专题: 生物分子模拟中的机器学习

生物大分子过渡态搜索算法及其中的机器学习*

杨建宇# 席昆# 竺立哲†

(香港中文大学 (深圳) 医学院, 瓦谢尔计算生物研究院, 深圳 518172)

(2023 年 8 月 13 日收到; 2023 年 9 月 9 日收到修改稿)

过渡态是物理化学家理解和调控生物大分子相关功能微观机制的关键. 因其存在时间极短, 难以被实验手段捕捉, 全面刻画其结构必须通过物理定律驱动的模拟计算搜索予以实现. 然而, 与化学反应过程只涉及少量原子不同, 生物大分子的功能性构象变化所涉的原子和坐标数量巨大, 搜索其过渡态将不可避免地遭遇维数灾难, 即反应坐标问题, 因而催生了多种应对策略和算法. 同时, 随着近年来新型机器学习算法的大量涌现和日臻成熟, 融入机器学习范式的过渡态搜索算法也已出现. 本文首先回顾和梳理过渡态搜索代表性算法的设计思想, 包括依赖集合变量的温和爬升动力学 (gentlest ascent dynamics, GAD)、有限温度弦方法 (finite temperature string, FTS)、快速断层扫描法 (fast tomographic)、基于旅行商的自动路径搜索算法 TAPS, 以及过渡路径采样法 (transition path sampling, TPS). 然后, 重点介绍 TPS 与强化学习融合而成的新型路径采样算法, 解析强化学习在其中的作用, 并厘清其适用场景. 最后, 我们提出一种将降维算法与 GAD 深度融合的新构想, 讨论研发可保留过渡态信息的新型降维算法的必要性及可行性.

关键词: 过渡态搜索, 温和爬升动力学, 路径算法, 强化学习, 生成模型

PACS: 87.10.Tf, 87.15.A–, 87.15.H–, 87.15.hp **DOI:** 10.7498/aps.72.20231319

1 引 言

生物分子实现功能时, 常伴随着结构的巨大转变, 即生物分子的功能性构象变化[1–3]. 利用实验方法, 往往只能获取上述转变过程前后重要的稳态结构, 如 X 射线 (X-ray macromolecular crystallography)[4]、核磁共振 (nuclear magnetic resonance, NMR)[5]、冷冻电子显微镜 (cryo-electron microscopy, cryo-EM)[6] 等; 或者揭示分子结构变化中的部分特征, 如荧光共振能量转移 (fluorescence resonance energy transfer, FRET) 可给出少数目标残基间的距离变化[7] 等. 因此, 仅依赖实验方法难以阐明生物分子转变过程的完整信息.

全原子 (all-atom) 分子动力学 (molecular dynamics, MD) 是从原子尺度全面描述生物分子动态行为的标准手段[8]. 但和化学反应仅涉及反应活性中心内的数十个原子不同, 构象变化所涉及的原子数目巨大, 极端情况下可包括溶质的全部原子, 甚至环境中脂类和溶剂分子的原子[9–36]. 众多的原子及其三维坐标带来了两个重要的瓶颈.

首先, 在计算效率方面, 复杂大分子百万级的原子数量意味着需要计算万亿级数量的原子间作用力, 即使在目前最优的通用硬件上, 人们所能完成的 MD 模拟时长也仅在微秒量级[8,37], 距离生物分子的实际功能性动力学行为毫秒级的发生时间仍有巨大差距. 为缓解该效率瓶颈, 数十年来, 人们发展了各类增强采样算法, 其中较有代表性的算法

* 国家自然科学基金 (批准号: 31971179) 和深圳市科技创新委员会 (批准号: JCYJ20200109150003938, RCYX20200714114645019) 资助的课题.

同等贡献作者.

† 通信作者. E-mail: zhulizhe@cuhk.edu.cn

包括副本交换[38–45], 选择性温度积分增强采样 (selective integrated tempering sampling)[46–49]、局部抬升 (local elevation)[50–53]、构象洪泛 (conformational flooding)[54–56]、元动力学 (metadynamics)[57–59]、高斯加速动力学[60–62] 等.

更为重要的是, 在数据分析层面, 尤其是在提取过渡态信息这类理论化学家最关心的问题上, 巨大的原子数量导致了维数灾难. 搜寻过渡态的结构或特征信息是准确刻画和解释所采样本中动力学机制的重中之重. 然而, 即使是在采样数据充足的情况下, 使用不恰当的分析手段 (即机器学习语境下的降维算法), 过渡态区域都将被扭曲以致相关信息丢失.

在已有大量模拟数据的场景中, 可借助 tICA (time-lagged independent component analysis)[63–65] 利用已有数据中蕴含的动力学信息进行降维, 或运用马尔可夫态模型 (Markov state models)[66–78] 等分析算法提取动力学信息来应对维数灾难, 并间接推测过渡态信息. 但这类算法中并不直接含有过渡态的定义, 因而超出了本文范畴. 对此类算法感兴趣的读者可参看其他综述[63,66,68,75–78].

在生物大分子模拟领域, 因其计算效率低下, 数据匮乏是常态, 因此人们对能高效搜寻过渡态的采样算法需求强烈. 但受限于维数灾难, 仅有以下两类采样策略可供选择.

1) 依赖 CV 的定向降维. 在不具备先验数据时, 依据直觉猜测少量有物理意义且可能重要的坐标, 即集合变量 (collective variable, CV), 强行定向降维到该预选的低维 CV 空间, 而后在 CV 空间内搜寻过渡态[79–95]. 代表性方法: 温和爬升动力学 (gentlest ascent dynamics, GAD)[79–81]、有限温度弦方法 (finite temperature string, FTS)[82–87]、快速断层扫描法 (fast tomographic, FT)[88–90]、基于旅行商的路径搜索 (travelling-salesman based automated path searching, TAPS)[91–95].

2) 非 CV 依赖的高维搜索. 事先不降维, 坚持在高维空间内完成采样和过渡态搜索过程, 事后再进行降维分析[96–101]. 代表性方法有过渡路径采样 (transition path sampling, TPS)[98–101].

尽管上述算法已在一定范围内取得成功, 但在面对复杂生物分子时, 仍面临诸多限制. 其中, 对于依赖 CV 的搜索算法, 最直接的问题便是如何从较高维度空间中选取合适的 CV; 而对于非 CV 依赖的路径采样算法, 则是计算资源消耗过大和有效采样率过低的问题.

近年来快速发展的机器学习及相关衍生算法 (如强化学习、生成式建模等), 已成功应用于解决诸多传统的复杂生物问题[102–112], 如生物结构预测及生物分子相互作用的研究[105], 或基于人工智能开发蛋白质从头设计算法[106], 或借助于机器学习实现蛋白质结构准确预测的 trRosetta 线上服务[107], 或实现生物分子冷冻电镜高分辨率结构重建的解析算法[108] 和蛋白质间相互作用位点的快速预测[109], 以及蛋白质与小分子、RNA 等复合物结构性质的预测[110,111]. 因此, 将机器学习与现有过渡态搜索算法进行有效融合, 有望成为未来过渡态搜索研究实现进一步突破的可行方向.

