计算机科学
COMPUTER SCIENCE

基于混合路径HMC的分子树空间采样方法

李晓鹏, 凌诚, 高敬阳

引用本文

李晓鹏, 凌诚, 高敬阳. 基于混合路径HMC的分子树空间采样方法[J]. 计算机科学, 2023, 50(12): 322-329.

LI Xiaopeng, LING Cheng, GAO Jingyang. Mixed Path HMC Sampling Methods for Molecular Tree Spaces [J]. Computer Science, 2023, 50(12): 322-329.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221100057

基于混合路径 HMC 的分子树空间采样方法

李晓鹏[1]　凌　诚[2]　高敬阳[1]
1 北京化工大学信息科学与技术学院　北京 100000
2 中海国际中心超威半导体　北京 100000
(752668528@qq.com)

摘　要　随着现代分子序列数据越来越丰富,描述物种间历史关系的树状拓扑空间也急剧扩大,系统发育树的可靠推断仍面临着巨大挑战。近年来,马尔可夫链蒙特卡洛算法(MCMC)家族中最先进的哈密顿马尔可夫蒙特卡洛(HMC)算法被证明可以应用于系统发育分析,可以避免传统 MCMC 算法中存在的大量随机游走行为,加快马氏链的混合。但在更为复杂的多模态发育树空间中,HMC 算法无法通过从其他模式中获得提议来逃离局部的高概率区域,为了提升算法的健壮性,文中提出了一种混合路径哈密顿马尔可夫蒙特卡洛(MPHMC)的优化方法。在不增加额外的计算成本的情况下,所提算法采样路径中添加针对离散参数的非 HMC 更新组件,与 HMC 确定性更新交替进行,进而在树空间中引入了拓扑变化更大的分支重排策略,能更自由地遍历整个后验分布的树空间。在 5 组经验数据集上进行实验,结果证明,MPHMC 方法能更好地从正确的后验分布中采样;在比较难采样的大数据集上运行时,HMC 单一路径的采样算法可能会失效,而 MPHMC 方法能获得比使用广泛的系统发育分析工具 Mrbayes(MCMC)高 14% 以上的采样效率。

关键词:MrBayes;树空间;哈密顿马尔可夫蒙特卡洛(HMC);多模态后验分布;混合路径
中图法分类号　TP399

Mixed Path HMC Sampling Methods for Molecular Tree Spaces

LI Xiaopeng[1], LING Cheng[2] and GAO Jingyang[1]
1 School of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100000, China
2 China Overseas International Center, Advanced Micro Devices, Inc(AMD), Beijing 100000, China

Abstract　With the increasing abundance of modern molecular sequence data and the dramatic expansion of the tree-like topological space describing historical relationships between species, reliable inference of phylogenetic trees continues to face enormous challenges. In recent years, the most advanced Hamiltonian Markov Monte Carlo(HMC) algorithm in the Markov Chain Monte Carlo(MCMC) family has been shown to be applicable to phylogenetic analysis, which can avoid the large amount of random walk behaviors present in traditional MCMC algorithms and speed up the mixing of Markov chains. However, in the more complex multimodal development tree space, the HMC algorithm cannot escape from the local high probability region by obtaining proposals from other modes. In order to improve the robustness of the algorithm, a hybrid path Hamiltonian Markov Monte Carlo (MPHMC) optimization strategy is proposed in this paper. Without adding additional computational cost, the algorithm samples paths with a non-HMC update component for discrete parameters, alternating with HMC deterministic updates, and introduces a branch rearrangement strategy with greater topological variation in the tree space, enabling freer traversal of the entire posterior distribution's tree space. Experiments on five empirical datasets demonstrate that the MPHMC method better samples from the correct posterior distribution, and the HMC single-path sampling algorithm may fail when run on larger datasets that are more difficult to sample, while the MPHMC method achieves a sampling efficiency gain over 14% than the widely used phylogenetic analysis tool, Mrbayes(MCMC).

Keywords　MrBayes, Tree space, HMC, Multimodal posterior distribution, Mixed path

1　引言

关联现存与灭绝物种,了解物种间的历史关系是计算分子学的一个重要目标,也是进化生物学的基础。树拓扑是研究物种间进化过程的重要数据结构,在系统发育分析中将遗传性状(表型或基因型)映射在一棵树上,两个分类群的同源特征存在于它们共同的祖先中[1]。因此,结合现存分子数据,通过追踪树上不断变化的物种特征,可以重建进化的历史

到稿日期:2022-11-07　返修日期:2023-03-14
基金项目:北京市自然科学基金(5182018)
This work was supported by the Natural Science Foundation of Beijing(5182018).
通信作者:高敬阳(gaojy@mail.buct.edu.cn)

路径。重建序列数据(D)之间的系统发育关系,既需要指定大量关于生物进化假设的连续参数,如位点间的置换速率 r、树分枝长度 b 等,又包含了离散树状拓扑结构 τ 等参数;同时随着基因组时代的到来,描述关系的备选树数量也随着序列数据的丰富呈指数级增长,因此可靠的树推断仍面临着一些挑战[2],模型的复杂性和计算需求之间存在着相应权衡。在过去的二十年中,生物学家广泛使用基于贝叶斯范式的发育分析软件[3-6],其树采样方法一直是经典的马尔可夫链蒙特卡洛(MCMC)算法,通常也被称作随机游走(Random-Walk Metropolis,RWM)算法[7-8]。RWM 算法的随机游走行为并不能很好地快速引导树搜索从低概率区域向高概率区域移动,导致马氏链收敛到后验平稳分布 $\pi(\tau,b,r|D)$ 效率低下,混合缓慢。

哈密顿马尔可夫蒙特卡洛(Hamiltonian Monte Carlo,HMC)采样是 MCMC 家族最先进的算法之一,其具备抑制马氏链随机游走的特点,在空间维数增加时拥有良好的扩展特性,很大程度地弥补了 RWM 算法的不足[9]。由于 HMC 算法基于哈密顿动力学方程,存在一定的应用限制:它只适用于在平滑的目标分布中采样,无法直接在系统发育分析的复合参数(τ,b,r)空间中采样,否则会导致高拒绝概率和较差的性能。

