

基于 DNA 序列的系统进化树构建

董安国^{1,2},高琳²,赵建邦²,呼加路²

(1 长安大学 理学院,陕西 西安 710064;2 西安电子科技大学 计算机学院,陕西 西安 710071)

【摘要】 【目的】利用生物 DNA 序列构建系统进化树,以分析生物的进化过程。【方法】利用最大似然估计原理,给出了构建系统发生树的模型和算法。分别建立了确定父节点的 ML 模型、确定进化时间的 ML 模型以及保守概率与相似度关系模型;并通过递推,逐步恢复生物进化的原过程。【结果】给定一系列生物 DNA 序列,通过本算法可以建立这些生物的系统进化树,以及生物各自父辈的 DNA 序列和进化时间。【结论】由本研究构建算法得到的生物进化关系合理、算法复杂度低,可以用于大规模生物群体的系统进化树构建。

【关键词】 DNA 序列;系统进化树;最大似然估计;进化时间

【中图分类号】 TP311.12 【文献标识码】 A 【文章编号】 1671-9387(2008)10-0221-06

Construction of phylogenetic tree based on DNA sequences

DONG An-guo^{1,2},GAO Lin²,ZHAO Jian-bang²,HU Jia-lu²

(1 College of Science,Changan University,Xi'an,Shaanxi 710064,China;

2 College of Computer Science and Technology,Xidian University,Xi'an,Shaanxi 710071,China)

Abstract: 【Objective】 The study constructed phylogenetic tree based on DNA sequence to analyze the evolutionary process of organism. 【Method】 A model and an algorithm were provided for constructing phylogenetic tree based on the principle of maximum likelihood estimate. Firstly,the ML model was given to determine father node,evolutionary time and the relationship between conserved probability and similarity. Then the process of organism evolution was recovered. 【Result】 Given a set of DNA sequences,phylogenetic tree can be constructed by our algorithm. Also,the father sequences and the evolutionary time are obtained from the evolutionary tree. 【Conclusion】 Phylogenetic trees obtained from our algorithm are rational. Furthermore,the algorithm is simple, and can be used to construct phylogenetic tree in large scale DNA sequences.

Key words: DNA sequence;phylogenetic tree;maximum likelihood estimate;evolutionary time

系统进化树已被公认为是一种生物进化模型,其建立几乎都是基于生物分子序列(氨基酸或 DAN)^[1]。系统发生树的构建从方法上可以分为 2 类^[2]:距离法和离散特征法。距离法首先构造一个矩阵,矩阵的元素代表每两个生物之间的距离,利用不同的聚类方法得到系统发生树。比较典型的方法有 UPGMA^[3]、NJ^[4] 和 Fitch-Margolias^[5],其中 UPGMA 只适用于系统进化树各个分支进化速率相同的情况;NJ 算法是由 Saitou 和 Nei 于 1987 年提出的,能较快地生成单一的亲缘树,缺点是构造的树不够准确,1988 年 Studier 等^[6]对其进行了改进,此后又出现了其他的改进算法^[7-10];当各个分支的突变率不同的时候,Fitch-Margoliash 法一般能够得到比 UPGMA 法更加准确的结果,是距离法建树中一

* 【收稿日期】 2008-04-02
【基金项目】 国家自然科学基金项目(60574039 和 30670486);陕西省自然科学计划项目(SJ08-ZT15);长安大学科技发展基金项目(07J04)
【作者简介】 董安国(1964—),男,浙江象山人,副教授,硕士,主要从事计算机算法设计及生物信息学研究。
E-mail: donganguo@zwwu.edu.cn

种相对较好的方法。距离法是基于距离的聚类算法,运算量小,但其不足是不能确定进化分支时间。离散特征法根据最优原则的不同,分为最大简约法(MP)和最大似然法(ML),其中 MP 方法旨在确定具有最少进化步骤的那棵树,ML 方法在给定序列进化模型的情况下,找到最有可能产生当前给定 DNA 序列的树。这类方法是对每种可能的树进行评价,找出系统进化树,其优点是得到的树中有进化时间。但当生物数量(m)很大时,树的数量将以指数规律增长,1986 年 MP 已被证明是困难问题^[11],而 ML 比 MP 的复杂程度更高^[1],所以在使用 ML 方法时,需要利用一些生物证据,排除一些不可能的树,减少搜索空间。因为似然函数比较复杂,求其最优解也比较困难,Woodhams 等在文献[12]中提出了 ML 建立系统进化树的近似算法,通过引入一个维数为 2^m 的向量,经转化后将似然函数变成一个二次多项式,利用代数方法求出近似解。

