生物物理学报　第二十卷　第二期　二〇〇四年四月
ACTA BIOPHYSICA SINICA　Vol.20　No.2　Apr. 2004

ortholog —— 概念、生物信息预测方法和数据库

陈作舟[1,2]，　朱　晟[2]，　薛成海[3]，　陈良标[2]
（1. 浙江大学生命科学院，杭州 310029; 2. 中国科学院遗传与发育生物学研究所，北京 100080;
3. 中国科学院自动化研究所 复杂系统与智能科学国家重点实验室，北京 100080）

摘要：orthologs 指起源于不同物种的最近的共同祖先的一些基因。orthologous 的基因，具有相近甚至相同的功能，由相似的途径调控，在不同的物种中扮演相似甚至相同的角色，因此在基因组序列的注释中，是最可靠的选择。orthologs 的生物信息预测方法主要有两类：系统发生方法和序列比对方法。这两类方法都是基于序列的相似性，但又各有特点。系统发生方法通过重建系统发生树来预测 orthologs，因此在概念上比较精确，但难于自动化，运算量也很大。序列比对方法在概念上比较粗糙，但简单实用，运算量相对较小，因此得到了较广泛的应用。

关键词：水平同源；垂直同源；基因重复；物种形成；系统发生；生物信息学
中图分类号：Q75

1　引　　言

随着人类基因组计划及其它基因组的测序计划得到越来越多的数据，人们把注意力逐渐转向这些序列的功能上来。为了使我们在进行耗时费力的实验之前，能够尽量准确地预测其功能，这些数量庞大的序列需要先进行生物学功能的数字化注释。

orthologous（ortholog 的形容词形式）的蛋白序列，被认为具有相近甚至相同的功能，被相似的途径调控，在不同的物种中扮演相似甚至相同的角色[1~3]，因此在基因组序列的注释中，是最可靠的选择[4]。尽管以上的假设还有待于实验的证明，但很多例子都支持以上假设[5,6]。如果已知蛋白 A 和 B 是一对 orthologs，当我们知道蛋白 A 的功能，我们就能比较有把握地预测 B 也有类似甚至相同的功能，这是核酸和蛋白序列电子注释（electronic annotation）的一种根本方法。

由于确认 ortholog 的过程耗时耗力，因此，已经被确认的 orthologs 少之又少[7]。所以，用生物信息的方法预测 ortholog 就成为了主流。

2　ortholog 及相关概念

ortholog 和 paralog 的概念最先由 Fitch 在 1970 年提出[8]，近年来已成为讨论的热点[5,9~11]。

2.1　ortholog 和 paralog 的基本概念

orthologs（垂直同源基因）：指一些基因，这些基因起源于这些基因所在物种的最近的共同祖先的一条基因[4,8,12]。

paralogs（水平同源基因）：指一些基因，这些基因由一个共同的祖先基因通过基因重复（duplication）而来[4,8,12]。

以上的定义是从系统发生（phylogenetic）的角度（即进化的角度）来定义 ortholog 和 paralog 的，这种定义已经被普遍接受[7,12]。由于 orthologous 的基因往往具有相同或相似的功能，因此有很多人试图用功能的定义来替代系统发生学的定义，认为 orthologs 是不同物种具有相同或相似功能的基因；又由于 orthologous 的基因在序列上有很大的相似性（similarity），又有人试图用序列的相似性来定义 ortholog。这两种定义都是错误的：因为这种关系不是绝对的——并非所有的 othologs 功能都相似，也并非不同物种中序列相似性最大的两条序列就一定是 orthologs[7]。

2.2　更加复杂的情况

orthologs 形成的原因是物种形成（speciation），而 paralogs 形成的原因是基因重复（duplication）。当物种形成和基因重复多次发生时，情况就会变得复杂，图 1 设计了一个假想的情况来

收稿日期：2003-07-10
基金项目：国家自然科学基金项目（90208022）
通讯作者：陈良标，电话：（010）62554807，
E-mail: lbchen@genetics.ac.cn