Dinh 等发展了概率路径哈密顿蒙特卡洛(PPHMC)[10],他们引入了 Louis Billera 等定义的连续树空间——Billera-Holmes-Vogtmann(BHV)树空间[11],这是以一种组合方式将欧几里德空间粘合在一起的单连通空间。Dinh 等证明了在 BHV 树空间中,PPHMC 保留了哈密顿动力学的良好理论性质,即时间可逆性、体积保持性和可遍历性。PPHMC 算法适用于 BHV 树空间中对树拓扑采样,相比较经典的 RWM 算法,不同马氏链状态间的移动不再是随机的,达到了加速树推断的效果。

PPHMC 理论上可在树空间中设置较长的采样路径,加快树空间的探索,同时也能够保持较小的误差,形成接受率高的长轨迹。但长路径需要付出大量额外的计算代价,如后验概率函数梯度计算,并当采样路径穿越复合体的边界时,算法需要以一定的标准重新对树拓扑计算似然概率,导致计算时间巨幅增加。除此之外,PPHMC 在包含 30 个或更多分类群的经验数据集中无法快速收敛到正确的分布,因为在复杂的多峰值后验分布中,单一 HMC 方法会失效,算法无法从其他模式中获得提议,从而逃离局部高概率区域。

本文提出新的方法,即混合路径哈密顿马尔可夫蒙特卡洛(MPHMC)算法,来应对上述问题,在采样路径中添加针对离散参数(τ)的非 HMC 更新组件,与 HMC 确定性更新交替进行,并用接受-拒绝机制确保更新都会收敛到它们的目标密度,权衡两种更新路径,在保证目标后验分布正确性的前提下,减少采样路径穿越复合体的边界,有效遏制了计算时间的巨幅增加;并且可以在树空间中引入拓扑变化更大的分支重排策略。这种混合类似于平行回火算法构建平行链,增加了局部峰值之间的跳跃机会,使得算法不再只依赖基于局部梯度信息的随机最近邻交换(NNI),能够更快更自由地遍历后验分布的树空间。我们在 5 组经验数据集上进行实验,结果证明,优化后的算法能更好地从正确的后验分布中采样;并且在更为丰富的数据集上运行时,HMC 单一路径的采样算法可能会失效,无法从正确的后验分布中采样,而 MPHMC 算法的采样效率依然比使用广泛的系统发育分析工具 Mrbayes 提升了 14%以上。

2 相关工作

2.1 BHV 树空间的定义

系统发育树(τ,b)是具有 N 个叶子节点的树拓扑图 τ,其一般假定为无根二分叉的结构(内部节点度为 3)。树中每个叶子节点包含着可观测的分子序列数据,而内部节点的数据是缺失的,只能依赖于叶子节点进行推断,其中利用了丰富的分子置换马尔可夫过程模型,包括相对简单的 JC69 模型[12]或者像广义时间可逆 GTR[13]之类的复杂模型。树中每条边 e 都关联着一个非负数 b(枝长,$b>0$),定义为平均每个位点(如核苷酸碱基对)置换的期望数目,用 $n=2N-3$ 表示这样一棵树的边数,令 T_N 是所有 N 个叶子系统发育树的集合。

最近邻交换(NNI)是改变树拓扑分支结构最常用的策略之一,具体的做法是将树的一条内部边压缩为零,然后将得到的 4 度节点以新的方式变为一条边和两个 3 度节点。NNI 可以通过这种方式来形式化彼此"接近"的树拓扑,以及彼此"遥远"的树拓扑。如果将树 τ_1 转换为 τ_2 只需要单次 NNI 移动,则两个树拓扑 τ_1 和 τ_2 称为相邻拓扑。Billera 等将这种形式化的 T_N 集合参数化为一种连续树空间,即 BHV 树空间[11],具体是将多个维数为 n 的 orthant 粘合在一起,形成了一个几何对象 orthant complex,每个 n 维 orthant 边界都是一组低维的 orthant faces,orthant complex 是欧几里得空间 orthants 的并集 X,任何两个 orthants 的交集是两个 orthants 共同的 face。因此,X 的每个状态都由一对(τ,b)表示。X 中 τ 拓扑结构索引是离散参数,而 b 枝长是空间中的连续分量。X 复合体是 $(2N-3)!!$ 个 n 维 orthants 沿着它们的共同面粘合在一起,因为 N 个叶子节点可以构成该数量级的不同树拓扑结构。

2.2 Mrbayes RWM 算法

贝叶斯方法的关键特征在于它是一种针对参数的概率分布概念,由给定的序列数据集合 D 以及对似然模型中参数的先验分布 $\pi_o(\tau,b)$,计算系统发育树拓扑和进化模型参数的概率 $\pi(\tau,b|D)$。算法在开始时可以假设每种树拓扑具有相等的先验概率,将序列信息和分子置换模型代入贝叶斯公式(1)计算每棵树的可能性。

$$\pi(\tau,b|D)=\frac{L(D|\tau,b)\times\pi_o(\tau,b)}{P[D]} \tag{1}$$

但问题在于,系统发育树包含树拓扑结构和置换模型中的诸多参数,边际似然概率 $P[D]$(贝叶斯范式(1)的分母部分)难以计算,无法直接得到系统发育树模型参数的后验分布,贝叶斯法很难直接用于计算系统发育树参数的后验分布。为方便后续实验比较,这里不考虑分子置换模型假设的参数推断。

一种突破性进展是借助于 RWM 算法,构造有效的 MCMC 采样器,间接地从后验分布 $\pi(\tau,b|D)$ 中抽取大量有效的样本,从而估计系统发育树的后验分布[14-15]。在 Mrbayes 程序中,RWM 算法从随机生成的初始树和一组模型参数值(包括初始枝长和分子进化模型参数)开始运行,随机提出建议对一个或多个模型参数进行更改,如一次局部的树分支交换等策略,使得在树参数空间中的每个点,在拓扑的形状、分支长度或进化参数值上都不同于其相邻点。其核心在于,Metropolis-Hastings算法要求对每个建议的状态计算建议比率 $q(\theta|$