随着基因组研究的深入,有人开始以生物体内的代谢路径为依据,通过图论的方法来建立系统进化树^[13-14]。

本研究通过计算 DNA 序列的相似度,利用最大似然估计的原理,建立系统进化树的模型,逆向逐步推断出两兄弟生物的父亲 DNA 序列和进化时间,旨在为分析生物进化过程提供参考依据。

1 变量说明及基本假设

文中涉及到很多数学推导,为此,先对变量和符号进行定义和说明。在分析过程中,用到了生物进化(变异)的一些机理,所以依据这些机理作一些基本假设。

1.1 变量说明

设 $N = \{ 'A', 'T', 'C', 'G' \}$ 表示基因集, $X = \{ X_1, X_2, \dots, X_m \}$ 表示给定的 m 个生物体(系统进化树中的叶子节点), $X_i = \{ x_{i1}, x_{i2}, \dots, x_{in} \}$, $x_{ij} \in N$, ($i=1, 2, \dots, m; j=1, 2, \dots, n$), X_i 表示第 i 个生物体的基因序列; Y 表示祖先生物体(系统进化树中的内部节点)的集合。图 1 为生物进化的基本单元结构,设生物体 A、B、C 的基因序列分别为 $A = (a_1, a_2, \dots, a_n)$, $B = (b_1, b_2, \dots, b_n)$, $C = (c_1, c_2, \dots, c_n)$; t_1, t_2 为进化时间(图中节点上的字母代表生物体,边上的字母代表进化时间,下文类同), p_1, p_2 为 C 相对 A、B 的保守概率(基因不发生变异的概率)。进化时间越长,生物发生变异的概率越大。 p_k 与 t_k ($k=1, 2$) 之间的关系由 Jukes-Cantor 模型^[15]确定,即:

$$P(t) = \begin{matrix} & A & C & G & T \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{pmatrix} p(t) & s(t) & s(t) & s(t) \\ s(t) & p(t) & s(t) & s(t) \\ s(t) & s(t) & p(t) & s(t) \\ s(t) & s(t) & s(t) & p(t) \end{pmatrix} \end{matrix},$$
$$\begin{cases} p(t) = (1 + 3e^{-4at})/4 \\ s(t) = (1 - e^{-4at})/4 \end{cases}.$$

显然,如果保守概率等于 1,表示下一代没有变异;保守概率等于 0.25,表示下一代没有遗传。所以, $0.25 < p_1, p_2 < 1$ 。

为了研究生物体之间的亲近关系,定义它们的

相似度如下: $R(A, B) = \frac{1}{n} \sum_{i=1}^n u_i$, 其中:

$$u_i = \begin{cases} 1, & \text{if } a_i = b_i; \\ 0, & \text{if } a_i \neq b_i. \end{cases}$$

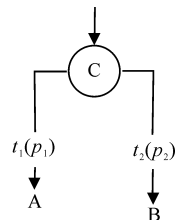


图 1 生物进化示意图

Fig. 1 Inlustration of speices evolution

1.2 基本假设

(1)在某一时间断面上存在的生物中,最相似的两个生物体是最后从它们的父亲分化出来的。

(2)将时间推进到父亲的时代,父亲与当时的某生物体或者与某些生物体的父辈具有大的相似性(即是同父兄弟)。

2 建立系统进化树的模型及算法

主要讨论 3 个问题:其一是以兄弟相对于其父亲的保守概率(对应着进化时间)为参数,建立父亲生物的 DNA 序列;其二是在当前生物中找出一对最后进化出来的兄弟,以它们的相似度最大为目标,确定二者相对于各自下一代的保守概率,即确定第 1 个问题中的参数;第三是给出了保守概率和相似度之间的关系,目的是减少第 2 个问题中优化模型的变量。

2.1 确定父亲生物的 ML 模型

根据生物进化关系,在每一个时间层面上,最相似的两个生物,一定是最近从某一生物(其父亲)进化来的。所以,构建系统进化树的过程就是从现有生物体中找出最相似的两个个体,通过它们估计出其父亲的基因序列,并将时间推移到其父亲的时代,

