

系统发育网络的构建与应用

程春花, 黄 原

(陕西师范大学生命科学学院, 陕西 西安 710062)

摘要: 系统发育网络目前主要应用于表达种上复杂的网状进化关系、种内个体及群体之间的关系以及相互矛盾数据集的系统发育分析结果 3 个方面。本文在综述网状进化现象的基础上, 介绍了目前用于构建系统发育网络的几种常用方法, 包括 T-Rex 软件包的网状图法, Tcs 软件中的统计简约法, Splitstrees4 的中间网法、二分裂法和邻接网法。

关键词: 网状进化; 系统发育网络; 网状图; 统计简约法; 中间网法; 二分裂法; 邻接网法

中图分类号: Q969 **文献标识码:** A

文章编号: 1000-7482(2008)03-0215-07

系统发育 (phylogeny) 或系统树 (phylogenetic tree) 用于表示任何分类单元的进化历史。目前大多数系统发育分析方法都假定生物类群是以树状分歧方式进化的, 即系统树是 1 个二岐 (bifurcating) 分支的树。但随着系统发育分析研究的深入和研究范围的扩大, 树状图显示出不能准确表达实际进化过程的情况, 如杂交起源的物种、基因水平转移、群体内的基因重组等。这些进化过程产生非树状的进化史, 而且, 即使实际的进化是树状的, 冲突或模糊信号的存在也使系统树不能合理地解释这些现象。因此, 近年来人们提出了构建可以表达性状冲突或网状进化过程的系统发育网络 (phylogenetic network), 以表示多个相关系统树的综合, 可以同时代表多个系统树, 暗示数据包含树状进化的程度, 并为可能的网状进化事件提供证据^[1]。

1 网状进化形式与机制

网状进化 (reticulate evolution) 是指进化过程不能充分地由树状进化模型表示的 1 种进化模型。网状进化表现出谱系之间缺乏独立性, 当 1 个网状进化事件发生时, 2 个或更多的独立进化谱系在一定水平相互联系起来。

网状进化关系在自然界普遍存在, 主要包括下述进化现象:

1.1 基因横向转移

基因横向转移 (gene lateral transfer) 是遗传物质从 1 个谱系向另 1 个谱系的直接转移, 在所有生物类群中都存在, 但在原核生物中最普遍。古细菌和真细菌可以通过基因横向转移获得新基因而提高适应新环境的能力。基因横向转移的主要机制是转化、接合和转导。

1.2 杂交

杂交物种的形成是另一种典型的网状进化。如果新种的染色体数与亲本一样, 这个过程为二倍体杂交, 如果新种的染色体数是双亲染色体的总数时, 称为多倍体杂交。杂交的主要机制是同

收稿日期: 2008-03-22

基金项目: 国家自然科学基金项目 (20670279, 30470238)

源多倍体杂交、异源多倍体杂交和双倍体杂交。通常同源多倍体并不引起网状事件,而植物中较常见的异源多倍体是网状杂交事件的主要来源^[3]。

1.3 遗传重组

重组涉及导致遗传物质新组合的所有过程,如种内地方群体间的微进化过程,包括异域群体间的遗传分化、由迁移引起的基因流等。重组产生谱系内的网状进化,影响多个不同水平的进化,并且是有性生殖自然群体内大量遗传多样性的来源^[4]。重组导致嵌套基因,使得基因的不同区域也许有不同的系统发育史,在系统发育重建过程中考虑重组是非常重要的,多种算法可用于重组的检测^[5, 6, 8]。

另外,包含宿主转移的寄生虫-宿主关系及替代生物地理 (vicariance) 和扩散生物地理过程,是非系统发育研究领域的网状关系^[2]。

2 构建系统发育网络的方法

网状进化关系包含闭合环 (cycle), 无法用树状图表示。网状图 (reticulogram) 或网络 (network) 是能够表现有机体相互之间有 1 条以上连通路径关系的图。Sneath (1975) 总结了当时已知的有关网状进化的生物学证据,并首次提出用改进的分支图来表示网状进化。但此后近二十几年里,有关网状进化关系重建方法的研究较少,只是近几年才又受到重视,“分类杂志” (Journal of Classification) 在 2000 年有 1 期刊登 6 篇论文讨论网状图的构建及其在进化上的应用^[2]。