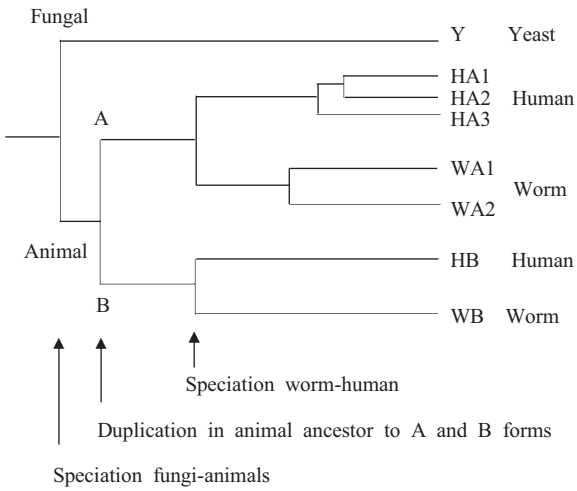

Fig.1 An imagined phylogenetic tree[13]

说明这个问题[13]：一个祖先的基因遗传给酵母（yeast）、线虫（worm）和人类（human），这个基因在“动物”形成之后，“线虫”（worm）和“人类”（human）分开之前发生了一次基因重复（duplication），形成A类和B类基因，在“线虫”和“人类”物种分开以后，各自又发生了独立的基因重复。这里，所有的“人类”基因（HA1、HA2、HA3、HB）和所有的线虫基因（WA1、WA2、WB）都是酵母基因Y的orthologs；反之亦然，“人类”的基因HA1、HA2、HA3和线虫的基因WA1、WA2互为orthologs，也称为co-orthologs [13]。而HA1、HA2、HA3互为in-paralogs，WA1和WA2也互为in-paralogs [4,13]（当然属于paralogs）。所谓in-paralog，指HA1、HA2、HA3基因形成的基因重复（duplication）发生在“人类”的物种形成之后，由于形成较晚，HA1、HA2、HA3之间相似度一般较高。相应地，HA1和HB是一对out-paralogs（当然也属于paralogs），这是因为基因重复（duplication）产生A和B的过程发生在“人类”和“线虫”的物种形成以前[4,13]。

当出现基因的横向转移（horizontal transfer）、趋同进化（convergence）、基因丢失（gene loss）或基因融合（gene fusion）时，这种关系就变得更为复杂，Fitch[12]曾对此作过详细的讨论。

2.3 ortholog的一些性质

(1) ortholog是相互的（reflexive）[12]，如图1中，Y是HA1的ortholog，反之亦然。

(2) ortholog是不可传递（non-transitive）的[12]，在图1中HA1是Y的ortholog，HB也是Y的ortholog，但HA1和HB却不是orthology的关系，而是一对paralogs。

(3) orthology不一定是一对一的关系（one-to-one relationship），它还可以是一对多(one-to-many)或多对多（many-to-many）的关系[7,14]。人们时常犯的一个错误就是，认为一个物种中每一条基因在另一个物种中都有一条orthologous的基因相对应[7]，事实上，这仅仅是其中的一种情况（一对一的关系）。如图1中，HA1、HA2、HA3是Y的orthologs，这显然是一对多的关系；HA1、HA2、HA3和WA1、WA2两组基因互为orthologs，则显然是多对多的关系。

3 ortholog的生物信息预测

从ortholog的定义我们可以看出，othology是指进化（系统发生）上的一种关系。由于过去的进化历史是无法重演的，因此，从绝对意义上说，orthology是无法用实验证实的。所以，我们只能用一些方法（如建立系统发生树）来推测序列在进化上的关系，然而不管何种方法，都是基于一个理念——进化上接近的序列，其相似度（包括序列、结构、功能等）也较大（当然这并不绝对）。所以，我们只能通过相似性的关系来推测orthology，而现今绝大多数数据都是序列数据，因此，现在的绝大多数方法都是基于序列的相似关系来推测orthologs。

3.1 推测ortholog的系统发生方法

由于orthology是一种系统发生关系，因此，最自然的方法莫过于重建系统发生树，再推测orthologs [4]。为了得到两个全基因组之间的orthologs，需要[4]：

(1) 将同源的基因聚类（clustering）

(2) 将每个聚类的同源结构域进行多序列联配（mutiple sequence alignment）

(3) 利用多序列联配的结果重建系统发生树（phylogenetic tree）

(4) 从这些“树”中抽取出orthologs

然而这个理想方法有着诸多困难，很难用计算机自动实现。其中第二步现在并没有很好的办法，其它三步虽然已有现成的办法，但其精确性仍有不足，所以最终得到的结果是不可靠的[4]。

Yuan等[15]曾做过类似的工作，可以作为一个很好的例子：

(1) 输入一条序列，用这条序列对一个数据库作blast[16]，序列相似度达到一定域值的所有序列被取

出（认为它们是一组同源序列）；

(2) 将得到的这组序列进行多序列联配（mutiple sequence alignment），利用的工具是 ClusteralW[17]；

(3) 利用多序列联配的结果，重建系统发生树（phylogenetic tree），同样也是用 ClusteralW[18]，得到的树称为基因树（gene tree）；