本文将首先回顾依赖 CV 的过渡态搜索算法的发展历程, 厘清其基本原理及潜存问题. 随后, 聚焦于非 CV 依赖的 TPS 路径采样算法, 着重介绍其融合了强化学习的最新版本. 最后, 探讨一种新型的过渡态搜索策略, 即结合生成模型和 GAD, 在保留原高维空间过渡态信息的低维空间内实现过渡态搜索. 完整的算法总结已展示于表 1 中.

2 依赖 CV 的过渡态搜索算法

如前所述, 为了准确阐明生物分子功能性动力学的微观机制, 需要在传统采样算法的基础上, 发展可获取上述转变过程过渡态信息的过渡态搜索算法, 包括依赖 CV[82–95] 和非 CV 依赖算法[96–101] 两大类. 对于依赖 CV 的算法, 需在缺乏对体系的先验数据和认知的条件下, 将高维相空间$\{x\}$“定向降维”至少量的依据经验或直觉定义的 CV 上 (arbitrary guess). 而后续的计算采样和过渡态搜索则发生在由这些 CV 构成的低维空间 (CV1, CV2, ···) 内 (图 1(a)).

低维 CV 空间中的过渡态搜索, 依照采样开始时的已知信息可分为非路径算法和路径算法. 非路径算法以 GAD 算法为代表, 而路径算法以 finite temperature string[82–87] 和快速断层扫描法[88–90] 为代表. 前者可在仅有一个稳定态已知时开启过渡态搜索, 而后者需事先已知至少两个稳定态, 通过寻找两个稳定态之间的最小自由能路径 (minimum free energy path, MFEP), 而后获得沿路径的自由能分布确定过渡态位置. 此外, 两者的区别

表 1 主要过渡态搜索算法的总结分类

Table 1. Classification of the algorithms for transition state searching.

过渡态搜索算法分类		代表性算法	参考文献	备注
传统方法	依赖CV	Gentlest ascent dynamics (GAD)	[79—81]	非路径方法
		Finite temperature string	[82—87]	路径方法
	预设低维	Fast tomographic	[88—90]	
	空间搜索	基于旅行商的路径搜索 TAPS	[91—95]	
融合AI	不依赖CV 高维空间搜索	Transition path sampling	[98—101]	
		Reinforcement path sampling	[113]	
	保留过渡态 信息的降维 低维空间搜索	融合生成模型及GAD 的过渡态搜索(待研发)	无	非路径方法

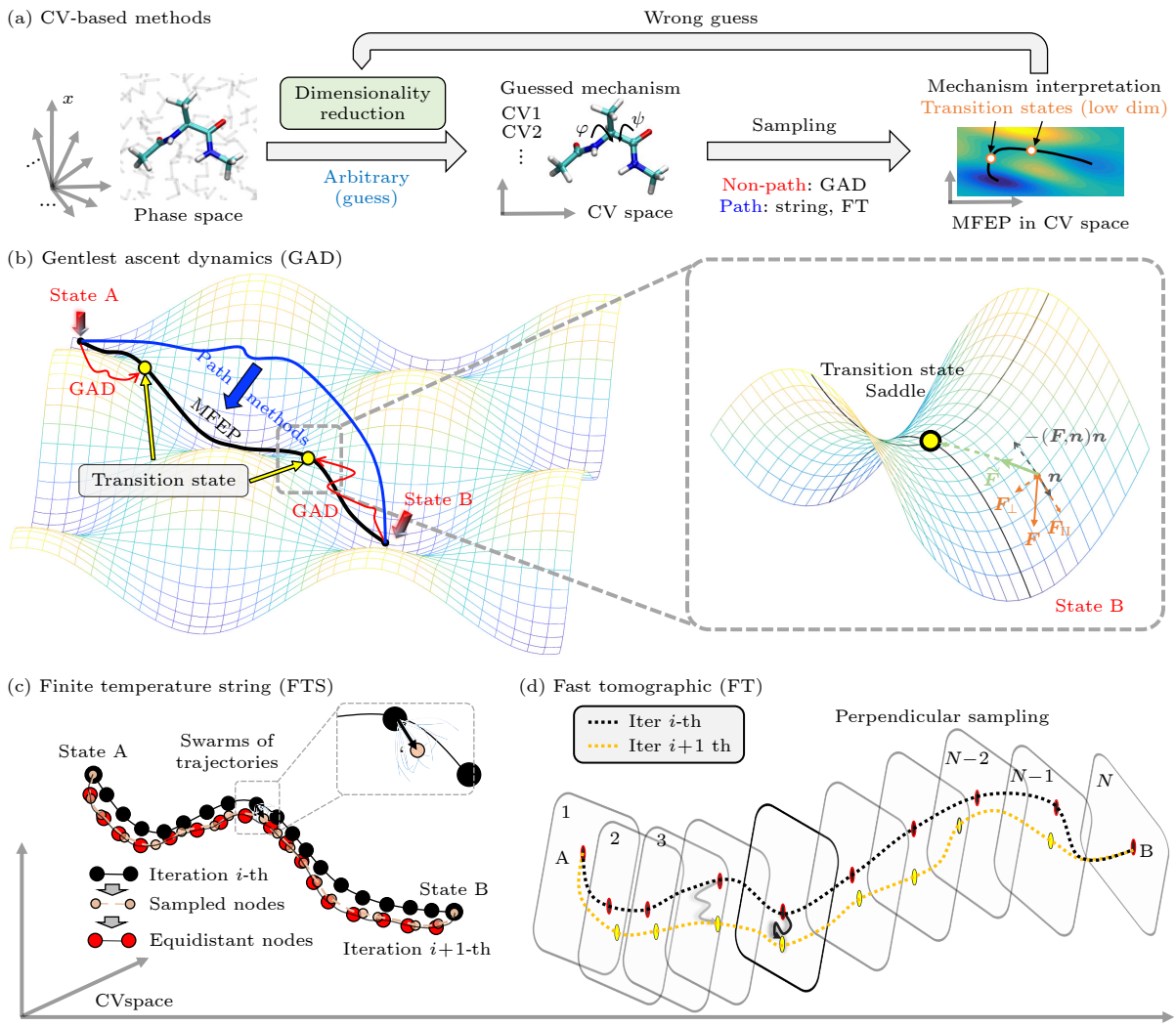

图 1 (a) 依赖集合变量的过渡态搜索示意图, 需由生物分子 (以丙戊酸二肽为例) 体系所在的高维相空间 (phase space) 选取少量集合变量 CV 强行“定向降维”, 后在此低维 CV 空间利用非路径类方法或路径方法, 找到过渡态 (Transition State), 并给出微观机制解释 (mechanism interpretation); (b) 非路径类的 GAD 算法原理示意图; (c), (d) 两类路径类搜索算法原理示意图

Fig. 1. (a) Illustration of the flow-chart of the collective variables (CVs) based transition state searching. A low dimensional space must be constructed with the CVs, which are arbitrary a priori guess about the mechanism. The transition state(s) is then determined by either the non-path or path methods. (b) The non-path method GAD. Path methods of (c) finite temperature string and (d) fast tomographic.

