等位基因频率资料的分支分类方法*

周厚高[1] 朱军[2]
(1 广西大学农学院，南宁 530005；2 浙江大学，杭州 310029)

摘要 等位基因频率资料是群体遗传学的重要资料，也是进化生物学的重要资料。本文讨论了利用此类资料重构生物类群系统发育的分支分类方法。常用的分支分类方法为俭约性方法和统计分支分类方法。俭约性方法中主要分为三类：数据转换方法、距离叠加树方法和频率空间重建系统发育的方法。同时对不同方法的特点和不足作了讨论。

关键词 等位基因频率资料；分支分类方法；系统发育
中图分类号 Q941.2

The Study of Cladistic Methods for Allele Frequency Data

Zhou Hougao[1] Zhu Jun[2]
(1 Agric. College, Guangxi Univ., Nanning 530005; 2 Zhejiang Univ., Hangzhou 310029)

Abstract Allele frequency data were very important for study of population genetics and evolutionary biology. There were many procedures had been proposed for reconstruction of phylogeny with this kind of data. The cladistic methods for allele frequency data were outlined and discussed in present paper. Parsimonious and statistic methods of cladistics were frequently used for allele frequency data. Parsimonious methods were composed by the three groups i. e data transformation, additive tree and reconstruction of phylogenetic of phylogeny in frequency space. The characteristics and limitations of the cladistic methods for allele frequency data were discussed in present paper.

Key words allele frequency data; cladistic method; phylogeny

等位基因频率资料（如等位基因酶、分子标记频率资料）是研究群体遗传学的主要资料来源，关于植物群体的遗传结构、遗传变异的证据主要是通过蛋白质电泳技术而来的。几十年的努力已经积累了大量的宝贵资料，是探讨生物类群系统发育、进化生物学的重要财富。进化生物学与群体遗传学的结合是系统发育研究的必然，在分子水平探讨进化的机理和系统发育重构是今后的发展方向。

利用等位基因频率资料重构系统发育的方法，应用最多的是数量分类学的聚类分析方法，其次是分支分类方法。本文将对应用于等位基因频率资料重构系统发育的分支分类方法作简要综述。

分支分类方法目前分为四类（俭约法、统计法、信息法、和谐法）[1]，其中俭约法和统计法在等位基因频率资料中使用较多，而后二类方法极少见使用。因此本文讨论前两类方法。

1 俭约分支分类方法在等位基因频率资料中的应用

俭约性方法应用于等位基因频率资料，必然与流行的 Wagner 方法[2,3]相一致，在逻辑上存在进化转换（步长）最小化和解释资料中同塑（平行、趋同、逆转）假设数最小化之间的对应。当等位基因以连续度量的频率形式出现时，为达到上述对应存在着相当多的困难。为克服这个困难在研究过程中发展了多种处理方法，常用的主要有三种：①数据变换，抛弃频率信息，转换为离散数据[4~6]；

* 国家自然科学基金资助项目（编号 39160010，39760010）。
第一作者：男，1962年生，博士，教授。
收稿日期：1999-10-10

②距离叠加树法或独立等位基因模型，将等位基因作为独立性状，以频率直接作俭约性分析[7,8]或构造距离最小的叠加树，而不考虑同一位点的多个等位基因频率之和为 1 的问题，只求等位基因频率变化最俭约或距离变化最俭约；③在等位基因频率空间重构系统发育树，以距离系数为依据，求解祖先类群等位基因频率分布，寻求距离变化的俭约性[9~12]。

1.1　数据变换方法

群体基因频率常受许多因素如遗传漂变、生境变化等的影响，经常处于变动状态，因此许多学者认为频率提供的信息不多，宁愿抛弃频率信息[13]，将等位基因频率数据离散化，形成缺失/存在的 0/1 编码。利用变换后的数据进行系统发育关系重构的方法主要有 3 种。

(1)直接利用 Wagner－Farris 方法进行分析，简便易行，有较多的软件可供利用。Swofford 等[10]认为此法面临着取样误差的严重错误，不能充分利用由电泳技术和血清技术揭示出的多态现象的宝贵财富。用 Wagner－Farris 方法直接分析 0/1 编码数据，经常造成祖先类群的某些位点无任何等位基因存在的现象。

(2) 转换系列分析技术 (TSA) (transformation series analysis)。由 Mickevich[13]发展并应用于研究实践[14,15]，试图通过迭代技术同时估计性状状态树和分支树。