(4) 将得到的基因树和美国国立生物技术信息中心的物种树（species tree）相比较，从而推测出 orthologs。这种算法称为 reconcile tree [19]；

Yuan 等的工作做的较早，近年来又有一些人对以上各步设计或其算法进行改进，特别是从“系统发生树”中抽取出 orthologs 的算法[20,21]。

3.2 推测 ortholog 的序列比对方法

3.2.1 inparanoid 和 COG

虽然推测 orthologs 系统发生方法理论上比较精确，但两个主要问题始终得不到很好的解决：以上工作还难于自动化，只有人工参与才有较好的效果；即使利用计算机自动运行，所需的运算能力也极为庞大[4]。因此，近年来人们发展了一类基于序列两两比对（pairwise alignment）的 orthologs 判别算法，这些算法理论上相对粗糙，但却简单实用，所需的运算能力也小得多。它的基本思想很简单，即两个基因组中彼此序列最为相似的一组基因作为一组 orthologs。虽然 Günter[7]对此有过批评，认为判断 orthologs 的唯一正确办法是系统发生（phylogenetic）方法，然而这类方法不失为折衷的好办法，而且已经得到了较广泛的应用[2,4,22,23]。本文以其中一个算法“inparanoid”[4]为例来详细说明其原理，见图 2。