重复这一过程,直至找出全部生物的共同祖先。

对图 1 所示的进化单元结构,以保守概率 p_1 和 p_2 为参数,通过极大似然估计得到序列 C ,即求 C ,使条件概率 $P(A,B/C)$ 达到最大,具体模型(ML)为:

设似然函数 $L(C, p_1, p_2)$ 表示给定序列 C 出现 $A、B$ 的概率,即:

$$L(C, p_1, p_2) = P(A, B | C, p_1, p_2) = \prod_{i=1}^n P(a_i, b_i | c_i, p_1, p_2) \tag{1}$$

其中, $P(a_i, b_i | c_i, p_1, p_2) =$

$$\begin{cases} p_1 p_2, & \text{if } c_i = a_i = b_i; \\ p_1 (1 - p_2) / 3, & \text{if } c_i = a_i \neq b_i; \\ (1 - p_1) p_2 / 3, & \text{if } c_i = b_i \neq a_i; \\ (1 - p_1) (1 - p_2) / 9, & \text{else.} \end{cases} \quad (i = 1, 2, \dots, n)$$

记 $l(c_i, p_1, p_2) = P(a_i, b_i | c_i, p_1, p_2)$, 则:

$$L(C, p_1, p_2) = \prod_{i=1}^n l(c_i, p_1, p_2).$$

使 $L(C, p_1, p_2)$ 达到最大,得到:

$$c_i(p_1, p_2) = \{z | l(z, p_1, p_2) = \max_{x \in N} (l(x, p_1, p_2))\}, \quad i = 1, 2, \dots, n. \tag{2}$$

由式(2)可知,如果已知 $p_1、p_2$,通过序列 $A、B$ 就可以确定其父亲 C 的基因序列,下节将就 $p_1、p_2$ 的计算展开讨论。

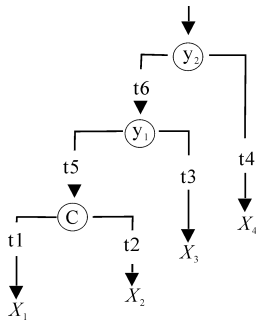


图 2 C 与 X_3 合并的生物进化示意图
Fig. 2 Sketch of C combine with X_3

设 S 是与 C 同时存在的生物体集合,与 C 的情况类似, S 中有的生物体也是其他生物体的父亲,其基因序列是关于相对其下一代保守概率的函数,所

$$R(W, C) = \begin{cases} \max_{p_1, p_2} \frac{1}{n} \sum_{i=1}^n v_i(p_1, p_2), & \text{if } W \notin Y \quad (*) \\ \max_{p_1, p_2, q_1, q_2} \frac{1}{n} \sum_{i=1}^n v_i(p_1, p_2, q_1, q_2), & \text{if } W \in Y \quad (**) \end{cases} \tag{3}$$

式中: $v_i(p_1, p_2) = \begin{cases} 1, & \text{if } c_i(p_1, p_2) = w_i; \\ 0, & \text{if } c_i(p_1, p_2) \neq w_i. \end{cases}$

$$v_i(p_1, p_2, q_1, q_2) = \begin{cases} 1, & \text{if } c_i(p_1, p_2) = w_i(q_1, q_2); \\ 0, & \text{if } c_i(p_1, p_2) \neq w_i(q_1, q_2). \end{cases}$$

式中: $w_i = w_i(q_1, q_2)$ 的计算与 c_i 类同,其中 $q_1、q_2$

2.2 确定保守概率的 ML 模型

如图 2 和图 3 所示,如果在 $\Sigma = \{X_1, X_2, X_3, X_4\}$ 中, X_1 和 X_2 的相关性最大,则 C 就是最先得到的一个父亲生物体,由式(2)可计算其 DNA 序列,但由于 C 到 $X_1、X_2$ 的进化时间未知,其保守概率不能确定,所以 C 的 DNA 暂时不能完全确定。考虑到 C 一定是由某一个更古老的生物体进化而来的个体,用 S 表示与 C 同时存在的生物体集合,则 C 一定与 S 中的某个生物(图 2)是兄弟,或与 S 中的某些生物体的前辈是兄弟(图 3)。如果是图 2 所示的情况,则在 $C、X_3、X_4$ 3 个生物体中, C 与 X_3 的相似度最大,即 C 与 X_3 本来就是兄弟,其相似度一定是很大的,根据最大似然估计的原理,可以使 $R(C, X_3)$ 达到最大来确定 C 中的参数 $p_1、p_2$,同时还用 $R(C, X_3)$ 表示其最大值。如果是图 3 所示的情况,则说明 X_3 和 X_4 是兄弟,表明 C 与 $X_3、X_4$ 的最大相似度 $R(C, X_3)$ 和 $R(C, X_4)$ 均小于 $R(X_3, X_4)$,这样第二个父亲生物体 D 是由 $X_3、X_4$ 得到的,此时 C 和 D 是兄弟,所以它们的相似度一定很大,根据最大似然估计的原理,可以使 $R(C, D)$ 达到最大来确定 C 和 D 中的参数。基于上述分析,可以得到以下的一般模型。