$\theta^*-)/q(\theta^*-|\theta)$(Hastings 比率),以便校正有偏差的建议,确保状态转移的可逆性,以获得正确的目标分布[16]。最后比较序列数据的两个备选历史,需要计算这两个新旧状态的后验概率,这使得难以计算的边际似然概率被抵消,新状态θ^*被接受的概率 α 为:

$$\alpha=\min\left(1,\frac{\pi(\theta^*)}{\pi(\theta)}\times\frac{q(\theta|\theta^*)}{q(\theta^*|\theta)}\right) \tag{2}$$

算法在运行足够长的时间后,马氏链最终会收敛到目标分布,但前期样本受初始状态的影响,可能会遵循非常不同的分布,因此需要舍弃前期一定数量的老化样本(burn-in samples),得到平稳样本集。我们期望马氏链快速度过老化期,并寻找到正确的系统发育解决方案集合,统计汇总每个样本的参数值,每棵树被访问的频率是该树后验概率的 MCMC 估计[17],因此可以用其来估计模型中每个参数的后验概率分布,这些样本序列通常也被称为典型集。

但实际中,Mrbayes(RWM)算法运行是基于对树状态空间中任一参数的随机变动,在采样前期会拒绝大量的样本,致使混合速度缓慢,而且采样期间遍历的树拓扑有限,尤其在具有更多分类群的大树空间中难以得到典型集。

2.3 HMC 确定性路径

HMC 算法在 MCMC 的基础上引入哈密顿系统的动力学方程(3),利用后验分布的梯度信息来确定参数 θ 随时间 t 的变化,从而抑制随机游走[11,18]。哈密顿量 $H(\theta,p)$是系统的势能 $U(\theta)$和动能 $K(p)$之和,并用它定义一个联合分布 $P(\theta,p)\rightarrow 1/Z*\exp(-H(\theta,p)/T)$(正则分布),$T$ 代表系统温度,Z 是归一化常数。哈密顿运动方程被用来构建新的候选状态,将该状态作为马尔可夫链的下一个状态。

$$\begin{aligned}&H(\theta,p)=U(\theta)+K(p),K(p)=\sum_{i=1}^{d}\frac{P_i^2}{2m_i}\\&\frac{\partial\theta_i}{\partial t}=\frac{\partial H}{\partial p_i}=\frac{\partial K(p)}{\partial p_i}\\&\frac{\partial p_i}{\partial t}=-\frac{\partial H}{\partial\theta_i}=-\frac{\partial U(\theta)}{\partial\theta_i},i=1,\cdots,d\end{aligned} \tag{3}$$

其中,$\boldsymbol{M}$ 是一个对称的正定"质量矩阵",在标准 HMC 中通常是单位矩阵的标量倍数,相比 RWM,它要求除了包含 d 维目标分布外,还需要人为设置同样维数的独立动量 p(通常设置为标准的正态分布)来构造哈密顿运动方程。由于独立性,在对目标分布和动量的联合分布 $P(\theta,p)$进行采样后,可以只关注感兴趣的目标分布 $U(\tau,b|D,\theta_o)$:

$$U\propto-\log L(D|\tau,b,\theta)-\log\pi_o(\tau)-\log\pi_o(b) \tag{4}$$

哈密顿动力学具有良好的性质,如可逆性、能量守恒和相空间体积不变。这些性质使得哈密顿力学可以生成有效的马尔可夫链,并且在这些链的迭代过程中,新生成的状态具有较高的接受率。Dinh 等由此发展了概率路径哈密顿蒙特卡洛(PPHMC)[10],并证明了 PPHMC 采样算法在树空间 X 中保留了哈密顿动力学的良好理论性质,即时间可逆性、体积保持性和可遍历性。需要注意的是,后验分布在每个欧氏分量的边界处仍然是不可求导的,可以先利用平滑代理函数将导数的不连续性换成势能(后验分布函数)的不连续性,从而使用 Afshar 等的"折射"方法来处理这种边界的不连续性[19],保证了程序在实践中能发挥出 HMC 算法的优势。

HMC 算法在后验分布函数的梯度引导下,可以在状态空间中提出距初始状态遥远的建议(沿着速度方向快速流动),并且保证了提议的高接受率。如图 1 中蓝色虚线表示 HMC 计算的确定性路径,每次穿越 orthant 时对应一次围绕零长度分支的随机最近邻交换(NNI),相比 RWM 混合速率得到提升;而且长轨迹(Multiple Leapfrog Steps)很大程度地提升了 HMC 的效率,这使得算法在树空间 X 中,以较高的概率接受离旧状态较远的新状态,从而在采样期间可以遍历到更多的树拓扑,更有效地探索系统发育树拓扑的后验分布。同时,长路径会大量增加树拓扑对数似然概率的计算次数(Log Likelihood Probabilities of the new state of the chain, LLPs)。

$$\begin{aligned}&L_i(x_i)=\sum_{x_j}p_{x_ix_j}(b_j)L_j(x_j)\times\sum_{x_k}p_{x_ix_k}(b_k)L_k(x_k)\\&L(\tau,b)=\prod_{s=1}^{S}\sum_{a}^{\Omega}\pi(a)L_r^s(a)\end{aligned} \tag{5}$$

其中,x_i 表示 i 节点观测到的核苷酸,k 和 j 分别为 i 节点的左右子节点,b 表示树拓扑的枝长参数。根据剪枝算法[20],内部节点$L_i(x_i)$等于两棵子树似然部分的乘积,在访问完树上所有内部节点后,计算根节点的概率向量时需要乘以该位点的核苷酸平衡频率 $\pi(a)$。

