生态环境学报 2013, 22(5): 767-773
Ecology and Environmental Sciences
http://www.jeesci.com
E-mail: editor@jeesci.com

外来入侵物种的基因签名及其遗传多样性聚类分析

谈承杰，朱平*

江南大学理学院, 江苏 无锡 214122

摘要：密码子的使用频率分布能够反映一定的生物特性，因而可作为一种基因签名。本文使用 CGR 方法来研究外来入侵物种不同组织序列的基因签名及遗传多样性聚类分析，首先得出了刺花莲子草(*Alternanthera pungens*)，紫茎泽兰(*Ageratina adenophora*)，水葫芦(*Eichhornia crassipes*)，微甘菊(*Mikania micrantha*)，土荆芥(*Chenopodium ambrosioides*)，一枝黄花(*Solidago canadensis*)等 6 种外来入侵植物的 31 条序列核苷酸字串长 *k*=1 到 *k*=6 的情况，并选取 *k*=3，即基因序列的密码子，作为生物特性的一个重要表达。并且构造序列间的 CGR 欧式距离，进而对外来入侵植物序列遗传多样性进行了聚类分析。通过对所获得的 6 种外来入侵植物的 31 条序列的基因签名，得出如下结果：CGR 是一种简便且计算量小的方法，且基于 CGR 方法的基因签名，具有典型的生物特性；入侵植物的基因序列在密码子的使用上是非均衡的，且物种亲缘关系近的，则基因签名相似越高；而且基因签名也揭示出了密码子的第三位碱基偏好使用碱基 T 的现象，与一般物种密码子第三位碱基偏好 G/C 情况有强烈反差。此外，从获得的 6 个物种的 31 条序列聚类谱系图可以直观看出，入侵植物间存在着一定的亲缘关系，遗传多样性较丰富。由于我们所建立的基于 CGR 方法的基因签名，不仅能够反映植物特性和进化关系，而且能揭示序列中密码子和碱基的偏好使用情况，因而该方法有利于对外来入侵物种的遗传多样性分析、风险评估及预防控制等提供科学依据。

关键词：外来入侵物种；基因签名；CGR；遗传多样性；聚类分析

中图分类号：Q16; Q786; O29　　**文献标志码：**A　　**文章编号：**1674-5906（2013）05-0767-07

引用格式：谈承杰，朱平. 外来入侵物种的基因签名及其遗传多样性聚类分析[J]. 生态环境学报, 2013, 22(5): 767-773.

TAN Chengjie, ZHU Ping. Genomic signature and cluster analysis of genetic diversity of alien invasive species [J]. Ecology and Environmental Sciences, 2013, 22(5): 767-773.

随着人类活动全球化，自然生态环境遭到了日益的破坏，森林、湿地、草地、岛屿、城市居民区等生态系统几乎无一幸免，部分地区生物多样性急剧丧失，造成了人类重大经济损失，严重影响了人类的可持续发展，生态环境破坏已然成为全球性问题，而生物入侵伴随着人类活动的其他因素，加剧了生态环境问题而成为全球关注的一个焦点[1-2]。入侵物种几乎涵盖包括哺乳类，鱼类，鸟类，两栖类，爬行类，甲壳类，节肢类，种子类，藻类，菠类，真菌，病毒，细菌以及其他微生物在内的生物类[3]，其中以植物入侵最为严重。外来入侵物种是指本存在于本地自然生态系统之外的，而通过人为或非人为的方式被引入到该生态系统中，且具有了一定的繁殖能力，对当地生态环境、生物多样性造成破坏，或危害人类生产生活的物种。而我国是世界上遭受外来物种入侵最为严重的国家之一，外来入侵物种近500种，每年造成的经济损失达千亿元。

DNA 是生物遗传物质的最基本成分，由 4 种碱基 A、C、G、T 按一定的排列顺序组成，而基因是 DNA 序列中具有生物功能的片段。利用基因签名来进行基因研究越来越多[4-10]，即从物种 DNA 序列中提取出可反映物种的分类和进化信息的模板作为基因签名。由于密码子的频数分布就具有这样的生物特性，因而可作为一种基因签名来表达。

