文章编号:1000-5862(2017)05-0470-06

基于随机聚类方法建模的序列分析

张 巍[1],王 洋[1],刘东宁[1],滕少华[1],张 莉[2],徐新爱[2]
(1. 广东工业大学计算机学院,广东 广州 510006;2. 南昌师范学院,江西 南昌 330032)

摘要:大数据下的系统发育估计是一个组合优化问题,在有限计算时间内,现有算法很难为大量序列数据的分析提供最优解. 基于前人启发式算法,提出了一种系统发育树随机聚类建树方法,可在较短时间内为系统发育过程产生的大规模序列数据提供所有具有进化意义的解及最优解,以揭示发育过程中的序列进化关系. 实验结果表明,该随机聚类方法是行之有效的,对生物计算及系统发育相关领域研究具有积极意义.

关键词:随机聚类算法;序列分析;系统发育
中图分类号:TP 391 文献标志码:A DOI:10.16357/j.cnki.issn1000-5862.2017.05.05

0 引言

系统发育分析发展至今,研究者们研究和开发了大量工具和方法. 作为系统发育主要的研究内容,序列比对和多序列比对分析旨在研究序列间的差异以揭示基因数据在分子水平上的进化关系.

系统发育研究的结果和推论是根据同源基因数据的相似性和差异性分析得到的,国内外主要的分析方法有邻接法(NJ)[1]、最大似然法(ML)[2]和最大简约法(MP)[3]等,多年来众多研究成果表明了这些方法的性能和效率. 但也有很多研究者致力于优化这些方法,或尝试寻找出其他高效的方法以促进生命科学的发展,这是由于当前每一种方法都存在着各自缺陷. 例如邻接法的邻接矩阵计算会导致忽略掉最优解,最大似然法的估值计算过程会导致大量的计算时间,最大简约法对数据的要求较高,而且计算结果在某些情况下会出现较大误差等[4]. 此外,这些方法往往需要对计算结果进行人工评估或编辑修改,未经处理的计算结果与实际现象的偏差较大.

考虑到系统发育是一个开放性的问题,近期研究表明,全面的序列关系进化分析在大的数据规模下会得到很多有价值的信息. 因此系统发育分析的结果并不应该是一个唯一的最优解,而应该是一组最优解的集合,其中包括所有其它的合理的次优解. 在这种思想下,系统发育分析应考虑到所有的序列进化关系,并在已验证的或假设的序列背景下进行择优选择.

另一方面,聚类分析是一种有效的数据分析统计方法[5-6],在生命科学中常用于对动植物的基因进行分类,以获取对种群固有结构的认识. 使用聚类方法进行启发式的分析具有一定的难度,因为序列匹配算法对序列相似度十分敏感,较小的序列数据集可能会产生唯一的计算结果,但是随着其他同源序列或者相似序列的加入,产生的结果很可能就不再唯一,而是包括其他置信度较高的计算结果. 这种现象可以用概率模型中相似概率值计算导致的相似的结果解释. 因此,本文提出了一种系统发育树随机聚类建树方法,可在较短时间内为大规模序列数据的系统发育过程提供所有具有进化意义的解及最优解,以揭示完整发育过程中的序列进化关系.

1 相关工作

序列比对是系统发育研究的主要方法,其本质是字符串的处理和模式识别,如使用贝叶斯概率模型的 MrBayes[7] 和 Beast[8],以及使用最大似然法来进行发育树分析的 GARLI[9] 和 IQPNNI[10-11] 等.

目前基于这些方法,已开发了许多实用的和有

收稿日期:2017-02-19
基金项目:国家自然科学基金资助项目(61402118,61673123),广东省科技计划(2015B090901016,2016B010108007),广东省教育厅项目(粤教高函[2014]97 号,粤教高函[2015]133 号),广州市科技计划(2016201604030034,201508010067,201604046017),江西省教育厅科技研究(GJJ151255)和南昌师范学院基金(15KJZD39)资助项目.
作者简介:张 巍(1964-),女,江西南昌人,教授,主要从事大数据、数据挖掘和协同计算方面的研究. E-mail:weizhang@gdut.edu.cn

