

【研究简报】

进化中蛋白质编码序列结构的随机化

葛伟文^① 贺福初^②(军事医学科学院^①卫生学环境医学研究所, 天津 300050; ^②放射医学研究所, 北京 100850)**关键词** χ^2 检验 基因功能与结构 进化 基因分析

核酸测序技术的发展, 为进化研究提供了丰富、精细的素材. 同源分析的结果, 使进化的概念更加直观. 但是, 现有的研究方法往往只利用核酸的一维信息, 未能穷尽序列资料所包含的广博内涵. 对核酸序列资料的充分利用呼唤引入更多更有效的方法.

按照构效学基本原理, 编码蛋白质的基因必然具有其独特的结构特征, 本文正是希望找到能够反映这种特征的指标, 并以此观察基因结构在进化过程中的变化事件及其趋势.

众所周知, 一条蛋白质编码序列上 4 种核苷酸在密码子 3 个位点上的分布是非随机性的^[1]. 这种定位的非随机性可能正是对应于编码功能的结构特征. 本文以 χ^2 检验反映这种

的同源异相的岩浆演化系列, 共同表明南秦岭曾经历古陆块的裂解过程, 并向显生宙洋盆转化. 鉴于南秦岭是扬子板块的重要组成部分, 这一认识提供的重要信息是晚元古代扬子板块可能经历过裂解作用.

(2) 基性岩墙群的发育是区域存在已固结大陆块的重要标志, 并表明该陆块已具有一定的刚性^[6]和相当的规模. 南秦岭区的同类基底岩块中不只武当地块中发育基性岩墙群, 其东延的随-枣地块中亦发育, 构造线一致, 自然而然的推论是南秦岭乃至其南的扬子区存在 800 Ma $\pm$ 以前固结的古陆块, 这与现今公认的区域地质相一致, 这个古陆块的固结可能是一次重大造山事件的产物^[4].

武当地块基性岩墙群的研究可能为扬子古陆壳的形成演化提供重要信息, 至于它与近年来国内外地学界关注的元古宙泛大陆的形成与裂解的关系尚待进一步探讨.

致谢 本工作为国家自然科学基金(批准号:49572145)资助项目.

参 考 文 献

- 1 张国伟, 张宗清, 董云鹏. 秦岭造山带主要构造岩石地层单元的构造性质及其大地构造意义. 岩石学报, 1995, 11(2): 101~114
- 2 张宗清, 张国伟, 付国民, 等. 秦岭变质地层年龄及其构造意义. 中国科学, D 辑, 1996, 26(3): 216~222
- 3 秦正永, 雷世和. 武当群成岩年龄新资料并兼论. 中国区域地质, 1996, 42(2): 176~185
- 4 Zhou Dingwu, Liu Liang, Hua Hong, et al. A discussion of the Middle and Late Proterozoic tectonic regime of Qinling-Dabie Orogenic Belt. Journal of Northwest University (in Chinese), 1996, 26: 133~141
- 5 程裕淇主编. 中国区域地质概论. 北京: 地质出版社, 1994. 170
- 6 Brian F, Windley B F. The Evolving Continents. New York: John Wiley and Sons, 1984

(1997-01-23 收稿, 1997-05-07 收修改稿)

非随机性特征,并比较不同组别间的 χ^2 值. 结果表明, χ^2 检验是一种判断编码基因结构特异性的理想方法. 在此基础上,本文以 χ^2 值为指标观察了假基因在物种内及谷氨酰胺合成酶基因在物种间的编码结构的改变,得出以下结论:蛋白质编码序列在横向(同源基因间)及纵向(物种间)进化中均存在结构的随机化. 最后,本文还对随机化的机制及其在生命起源中的意义进行了粗略探讨.

1 材料与方法

核酸序列数据通过 Internet 取自 GenBank 数据库.

从序列起始位点,以 3 为重复周期,将碱基编码为 1, 2, 3, 1, 2, 3, ..., 以此重复;分别计算位点 1 上 G, A, T, C 的出现次数,计为 g_1 , a_1 , t_1 及 c_1 , 同样计算其他位点上的这些数据,列成 4 碱基 $\times$ 3 位点联列表,如表 1 所示.

