使用基因组数据进行贝叶斯物种分化时间估计

朱天琪^{1,2,3*}

1. 中国科学院国家数学与交叉科学中心, 北京 100190;

2. 中国科学院数学与系统科学研究院, 中国科学院随机复杂结构与数据科学重点实验室, 北京 100190;

3. 中国科学院数学与系统科学研究院应用数学研究所, 北京 100190

* 联系人, E-mail: zhutq@amss.ac.cn

收稿日期: 2019-01-30; 接受日期: 2019-03-06; 网络版发表日期: 2019-04-15

国家自然科学基金(批准号: 31671370, 31301093, 11201224, 11301294)、中国科学院青年创新促进会基金(批准号: 2015080)资助。

摘要 随着测序技术的发展, 基因组数据呈井喷式增长. 大量的基因组数据使得我们可以非常准确的估计出进化距离, 即进化速率和物种分化时间的乘积. 在很多情况下, 我们想要了解的是物种分化的绝对时间. 然而分子数据并没有分别提供进化速率和分化时间的信息, 我们可以使用贝叶斯方法借助化石、先验等外部信息进行估计. 近年来, 进化模型和计算方法方面研究取得的进展使得我们可以在复杂模型下分析多基因数据. 能够分析复杂模型的贝叶斯MCMC方法也成为了主流的物种分化时间估计方法. 本文主要介绍贝叶斯物种分化时间估计的框架以及相关工作近年来的进展.

关键词 贝叶斯物种分化时间估计, MCMC算法, 分子钟模型, 先验, 化石校准

1 贝叶斯方法和贝叶斯物种分化时间估计问题简介

1.1 贝叶斯方法和马尔科夫链蒙特卡洛方法(MCMC)

贝叶斯统计的一个主要特征是随机分布来描述不确定性, 例如把参数看做一个随机变量而非常数. 假定我们对一个参数 θ 感兴趣, 通过之前的经验积累, 可以在分析数据前给 θ 一个分布 $f(\theta)$, 我们称这个分布为先验分布, 它反映了我们对参数的主观认知. 给定参数的情况下数据 D 的概率 $f(D|\theta)$ 称为似然函数. 根据贝叶斯定理, 有

$$f(\theta|D) = \frac{f(D|\theta)f(\theta)}{f(D)} = \frac{f(D|\theta)f(\theta)}{\int f(D|\theta)f(\theta)d\theta}, \quad (1)$$

其中 $f(\theta|D)$ 是参数 θ 的后验分布, 即在给定数据的情况下参数的 θ 的分布. 参数的后验分布结合了先验的信息和数据的信息. (1)式中的 $f(D)$ 称为边际似然函数, 它是一个归一化常数. 在参数个数多于一个的时候, 计算 $f(D)$ 需要计算高维积分, 是比较困难的.

马尔科夫链蒙特卡洛方法(Markov Chain Monte Carlo), 简称MCMC, 产生于19世纪50年代早期, 是在贝叶斯理论框架下, 通过计算机进行模拟的蒙特卡洛方法(Monte Carlo). MCMC方法的基本思想是通过构造一条马尔科夫链, 使其平稳分布为待估参数的后验分布(即(1)式中的 $f(\theta|D)$), 通过这条马尔科夫链产生后验分布的样本. 当马尔科夫链达到平稳后, 所得到的样本可用于恢复后验分布. 使用MCMC方法能够避

引用格式: 朱天琪. 使用基因组数据进行贝叶斯物种分化时间估计. 中国科学: 生命科学, 2019, 49: 472–483

Zhu T Q. Bayesian molecular dating with genomic data (in Chinese). Sci Sin Vitae, 2019, 49: 472–483, doi: 10.1360/N052018-00224

免计算边际似然函数 $f(D)$, 通常它是一个高维积分, 因而MCMC方法能够用于分析复杂模型. 关于MCMC方法在贝叶斯物种分化时间估计问题中的应用的细节会在1.2节和第4节中涉及.

1.2 贝叶斯物种分化时间估计问题

物种分化时间的估计一直是一个棘手的问题, 主要原因在于分子序列只提供了进化距离的信息, 并没有分别提供进化速率和进化时间的信息, 其中进化距离是进化速率和物种分化时间的乘积. 简单来说, 这就类似于解方程时只有一个方程但是有两个未知数, 不同的时间和速率的组合能够得到相同的进化距离, 我们称之为不可识别性. 在非常简单的进化模型下, 可以使用非贝叶斯方法进行物种分化时间估计^[1~5], 但是这些方法普遍缺乏对进化过程中各种不确定因素的考虑. 当考虑较复杂的进化模型时, 我们只能借助先验、化石等外在信息, 使用贝叶斯方法才能得到绝对的分化时间和进化速率^[6~9].

贝叶斯方法是在2000年左右被引入分子物种分化时间估计领域的^[6,10,11]. 由于贝叶斯方法能够综合考虑化石、分子数据甚至形态学数据等多种信息来源, 它逐渐发展成为主要的物种分化时间估计方法, 也是近年来物种分化时间估计相关研究的热点之一^[8,9,12,13].

在贝叶斯框架下同时考虑形态学特征和分子数据来进行物种分化时间推断的方法称为全证据分析(total evidence dating)或综合分析. 全证据分析是一种端点定年(tip-calibration)法. 在全证据分析中, 化石不仅包含形态学特征信息, 还包含其在系统发育树中的位置信息. 在一次分析中可以同时估计化石信息, 树拓扑信息, 物种分化时间和其他相关参数. 虽然端点定年法具有不需要对节点进行化石校准等优势, 但这种方法也存在一定局限性. 本文主要集中介绍节点定年(node-calibration)方法, 关于端点定年方法的更多内容可以参看^[14,15]或者综述文章^[16,17].

贝叶斯估计是具有统计一致性的, 即在通常的贝叶斯分析中, 随着数据量增大, 先验对后验的影响会减弱, 参数后验分布的均值会收敛到参数真值. 但是由于时间和速率的不可识别性, 和通常的贝叶斯分析不同, 在带有化石校准的物种分化时间估计问题中, 即使数据量趋于无穷, 先验仍对后验有持续影响^[8,13,18,19]. 在贝叶斯分析中先验如何设置本就是一个

困难的问题, 而在物种分化时间估计问题中, 先验的选取和设置就更需谨慎^[20,21]. 此外, 在通常的贝叶斯分析中, 后验分布的方差是收敛到0的, 也就是后验分布会收敛到一个单点分布, 不再有任何不确定性. 而在物种分化时间估计问题中, 由于化石不确定性的存在, 即使使用大量数据, 后验不确定性都一直存在.