目前用于构建系统发育网络的大多数方法类似于构建系统树的简约法和距离法。对于简单性状组成的数据集,大多数系统发育网络方法将产生大致一样的结果,但是随着性状复杂性的增加,这些方法形成的系统发育网络将有较大的差异。以下简单介绍一些较常用的构建方法的原理及其相应的软件。

2.1 网状图法

这是一类基于距离法的网络构建方法,最早由 Legendre 和 Makarenkov 提出^[2, 7]。网状图构建算法是:首先通过距离矩阵推断出 1 个系统发育树,然后将网状分支 1 次 1 个逐渐添加到系统树上,添加网状分支是基于最小化的 1 个最小二乘丢失函数或是 1 个加权的最小二乘丢失函数 (minimizing a least-squares or a weighted least-squares loss function),当到达 1 个拟合优度标准 (goodness of fit criteria) 的最小值时,网状分支的添加终止,有 4 个这样的拟合准则, Q1、Q2、AIC 和 MDL,每一个标准考虑了最小二乘丢失函数和网状图的分支总数,这个算法需要 $O(kn^4)$ 次以添加 k 个网状分支^[2, 7]。这个方法包含在软件包 T-Rex 中且已用于多个生物学问题,如杂交、水平基因转移和生物地理网等^[2]。

2.2 统计简约法

该法根据性状状态差异的大小依次地连接相应的分类单元,直到达到简约上限 (parsimony limit)^[5, 8, 9]或是已经连接了所有的单倍型 (haplotype),这是简约法作为一个可靠系统发育分析法的限度,默认值是 95%。这个方法可用于假定存在的重组和同型性状的鉴定^[8],性状冲突的表达类似简约树 (如每个分支代表 1 个特定的性状状态变化),但是由于某些可能存在的性状变化在网上没有表示出来,所以分析是不完全的。同样,生物学的解释也类似简约树,祖先节点可清楚地推断出来,而且,群体遗传中的溯祖理论 (coalescent theory) 允许无外类群的系统树根的推断^[9]。此方法适合较简单的性状变化,随着性状复杂性的增加,这个方法将产生几个不联系的网而不是 1 个单一网,这意味着将有很多错误的负值^[10] (FN, 即存在于构建的系统发育网络上的

分支, 而不存在于用同样的数据构建的系统树上) 的存在。此方法包含在 Tcs 程序中。

2.3 中间网

这是 1 个基于性状的用于二态数据的方法, 由 Bandelt 等人提出, 目标是同时表达分类单元间的所有性状状态差异, 这样可以保证包含所有的最简约树。中间网^[12]的算法是: 序列首先被转化成由二态性状 (binary character) 组成的矩阵, 二态性状采取 0 和 1 值, 进而确定哪些性状是相容的, 或是不相容的。相容的所有性状形成 1 个小组 (clique), 进一步形成系统发育树; 不相容的性状形成另 1 个反小组 (anticlique), 是形成中间网空间维数的根源。随着所研究序列的增多, 不相容位点也会增多, 必然会使网的空间维数增加而不易于分析。中间网法衍生出了许多不同方法, 包括简化中间网法 (reduced median network), 中间连接网 (median-joining network), 贪婪简化中间网 (greedy reduced median network)^[9]和一致网法 (consensus network)^[11]等。

2.4 二分裂法

这是一个可以将性状变化以距离或简约原则转化成一系列微相容组合的方法, 目标是将所研究分类单元的所有可能的相容组合表现在系统发育网络上, 最早由 Bandelt 和 Dress 提出。

