·综 述·

直系同源基因的识别方法与数据库

杨 婧, 黄 原*, 汪晓阳

(陕西师范大学 生命科学学院, 中国陕西 西安 710062)

摘 要: 直系同源(orthology)是指由于物种形成事件而享有共同祖先的基因之间的关系, 直系同源基因之间通常具有相似的结构和生物学功能. 由于基因组和转录组序列的快速积累, 精确的识别直系同源基因有助于功能基因的注释, 比较和进化基因组学研究. 综述了现有的识别直系同源基因的主要方法, 并列举了由此构建的数据库. 这些方法可以归纳为三大类, 第一类是基于序列相似性的方法, 具有识别速度快以及灵敏度高等优点; 第二类是基于构建系统发育树的方法, 具有准确性高和信息量大等优点; 第三类是将上述两种方法结合起来的混合方法, 更好地平衡了灵敏性和准确性. 最后总结了识别过程所面临的问题.

关键词: 直系同源; 直系同源识别; 数据库

中图分类号: Q953 **文献标识码**: A **文章编号**: 1007-7847(2013)03-0274-04

Methods and Databases for Identification of Orthologs

YANG Jing, HUANG Yuan*, WANG Xiao-yang

(*College of Life Sciences, Shaanxi Normal University, Xi'an* 710062, *Shaanxi, China*)

Abstract: Orthologous genes are those derived from a common ancestor through speciation, and typically retain similar architecture and biological function. Because of rapid accumulation of genomic and transcriptomic sequence, automated identification of orthology can facilitate functional annotation, and studies on comparative and evolutionary genomics. The main methods of orthologs prediction and corresponding databases constructed with these methods were briefly reviewed here. These methods can be grouped into three kinds, the first is similarity-based method, it has high sensibility and fast speed; the second is tree-based method, it is precise and informative; the third is hybrid method, it is the optimal trade-off between precision and sensibility. Finally the problems faced by the recognition process were summarized.

Key words: orthology; ortholog identification; database

(*Life Science Research*, 2013, 17(3): 274~277)

近年来, 随着第二代测序技术的快速发展, 测序数据急剧增加. 因此, 需要更多的生物信息学方法对其进行分析研究. 由于直系同源基因之间具有相似的生物学功能, 它的识别对于分析生物的系统进化以及预测新基因与蛋白质的功能至关重要.

基因之间的同源关系主要包括两种类型: 直系同源(orthology)是指由于物种形成事件, 从共同祖先进化而来的基因, 通常具有相同或者相似的基因功能, 这个术语最早是由 Walter Fitch 在 1970 年首先提出的; 旁系同源(paralogy)是指由基因复制而分离的同源基因, 同时基因复制通常伴随着基因功能的分化, 因此旁系同源基因会进化出不同的功能[1, 2].

1 直系同源基因的基本特性和应用

直系同源基因是分布于两种或两种以上物种基因组中, 由于物种形成事件而享有共同祖先的

收稿日期: 2012-12-10; **修回日期**: 2013-01-10

基金项目: 国家自然科学基金资助项目(31172076; 30970346)

作者简介: 杨婧(1988-), 女, 陕西宝鸡人, 硕士研究生, 主要从事分子系统学研究; * 通讯作者: 黄原(1962-), 男, 甘肃天水人, 陕西师范大学生命科学学院教授, 博士, 主要从事昆虫分子进化和分子系统学研究, Tel: 029-85310271, E-mail: yuanh@snnu.edu.cn.

同源序列, 通常认为直系同源的序列具有相似的结构和生物学功能[3], 功能高度保守乃至于近乎相同, 甚至其在近缘物种可以相互替换, 通常是编码生命活动必需的关键性调控蛋白、酶或辅酶的基因[4]. 此外, 许多直系同源基因还具有序列变化速度与进化距离相当, 调控途径相似并且能够重现物种的进化历史等特征.

