EasyCodeML 在适应性进化分析中的应用

Application of EasyCodeML on Adaptive Evolution Analysis

许瑜婷, 高芳銮*

植物保护学院, 福建农林大学, 福州 350002;

*通讯作者邮箱: raindy@fafu.edu.cn

引用格式: 许瑜婷, 高芳銮. (2021). EasyCodeML 在适应性进化分析中的应用. *Bio-101* e1010651. Doi: 10.21769/BioProtoc.1010651.

How to cite: Xu, Y. T. and Gao, F. L. (2021). Application of EasyCodeML on Adaptive Evolution Analysis. *Bio-101* e1010651. Doi: 10.21769/BioProtoc.1010651. (in Chinese)

摘要: 自然选择是生物进化的动力, 在分子水平上检测正选择作用, 进而识别重要的氨基酸位点, 这对于揭示生物适应性进化机制具有重要的意义。在适应性进化分析中, PAML 软件包中的 CodeML 子程序是应用最为广泛的一款工具, 但由于其晦涩难懂的参数配置及命令行式的操作让不少新用户望而却步。本文应用新开发的 CodeML 可视化分析工具—EasyCodeML 对示例数据进行选择压力分析, 并给出 CodeML 分析基本流程, 包括格式转换、参数配置、结果解读和绘图等步骤, 可为不同物种开展相关分析提供参考。

关键词: 适应性进化, EasyCodeML, 自然选择, 正选择位点, 似然率检验

研究背景

在 DNA 水平上检测正选择, 并找出对应的氨基酸位点是适应性进化分析中不可或缺的一项重要内容。在基于密码子 (Codon-based) 进化模型的生物软件和资源中, PAML 包中的 CodeML 程序目前使用最为广泛 (Yang, 2007)。根据序列中不同位点或系统发育树中不同分支上 ω 值 (非同义替代与同义替换的比值, 即 dN/dS 或 Ka/Ks) 不同, CodeML 中的密码子模型可以分为枝模型 (Branch model)、位点模型 (Site model)、枝位点模型 (Branch-site model) 和进化枝模型 (Clade model) 四大类 (Yang and Nielsen, 1998 and 2002; Forsberg and Christiansen, 2003; Bielawski and Yang, 2004;

Zhang *et al.*, 2005)。

位点模型 (Site model) 常用于检测核苷酸序列中位点的正选择 ($\omega > 1$) 作用, 该模型假设数据集中不同密码子的选择压不同, 而支系 (Lineage) 间的选择压力相同。位点模型中共有 8 个不同假设的模型 (表 1), 其中 M0 和 M3、M1a 和 M2a、M7 和 M8 以及 M8a 和 M8 为四对巢式模型 (Nested model), M0、M1a、M7 和 M8a 为零假设模型 (null model), M3、M2a 和 M8 为备选假设模型 (alternative model)。除 M0 vs. M3 (不是用于检测正选择) 外, 如果备选模型显著优于零假设模型 (Likelihood ratio test, $p < 0.05$), 表明存在正选择作用 (Yang *et al.*, 2000)。检测正选择作用的巢式模型中, M7 vs. M8 是检测正选择最为常用的模型对, 但由于 M8a vs. M8 得到的假阳性结果更低, 现在也渐渐被使用。当数据中检测到正选择作用后, 可以应用贝叶斯经验贝叶斯法 (Bayes empirical Bayes, BEB) 进一步鉴定可能受正选择作用的氨基酸位点 (Yang *et al.*, 2005)。

表 1. 位点模型 (Site model) 中不同模型的背景假设

模型	背景假设
M0 (单一率)	假设序列中所有位点具有相同的 ω 值
M1a (近中性)	假设位点中仅有保守位点 ($0 < \omega < 1$) 和中性位点 ($\omega = 1$) 而没有正选择位点 ($\omega > 1$) 存在, 这两类位点的比率分别为 p_0 和 p_1 , 其对应的 ω 值分别为 ω_0 、 ω_1 ;
M2a (正选择)	在 M1 基础上增加了第三类 ω 值, 即假设位点中除了保守位点和中性位点外, 还存在正选择位点 ($\omega > 1$), 这三类位点的比率分别为 p_0 、 p_1 和 p_2 , 其对应的 ω 值分别为 ω_0 、 ω_1 和 ω_2
M3 (离散)	该模型假设所有的位点 ω 值呈简单的离散分布;
M7 (beta)	该模型假设所有位点的 ω 属于矩阵 (0, 1) 并呈 beta 分布
M8 (beta & ω)	该模型在 M7 基础上增加另一类 ω 值 ($\omega > 1$)
M8a (beta & $\omega = 1$)	该模型与 M8 模型类似, 但将 ω 值固定为 1 ($\omega = 1$)