还有, 前者采样过程是主动“爬山”(即向高能区域运动, 图 1(b) 左红), 而后者是先通过施加外力促使分子强行翻山越岭得到能量过高的初始路径 (图 1(b) 左蓝), 再设法使路径“整体下山”, 落入附近的最优路径 MFEP (图 1(b) 左黑).

2.1 非路径类过渡态搜索

GAD 是非路径类过渡态搜索的代表性算法, 在预设的低维 CV 空间, 从亚稳态或任意状态出发, 可在低维势能面空间内, 直接完成过渡态搜索[79–81]. 如图 1(b) 所示, 此算法的原理为由低维势能面空间内的任意一点出发, 根据以下规则:

$$\dot{\boldsymbol{x}} = \boldsymbol{F}(\boldsymbol{x}) - 2\widetilde{\boldsymbol{F}}, \tag{1a}$$

$$\gamma\dot{\boldsymbol{n}} = -H\boldsymbol{n} + (\boldsymbol{n},\ H\boldsymbol{n})\,\boldsymbol{n}, \tag{1b}$$

来确定每轮迭代时移动至下一步的位移方向, 即沿势能函数梯度变化率的最小方向进行小步长移动, 最终收敛于鞍点位置 (即过渡态). 其中 $\widetilde{\boldsymbol{F}} = (\boldsymbol{F}(\boldsymbol{x}), \boldsymbol{n})\,\boldsymbol{n}$, $\boldsymbol{F}(\boldsymbol{x})$ 为分子体系在根据当前低维 CV 空间内的势能梯度计算得到的作用力; 而 $\boldsymbol{n}$ 被设定为趋近于势能函数海森矩阵最小特征值对应的特征向量, 即指向曲率最小方向, 其需要基于 (1b) 式反复迭代达到收敛, 在此期间, γ 则控制 H 对 $\boldsymbol{n}$ 变化的影响能力, 以此消除势能函数中的噪音. 简单而言, (1) 式的规则将引导分子不断沿势能坡度最缓的方向逆势攀登, 直至收敛停滞于过渡态.

2.2 基于路径优化的过渡态搜索

对于基于路径优化进行过渡态搜索的算法, 根据其输入不同, 可主要分为两类: 1) 需要高质量预选集合变量 CV 的路径优化算法, 包括 finite temperature string[82–87] 和快速断层扫描法[88–90]; 2) 基于路径集合变量 (path collective variable, PCV) 的路径优化算法, 即基于 TAPS 算法[91–95], 此方法中避免了高质量预选集合变量的困境, 可高效且快速找到最优转变路径. 当构建完路径优化的低维空间后, 需要从目标系统的两个稳定态结构出发, 产生一条较为粗糙的转变路径[114–116], 而后对此路径进行迭代优化 (路径整体下山), 并最终收敛于最优路径 (MFEP)[82–95]; 继而便可通过计算 MFEP 的自由能图景, 准确给出微观转变机制和过渡态信息[57–59,117].

2.2.1 Finite Temperature String

当基于传统的增强采样算法 (如 steered MD, climber MD, targeted MD 等[114–116]) 快速得到描述目标生物分子过程的转变路径后, 前人发现还需要通过选取合适的集合变量信息, 来构建低维空间和完成对初始转变路径的进一步优化, 从而得到最优路径, 即最小自由能路径 (minimum free energy path, MFEP). 作为研究此类问题中的代表算法, finite temperature string 的优化策略[82–87] 较为简洁 (以 swarms-of-trajectories 版本为例[87]), 见图 1(c). 通过对连接转变路径 (由 State A 到 State B) 的所有节点, 依次分别完成大量 (swarms) 非常短时长的随机初始速率 MD 采样后, 在预选的低维空间对采样结果聚类, 找到出现概率最高的构象, 作为代表性的采样节点 (图 1(c) 中 sampled node). 这样做是为了在路径上各节点附近做非常局部的采样, 从而估计各节点目前所在位置的自由能梯度, 等效于让各节点沿着当前所在位置的自由能梯度最大方向稍作移动 (下山), 类似于势能最小化问题中的最速下降法; 通过再优化节点分布来保证相邻节点间距离相近 (equidistant nodes, 图 1(c)), 进而得到新一轮的转变路径.

通过不断重复上述迭代策略, 路径将最终收敛到达最小自由能路径 MFEP. 最终便可通过伞形采样等[117] 方法获取沿此 MFEP 的自由能景观 (free energy landscape)[82–87], 进而给出微观机制解释和得到相应的过渡态信息.

2.2.2 快速断层扫描法

快速断层扫描法与前述的 finite temperature string 方法较为相似, 亦需基于经验或随机预选取集合变量来构建低维空间[88–90], 而后在此低维空间进行路径搜索, 找到 MFEP, 如图 1(d) 所示:

首先, 在选定的低维度空间内, 均匀选取转变构象 (每个构象称为节点, 共 N 个节点) 来代表初始转变路径 (由 State A 到 State B); 随后, 对于每个节点, 都在垂直于当前路径的超平面空间内进行相同时长的 MD 模拟采样, 在采样过程中还需引入 SHAKE 算法[118] 以避免其离超平面空间过远, 同时, 结合自适应偏势 MD 方法 (adaptively biased molecular dynamics, ABMD)[119] 来提高其采样效率; 接着, 针对每个节点的采样轨迹, 直接将采样的终态结构进行连接, 保存为新的转变路径

(如图 1(d) 中黑色虚线代表的第 i 轮结果和黄色虚线代表的第 i+1 轮结果). 按照上述流程反复迭代, 将最终得到 MFEP, 及相应自由能景观分布, 从而阐明其微观转变机制并确定目标过渡态信息.

2.2.3 基于旅行商的自动化路径搜索算法

在基于集合变量的搜索算法中, 还存在一种基于路径集合变量 PCV 的新型算法[120], 即基于旅行商问题的自动路径搜索算法 (TAPS). TAPS 巧妙地避开了其他路径优化算法中集合变量的选取问题, 同时基于并行化和 GPU 加速, 快速得到较高维度空间中的最优路径 (MFEP), 给出相应的微观转变机制和过渡态信息 (图 2)[91–95].