效的工具来解决生物学问题,例如 BLAST[12] ,这是一套部署在 NCBI 网站上为在蛋白质和 DNA 分析提供帮助的工具集; PAUP*[13] ,一个用于构建进化树(系统发育树) 和分析这些数据相关性的软件,包含许多分子进化的模型与方法; MEGA[14] ,一套可以自动和手动进行序列比对,构建分子系统进化树,在网络数据库中挖掘信息、估计分子进化的速率和测试进化假说的整合性工具集. 这些工具经过时间检验取得了较大的成功,同时也有许多新的具有其他特性的工具,如 RAxML[15] 是一款基于最大似然法建立进化树的软件,可以处理包括数千种生物和几百万条序列数据在内的大规模的序列数据,以及 ppfold[16] ,一个用于预测 RNA 2 级结构的多线程 Pfold 算法程序.

这些主流的序列比对程序会计算得到一个最优结果,而且可以保证这个结果在众多结果中其置信度最高,但是并不能说明其它结果的置信度低到可以忽略. 因为这些算法忽略了在计算过程中产生的偏差细节,并且有研究者指出这种模式的系统发育分析并不具有进化意义[17] . 同时,这些算法和程序并没有涉及次优解的选择和聚类算法在概率模型中的性能表现. 在分子进化计算中将聚类方法和概率模型相结合仍是一个未知的领域.

为体现进化意义,本文借鉴了上述算法中处理序列间关系的数值表征方法,并结合模糊聚类方法和概率模型来进行系统发育关系分析.

2　模型与算法

2.1　概率模型

本文使用一种改进的概率模型来计算序列数据的系统发育关系,采用概率值作为模型参数. 概率模型具有随机模型的性质,作为参数的概率值可以作为模型的变量,同时在模糊聚类分析中,概率值可以作为分类标准的阈值参与计算. 由于不同的模糊相似矩阵会产生不同的分类结果,即使采用相同的模糊相似矩阵,不同的阈值也会产生不同的分类结果,所以考虑到不同数据规模下的算法伸缩性、对噪声数据的处理能力和对数据输入顺序的不敏感性,使用概率值作为模型参数是非常合适的. 因此本文使用随机概率模型来形式化序列之间的关系,利用概率值来综合表征各参数的数值.

2.2 算法

本文参考动态规划来快速地处理序列并表征为相应数值. 通过使用一个典型的差距罚分和替代矩阵来处理序列数据和进行概率计算,细节参考 Needleman-Wunsch 算法和 Smith-Waterman 算法. Needleman-Wunsch 算法由 S. B. Needleman 和 C. D. Wunsch 提出,已经在研究中被证明是高效可靠的,并具有多个改进的版本[18-20] . Smith-Waterman 算法是类似 Needleman-Wunsch 算法的动态编程方法,但 Smith-Waterman 算法不需要进行全局序列匹配,只需要局部匹配就可以计算输出结果.

目标是构建一个模型来分析序列数据之间的进化关系. 该模型应该满足: 1) 使用概率值来表达序列之间的相似关系; 2) 可以为大规模数据提供最佳的解决方案; 3) 能够适应不同物种间的序列数据,即不确定的数据情况; 4) 保证同源或相似序列数据之间差异的敏感性. 前面章节中提到的一些进化模型已经针对这些方面进行了尝试,与这些模型相比,随机聚类方法在处理不确定的数据类型和大数据规模情况下表现出较好的性能. 随机聚类方法具有明显的马尔可夫性质,这意味着计算过程更侧重于寻找所有可能的解,而不是唯一的解,而在一般情况下,全局最优算法通常需要一个较长的计算时间和较复杂的计算模型,并且不具有启发性. 在不考虑启发式算法的情况下,全局最优算法可以做到更快、更准确,但对于大规模的序列数据或异常的序列,一个可以应对各种情况的综合性算法是必要的. 针对这一点,本文提出了一种改进的随机聚类方法来处理不确定序列数据的系统发育问题,流程图如图 1 所示. 算法的终止条件是聚类的解集呈现出收敛状态. 随机步长因子是随机产生的,用于模拟进化过程中各种影响,并调整系统发育间的关系. 与其他系统发育方法不同的是,本文使用一个解集来记录计算过程中出现的临时最优解,这意味着本文的方法在不同情况下会输出一个或多个解.