使用贝叶斯方法进行物种分化时间估计的框架如下:

令 $D=\{D_1, D_2, \dots, D_L\}$ 为序列数据, 其中 D_i 是基因位点 i 上对齐了的序列, L 为基因个数, θ 为进化模型中的参数, $\mathbf{t}=\{t_1, t_2, \dots, t_{s-1}\}$ 为 $s-1$ 个物种分化时间. $\mathbf{r}=\{\mathbf{r}_i\}$ 为进化速率, 其中向量 $\mathbf{r}_i$ 是基因位点 i 上的进化速率, 具体形式由分子钟模型决定. 根据贝叶斯公式, 模型参数、时间及速率的联合后验密度函数为

$$f(\theta, \mathbf{t}, \mathbf{r} | D) = \frac{f(D | \mathbf{t}, \mathbf{r}, \theta) f(\mathbf{r} | \mathbf{t}, \theta) f(\mathbf{t} | \theta) f(\theta)}{f(D)}. \quad (2)$$

在上式中, $f(\theta)$ 是模型中参数的先验, $f(\mathbf{t} | \theta)$ 是分化时间的先验, 根据(9)式通过一个生灭过程进行设置, 在第3.1和3.2节会详细介绍. $f(\mathbf{r} | \mathbf{t}, \theta)$ 是速率先验, 进化速率不仅在不同的基因位点之间可能有差异, 在基因树的各个树枝上也可能有差异. 第3.4节将详细回顾基因位点间进化速率先验设置的两种主要形式: 复合狄利克雷先验和独立同分布形式的先验. 基因树各个树枝上的速率先验由分子钟模型决定, 在第2节将回顾常用的分子钟模型: 严格分子钟、宽松分子钟和局部分子钟, 并在第3.3节回顾在不同分子钟模型下基因树树枝上速率先验的设置. 数据的似然函数为

$$f(D | \mathbf{t}, \mathbf{r}, \theta) = \prod_i f(D_i | \mathbf{t}, \mathbf{r}_i, \theta), \quad (3)$$

其中 $f(D_i | \mathbf{t}, \mathbf{r})$ 是第 i 个基因位点上的Felsenstein似然函数^[22]. 在使用软件分析数据时, 似然函数的计算最耗费时间, 可以使用近似算法来加速计算^[6,23,24].

(2)式的分母 $f(D)$ 称为边际似然函数, 是数据的概率, 它是一个高维积分. 仅在非常特殊的情况下, $f(D)$ 可以解析或数值计算. 使用马尔科夫链蒙特卡洛方法(Markov Chain Monte Carlo, MCMC)可以避免计算 $f(D)$, 这也是贝叶斯方法能够处理复杂模型的原因. 使用MCMC方法时, 用于抽样的马氏链的状态包含模型中的参数 θ , 分化时间 $\mathbf{t}$ 和各个位点各个树枝上的进化速率 $\mathbf{r}$. 在当前状态 $(\theta, \mathbf{t}, \mathbf{r})$ 下, 下一步拟访问的新状态

(θ^*, t^*, r^*) 由建议密度函数(proposal density) $q(\theta^*, t^*, r^* | \theta, t, r)$ 决定. 新状态以概率 α 被接受, 其中

$$\alpha = \min \left\{ 1, \frac{f(D | t^*, r^*, \theta^*)}{f(D | t, r, \theta)} \times \frac{f(r^* | t^*, \theta^*)f(t^* | \theta^*)f(\theta^*)}{f(r | t, \theta)f(t | \theta)f(\theta)} \times \frac{q(\theta, t, r | \theta^*, t^*, r^*)}{q(\theta^*, t^*, r^* | \theta, t, r)} \right\} \quad (4)$$

α 又称为接受比, 它其实是似然比、先验比和建议比的乘积. 在MCMC算法的具体实施过程中, 可先对 t^* 进行建议, 需注意分化时间之间不能相互冲突, 再依次对每个位点上的速率 r_i^* 进行建议, 最后对模型中的参数 θ^* 进行建议.

2 分子钟模型

2.1 严格分子钟模型

严格分子钟模型, 简称为分子钟, 是由Zuckerkandl和Pauling首次提出的^[25], 而Kimura首先提出了使用中性的理论来解释分子钟^[26]. 20世纪60年代分子钟假设的提出和中性理论的提出是20世纪进化生物学中的重大事件. Zuckerkandl等人在研究中发现, 不同物种中的某一蛋白质中氨基酸替代数目和物种分歧时间呈线性关系, 并将这种关系称为分子钟. 之后Kimura^[26]以及King等人^[27]的工作完善了分子钟假设. 在分子钟假设(或称严格分子钟模型)下, 进化速率在整棵进化树上是一个常数. 当进化主要由中性突变主导时, 我们可以认为在不同的物种或者谱系(lineage)之间, 突变速率大致是一个常数, 因而分子钟假设成立.

通常在分析群体数据或分歧时间较浅的近缘物种数据时, 可以假定分子钟成立. 例如对人、尼安德特人、黑猩猩和大猩猩的线粒体基因组数据分析表明, 人和黑猩猩的进化速率大致相同^[16]. 我们可以使用序列间差异小于5%作为一个非常粗略的判断准则, 更加合理的做法是使用似然比检验检测分子钟是否成立, 可参考文献^[28]第7.2节.

2.2 局部分子钟与宽松分子钟

分子钟模型虽然简单便于计算, 但是大多数时候, 特别是考虑远缘物种数据时, 严格分子钟都不成立(如^[29-32]). 通常世代的时间间隔、群体大小、基础代谢速率、强自然选择等都是可能引起系谱间进化速率

差异的原因^[16], 但是引起速率差异的进化机制以及这些因素如何影响进化速率并不十分清晰^[12,33,34].

局部分子钟首先被Kishino和Hasegawa引入^[35], 作为解决严格分子钟不成立时的方法. 局部分子钟假设进化树上的不同部分有不同的进化速率, 这里部分指的是一些树枝的集合, 而同一部分内的进化速率是相同的. 在使用局部分子钟时, 部分的划分和每个部分的速率是任意指定的, 这种不确定性也成为局部分子钟模型的主要缺点^[28].

考虑进化速率差异的宽松分子钟模型^[6,7,13]是和实际进化过程更加贴近的模型. 在贝叶斯物种分化时间估计问题中, 分子钟模型实际上体现在速率先验的设置上. 自相关速率模型和独立速率模型是常用的两个宽松分子钟模型, 大部分主流贝叶斯物种分化时间估计软件都提供了这两个分子钟模型供用户选择. 自相关速率模型中, 假定进化树上祖先和后代的进化速率是自相关的. 而在独立速率模型中假定每个树枝上的进化速率独立并来自相同的分布(例如对数正态分布). 这两个速率模型会在介绍速率先验的部分进行详细介绍.