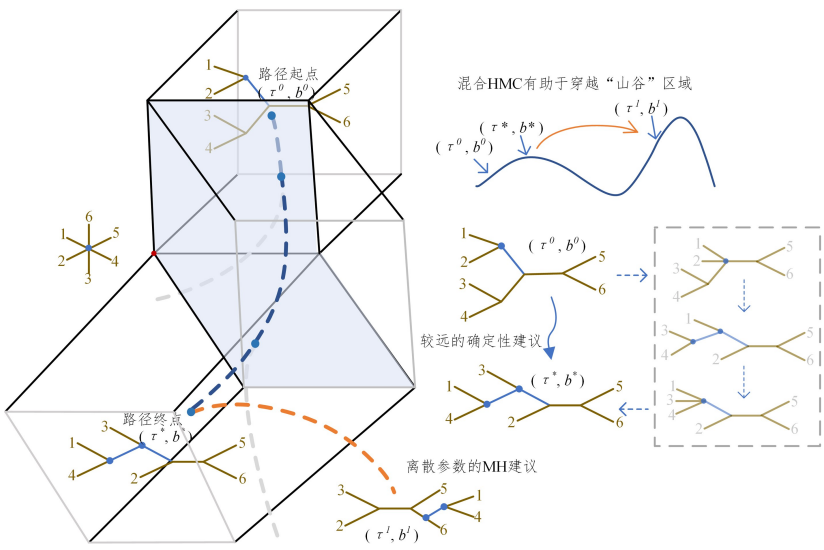


图 1 树空间中 MPHMC 的计算路径(电子版为彩图)

Fig. 1 MPHMC computational path in tree space

系统发育树的梯度计算和计算似然函数一样，先通过后序遍历对 LLPs 进行更新并存储，同时通过前序遍历，将任意节点上的 LLPs 展开为其后序和前序部分似然向量的内积，这样方便在采样序列数的线性时间内计算似然函数对所有特定于分支的参数的梯度[21]。

此外，HMC 在实践中同样需要进行 MH 校正，因为在程序中用离散化数值来模拟连续时间的积分计算存在一定误差。随着辛积分器（Symplectic Integrator）的发展，HMC 计算变得易于处理，蛙跳积分器（Leapfrog）就是一组通用解决方案的特定实例（Euler 方法的修改升级版）。

$$\begin{aligned} p_i\left(t+\frac{\epsilon}{2}\right)&=p_i(t)-\left(\frac{\epsilon}{2}\right)\frac{\partial U}{\partial \theta_i}(\theta(t))\\ \theta_i(t+\epsilon)&=\theta_i(t)+\epsilon\frac{p_i\left(t+\frac{\epsilon}{2}\right)}{m_i}\\ p_i(t+\epsilon)&=p_i\left(t+\frac{\epsilon}{2}\right)-\left(\frac{\epsilon}{2}\right)\frac{\partial U}{\partial \theta_i}(\theta(t+\epsilon)) \end{aligned} \tag{6}$$

式(6)不是依次对动量和位置（即参数）向量进行整步长模拟，而是先对动量变量进行半步更新，然后用更新后的动量向量对位置向量进行整步长更新，最后根据新的位置向量再对动量向量进行半步长更新。在一个更新步骤中，算法进行 L 步 leapfrog 近似，因此除了采样代数，还需要用户指定ϵ和 L 两个超参数，从而确保 HMC 的正确性[9,22]。

3 MPHMC 算法设计

3.1 PPHMC 的多模态问题

当现存数据具有 30 个或更多分类群数目时（$N>30$），会出现难以采样的多峰值后验分布，并且随着树空间变大，该问题会更严重[23]。类似地，在具有更多位点的大数据集中，峰趋向变得更高，谷变得更深，使得在峰之间跨越更加困难[2]。在这些复杂的后验概率分布中，HMC 方法也会失效，梯度信息也无法检测到多峰[24]，初始状态的位置很大程度影响了 HMC 方法的效率，单一 HMC 算法在树空间 X 中很容易陷入局部高概率区域，它只能通过从其他模式中获得提议来走出这个“陷阱”，但算法的提议机制使这种情况几乎不可能发生。设 $p(0)$ 为该哈密顿系统初始时刻的动能，那么 $2K(p(0))=\|M^{-1/2}p(0)\|_2^2\sim\chi_d^2$，$M^{-1/2}p(0)\sim N(0,I)$，$\chi_d^2$ 表示自由度为 d 的卡方分布，在后验分布 π 中，跨越被深谷隔开的局部峰的概率有一个切比诺夫边限[25]。

$$\begin{aligned} P(K(p(0))>\Delta)&=P(\chi_d^2>2\Delta)\leqslant\left(\frac{2\Delta}{d}\right)^{d/2}e^{\frac{d}{2}-\Delta}\\ \Delta&=\max_{0\leqslant s\leqslant t}U\{\theta(s)\}-U\{\theta(0)\}\\ &=-\log\frac{\min_{0\leqslant s\leqslant t}\pi(\theta(s))}{\pi(\theta(0))} \end{aligned} \tag{7}$$

其中，Δ 是沿着路径 s 的最大势能增量，随着 Δ 的增加，跨越局部峰的概率呈指数级下降。因此，PPHMC 从尖峰分布和多峰分布中采样会停滞在某些局部峰值，无法快速收敛到正确的分布。

3.2 算法优化实现

在不增加额外计算成本的情况下，HMC 算法中实现粒子峰值跨越常见的启发式策略是提高系统的初始动能，模拟的粒子具备更高的初始动能，有可能会跨越高势能障碍。一种直接有效的提高动能的方法是在轨迹前半部分的每个跃迁步骤中，动量都乘以某个略大于 1 的因子（例如 ξ），而在轨迹的后半部分，每个跃迁步骤中动量再除以相同的因子 ξ，在轨迹的前半段，系统能量逐渐增加，有可能跨越包含马尔可夫链当前状态的函数 $U(x)$ 的局部山峰。而在后半段能量减小，粒子最终可能处于一个与初始不同的状态中，轨迹的终点被认为是马尔可夫链的下一个状态的候选点[11]。需要注意因子 ξ 不宜过大，过大的 ξ 会导致系统能量最终大于初始状态。