计算表中每个观察值与按完全随机原则推算出的理论值(所在行和 $\times$ 所在列和/总和)

之间的差值,并以 χ^2 值表达序列中碱基组成和排列与完全随机状态之间的差异: $\chi^2 = \sum ((\text{观察值} - \text{理论值})^2 / \text{理论值})$, 由国际通用的 SAS6.02 版本统计软件包执行运算.

为了说明重复周期对 χ^2 值意义的影响,我们也计算序列在其他重复周期中的 χ^2 值,此时就将碱基按 1, 2, 3, ..., n , 1, 2, 3, ..., n , ... 编号,并将数据列成 4 碱基 $\times$ n 位点联列表,同理计算 χ^2 值.

2 结果

2.1 χ^2 检验识别三联密码子规则及序列的编码能力

为了考察 χ^2 检验鉴别编码序列及三联密码子结构的能力,本文以大肠杆菌核糖体蛋白编码区(CDS)及相同长度的一条由计算机按完全随机原则产生的随机序列为对象,计算了不同碱基数重复周期下序列的 χ^2 值,结果列于表 2. 可以看到, CDS 的 χ^2 值在 3 及 3 的倍数碱基的重复周期中显著较高,而在其他碱基数的周期中则很低,说明编码序列的核苷酸组成及排列在三联体结构中表现出高度的特异性,而在非三联体周期中则极度地随机化;在完全随机序列中, χ^2 值没有周期性特征,并且都比较小. 此结果表明,较高 χ^2 值鉴别出了编码序列的一种结构特征,这种特征来自于编码功能所要求的各种条件的制约,是序列编码功能的结构基础. 本文随后出现的 χ^2 值都是所分析序列以 3 碱基为重复周期时得出的计算值.

为了进一步观察 χ^2 值判断序列编码结构的能力,本文比较了基因外显子与内含子

表 1 某一序列的 4 碱基 $\times$ 3 位点联列表

碱基	G	A	T	C	总和
位点 1	g_1	a_1	t_1	c_1	N_1
位点 2	g_2	a_2	t_2	c_2	N_2
位点 3	g_3	a_3	t_3	c_3	N_3
总和	G	A	T	C	N

表 2 核酸不同重复周期的 χ^2 值比较

重复周期碱基	序列	
	大肠杆菌核糖体蛋白 CDS	完全随机序列
2	0.03	1.25
3	369.57	10.73
4	0.25	4.21
5	0.34	4.17
6	369.70	14.85
7	0.29	16.78
8	1.07	15.53
9	369.88	26.77
10	0.64	17.89

(图 1)、mRNA 翻译区与两端非翻译区(表 3)、翻译外显子与非翻译外显子(图 2)、蛋白质编码序列与 tRNA 基因及 rRNA 基因(表 4)序列的 χ^2 值.

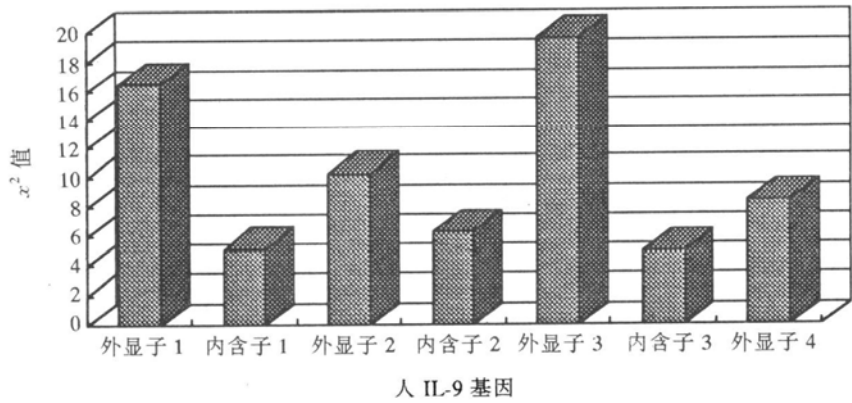


图 1 外显子与内含子的 χ^2 值比较

表 3 mRNA 翻译区与非翻译区的 χ^2 值比较

基因序列	χ^2 值		
	5'-非翻译区	翻译区	3'-非翻译区
人 IL-2	8.71	43.31	3.02
人 IL-5	4.78	15.30	3.36
人 IL-8 受体	6.17	111.42	16.07
果蝇核层蛋白 Dm0	5.09	244.59	7.71
人 c-fos	7.67	60.69	0.83
人 ATP 酶	3.64	98.33	12.60
人 EPO	2.11	28.81	4.85
小鼠 EPO 受体	5.63	39.19	7.52
牛 IL-4	3.52	33.27	8.32