外来植物的入侵，不仅打破原有生态系统的平衡，而且降低生物多样性，因而是最为严重的生态多样性问题。例如紫茎泽兰，是一种世界公认的恶性有毒杂草，原产地本是中美洲，而迅速蔓延亚洲多国，并在我国泛滥，造成湿地系统，农田系统，森林系统等极大危害和重大经济损失。鉴于植物的入侵给我国造成的巨大经济损失，本文重点研究的是我国外来入侵物种中的重点植物，包括刺花莲子草，紫茎泽兰，凤眼莲(又名“水葫芦”)，微甘菊，土荆芥，一枝黄花等 6 种植物。基因序列的密码子使用频率可作为一种基因签名，CGR 模式对核苷酸三联体，即密码子的频数

基金项目：国家自然科学基金项目(11271163)；环保公益性行业科研专项(200909070)；中央高校基本科研业务费专项资金资助（JUSRP51317B）

作者简介：谈承杰(1988 年生)，男，硕士研究生，E-mail: zptl2002@yahoo.com.cn

*通讯作者：朱平(1962 年生)，女，教授，博士，硕士生导师。E-mail:zhuping@jiangnan.edu.cn

收稿日期：2013-01-27

评估是高效的[11],本文使用 CGR 方法来评估外来入侵物种不同组织序列的基因签名分析。CGR 是一个单位正方形，基因序列通过反复迭代而映射描绘出来。首先得到了核苷酸字串长 k=1 到 k=6 的情况，以 k=1 的情况进行了简单分析，并选取 k=3，即基因序列的密码子作为生物特性的一个重要反映。而遗传多样性是生物遗传信息的总和，构成了生物多样性的核心部分，是生态多样性和物种多样性的基础。本文构造序列间的 CGR 欧式距离，求得了序列间的欧式距离矩阵，对外来入侵物种序列进行了聚类分析，以探讨植物遗传多样性问题。基于 CGR 方法的基因签名，具有典型的生物特性。本文的研究反映了 CGR 是一种简便且计算量小的方法，并可对外来入侵物种进行有效的遗传多样性聚类分析。本文基于 CGR 方法的基因签名能够有效反映植物生物特性，对外来入侵植物的遗传多样性研究、风险评估及预防控制提供一定的理论支持。

1 材料与方法

1.1 基于 CGR 方法的基因签名

CGR(chaos game representation)方法是一种可将基因序列密码子迭代映射到坐标空间的方法。

CGR 图形构造：将 4 个碱基分别对应于单位正方形中各顶点：A(0,0)、T(1,0)、G(1,1)、C(0,1)。则对长度为 N 的基因序列,每一个碱基对应的坐标公式[12]为：

$$\mathrm{CGR}_i = \frac{g_i + \mathrm{CGR}_{i-1}}{2}, i = 1,2...,N \qquad (1)$$

其中，初始点为正方形的中心点，即 $\mathrm{CGR}_0 = (0.5,0.5)$，$g_i$ 表示序列第 i 个碱基对应的顶点坐标。

根据 CGR 图形构造原理，我们可以认为对于一个碱基，单位平面即被划分为 2×2 等格；对于碱基对，则第二位碱基在 4 个等格中继续按照最初的碱基排列顺序进行 2×2 等格的划分；对于三联体碱基的串，则第三位碱基在 16 个等格中进行同样的操作；对于长度为 n 的串，则单位平面被划分为 $2^n \times 2^n$ 等格。

本文研究的是基因序列三联体密码子(k=3)的使用分布情况，一个密码子由 3 个碱基构成，因而需要将单位区域划分为 $2^3 \times 2^3$=64 个等格，如下图 1 所示。