在正则分布 $P(\theta,p)$ 中利用哈密顿运动方程的特性，可以将系统温度、时间等变量的改变进行等值转换，转换为改变粒子质量、模拟步长等参数，例如：改变 $T=1$ 为 $T=\xi$，等价于 $\widetilde{U}(x):=\xi^{-1}U(x)$，根据哈密顿运动方程：

$$\begin{aligned} \frac{\mathrm{d}\theta}{\mathrm{d}t}&=v\\ \frac{\mathrm{d}v}{\mathrm{d}t}&=-M^{-1}\frac{\partial\widetilde{U}}{\partial\theta}=-(\alpha M)^{-1}\frac{\partial U}{\partial\theta} \end{aligned} \tag{8}$$

方程具有$(M,\widetilde{T}=\xi)\leftrightarrow(\widetilde{M}=\xi M,T=1)$的等值变换。因此前面所述的动量直接伸缩变化可以转化为粒子质量、模拟步长的伸缩变换。针对这种变换方法，Park 开发了一种更为有效的参数调整策略[25]。这种调整策略只适用于局部的系统发育树空间，在不同的 orthants 中后验分布可能大不相同，算法无法保证局部梯度可以提供正确的全局树修改建议，从整体上看，改变系统能量无法实现树拓扑峰值间的跨越。两个高似然率树拓扑之间很容易被较低似然率的其他树拓扑隔开，后验分布尖峰与相应较长的深谷会导致邻域空间较小的树拓扑分支交换策略无法越过低概率区域，可能停滞在树拓扑某些局部峰值上[26]。

除了调整系统能量，在 HMC 组件对树拓扑枝长变量采样的同时，MPHMC 算法加入非 HMC 更新，对树拓扑离散参数单独采样，引入更为大胆的树拓扑重排建议来加速树空间探索，如随机子树剪接算法 rSPR（random-Subtree Prune and Regraft）、SPR 策略和 MCMC 中其他大多数树拓扑变化策略（如 NNI）之间都存在对应关系。如图 1 所示，假设(π^0,b^0)为采样过程中某次迭代初始时刻的粒子状态，蓝色虚线表示 HMC 组件 leapfrog 更新计算的确定性路径，分支长度的连续变化产生拓扑变化的提议，两次穿越 orthant 边界需要能量补偿，耗费两次额外的能量评估。假设路径终端粒子的新状态为(π^*,b^*)，此时局部梯度信息无法获取跨越树拓扑峰值的提议；橙色虚线表示对树拓扑离散参数的 MH 更新，在路径终端树拓扑相邻的空间进行探索，生成更多的邻居加速树空间探索，获取其他模式的树拓扑提议。这种分裂非单一的混合路径有助于减少 HMC 组件承载的额外计算负担，以低计算成本更快地探索目标分布，即在保持所有变量的期望联合分布不变的同时[11,27]，涉及随机分支重排提议的组合策略大大减少了树空间 X 中采样路径直接穿越复合体边界的次数，有效遏制了计算时间的巨幅增加。否则，依赖于每次迭代较长的单一概率路径，会耗费比 RWM 算法更大的计算代价，还会导致模拟哈密顿系统较大的离散化误差[10]，获得错误的后验分布典型集。

MPHMC 树采样方法的相关核心函数伪代码如算法 1 所示。算法 1 的思想类似于平行回火算法，构建了平行链以增加局部峰值之间的跳跃机会[28]。但平行回火算法依赖于多个 HMC 采样器并行工作，每条链都针对不同的回火分布，

其中一个取样器从目标概率密度中取样(冷链),两个相邻的链试图交换它们当前的状态,这种交换可以根据 MH 标准被概率性地接受或拒绝,每条链都需要相同的计算量,只能在并行机器或网络空间站上实现,否则将耗费大量额外计算时间。而 MPHMC 方法额外消耗较少的计算资源,就可以在一定程度上消除随机游走的同时,也有助于避免采样器在局部峰停滞过长时间。

算法 1 系统发育树空间混合采样 $\pi(\tau,b,p)\propto exp(-U(\tau,b|D,\theta_0))$

```
1. def MPHMC(τ,b,ϵ,L|U)
2.     τ←RWM_rSPR(τ,L|U,b);
3.     p0～N(0,Id);
4.     τ0←τ;b0←b;
5.     for i←1 to L do
6.       τ,b,p← LeapFrog_Refractive(τ,b,p,ϵ|U);
7.     ϵ end
8.     τ,b← MH_correction(τ,τ0,b,b0,p,p0|U,K);
9.     ϵ returnτ,b;
10. end
11. def RWM_rSPR(τ,L|U,b)
12.    for j←1 to L do
13.      τ1←rSPR(τ);
14.      τ← MH_correction(τ1,τ,b,b,0,0|U);
15.    end
16.    return τ;
17. end
18. def LeapFrog_Refractive(τ,b,p,? |U)
19.    p←p←ϵ∇U(τ,t)/2;
20.    b1←b+ϵ p;
21.    if 树分支跨越边界的节点集合 X(τ,b1,p)≠? then
22.      t←0;
23.      while X≠? do
24.        //检测每次树分支跨越边界事件
25.        //将步长分割为更小的时间片te
26.        //依次更新每个零长度分支
27. t←t+te;
28. b←b+(ϵ−t)p;
29. τ1,b← NNI_UpdateEvent(τ,b,p,te);
30. pe←−pe;
31. ΔE←U(τ1,b)−U(τ,b)
32.        if ‖pe‖²>2ΔE then
33. pe← √(‖pe‖²−2ΔE);
34. τ←τ1;
35.        end
36.      end
37.      p←p−ϵ∇U(τ,b)/2;
38.      return τ,b,p;
39. end
40. def MH_correction(τ,τ0,b,b0,p,p0|U,K)
41.    E0←U(τ0,b0)+K(p0);
42.    E←U(τ,b)+K(p);
43.    if min(0,log(Uniform(0,1)))≥−E+E0 then
44.      τ,b←τ0,b0;
45.    return τ,b;
46. end
```

