拟南芥基因密码子偏爱性分析

范三红* 郭蔼光 单丽伟 胡小平
(西北农林科技大学, 陕西省农业分子生物学重点实验室, 杨凌 712100)

摘要 密码子偏爱性对外源基因的表达强度有一定影响, 特别是编码蛋白质 N 端 7~ 8 个氨基酸残基的密码子. 通过对拟南芥染色体中 26 827 个蛋白质对应的基因密码子进行分析, 得到了编码氨基酸的 61 种密码子在拟南芥中的使用频率, 并与大肠杆菌和哺乳动物进行了比较, 结果表明三者间的密码子偏爱性有较大差异. 这一分析结果对于动物基因在植物中的表达, 及植物基因在微生物中的表达具有一定指导意义. 同时提供了一种直接以 XML 文档为数据源解析巨型 XML 格式染色体数据的方法.

关键词 拟南芥, 密码子偏爱性, XML
学科分类号 Q811.4

密码子是核酸携带信息和蛋白质携带信息间对应的基本规则, 是生物体内信息传递的基本环节. 编码 20 种氨基酸的密码子共 61 种. 对于不同的物种来说, 不同密码子在基因中出现的频率有很大不同. 一方面, 对同一种氨基酸而言, 编码该氨基酸的不同密码子的比率在不同物种中有差异; 另一方面, 对同一个密码子而言, 在不同的物种中出现的频率有所不同. 这种现象会导致外源基因在宿主细胞中的表达量降低. 通过改造目的基因的密码子, 提高动物基因在细菌中的表达已有不少成功的例子, 但关于植物密码子偏爱性的研究还少见报道. 本文以模式植物拟南芥为对象[1], 分析了拟南芥基因密码子偏爱性, 希望能够对动物基因在植物中的表达及植物基因在细菌中的表达提供参考.

XML (extensible markup language) 是一种用于描述结构化数据的可扩展的标注语言, 是网络应用程序数据交流及商业通讯中可交换数据格式的标准[2,3]. 由于其本身具备很强的描述能力和很好的灵活性, 在生物信息学领域的信息表述和数据交换中已得到广泛的应用, 而且这种趋势还在不断增强[4]. TIGR 是国际权威的基因组研究中心之一, 2001 年他们提供了一种用 XML 格式的拟南芥染色体数据, 从 TIGR 的 XML 格式染色体数据可以充分体现出 XML 在复杂生物学数据描述中的灵活性和巨大潜力. 然而每一个染色体的数据都是一个巨型的 XML 格式文件, XML 染色体文档的复杂性和超长序列字符串的存在使许多 XML 操作软件都不能正常运行. 本文通过 DELPHI 提供的 XML 文件接口实现了对 TIGR 染色体数据的处理, 并在此基础上完成了对拟南芥基因密码子偏爱性的统计.

1 数据和分析方法

1.1 数据

本文以 TIGR 的染色体数据为基础 (FTP://FTP.TIGR.ORG.). TIGR 的 XML 格式染色体数据中各种元件的关系结构如图 1 所示. 从图 1 中可以看出, 一个 ASSEMBLY (装配染色体) 元件包括基因列表 (GENE_LIST) 元件、装配序列 (ASSEMBLY_SEQUENCE) 元件等. GENE_LIST 元件包括蛋白质编码 (PROTEIN_CODING) 基因和 RNA 基因 (RNA_GENE). 蛋白质编码基因 (TU) 包括基因信息 (GENE_INFO)、基因模型 (MODEL)、转录序列 (TANSCRIPT_SEQUENCE) 等. 基因模型包括内含子、外显子等. 总的来说, TIGR 的 XML 格式的染色体数据是一个接近完备的染色体表述形式, 通过 XML 格式的描述使研究者更容易从全局分析一个基因在整个基因组中的位置及作用. 本文中, 人和大肠杆菌的密码子偏爱性数据来自 Codon usage database[5] (http://www.kazusa.or.jp/codon).