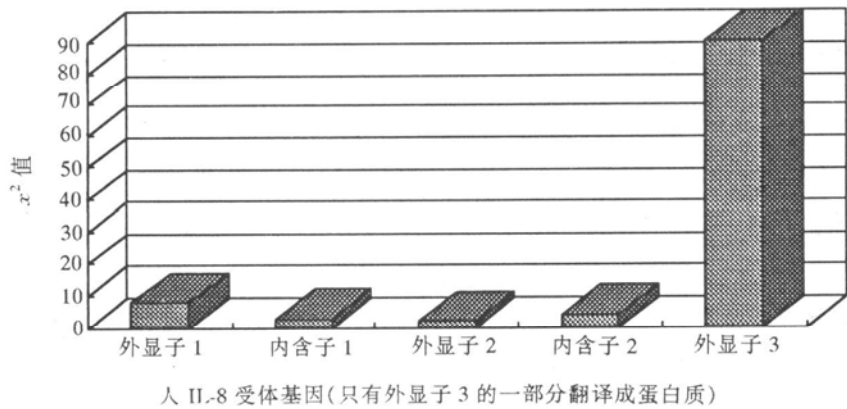


图 2 基因翻译区与非翻译区的 χ^2 值比较

表 4 mRNA 翻译区、tRNA、rRNA 序列的 χ^2 值比较

mRNA 翻译区		tRNA		rRNA	
基因	χ^2 值	基因	χ^2 值	基因	χ^2 值
牛 IL-2	16.09	Gln(UUG)	5.14	小麦叶绿体 23S	6.25
牛 IL-2 受体	26.30	Ser(GCU)	9.85	16S	5.64
果蝇核纤层 Dm0	244.59	Ser(UGA)	5.54	5S	8.03
果蝇 RNA 聚合酶 II 延伸因子	113.51	Gly(GCC)	5.75	4.5S	4.28
大鼠心脏细胞色素氧化酶 IV c 亚基	16.30	fMet(CAU)	7.58	人线粒体 16S	7.42
人组蛋白 H3	53.68	Thr(GGU)	2.39	12S	8.03
人组蛋白 H1	81.64	Glu(UCC)	3.88	醋酸杆菌 16S	10.76
人 GM-CSF 受体	14.68	Tyr(GUA)	6.93	梭状芽孢杆菌 16S	12.32
大肠杆菌 S-腺苷甲硫氨酸合成酶	78.11	Cys(GCA)	5.14	人 28S	4.75
猪 GM-CSF	31.63	Arg(UCU)	4.37	18S	3.87

通过上述比较可以看出,较高的 χ^2 值往往对应于序列的编码特性,而编码蛋白质的序列也往往具有较高的 χ^2 值. 表明,基于密码子三联体原则进行的核酸结构的 χ^2 检验可以反映核酸的蛋白质编码性能,透过序列的一维信息,揭示核酸结构与功能间的相关性. χ^2 值可以作为一个量化的、联系核酸功能与结构的客观指标.

2.2 χ^2 值分析揭示蛋白质编码基因在进化过程中的结构随机化

2.2.1 蛋白质编码序列结构特异性在横向进化过程中呈现进行性退化

我们在比较功能性基因与对应假基因之间的结构特异性时观察到(表 5),假基因的 χ^2 值都低于对应的功能性基因,说明核酸链的结构在假基因中发生随机化,这种变化与假基因已丧失蛋白质编码功能的事实是相符的,反映出基因序列特异性在横向进化(即物种内部同源基因的衍化)过程中的退化行为.

结果还表明, χ^2 值能反映基因功能及相应结构的动态变化,随着基因蛋白质编码功能的丧失,基因序列结构的特异性也随之丧失,表现为 χ^2 值的不断下降.