3 时间和速率的先验

3.1 时间先验设置

生灭过程^[36]可用来描述物种的分化、灭绝以及物种抽样的过程, 是用于设置时间先验的数学工具^[37]. 生灭过程的重要数学性质可参考Nee等人的文章^[38]. 根据生灭过程假设, 在一个非常小的时间段 Δt 内, 物种个数增加一个即发生一次物种分化的概率为 $\lambda \Delta t$, 物种个数减少一个即一个物种灭绝的概率为 $\mu \Delta t$, 其中 λ 和被 μ 称之为出生率(birth rate)和死亡率(death rate). 发生多于一次物种分化或灭绝的概率可以忽略.

假定祖先物种出现在 t 时刻之前, 当前(即时刻0)存在的物种数为 S , 存在于样本中的物种数为 s , 则每个物种被抽到样本中的概率为 $\rho = s/S$, 其中 ρ 称为样本比例(sample fraction). 这里仅考虑在现存物种中抽样, 不考虑历史过程中抽样的情况. 出生率 λ , 死亡率 μ 以及样本比例 ρ 决定了树的形状. 一般来说, 当 ρ 比较小时, 由生灭过程生成的树在靠近当前时刻的枝长较长, 而当 ρ 比较大时, 靠近根结点的枝长较长^[8].

在带有化石校准的物种分化时间估计问题中, 时间的先验还依赖于化石信息. 假设根节点时间为 t_1 , 非根节点的集合 $\mathbf{t}_{-1}=(\mathbf{t}_C, \mathbf{t}_{-C})$, 其中 $\mathbf{t}_C$ 是带有化石校准的时间点集合, $\mathbf{t}_{-C}$ 是不带化石校准的时间点集合. 给定根结点信息、化石信息, $\mathbf{t}_{-1}$ 的联合先验为

$$\begin{aligned} f(\mathbf{t}_{-1} | t_1, s, C) &= f(\mathbf{t}_{-C}, \mathbf{t}_C | t_1, s, C) \\ &= f_{BD}(\mathbf{t}_{-C} | \mathbf{t}_C, t_1, s) f(\mathbf{t}_C | C), \end{aligned} \quad (5)$$

其中 $f(\mathbf{t}_C | C)$ 是 $\mathbf{t}_C$ 的联合先验, 依赖于化石校准信息, 后文会详细阐述带有化石信息的节点如何设置时间先验. 根据条件概率公式

$$f_{BD}(\mathbf{t}_{-C} | \mathbf{t}_C, t_1, s) = f_{BD}(\mathbf{t}_{-1} | t_1, s) / f_{BD}(\mathbf{t}_C | t_1, s), \quad (6)$$

其中 $f_{BD}(\mathbf{t}_{-1} | t_1, s)$ 和 $f_{BD}(\mathbf{t}_C | t_1, s)$ 由生灭过程给出. 根结点先验的设置取决于在该节点上是否有化石信息, 若无化石信息, 由生灭过程决定, 否则根据化石信息来设置先验.

综上所述, 分化时间的联合先验需根据化石信息和带有物种抽样(species sampling)的生灭过程共同决定, 即

$$\begin{aligned} f(\mathbf{t} | s, C) &= f(t_1, t_2, \dots, t_{s-1} | s, C) \\ &= f(t_1 | s) f_{BD}(\mathbf{t}_{-C} | \mathbf{t}_C, t_1, s) f(\mathbf{t}_C | C) \\ &= f(t_1 | s) \frac{f_{BD}(\mathbf{t}_{-1} | t_1, s)}{f_{BD}(\mathbf{t}_C | t_1, s)} f(\mathbf{t}_C | C). \end{aligned} \quad (7)$$

关于公式(5), (6)以及根结点先验的计算细节参看文献[8].

Stadler等人提出了使用化石化的生灭过程(fossilized birth-death (FBD) process)作为时间先验^[14,39]. FBD模型是一个描述物种分化、灭绝以及导致现生物种及灭绝物种产生的化石作用的过程. 在FBD模型中, 参数除了出生率 λ , 死亡率 μ 以及样本比例 ρ 之外, 还有化石产生的速率 ψ . FBD模型所使用的是端点定年(tip-calibration)方法, 端点定年方法把化石看作是分类单元并放置于进化树的端点(即叶结点)处. 与传统的节点定年(node-calibration)方法(3.2节会做更详细介绍)相比, 端点定年法不需要对每个节点进行化石校准. 此外, FBD模型能够更充分的利用所有可用的化石信息, 特别是在一个节点上存在多个化石的情况下. 但同时使用FBD过程作为时间先验也存在一些局限性, dos Reis等人^[16]对此作了详细的阐述.

3.2 化石先验设置

在物种分化时间估计问题中, 化石扮演了非常重要的角色^[40,41]. 在用于估计物种分化时间的公式(7)中, $f(\mathbf{t}_C | C)$ 就是给定化石信息后时间结点的先验, 简称为化石先验. 除了通常的化石记录形式之外, 重要的地质事件(如地峡的闭合、板块的破碎等)也可以作为时间的校准^[42]. 在早期的分析中, 化石对结点的校准是一个固定的时间^[43]. 但是化石本身具有不确定性, 而现在的分析中通常都需要考虑这种不确定性, 所以化石校准以一个区间的形式存在. 在区间中指定化石的上界、下界或同时指定上下界^[44]. 通常化石能够提供一个有效的下界信息, 但是如果把该结点所能获得的最古老的化石作为结点时间的上界是不合理的^[45,46]. 如果地层中缺失某个化石物种的信息, 可以用这个地层年代作为化石校准的上界, 但是这种上界包含的信息比较弱.

如果校准化石将结点的分化时间严格限制在区间 $[t_L, t_U]$ 内, 这样的校准信息称为硬边界(hard bounds), 在早期的物种分化时间估计问题中都使用了硬边界. 硬边界可以是双边校准, 也可以是单边校准. 近年来被广泛使用的软边界(soft bounds)一种新的设置化石先验的办法^[7,8,47]. 使用软边界时, 以一个小概率(通常 $<10\%$)允许真实物种分化时间在边界之外. 通过这种形式, 对时间结点 t_k 设置先验 $f(t_k | C)$. Yang等人^[8]指出了使用软边界相对于硬边界的优势. 当不同结点上的校准化石记录相互冲突或者和分子数据相冲突时, 使用软边界可以依靠分子数据来修正较差的化石先验, 使得后验估计时间更加合理. 而如果在这种情况下使用硬边界, 即使数据量很大, 后验时间估计也会极大地限制在化石先验区间内. 此外, 当化石与分子数据冲突时, 使用硬边界容易使后验分布收敛到错误的地方, 同时后验精度极高, 产生“假精确”的后验估计, 而使用软边界可以很大程度上避免这种“假精确”现象. 再次, 使用硬边界时, 当化石上界信息量不足时, 通常需要使用较安全但是时间过于古老的上界, 这样的设置可能引起后验估计的误差, 而使用软边界就避免了这样的问题.