具体来讲, 在使用 TAPS 方法时, 需提供目标生物系统的两个稳态结构和连接其转变过程的初始路径; 而后从初始路径中确定转变过程中变化较大的所有结构域, 并以这些结构域的重原子 (图 2(a) 中丙戊酸二肽结构中以球形显示的原子) 为参考, 通过计算构象间均方根位移偏差 (root mean square distance, RMSD) 来评估构象差异, 并从初始路径中在保证相邻构象间适度的差异基础上, 均匀选取构象 (即节点) 来代表整个转变过程; 接着, 基于此少量节点组成的转变路径, 便可利用 PCV 的计算公式得到二维的路径集合变量低维空间: 即 PCV-s 和 PCV-z. 其中, 对于任意构象 $\boldsymbol{x}$, 参照目标路径计算得到的 PCV-s 代表其沿路径

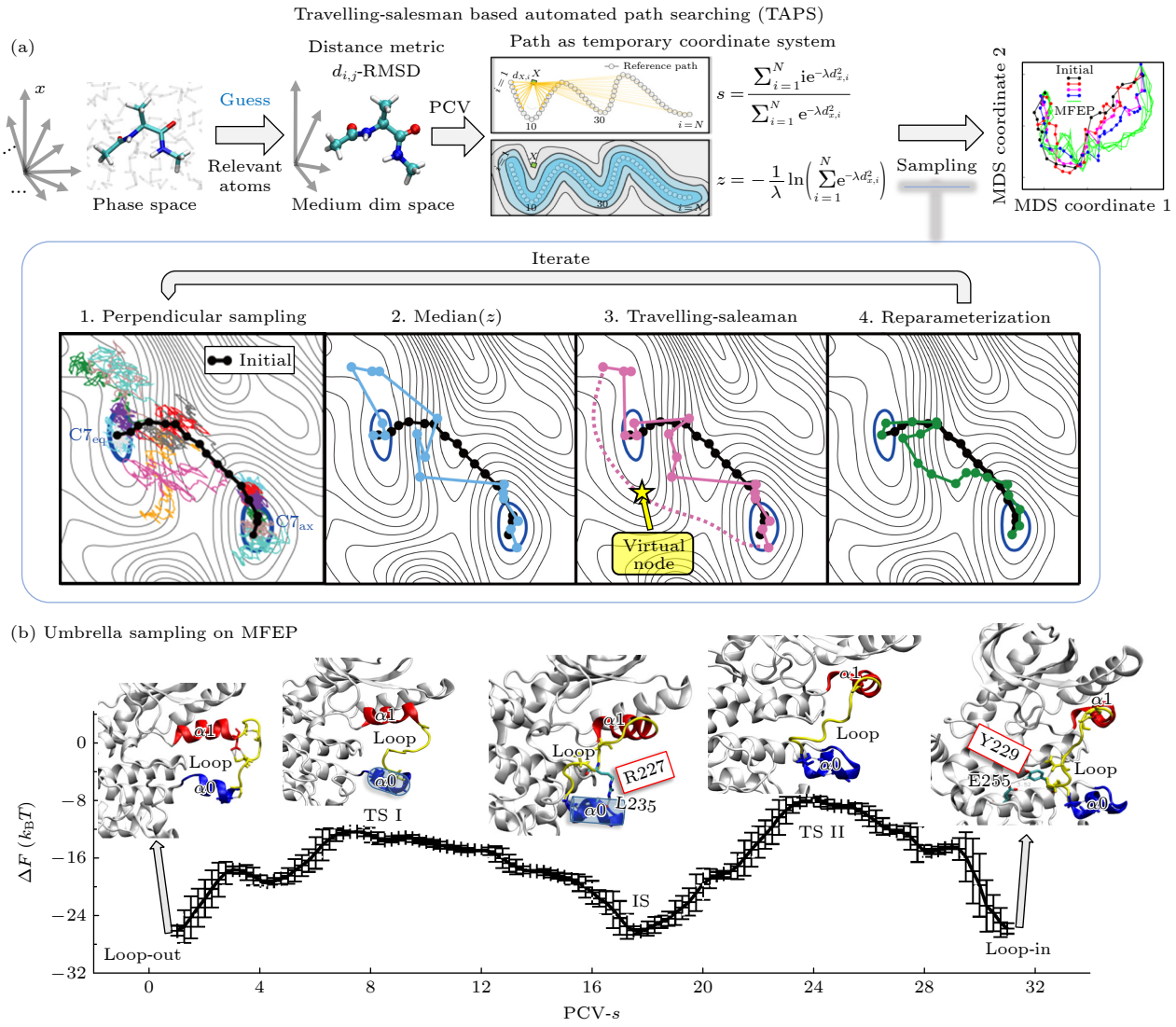

图 2 (a) PCV 构建[120] 和 TAPS Method[91–95,121] 算法原理示意图; (b) 基于伞形采样方法得到的 TAPS 算法确定的 MEK1 由 Loop-Out 到达 Loop-In 转变过程最小自由能路径 (MFEP) 的自由能图景及相应的微观转变机制[92]

Fig. 2. (a) Illustration for the construction of PCV and the flow-chart of the TAPS method; (b) TAPS revealed the free energy landscape and the transition states for the transition from the Loop-Out state of MEK1 to its Loop-In state[92].

方向的投影位置; 而 PCV-z 表示其距离参考路径的平均距离, 见图 2(a)[120]. 通过在此路径集合变量空间内, 快速完成路径搜索, 将最终确定目标转变过程的最优路径 (MFEP), 如图 2(a) 中基于多维度标度方法 (multidimensional scaling method, MDS)[122] 得到的二维路径搜索过程展示, 从黑色的初始路径快速搜索到达绿色的最优路径 (MFEP).

此处以丙戊酸二肽由 $C7_{eq}$ 到 $C7_{ax}$ 的转变为例, 完整展示 TAPS 进行路径优化的主要过程, 包括以下四步 (见图 2(a) 中下方白色框内的 TAPS 迭代流程).

步骤 1 基于转变路径节点间结构差异 ($\boldsymbol{d}_{\boldsymbol{x},i}$) 和节点编号 ($i = 1, 2, \cdots, N$) 信息, 利用 PCV[120] 构建路径优化的二维空间: 沿路径方向, PCV-s ((2a) 式) 和垂直于路径方向, PCV-z ((2b) 式), 而后从每个节点出发做采样, 采样时在 PCV-s 方向加入限制偏势, 阻止分子在平行于当前路径的方向运动, 但允许其在垂直于当前路径的超平面内任意运动; 同时, 为了后续步骤 4 补入节点时能有更多候选构象, 在 PCV-s 进行元动力学 (well-tempered metadynamics[123]) 采样.

$$s = \frac{\sum_{i=1}^{N} i\mathrm{e}^{-\lambda d_{\boldsymbol{x},i}^2}}{\sum_{i=1}^{N} \mathrm{e}^{-\lambda d_{\boldsymbol{x},i}^2}}, \tag{2a}$$

$$z = -\frac{1}{\lambda}\ln\left(\sum_{i=1}^{N} \mathrm{e}^{-\lambda d_{\boldsymbol{x},i}^2}\right), \tag{2b}$$

步骤 2 对于每个节点的采样轨迹, 通过获取最接近轨迹 PCV-z 中位值的结构, 并按照上轮编号连接为新的转变路径 (蓝色实线).

步骤 3 经步骤 1 非局部的垂直空间采样后, 节点顺序很可能已发生改变需要重排. 本算法将节点重排转化为旅行商问题[121], 并通过插入虚拟点 (即与其他任何节点间的距离为零) 来将旅行商问题的闭环解转化为节点顺序编号.

步骤 4 去除转变路径范围外节点, 并在距离较远的相邻节点间补入新节点.