4 实验

4.1 实验数据

本文重点在经验数据集上比较 3 种采样器的表现(Mrbayes-RWM,PPHMC,MPHMC)和分析汇总结果,因为经验数据的树重建更具挑战性和一般性[29-30]。实验使用了 5 组被广泛验证的核苷酸经验数据集,数据特征如表 1 所列。

表 1 dataset1－dataset5 不同的数据特征

Table 1 Different data characteristics of dataset1－dataset5

	序列名称	分类群	位点数
dataset1	Trichophora 18S rDNA	26	1546
dataset2	Euhemiptera 18S rDNA	33	2238
dataset3	metazoan 18S rDNA	111	1506
dataset4	eukaryotic 18S rDNA	234	1790
dataset5	Bacteria-Archaea-Eukaryota 23-28S rDNA	288	3386

这些数据集的具体信息可在文献[31-33]中查阅,并可从 TreeBASE[1)] 中获取。在 dataset1－dataset5 上进行实验,结果证明 MPHMC 能获取到正确的典型集(表),确保 MPHMC 采样的正确性并获取良好的发育树,然后重点是在采样困难的数据集 dataset3－dataset5 上,比较 3 种采样器(Mrbayes-RWM,PPHMC,MPHMC)的收敛速度。

4.2 实验内容

实验目的是评估算法对发育树的采样能力,而不是从这些经验数据集中推断出完美的系统发育树,因此实验中只考虑用固定的、较少的代数(ngen)。所有实验不使用 Metropolis 耦合[5,34],只运行单链模式;核苷酸置换模型统一选择 JC69 模型,并且忽略不同位点间的比率变化,这样的设定消除了一些置换模型参数,使后验分布更难以采样[35]。除此之外,我们还尝试消除其他影响因素,以便专注于树空间采样的性能,如参照 Mrbayes3.1.2 的默认设置,在树拓扑上采用无限制的均匀先验、不考虑树拓扑上分支间的进化速率变化、在分支长度上采用无约束的指数先验、汇总结果时将前 25% 的样本作为"老化"丢弃等。实验中假设 Mrbayes 程序的运行结果是正确可靠的,并将其作为标准。

为公平地对比性能,实验中将 PPHMC 和 MPHMC 的采样代数转换为树拓扑对数似然概率的计算次数(LLPs),LLPs 计算次数与 Mrbayes 程序的采样代数基本一致。PPHMC 和 MPHMC 以 Mrbayes 中采用的随机初始树作为采样起点。超参数 L 和 ϵ 的设定会影响算法的采样效率,L 太大会使探索路径回环,导致大量无效的 LLPs 计算,ϵ 太大会使离散化步骤误差影响采样器失效[9],在这里统一设置步长 $\epsilon\approx0.001$ 和步数 $L\approx20$(固定分布中随机选择),不详细讨论 ϵ 和 L 对树空间采样效率的影响。此外,MPHMC 同样参考 Dinh 等对平滑阈值的设置,$delta=0.001$。

4.3 结果分析

本文使用 Tracer v1.7.2[36],在 dataset1－dataset5 上对 3 种贝叶斯系统发育程序的 MCMC 运行结果进行了跟踪分析(见表 2－表 5)。

1) http://www.treebase.org:在线的系统发育数据存储库

表 2 在 5 组数据集上后验分布的对数似然平均值对比

Table 2 Comparison of log-likelihood means of PDs on five datasets

Methods	Log-Likelihood(LL)				
	DS1	DS2	DS3	DS4	DS5
MB-RWM	−7962.62(±5.28)	−18913.28(±6.36)	−56191.66(±64.27)	−183513.28(±920.19)	−424759.78(±20.17)
PPHMC	−8047.02(±5.92)	−19049.70(±23.29)	−60801.50(±280.10)	−214909.66(±489.51)	−500428.47(±489.51)
MPHMC	**−7961.76(±5.28)**	**−18916.45(±6.22)**	**−56014.48(±63.06)**	**−182380.47(±71.05)**	**−421753.18(±195.32)**

表 3 在 5 组数据集上后验分布的树分支总长平均值

Table 3 Means of tree branch lengths of PDs on five datasets

Methods	the tree length(TL)				
	DS1	DS2	DS3	DS4	DS5
MB-RWM	0.676(±0.021)	1.815(±0.033)	8.200(±0.077)	30.584(±0.566)	42.364(±0.137)
PPHMC	0.694(±0.022)	1.848(±0.032)	9.181(±0.105)	41.341(±0.305)	57.243(±0.269)
MPHMC	**0.676(±0.021)**	**1.816(±0.034)**	**8.163(±0.079)**	**29.787(±0.152)**	**41.370(±0.165)**

表 4 在 5 组数据集上 MCMC 运行的有效样本比例

Table 4 Proportion of valid samples for MCMC runs on five datasets

Methods	ESS per sample of TL				
	DS1	DS2	DS3	DS4	DS5
MB-RWM	1.07×10^{-1}	7.50×10^{-2}	2.00×10^{-3}	3.00×10^{-4}	3.00×10^{-3}
PPHMC	2.87×10^{-1}	1.23×10^{-1}	1.00×10^{-3}	1.00×10^{-3}	1.00×10^{-3}
MPHMC	$\mathbf{2.23\times10^{-1}}$	$\mathbf{1.30\times10^{-1}}$	$\mathbf{8.50\times10^{-2}}$	$\mathbf{8.90\times10^{-2}}$	$\mathbf{1.00\times10^{-3}}$