软件运行环境及信息

1. EasyCodeML 程序为 Java 编写，支持跨平台使用。运行前需确保操作系统中已安装了 JAVA 运行环境 (JRE 1.6 或更高版本)。JRE 最新版本可从 JAVA 官网下载，下载链接为: <https://www.java.com/en/download/manual.jsp>
2. EasyCodeML v1.41 <https://github.com/BioEasy/EasyCodeML>
3. 本文示例数据可以访问作者永硕网盘获取 (<http://raindy.ys168.com/>); EasyCodeML 软件子目录 Example 中的 Example2.pml 和 Example2.tre 文件与示例数据相同。

实验步骤

1. 分析前数据处理

1.1 比对序列文件

用于选择压力的序列文件必须是比对后的 PAML 格式，扩展名为 PML。非 PAML 格式的序列文件，可以通过 EasyCodeML 工具栏下的 "Sequence Format Converter" 进行转换，通过拖曳操作将不同的序列格式转化为 PAML 格式 (图 1)。格式转换前，需要检测编码蛋白的核苷酸序列长度是否为 3 的倍数。除外，序列中的终止密码子也必须删除，否则容易产生假阳性结果。

图 1. 比对序列格式转换

1.2 系统发育树文件

用于适应性进化分析的树文件必须是不带任何数值的 Newick 格式。如果树文件中的数值未被清除，可以使用工具栏下的 "Tree Cleaner" 工具进行过滤 (图 2)。

图 2. 树文件过滤器

2. 数据加载及分析参数配置

EasyCodeML 分析所需要的数据文件准备就绪后，可在 **Preset Mode** (预置模式) 中"Setup"标签下加载对应的数据文件，并配置主要参数 (图 3)。

图 3. EasyCodeML 参数配置界面

2.1 选择一个本地文件夹作为工作目录，可以直接通过拖曳进行操作。

2.2 选择 **Site Model** 作为分析模型 (默认)。

- 2.3 载入已转为 PAML 格式的比对序列文件。
- 2.4 载入已过滤掉数值的 Newick 格式树文件。
- 2.5 "nt"为线程数，如果计算机资源支持多线程，可以将"nt"选项后的数值调至系统支持的最大值，但一般不超过 7 (Site Model 中总共有 7 个模型)。
- 2.6 物种对应的密码子可以通过"icode"选项下拉菜单进行选择，示例数据为标准密码子，故使用默认密码子 (0: standard)。
- 2.7 通过"Check"检查树文件中类别名称与比对序列的名称是否一致，当弹出 "SeqFile And TreeFile is Ready"时，说明两者名称一致，可以进行下一步的 CodeML 分析。

3. CodeML 分析

启动 CodeML 分析前，需要点击"Save Current Profile"以使前面配置的参数信息生效。点击"Run CodeML"即可开始 CodeML 分析，其运行时间主要取决于数据中序列长度和数量。当 CodeML 运行完毕，程序会自动弹出"Running Finished"提示，在 LRT 标签下将对数似然法 (log-likelihood value, lnL) 和参数个数 (the number of parameter, np) 等关键参数显示在界面中(图 4)。

Likelihood Ratio Test (LRT)				
Current Model: Site model (M0 vs. M3, M1a vs. M2a, M7 vs. M8, M8a vs. M8)				
Test Statistic:	M0 vs. M3: 2 Δ lnL=	77.164283999999	d.f. = 4	P-value = 0.000000000
	M1a vs. M2a: 2 Δ lnL=	29.148333999999	d.f. = 2	P-value = 0.000000468
	M7 vs. M8: 2 Δ lnL=	37.031211999999	d.f. = 2	P-value = 0.000000009
	M8a vs. M8: 2 Δ lnL=	27.561433999999	d.f. = 1	P-value = 0.000000152
Summary of Results: /Users/raindy/Desktop/Demo/SM/SM_res.xls				
Export View				