最终, 通过不断重复迭代上述 1—4 步的路径优化过程, 将最终搜索到 MFEP 并结合伞形采样等算法[117] 得到沿 MFEP 的自由能景观分布, 进而给出微观转变机制解释和确定相应的过渡态信息.

以 TAPS 对丝裂原激活蛋白激酶激酶 (MEK1) 由 Loop-Out 状态转变为 Loop-In 状态的研究为例 (图 2(b)), 实验发现其在传递生物信号中时需经历 Loop-Out 态到 Loop-In 态的转变, 即两个 α 螺旋 (α0 和 α1) 的局部翻转以及连接螺旋的 Loop 进入激活口袋; 利用 TAPS 方法同时考察上述过程中涉及的所有重要残基, 在较短的采样总时间 (短于 32.6 ns) 内便得到了 MFEP(图 2(a) 最右侧的 MDS 结果内的绿色线)[92]; 沿收敛的 MFEP 进一步得到了相应的自由能图景 (图 2(b)), 进而获得了主要转变机制和两个关键过渡态结构 (TS Ⅰ和Ⅱ). 此研究所新发现的 R227:L235 及 Y229:E255 极性接触作用, 也被成功用于解释实验关于 R227 或 Y229 的点突变造成 MEK1 无法激活的现象[124,125].

尽管 TAPS 算法巧妙地规避了预选 CV 空间定向降维带来的试错成本, 但仍需选择计算 RMSD 所需的原子集作为输入信息. 这意味着在复杂大分子的过渡态搜索中, 即便 TAPS 的整体效率相比依赖 CV 的方法已有大幅提升, 它仍在事先对所研究构象变化的机制做出了一定假设.

3 基于路径采样的过渡态搜索

目前所有算法中, 只有以 TPS 为代表的路径采样方法在事先对构象变化机制未作任何假设, 因为 TPS 将构象转变路径直接定义在了高维相空间内. 传统 TPS 通过大量随机的不外加偏执势的无偏采样, 得到一个过渡路径系综 (transition path ensemble, TPE), 见图 3(a). 最终通过对 TPE 的后处理分析, 选取合适的集合变量以描述过渡态[98–101] (图 3(b) 左); 最近, 通过引入强化学习范式 (reinforcement learning), 该方法实现了自适应无偏采样 (图 3(b) 右), 并采用符号回归 (symbolic regression) 完成机制解析[113,126].

3.1 过渡路径采样

3.1.1 相空间中过渡态的定义 committor probability

由于 TPS 中的路径直接定义在相空间, 相应地过渡态也无法直接套用低维空间中的鞍点 (saddle) 来具象地表征. 假设我们能通过某些 CV

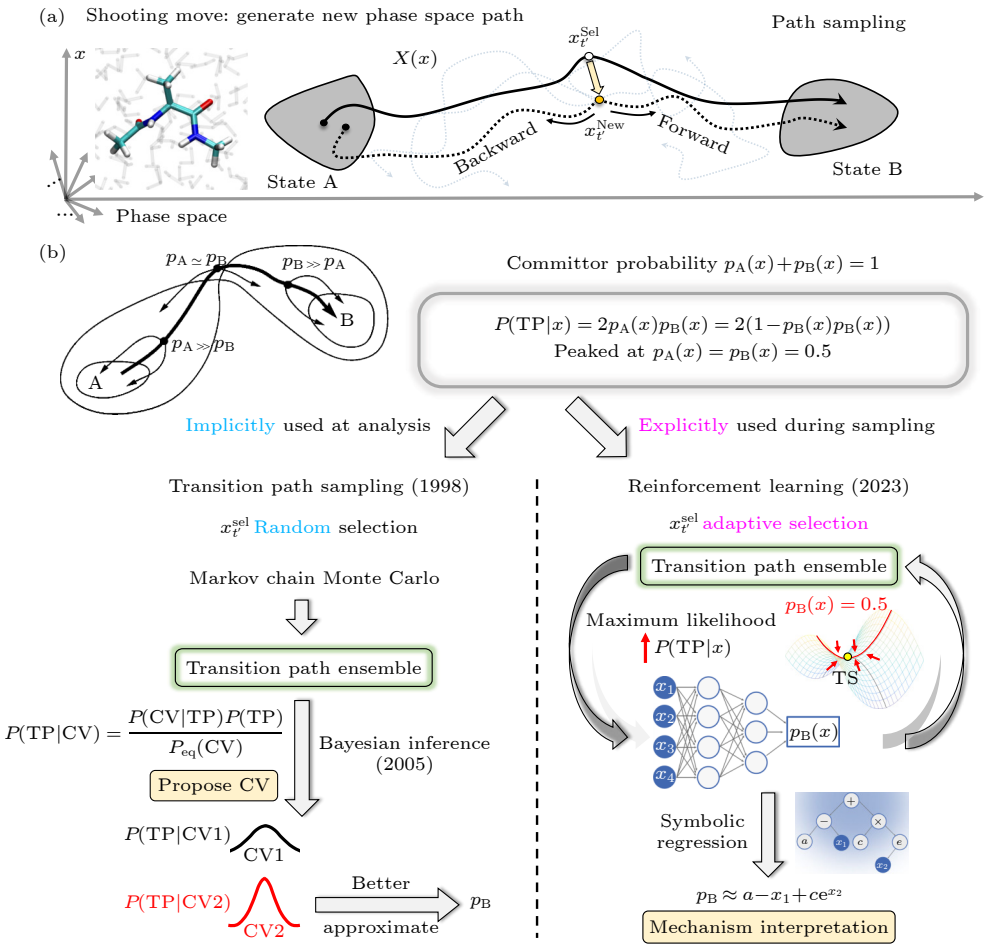

图 3 路径采样算法的基本原理示意图 (a) 路径采样中生成新相空间路径的 shooting move; (b) 传统过渡路径采样 (左侧) 的随机蒙特卡罗采样与过渡态分析原理[98–101], 融合强化学习的路径采样 (右侧) 在学习过程中不断促进采样起始点选择向过渡态集中[113]

Fig. 3. Schematics of path sampling methods. (a) Shooting move: select a phase space point on the current path, make a small perturbation to this point (redraw random initial velocities) and perform a set of simulations. (b) Path sampling is built upon the committor probability p_B. The traditional transition path sampling (left)[98–101] selects shooting points randomly and uses Monte Carlo for sampling; the transition state is characterized through post-analysis: choosing the CVs with the highest and narrowest distribution of $P(\mathrm{TP}|\mathrm{CV})$; the new reinforcement path sampling (right)[113] chooses shooting points adaptively and directly learns the committor probability p_B with maximized $P(\mathrm{TP}|x)$. Symbolic regression of p_B is used for mechanism interpretation.

定义出两个稳定态 A 和 B (并同时假设 A 和 B 中间不存在第 3 个稳定态 C), 那么 A 和 B 之间的过渡态就能通过 committor probability 来定义.