* 通讯联系人.
Tel: 029-7092262, E-mail: sanhongfan@yahoo.com.cn
收稿日期: 2002-08-12, 接受日期: 2002-11-20

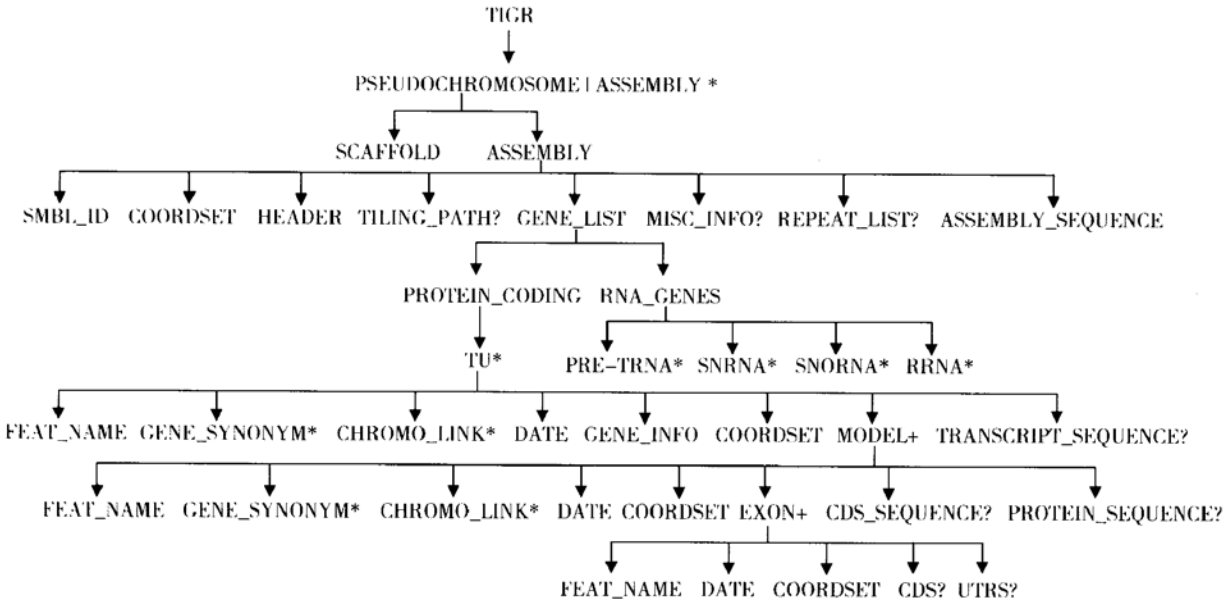

Fig. 1 The major elements and mutual relationship in TIGR's chromosome data

1.2 数据分析方法

为了对 XML 格式染色体数据的操作，首先利用 BORLAND DELPHI 提供的以 XML 数据为数据源的数据库技术，将 XML 格式染色体数据转化成数据库形式. DELPHI 提供了实用工具 XMLMAPPER，该工具可轻松建立数据库字段和 XML 元件之间的对应关系. 然后以这种关系为基础，将 XML 格式染色体数据中所有的蛋白质编码基因信息读入数据库，根据每一个基因在染色体上的位置抽取每一个基因对应的核酸序列. 基因模型中记录了每一个基因的内含子和外显子在染色体上的位置，根据基因模型的信息再拼接出每一个基因的编码区序列. 翻译每一个编码区序列，并统计各种密码子的使用频率[6]. 数据分析的基本流程如图 2 所示.