Mickevich[13]的特征状态转换系列分析 (TSA) 并不是研究等位基因的系统发育即等位基因的进化树，而是将一个位点作为一个性状，而将该位点等位基因的不同组合作为性状状态。为此 Mickevich 和 Mitter[14~15]提出了两个方法：共有等位基因模型(shared allele model)和最小转换模型(minimum turnover model) 来排序等位基因组合。

Swofford 和 Berlocher[10]认为 TSA 在概念上和数学方法上都是失败的。首先，TSA 不能直接利用基因的频率数据，损失了大量系统发育信息；其次，TSA 在实践中也存在困难：增加抽样样本的大小，等位基因组合将越来越多以致 TSA 分析将失去其提供进化信息的能力。

(3) 质量亨利希分析 (qualitative Hennigian analysis) 方法[6]。采用 Hennig[16]的常规方法处理等位基因频率资料，引入外群对性状作极性分析，用同远态进行归类。该方法的第一个缺点是[10]将等位基因频率资料转换为 0/1 编码，受抽样误差影响，将形成误解的分类结果；第二，外群比较并不能揭示出等位基因的原始性与衍生性，因而不能正确的使用等位基因在系统发育中的重要意义；第三，存在最普遍即为原始的外群判别原则应用于等位基因的性状极性分析，从群体遗传学的观点看，是不确切的[17]。因此在等位基因频率资料中引入外群分析是不宜的。

1.2　距离叠加树方法

这是一类通过拟合分支长度于遗传距离矩阵的构建系统发育树的方法。叠加树方法采用的大部分相似性系数和成聚方法同于数量分类学所采用的，不同之处在于二者的前提假设和由这些研究结果引出的结论[18]。叠加树方法的应用范围可以分为三类[19]：第一类是将叠加距离矩阵转换为超度量距离 (ultrametric distance) 矩阵，然后应用聚类分析获得树状图，如当前祖先法 (present－day ancestor method)[20]和邻接法 (neighbor－joining method)[21]；第二类是应用对 4－类群树的叠加定义，将具有最大残差 (largest fraction) 一致性的类归并在一起的方法，其中有 Sattath 和 Tversky[22]以及 Fitch[23]的方法；第三类为 Wagner 方法[24]，在实际的应用中，许多研究人员采用此法处理连续性状资料。许多定量俭约法程序如 Hennig－78、Hennig－86，能从连续变量数据获得 Wagner 树[25]。这种算法首先把数据矩阵转换成曼哈顿距离矩阵，然后由此构建一个主网络，优化主网络，产生最俭约的分支树[26,27]。根据曼哈顿距离的性质，这种距离俭约性事实上是频率变化俭约性。

Farris[5]对这类通过拟合分支长度于遗传距离矩阵构建系统发育树的叠加树法提出了批评：(1)应用最广的距离度量 (Nei 氏距离[28]) 是非度量性的，歪曲了三角不等原理，因此不能看作是进化路径长度的度量，由此推导出的进化最小结果是不可靠的；(2) 即使是采用了度量性的系数 (如 Rogers 距离[11]、曼哈顿距离等)，所形成的结果不能在等位基因频空间构建系统发育树，不能反映等位基因组成和频率分布的动态变化；(3)将分类群对偶的性状资料浓缩为一个距离值导致了信息的丢失 (information loss)，特别是有关特定等位基因在类群间的分布信息的丢失。

部分叠加树方法是建立在独立等位基因模型（independent allele model）之上的，由距离最俭约引出进化最俭约的一类方法，用分枝的长度拟合遗传距离矩阵。独立等位基因模型是一种对等位基因频率资料处理方法，该法将每个等位基因作为一个性状，以其频率作为观察值，进行距离计算[29,30]。这种距离转换可能在欧氏空间进行，也可以通过曼哈顿距离转换，直接比较基因频率的变换，还可以通过其它遗传距离进行转换。

独立等位基因模型的最大弱点就是将等位基因频率变化作为一个独立的量来利用。由于等位基因频率至少与其同位点的其它等位基因相依赖，很明显不满足独立性的原则，而是满足同位点所有等位基因频率和为1的原则（additivity requirement），因此许多研究者是以基因位点作为研究的性状[11,14]。独立等位基因模型的另一个弱点是：当一个位点具有两个或两个以上等位基因时，祖先类群（HTU）位点上的等位基因频率之和不为1，也即正如Farris[5]指出的不能在等位基因频率空间重构进化树。优化处理时，很多情况下将产生HTU某个位点不存在等位基因的情况，因此其分支长度缺乏进化的意义[14,15]。