对相空间中的任一点, 都可以从其出发运行大量 MD 模拟并统计其中有多少比率分子是在抵达稳态 B 之前到达了 A, 另有多少比率相反在到达了 A 之前抵达了 B. 这两种比率 p_A 和 p_B 就是这一点对稳态 A 和 B 的 committor probability. 显然在不存在第 3 个稳态的前提下 $p_A+p_B=1$. 相应地, 过渡态则可以定义为由相空间内所有 $p_A=p_B=0.5$ 的点所组成的集合. 同时, 依据过渡路径理论 (transition path theory)[96], 我们知道对相空间中的任一点 $\boldsymbol{x}$ 而言, 它是属于连接 A 和 B 反应路径, 即过渡路径 (transition path, TP) 的其中一点的条件概率是

$$P(\mathrm{TP}|\boldsymbol{x})=2p_A(\boldsymbol{x})p_B(\boldsymbol{x})=2(1-p_B(\boldsymbol{x}))p_B(\boldsymbol{x}). \quad (3)$$

而此条件概率在过渡态上 $p_A=p_B=0.5$ 时将达到其峰值, 即过渡态上的点是所有相空间点中最有可能属于某条反应路径的. 这一点对路径采样算法至关重要.

3.1.2 Shooting move 新相空间路径的生成

假设已利用传统增强采样算法 (如 climber method/steered MD/targeted MD 等[114–116]) 得到一条连接 A 到 B 的转变路径, 便可以在此转变路径中抽选一个点 $\boldsymbol{x}^{\mathrm{sel}}$; 随后, 对 $\boldsymbol{x}^{\mathrm{sel}}$ 做出微扰 $\Delta\boldsymbol{x}$

(典型做法为根据给定温度的麦克斯韦-玻尔兹曼随机重置所有分子的初始速率), 而后以 $\boldsymbol{x}^{\rm new}=\boldsymbol{x}^{\rm sel}+\Delta\boldsymbol{x}$ 为新的初始条件进行多次无偏 MD 模拟采样. 其中, 每次 MD 模拟采样的终止条件为此采样路径到达了目标态 A 或 B 中的一个; 当这些轨迹中既有到达过 A 也有到达过 B 态时, 将到达过 A 态的任意路径和到达过 B 态的任意路径连接便成为由 A 态到达 B 态的转变路径. 该过程被称为 shooting move (图 3(a))[127].

路径采样过程就是不断迭代选定 $\boldsymbol{x}^{\rm sel}$, 而后进行 Shooting 的过程. 经过迭代最终会得到从 A 到 B 转变的路径系综 TPE[128,129]. 但传统 TPS 和其强化学习新版本在 $\boldsymbol{x}^{\rm sel}$ 的选择策略上有所不同.

3.1.3 过渡路径采样的 shooting move 策略

在原版 TPS 中, $\boldsymbol{x}^{\rm sel}$ 的选择是完全随机的. 同时, shooting move 的迭代是马尔科夫链蒙特卡罗的串行过程 (图 3(b) 左). 因此, TPS 天然欠缺并行化能力.

3.1.4 从路径系综中提取过渡态信息

经 shooting move 迭代得到路径系综后, 传统 TPS 需要用户自行定义 CV 来帮助解释其中蕴含的机制、提取过渡态信息. 根据 (3) 式, 如果所选的 CV 能够较好地表征过渡态, 即无限趋近 $p_{\rm B}$, 那么 $P({\rm TP}|{\rm CV})$ 应该呈现窄而高的分布. 但由于 $P({\rm TP}|{\rm CV})$ 无法直接计算, 需要通过贝叶斯推测间接计算:

$$P({\rm TP}|{\rm CV})=\frac{P({\rm CV}|{\rm TP})P({\rm TP})}{P_{\rm eq}({\rm CV})},\tag{4}$$

其中 $P({\rm CV}|{\rm TP})$ 可直接从 TPE 计算获得, $P({\rm TP})$ 需经额外长时间无偏采样算出, 而 $P_{\rm eq}({\rm CV})$ 是 CV 上的平衡态分布, 也需通过额外的伞形采样获得. 在用户选择的 CV 中, 以 $P({\rm TP}|{\rm CV})$ 分布最窄最高者最能表征过渡态和 A 到 B 的转变机制[98–101].

3.2 基于强化学习的路径采样

仔细分析原版 TPS 的后处理分析过程, 不难看出其对蒙特卡罗迭代采样结果的要求较高, 需确保所得 TPE 在过渡态附近有充足样本, 但由于其 $\boldsymbol{x}^{\rm sel}$ 的选择是完全随机, 这在面临较大的生物分子体系时是难以实现的.

因此, Jung 等[113] 于近期开发了基于强化学习 (reinforcement learning) 的路径采样算法. 与原版 TPS 仅在数据处理分析阶段隐性地使用 (4) 式不同, 新框架直接将 $P({\rm TP}|\boldsymbol{x})$ 用作了强化学习中的目标函数 (通过最大似然估计将其最大化), 用以训练以深度神经网络表达的 committor probability $p_{\rm B}$ (图 3(b) 右). 因此, 在此强化学习过程中, $P({\rm TP}|\boldsymbol{x})$ 的最大化意味着算法会自适应地选择 $\boldsymbol{x}^{\rm sel}$, 自发将其聚焦至过渡态附近 (即 $p_{\rm B}=0.5$, 图 3(b) 红线).

而后续对转变机制的解释, 即神经网络 $p_{\rm B}$ 物理含义的挖掘则可通过符号回归 (symbolic regression) 达成, 将 $p_{\rm B}(\boldsymbol{x})$ 的神经网络表达为容易理解的简单解析式[125,126].

3.3 路径采样算法的适用场景

值得强调的是, 无论是传统 TPS 还是强化学习路径采样, 二者的理论基础都是 $p_{\rm A}+p_{\rm B}=1$, 即不允许稳态 A 和 B 之间有第 3 个稳定态存在. 这意味着路径采样只能处理单个能垒, 即只能表征单个过渡态. 然而, 生物大分子的运动复杂, 亚稳定态数量众多, 很难保证已知的两个稳定态之间只有一个能垒. 这也限制了路径采样在生物大分子模拟中的应用.

4 融合 GAD 与降维算法的可能方案

经过对上述算法的简单回顾, 可以看出近年来依赖 CV 的路径搜索算法和非 CV 依赖的路径采样算法都已呈现与计算机科学和机器学习算法深度融合迈向自动化的发展趋势, 但依赖 CV 的 GAD 方法尚无相似案例可循. 我们推测一个可能的发展方向是将 GAD 在低维空间搜索过渡态的能力与降维算法结合起来. 自然地, 这对降维算法的性能提出了新的要求. 因此, 有必要先对现有降维算法的设计思想进行简要梳理.