整个过程包括以下步骤（序号对应图中序号）：

(1) 输入两个 fasta 格式的文件，每一个文件包含一个物种的蛋白质组（这个物种的非冗余的所有蛋白序列），分别记作 A 和 B，将这两个文件的序

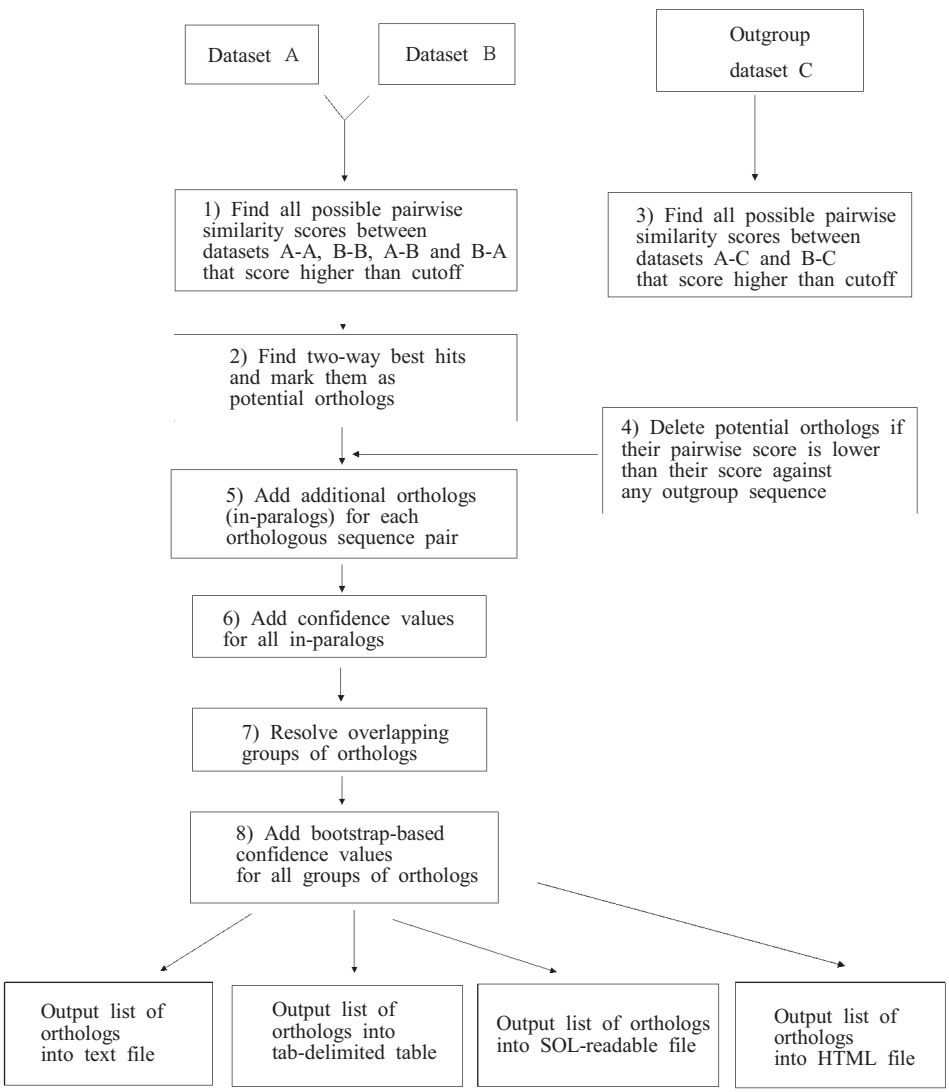

Fig.2 The process of inparanoid to get orthologs

列进行两两比对（pairwise alignment），得到两两比对的分值（用以衡量两个序列的相似度）。这个过程可以用 blast[16]实现，也可以用其它序列比对工具实现。

(2) 取出相互之间分值最高的所有“序列对”，作为“基本 orthologs”。

(3) 如果用户选择了“outgroup”的选项，那么用户还将输入另一个物种的蛋白质组 C，这个蛋白质组所属的物种离 A、B 的亲缘关系比 A 和 B 之间更远。然后做 A 和 C，以及 B 和 C 的序列两两比对（pairwise alignment）。

(4) 由于 C 是 outgroup，它的亲缘关系离 A 和 B 较远，因此，它和 A 或 B 的 ortholog 之间的相似度，理论上要比 AB 之间的小，换言之，若 AB 之间的一对“基本 orthologs”，例如 A1-B1 能在 C 中找到一条序列 C1，它和 A1 或 B1 的相似度比 A1-B2 之间更大，则说明 A1-B1 并不是一对真正的 orthologs，因此要把这样的 A1-B1 对去掉。这样做可以去掉假阳性（false positive）的 orthologs，但同时增加了大约一半的运算量。

(5) 对每一个基本的 orthologs（如 A1-B1），加入其 in-paralogs。如前所述，orthologs 并不一定是一对一的关系，还可以是一对多或多对多的关系。如图 1 中 HA1、HA2、HA3 和 WA1、WA2 就是一组多对多的 orthologs。假如 HA1 和 WA1 由于序列最为相似，作为一对“基本 orthologs”，则 HA2、HA3、WA2 就是 HA1-WA1 的 in-paralogs。其原理如图 3，图中黑点表示蛋白组 A 中的序列，圆圈表示蛋白组 B 中的序列，彼此距离最近（相似度最高）的一对序列 A1-B1 作为一对“基本 orthologs”。由于 in-paralogs 是物种形成（speciation）后由基因重复（duplication）形成的序列，形成的时间比 A1-B1 分开的时间晚，因此相似度也更高（当然并不绝对）。因此，将 A 中的一些序列取出作为 A1-B1 的 in-paralogs，这些序列满足：它们离 A1 的距离（衡量相似度）比 B 中的任一条序列更接近。

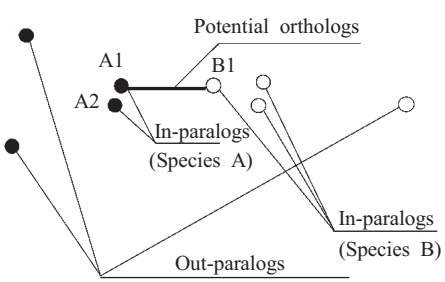

Fig.3 The prediction of in-paralogs

(6) 加入 in-paralogs 的可置信度。推测得到的 in-paralogs 序列，有的和 A1-B1 相似度高，有的则较低。因此，需要一个值来衡量一个 in-paralog 和“基本 orthologs”（如图 3 中的 A1-B1）的相似程度，称为置信度，任意序列 A2 的置信度是 confidence for A2＝100％(scoreA1A2−scoreA1B1)/(scoreA1A1−scoreA1B1)，其中 confidence for A2 指 A2 的置信度，scoreA1A2 指序列 A1 和 A2 进行序列比对的分值（衡量 A1 和 A2 的相似程度），余者类推。