1.3 等位基因频率空间构建系统发育的方法

针对数据转换和叠加树方法的不足，最近从不同的途径、采用不同的距离系数，发展了在等位基因频率空间重构系统发育的新方法[19,20]，是等位基因频率资料重构系统发育俭约性方法研究的新发展。

这类方法的主要特点是：①对于等位基因频率资料，反对0/1编码，充分利用频率信息；②以位点作为性状；③在等位基因频率空间构建系统发育树，为每个假设祖先类群（HTU）确定等位基因组成和频率，位点等位基因频率和为1；④优化HTU等位基因频率来拟合分枝长度以达到进化长度最俭约的目的；⑤均形成无根树；⑥两者的计算均十分复杂。

这类方法仅有两个，一个是Rogers[9~11]提出的，称为双曲线拟合法（hyperboloid approximation procedure，HAP），另一个是由Swofford和Berlocher[12]发展的，称为线性规划法。二者的区别在于：①双曲线拟合法采用Rogers遗传距离系数，同时可以推广到其它欧氏距离、非欧氏距离（如Cavalli-Sforza和Edwords的弧距离[7]），甚至推广到非度量的距离系数（如Nei氏距离）；线性规划法采用曼哈顿系数，这是一个在分支系统学中广泛应用的距离系数，由于其在分支树的可加性，因而具有较好的性质；②双曲线拟合法采用微分方程来确定插入类群与其它相邻类群（无根树）的距离，并使距离之和最小化；线性规划方法用线性规划技术优化无根树；③双曲线拟合法的计算较线性规划法更为复杂。

两个方法存在的缺点：① 祖先类群的杂合性高。Rogers[10]的研究表明，这些方法形成的分支树，有一个明显的趋势，即祖先类群（HTUs）越靠近树的下部，基因杂合性越大，通过杂合性与树根的距离的回归分析表明了这种明显的相关关系。Rogers[9~10]承认这是其方法的一个不足之处。对不同的系数，这种偏差现象是不同的。② 优化方法由于为了满足树长最大俭约的要求，在确定祖先类群（HTUs）各位点的等位基因的组成和其频率时，完全为适应和满足数学方法的需要，而不考虑遗传、进化的规律。③ 由此造成了等位基因组成的随意性，而不考虑现存类群中等位基因的分布特点。

造成这种后果的原因，在于：①过分强调了频率的重要性，没有从这种数量特征中看到等位基因的质量特性。等位基因频率资料不仅是一个数量性状，同时更重要的是，它也是一个质量性状。②过分依赖于数学的技术，数学的技术只有在赋予了一定的生物学的、遗传学的、进化论的含义才具有意义。因此，为数学运算过程设置一个最小化的要求，多次迭代的结果，使祖先类群位点的等位基因数目增加，从而使群体杂合性增加。

在频率空间重构系统发育方法中，究竟何种距离系数更优越，有的学者作了一定的研究。Swofford和Berlocher[12]认为曼哈顿距离（Manhattan distance）更具良好的性质，而Rogers[11]认为不然。Rogers根据：①度量性特性；②使祖先群（HTUs）杂合性接近现存类群的能力；③收敛于一个稳定群的能力；④运算时间这4个指标来判断各个系数的优劣[10]。在度量性的距离系数中，修正的Cavalli-Sforza和Edward弦距离[10]似乎是最好的，弧距离[10]。其次，因为它受群体群杂合性程度影响较大；

Rogers 遗传距离不适宜用于这类方法，Rogers 遗传距离系数产生的祖先类群的杂合性远远比其它系数更偏离现存（顶端）类群的杂合性。进一步的研究表明[11]，用于等位基因频率空间构建进化树时，大多数的距离系数均有高估遗传距离的趋向。

2　统计分支分类方法在等位基因频率资料中的应用

第一个用标准的统计推断方法估计系统发育树的工作是 Edwards 和 Cavalli－Sforza[31]的研究（1964，也见 Cavalli－Sforza 和 Edwards[7]），探讨连续变量如等位基因频率资料和数量性状的构树方法。