生命科学领域的多个学科（包含功能基因组学、基因组注释、分子系统发生学、进化生物学等）的研究依赖于直系同源基因类群的识别, 例如物种新发现基因的功能预测、生物体之间基因注释的转移、系统发生关系的构建及重现基因的进化历史等. 在系统发生分析方面, 由于系统发育树的构建需要不同群体间的直系同源基因, 故直系同源基因的识别对系统发生分析是至关重要的. 在进化生物学方面, 完整的直系同源基因的识别方法, 能够重现基因的进化过程[5].

目前发现的功能千差万别的基因最初都是由少量祖先基因通过基因加倍、变异和功能域重组产生的. 因此, 通过基因序列的比较, 可从同一物种或不同物种中找到同源的基因成员. 随着基因组数据的增加, 鉴定和区分这些具有相同或者不同功能的同源基因成为功能基因组研究的一个重要内容. 基因组注释依赖于精确的直系同源基因的识别[6], 并有助于预测新基因的功能以及对控制元件进行识别[7]. 同时, 识别直系同源基因可以帮助重建进化历史, 了解垂直遗传关系和谱系特有的基因丢失以及基因水平转移.

2　直系同源基因识别的方法

直系同源关系是进化历史形成的, 无法通过具体的实验来鉴别, 只能通过生物信息学的方法, 从序列差异性上来推测不同物种之间的直系同源关系. 所有预测直系同源方法的理念是: 在生物进化过程中越相近的基因, 其序列结构与功能相似性程度就越高. 现有的识别直系同源基因的方法大致可分为三类: 一是比较序列相似性来识别直系同源基因; 二是通过构建系统发育树来识别直系同源关系; 三是基于序列相似性和系统发育树的混合方法.

2.1　基于序列相似性的方法

基于序列相似性的方法适于从两个或者多个全基因组或蛋白质组中推断直系同源关系, 主要依赖于通过所有序列计算出的成对序列的相似性得分. 这一方法需要首先计算出成对序列相似性的得分, 并使用聚类算法来识别直系同源类群[8]. 表 1 归纳了基于序列相似性方法的直系同源识别的数据库资源. 基于序列相似性的方法具有识别速度快以及灵敏度高等优点, 它适用于比较大量物种间蛋白质的进化关系[9]. 但该方法错误较多, 尤其是当基因缺失发生时, 就会漏掉一些复制事件, 从而预测出更多的直系同源基因, 使得一个聚类中包含很多的旁系同源关系.

2.2　基于系统树的方法

基于系统树的方法是根据系统树来推断直系同源和旁系同源关系. 这一方法首先要收集同源基因, 进行多序列比对, 再构建系统树, 进行基因树和物种树的调和. 然后, 根据已知的外群建立有根数, 确定节点, 区分物种形成事件和复制事件[17], 判断出直系同源和旁系同源关系. 表 2 归纳了基于系统发育树的方法的直系同源识别的方法和资源.

现有的基于系统树的识别方法具有错误少、直观性好以及信息量大等优点, 而且呈现了基因家族的进化历史[18]. 但是该方法也存在一些缺点: 第一, 现存的系统树的构建方法很难产生完整的可信树, 调和树的方法识别直系同源和旁系同源关系需要比较基因树和物种树, 但构建精确的基因树和物种树对研究者来说是一项巨大的挑战; 第二, 基于系统树的方法需要大量时间以及庞大的计算; 第三, 系统树的构建和多序列的比对方法跟不上大量增长的序列数据; 第四, 基于系统树的方法需要合适的有根树, 若树根选择不当, 会造成系统树结构的巨大变化, 影响直系同源的识别结果[19].

2.3　混合方法

此外, 还有一些数据库结合了序列相似性和系统树的方法来识别直系同源关系. 例如 Ensembl Compara; OPM (OrthoParaMap) 以及 PhyOP (Phylogenetic orthology and paralogy) [25], Orthoinspector[26]等. 混合的直系同源基因识别的方法, 弥补了基于系统树或序列相似性的方法的几个缺点, 计算时使用两种方法, 更有可能提供丰富的系统发生的内容和功能关系.