3.3 分子钟模型与树枝上速率先验的设置

严格分子钟模型假设整个进化树上的进化速率 r

一致, 仅需根据情况对 r 指定一个先验分布即可, 指数分布或者均匀分布是常用的速率先验分布. 当分子钟不成立时, 情况要复杂很多. 自相关或者树枝上独立的速率分布是两种常用的速率分布.

Thorne等人^[6]和Kishino等人^[10]引入了使用几何布朗运动(geometric brownian motion, GBM)来设置树枝上的速率先验的方法. 假设在某一树枝上 t 时刻的进化速率为 r_t , 树枝连接的祖先节点上的进化速率为 r_A . 在自相关模型中, 假设 r_t 服从几何布朗运动, 或等价的, $y_t = \log(r_t)$ 服从布朗运动. 计算在该树枝上的平均进化速率需要通过对时间 t 积分来做平均, 这样的计算是非常复杂的. Kishino等人^[10]使用树枝的两个结点上的进化速率平均值作为整个树枝上的平均进化速率; 而Rannala等人^[13]使用了树枝中点上的进化速率来代替. 在几何布朗运动中, 有一个描述速率漂移(rate drift)的参数 σ^2 , 同时 r_A 也是一个参数. 在实际分析中, 需要对 σ^2 和 r_A 分别设置一个超先验(hyper-prior), 常用的形式是Gamma分布.

Rannala等人^[13]提出了树枝上独立同分布(i.i.d.)的速率先验, 这种速率先验模型考虑了系谱之间进化速率的差异, 但是没有考虑它们之间的相关性. 树枝上独立同分布的速率模型是大多数主流物种分化时间估计软件中使用的方法, 如MCMCTree^[48]和Beast^[7,49]软件中均提供独立同分布速率先验的选项. 在树枝上独立同分布的速率先验模型下, 在每个树枝上进化速率是固定的. 树枝上的进化速率与其他树枝上的进化速率独立, 并且每个树枝上的进化速率都服从对数正态分布Lognormal($\log(\mu) - \sigma^2/2$, σ^2), 其均值为 μ , 方差为 $\mu^2(\exp(\sigma^2) - 1)$. 与自相关速率模型相似, 在使用过程中, 需对两个参数 μ 和 σ^2 设置超先验.

3.4 基因位点间进化速率的先验

当使用基因组数据进行物种分化时间估计时, 每个基因位点(locus)上有一棵基因树. 当考虑进化速率的差异时, 我们既需要考虑一棵基因树上不同树枝的进化速率差异(即速率模型或分子钟模型), 也需要考虑进化速率在不同基因位点上的差异. 假定位点 i 的进化树上每个树枝的进化速率为 r_{ij} , $i=1, 2, \dots, L$, $j=1, 2, \dots, 2s-1$, 其中 L 为基因位点个数. r_i 是对这个位点上进化速率的一个平均衡量, 称为位点上的平均进化速率. 考虑位点间进化速率的差异时, 最自然的方法是让 $\{r_i\}$

独立同分布, 这也是相当长一段时间内主流物种分化时间估计软件, 如MrBayes, BEAST和MCMCTree的旧版本中都唯一使用这一先验. 例如选择在宽松分子钟模型下进行物种分化时间估计时, 若使用独立同分布的速率模型, 则 $r_{ij}(j=1, 2, \dots, 2s-1)$ 独立, 服从均值为 r_i 的Gamma或对数正态分布. 若使用自相关的速率模型, r_i 是树根上的进化速率, 也是基因树上任意位置进化速率 r_t 的期望.

众所周知, 使用贝叶斯方法时, 先验的设置是非常关键并有一定难度的. 近年来, 对物种分化时间估计问题中先验的研究集中在时间先验上, 直到dos Reis等人^[50]发现之前唯一被使用的基因位点间独立同分布的速率先验是有问题的. 在通常的贝叶斯分析中, 随着数据量的增大, 先验对后验的影响减弱. 但是在带有化石校准的物种分化时间估计问题中, 如果先验设置不合理, 后验估计值会收敛的错误的地方, 并且随着数据量的增大, 错的更加厉害. 同时后验置信区间长度非常短, 造成了估计的“假精确”, 即估计时间错误且确信程度极高.

dos Reis等人^[50]使用了基因位点间独立同分布的速率先验分析了人和黑猩猩的分歧时间, 其后验分布如图1(文献[50]中的图3)所示. 根据化石信息, 这个时间应该在5.7~10 My左右^[51], 而分子生物学研究表明进化速率应该在 10^{-9} 个突变 site⁻¹ year⁻¹附近. 在分析中使用了快速先验G(2, 2) (均值 10^{-8} 次突变 site⁻¹ year⁻¹), 适中先验G(2, 20) (均值 10^{-9} 次突变 site⁻¹ year⁻¹)和慢速先验G(2, 200) (均值 10^{-10} 次突变 site⁻¹ year⁻¹). 当数据包含300个位点时, 若使用快速先验, 后验时间估计值非常年轻, 为3.3(2.9, 3.6) MA (后验均值及95%后验置信区间); 而若使用慢速先验, 后验时间估计值过于古老, 为26.4(24.7, 28.2), 两者相差近10倍. 在这两种先验设置不适当的情况下, 后验均值都超出了化石限定的时间区间(图中虚线). 引起这个现象的原因是随着数据量的增加, 基因位点间平均进化速率 $\bar{r} = \sum r_i / L$ 的先验方差收敛到0, 从而先验主导了后验.

dos Reis等人^[50]提出使用复合狄利克雷先验作为解决方案. 复合狄利克雷(Compound Dirichlet)先验是一个两步骤式的先验. 在贝叶斯系统发育分析中, 类似的依赖于狄利克雷分布的两步骤式先验也被用于解决先验对后验影响过强的问题^[52,53]. 在使用复合狄利克雷先验时, 先对位点间平均进化速率 $\bar{r}$ 设置一个Gam-