2.2.2 纵向物种进化过程中基因序列结构的进行性随机化

为了观察物种进化过程中其具体基因在结构特异性上的变化趋势,我们利用 χ^2 检验比较分析了不同生物谷氨酰胺合成酶基因的结构特征. 谷氨酰胺合成酶基因是一个古老的基因,在所分析过的现存的所有物种中都存在,并且彼此间存在良好的同源性^[2]. 表 6 列举了多种生物谷氨酰胺合成酶基因的 χ^2 值.

表 5 功能性基因与对应假基因的 χ^2 值比较

基因		χ^2 值
人 IL-9 受体	功能基因	50.22
	假基因(10 号染色体)	38.35
	假基因(16 号染色体)	13.37
人谷氨酰胺合成酶	功能基因	61.14
	假基因	8.50
小鼠细胞质骨架 γ -肌动蛋白	功能基因	109.26
	假基因	15.33
	γ -119-假基因	20.01
	γ -36-假基因	7.30
大鼠 S-腺苷甲硫氨酸脱羧酶	功能基因	22.85
	假基因	7.15

表 6 不同物种谷氨酰胺合成酶基因 χ^2 值比较

分类	物种	χ^2 值
原核生物	豆科植物根瘤菌 <i>R. leguminosarum</i>	269.37
	豆科植物根瘤菌 <i>B. japonicum</i>	193.05
	醋酸盐梭状芽胞杆菌 <i>C. acetobutylicum</i>	184.33
	硫杆菌 <i>T. ferrooxidans</i>	173.92
	鼠伤寒沙门氏菌 <i>S. typhimurium</i>	132.63
	大肠杆菌 <i>E. coli</i>	105.97
	酵母	76.02
真核生物	豌豆	67.47
	小鼠	66.63
	中国仓鼠	59.34
	人	61.14

研究发现,在物种进化过程中,谷氨酰胺合成酶基因的 χ^2 值不断下降,表明基因不断丢失其结构特异性和信息载荷,这在假基因的 χ^2 值分析中已经观察到,是一种基因的退化行为。由于在至今所发现的最古老生物的化石中就发现有谷氨酰胺合成酶分子,而即使在动物细胞中,谷氨酰胺合成酶仍负责将体内有毒的氨转变成无毒的谷氨酰胺并加以贮存与转运,因而始终具有极其重要、不可或缺的生物意义。所以谷氨酰胺合成酶基因结构的退化行为很难以酶功能重要性的衰退来解释,而可能代表着一种具有普遍意义的事件,即在物种个体日益完善复杂的同时,其具体基因结构及功能却发生着进行性的随机化和退化。

3 讨论

自发突变的随机效果及其持久性可能解释序列结构的进行性随机化。然而有关基因结构进行性随机化的观点还需要回答一个复杂的问题,即最原始基因的高度特异性结构是如何获得的?

蛋白质编码基因之所以具有结构的特异性,归根到底来自于其为蛋白质编码这一功能特点。研究也表明,不编码蛋白质的 tRNA 和 rRNA 则不具有 χ^2 值所反映的结构特异性。此结果提示我们,最原始的蛋白质编码基因在其起源过程中,受对应蛋白质的制约,而不是独立于蛋白质仅依靠核苷酸之间的随机接合形成的。随机接合形成的核苷酸链相当于我们在计算机上模拟生成的随机链,必定具有极低的 χ^2 值和极随机的结构特性,很难从中选择出具有蛋白质编码功能的基因序列。基于此,我们大胆假设,至少对于蛋白质编码基因而言,核酸序列的合成受所编码氨基酸链的指导,即先有蛋白质,后有为其编码的核苷酸链。文献[3]曾经在试管中观察到氨基酸与核苷酸及核苷酸组合之间存在天然对应的化学亲和力。这种亲和力可能介导了在生命完善以前所发生的在蛋白质指导下的蛋白质编码基因的序列合成。

致谢 本文数据的取得及思路发展,得到本院生物工程研究所梁龙博士的热心指导和帮助,特此感谢。本工作为国家自然科学基金(批准号:39570275)及杰出青年科学基金(批准号:39625014)资助项目。