3　问题与展望

由于基因组序列的快速积累, 使得研究者从新基因组中获取更多的功能和进化信息变得极富

表 1 基于序列相似性识别直系同源关系的方法和数据库

Table 1 Methods and databases based on sequence similarity for identification of orthologs

Resources	Methods and traits	Websites
COG (clusters of orthologous groups)	较早的一个识别直系同源的数据库,是通过对完整的原核生物的蛋白质序列大量比较而来的,现在已经扩展到包含 630 个完整的基因组.	http://www.ncbi.nlm.nih.gov/COG/
eggNOG (evolutionary genealogy of genes: non-supervised orthologous groups)	对直系同源类群进行了功能描述和功能分类的注释;包含了 1 133 个物种的直系同源类群.	http://eggnog.embl.de
OrthoMCL-DB	使用了 Markov 聚类方法;OrthoMCL-DB version 5 包含了 150 种大部分的真核生物基因组以及一些细菌和古生菌.	http://www.orthomcl.org/
Inparanoid/multiparanoid	可以区分 in-paralogs 和 out-paralogs,最新的 7.0 版本包含了 99 种真核生物并将大肠杆菌作为外群[10].	http://InParanoid.sbc.su.se
OMA(orthologous matrix)	目前包含基因组最多的数据库, 包含 1320 个真核和原核基因组.	http://omabrowser.org
RoundUp	使用了互为最小距离 (reciprocal smallest distance, RSD)算法;版本 3 包含 1807 个基因组.	http://roundup.hms.harvard.edu/
OrthoDB(the hierarchical catalog of orthologs)	主要针对真核生物蛋白编码基因的直系同源关系;结合了 GO 和 InterPro 对直系同源类群进行描述;包含了 48 种脊椎动物、33 种节肢动物、73 种真菌以及 12 种后生动物[11].	http://cegg.unige.ch/orthodb
MSOAR	结合了序列相似性和基因组重排的方法来识别[8].	http://msoar.cs.ucr.edu
OrthoSelect	是一个在 ESTs 数据库中查询直系同源类群的工具;它可以通过安装本地软件来使用或者通过网站来查询;准确性高,可以处理各种来源的 cDNA 序列[12].	http://gobics.de/fabian/orthoselect.php
P-POD(princeton protein orthology database)	该数据库链接了相关人类疾病信息的数据库;给生物学家研究他们感兴趣的基因的进化历史提供帮助,从而能够快速地获得该基因的进化信息以及其它相关信息[13].	http://ortholog.princeton.edu
BLASTO	整合了许多直系同源数据库, 包括 NCBI 的 COG、EOG、OrthoMCL, MultiParanoid 和 TIGR 的 EGO;利用直系同源基因的信息对个体进行序列的功能推断和进化研究[14].	http://oxytricha.princeton.edu/BlastO
Proteinortho	主要包含 NCBI 中的 717 个真细菌基因组[15].	http://www.bioinf.uni-leipzig.de/Software/proteinortho/
QuartetS-DB	使用了基于序列相似性的 QuartetS 算法来识别;提供了 1 621 个全基因组(包含 1 365 种细菌,92 种古生菌以及 164 种真核生物)的直系同源的预测[16].	http://applications.bioanalysis.org/quartetsdb