这个算法并不是选择距离之和最小的拓扑结构为最优系统发育, 而是每一步排除最不可能的那个结构, 所形成的网状图称为二分裂图 (splits graph), 包含分类单元的所有可能拓扑结构, 正因为这个原因, 所以此图包含由分裂边 (splits) 组成的一系列等长平行边, 这些平行边的长度代表给定种群某一组合的隔离指数 (isolation index)^[13]。同样地, 针对较简单的性状冲突的分析结果较好, 随着性状冲突复杂性的增加, 将产生无信息的网状分支。因此, 1 个多歧分支 (multifurcating) 究竟是代表不充足的信息或是过多的冲突信息, 是无法从所建的系统发育网络上得到解释的。当性状冲突的复杂性降低时, 此法类似中间网法^[13, 14]。

2.5 邻接网法

邻接网法 (NeighborNet) 是 1 个基于距离法的方法^[1, 15], 由 Bryant 和 Moulton 提出, 可用于较复杂性状冲突的表达, 且随着性状复杂性的增加分析效果较好, 这也使得这个方法成为分析复杂进化模式的最好选择。1 个系统发育网络至多表达数据中的所有性状冲突, 这对任一网络构建方法来说都是至关重要的, 因为这是避免错误推断所必不可少的, 也是系统发育网络用于统计检验需要的, 邻接网法符合这个特性, 如果在邻接网法中输入的是网络距离 (reticulate distance) 参数, 那么将产生 1 个精确代表这个距离的网络, 如果输入叠加距离 (additive distance), 那么将产生 1 个系统树, 即邻接网法是统计上一致的方法^[1]。目前还有一种类似邻接网法的 Qnet 法, 通过计算数据中的一系列环状权重的分裂边而构建 1 个系统发育网络, 如果距离数据不充足或是基于性状数据时可以采用这个方法^[16]。

以上是目前常用的网络构建方法, 网中的多歧分支表达了所研究数据的性状冲突, 包含 3 种含义: ①代表数据中的不确定性, 如抽样误差和系统误差。②代表进化中的相似事件, 通常是由趋同, 平行和回复突变形成的同型性状。③代表系统发育史上的同源事件, 如重组、杂交和水平基因转移^[9]。但是如何区分这 3 种多歧分支还没有确定的方法, 所以这些方法构建的网络只能称为代表性状的系统发育网络, 或是二分裂网络^[17, 18], 如上述的中间网法, 二分裂法和邻接网法构建的系统发育网络。相应地, 能够精确表达真实进化史的网络称为网状进化网 (reticulate networks)^[17, 18], 如杂交和重组网, 这些都包含在软件包 Splitstree4 中。

3 系统发育网络的应用

3.1 表达种上网状进化关系

杂交和水平基因转移是网状进化的主要机制，它们属于谱系间的网状进化事件，也是物种多样性的主要来源。但是所有的系统发育树都有 1 个本质的缺点，仅能处理垂直遗传。进化生物学家普遍认为大多数微生物的进化史是镶嵌的，存在大量的水平遗传，只有通过 1 个合适的网络模型，才能同时处理垂直和水平遗传，以便更好的理解微生物的生物特性和动态变化^[19]。同样地，植物中普遍存在的杂交现象也只能通过合适的系统发育网络来表达^[20]，如图 1 示（由 supernetwork 法构建的十字花科多个种的杂交网），这种普遍存在于植物世界里的异种杂交形成新种的现象在数量繁多、新种不断涌现的昆虫世界里是否也存在呢？2006 年 6 月 15 日英国《自然》杂志报道了 1 例自然杂交现象，即 *Heliconius heurippa* 是 *H. melpomene* 和 *H. cydno* 在自然状态下杂交所形成，并通过实验得到验证，但此现象在昆虫世界里毕竟还是罕见的^[21]。