参 考 文 献

- 1 Staden R. Measurement of the effects that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Research*, 1984, 12(1): 551~567

【研究简报】

JAK3 介导白细胞介素-2 对 c-myc
基因表达的诱导朱锦芳^① 鲁林荣^① 季红斌^① 郑仲承^①徐静芬^① J. Thèze^② 刘新垣^①

(①中国科学院上海生物化学研究所, 上海 200031; ②Pasteur Institute, Paris)

关键词 白细胞介素-2(IL-2) JAK3 (Janus Kinase 3) c-myc 基因 信号转导

白细胞介素-2 (IL-2) 与其靶细胞表面的 IL-2 受体 (IL-2R) 相互作用并诱导 c-myc 基因表达, 这一过程与 IL-2 刺激 T 淋巴细胞增殖有关, 但其具体机制与过程还不明了^[1]. IL-2 受体由 α , β , γ 3 亚基组成. 其中 α 亚基可提高受体与 IL-2 的亲合力, 而 β , γ 亚基与信号转导有关^[2]. 有些实验认为, 与 IL-2R γ 胞内域结合的非受体型酪氨酸激酶 JAK3 可能介导了 IL-2 对 c-myc 基因表达的诱导, 但缺乏直接证据^[3]. 为了直接验证 JAK3 参与了这一过程, 我们构建了一个嵌合受体基因 IL-2R $\alpha/\gamma/\Delta$ NJAK3, 其胞外域来自白细胞介素-2 受体 α 亚基 (IL-2R α), 跨膜部分来自 IL-2R γ , 胞内域来自 JAK3 催化功能域 (Δ NJAK3). 将此嵌合受体转入已经稳定表达 IL-2R β 的小鼠成纤维细胞 L929 (L929 β) 中, 筛选到稳定表达 IL-2R β 和 IL-2R $\alpha/\gamma/\Delta$ NJAK3 的细胞株 L929 $\beta + \alpha/\gamma/\Delta$ NJAK3. IL-2 作用于此转基因细胞株可以诱导 c-myc 基因表达. 由于在这个系统中, 没有 IL-2R γ 胞内域, 从而排除了结合于 IL-2R γ 胞内域的其他分子的干扰, 因此这个实验结果为 JAK3 在 IL-2 诱导 c-myc 基因表达中起重要作用提供了直接证据.

1 材料与方法

(i) 嵌合受体表达质粒构建. 质粒 pSVK3IL-2R $\alpha/\gamma/\gamma$, pSVK3IL-2R α 及 pSVLIL-2R β 为本实验室构建, pBSKJAK3 由 Dr. Ihle 惠赠. 从 pSVK3IL-2R $\alpha/\gamma/\gamma$ 中切出编码 IL-2R α 胞外域和 IL-2R γ 跨膜部分的基因片段, 将它克隆入 pUC6S 并鉴定插入方向, 得到重组子 pUC6S α/γ . 从 pBSKJAK3 中切出编码 JAK3 催化功能域的基因片段, 将它克隆入 pUC6S α/γ , 形成杂合受体基因 $\alpha/\gamma/\Delta$ NJAK3. 最后将杂合受体基因 $\alpha/\gamma/\Delta$ NJAK3 切出并克隆入表达载体 pSVK3. 用 IL-2R α 胞外域替代 IL-2R γ 胞外域是为了提高杂合受体与 IL-2 的亲合力, 并方便筛选.

(ii) 细胞培养及单克隆筛选. 小鼠成纤维细胞 L929 用 DMEM 培养液在 37℃, 5% CO₂ 培养箱中培养. 细胞转染按照 LipofectAMINE 的产品说明 (LIFE TECHNOLOGIES). 转染稳定株时用含抗药筛选标记的质粒 pCDNA3 或 pDR2 与表达受体的质粒共转染, 48 h 后, 分

2 Kumada Y, Benson D R, Hillemann D, et al. Evolution of the glutamine synthetase gene, one of the oldest existing and functioning genes. Proc Natl Acad Sci USA, 1993, 90(7): 3 009~3 013

3 Calvin M. Chemical evolution. Chem Br, 1969, 5(1): 22~28

(1997-03-27 收稿, 1997-07-28 收修改稿)