然后对序列密码子的所在区域着色，密码子出现的次数越多，则区域内颜色越深，这样得到的图形就可作为一种反映生物特性的基因签名。

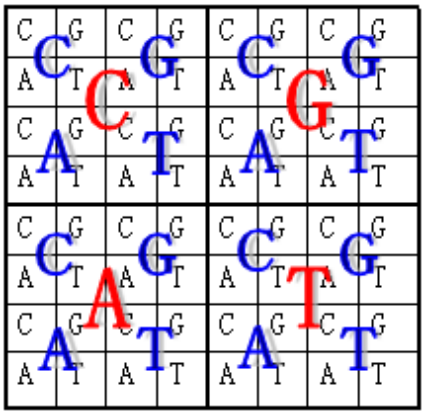

图 1 三联体密码子 CGR 图形

Fig.1 Figure of triplet codon based on CGR

1.2 材料来源

本文主要选取了 6 种入侵影响范围广和造成经济损失较严重的植物：刺花莲子草(*Alternanthera pungens*)，紫茎泽兰(*Ageratina adenophora*)，水葫芦(*Eichhornia crassipes*)，微甘菊(*Mikania micrantha*)，土荆芥(*Chenopodium ambrosioides*)，一枝黄花(*Solidago canadensis*)。由于目前基因测序的序列有限，本文从 Genbank/NCBI 中选取了 3 条刺花莲子草序列，序列号分别为 AY270054、AY514795、AY950665；4 条紫茎泽兰序列，序列号分别为 AY954290、AY954289、AY576867、AF374909；7 条水葫芦序列，序列号分别为 AF069215、AB040212、X84126、AY832131、ECU41599、ECU41574、ECR247090；2 条微甘菊序列，序列号分别为 AY270024、AF501627；3 条土荆芥序列，序列号分别为 DQ006049、DQ005963、DQ006134；3 个一枝黄花品种，3 条加拿大一枝黄花序列，序列号分别为 D49486、D49485、U97646；4 条 *Solidago fistulosa* 序列，序列号分别为 AF477731、AF477730、AF477667、AF477666；5 条 *Solidago sempervirens* 序列，序列号分别为 AF477732、AF477668、AF506905、AF506913、AF506911。

1.3 基于 CGR 的欧式距离

为深入分析序列间的相关性，本文采用欧式距离来度量物种序列间的 CGR (k=3) 签名。公式如下：

$$d(s_i, s_j) = \sqrt{\sum_{m=1}^{2^3}\sum_{n=1}^{2^3}(x_{m,n} - y_{m,n})^2} \qquad (2)$$

式(2)中，$d(s_i, s_j)$ 表示序列 s_i 和 s_j 的欧式距离，$x_{m,n}, y_{m,n}$ 表示两序列 CGR 的密码子频数。

2 结果与分析

2.1 碱基含量分析

根据 CGR 方法的基因签名，本文求得序列核苷酸字串长 k=1 到 k=6 的基因签名。从 k=1 的情况，即序列的碱基含量情况，我们知道刺花莲子草，紫茎泽兰，水葫芦，微甘菊和土荆芥序列的碱基 A+T 的含量几乎都超过了 50%，除了水葫芦的第 1,4 条序列和土荆芥第 2 条序列的碱基 G+C 含量超过了 50%，则知不同种类植物序列的碱基使用是非均衡的；在一枝黄花品种中，加拿大一枝黄花和 Sempervirens 一枝黄花序列的碱基 A+T 的含量较碱基 G+C 含量要高，而 Fistulosa 一枝黄花序列的碱基 A+T 含量与碱基 G+C 含量相当，可见同类植物序列中也存在着碱基非均衡使用情况。

2.2 CGR 方法的基因签名图

根据 GENSTYLE 网站的作图功能，以核苷酸字串长 k=3 时作为基因序列特征的一个重要反映，列出 6 种植物的 31 条基因序列的三联体密码子使用情况二维图，即基因签名。

在图 2 中，刺花莲子草的第 2 条序列基因签名与其他 2 条有很大的不同，可能是物种在组织进化过程中导致的变异。但其和图 3 紫茎泽兰的基因签名却有着局部的相似性，如刺花莲子草的第 1,3 条序列基因签名和紫茎泽兰的第 1，3，4 条序列的基因签名。图 4 水葫芦第 2 条序列的基因签名与图 5 中薇甘菊的第 1 条序列的基因签名也较相似，说明两物种可能存在一定的进化关系。水葫芦的第 5 条序列的基因签名与图 6 中土荆芥第 3 条序列的基因签名较为相似,说明水葫芦与土荆芥可能存在着一定的进化关系。

在图 7、图 8 和图 9 一枝黄花品种的基因签名中，加拿大一枝黄花第 3 条序列与 Fistulosa 一枝黄花第 4、5 条序列及 Sempervirens 一枝黄花第 2 条序列间的基因签名非常相似，Fistulosa 一枝黄花第 1、2 条序列和 Sempervirens 一枝黄花第 1 条序列也非常相似，其他序列也存在着局部的相似性。这说明，同类物种下的基因签名是相似的。