针对基因频率资料或数量性状的统计分支分类方法研究的文献不多，但十分复杂难懂[7,31~41]，这些方法少为人所知，更少为人所用，一部分原因是数学原理的坚深，一部分原因是计算困难。

对遗传模型（进化模型）的依赖是统计分支分类方法的特点之一。从 Edwards 和 Cavalli－Sforza[31]开始，最主要的假设是每个性状的进化独立地按照布朗运动（Brownian motion）过程进行，群体的平均表型以无限的线性方式随机扩散。

从生物学意义上考察，有二个生物学过程可导致这种进化：①随机遗传漂变十分接近布朗运动，不过在不同的等位基因频率尺度，这种扩散速率不同[36]；②选择系数的随机变化也接近布朗运动。同时假定，每个性状的布朗运动方差是一样的。

统计分支分类方法能对系统发育的各种参数作统计推断，是该法的主要优点。最大似然法能计算估计值的方差，故能对假设进行似然比检验（likelihood ratio test）。该法对估计值也能给出置信区间[36]。分类学家早就希望对所形成的系统发育结果能在统计的框架之下，给出其可信的程度[4]。因此，统计分支分类方法将是一类有发展前途的方法，特别是随着分子生物学数据的大量积累，对这方面方法的要求将越来越迫切，目前已在 DNA 序列，蛋白质氨基酸序列等资料发展了一些似然模型[42]。

统计分支分类方法力图避免其它方法的局限性。与距离矩阵法不同，该方法试图充分有效利用所有资料，而不是将资料浓缩为距离的集合。与俭约法不同在于为进化的概率模型采用了标准的统计方法[35,42]。

但统计分支分类方法也存在诸多的不足。首先，计算量大而复杂是它不易推广的原因；其次，进化模型建立在布朗运动的假设之上，也颇有争议。与布朗运动相接近的两个生物学现象已如上述，但是生物进化过程中的许多遗传变异不满足布朗运动假设，同时，布朗运动假设下的等位基因是属于中性突变基因，而现在的研究证明，许多的重要基因的活动是受选择的制约的；第三，各分支进化等速的假设也是不完全合理的；第四，从求解的数学方法而论，统计分支系统学的目的之一是通过一定的树枝长度改变，产生特定树状图的资料的似然值最大化，最后通过比较不同树状图的似然函数值，将具有最大似然值的分支图看作最佳估计。然而，为实现这一目的的登山法不能保证获得最大的似然值，而常常只能寻找到当前最大值；第五，统计分支分类方法以获得似然值最大的树为目的，以似然值作为判断标准，因而对等位基因在分支树上的动态和分布不能反映；第六，统计分支分类方法也不能在祖先类群等位基因频率空间重构系统发育，不能反映等位基因的频率变化。

参　考　文　献

1　徐克学. 数量分类学. 北京：科学出版社，1994

2　Kluge A G，J S Farris. Quantitative phyletics and the evolution of *Arurans*. Syst Zool，1969，18：1～32

3　Farris J S. Methods for computing Wagner trees. Syst Zool，1970，19：83～92

4　Throckmorton L H. Molecular phylogenetics. 221～239. In：Rhomberger J A，eds. Biosystematics in agriculture. New York：Wiley & sons，1978

5　Farris J S. Distance data in phylogenetic analysis. In：Funk V A，Brooks D R ，eds. Advances in Cladistics，2. New York：Columbia Univ Press，1981

6　Patton J C，Avise J C. An empirical evolution of qualitative Hennigian analysis of protein electrophoretic data. J. Mol. Evol.，1983，19：244～254