表 2 基于系统发育树识别直系同源的方法和数据库

Table 2 Methods and databases based on trees for identification of orthologs

Resources	Methods and traits	Websites
TreeFam	使用了多种系统发生的方法识别直系同源; 新版本包含 25 种全动物基因组序列,外加 4 种植物和真菌的外群.	http://www.treefam.org
LOFT(levels of orthology from trees)	构建了各种分层类群, 强调直系同源和旁系同源不同水平的关联性;适用于大规模的系统发生分析.	http://www.cmbi.ru.nl/LOFT
RIO(resampled inference of orthologs)	使用 SDI(speciation duplication inference)算法来推断物种形成和复制事件;根据自举多次重新抽样基因树,以评估识别结果的可靠性[20].	http://rio.janelia.org/
PhylomeDB	使用了高质量的系统发生方法包括进化模型测试和比对校正;最新版本从各种有机体中选取了包含人、酵母甚至细菌基因组[21].	http://phylomeDB.org
HCOP(HGNC comparison of orthology predictions)	结合了多种识别数据库,提高了预测的准确性;提供了有用的一站式的资源来总结、比较、获得各种人类和鼠的直系同源数据[22].	http://www.genenames.org/hcop
PHOG (phyloFacts orthology group)	使用 PhyloFacts 的蛋白家族系统发生信息, 可以预测 super-orthologs.	http://phylogenomics.berkeley.edu/phog/
OrthologID	可以识别直系同源类群以及查找单个基因的直系同源基因;现在包含一些植物全基因组,以单细胞生物衣藻为外群构建系统发育树[23,24].	http://nypg.bio.nyu.edu/orthologid/

有挑战性. 最大挑战不是缺少直系同源关系识别的方法, 而是过多的方法和数据库分别满足了部分人的需要, 这样的不均衡性成为现在主要的障碍. 另外, 缺少统一标准和格式也使得整合或者比较这些不同的直系同源数据集变得相对困难[27]. 直系同源基因来源于物种形成进化历史, 所以它只能通过推测而识别; 但基因水平转移、基因丢失、基因融合和基因分裂[28]等伴随物种形成过程发生的大量进化事件, 可能导致错误的直系同源关系的预测; 此外, 不同谱系的基因丢失也可能导致错误的直系同源预测. 解决方法是选取更多的样本来构建基因家族的进化历史.

基因组学研究已经进入后基因组时代, 研究重心从结构基因组学过渡到功能基因组学. 面对大规模测序产生的大量信息学数据, 基因的识别, 特别是直系同源基因的识别, 是获得序列功能注释的首要条件. 然而, 一个全自动的、更精确的识别直系同源基因的程序对比较基因组而言仍是一个近期内难以实现的目标.

参考文献(References):

[1] SONNHAMMER E L, KOONIN E V. Orthology, paralogy and proposed classification for paralog subtypes[J]. Trends in Genetics, 2002, 18(12): 619-620.

[2] GABALDÓN T, DESSIMOZ C, HUXLEY-JONES J, *et al*. Joining forces in the quest for orthologs[J]. Genome Biology, 2009, 10(9): 403.

[3] CHEN F, MACKEY A J, VERMUNT J K, *et al*. Assessing performance of orthology detection strategies applied to eukaryotic genomes[J]. PLoS One, 2007, 2(4): e383.

[4] 潘增祥, 许丹, 张金璧, 等. 基于直向同源序列的比较基因组学研究[J]. 遗传(PAN Zeng-xiang, XU Dan, ZHANG Jin-bi, *et al*. Reviews in comparative genomic research based on orthologs[J]. Hereditas), 2009, 31(5): 457-463.

[5] CONTE M G, GAILLARD S, DROC G, *et al*. Phylogenomics of plant genomes: a methodology for genome-wide searches for orthologs in plants[J]. BMC Genomics, 2008, (9): 183.

[6] ALTENHOFF A M, DESSIMOZ C. Phylogenetic and functional assessment of orthologs inference projects and methods[J]. PLoS Computational Biology, 2009, 5(1): e1000262.

[7] SENNBLAND B, LAGERGREN J. Probabilistic orthology analysis[J]. System Biology, 2009, 58(4): 411-424.

[8] SHI G, PENG M C, JIANG T. MultiMSOAR2. 0: an accurate tool to identify ortholog groups among multiple genomes[J]. PLoS One, 2011, 6(6): e20892.

[9] TRACHANA K, LARSSON T A, POWELLL S, *et al*. Orthology prediction methods: a quality assessment using cruated protein families[J]. Bioessays, 2011, 33(10): 769-780.