本文提出的随机聚类算法中除了概率参数外还有 2 个参数: 1) 经验参数(从大量的实验中分析获得,在陌生环境中可以作为指导参数) ; 2) 阈值参数(随机产生以模拟进化压力的影响,适当的阈值参数的设定可以模拟多序列进化的关系,从而确保模型可以综合考虑进化速率和进化压力的异质性) . 选取 GenBank 数据库中的数据重复进行了大量的实验来模拟数据间的进化过程,以确定这些参数的选择. 具体过程及相应定义如下.

定义 1　若 S 是一个序列,则 $|S|$ 表示 S 中的字符长度,$S[i]$ 表示序列中的第 i 个字符. 如果序列 S 与序列 T 相同,必须满足如下条件: (i) $|S| = |T|$;

(ii) $S'[i] = T'[j]$ $(0 < i \leqslant |S|)$.

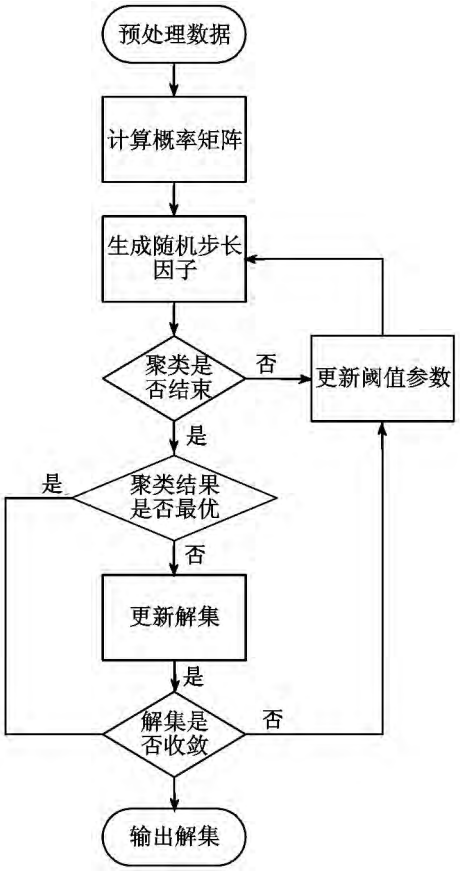

图 1 随机聚类算法的流程图

定义 2 若 x 和 y 是 2 个字符,则 $\sigma(x,y)$ 表示 x 和 y 字符在进行比较时所得的分值,σ 成为一个计分函数,包括当 x 为空字符或 y 为空字符. 在序列中空字符用于表示这个该位置可能缺失一个未知的字符.

定义 3 若 S 和 T 是 2 个序列,S 和 T 的一个相似性比较 A 可以用 S' 和 T' 来表示,其中: (i) $|S'| = |T'|$; (ii) 将 S' 和 T' 中的空字符排除后所得的序列分别和 S、T 相同.

相似性比较 A 就是 S' 和 T' 中字符意义比对,相似性比较 A 的得分 S_{core} 可以表示为

$$S_{core} = \sum \sigma(S'[t], T'[t]), t = |S'| = |T'|. \quad (1)$$

定义 4 对于 2 个序列 S 和 T,它们的最优相似性比较 A 是指在 S 和 T 的所有相似性比较中得分最高的一个,序列相似性比较算法的主要目标是如何寻找序列间的最优相似性比较.

定义 5 如果 2 个序列 S 和 T 均输入某个字符集 Ω,对 Ω 中的任何元素和空符号,他们两两之间均有一个计分值,用计分函数 $\sigma(x,y)$ 表示,$F(i,j)$ 表示序列 S 的前缀 $S[1]S[2]\cdots S[i-1]S[i]$ 和序列 T 的前缀 $T[1]T[2]\cdots T[j-1]T[j]$ 之间的最优相似性比较的得分,则有

$$F(i,j) = \mathrm{MAX}\begin{cases} 0, (i=0 \text{ or } j=0), \\ F(i-1,j-1) = \sigma(S'[i], T'[j]), \\ F(i-1,j) = \sigma(-, j), \\ F(i,j-1) = \sigma(i, -). \end{cases} \quad (2)$$

通过公式(2)可以得到一个如下的得分矩阵,如表 1 所示.