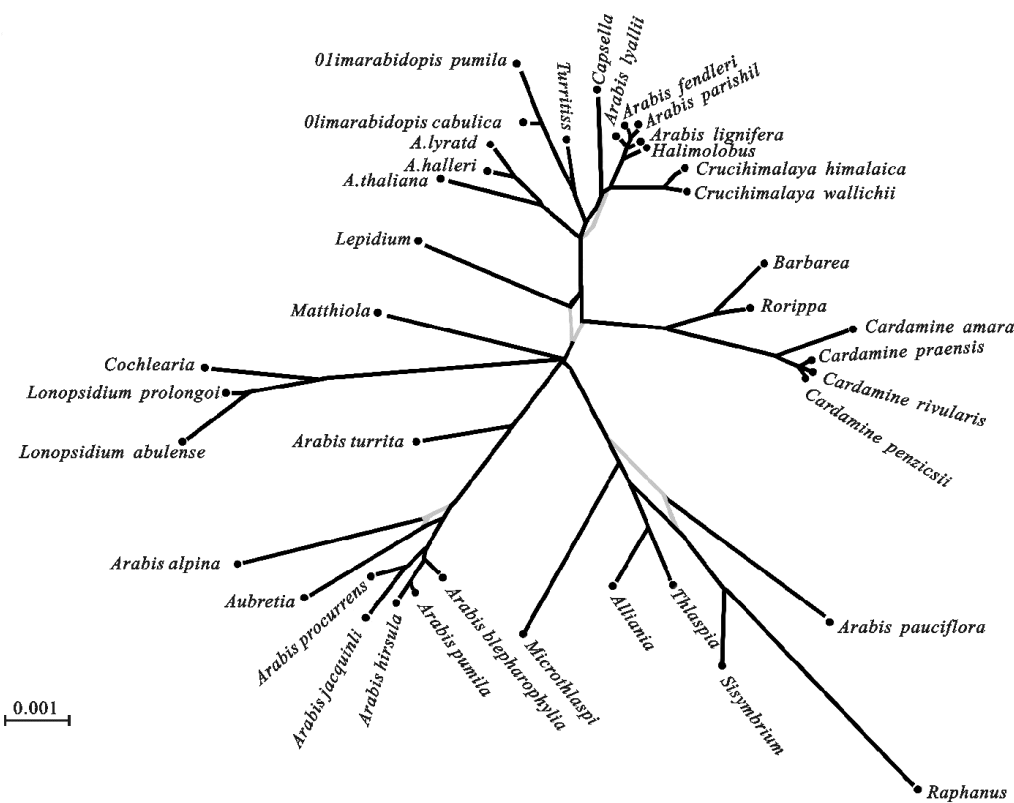


图 1 十字花科多个种的杂交网图，引自文献 20，图中的淡色环表示推测的杂交事件
(The hybridization network of the brassicaceae constructed by supernetwork, coming from reference 20, the light circles represent possible hybridization events)

3.2 表达种内个体及群体之间的关系

系统发育网络可用于由重组引起的谱系内微进化的分析。这类进化事件发生在有性生殖群体内，因为形成配子过程中，双亲贡献给后代的基因组各自发生了不同的减数分裂重组，导致后代来自两个亲本的基因组具有不同的进化史，也就是所谓的性重组 (sexual recombination)，从而

导致群体水平的微进化。种内个体的分歧度较种间个体低，包含较少的性状变化，因此，在进行种内群体之间系统发育分析时，降低了传统系统发育分析法的能力^[22]；而且起源于重组的网状进化事件是不能用系统树来说明的，只能用网状图表达^[10, 18, 23]，图 2 示就是表示 1 个群体 9 个单倍型之间进化关系的网络（a-i 代表黑腹果蝇乙醇脱氢酶的 9 个单倍型，该重组网的数据集来自这 9 个单型组成的二态性状，由软件包 *splitstree4* 中的重组网法构建，具体内容参考文献 18）。