[10] ÖSTLUND G, SCHMITT T, FORSLUND K, *et al*. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis[J]. Nucleic Acids Research, 2010, 38(suppl 1): 196-203.

[11] WATERHOUSE R M, ZDOBNOV E M, TEGENFELDT F, *et al*. OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011[J]. Nucleic Acids Research, 2011, 39(suppl 1): 283-288.

[12] SCHREIBER F, PICK K, ERPENBECK D, *et al*. OrthoSelect: a protocol for selecting orthologous groups in phylogenomics[J]. BMC Bioinformatics, 2009, (10): 219.

[13] HEINICKE S, LIVSTONE M S, LU C, *et al*. The princeton protein orthology database (P-POD): a comparative genomics analysis tool for biologists[J]. PLoS One, 2007, 2(8): e766.

[14] ZHOU Y, LANDWEBER L F. BLASTO: a tool for searching orthologous groups[J]. Nucleic Acids Research, 2007, 35(Suppl 2): 678-682.

[15] LECHNER M, FINDEIB S, STEINER L, *et al*. Proteinortho: detection of (Co-)orthologs in large-scale analysis[J]. BMC Bioinformatics, 2011, 12: 124.

[16] YU C, DESAI V, CHENG L, *et al*. QuartetS-DB: a large-scale orthology database for prokaryotes and eukaryotes inferred by evolutionary evidence[J]. BMC Bioinformatics, 2012, 13(1): 143.

[17] VAN DER HEIJDEN R T, SNEL B, van NOORT V, *et al*. Orthology prediction at scalable resolution by phylogenetic tree analysis[J]. BMC Bioinformatics, 2007, (8): 83.

[18] FU Z, JIANG T. Clustering of main orthologs for multiple genomes[J]. Journal of Bioinfomatics and Computational Biology, 2008, 6(3): 573-584.

[19] SJÖLANDER K, DATTA R S, SHEN Y, *et al*. Ortholog identification in the presence of domain architecture rearrangement[J]. Briefings in Bioinformatics, 2011, 12(5): 413-422.

[20] ZMASEK C M, EDDY S R. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs[J]. BMC Bioinformatics, 2002, (3): 14.

[21] HUERTA-CEPAS J, CAPELLA-GUTIERREZ S, PRYSZCZ L P, *et al*. PhylomeDB v3. 0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions[J]. Nucleic Acids Research, 2011, 39(Suppl 1): 556-560.

[22] WRIGHT M W, EYRE T A, LUSH M J, *et al*. HCOP: The HGNC comparison of orthology predictions search tool[J]. Mammalian Genome, 2005, 16(11): 827-828.

[23] CHIU J C, LEE E K, EGAN M G, *et al*. OrthologID: automation of genome-scale ortholog identification within a parsimony framework[J]. Bioinformatics, 2006, 22(6): 699-707.

[24] EGAN M, LEE E K, CHIU J C, *et al*. Gene orthology assessment with orthologID [J]. Methods in Molecular Biology, 2009, 537: 23-38.

[25] KUZNIAR A, VAN HAM R C, PONGOR S, *et al*. The quest for orthologs: finding the corresponding gene across genomes[J]. Trends in Genetics, 2008, 24(11): 539-551.

[26] LINARD B, THOMPSON J D, POCH O, *et al*. OrthoInspector: comprehensive orthology analysis and visual exploration[J]. BMC Bioinformatics, 2011, (12): 11.

[27] KIM K M, SUNG S, CAETANO-ANOLLÉS G, *et al*. An approach of orthology detection from homologous sequence under minimum evolution[J]. Nucleic Acids Research, 2008, 36(17): e110.

[28] KRISTENSEN D M, WOLF Y I, MUSHEGIAN A R, *et al*. Computational methods for gene orthology inference[J]. Briefings in Bioinformatics, 2011, 12(5): 379-391.