在这些基因签名中，或多或少的都能找到它们局部相似性的地方。此外，以上基因签名中都出现了颜色深黑的块，即代表该密码子出现的次数很多，且也都出现了许多颜色较浅的块，说明物种基因在密码子的使用上存在着普遍的非均衡现象。特别地，通过观察本文这些着色较深的密码子，发现密码子的第三位碱基较偏向使用碱基 T，而使用碱基 C 的最少，与以往对基因密码子偏好性研究[13-16]的一般结论不同。

通过对以上序列的分析，得出 6 个物种的基因签名都是具有一定的生物特性的，若物种亲缘关系近，则签名相似，且发现不同物种之间的基因签名也存在着局部的相似性，可能和物种的遗传进化有关。6 个外来入侵物种的基因签名还反映了序列密码子的非均衡使用现象，导致进化的可能与人类的活动有关。

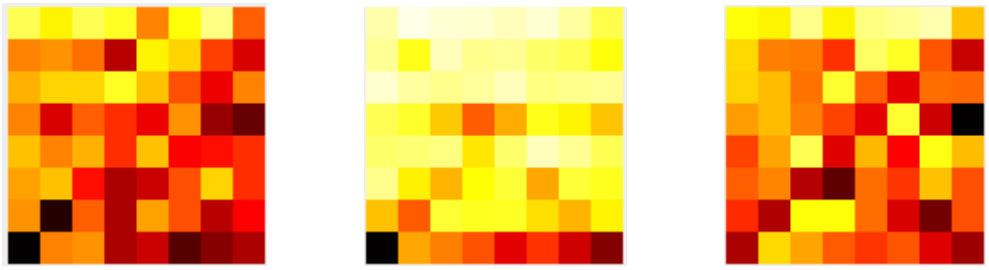

图 2 3 条刺花莲子草序列的基因签名

Fig.2 Signatures on 3 sequences of *Alternanthera pungens*

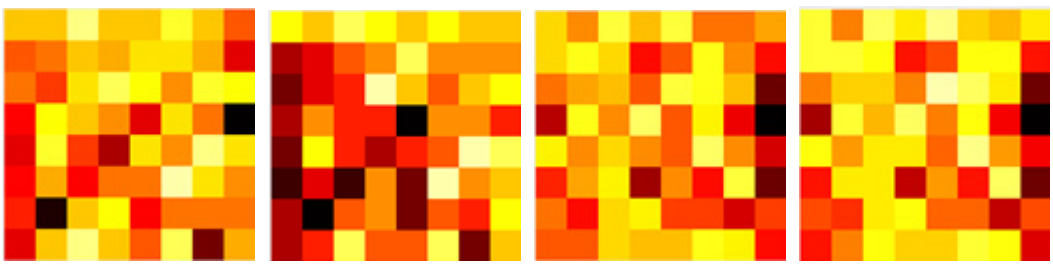

图 3 4 条紫茎泽兰序列的基因签名

Fig.3 Signatures on 4 sequences of *Ageratina adenophora*

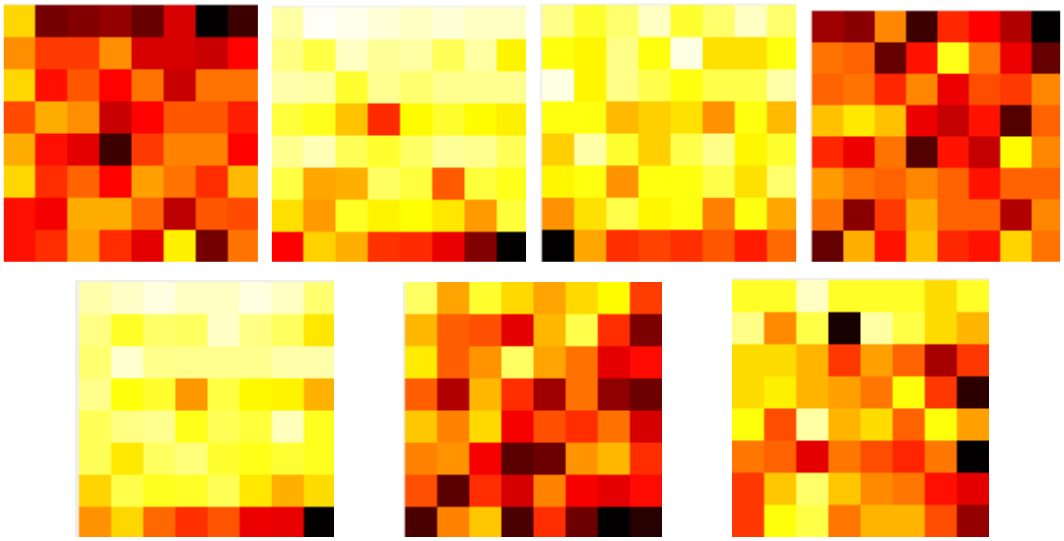