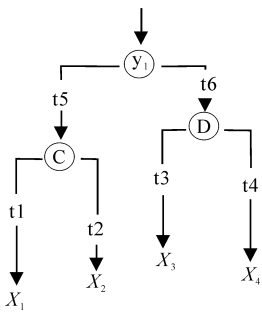


图 3 C 与 $X_3、X_4$ 父亲 D 合并的生物进化示意图
Fig. 3 Sketch of C combine with D

以计算两个父亲生物的相似度时,将包含 4 个参数。当 $W \in S$,计算 W 与 C 的相似度:

表示 W 对其下一代的基因保守概率。

经过计算,即可得到 C 与 S 中所有生物体的相似度,而 S 中所有叶子节点之间的相似度在事先可以求得,经比较,找出相似度最大的两个生物体,设为 $A1$ 和 $A2$,此时可按图 1 所示的进化单元结构得

出 A1 和 A2 的父亲序列,同时确定 A1 和 A2 相对于其下一代的基因保守概率,具体方法如下:

情况 1:如果 $A1, A2 \notin Y$,则由式(2)可计算它们父亲的 DNA 序列;

情况 2:如果 $A1 \in X, A2 \in Y$,则 $R(A1, A2)$ 的计算式为式(3)中的 $(*)$,使 $R(A1, A2)$ 达到最大,求出的 p_1, p_2 即是 A2 相对于其下一代的基因保守概率,代入式(2),可求出 A2 的序列,进而根据图 1 所示的生物进化单元结构,得到 A1 和 A2 上一代生物体的 DNA 序列。

情况 3:如果 $A1 \in Y, A2 \in Y$,则 $R(A1, A2)$ 的计算式为式(3)中的 $(**)$,使 $R(A1, A2)$ 达到最大,求出的 $(p_1, p_2), (q_1, q_2)$ 分别是 A1 和 A2 相对于其下一代的基因保守概率,代入式(2),即可求出 A1 和 A2 的序列,进而根据图 1 所示的生物进化单元结构,得到 A1 和 A2 上一代生物体的 DNA 序列。

本节式(3)关于 $R(W, C)$ 的计算是一个优化问题,在构建系统进化树的整个过程中,需要多次计算 $R(W, C)$,对此,本研究将采用穷举搜索最优解。为了提高搜索效率,下文给出了 p_1, p_2 的关系(q_1, q_2 的关系类同),从而将二维搜索转化为一维搜索,四维搜索转化为二维搜索。

2.3 保守概率与相似度的关系

如图 1 所示,保守概率 p_1, p_2 与 A、B 的相似度 $R(A, B)$ 存在一定的关系,这种关系可通过定理 1 给出。

定理 1:由图 1 所示的生物进化单元结构可知,序列 A、B 由 C 进化而来, $f(A, B)$ 表示序列 A、B 对应位置上的基因相同的概率,则:

$$4p_1p_2 - p_1 - p_2 + 1 - 3f(A, B) = 0。$$

证明:设 a, b, c 分别表示序列 A、B、C 在对应位置上的基因,由于

$$f(A, B) = P(a=b) = \sum_{x \in N} P(a=b|c=x)P(c=x)。$$

(4)