图 4. 似然率检验结果

4. 结果分析及绘图

LRT 结果显示，M1a vs. M2a、M7 vs. M8 和 M8a vs. M8 三对巢式模型的 p 值均小 0.05，说明备选模型显著优于零假设模型，表明存在正选择作用。通过"Export"模块可将主要结果及参数导出，并自动生成为准发表级的 Excel 表格(图 5)。

14 L 0.603,133 G 0.961*,228 S 0.972*,230 F 0.996**,231 V 0.993**,232 S 0.567,233 K

Site model (SM)	Model	np	Ln L	Estimates of parameters	Model compared	LRT P-value	Positive sites
M3	15	-1876.512	p: 0.94718 ω: 0.04809	0.00473			
M0	11	-1915.094	ω: 0.07365 ω0: 0.30409	0.33208 0.02642	M0 vs. M3	0.000000000	Not Allowed
M2a	14	-1877.941	p: 0.81874 ω: 0.00000	0.15483 1.00000			
M1a	12	-1892.515	p: 0.82898 ω: 0.00000	0.17102 1.00000	M1a vs. M2a	0.000000468	Not Allowed
M8	14	-1878.735	p0=0.9668 p1=0.033 ω=19.65767	q=0.16652			14 L 0.603,133 G 0.961*
M7	12	-1897.250	p=0.00500 p0=0.8288 p1=0.171 ω=1.00000	q=0.00637	M7 vs. M8	0.000000009	Not Allowed
M8a	13	-1892.515	p1=0.171 ω=1.00000		M8a vs. M8	0.000000152	Not Allowed

图 5. EasyCodeML 自动生成的准发表级数据表

如图 5 所示，表中最后一列"Positive Sites"是由 BEB 法鉴定可能受正选择作用的氨基酸位点。虽然 M2a 中也检测了正选择作用，但其灵敏度未能鉴定受正选择压力的氨基酸点。相比之下，M8 模型更为灵敏，示例数据中共找到了 9 个显著的正选择位点(后验概率 ≥ 0.95)。

在位点模型中，模型对 M0 vs. M3 虽然不是用于检测正选择，但 M3 模型中生成的 rst 文件，可用于 dN/dS 的滑窗分析(Sliding windows analysis) 以直观显示序列中不同位点的选择压力变化。如图 6 所示，可以提取倒数第二列数据进行绘图，横坐标为比对序列 Codon 位置，纵坐标为 dN/dS 比值，预览图如 Gao *et al.* (2017) 一文的 Figure 3 所示，本文不再赘述。

rst

Naive Empirical Bayes (NEB) probabilities for 3 classes & postmean_w & P(w>1)
(amino acids refer to 1st sequence: AF335467)

1	M	0.99329	0.00671	0.00000	(1)	0.129	0.007
2	K	0.99116	0.00884	0.00000	(1)	0.147	0.009
3	V	0.99020	0.00980	0.00000	(1)	0.155	0.010
4	K	0.99116	0.00884	0.00000	(1)	0.147	0.009
5	V	0.97775	0.02224	0.00001	(1)	0.258	0.022
6	L	0.97850	0.02149	0.00000	(1)	0.252	0.021
7	S	0.99513	0.00487	0.00000	(1)	0.114	0.005
8	L	0.99524	0.00476	0.00000	(1)	0.113	0.005
9	L	0.97850	0.02149	0.00000	(1)	0.252	0.021
10	V	0.97775	0.02224	0.00001	(1)	0.258	0.022
11	P	0.97163	0.02834	0.00003	(1)	0.310	0.028
12	A	0.97807	0.02192	0.00001	(1)	0.256	0.022
13	L	0.97850	0.02149	0.00000	(1)	0.252	0.021
14	L	0.30832	0.69057	0.00111	(2)	5.874	0.692
15	V	0.97775	0.02224	0.00001	(1)	0.258	0.022
16	A	0.98809	0.01191	0.00000	(1)	0.172	0.012
17	G	0.99810	0.00190	0.00000	(1)	0.089	0.002
18	A	0.97766	0.02233	0.00001	(1)	0.259	0.022
19	A	0.97766	0.02233	0.00001	(1)	0.259	0.022