图 1 使用独立同分布速率先验得到的人-黑猩猩物种分化时间的后验估计值。所使用的3个先验为: (1)快速先验 $\mu_r \sim G(2, 2)$ (空心菱形), 先验均值为1; (2)适中先验 $\mu_r \sim G(2, 20)$ (空心圆圈), 先验均值为0.1; (3)慢速先验 $\mu_r \sim G(2, 200)$ (实心圆圈), 先验均值为0.01。当使用快速先验时, 后验估计值随着使用基因位点个数的增加而减少; 而使用慢速先验时, 时间后验估计值随着使用基因位点个数的增加而变大。当使用300个基因位点时, 在这两种情况下, 后验估计值都超出了化石边界(水平虚线)。横坐标为位点个数, 时间单位为My。

Figure 1 Posterior estimation of the human-chimpanzee divergence times under the i.i.d. prior for locus rate. The three rate priors used are (1) A fast rate, $\mu_r \sim G(2, 2)$ (diamonds), with prior mean 1 (substitutions per site per 10^8 years); (2) A medium rate $\mu_r \sim G(2, 20)$ (empty circles), with prior mean 0.1; and (3) A slow rate, $\mu_r \sim G(2, 200)$ (black circles), with prior mean 0.01. When the prior rate is too fast, the estimated time becomes younger as L is increased. On the other hand, when the prior rate is too slow, the time becomes older with increased L . In both cases, the posterior times are outside the fossil bounds (dashed lines) when $L=300$ loci are used. The x axis is the number of loci used and the time unit is My

ma先验, 再把进化速率在所有基因位点上的总和 $\sum_{i=1}^L r_i = L\bar{r}$ 根据狄利克雷分布分配到各个位点。经过这样的设置后, 基因位点间平均进化速率和各个基因位点上的进化速率 $\{r_i\}$ 先验方差均收敛到一个常数, 避免了先验主导后验。

4 一个例子

以人和猩猩12s rRNA基因数据为例来说明如何使用贝叶斯方法进行物种分化时间估计。数据包含1个长度为 $n=948$ bp的基因, 在其中 $x=90$ 个位点上发生了突变。假设严格分子钟成立且序列的进化遵循JC69

模型^[54]。由于严格分子钟假设, 进化速率不依赖于时间。在JC69模型的假设下, 似然函数仅以 $d=2rt$ 为自变量, 因此模型中并无额外的模型参数 θ , 所有需要考虑的参数只有两个: 整个树上共享的进化速率 r 和两个物种的分化时间 t 。在这些假设下, 公式(2)变为

$$f(t, r | D) = \frac{f(D | t, r) f(r) f(t)}{\iint f(D | t, r) f(r) f(t) dr dt}. \quad (8)$$

在JC69模型下, 似然函数为

$$f(D | r, t) = \left(\frac{3}{4} - \frac{3}{4} e^{-8rt/3} \right)^x \left(\frac{1}{4} + \frac{3}{4} e^{-8rt/3} \right)^{n-x}. \quad (9)$$

在分化时间结点上, 我们使用化石校准信息(即化石先验) $0.112 < t < 0.337$ (时间单位: 100 My)^[55]。在实际操作中对化石使用一个软边界^[8], 允许分化时间有5%的可能性超出化石边界, 其先验密度函数为

$$f(t) = \begin{cases} 0.025 \frac{\theta_1}{t_L} \left(\frac{t}{t_L} \right)^{\theta_1-1} & t \leq t_L \\ \frac{0.95}{t_U - t_L} & t_L < t < t_U \\ 0.025 \theta_2 \exp\{-\theta_2(t - t_U)\} & t \geq t_U \end{cases}, \quad (10)$$

其中 $\theta_1=18.9156$, $\theta_2=168.889$ 。时间先验(10)的密度函数如图2A所示。在分子钟假设下, 各个树枝共享进化速率 r , 并且不依赖于时间 t 。我们分别使用指数分布Exp(10)和逆Gamma分布InvG(3, 0.2)作为速率先验, 其中InvG(α, β)的密度函数为

$$f(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{-\alpha-1} \exp\left(-\frac{\beta}{r}\right). \quad (11)$$

这两个速率先验的均值都为0.1(即 10^{-9} 次突变/位点/年), 方差为0.01, 两个速率先验的密度函数如图2B所示。使用不同速率先验得到的分化时间的后验密度函数如图2C所示。虽然后验密度的形状不尽相同, 但是得到的后验均值和置信区间都相同, 均为0.2637 (0.1325, 0.3420), 这主要是由于两个速率先验在尾部分布接近。

更一般的处理方法是使用MCMC方法得到(近似)后验分布。在这个例子中, 我们使用对称的建议密度函数, 即建议比为1, 所以(4)被简化为

$$\alpha = \min \left\{ 1, \frac{f(D | t^*, r^*) f(r^*) f(t^*)}{f(D | t, r) f(r) f(t)} \right\}. \quad (12)$$

图 2 带有化石校准的物种分化时间估计问题中的时间先验, 速率先验和时间估计后验密度函数. A: 带有软边界的化石先验 $0.112 < t < 0.337$ (100 My) 密度函数, 先验时间小于0.112和大于0.337的概率均为2.5%. B: 两个速率先验的密度函数. 实线为 $G(1, 10)$, 虚线为 $InvG(3, 0.2)$. 两个先验的先验均值均为0.1(即 10^{-9} 次核苷酸替代/位点/年), 方差均为0.01. C: 使用 $G(1, 10)$ 和 $InvG(3, 0.2)$ 得到的分化时间后验密度函数. 使用两种速率先验, 时间后验均值均为0.2637, 95%后验置信区间均为(0.1325, 0.3420) (网络版彩图)

Figure 2 Time prior, rate priors, and time posteriors in the problem of divergence time estimation with fossil calibrations. A: The density of the fossil prior with soft bound $0.112 < t < 0.337$ (100 My). The probabilities that the divergence time is less than 0.113 and larger than 0.337 are 2.5% each. B: The densities of two rate prior: $G(1, 10)$ (solid line) and $InvG(3, 0.2)$ (dash line). Both rate priors are with mean 0.1 (i. e., 10^{-9} substitutions per site per year) and standard error 0.01. C: The densities of posterior time estimations with rate priors $G(1, 10)$ and $InvG(3, 0.2)$. With both rate priors, the posterior mean is 0.2637 with 95% HPD CI to be (0.1325, 0.3420) (color online)

在程序运行过程中, MCMC链访问过的状态会被记录, 通过对这些状态进行简单的统计处理即可得到

参数的(近似)后验分布. 为了保证MCMC链抽样是从平稳分布中抽出, 我们需要舍弃最初的一些样本, 这个过程称为burnin. 使用MCMC方法对这个数据进行分析的简要步骤如下:

- (1) 设置 t 和 r 的初始值, 例如 $t=0.2$, $r=0.1$.
- (2) 在每个循环中依次进行如下步骤:
 - (i) 对 t^* 进行建议, 例如使用宽度为0.05均匀分布滑动窗口;
 - (ii) 对 r^* 进行建议, 例如使用宽度为0.01均匀分布滑动窗口;
 - (iii) 根据公式(12)计算接受比, 并以概率 α 接受新状态. 如果新状态被接受了, 那么令 $(t, r)=(t^*, r^*)$, 否则保持 (t, r) 不变;
 - (iv) 每 k 迭代次记录下 (t, r) , 其中 k 是抽样频率.
- (3) 总结数据. 将最初的 B 个样本抛弃. 在剩余的样本中, 将 (t, r) 中的 t 抽出, 来恢复 t 的分布.