当 $x = 'A'$ 时,

$$\begin{aligned} P(a=b|c='A') &= P(a=b='A'|c='A') + \\ &P(a=b='T'|c='A') + P(a=b='C'|c='A') + \\ &P(a=b='G'|c='A'), \\ P(a=b='A'|c='A') &= p_1p_2。 \end{aligned}$$

$$\text{所以, } P(a=b|c='A') = p_1p_2 + \frac{(1-p_1)(1-p_2)}{3}。$$

同理, $P(a=b|c='T') = P(a=b|c='C') =$

$$P(a=b|c='G') = p_1p_2 + \frac{(1-p_1)(1-p_2)}{3},$$

$$\begin{aligned} \text{代入式(4)得: } f(A, B) &= \sum_{x \in N} P(a=b|c=x)P(c=x) = \\ &p_1p_2 + \frac{(1-p_1)(1-p_2)}{3}, \end{aligned}$$

$$\text{即 } 4p_1p_2 - p_1 - p_2 + 1 - 3f(A, B) = 0。$$

由 $R(A, B)$ 的定义可知,其表示 A、B 在对应位置上基因相同的频率,当序列长度 n 较大时,由大数定律可知, $f(A, B) \approx R(A, B)$, 所以有:

$$4p_1p_2 - p_1 - p_2 + 1 - 3R(A, B) \approx 0。 \quad (5)$$

式(5)表明,当生物 A、B 已知时, p_1, p_2 的关系是确定的,所以在计算式(3) $(*)$ 时,只要对 p_1 进行穷举,即可求出最优解;计算式(3) $(**)$ 时,只要对 p_1, q_1 进行穷举即可;由于 p_1, q_1 的值为 $0.25 \sim 1$, 设定计算精度为 0.01 ,在最不利情形下,对式(3)要计算 5 625 次,每次的运算量为 $O(n)$,所以总运算量为 $O(nm^2)$ 。

3 实验验证

利用 Matlab7.1 对本研究的算法进行了仿真试验研究,试验使用的电脑是 Intel Celeron(R) 2.26 GHz,内存为 256 M。以 20 种有胎盘哺乳动物的线粒体全基因组(mtDNA)作为分子数据。试验数据来源于 NCBI^[16],其名称和索取号列于表 1。

利用多序列比对算法 CLUSTALW^[17],对上述 20 种有胎盘哺乳动物的 mtDNA 序列进行比对,然后通过本算法构建出系统进化树(图 4)。

灵长动物、原蹄动物和啮齿动物三者中究竟哪两类之间的亲缘关系更近一些,不同的分子数据和分析方法支持不同的结果。本研究结果支持灵长动物与原蹄动物之间的关系更近一些。图 5 是文献[10]的研究结果,在进化关系上,图 4 和图 5 几乎相同,只有鸭嘴兽的亲缘关系有所不同。由于在现存的哺乳动物中,鸭嘴兽是最古老、最原始、最低等的种类,因此图 4 更合理地体现了这 20 种生物体的进化关系。此外,与图 5 相比,图 4 还能给出生物体的进化时间。

表 2 是对系统进化树(图 4)的定量表达,例如,3 号(黑猩猩)和 4 号(倭黑猩猩)生物体的父亲是 21 号生物体(他是历史上曾经存在过的生物体,本研究可以得到他的基因序列,这里没有列出),6 号(智人)和 21 号生物体的父亲是 24 号生物体,表 2 与图 4 一样,表达了 20 种生物体的进化历史,表 2 中的进化时间代表图 4 中相应分支的长度。

表 1 20 种有胎盘哺乳动物的名称及 GenBank 索取号

类别 Class	序号 ID	物种名 Species	简称 Shortened form	索取号 Index ID
灵长类 Primates	3	黑猩猩 <i>Pan troglodytes</i>	comChim	D38116
	4	倭黑猩猩 <i>Pan paniscus</i>	pigChim	D38113
	5	大猩猩 <i>Gorilla gorilla</i>	Gonlla	D38114
	6	智人 <i>Homo sapiens</i>	Human	V00662
	7	马来猩猩 <i>Pongo pygmaeus</i>	Orangutan	D38115
	8	白掌猿 <i>Hylobates lar</i>	Gibbon	X99256
	9	埃及狒狒 <i>Papio hamadryas</i>	Baboon	Y18001
原蹄动物 Ferungulates	10	长须鲸 <i>Balenoptera physalus</i>	Fin whale	X61145
	11	蓝鲸 <i>Balenoptera musculus</i>	Blue whale	X72204
	12	牛 <i>Bos taurus</i>	Cow	V00654
	13	斑海豹 <i>Phoca vitulina</i>	Harbor seal	X63726
	14	灰海豹 <i>Halichoerus grypus</i>	Gray seal	X72004
	15	猫 <i>Felis catus</i>	Cat	U20753
	16	马 <i>Equus caballus</i>	Horse	X79547
啮齿动物 Rodents	17	白犀牛 <i>Ceratotherium simum</i>	White rhino	Y07726
	18	大鼠 <i>Rattus norvegicus</i>	Rat	X14848
其他 Outgroup	19	小鼠 <i>Mus musculus</i>	Mouse	V00711
	1	北美负鼠 <i>Didelphis virginiana</i>	Opossum	Z29573
	2	岩大袋鼠 <i>Macropus robustus</i>	Wallaroo	Y10524
	20	鸭嘴兽 <i>Ornithorhyncus anatinus</i>	Platypus	X83427