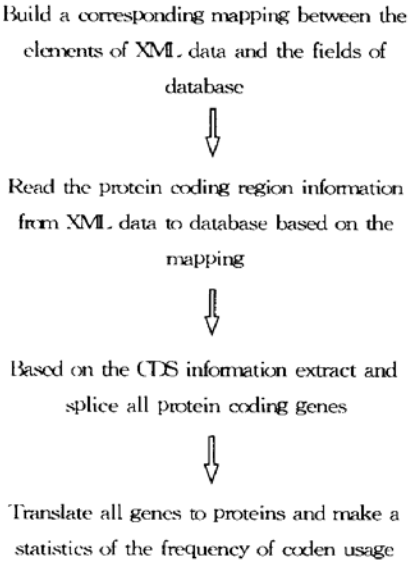

Fig. 2 The flow chart of data analysis

2 结果与分析

拟南芥蛋白质编码基因中各种密码子的出现次数，及编码同一种氨基酸的不同密码子出现的百分率如表 1 所示. Count 为每种密码子在整个拟南芥基因组中出现的次数. Percentage 表示编码同一种氨基酸的不同密码子所占的百分率. 从表 1 中可以看出，在拟南芥中，终止密码子 UAG 出现的次数要远低于其他两种终止密码子. 编码 Arg 的 6 种密码子中 AGA 出现率为 35.69%，而 CGC 出现率只有 6.88%. 编码 Cys 的偏爱密码子为 UGU，编码 Asp 的偏爱密码子为 GAU，编码 His 的偏爱密码子为 CAU.

拟南芥、大肠杆菌和人的密码子使用频率的比较如表 2 所示. *Ara*(%)、*E. coli*(%)、Human(%) 分别表示拟南芥、大肠杆菌和人的一种密码子出现次数分别占该物种总密码子出现次数的百分比. *Ara*/hum、*Ara*/*Eco*、*Eco*/hum 分别表示 3 个物种每种密码子出现百分比的比值. 通过表 2 可以发现，拟南芥和大肠杆菌、大肠杆菌和人的密码子偏爱性有显著的差异. 拟南芥和大肠杆菌的密码子出现频率比值中有 20 种密码子的值大于等于 2.0 或小于等于 0.5；大肠杆菌和人的密码子出现百分率比值中有15种密码子的值大于等于2.0或小于等

Table 1 Count of every codon and percentage of codons for every amino acid in *Arabidopsis thaliana*

Amino acid	Codon	Count	Percentage	Amino acid	Codon	Count	Percentage
Ala(A)	GCT	299 447	43.02	Pro(P)	CCT	202 463	38.14
	GCC	107 596	15.46		CCC	57 681	10.87
	GCA	193 372	27.78		CCA	179 059	33.73
	GCG	95 589	13.73		CCG	91 598	17.26
Cys(C)	TGT	123 400	60.45	Gln(Q)	CAA	220 367	57.20
	TGC	80 725	39.56		CAG	164 898	42.80
Asp(D)	GAT	416 904	68.74	Arg(R)	CGT	96 508	16.11
	GAC	189 628	31.26		CGC	41 193	6.88
Glu(E)	GAA	394 403	52.39		CGA	71 716	11.97
	GAG	358 436	47.61		CGG	54 207	9.05
Phe(F)	TTT	253 598	53.04		AGA	213 787	35.69
	TTC	224 519	46.96		AGG	121 678	20.31
Gly(G)	GGT	237 279	33.60	Ser(S)	TCT	280 575	28.16
	GGC	98 311	13.92		TCC	121 407	12.19
	GGA	258 143	36.55		TCA	205 984	20.68
	GGG	112 560	15.94		TCG	101 532	10.19
His(H)	CAT	158 382	62.51		AGT	162 022	14.49
	CAC	95 002	37.49		AGC	124 706	12.52
Ile(I)	ATT	244 028	41.18	Thr(T)	ACT	193 115	34.00
	ATC	201 460	34.00		ACC	111 456	19.63
	ATA	147 120	24.83		ACA	178 480	31.43
Lys(K)	AAA	351 259	49.31		ACG	84 877	14.95
	AAG	361 077	50.69	Val(V)	GTT	301 496	40.42
Leu(L)	TTA	149 050	14.12		GTC	137 960	18.49
	TTG	239 171	22.65		GTA	115 334	15.46
	CTT	267 577	25.34		GTG	191 171	25.63
	CTC	174 083	16.49	Trp(W)	TGG	140 546	100.00
	CTA	114 224	10.82	Tyr(Y)	TAT	170 130	53.35
	CTG	111 680	10.58		TAC	148 772	46.65
Met(M)	ATG	271 269	100.00	STOP	TAA	10 564	36.30
Asn(N)	AAT	260 432	53.08		TAG	6 279	21.58
	AAC	230 188	46.92		TGA	12 256	42.12

Table 2 Comparison of codon preference among Human, *E. coli* and *Arabidopsis thaliana*