图 4 7 条水葫芦序列的基因签名

Fig.4 Signatures on 7 sequences of *Eichhornia crassipes*

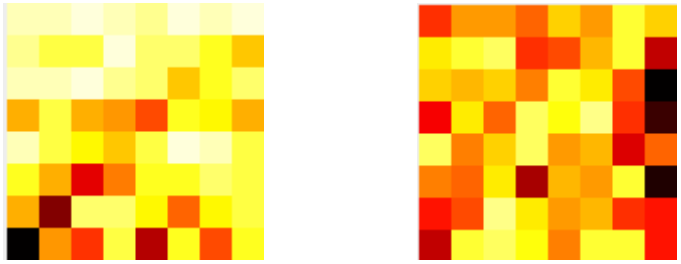

图 5 2 条微甘菊序列的基因签名

Fig.5 Signatures on 2 sequences of *Mikania micrantha*

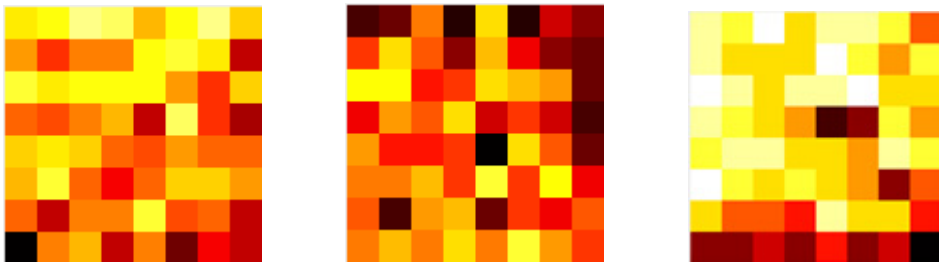

图 6 3 条土荆芥序列的基因签名

Fig.6 Signatures on 3 sequences of *Chenopodium ambrosioides*

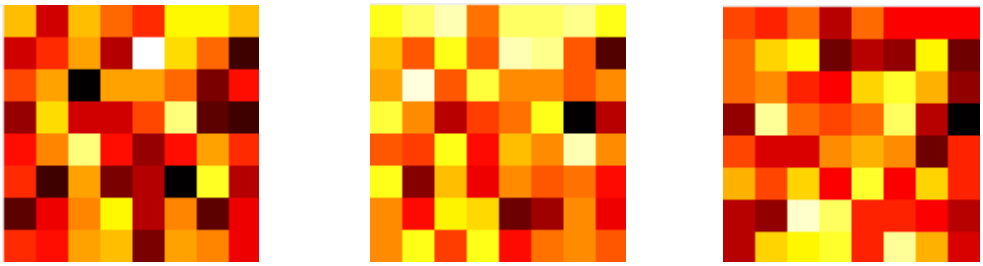

图 7 3 条加拿大一枝黄花序列的基因签名

Fig.7 Signatures on 3 sequences of *Solidago canadensis*

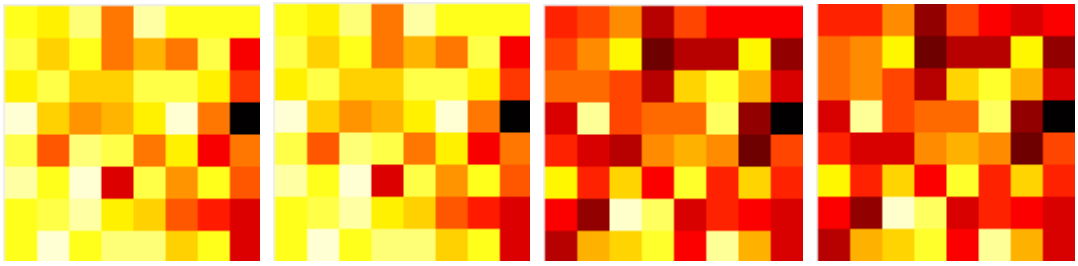

图 8　4 条 Fistulosa 一枝黄花序列的基因签名

Fig.8　Signatures on 4 sequences of *Solidago fistulosa*

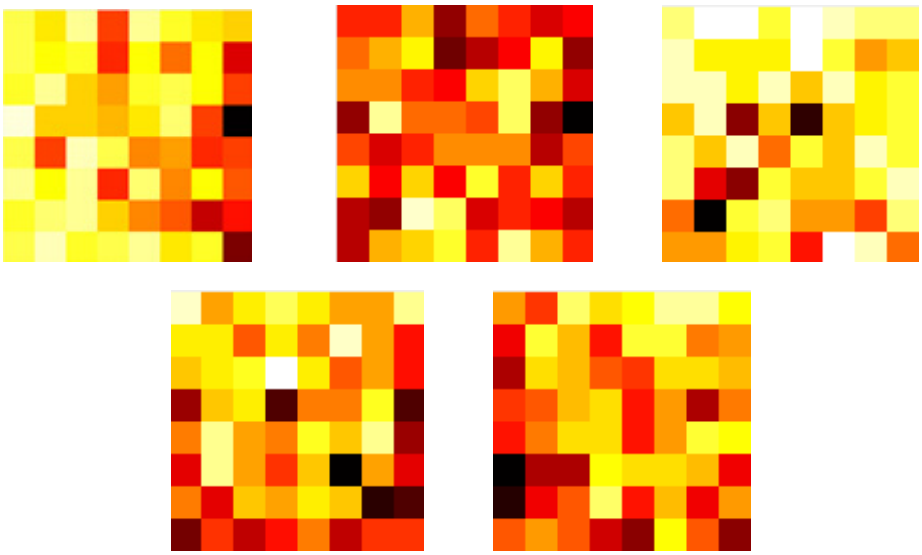

