量。GC 含量往往反映了方向性突变压的强弱，密码子各位上的 GC 含量可以用来统计 GC 含量和密码子用法的关联，尤其是同义密码子第三位上的 GC 含量和密码子偏好性有密切关系。

（4）有效密码子使用个数（Effective Number of Codons，ENC 或 Nc）。由 Wright 于 1990 年提出，首先用式（1－1）定义密码子用法的纯合性：

$$\hat{F}_a = (n_a \sum_{1}^{k} p_i^2 - 1)/(n_a - 1) \tag{1-1}$$

式中：n_a 是氨基酸实际使用的密码子的观察数量，即该氨基酸实际出现次数；p_i 是第 i 个密码子的频率，即 n_i/n_a；k 是编码目标氨基酸的同义密码子种类数目。根据同义密码子个数，氨基酸可分为 5 种类别，分别由 1、2、3、4、6 个密码子编码，对某一类别 r 个密码子编码家族，其平均用法由式（1－2）定义：

$$\overline{\hat{F}_r} = \frac{1}{n_{RC}} \sum_{a \in RC} \hat{F}_a \tag{1-2}$$

式中：n_{RC}是属于该类别的氨基酸数目，则 ENC 由式（1－3）计算：

$$\hat{N}_c = 2 + (9/\overline{\hat{F}_2}) + (1/\overline{\hat{F}_3}) + (5/\overline{\hat{F}_4}) + (3/\overline{\hat{F}_6}) \tag{1-3}$$

ENC 的取值范围从 20～61，当密码子偏好性达到最大程度，每个氨基酸仅有 1 个密码子编码时，ENC 值为 20，当所有密码子平均使用时，ENC 值为 61。

（5）最优密码子频率（Frequency of Optimal codons，Fop）。该指标是指最优密码子和其同义密码子间的比值，和 CAI 一样，前提是需要已知在高表达的基因中的最优密码子。Fop 的值域为 0（没有最优密码子被使用）～1（密码子偏好性最大，只有最优密码子被使用）。

（6）密码子偏好性指数（Codon Bias Index，CBI）。这是另一种直接估计密码子偏好性的指标，同 Fop 相似，需已知在高表达基因中的最优密码子。CBI 的值域也是 0～1，当密码子偏好性最大时，CBI 值为 1，当同义密码子平均使用时，CBI 值为 0。值得注意的是，该指标允许事先设定的最优密码子使用次数比平均使用次数还少的情况出现，此时 CBI 值为负值。

1.2.4 密码对偏好使用的研究现状

密码子上下文关系（Codon Context）是密码子非随机使用研究的重要内容之一。本文的研究集中于密码对的偏好性使用。最近 20 年来，在密码对非随机使用方面，一些学者尤其是生物信息学家在基于实验数据和理论计算分析两方面进行了初步的探讨，并取得了一些有意义的结果[80-85]。研究方法大致可归为两大类：实验方法和统计方法。下面对这些方法进行简要的回顾。

从核苷酸序列出发来研究密码对的使用一直是一个难点。国内有关这方面的报道还非常少。早期 Folley 等人主要是集中在实验基础上进行分析，提出密码对的使用偏好性影响翻译的延伸效率[82]。Gutman、Gurvich、Irwin 等人[83-85]后来的分析方法主要是利用统计学的方法。最近 Buchan 等人运用聚类分析、多变量分析研究了密码对的使用[86]。实验证明，这些方法都得到了不错的结果。例如：发现影响密码对搭配的因素包括密码子的偏好性、二肽的偏好性、二核苷酸的偏好性、密码子的前后文关系、序列的 GC 含量以及转录期间翻译调节信号的潜在驱动力等。总之，近十几年来研究者在密码对非随机使用方面做了初步的研究，取得了一定的成果。总结密码对使用研究的发展历程大体可以分为以下几个方面。

1. 基于实验的方法

已知翻译的延伸速率是不连续的，并且受密码子上下文关系的影响，然而有关翻译速率与密码子上下文之间相互影响的内在机制知之甚少。在 20 世纪 90 年代，研究者提出翻译的速率反映了物种特异的密码子使用与其同源的 tRNA 丰度的关联[48,49]。后来，Smith 等人的实验结果表明参与翻译的核糖体表面 A 和 P 位点相邻 tRNA 的兼容性影响着翻译的速率[64]。1986 年，Kolaskar 和 Reddy[75]分析了原核生物系统的 278 个蛋白质编码序列中密码对的使用，指出相邻密码子之间存在约束，这些约束可能包含 mRNA 三维结构的信息并调控翻译效率。1989 年，Gutman 和 Hatfield[83]基于大肠杆菌 237 个基因中密码对观察频数与理论频数的比较发现：①有些密码对的使用频数偏离其理论频数甚远，即偏好的或稀有的密码对；②密码对的偏好使用具有方向性，如密码对 A－B 与密码对 B－A 的偏好是无关的；③高表达基因避免使用偏好密码对，而低表达基因却避免使用稀有密码对。这些结果暗示至少部分密码对的偏好性与翻译过程有关。因此推测密码对的使用与其同源的两相邻 tRNA 的丰度和结构协同进化去控制翻译速率。1995 年 Irwin 等人通过实验验证了翻译偏好密码对的速率低于非偏好密码对，证明密码对的使用与翻译效率的确存在关联[84]。2005 年，Gurvich 等人的实验证明在大肠杆菌中，基因表达水平影响核糖体在稀有密码对 AGG－