图 6. 不同位点的 dN/dS 值

常见问题及解决方法

在分析过程中，EasyCodeML 用户可以会遇到一些如下问题：

1. 为什么 EasyCodeML 无法运行，解压缩包死循环？

EasyCodeML.jar，是主程序文件，正常情况下，直接双击即可运行。但如果*.jar 文件被识别为压缩包，则程序会被系统关联的解压缩软件进行解压缩处理。

解决方法：是去除*.jar 与解压缩软件的关联，比如 WinRAR 之类的，在参数设置中去除 jar 的关联，重新关联为 JAVA 程序即可。

2. 为什么找到的正选择位点在原始比对序列中找不到？

在 EasyCodeML 中，默认启用了 Clean data 功能(Clean data = 1)，即：比对序列中如带有 gap，则 gap 所对应的一列序列被忽略不进行分析，从而导致位置发生偏移。

解决方法：(1) 需要重新运行 CodeML，分析前将 Clean data 选项前的勾选取消，即 Clean data = 0，此时数据含有 gap 运行 CodeML 分析，其运算量也相应增大；(2) 分析前将带 gap 的同一列碱基全部手工删除，找到正选择位点的氨基酸位置后，再还原对应到原始比对序列上。

小结与建议

EasyCodeML 以 CodeML 为内核，可以实现 CodeML 分析界面化、自动化操作，生成准发表级的数据表格。应用 EasyCodeML 进行适应性进化分析，简单易用。但受内核所限，该软件仅适用于中、小型数据集。对于多基因的 CodeML 分析，推荐使用 LMAP 软件 (Maldonado *et al.*, 2016)；对于基因组级别的数据，则推荐使用软件 PosiGene (Sahm *et al.*, 2017) 和 GWideCodeML (Macías *et al.*, 2020)。

致谢

EasyCodeML 自 2015 年首次推出测试版，时至今日最新的 1.4 版，期间历经无数次的反复调试。在此，特别感谢华南农业大学陈程杰博士的辛勤付出和广大 EasyCodeML 用户的反馈意见使得程序日臻完善。

竞争性利益声明

作者声明没有利益冲突

参考文献

1. Bielawski, J. P. and Yang, Z. (2004). [A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution](#). *J Mol Evol* 59(1): 121-132.
2. Forsberg, R. and Christiansen, F. B. (2003). [A codon-based model of host-specific selection in parasites, with an application to the influenza A virus](#). *Mol Biol Evol* 20(8): 1252-1259.
3. Gao, F., Zou, W., Xie, L. and Zhan, J. (2017). [Adaptive evolution and demographic history contribute to the divergent population genetic structure of potato virus Y between China and Japan](#). *Evol Appl* 10(4): 379-390.
4. Macías, L. G., Barrio, E. and Toft, C. (2020). [GWideCodeML: a python package for testing evolutionary hypotheses at the genome-wide level](#). *G3: Genes Genom Genet* 10(12): 4369-4372.
5. Maldonado, E., Almeida, D., Escalona, T., Khan, I., Vasconcelos, V. and Antunes, A. (2016). [LMAP: lightweight multigene analyses in PAML](#). *BMC bioinformatics* 17: 354.
6. Sahm, A., Bens, M., Platzner, M. and Szafranski, K. (2017). [PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes](#). *Nucleic Acids Res* 45(11): e100.
7. Yang, Z. (2007). [PAML 4: phylogenetic analysis by maximum likelihood](#). *Mol Biol Evol* 24(8): 1586-1591.
8. Yang, Z. and Nielsen, R. (1998). [Synonymous and nonsynonymous rate variation in nuclear genes of mammals](#). *J Mol Evol* 46(4): 409-418.
9. Yang, Z. and Nielsen, R. (2002). [Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages](#). *Mol Biol Evol* 19(6): 908-917.
10. Yang, Z., Nielsen, R., Goldman, N. and Pedersen, A. M. (2000). [Codon-substitution models for heterogeneous selection pressure at amino acid sites](#). *Genetics* 155(1): 431-449.
11. Yang, Z., Wong, W. S. and Nielsen, R. (2005). [Bayes empirical Bayes inference of amino acid sites under positive selection](#). *Mol Biol Evol* 22(4): 1107-1118.
12. Zhang, J., Nielsen, R. and Yang, Z. (2005). [Evaluation of an improved branch-site](#)

[likelihood method for detecting positive selection at the molecular level](#). *Mol Biol Evol* 22(12): 2472-2479.