(7) 解决一些 orthologs 组相互重叠的问题，达到一定域值时，将两个 orthologs 组合并。

(8) 用 bootstrap 的方法计算每一个 orthologs 的可置信度，其算法见原文[4]和一些相关文献[24]。

(9) 以四种格式输出结果：Text, 以 Tab 键分隔的表, 能够被 SQL 语句阅读的表, HTML。

以上是 inparanoid 的基本算法，它通过对蛋白序列两两比对（pairwise alignment）结果的分析，得出 orthologs 组，包括 in-paralogs。读者可以通过访问网页 http://www.cgb.ki.se/inparanoid/ 来访问其数据库和免费下载这个工具。和它类似的还有 COG 数据库,其全称是 Clusters of Orthologous Groups of Proteins[2](http://www.ncbi.nih.gov/COG)。它和 inparanoid 很类似，但各有侧重：inparanoid 执行两个物种的比较，基本原理是两两最佳比对（two-way best match），而 COG 数据库是多个物种的 orthologs，一个 orthologs 组至少有三个物种（three-way best match）(三条序列互为最佳匹配序列)，还有多达几十个物种的；inparanoid 的优势在于它能够找出 in-paralogs。

3.2.2 TOGA 和 HomoloGene

inparanoid 和 COG 使用的对象是蛋白质组。但是由于真核生物的基因组特别庞大，得到测序的真核生物基因组并不多，所以对于真核生物，如果是仅从已知的蛋白序列和从基因组中推测的蛋白序列（predicted proteins），其物种数并不多。另外，我们当前用的所谓某一物种全基因组的蛋白序列，其实大部分是从基因组中推测的蛋白序列（predicted proteins），但是现有的基因预测程序又有明显的不足[25]，因此这些蛋白质组中预测的序列，并不怎么可靠，得到的 orthologs 也就不怎么可靠了。

正是由于使用对象（蛋白质组）有以上两个不足，因此人们考虑使用 EST（expressed sequence tag）和 mRNA 的序列。这有两个优点：一是这些序列是直接测序得到的，不像很多蛋白序列是计算机预测的，相对比较可靠；二是这些序列非常之多，dbEST（http://www.ncbi.nlm.nih.gov/dbEST/index.html）中的 EST 序列已超过 16 000 000 条。这正好弥补了蛋白质组的以上两个缺点。然而 EST 也有其自身的缺点：一是 EST 往往带有 3' 或 5' 端的 UTR，可能会对预测 orthologs 有所影响，不过 Makałowski 等[26]发现人和鼠的 orthologs，在 UTR 处也比较保守，这使得我们有可能找到核酸序列的 orthologs[22]；另一个缺点是 EST 是 cDNA 片断，并且测序中的错误也比较多，这就要求先要对 EST 进行预处理，如去除载体序列和拼接等。

一个例子是 TOGA（TIGR orthologous gene alignments）[22]，首先将 dbEST（NCBI 的 EST 数据库）的所有 EST 进行预处理，包括去除载体序列，去除污染的细菌的序列，去除 poly-A/T 的尾巴，然后将这些 EST 加上 EGAD（由 TIGR 管理，http://www.tigr.org/tdb/egad/egad.html）的转录序列，进行拼接。如果两个序列有 40 个氨基酸以上的重叠，重叠部分的一致率大于 95%，且非重叠部分小于 20 个氨基酸，则将这两个序列拼接起来[27,28]，拼接所用的工具是 CAP3[29]。然后将不同物种拼接好的序列进行两两比对（BLASTN），然后把 n 个物种的 n 条序列作为一组 orthologs，这 n 条序列必须满足以下条件：$n \geqslant 3$ 并且任意两条序列都是两两最佳比对（three-way best match）；n 条序列分别来自 n 个不同的物种；任意两条序列的 BLASTN 的 E 值必须小于 10^{-5}。

HomoloGene（http://www.ncbi.nlm.nih.gov/HomoloGene/）的算法和以上几种方法也很类似，所不同的是 HomoloGene 取的既有 UniGene（http://www.ncbi.nlm.nih.gov/UniGene）的 EST 序列又包括基因组中预测的序列，另外它只要求two-way best match，并不要求 three-way best match。

参考文献：

[1] Makarova K, Aravind L, Galperin M, Grishin N, Tatusov R, Wolf Y, Koonin E. Comparative genomics of the archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. *Genome Res*, 1999,9: 608~628

[2] Tatusov R, Galperin M, Natale D, Koonin E. The cog database: a tool for genome-scale analysis of protein functions and evolution. *Nucl Acids Res*, 2000,28:33~36

[3] Gelfand M, Koonin E, Mironov A. Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucl Acids Res*, 2000,28:695~705

[4] Remm M, Storm C, Sonnhammer E. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 2002,314:1041~1052