AGG 和 AGA - AGA 位点的移码效率[85]。

2. 基于单个基因组的统计分析方法

2003 年 Boycheva 等人[87]运用统计学方法，分析了大肠杆菌全基因组中密码对的使用，再次证明密码对的使用是非随机的。同时，对大肠杆菌 334 个高表达基因和 303 个低表达基因的分析发现两组基因中密码对的使用存在明显差异。对终止密码对（基因末端有义密码子与终止密码子的组合）的分析发现，偏好使用的终止密码对含有终止密码子 UGA，而避免使用的终止密码对含有终止密码子 UAA。这个结果暗示密码对的使用与翻译终止效率有关。

2005 年 Moura 等人[88]基于统计学开发了密码对分析软件包 ANACONDA 1.0，此软件包的体系结构如图 1.1 所示。

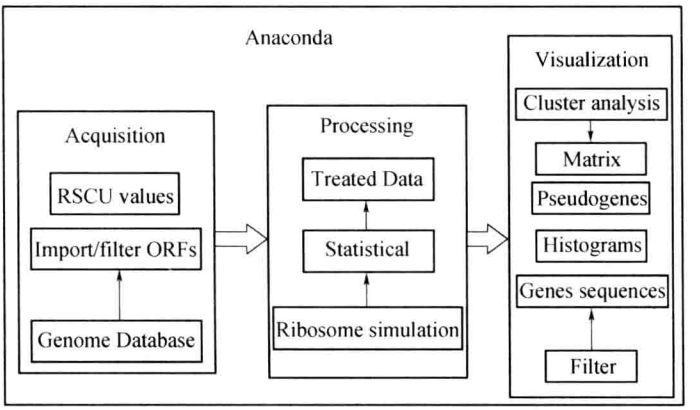

图 1.1　Anaconda 的生物信息系统体系结构[89]

它包括三个模块，首先，数据采集模块（Acquisition）下载 BLAST 格式基因组数据并对数据进行筛选后存放到本地数据库；其次，数据处理模块（Processing）运用核糖体仿真算法对此数据库进行处理，统计基因组中密码对出现频率，由密码对观察值与理论值的关系确定密码对的偏好性；第三，可视化模块（Visualization）将密码对偏好程度以 61 × 64 的可视化彩色图形式显示出来，以不同的颜色代表密码对偏好的程度，可直观地体现密码对的偏好分布。另外，此数据处理模块还包括一系列进行数据可视化分析的工具，如计算同义密码子相对使用频率、密码子适应性指数，密码子第三位点 GC 含量，有效密码子数目等。由此研究人员分析了酿酒酵母、非洲酒裂殖酵母、白念珠菌及大肠杆菌基因组中密码对的使用。发现密码对的偏好使用是物种特异的，并对上述四个物种部分偏好与稀有密码对的组合模式进行了比较分析，从突变和选择压力两方面对密码对偏好的影响进行了讨论，证明突变的偏好对密码对的压力主要反映在偏好性较低的密

码对中。

3. 基于比较基因组学的方法

异同的比较,历来是遗传学以及整个生物学的重要方法。比较基因组学是基因组学的重要分支,是随人类基因组计划,特别是随人类与其他生物基因组的大规模序列分析发展起来的新科学,现已成为研究生物基因组的最重要策略与手段之一。它通过对不同物种的基因组数据进行比较分析,揭示彼此的相似性和差异性,了解不同物种间进化上的差异。如基因位置的比较、基因编码区长度或外显子数的变异[89-110]、基因组上非编码区的比例、进化关系较远的物种间高度保守区域的比较分析等[111-122]。同时随着密码对偏好使用研究的深入,利用比较基因组学方法对不同进化水平的基因组进行比较分析,成为当前有关密码对研究的重要内容之一,并取得了有意义的结果。

例如:2006 年 Buchan 等人[87]在统计学基础上用比较基因组学方法分析了 17 个物种基因组中密码对的使用,发现稀有密码子的组合通常在真核生物中是避免的,而在原核生物中是偏好的;在单细胞的真核生物中,tRNA 的间接选择对密码对的偏好有显著影响。

2007 年,在 ANACONDA1.0 的基础上,Moura 等人[123]对此软件进行了改进,开发了 ANACONDA2.0,对包括细菌、古菌和真核生物的 119 个物种进行比较分析。ANACONDA2.0 的分析流程是:首先由 ANACONDA2.0 将单个物种 $61 \times 64 = 3904$ 的密码对偏好分布二维彩色图转换为 3904×1 的分布,对 119 个物种得到 3904×119 的分布图,图中可以对 119 个物种密码对的偏好性直观地进行比较。其次,为了探讨基因组复制偏好性与 mRNA 翻译偏好对密码对偏好的影响,ANACONDA2.0 也提供对全基因组二核苷酸的偏好性进行可视化的图示分析。利用此软件,Moura 等人对 119 个物种分析之后提出细菌和古菌 mRNA 一级结构进化主要来自翻译机制的约束,而真核生物密码对偏好的重要因素是 DNA 甲基化和三核苷酸重复。

最近,Tats 等人[124]运用统计的方法研究了 138 个物种密码对的使用,提出了相对密码对使用频率(RDCU)的计算方法,此方法基于四组密码对的观察频数与理论频数之比,分别为相邻密码子(1~2)、间隔一个密码子(1~3)、间隔两个密码子(1~4)和间隔 8 个密码子(1~10)的密码对组合。然后基于同义密码子相对使用频率和相对密码对使用频率,运用 Spearman 相关系数分别计算了物种之间密码子和密码对的关联,由此得到密码子与密码对的分组,进而得到同义密码子相对使用频率和相对密码对使用频率的进化树。结果发现在古菌、细菌和真核生物中部分紧邻密码对的偏好性十分保守,最普遍避免的密码对模式 nnUAnn、nnGGnn、nnGnnC、nnCGCn、GUCCnn、CUCCnn、nnCnnA 和 UUCGnn。最