表 1 公式(2)的得分矩阵

$T\backslash S$	1	2	…	$i-1$	i
1	$F(1,1)$	$F(2,1)$			
2	$F(1,2)$	$F(2,2)$			
…					
$j-1$				$F(i-1,j-1)$	$F(i,j-1)$
j				$F(i-1,j)$	$F(i,j)$

通过得分矩阵进行动态规划回溯分析的算法过程为

```
for( i = |S'|, j = |T'|; i > 0 && j > 0; )
{
if( M[i,j] = M[i-1,j-1] + d_eta(S[i], T[j]) )
{
    i--;
    j--;
}
else if( M[i,j] = M[i-1,j] + d_eta(S[i], -) )
{
    i--;
    insert('-', T, j);
}
else if( M[i,j] = M[i,j-1] + d_eta(-, T(j-1)) )
{
    j--;
    insert('-', S', i);
}
}
```

其中 $insert(a,b,c)$ 函数表示在一个序列 b 的第 c 个位置插入一个字符 a.

假设 $|S| = m$, $|T| = n$,则算法复杂度为 $O(m \times n)$,绝大部分计算时间消耗在计算得分矩阵上. 虽然较大的计算量会对计算过程和结果质量产生一定影响,例如对于 2 条长度为 n 的序列,有 $C_{2n}^{n} = (2n)!/((n!) \times 2) \approx 2^{2n}/(n \times \sqrt{\pi})$ 种全局匹配模式,而 n 条序列可以构成 $((2m-3)!!)/(2m-2(m-2)!!)$ 种可能的二叉树拓扑结构,但是其中绝大部分是无意义的. 因此引入随机步长因子来模拟

进化压力的影响和其他可能对进化造成影响的情况,从而尽可能地重复模拟各种进化过程的情况,为各数据集提供一个稳定可信的解集.

随着序列数据规模的增加,同源序列部分的比例呈下降趋势,如图 2 所示. 采用模糊聚类方法的目的就是为了处理这一不确定的冗余情况. 因为当数据规模超过 100 条时,真正有效的参与计算的数据量仅占序列数据的 1%,如何消除或减少剩下的 99% 的冗余数据对计算过程来说至关重要. 模糊聚类方法中适当的参数设置可以有效地降低冗余数据参与计算的可能. 为了增强聚类的启发性来获得更高的置信度,本文采用随机数发生器来生成随机浮点数,利用不同的随机扰动来模拟各种可能的背景环境,从而得到一个完整的和全面的结果. 对于每一个数据集,利用随机聚类计算得到收敛结果的置信度进行排序,然后根据这个次序进行折衷的处理. 此外,对于每一个数据集,本文与经典的系统发育方法: 最大似然法(ML) 和简约法(MP) 进行了比较试验,并使用相同的数据和相同的实验条件. 由于最大似然法通常可以得到置信度较高的计算结果但计算速度较慢,简约法计算速度很快但计算结果置信度相对较低,本文将侧重与最大似然法的结果比较和简约法的速度比较.

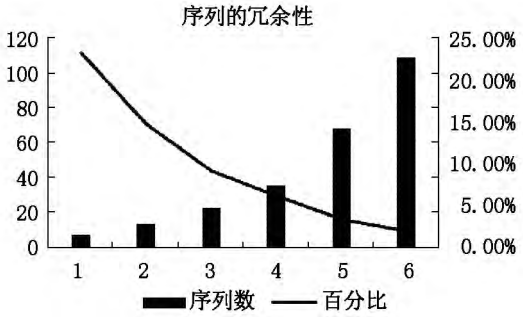

图 2　序列数据的冗余情况

3　实验与性能分析

下面与经典系统发育方法(ML 法和 MP 法) 的性能进行比较. 为了确保公平,实验中采用相同的软件编写 ML 法和 MP 法的代码并满足以下条件: (i) 所有方法的数据都经过相同的预处理; (ii) 采用的 DNA 序列数据长度至少为 4 的倍数; (iii) 空白和未知字符的比例均小于序列长度的 50%. DNA 序列数据的长度范围在 976 到 61 199 之间,蛋白质序列数据的长度在 126 和 22 426 之间.