图 9　5 条 Sempervirens 一枝黄花序列的基因签名

Fig.9　Signatures on 5 sequences of *Solidago sempervirens*

2.3　遗传多样性聚类分析

根据公式(2)，我们利用 Microsoft Visual C++6.0 计算出 31 条序列 CGR 间的欧式距离，得到了 CGR 间的 31×31 欧式距离矩阵，总体计算时间复杂度为 $O(10^3)$，计算量较小。并利用 SPSS 18.0 软件 Hierarchical Cluster 功能 Ward 方法来构建序列间的聚类谱系图，如下图 10 所示。

由上图，我们可以清晰地观察到刺花莲子草序列与水葫芦的两条序列(编号9,12)相关度高；紫茎泽兰两条序列(编号6,7)与薇甘菊序列、土荆芥和一枝黄花序列相关度也较高；薇甘菊序列，土荆芥序列和一枝黄花序列也存在着较高的关联度；水葫芦的一条序列(编号10)与土荆芥的一条序列(编号17)相关度很高；而同属一枝黄花的加拿大一枝黄花序列与其他两种植物序列相关度不是很高。总体来看，刺花莲子草与水葫芦亲缘关系近，而其他4种植物亲缘关系最为接近。说明不同植物的组织基因序列间存在着一定的相似性，即亲缘关系近，因而它们的组织存在一定的相似功能。同类植物的不同组织基因序列也存在着一定的不相似性，即亲缘关系较远。以上植物呈现的遗传