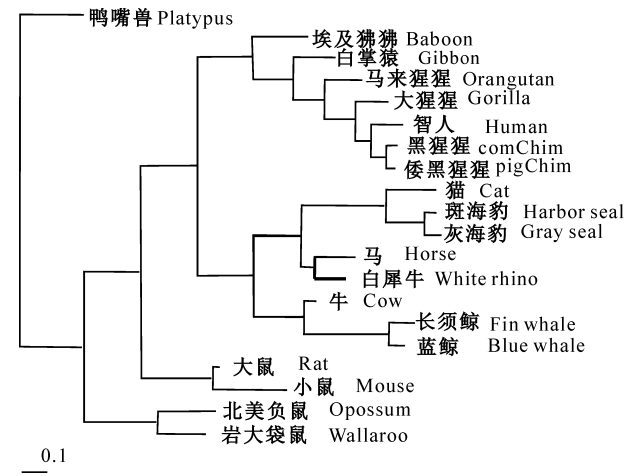


图 4 本研究建立的 20 种有胎盘哺乳动物的 mtDNA 系统进化树

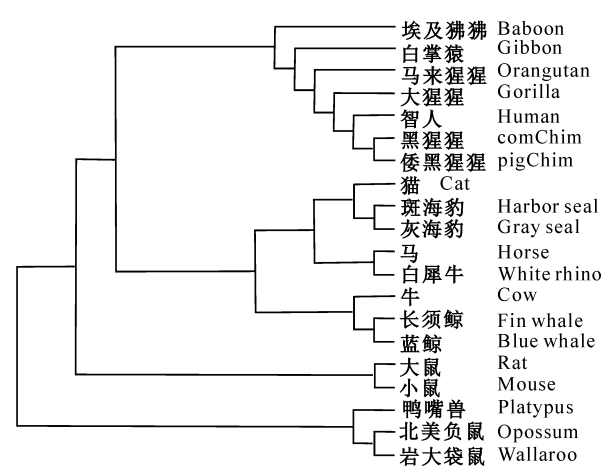


图 5 文献[10]建立的 20 种有胎盘哺乳动物的 mtDNA 系统进化树

Fig. 4 The phylogenetic trees of 20 kinds of mammals with placenta created by proposed algorithm

Fig. 5 The phylogenetic trees of 20 kinds of mammals with placenta created by references [10]

表 2 由 20 种有胎盘哺乳动物 mtDNA 系统进化树得到的父子关系与进化时间

父亲序号 Father ID	儿子序号 Son ID	保守概率 Conserve Probability	进化时间 Evolutionary time	父亲序号 Father ID	儿子序号 Son ID	保守概率 Conserve Probability	进化时间 Evolutionary time
21	3	0.974	0.038	30	15	0.874	0.194
	4	0.98	0.029		22	0.90	0.152
22	13	0.971	0.042	31	9	0.866	0.208
	14	0.96	0.059		28	0.90	0.152
23	10	0.933	0.100	32	12	0.974	0.038
	11	0.96	0.059		23	0.80	0.322

续表 2 Continued table 2

父亲序号 Father ID	儿子序号 Son ID	保守概率 Conserve Probability	进化时间 Evolutionary time	父亲序号 Father ID	儿子序号 Son ID	保守概率 Conserve Probability	进化时间 Evolutionary time
24	6	0.924	0.115	33	27	0.967	0.049
	21	0.96	0.059		30	0.80	0.322
25	5	0.922	0.118	34	32	0.873	0.196
	24	0.96	0.059		33	0.88	0.184
26	7	0.911	0.134	35	1	0.858	0.221
	25	0.92	0.120		2	0.88	0.184
27	16	0.897	0.156	36	31	0.836	0.258
	17	0.92	0.120		34	0.86	0.218
28	8	0.897	0.157	37	29	0.826	0.2765
	26	0.92	0.120		36	0.86	0.218
29	18	0.987	0.020	38	35	0.830	0.283
	19	0.82	0.286		37	0.85	0.219