由于 ML 法和 MP 法是 2 种公认的高性能建树方法,首先使用测试数据集来比较 RCM 法和 ML 法、MP 法的时间性能. 每一组实验均都运行 10 次 RCM 法、ML 法和 MP 法并取计算的时间均值作为比较的标准. 图 3 显示了 RCM 法与 ML 法对比的时间分布(图 3(a) 和图 3(c)) ,RCM 法与 MP 法对比的时间分布(图 3(b) 和图 3(d)) . 在 DNA 和蛋白质数据集中使用 RCM 法均比使用 ML 法更加快速,并可以保持与 MP 法相近的速度. 此外,RCM 法生成的解集中包含更多可能的最优解,经过手工处理后与 ML 法和 MP 法相比会得到更好的结果.

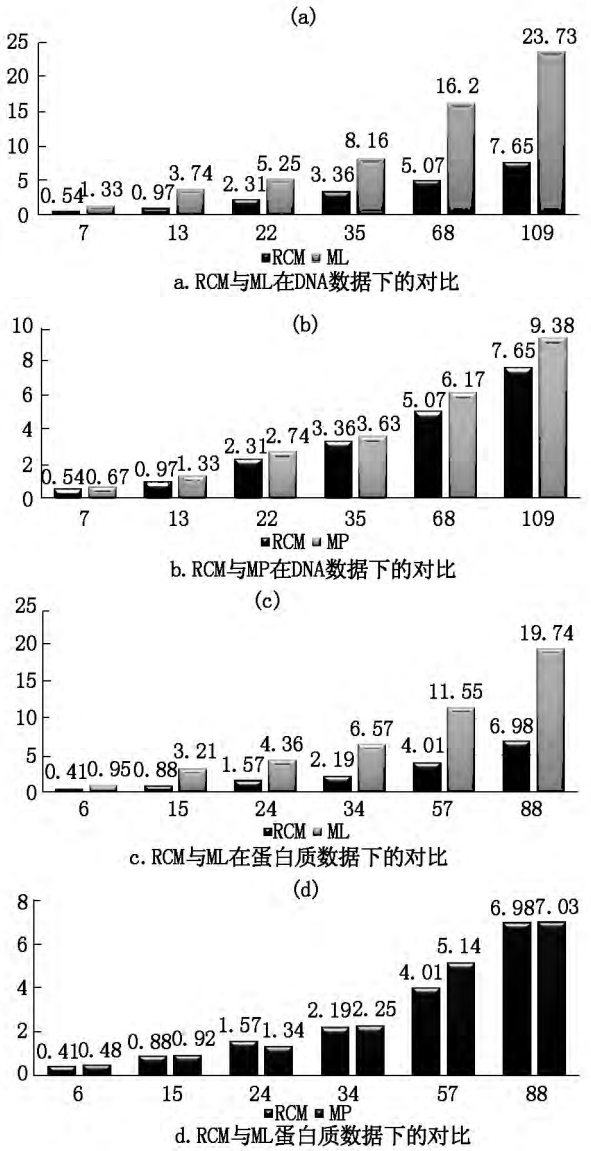

图 3　随机聚类方法(RCM) 在各数据集下的性能

当数据规模超过某个阈值时,输出的解集将可能会包含多个最优结果,如图 4 所示. 不同的建树结果之间的差异是相似序列的概率计算导致的. 所以当 2 个序列非常相似时,容易被分到相同的分支中. 在某些情况下,相似的序列数据会由于相近的概率

值而被分配到不同的分支上,从而导致较大的分歧错误. 如图 5 所示,分歧发生在树的末端可能会产生较小的影响,但当分歧发生靠近树的根部时,就有必要进行手动修改以确保序列分析的正确性.