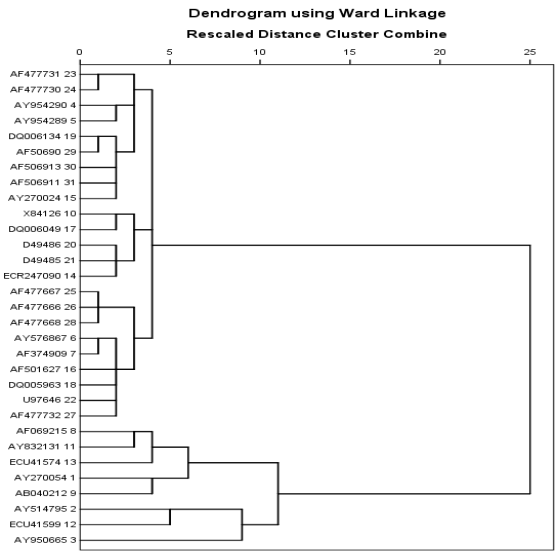

图 10　植物序列聚类谱系图

Fig.10　Dendrogram of clustering for 31 sequences

多样性较丰富，表明物种对我国环境适应性变的越强，该研究势必对外来入侵物种的遗传多样性、起源和进化、分类，风险评估及预防控制提供一定的理论支持。

3 结论

由于人类活动的不约束性，导致了生物入侵的泛滥，并造成了全球生态和经济的巨大损失。人们已经认识到生物入侵过程实际上是一个受现代人类活动影响的物种快速进化的过程[17-19]，因而要想防治生物入侵，必须规范人类活动。本文从我国外来入侵物种出发，选取了入侵危害最为严重的6种植物，使用基于CGR方法的基因签名，得出了31条序列核苷酸字串长k=1到k=6的情况。因基因序列核苷酸密码子(k=3)的CGR具有典型的生物特性,便将密码子使用频率分布作为本文基因序列的基因签名。通过它们的碱基含量及基因签名的对比分析，得出基因序列在碱基的使用上是非均衡的，物种亲缘关系近，则基因签名也相似。且不同的植物的基因签名也表现出局部的相似性，说明两者之间具有一定的进化关系。特别的，基因签名也揭示出了密码子的第三位碱基偏好使用碱基T的现象，与一般物种密码子第三位碱基偏好G/C情况有强烈反差。我们构造了序列间的CGR欧式距离矩阵，对外来入侵物种序列进行了有效的遗传生物多样性聚类分析。由聚类谱系图可以直观看出，入侵植物间存在着一定的亲缘关系，且遗传多样性较丰富。由于我们所建立的基于CGR方法的基因签名能够反映植物特性和进化关系，因而该方法有利于对外来入侵物种的遗传多样性分析、风险评估及预防控制等提供科学依据。

对于我国的生物入侵现状，研究者也提出了许多的建议[20-22]。我们认为最重要的是加强对外来物种的生物特性、遗传进化等的研究，建立生物入侵数学模型、外来物种的数据库和风险评估机制，并与国内及国外研究人员展开交流合作，实现入侵生物资源共享，为我国实施生态保护提供科学依据。

参考文献:

[1] DUKES J A, MOONEY H A. Does global change increase the success of biological invaders[J].Trends in Ecology and Evolution, 1999, 14: 135-139.

[2] VITOUSEK P M, DANOTONIO C M, LOOPE L L, et al. Introduced species: a significant component of human-caused global change[J]. New Zealand Journal of Ecology, 1997, 21:1-16.

[3] XU H G,OIANG S,HAN Z, et al. The distribution and introduction pathway of alien invasive species in China[J].Biodiversity Science, 2004, 12(6):626-638.

[4] BERNARD F, GENSTYL E. exploration and anaiysis of DNA seguences with genomic signature[J]. Nucleic Acids Research, 2005, 33: 512-515.

[5] SAMUEL K, CHRIS B. Dinucleotide relative abundance Extremes: a genomic signature[J]. TIG July, 1995, 11(7): 283-290.

[6] 管维红, 朱平, 张立亭, 等. 关于入侵物种三联密码子的基因签名[J]. 电子测量技术, 2007,30(4):40-44.

[7] CHRISTINE D. Detection and characterization of horizontal transfers in prokaryotes using genomic signature[J]. Nucleic Acids Research, 2005, 33(1):1-12.

[8] PANDIT A, SINHA S. Using genomic signatures for HIV-I subtyping[J]. BMC Bioinformatics, 2010, 11(Suppl 1): S26.

[9] Bohlin J. Genomic signatures in microbes-properties and Applications[J]. The Scientist World Journal, 2011, 11: 715-725.