4 总 结

利用最大似然估计的原理,建立了 2 个 ML 模型,以这 2 个模型为基础,构造了一种递推算法,该算法可以从现有的生物群体开始,逆向逐步恢复这些生物的进化过程。在每一步的递推中都要涉及到一个优化计算来确定保守概率,为了提高运算效率,给出了保守概率与相似度的关系,将 4 个变量的优化问题转化为 2 个变量;考虑到决策变量(保守概率)的取值为 0.25~1,所以对优化问题进行求解时,本研究采用了穷举法,这在一定程度上影响了算法效率。本算法是针对 DAN 序列设计的,但就其原理而言,可以应用到蛋白质序列中,下一阶段作者将进行这方面的研究。

[参考文献]

[1] Chor B, Tulle T. Maximum likelihood of evolutionary trees: hardness and approximation [J]. Bioinformatics, 2005, 1(21): i97-i106.

[2] 常 青,周开亚. 分子进化研究中系统进化树的重建 [J]. 生物多样性, 1998, 1(6): 55-62.

Chang Q, Zhou K Y. Phylogeny reconstruction in the study of molecular evolution [J]. Biodiversity Science, 1998, 1(6): 55-62. (in Chinese)

[3] Dave Thomas. Example calculation of phylogenies: the UPGMA method [EB/OL]. [2008-02-13]. <http://www.illTLSr.org/upgma/htm>.

[4] Saitou N, Nei M. The neighbor-joining method: a new method for reonsmacting phylogenetic trees [J]. Molecular Biology and Evolution, 1987, 4(4): 406-425.

[5] Chris. Fitch-Margoliash algorithm for calculating the branch lengths [EB/OL]. [2008-02-13]. <http://www.bioinfo.rpi.edu/~bystre/Courses/bio14540/lecture12/sld002/htm>.

[6] Studier J A, Keppler K J. A note on the neighbor-joining method of Saitou and Nei [J]. Mol Biol Evol, 1988, 5: 729-731.

[7] Pearson W R, Robins G, Zhang T. Generalized neighbor-joining: more reliable phylogenetic tree reconstruction [J]. Molecular Biology and Evolution, 1999, 16(6): 806-816.

[8] William J B, Nicholas D S, Aaron L H. Weighted neighbor-joining: a likelihood based apntvach to distance based phylogeny reconstruction [J]. Molecular Biology and Evolution, 2000, 17(1): 189-197.

[9] Ranwez V, Gascuel O. Improvement of distancebased phylogenetic methods by a local maximum likelihood approach using triplets [J]. Molecular Biology and Evolution, 2002, 19(11): 1952-1963.

[10] 李 斌,何红波,李义兵. 基于 DNA 序列 LZ 复杂性距离的系统进化树重构 [J]. 高技术通讯, 2006, 16(5): 506-510.

Li B, He H B, Li Y B. Phylogenetic tree reconstruction based on DNA sequence LZ complexity distance [J]. High Technology Letters, 2006, 16(5): 506-510. (in Chinese)

[11] Day W, Sankoff D. The computational complexity of inferring phylogenies by compatibility [J]. Syst Zool, 1986, 35: 224-229.

[12] Woodhams M D, Hendy M D. Reconstructing phylogeny by quadratically approximated maximum likelihood [J]. Bioinformatics, 2004, 20: i348-i354.

[13] Heymans M, Singh A K. Deriving phylogenetic trees from the similarity analysis of metabolic pathways [J]. Bioinformatics, 2003, 19: i138-i146.

[14] Oh S, Joun J G, Chang J H, et al. Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks [J]. BMC Bioinformatics, 2006, 7: 284.

[15] Chor B, Hendy M D, Snir S. Maximum likelihood jukes-cantor triplets: Analytic solutions [J]. Molecular Biology and Evolution, 2006, 23(3): 626-632.

[16] NCBI. NCBI Entrez Gene Database [EB/OL]. [2008-02-12]. <http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>.

[17] Thompson J D, Higgins D G, Gibson T J. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice [J]. Nucleic Acids Res, 1994(22): 4673-4680.