Amino acid	codon	*Ara* (%)	*E. coli* (%)	Human (%)	*Ara*/hum	*Ara*/*Eco*	*Eco*/hum
Ala(A)	GCT	2.69	1.70	1.86	1.45	1.58	0.91
	GCC	0.97	2.42	2.8	0.34	0.40	0.85
	GCA	1.74	2.12	1.6	1.09	0.82	1.33
	GCG	0.86	3.01	0.76	1.13	0.29	3.96
Cys(C)	TGT	1.11	0.52	0.99	1.12	2.13	0.53
	TGC	0.73	0.61	1.22	0.60	1.20	0.50
Asp(D)	GAT	3.74	3.27	2.23	1.68	1.14	1.47
	GAC	1.70	1.92	2.60	0.65	0.89	0.74
Glu(E)	GAA	3.54	3.92	2.91	1.22	0.90	1.35
	GAG	3.22	1.87	4.08	0.79	1.72	0.46
Phe(F)	TTT	2.28	2.21	1.69	1.35	1.03	1.31
	TTC	2.02	1.61	2.04	0.99	1.25	0.79
Gly(G)	GGT	2.13	2.55	1.08	1.97	0.84	2.36
	GGC	0.88	2.72	2.29	0.38	0.32	1.19
	GGA	2.32	0.95	1.63	1.42	2.44	0.58
	GGG	1.01	1.13	1.64	0.62	0.89	0.69
His(H)	CAT	1.42	1.25	1.04	1.37	1.14	1.20
	CAC	0.85	0.93	1.49	0.57	0.91	0.62
Ile(I)	ATT	2.19	2.98	1.57	1.39	0.73	1.90

续 表

Amino acid	codon	*Ara*(%)	*E. coli* (%)	Human (%)	*Ara*/hum	*Ara*/*Eco*	*Eco*/hum
	ATC	1.81	2.37	2.15	0.84	0.76	1.10
	ATA	1.32	0.68	0.71	1.86	1.94	0.96
Lys (K)	AAA	3.15	3.53	2.40	1.31	0.89	1.47
	AAG	3.24	1.24	3.29	0.98	2.61	0.38
Leu (L)	TTA	1.34	1.43	0.72	1.86	0.94	1.99
	TTG	2.15	1.30	1.25	1.72	1.65	1.04
	CTT	2.40	1.19	1.27	1.89	2.02	0.94
	CTC	1.56	1.02	1.94	0.80	1.53	0.53
	CTA	1.03	0.42	0.69	1.49	2.45	0.61
	CTG	1.00	4.84	4.02	0.25	0.21	1.20
Met (M)	ATG	2.44	2.64	2.23	1.09	0.92	1.18
Asn (N)	AAT	2.34	2.06	1.67	1.40	1.14	1.23
	AAC	2.07	2.14	1.95	1.06	0.97	1.10
Pro (P)	CCT	1.82	0.75	1.73	1.05	2.43	0.43
	CCC	0.52	0.54	2.00	0.26	0.96	0.27
	CCA	1.61	0.86	1.67	0.96	1.87	0.51
	CCG	0.82	2.09	0.70	1.17	0.39	2.99
Gln (Q)	CAA	1.98	1.46	1.18	1.68	1.36	1.24
	CAG	1.48	2.84	3.46	0.43	0.52	0.82
Arg (R)	CGT	0.87	2.00	0.47	1.85	0.44	4.26
	CGC	0.37	1.97	1.09	0.34	0.19	1.81
	CGA	0.64	0.38	0.63	1.02	1.68	0.60
	CGG	0.49	0.59	1.19	0.41	0.83	0.50
	AGA	1.92	0.36	1.15	1.67	5.33	0.31
	AGG	1.09	0.21	1.14	0.96	5.19	0.18
Ser (S)	TCT	2.52	1.04	1.46	1.73	2.42	0.71
	TCC	1.09	0.91	1.74	0.63	1.20	0.52
	TCA	1.85	0.89	1.17	1.58	2.08	0.76
	TCG	0.91	0.85	0.45	2.02	1.07	1.89
	AGT	1.46	0.98	1.19	1.23	1.49	0.82
	AGC	1.12	1.52	1.93	0.58	0.74	0.79
Thr (T)	ACT	1.73	1.03	1.28	1.35	1.68	0.80
	ACC	1.00	2.20	1.92	0.52	0.45	1.15
	ACA	1.60	0.92	1.48	1.08	1.74	0.62
	ACG	0.76	1.37	0.62	1.23	0.55	2.21
Val (V)	GTT	2.71	1.98	1.09	2.49	1.37	1.82
	GTC	1.24	1.43	1.46	0.85	0.87	0.98
	GTA	1.04	1.16	0.70	1.49	0.90	1.66
	GTG	1.72	2.44	2.89	0.60	0.70	0.84
Trp (W)	TGG	1.26	1.39	1.28	0.98	0.91	1.09
Tyr (Y)	TAT	1.53	1.74	1.20	1.28	0.88	1.45
	TAC	1.34	1.22	1.56	0.86	1.10	0.78
Stop	TAA	0.09	0.20	0.07	1.29	0.45	2.86
	TAG	0.06	0.03	0.06	1.00	2.00	0.50
	TGA	0.11	0.10	0.13	0.85	1.10	0.77