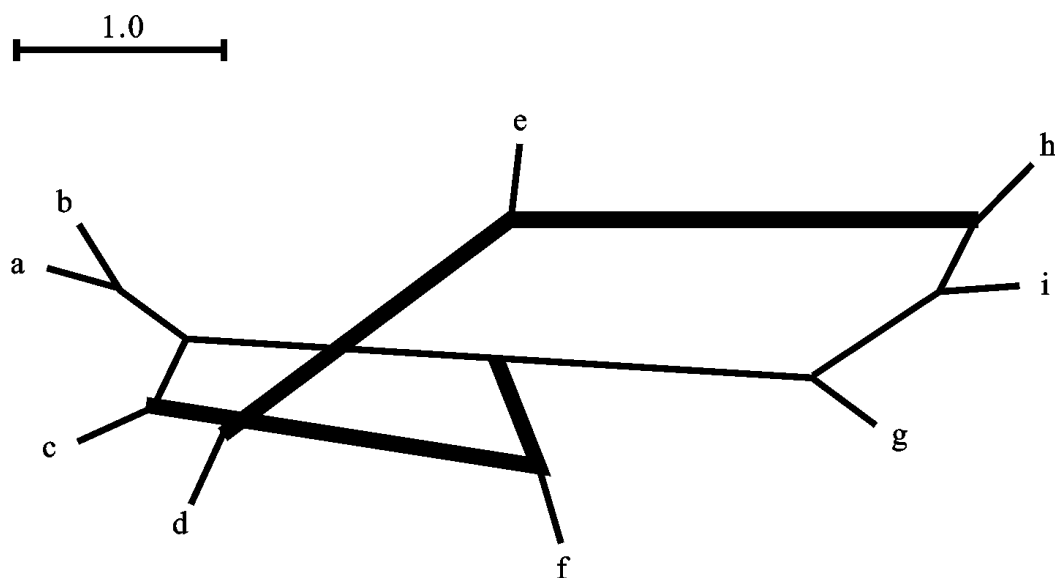


图 2 黑腹果蝇乙醇脱氢酶位点 9 个单倍型的重组网，深蓝色（加粗）分支代表重组事件
(The recombination network of 9 haplotypes of the alcohol dehydrogenase locus from *Drosophila melanogaster*, the dark blue branches represent recombination events)

3.3 表达相互矛盾的数据集的系统发育分析结果

位点和谱系间进化速率的变异是产生错误系统发育推断的重要原因之一，如 Kolaczowski 和 Thornton 在分析变异进化模型时，提出简约法优于似然值法。Huson 和 Bryant 重复了他们的实验，不过采用的方法还包含 1 种网络构建法（二分裂法），结果证实网络构建法在分析矛盾数据时明显优于简约法和似然法^[17]。

4 系统发育网络的意义

进化是生物学中的核心问题，系统发育网络用于研究复杂的进化关系，这是物种多样性等复杂问题的根源，随着网络构建法的改进，系统发育网络必然会成为生物学各个领域的重要分析工具，因为系统发育网络可以对已知现象提供可能的证据且为探索未知现象提供可能的线索。

昆虫纲是整个动物界中最大的 1 个类群。地球上有 150 余万种动物，昆虫有 100 多万种，占动物总数的 4/5，每年还陆续发现约 0.5~1.0 万新种。如此多样的类群潜在着怎样的生命秘密或是生命启示呢？通过系统发育网络的方法也许或多或少可以提供一些有用的参考价值吧。