表 5 Mrbayes 在 5 组数据集上分别与 PPHMC，MPHMC 的总结树(MCCT)的距离比较

Table 5 Comparison of Mrbayes' tree distances to the MCCTs of PHMC and MPHMC

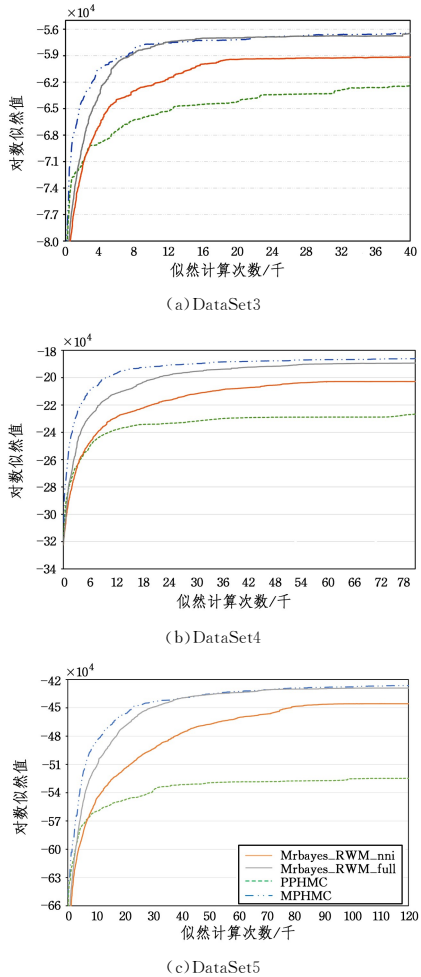
	PPHMC		MPHMC	
	RF	Quartet	RF	Quartet
DS1	0.256	0.222	**0.145**	**0.000**
DS2	0.513	0.032	**0.467**	**0.002**
DS3	4.332	0.443	**2.739**	**0.216**
DS4	27.727	0.584	**12.374**	**0.196**
DS5	37.920	0.619	**14.019**	**0.154**

表 2 列出了 Mrbayes-RWM，PPHMC，MPHMC 在 5 组数据集上 MCMC 模拟后验分布的对数似然平均值。表 3 列出了 Mrbayes-RWM，PPHMC，MPHMC 在 5 组数据集上 MCMC 模拟后验分布的树分支总长平均值. 表 4 列出了 Mrbayes-RWM，PPHMC，MPHMC 在 5 组数据集上 MCMC 运行的有效样本比例。表 2、表 3 中，括号中的数值均为标准差。

除此之外，还将 PPHMC，MPHMC 与 Mrbayes 程序运行结果的树集总结成一棵最大进化枝可信度树(MCCT)：进化枝支持度的乘积最大化树，用对进化枝的支持度(分支后验概率)表示分支标签，用所有树分支的平均值表示分支长度[37]。并且用一些广泛使用的树距离(例如 Robinson-Foulds(RF)距离、Quartet 距离等)来对结果进行进一步分析(见表 5)，这些距离在树的构建中具有较大的价值空间[38-39]。总体而言，表 2－表 5 的数据表明，MPHMC 和 MrBayes 收敛到相似的后验分布。

4.4 对比评估

本文主要在数据集 dataset3－dataset5 上对 3 种采样器的混合速度进行评估(见图 2)，因为在较为简单的小数据集 dataset1－dataset2 上，RWM 和 HMC 方法在树空间中的探索效率差异不大。根据树拓扑搜索策略，将 Mrbayes 程序设定为两种形态，full 指采用了 Local，TBR，NNI 等建议来加快 Mrbayes 程序对较大发育树空间的探索，与此相对 nni 是只设定了 NNI 的树变换。



(a) DataSet3

(b) DataSet4

(c) DataSet5

图 2 Mrbayes-RWM，PPHMC，MPHMC 方法在数据集 dataset3－dataset5 上运行的似然计算次数与对数似然值关系

Fig. 2 Number of likelihood calculations versus log-likelihood values for Mrbayes-RWM，PPHMC，MPHMC methods run on datasets 3－datasets5

从图 2 中可以观察到，在所有数据集上，MPHMC 采样器的混合速度均最快，其表现明显优于 PPHMC 和 Mrbayes_nni 方法，同时在数据集 dataset4 上的速度明显快于全树拓扑搜索策略的 Mrbayes 程序。总体而言，MPHMC 比传统的 RWM 方法能更快、更有效地探索树空间，具有巨大的开发潜力。

结束语 本文针对 PPHMC 算法在较长分子序列或多分类群数目的经验数据集上采样性能低的问题，提出了 MPHMC 方法，混合对参数子集的 HMC 更新与其他 MH 更新，类似于平行回火算法，构建了平行链，增加了后验分布空间中局部峰值之间的跳跃机会。与独立设置相比，这种跳跃可以使粒子在状态空间中更自由，可以高效、正确地获取后验分布典型集，自然地扩展了 Dinh 等提出的树空间哈密顿蒙特卡洛采样。结果表明，系统发育分析中使用 MPHMC 树采样算法，可以实现更好的混合和更快的收敛。

未来仍然有大量的工作需要去做，包括对算法中能量函数梯度计算部分的进一步优化，可借助并行计算缩短运行时间，将算法扩展到更复杂的分子进化模型和树模型，或者在其他更灵活的树描述空间中实现高效的 HMC 采样算法等。

参 考 文 献

[1] KAPLI P, YANG Z, TELFORD M J. Phylogenetic tree building in the genomic age[J]. Nature Reviews Genetics, 2020, 21(7): 428-444.

[2] YANG Z. Computational molecular evolution[M]. Oxford: Oxford University Press, 2006.

[3] HUELSENBECK J P, RONQUIST F. MRBAYES: Bayesian inference of phylogenetic trees[J]. Bioinformatics, 2001, 17(8): 754-755.

[4] RONQUIST F, HUELSENBECK J P. MrBayes 3: Bayesian phylogenetic inference under mixed models[J]. Bioinformatics, 2003, 19(12): 1572-1574.

[5] 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.

[6] DRUMMOND A J, RAMBAUT A. BEAST: Bayesian evolutionary analysis by sampling trees[J]. BMC Evolutionary Biology, 2007, 7(1): 1-8.

[7] METROPOLIS N, ROSENBLUTH A W, ROSENBLUTH M N, et al. Equation of state calculations by fast computing machines[J]. The Journal of Chemical Physics, 1953, 21(6): 1087-1092.