4.1 现有降维算法

降维是无监督机器学习的传统分支, 其在生物分子模拟中的广泛应用已有综述阐明[130], 此处不再赘述. 但在目前众多的降维算法中, 显式利用时间序列信息, 即动力学信息, 进行降维的仅有时间结构独立成分分析 (time-lagged independent components analysis, tICA) 方法[63–65]. 但经 tICA 降维所得的低维 tIC 空间已被限定只能是原高维空间的线性组合, 而能够表征跃迁过程和过渡态的

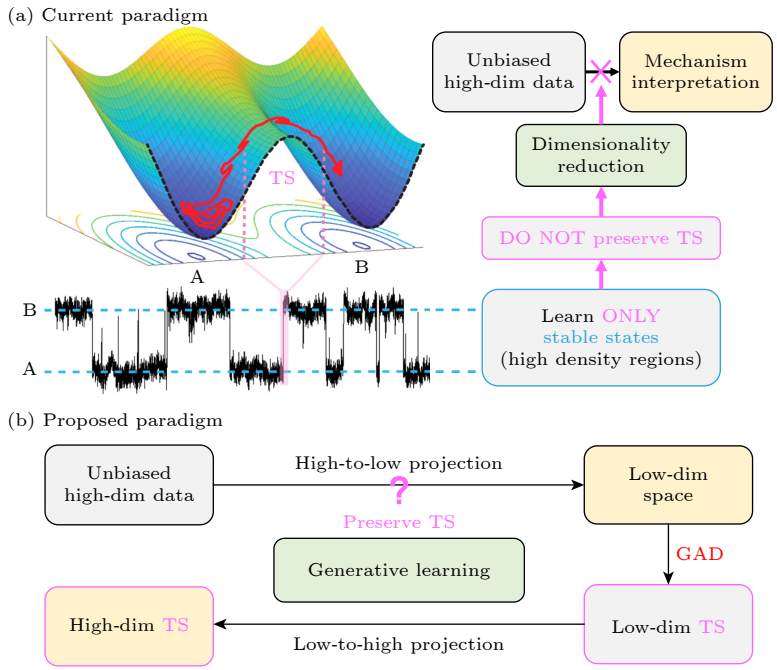

图 4 物理化学家需要怎样的降维算法 (a) 现有降维算法范式不保留过渡态信息, 不利于机制解析; (b) 可能的替代范式, 基于生成模型研发可保留过渡态信息的可逆降维算法, 并与低维空间搜索过渡态的 GAD 联用

Fig. 4. Requirements on dimensionality reduction algorithms by physical chemists. (a) Current paradigm for dimensionality reduction and the main difficulties for the transition state searching. (b) Proposed alternative paradigm for transition state searching: combine dimensionality reduction that preserves transition state information with GAD.

坐标很可能是原高维坐标的非线性函数. 其他现存降维算法, 因在降维过程中, 只关注保留高密度区域信息 (即稳定态信息), 常会将高维空间过度扭曲以致过渡态信息丢失 (图 4(a)). 因此, 现存降维算法都无法与 GAD 联用.

4.2 基于生成模型的可逆降维及过渡态搜索

近年来, 可逆神经网络和生成模型的发展, 为研发能够保留过渡态信息的新型降维算法提供了良好契机. 首先, 通过可逆神经网络, 我们可以期望利用深度学习训练出一个可以进行双向映射的生成模型, 即在将高维的全原子轨迹信息映射到某一低维空间的同时, 拥有把生成的低维空间样本逆投影回原空间的能力. 这样便可利用 GAD 在低维空间搜得鞍点结构, 再经逆投影自动得到完整的高维过渡态结构.

当然, 这一构想的实现难点是必须保证在降维过程中, 低维空间保有和原高维空间一致的动力学特征以及概率密度信息, 即保留过渡态信息. 这里我们建议参考 tICA 中直接使用动力学信息进行降维的做法. 此外, 为保障 GAD 在低维空间的顺利运行, 该生成模型应能为低维空间自动拟合出连续可导的自由能面.

5 结 论

生物分子功能机制的有效调控有赖于对其转变过程微观机制的全面考察, 其中以获取其主要转变路径中的过渡态信息最为关键. 当预设静态集合坐标较为容易、可强行定向降维时, 前人开发的 GAD 算法、finite temperature string 和快速断层扫描法, 已成功阐明了诸多生物过程的微观转变机制, 但当面对复杂转变过程时, 仍易出现预设集合变量常不合理, 需要消耗大量资源试错. 近年出现的基于旅行商的自动路径搜索算法 TAPS, 则有效避免了集合变量的预设问题, 还在并行化和 GPU加速的基础上, 提升了自动化程度和过渡态搜索效率.

在完全无需事前降维、不依赖集合变量的路径采样类算法中, 也已出现了通过融入强化学习思想实现自适应的高效率采样及过渡态分析优秀变体. 但只能处理单个能垒和过渡态搜寻的特点限制了这类算法在生物分子模拟中的应用.

因此, 研发可保留过渡态信息的新型降维算法

或是将机器学习进一步融入过渡态搜索的可行方向. 在此, 我们建议基于生成模型研发此种高质量降维方法, 并将之与 GAD 联用, 从而做到从任意状态出发, 快速捕捉其周围的过渡态信息.

参考文献

[1] Edman L, Földes-Papp Z, Wennmalm S, Rigler R 1999 *Chem. Phys.* **247** 11

[2] Evenäs J, Malmendal A, Thulin E, Carlström G, Forsén S 1998 *Biochemistry* **37** 13744

[3] Hanson J A, Duderstadt K, Watkins L P, Bhattacharyya S, Brokaw J B, Chu J W, Yang H 2007 *Proc. Natl. Acad. Sci. USA* **104** 18055

[4] Moffat K 1989 *Annu. Rev. Biophys. Chem.* **18** 309

[5] Huang C, Kalodimos C G 2017 *Annu. Rev. Biophys.* **46** 317

[6] Weissenberger G, Henderikx R J M, Peters P J 2021 *Nat. Methods* **18** 463

[7] Clegg R M 1995 *Curr. Opin. Biotechnol.* **6** 103

[8] Karplus M, McCammon J A 2002 *Nat. Struct. Biol.* **9** 646

[9] Hollingsworth S A, Dror R O 2018 *Neuron* **99** 1129

[10] Bernèche S, Roux B 2001 *Nature* **414** 73

[11] Khafizov K, Perez C, Koshy C, Quick M, Fendler K, Ziegler C, Forrest L R 2012 *Proc. Natl. Acad. Sci. USA* **109** E3035

[12] Li J, Shaikh S A, Enkavi G, Wen P C, Huang Z, Tajkhorshid E 2013 *Proc. Natl. Acad. Sci. USA* **110** 7696

[13] Dror, R O, Green H F, Valant C, Borhani D W, Valcourt J R, Pan A C, Arlow D H, Canals M, Lane J R, Rahmani R, Baell J B, Sexton P M, Christopoulos A, Shaw D E 2013 *Nature* **503** 295

[14] Wacker D, Stevens R C, Roth B L 2017 *a Cell* **170** 414

[15] Wacker D, Wang S, McCorvy J D, Betz R M, Venkatakrishnan A J, Levit A, Lansu K, Schools Z L, Che T, Nichols D E, Dror R O, Roth B L 2017 *Cell* **168** 377