为了避免数值误差, (12)中计算接受比时应该比较新旧状态下的对数似然函数值 l^* 和 l , 若 $l^* < l$, 则 $\alpha = e^{l^* - l}$.

5 后验估计的不确定性

在通常的贝叶斯分析中, 随着数据量增大, 后验均值应该收敛到参数真值, 后验置信区间长度收敛到0, 也就是说, 参数的后验分布收敛到一个单点的分布, 不存在不确定性. 但是由于物种分化时间估计问题中时间和速率是不可识别的, 即似然函数只是进化距离(为时间和速率的乘积)的函数. 在这个非通常性的贝叶斯分析中, 即使数据量趋于无穷, 不确定性还将永远存在于后验分布中. 对不确定性我们通常用后验方差或者后验置信区间长度的平方来衡量, 这两种衡量方式也是等价的. 为避免混淆, 在这一节中, 作者将locus称为基因位点, site称为位点.

5.1 严格分子钟模型下的无穷位点理论

Yang等人^[8]研究了分子序列仅含有一个基因位点(locus)时、严格分子钟模型下物种分化时间后验估计的渐近行为, 称之为无穷位点(infinite-site)理论. 假设速率和时间的先验分别为 $g(r)$ 和 $f(t_1, t_2, \dots, t_{s-1})$, 其中 s 为物种个数, 那么第 j 个分化时间 t_j 的后验密度函数

$$f(t_j | d_1, d_2, \dots, d_{s-1}) \propto g\left(\frac{d_j}{t_j}\right) f\left(\frac{d_1 t_j}{d_j}, \frac{d_2 t_j}{d_j}, \dots, \frac{d_{s-1} t_j}{d_j}\right) \left(\frac{d_j}{t_j}\right)^{2-s} \frac{1}{t_j}. \quad (13)$$

式中 d_i 表示结点 i 为祖先结点的树枝的枝长, 以平均每位点(site)上发生的核苷酸替代数来恒量. 从上式中可以看出, t_j 的后验分布并没有收敛到一个单点值, 而是收敛到了一个一维的分布. 如果把所有的分化时间的后验均值作为横坐标, 真实分化年代或后验置信区间长度作为纵坐标, 那么这些散点落在一条直线上^[8,19]. 进行实际数据分析时, 我们也可以对(后验均值, 后验置信区间长度)画散点图, 以检测数据量是否达到饱和.

5.2 严格分子钟模型下的有限位点理论

dos Reis等人^[19]通过解析计算和数值模拟研究了严格分子钟下的有限位点(finite-site)理论. 在研究中假设数据包含一个基因位点, 长度为 N , 那么后验不确定性以 $1/N$ 的速度收敛到数据量为无穷时的不确定性, 即

$$w^2 - w_\infty^2 \propto 1/N, \quad (14)$$

其中 w 是时间估计的后验置信区间的长度, w_∞ 是数据量为无穷时的后验置信区间长度, 后验置信区间长度平方是不确定性的一个度量. 根据有限位点理论, $u_F = w_\infty^2 / w^2$ 是由化石带来的不确定性, 而 $u_S = 1 - u_F$ 是由有限的数据量带来的不确定性. 随着数据量趋于无穷, (14)右侧趋于0, u_F 收敛到1而 u_S 收敛到0, 此时所有不确定性均来自化石.

5.3 宽松分子钟模型下多基因位点的有限位点理论

在基因组时代, 越来越多的数据分析都采用多基因数据. 同时, 囿于严格分子钟模型的各种限制, 宽松分子钟模型更被分析者所青睐. Zhu等人^[55]将之前的有限位点理论拓展到了宽松分子钟、使用多基因位点数据的情形. 通过对例子的解析分析, 他们预测

$$w^2 = w_\infty^2 + \frac{a}{L} + \frac{b}{NL}. \quad (15)$$

式中 L 是基因位点的个数, N 是每个基因位点的长度, a 和 b 是与 N 和 L 无关的常数. 通过数值模拟和真实数据分析也验证了(15)式的正确性. 在实际数据分析中, 每

个基因位点的长度不一定相同, 但(15)式所呈现的这种线性关系依然成立(见文献^[55]图4).

(15)式将后验估计的不确定性划分为3个部分. 第一个部分 $b/(NL)$ 是每个基因位点上基因树枝长估计的误差引起的不确定性, 这种不确定性是由有限的序列长度导致的, 当序列长度或者基因位点个数趋于无穷时, 这一项不确定性以速率 $1/N$ 收敛到0. 第二个部分 a/L 是由基因位点上进化速率的差异以及各个谱系上进化速率的差异引起的, 当所使用数据的基因位点数趋于无穷时, 这一项不确定性以速率 $1/L$ 收敛到0, 同时第一部分 $b/(NL)$ 也收敛到0. 最后一项 w_∞^2 是由化石不确定性引起的, 即使使用再多的数据, 这一项不确定性将持续存在.

Zhu等人^[55]对类人猿数据的分析表明, 想要增加后验估计的精度, 在数据总量相同的情况下, 增加基因位点数 L 比增加基因位点长度 N 的效果要明显显著. 这项研究强调了使用多基因位点数据的重要性. 但是, 即使使用大量的数据, 由于化石信息存在不确定性, 后验时间估计仍会存在大量的不确定性; 同时, 增加化石的使用量也能够大幅降低后验估计的不确定性.

6 贝叶斯物种分化时间估计软件

由于贝叶斯物种分化时间估计问题的复杂性, 使用现有软件进行相关分析几乎是数据分析的唯一选择. 本节将常用的贝叶斯物种分化时间估计软件进行总结.