7　Cavalli－Sforza L L，Ewards A W F. Phylogenetic analysis：model and estimation procedures. Evolution，1967，21：550～570

8 Farris J S. The retention index and homoplasy excess. Syst Zool, 1990, 38: 406～407
9 Rogers J S. Deriving phylogenetic trees from allele frequencies. Syst Zool, 1984, 33: 52～63
10 Rogers J S. Deriving phylogenetic trees from allele frequencies: A comparison of nine genetic distances. Syst Zool, 1986, 35: 297～310
11 Rogers J S. A comparison of the suitability of the Rogers, modified Rogers, Manhattan, and Cavalli～Sforza and Ewards distances for inferring phylogenetic trees from allele frequencies. Syst Zool, 1991, 40 (1): 63～73
12 Swofford D L, Berlocher S H. Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. Syst Zool, 1987, 36: 293～325
13 Mickevich M F. Transformation series analysis. Syst. Zool., 1982, 31: 461～478
14 Mickevich M F, Mitter C. Treating polymorphic characters in systematics: A phylogenetic treatment of electrophoretic data. In Funk V A, Brooks D R, eds. Advances in Cladistics: Proceedings of the first meeting of the Willi Hennig Society. New York: Bronx, 1981. 45～60
15 Mickevich M F, Mitter C. Evolutionary patterns in allozyme data: A systematic approach. In: Platnick N I, Funk V A, eds. Advances in Cladistics. Volume 2. New York: Columbia Univ Press, 1983. 169～176
16 Hennig W. Phylogenetic Systematics. Ann Rev Entomol, 1966, 10: 97～116
17 Waterson G A, Guess H A. Is the most frequent allele the oldest? Theor Biol, 1977, 11: 141～160
18 Sneath P H A, Sokal R R. Numerical Taxonomy. San Francisco: Freeman, 1973
19 钟扬，李伟，黄德世. 分支分类的理论与方法. 北京：科学出版社，1994
20 Klotz L C, Blanken R L. A practical method for calculating evolutionary trees from sequence data. J Theoret, Biol, 1981, 91: 261～272
21 Saitou N, Nei M. The neighbor～joining method: A new method for reconstruction phylogenetic trees. Molecular Biology and Evolution, 1987, 4: 406～425
22 Sattath S, Tversky A. Additive similarity trees. Psychometrika, 1977, 42: 319～345
23 Fitch W M. A non～sequential method for constructing trees and hierachical classification. J Mol Evol, 1981, 18: 30～37
24 Farris J S. Estimating phylogenetic trees from distance matrices. Am Nat, 1972, 106: 645～668
25 Farris J S. HENNIG86: Program and documentation. Port Jefferson Station, New York, 1988
26 Brooks D R. Quantitative parsimony. In: Duncan T, Stuessy T F, eds. Cladistics. New York: Columbia University Press, 1984. 119～132
27 赵铁桥. 系统生物学的概念和方法. 北京：科学出版社，1995
28 Nei M. Genetic distance between populations. Am Natur, 1972, 106: 283～292
29 Buth D G. Biochemical systematics of the cyprinid genus *Notropsis* I. The subgenus Luxilus. Biochem Syst Ecol, 1979, 7: 69～79
30 Simon C M. Evolution of periodical cicadas: Phylogenetic inferences based upon allozyme data. Syst Zool, 1979, 28: 22～39
31 Edwards A F, Cavalli－Sforza L L. Reconstruction of evolutionary tree. In: Heywood V H, McNeil J, eds. Phenetic and Phylogenetic classification. Syst Ass Pub, 1964. 67～76
32 Edwards A W F. Estimation of the branch～points of a branch～diffusion process. J Roy Statist Soc B, 1970, 32: 155～714
33 Felsenstein J. Maximum likelihood estimation of evolutionary trees from continuous characters. Am J Human Genet, 1973a, 25: 471～492
34 Felsenstein J. The Alternative methods of phylogenetic inference and their interrelationship. Syst Zool, 1979, 28: 49～62
35 Felsenstein J. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. Biol J Linn Soc, 1981a, 16: 183～196
36 Felsenstein J. Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates and testing hypotheses. Evolution, 1981b, 35: 1229～1246
37 Felsenstein J. A statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. In: Duncan T and Stuessy T F, eds. Cladistics. New York: Columbia Univ Press, 1984. 169～191
38 Felsenstein J. Phylogenies and quantitative characters. Ann Rev Ecol Syst, 1988, 19: 445～471
39 Cavalli－Sforza L L, Piazza A. Analysis of evolutionary rates, independence and treeness. Theor Pop Biol, 1979, 8: 127～165
40 Astolfi P, Piazza A, Kidd K K. Testing of evolutionary independence in simulated phylogenetic trees. Syst Zool, 1978, 27: 391～400
41 Navidi W C, et al. Phylogenetic inference: Linear invariance and maximum likelihood. Biometrics, 1993, 49 (2): 543～555
42 Weir B S. Genetic Data Analysis Methods for Discrete Data. Sunderland M A, USA: Sinauer Associates, Inc. 1990

（责任编辑 梁 健）