[8] HASTINGS W K. Monte Carlo sampling methods using Markov chains and their applications[J]. Biometrika, 1970, 57(1): 97-109.

[9] NEAL R M. MCMC using Hamiltonian dynamics[J]. Handbook of Markov Chain Monte Carlo, 2011, 2(11): 2.

[10] DINH V, BILGE A, ZHANG C, et al. Probabilistic path hamiltonian monte carlo[C]// International Conference on Machine Learning. PMLR, 2017: 1009-1018.

[11] BILLERA L J, HOLMES S P, VOGTMANN K. Geometry of the space of phylogenetic trees[J]. Advances in Applied Mathematics, 2001, 27(4): 733-767.

[12] JUKES T H, CANTOR C R. Evolution of protein molecules[J]. Mammalian Protein Metabolism, 1969, 3: 21-132.

[13] YANG Z. Estimating the pattern of nucleotide substitution[J]. Journal of Molecular Evolution, 1994, 39(1): 105-111.

[14] MAU B, NEWTON M A. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo[J]. Journal of Computational and Graphical Statistics, 1997, 6(1): 122-131.

[15] YANG Z, RANNALA B. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method[J]. Molecular Biology and Evolution, 1997, 14(7): 717-724.

[16] BROMHAM L, DUCHÊNE S, HUA X, et al. Bayesian molecular dating: opening up the black box[J]. Biological Reviews, 2018, 93(2): 1165-1191.

[17] YANG Z, RANNALA B. Molecular phylogenetics: principles and practice[J]. Nature Reviews Genetics, 2012, 13(5): 303-314.

[18] BETANCOURT M. A conceptual introduction to Hamiltonian Monte Carlo[J]. arXiv: 1701. 02434, 2017.

[19] AFSHAR H M, DOMKE J. Reflection, Refraction, and Hamiltonian Monte Carlo[C]// NIPS. 2015: 3007-3015.

[20] FELSENSTEIN J. Evolutionary trees from DNA sequences: a maximum likelihood approach[J]. Journal of molecular evolution, 1981, 17(6): 368-376.

[21] JI X, ZHANG Z, HOLBROOK A, et al. Gradients do grow on trees: a linear-time O(N)-dimensional gradient for statistical phylogenetics[J]. Molecular Biology and Evolution, 2020, 37(10): 3047-3060.

[22] BOU-RABEE N, SANZ-SERNA J M. Geometric integrators and the Hamiltonian Monte Carlo method[J]. Acta Numerica, 2018, 27: 113-206.

[23] WHIDDEN C, MATSEN IV F A. Quantifying MCMC exploration of phylogenetic tree space[J]. Systematic Biology, 2015, 64(3): 472-491.

[24] MANGOUBI O, PILLAI N S, SMITH A. Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities? [J]. arXiv: 1808. 03230, 2018.

[25] PARK J. Sampling from multimodal distributions using tempered Hamiltonian transitions[J]. arXiv: 2111. 06871, 2021.

[26] LAKNER C, VAN DER MARK P, HUELSENBECK J P, et al. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics[J]. Systematic Biology, 2008, 57(1): 86-103.

[27] ZHOU G. Metropolis Augmented Hamiltonian Monte Carlo[J]. arXiv: 2201. 08044, 2022.

[28] SUN S, SHEN Y. "Parallel-Tempering"-Assisted Hybrid Monte Carlo Algorithm for Bayesian Inference in Dynamical Systems[C]// UK Workshop on Computational Intelligence. Cham: Springer, 2019: 357-368.

[29] HUELSENBECK J P. Performance of phylogenetic methods in simulation[J]. Systematic Biology, 1995, 44(1): 17-48.

[30] NEI M, TAKEZAKI N, SITNIKOVA T. Assessing molecular phylogenies[J]. Science, 1995, 267(5195): 253-255.

[31] XIE Q, BU W, ZHENG L. The Bayesian phylogenetic analysis

of the 18S rRNA sequences from the main lineages of Trichophora(Insecta:Heteroptera:Pentatomomorpha)[J]. Molecular Phylogenetics and Evolution,2005,34(2):448-451.

[32] XIE Q,TIAN Y,ZHENG L,et al. 18S rRNA hyper-elongation and the phylogeny of Euhemiptera(Insecta:Hemiptera)[J]. Molecular Phylogenetics and Evolution,2008,47(2):463-471.

[33] XIE Q,WANG Y,LIN J,et al. Potential key bases of ribosomal RNA to kingdom-specific spectra of antibiotic susceptibility and the possible archaeal origin of eukaryotes[J]. PloS one,2012,7(1):e29468.

[34] GEYER C J. Markov Chain Monte Carlo Maximum Likelihood [M]. American Cancer Society,2005.

[35] NYLANDER J A A,FREDRIK R,HUELSENBECK J P,et al. Bayesian Phylogenetic Analysis of Combined Data[J]. Systematic Biology,2004,53(1):47-67.

[36] RAMBAUT A,DRUMMOND A J,XIE D,et al. Posterior summarization in Bayesian phylogenetics using Tracer 1.7[J]. Systematic Biology,2018,67(5):901-904.

[37] SUKUMARAN J,HOLDER M T. DendroPy:a Python library for phylogenetic computing[J]. Bioinformatics,2010,26(12):1569-1571.

[38] HUERTA-CEPAS J,SERRA F,BORK P. ETE 3:reconstruction,analysis,and visualization of phylogenomic data[J]. Molecular Biology and Evolution,2016,33(6):1635-1638.

[39] SMITH M R. Robust analysis of phylogenetic tree space[J]. Systematic Biology,2022,71(5):1255-1270.

LI Xiaopeng, born in 1998, postgraduate. His main research interest is computational phylogenetics.

GAO Jingyang, born in 1966, Ph.D, professor, Ph.D supervisor. Her main research interests include artificial intelligence and bioinformatics.

(责任编辑:喻藜)