6.1 MCMCTree

MCMCTree是PAML中的子软件包^[48]. MCMCTree中的化石使用软边界形式, 用户可以指定化石的先验分布形式和先验在软边界外的概率. 软件支持多种分子钟模型及多种复杂进化模型, 允许进化速率在位点(site)上、谱系间以及基因位点(locus)间存在差异. MCMCTree的新版本中对基因位点间的进化速率使用了复合狄利克雷分布, 因此时间后验估计对速率先验参数不敏感, 从而估计更加精确和具有稳健性, 对用户也更加友好. MCMCTree还可以选择对似然函数进行快速近似计算, 可以分析较大规模的基因组数据.

MCMCTree支持PHYLIP格式序列或者模式计数格式(site pattern counts), 当做近似计算时, 也支持编码蛋白质DNA序列格式. 当使用NEXUS格式时, 序列和树信息会被读入, 而命令块信息会被忽略.

6.2 BEAST

BEAST是最常用的贝叶斯MCMC分析综合软件之一^[49]. 与大多数贝叶斯物种分化时间估计软件不同, 在分析中BEAST使用MCMC链访问进化树所构成的空间, 因此使用BEAST分析时无需指定物种树. 软件提供了全面的进化模型和严格分子钟、宽松分子钟、局部分子钟等分子钟模型. BEAST还包含了形态特征模型, 可以用于综合分析. BEAST支持NEXUS以及FASTA格式序列, 以及可以读入BEAST XML文件作为输入.

6.3 MrBayes

Mrbayes是用于进行贝叶斯系统发育分析和进化模型选择的软件包^[56], 是最常用的贝叶斯分析软件. 发表于2003年的Bayes3是全球最高引100篇文章之一(截止到2017年, 数据来自web of science, 分析来自微信公众号iNature)^[57]. 在使用MrBayes进行物种分化时间估计时, 物种树可以固定, 或者限制在某些可能的树集合上, 或者遍历所有可能的标记树(label tree). 软件所提供的分子钟模型包括严格分子钟和3种宽松分子钟: Thorne-Kishino 2002 (TK02), 复合泊松过程(compound Poisson process, CPP), 和独立Gamma速率(independent gamma rates, IGR)模型. 若结点上存在校准化石, 则结点的时间先验被严格限制在化石指定的时间区间内. Mrbayes也可用于全证据分析. Mrbayes对序列格式有特殊要求, 需要在PAUP*中对序列格式

进行相应编辑.

6.4 PhyloBayes

PhyloBayes是一个使用蛋白质序列进行贝叶斯MCMC分析的软件包, 主要用于系统发育重建和系统分化时间估计^[58,59]. PhyloBayes基于被称为CAT的概率模型进行计算, 该模型考虑了蛋白质进化的位点特异性(site-specific)特征. 在分子钟模型方面, PhyloBayes提供了自相关和非自相关的分子钟模型. 根据化石信息, 用户可以选择软边界或者硬边界的形式. 似然函数的数值计算使用了数据扩增的方法来加速, 同时软件支持并行计算, 因而适合分析大规模的多基因数据. PhyloBayes对序列要求比较宽松, 能读入DNA, RNA序列和蛋白质序列. 支持PHYLIP格式序列.

6.5 DPPDiv

DPPDiv在固定的物种树下估计物种分化时间和进化速率^[60]. 软件提供了全面的进化模型供用户选择. 分子钟模型包含严格分子钟, 狄利克雷宽松分子钟和独立速率模型. 时间先验模型包括生灭过程模型和均匀分布模型. 软件还有两个快速近似分析版本FastDPPDiv和FDPPDIV. 支持PHYLIP和NEXUS序列格式.

6.6 Multidivtime

Multidivtime是第一个贝叶斯物种分化时间估计软件^[6]. 树枝上进化速率的几何布朗运动模型也是这个软件首次使用的. 时间先验的设置方法是在随机有根树上使用狄利克雷分布来划分枝长(此处枝长指绝对时间间隔). Thorne等人^[6]使用多元正态分布去近似似然函数, 以减少似然函数昂贵的运算代价. 支持PHYLIP格式DNA序列或者氨基酸序列.

致谢 感谢英国伦敦大学杨子恒教授对文章提出的建设性意见.

参考文献

- 1 Xia X, Yang Q. A distance-based least-square method for dating speciation events. *Mol Phylogenet Evol*, 2011, 59: 342–353
- 2 Tamura K, Battistuzzi F U, Billings-Ross P, et al. Estimating divergence times in large molecular phylogenies. *Proc Natl Acad Sci USA*, 2012, 109: 19333–19338