[10] RYAN K VAN LAAR. Genomic signatures for predicting survival and adjust chemotherapy benefit in patients with non-small-cell lung cancer[J]. BMC Medical Genomics, 2012, 5:30.

[11] GOLDMAN N. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences[J]. Nucleic Acids Research, 1993, 21:2487-2491.

[12] JEFF REY H. Chaos game representation of gene structure[J]. Nucleic Acids Research, 1990, 18:2163.

[13] 孔娟娟, 朱平. 人类p53肿瘤蛋白的偏好性分析及其应用[J]. 计算机应用研究, 2011, 28(8):2987-2990.

[14] 谈承杰, 朱平. 抑癌基因 p53 密码子偏好性分析及其突变致癌预测[J]. 计算机与应用化学, 2012, 29(11):1299-1303.

[15] 吴宪明, 吴松锋, 任达明, 等. 密码子偏性的分析方法及相关研究进展[J]. 遗传, 2007,29(4):420-426.

[16] 石秀凡, 黄京飞, 柳树群, 等. 人类基因同义密码子偏好的特征以及与基因GC含量的关系[J]. 生物化学与生物物理进展, 2002,29(3):411-414.

[17] 唐旭清, 马保, 金英花, 等. 入侵物种沿青藏铁路扩散的仿真研究, 系统仿真学报, 2012, 24(12): 2556-2661.

[18] Hanfling B, Kollmann J. An evolutionary perspective of biological invasions[J]. Trends in Ecology and Evolution, 2002,17:545-546.

[19] Lee CE.. Evolutionary genetics of invasive species[J].Trends in Ecology and Evolution, 2002,17, 386-391.

[20] 杨秀娟, 张树苗. 生物入侵对生物多样性的影响[J]. 林业调查规划, 2005,30(1):36-38.

[21] 王丰年. 外来物种入侵的历史,影响及对策研究[J]. 自然辩证法研究, 2005, 21(1):77-81.

[22] 丁晖, 徐海根, 强盛, 等. 中国生物入侵的现状与趋势[J]. 生态与农村环境学报, 2011, 27(3):35-41.

Genomic signature and cluster analysis of genetic diversity of alien invasive species

TAN Chengjie, ZHU Ping

School of Science, Jiangnan University, Wuxi, Jiangsu 214122, China

Abstract: Alien invasive species had been a global eco-environmental problem, resulting in huge economic loss. The distribution of codon usage could reflect some biological characteristics, therefore it could be a genomic signature. The paper proposed a genomic signature-based approach, using CGR with word length of k=3, which could be applied to analyse the sequences of alien invasive species and cluster analysis of genetic diversity. Firstly, we gained these cases of nucleotide word lengths (k=1 to 6) on 31 sequences of 6 alien invasive plants, which were alternanthera pungens, ageratina adenophora, eichhornia crassipes, mikania micrantha, chenopodium ambrosioides and solidago, respectively, and chosen word length of k=3, also was the codons, which be as the important biological expression. And then we constructed the Euclidean distance of CGR, and finally gained the dendrograms of alien invasive species. And these results were shown by analyzing these genomic signatures of 31 sequences of 6 alien invasive, respectively. CGR was a handy method with small calculation. And the genomic signature based on CGR could reflect some typical biological properties. The usage of codon in the gene sequence of alien invasive plant was unbalanced. The more closed genetic relationship of alien invasive species were, the more similar their genomic signatures were. Among these genomic signatures, a especial phenomena was revealed, and that was the third base of codons of gene sequence preferred to using nucleotide T, which was very different to these common species preferring to using nucleotide G or C. What's more, from the dendrograms of clustering of 31 sequences of 6 alien invasive plants, we could see clearly alien invasive plants had some relationships, and genetic diversity was abundant. By using this constructed method of genomic signature based on CGR, not only efficiently reflected these biological characteristics and evolutionary relationship of alien invasive plants, but also revealed the usage condition of codon and base in the codon of alien invasive plants. Therefore, the structured method which was benefit to offering scientific basis for analysis of genetic diversity, risk assessment, prevention and control of alien invasive plants.

Key words: alien invasive species; genomic signature; CGR; genetic diversity; cluster analysis