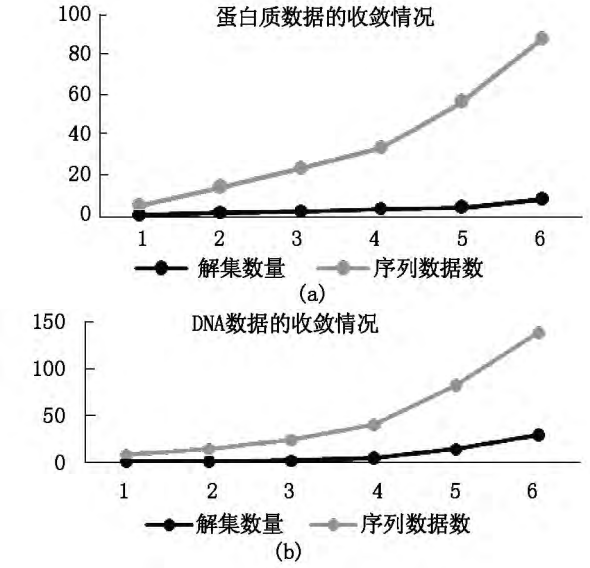

图 4 不同数据规模下的解集收敛情况.

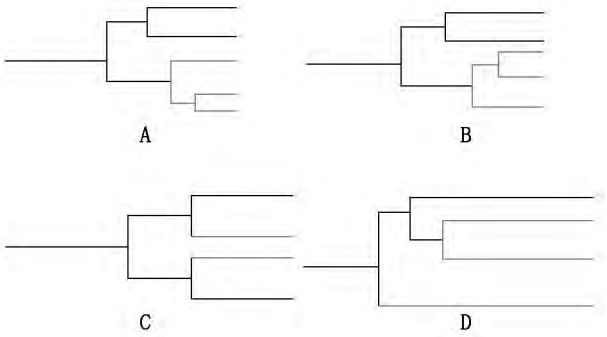

图 5 发育树上不同区域的分歧情况

综上所述,基于各条件和数据集下的对比实验分析,表明了 RCM 法相对于 ML 法具有较高的速度优势,对于 ML 法会得到更完整的解. 由于 RCM 法运行时间取决于聚类规模,本文认为采用这种建树策略的树生成算法更有效. 本文将成熟的系统发育方法和优化技术结合成一个快速和有效的建树算法. RCM 法的实现重点在于 2 个因素: 1) 模糊聚类策略(有助于摆脱冗余计算); 2) 随机步长因子的引入(为建树过程和系统发育分析提供了更多合理的最优解,从而构成输出的解集). 在计算所用数据量相同的条件下,使用 RCM 法的模糊聚类和随机模拟操作与重复使用 ML 法或 MP 法相比会得到更高置信系数的进化树. 如果考虑经验参数和阈值参数的选择与引入,则会进一步提高这种方法的性能. 发生在建树过程中局部最优时,随机扰动有助于增加局部最优解,然后通过聚类操作,可能找到新的和更佳的局部最优解. L. S. Vinh 等[21] 已经讨论了 ML 法中关于随机性和确定性搜索建树的组合可能.

4 结论

本文提出的随机聚类启发式发育方法可以动态地建立解集,避免淘汰掉潜在的最优解. 通过对相近概率引起冲突的处理以及次优解的讨论,可以避免典型的分歧错误,从而生成更可信的进化树. 尽管本文提出的算法并不能解决所有的问题,某些特殊的数据集可能会对算法产生较大的影响(例如用于计算的序列数据过于相似,就有可能会产生一个发散的解集,从而不能得到有效的最优解集),但是在处理正常类型的数据时,随机聚类算法将稳定可靠的产生高质量的结果.

本文提出的算法综合了聚类方法和随机概率模型的特性,可以容易地结合其他的方法扩展成特殊的模型,如考虑到回溯的反复进化过程,或利用超级计算机针对大数据建立完整的系统分析等. 在下一步工作中,将增加系统发育评分功能完善随机聚类算法,并通过研究隔代序列信息的功能可以分析序列后代和序列之间的亲缘关系,对可能的位点进行匹配,使随机聚类过程中序列向更高置信度的匹配结果收敛.