参考文献

- [1] Bryant D, Moulton V, Spillner A. Consistency of the Neighbor-Net Algorithm[J]. *Algorithms Mol. Biol.*, 2007, 2: 8.
- [2] Legendre P, Makarenkov V. Reconstruction of biogeographic and evolutionary networks using reticulograms[J]. *Syst. Biol.*, 2002, 51(2): 199-216.
- [3] Linder C R, Moret B M, Nakhleh L, *et al.* Network (reticulate) evolution: biology, models, and algorithms[C]. A tutorial presented at the Ninth Pacific Symposium on Biocomputing (PSB2004).
- [4] Pérez-Losada M, Porter M L, Tazi L, *et al.* New methods for inferring population dynamics from microbial sequences[J]. *Infect. Genet. Evol.*, 2007, 7(1): 24-43.
- [5] Cassens I, Mardulyn P, Milinkovitch M C. Evaluating intraspecific "network" construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach[J]. *Syst. Biol.*, 2005, 54(3): 363-372.
- [6] Posada D, Crandall K A. Evolution of methods for detecting recombination from DNA sequences: Computer simulations[J]. *Proc. Natl. Acad. Sci. USA.*, 2001, 98(24): 13757-13762.
- [7] Makarenkov V, Legendre P. From a phylogenetic tree to a reticulated network[J]. *J. Comput. Biol.*, 2004, 11(1): 195-212.
- [8] Teiji S, Sasabe M. Utility of nuclear allele networks for the analysis of closely related species in the genus *Carabus*, subgenus *Ohomopterus*[J]. *Syst. Biol.*, 2006, 55(2): 329-344.
- [9] Morrison D A. Networks in phylogenetic analysis: new tools for population biology[J]. *Int. J. Parasitol.*, 2005, 35(5): 567-582.
- [10] Nakhleh L, Sun J, Warnow T, *et al.* Towards the development of computational tools for evaluating phylogenetic network reconstruction methods[J]. *Pac. Symp. Biocomput.*, 2003, 315-326.
- [11] Holland B, Delsuc F, Moulton V. Visualizing conflicting evolutionary hypotheses in large collections of trees: using consensus networks to study the origins of placentals and hexapods[J]. *Syst. Biol.*, 2005, 54(1): 66-76.
- [12] Bandelt H J, Macaulay V, Richards M. Median networks: speedy construction and greedy reduction, one simulation and two case studies from human mtDNA[J]. *Mol. Phylogenet. Evo.*, 2000, 16: 8-28.
- [13] Winkworth R, Bryant D, Lockhart P, *et al.* Biogeographic interpretation of splits graphs: least squares optimization of branch lengths[J]. *Syst. Biol.*, 2005, 54(1): 56-65.
- [14] Dress A W, Huson D H. Constructing splits graphs[J]. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 2004, 1(3): 109-115.
- [15] Bryant D, Moulton V. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks[J]. *Mol. Bio. Evol.*, 2004, 21: 255-265.
- [16] Grünwald S, Forslund K, Dress A, *et al.* QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets[J]. *Mol. Biol. Evol.*, 2007, 24(2): 532-538.
- [17] Huson D H, Bryant D. Application of phylogenetic networks in evolutionary studies[J]. *Mol. Biol. Evol.*, 2006, 23(2): 254-267.
- [18] Huson D H, Kloepper T H. Computing recombination networks from binary sequences[J]. *Bioinformatics*, 2005, 21 (Suppl 2): ii159-165.
- [19] Kunin V, Goldovsky L, Darzentas N, *et al.* The net of life: Reconstructing the microbial phylogenetic network[J]. *Genome Res.*, 2005, 15(7): 954-959.
- [20] McBreen K, Lockhart P J. Reconstructing reticulate evolutionary histories of plants[J]. *Trends Plant Sci.*, 2006, 11(8): 398-404.
- [21] Mavárez J, Salazar C A, Bermingham E, *et al.* Speciation by hybridization in *Heliconius* butterflies[J]. *Nature*. 2006, 441(7095): 868-871.
- [22] Posada D and Crandall K A. Intraspecific gene genealogies: trees grafting into networks[J]. *Trends. Ecol. Evol.*, 2001, 16(1): 37-45.
- [23] Gusfield D, Eddhu S, Langley C. Optimal efficient reconstruction of phylogenetic networks with constrained recombination[J]. *J. Bioinform. Comput. Biol.*, 2004, 2(1): 173-213.

Construction and Application of Phylogenetic Network

CHENG Chun-hua, HUANG Yuan

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

The phylogenetic networks mainly applied to display complicated reticulations above species, display the relationships between intraspecific individuals and among populations, and display the results of phylogenetic inference of contradicted data sets. After having explained the background knowledge of phylogenetic network and reticulate evolution, we introduced several methods applying to construct phylogenetic networks in widely used softwares, such as Reticulograms in T-Rex, Statistical parsimony in Tcs, Median networks, Split decomposition and Neighbour-net in Splitstrees4.

Key words: Reticulation; Phylogenetic networks; Reticulograms; Statistical parsimony; Median networks; Split decomposition; Neighbour-net