Data with [ ] means that the rate is ≥2.0 or ≤0.5.

于0.5. 拟南芥和人的密码子偏爱性也有显著不同，但拟南芥和大肠杆菌及大肠杆菌和人之间的差异要小，有9种密码子在使用频率上有较大差异.

3 讨 论

TIGR拟南芥染色体中蛋白质编码基因大多是通过生物信息学方法预测的结果，因而以此数据为基础分析拟南芥基因密码子偏爱性具有一定偏差. 但在目前通过实验验证的蛋白质序列还不充分的条件下，使用TIGR的染色体数据分析密码子偏爱性还是切实可行的. 另外，每一个基因的表达会受到多种因素的影响，如基因在染色体中的位置，上游的顺式作用元件，转录后加工，mRNA的稳定性等都会影响基因的表达，是否满足密码子的偏爱性只是影响目的基因表达的因素之一. 利用植物反应器生产药用蛋白、抗体、疫苗是基因工程研究的新亮点，然而，通过密码子的改造提高目的蛋白在植物中表达的研究还少见报道，本文的分析结果可为动物基因在植物中的表达提供参考.

参 考 文 献

1 The *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 2000, **408** (6814): 796~815

2 Mackenzie D. New language could meld the web into a seamless database. Science, 1998, **280** (5371): 1840~1841

3 Lawrence S, Giles C L. Accessibility of information on the web. Nature, **400** (6740): 107~109

4 Vaysseix G, Barillot E. XML, bioinformatics and data integration. Bioinformatics, 2001, **17** (2): 115~125

5 Nakamura Y, Gojobori T, Ikemura T. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucl Acids Res, 2000, **28** (1): 292

6 范三红，张大鹏. Manipulating TIGR's chromosome data in XML format. 西北大学学报（自然科学版），2002, **32** (增刊): 105~109
Fan S H, Zhang D P. J Northwest University (Natural Science Edition), 2002, **32** (supplement): 105~109

Analysis of Genetic Code Preference in *Arabidopsis thaliana*

FAN San-Hong*, GUO Ai-Guang, SHAN Li-Wei, HU Xiao-Ping

(Northwest Sci-Tech University of Agriculture and Forestry, Key Laboratory of Agricultural Molecular Biology in Shaanxi Province, Yangling 712100, China)

Abstract The frequency of codon usage often affects the expression of foreign gene in transgenic research. A statistics of the frequency of codon usage in *Arabidopsis thaliana* was made, and a direct comparison of genctic code preference among *Arabidopsis thaliana*, *Homo sapiens* and *Escherichia coli* was carried out. The results show that *Arabidopsis thaliana* like *Homo sapiens* its frequency of codon usage is obviously different from *Escherichia coli*, and there is a considerable difference between *Arabidopsis thaliana* and *Homo sapiens*. The data will give some suggestions to those researchers who want to introduce a animal gene into plant or a plant gene into bacteria.

Key words *Arabidopsis thaliana*, code preference, XML

* Corresponding author. Tel: 86-29-7092262, E-mail: sanhongfan@yahoo.com.cn

Received: August 12, 2002 Accepted: November 20, 2002