5 参考文献

[1] Howe K, Bateman A, Durbin R. QuickTree: building huge neighbour-joining trees of protein sequences [J]. Bioinformatics, 2002, 18(11): 1546-1547.

[2] Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population [J]. Molecular biology and evolution, 1995, 12(5): 921-927.

[3] Kolaczkowski B, Thornton J W. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous [J]. Nature, 2004, 431(7011): 980-984.

[4] Seo T K, Kishino H. Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences [J]. Systematic biology, 2009, 58(2): 199-210.

[5] 高灵渲,张巍,霍颖翔,等. 改进的聚类模式过滤推荐算法 [J]. 江西师范大学学报: 自然科学版, 2012, 36(1): 106-110.

[6] 韩娜,滕少华,房小兆. 基于哈达玛变换的多元时间序列聚类研究 [J]. 计算机工程与设计, 2012, 33(3):

983-986,1021.

[7] Ronquist F,Teslenko M,van der Mark P,et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space [J]. Systematic biology,2012,61(3):539-542.

[8] Drummond A J,Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees [J]. BMC evolutionary biology,2007,7(1):214.

[9] Zwickl D J. GARLI: genetic algorithm for rapid likelihood inference [EB/OL]. [2016-12-27]. See http://www.bio.utexas.edu/faculty/antisense/garli/Garli.html,2006.

[10] Minh B Q,Vinh L S,Von Haeseler A,et al. pIQPNNI: parallel reconstruction of large maximum likelihood phylogenies [J]. Bioinformatics,2005,21(19):3794-3796.

[11] Schmidt H A,von Haeseler A. Phylogenetic inference using maximum likelihood methods [J]. The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing,2009(2):512-522.

[12] Altschul S F,Gish W,Miller W,et al. Basic local alignment search tool [J]. Journal of Molecular Biology,1990,215(3):403-410.

[13] Swofford D L,Documentation B. Phylogenetic analysis using parsimony [M]. IL: Illinois Natural History Survey,Champaign,IL,1991.

[14] Tamura K,Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees [J]. Molecular biology and evolution,1993,10(3):512-526.

[15] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies [J]. Bioinformatics,2014,30(9):1312-1313.

[16] Sükösd Z,Knudsen B,Kjems J,et al. PPfold 3.0: fast RNA secondary structure prediction using phylogeny and auxiliary data [J]. Bioinformatics,2012,28(20):2691-2692.

[17] Löytynoja A,Vilella A J,Goldman N. Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm [J]. Bioinformatics,2012,28(13):1684-1691.

[18] Markova-Raina P,Petrov D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes [J]. Genome research,2011,21(6):863-874.

[19] Pruesse E,Peplies J,Glöckner F O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes [J]. Bioinformatics,2012,28(14):1823-1829.

[20] Güyer T,Atasoy B,Somyürek S. Measuring disorientation based on the Needleman-Wunsch algorithm [J]. The International Review of Research in Open and Distributed Learning,2015,16(2):316-322.

[21] Vinh L S,von Haeseler A. IQPNNI: moving fast through tree space and stopping in time [J]. Molecular biology and evolution,2004,21(8):1565-1571.

The Sequence Analysis Method Based on Random Clustering Model

ZHANG Wei[1],WANG Yang[1],LIU Dongning[1],TENG Shaohua[1],ZHANG Li[2],XU Xinai[2]

(1. School of Computer Science and Technology,Guangdong University of Technology,Guangzhou Guangdong 510006,China;
2. Nanchang Normal University,Nanchang Jiangxi 330032,China)

Abstract: Large phylogeny estimation is a combinatorial optimization problem that no future computer will ever be able to solve exactly in practical computing time. Here,a tree constructing approach has been reported,the random clustering method,involving several pruning of trees that are used to provide the optimal solution and near-optimal solution with evolutionary significances,to reveal the complete evolutionary relationships based on basis of previous studies. The experiments show the correctness and efficiency of our method,and the significances to biocomputing and phylogenetic analysis.

Key words: random clustering algorithm; sequence analysis; phylogeny

(责任编辑: 冉小晓)