- 3 Paradis E. Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. *Mol Phylogenet Evol*, 2013, 67: 436–444
- 4 Fourment M, Holmes E C. Novel non-parametric models to estimate evolutionary rates and divergence times from heterochronous sequence data. *BMC Evol Biol*, 2014, 14: 163–172
- 5 Ho S Y W, Duchêne S. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol Ecol*, 2014, 23: 5947–5965
- 6 Thorne J L, Kishino H, Painter I S. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*, 1998, 15: 1647–1657
- 7 Drummond A J, Ho S Y W, Phillips M J, et al. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 2006, 4: e88
- 8 Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol*, 2006, 23: 212–226
- 9 Lepage T, Bryant D, Philippe H, et al. A general comparison of relaxed molecular clock models. *Mol Biol Evol*, 2007, 24: 2669–2680
- 10 Kishino H, Thorne J L, Bruno W J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol*, 2001, 18: 352–361
- 11 Thorne J L, Kishino H. Divergence time and evolutionary rate estimation with multilocus data. *Syst Biol*, 2002, 51: 689–702
- 12 Drummond D A, Raval A, Wilke C O. A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol*, 2006, 23: 327–337
- 13 Rannala B, Yang Z. Inferring speciation times under an episodic molecular clock. *Syst Biol*, 2007, 56: 453–466
- 14 Stadler T. Sampling-through-time in birth–death trees. *J Theor Biol*, 2010, 267: 396–404
- 15 Gavryushkina A, Heath T A, Ksepka D T, et al. Bayesian total-evidence dating reveals the recent crown radiation of penguins. *Syst Biol*, 2016, 64: syw060
- 16 dos Reis M, Donoghue P C J, Yang Z. Bayesian molecular clock dating of species divergences in the genomics era. *Nat Rev Genet*, 2016, 17: 71–80
- 17 Donoghue P C J, Yang Z. The evolution of methods for establishing evolutionary timescales. *Phil Trans R Soc B*, 2016, 371: 20160020
- 18 Britton T. Estimating divergence times in phylogenetic trees without a molecular clock. *Syst Biol*, 2005, 54: 500–507
- 19 Dos Reis M, Yang Z. The unbearable uncertainty of bayesian divergence time estimation. *J Syst Evol*, 2013, 51: 30–43
- 20 Inoue J, Donoghue P C J, Yang Z. The impact of the representation of fossil calibrations on bayesian estimation of species divergence times. *Syst Biol*, 2010, 59: 74–89
- 21 Warnock R C M, Parham J F, Joyce W G, et al. Calibration uncertainty in molecular dating analyses: There is no substitute for the prior evaluation of time priors. *Proc R Soc B-Biol Sci*, 2015, 282: 20141013
- 22 Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol*, 1981, 17: 368–376
- 23 Guindon S. Bayesian estimation of divergence times from large sequence alignments. *Mol Biol Evol*, 2010, 27: 1768–1781
- 24 dos Reis M, Yang Z. Approximate likelihood calculation on a phylogeny for bayesian estimation of divergence times. *Mol Biol Evol*, 2011, 28: 2161–2172
- 25 Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. *Evol Genes Protein*, 1965, 97: 97–166
- 26 Kimura M. Evolutionary rate at the molecular level. *J Genet Mol Biol*, 1968, 18: 219–225
- 27 King J L, Jukes T H. Non-darwinian evolution. *Science*, 1969, 164: 788–798
- 28 Yang Z H. Computational Molecular Evolution. Oxford: Oxford University Press, 2006
- 29 Langley C H, Fitch W M. An examination of the constancy of the rate of molecular evolution. *J Mol Evol*, 1974, 3: 161–177
- 30 Yoder A D, Yang Z. Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol*, 2000, 17: 1081–1090
- 31 Hasegawa M, Thorne J L, Kishino H. Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes Genet Syst*, 2003, 78: 267–283
- 32 Springer M S, Murphy W J, Eizirik E, et al. Placental mammal diversification and the cretaceous-tertiary boundary. *Proc Natl Acad Sci USA*, 2003, 100: 1056–1061
- 33 Bromham L, Penny D. The modern molecular clock. *Nat Rev Genet*, 2003, 4: 216–224
- 34 Ho S Y W. The changing face of the molecular evolutionary clock. *Trends Ecol Evol*, 2014, 29: 496–503
- 35 Kishino H, Hasegawa M. Converting distance to time: Application to human evolution. *Methods Enzymol*, 1990, 183: 550–570
- 36 Kendall D G. On the Generalized "Birth-and-Death" Process. *Ann Math Statist*, 1948, 19: 1–15
- 37 Yang Z, Rannala B. Bayesian phylogenetic inference using DNA sequences: A markov chain monte carlo method. *Mol Biol Evol*, 1997, 14: 717–

- 38 Nee S, May R M, Harvey P H. The reconstructed evolutionary process. *Phil Trans R Soc Lond B*, 1994, 344: 305–311
- 39 Heath T A, Huelsenbeck J P, Stadler T. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc Natl Acad Sci USA*, 2014, 111: E2957–E2966
- 40 Donoghue P C J, Benton M J. Rocks and clocks: Calibrating the tree of life using fossils and molecules. *Trends Ecol Evol*, 2007, 22: 424–431
- 41 Ho S Y W, Phillips M J. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst Biol*, 2009, 58: 367–380
- 42 Goswami A, Upchurch P. The dating game: A reply to heads. *Zool Scr*, 2010, 39: 406–409
- 43 Graur D, Martin W. Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision. *Trends Genets*, 2004, 20: 80–86
- 44 Donoghue P, Benton M, Yang Z H, et al. Calibrating and constraining the molecular clock. *J Vertebr Paleontol*, 2009, 29: 89a–89a
- 45 Reisz R R, Müller J. Molecular timescales and the fossil record: A paleontological perspective. *Trends Genets*, 2004, 20: 237–241
- 46 Benton M J, Donoghue P C J. Paleontological evidence to date the tree of life. *Mol Biol Evol*, 2007, 24: 26–53
- 47 Warnock R C M, Yang Z, Donoghue P C J. Exploring uncertainty in the calibration of the molecular clock. *Biol Lett*, 2012, 8: 156–159
- 48 Yang Z. Paml 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 2007, 24: 1586–1591
- 49 Bouckaert R, Heled J, Kühnert D, et al. Beast 2: A software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 2014, 10: e1003537
- 50 Dos Reis M, Zhu T, Yang Z. The impact of the rate prior on bayesian estimation of divergence times with multiple loci. *Syst Biol*, 2014, 63: 555–565
- 51 Benton M, Donoghue P, Asher R. Calibrating and constraining molecular clocks. In: Near T, ed. *The Timetree of Life*. New York: Oxford University Press. 2009, 35–86
- 52 Rannala B, Zhu T, Yang Z. Tail paradox, partial identifiability, and influential priors in bayesian branch length inference. *Mol Biol Evol*, 2012, 29: 325–335
- 53 Zhang C, Rannala B, Yang Z. Robustness of compound dirichlet priors for bayesian inference of branch lengths. *Syst Biol*, 2012, 61: 779–784
- 54 Jukes T H, Cantor C R. Evolution of protein molecules. *Mamm Protein Metab*, 1969, 21–132
- 55 Zhu T, Dos Reis M, Yang Z. Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. *Syst Biol*, 2015, 64: 267–280
- 56 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. *Syst Biol*, 2012, 61: 539–542
- 57 Ronquist F, Huelsenbeck J P. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 2003, 19: 1572–1574
- 58 Lartillot N, Lepage T, Blanquart S. Phylobayes 3: A bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 2009, 25: 2286–2288
- 59 Lartillot N, Rodrigue N, Stubbs D, et al. Phylobayes mpi: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol*, 2013, 62: 611–615
- 60 Heath T A, Holder M T, Huelsenbeck J P. A dirichlet process prior for estimating lineage-specific substitution rates. *Mol Biol Evol*, 2012, 29: 939–955

Bayesian molecular dating with genomic data

ZHU TianQi^{1,2,3}

1 National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;

2 Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China;

3 Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

With the advancement of the sequencing technology, the amount of genome data increases explosively. Genome data enable us to estimate the evolutionary distance accurately, which is the product of divergence time and evolutionary rate. However, molecular data do not provide the information about times and rates separately. By combining prior information about times and rates, fossil calibrations which provide information about absolute times, and information from molecular data, Bayesian MCMC methods make it possible to estimate the absolute divergence time as well as the evolutionary rate. With the development of phylogenetic models and computational methods in recent years, multiple loci data can be analyzed under complex models. Nowadays, Bayesian methods have become the prevailing dating methods. In this article, we introduce the framework of Bayesian dating, and the progress made in the related area.

Bayesian dating, MCMC algorithm, molecular clock models, prior, fossil calibrations

doi: [10.1360/N052018-00224](https://doi.org/10.1360/N052018-00224)