[5] Gerlt J, Babbitt P. Can sequence determine function? *Genome Biol*, 2000,1:5

[6] Gerlt J, Babbitt P. Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem*, 2001,70:209~246

[7] Thei β en G. Othology: secret life of genes. *Nature*, 2002, 415:741

[8] Fitch W. Distinguishing homologous from analogous proteins. *Syst Zool*, 1970,19:99~113

[9] Koonin E. An apology for orthologs —— or brave new memes. *Genome Biol*, 2001,2(4):COMMENT 1005.1~1005.2

[10] Petsko G. Homologuephobia. *Genome Biol*, 2001,2 (2): COMMENT 1002.1~1002.2

[11] Jensen R. Orthologs and paralogs —— we need to get it right. *Genome Biol*, 2001,2(8):INTERACTION 1002.1~1002.3

[12] Fitch W. Homology: a personal view on some of the problems. *Trends Genet*, 2000,16:227~231

[13] Sonnhammer E. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*, 2002,18:619~620

[14] Storm C, Sonnhammer E. Automated ortholog inference from phylogenetic trees and caculation of ortholog reliability. *Bioinformatics*, 2002,18:92~99

[15] Yuan Y, Eulenstein O, Vingron M, Bork P. Towards detection of othologues in sequence databases. *Bioinformatics*, 1998,14:285~289

[16] Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. *J Mol Biol*, 1990,215:403~410

[17] Higgins D, Thompson J. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*, 1996,266:382~402

[18] Saito N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenitic trees. *Mol Biol Evo*, 1987,4: 406~425

[19] Page R, Charleston M. From gene to organismal phylogeny:

Rconciled trees and the gene tree/species tree problem. *Mol Phylogenet Evol*, 1997,7:231~240

[20] Page R. GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 1998,14:819~820

[21] Storm C, Sonnhammer E. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 2002,18:92~99

[22] Lee Y, Sultana R, Pertea G, Cho J, Karamycheva S, Tsai J, Parvizi B, Cheung F, Antonescu V, White J, Holt I, Liang F, Quackenbush J. Cross-referencing eukaryotic genomes: TIGR orthologous gene alignments (TOGA). *Nucl Acids Res*, 2002,12:493~502

[23] Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. *J Comp Biol*, 2000,7: 203~214

[24] Zharkikh A, Li W. Estimation of confidence in phylogeny: the complete-and-partial bootstrap techniques. *Mol Phylogenet Evol*, 1995,4:44~63

[25] Guigo R, Agarwal P, Abril J, Burset M, Fickett J. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res*, 2000,10:1631~1642

[26] Makałowski W, Boguski M. Evolutionary parameters of the transcribed mammalian genome: An analysis of 2 820 orthologous rodent and human sequences. *Proc Natl Acad Sci*, 1998,95:9407~9412

[27] Liang F, Holt I, Pertea G, Karamycheva S, Salzberg S, Quackenbush J. An optimized protocol for analysis of EST sequences. *Nucleic Acids Res*, 2000,28:3657~3665

[28] Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana, R, White J. The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acid Res*, 2001,29: 159~164

[29] Huang X, Madan A. CAP3: A DNA sequence assembly program. *Genome Res*, 1999,9:868~877

ORTHOLOG —— CONCEPT, BIOINFORMATIC INFERENCES AND DATABASES

CHEN Zuo-zhou[1,2], ZHU Sheng[2], XUE Cheng-hai[3], CHEN Liang-biao[2]

(*1. College of Life Sciences, Zhejiang University, Hangzhou 310029, China;*
2. Institute of Genetics and Developmental Biology, The Chinese Academy of Sciences, Beijing 100080, China;
3. Key Laboratory of Complex Systems and Intelligence Science,
Institute of Automation, The Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Orthologs are genes in different species that originate from a single gene in the last common ancestor of these species. Orthologous genes are suggested to share similar functions, be regulated by similar biochemical pathways and play similar roles in different species. Thus, it is the best choice to use orthologous genes when annotating newly discovered genes. There are mainly two categories of algorithms for predictions of orthologs: phylogenetic algorithms and sequence comparison algorithms. Both of them are based on sequence similarities, whereas they have their own characteristics. Phylogenetic ways predict orthologs by reconstructing phylogenetic trees. As a result, they are conceptually accurate, but hard to automate, and demanding huge amount of computational resources. In contrast, the latter methods are conceptually less accurate but not as complex and require less computational resources, therefore, widely used.

Key Words: Ortholog; Paralog; Duplication; Speciation; Phylogeny; Bioinformatics