[16] McCorvy J D, Butler K V, Kelly B, Rechsteiner K, Karpiak J, Betz R M, Kormos B L, Shoichet B K, Dror R O, Jin J, Roth B L 2018 *Nat. Chem. Biol.* **14** 126

[17] Provasi D, Artacho M C, Negri A, Mobarec J C, Filizola M 2011 *PLoS Comput. Biol.* **7** e1002193

[18] Cordero-Morales J F, Jogini V, Lewis A, Vásquez V, Cortes D M, Roux B, Perozo E 2007 *Nat. Struct. Mol. Biol.* **14** 1062

[19] Fields J B, Németh-Cahalan K L, Freites J A, Vorontsova I, Hall J E, Tobias D J 2017 *J. Biol. Chem.* **292** 185

[20] Groban E S, Narayanan A, Jacobson M P 2006 *PLoS Comput. Biol.* **2** e32

[21] Liu Y, Ke M, Gong H 2015 *Biophys. J.* **109** 542

[22] Delemotte L, Tarek M, Klein M L, Amaral C, Treptow W 2011 *Proc. Natl. Acad. Sci. USA* **108** 6109

[23] Lindorff-Larsen K, Piana S, Dror R O, Shaw D E 2011 *Science* **334** 517

[24] Snow C D, Nguyen H, Pande V S, Gruebele M 2002 *Nature* **420** 102

[25] Dror R O, Arlow D H, Maragakis P, Mildorf T J, Pan A C, Xu H, Borhani D W, Shaw D E 2011 *Proc. Natl. Acad. Sci. USA* **108** 18684

[26] Dror R O, Pan A C, Arlow D H, Borhani D W, Maragakis P, Shan Y, Xu H, Shaw D E 2011 *Proc. Natl. Acad. Sci. USA* **108** 13118

[27] Gu Y, Shrivastava I H, Amara S G, Bahar I 2009 *Proc. Natl. Acad. Sci. USA* **106** 2589

[28] Latorraca N R, Fastman N M, Venkatakrishnan A J, Frommer W B, Dror R O, Feng L 2017 *Cell* **169** 96

[29] Stelzl L S, Fowler P W, Sansom M S, Beckstein O 2014 *J. Mol. Biol.* **426** 735

[30] Buch I, Giorgino T, De Fabritiis G 2011 *Proc. Natl. Acad. Sci. USA* **108** 10184

[31] Liang R, Swanson J M J, Madsen J J, Hong M, DeGrado W F, Voth G A 2016 *Proc. Natl. Acad. Sci. USA* **113** E6955

[32] Suomivuori C M, Gamiz-Hernandez A P, Sundholm D, Kaila V R I 2017 *Proc. Natl. Acad. Sci. USA* **114** 7043

[33] Tajkhorshid E, Nollert P, Jensen M Ø, Miercke L J W, O'Connell J, Stroud R M, Schulten K 2002 *Science* **296** 525

[34] Watanabe A, Choe S, Chaptal V, Rosenberg J M, Wright E M, Grabe M, Abramson J 2010 *Nature* **468** 988

[35] Dedmon M M, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson C M 2005 *J. Am. Chem. Soc.* **127** 476

[36] Nguyen H D, Hall C K 2004 *Proc. Natl. Acad. Sci. USA* **101** 16180

[37] Levitt M 1983 *J. Mol. Biol.* **168** 595

[38] Sugita Y, Okamoto Y 1999 *Chem. Phys. Lett.* **314** 141

[39] Rhee Y M, Pande V S 2003 *Biophys. J.* **84** 775

[40] Zhang W, Wu C, Duan Y 2005 *J. Chem. Phys.* **123** 154105

[41] Zhou R 2006 *Protein Folding Protocols* (Humana Totowa, NJ: Springer) pp205–223

[42] Sindhikara D, Meng Y, Roitberg A E 2008 *J. Chem. Phys.* **128** 024103

[43] Buchete N V, Hummer G 2008 *Phys. Rev. E* **77** 030902

[44] Rosta E, Hummer G 2009 *J. Chem. Phys.* **131** 165102

[45] Stelzl L S, Hummer G 2017 *J. Chem. Theory Comput.* **13** 3927

[46] Yang L J, Gao Q Y 2009 *J. Chem. Phys.* **131** 214109

[47] Yang L, Liu C W, Shao Q, Zhang J, Gao Y Q 2015 *Acc. Chem. Res.* **48** 947

[48] Yang Y I, Zhang J, Che X, Yang L J, Gao Y Q 2016 *J. Chem. Phys.* **144** 094105

[49] Yang Y I, Shao Q, Zhang J, Yang L J, Gao Y Q 2019 *J. Chem. Phys.* **151** 070902

[50] Huber T, Torda A E, Van Gunsteren W F 1994 *J. Comput. -Aided Mol. Des.* **8** 695

[51] Wada T, Kuroda K, Yoshida Y, Ogasawara K, Ogawa A, Endo S 2006 *Neurosurg. Rev.* **29** 242

[52] Hansen H S, Hünenberger P H 2010 *J. Comput. Chem.* **31** 1

[53] Perić-Hassler L, Hansen H S, Baron R, Hünenberger P H 2010 *Carbohydr. Res.* **345** 1781

[54] Grubmüller H 1995 *Phys. Rev. E* **52** 2893

[55] Schulze B G, Grubmüller H, Evanseck J D 2000 *J. Am. Chem. Soc.* **122** 8700

[56] Bouvier B, Grubmüller H 2007 *Biophys. J.* **93** 770

[57] Barducci A, Bonomi M, Parrinello M 2011 *WIREs Comput. Mol. Sci.* **1** 826

[58] Tiwary P, Parrinello M 2013 *Phys. Rev. Lett.* **111** 230602

[59] Bussi G, Laio A 2020 *Nat. Rev. Phys.* **2** 200

[60] Miao Y, Feher V A, McCammon J A 2015 *J. Chem. Theory Comput.* **11** 3584

[61] Miao Y, McCammon J A 2017 *Annu. Rep. Comput. Chem.* **13** 231

[62] Wang J, Arantes P R, Bhattarai A, et al. 2021 *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **11** e1521

[63] Naritomi Y, Sotaro F 2011 *J. Chem. Phys.* **134** 065101

[64] Schwantes C R, Vijay S P 2013 *J. Chem. Theory Comput.* **9** 2000

[65] Perez-Hernandez G, Paul F, Giorgino T, Fabritiis G D, Noé

F 2013 *J. Chem. Phys.* **139** 015102
[66] Bowman G R, Huang X H, Pande V S 2009 *Methods* **49** 197
[67] Metzner P, Noé F, Schütte C 2009 *Phys. Rev. E* **80** 021106
[68] Pande V S, Beauchamp K A, Bowman G R 2010 *Methods* **52** 99
[69] Prinz J H, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera J D, Schütte C, Noé F 2011 *J. Chem. Phys.* **134** 174105
[70] Kellogg E H, Lange O F, Baker D 2012 *J. Phys. Chem. B* **116** 11405
[71] Yao Y, Cui R Z, Bowman G R, Silva D A, Sun J, Huang X H 2013 *J. Chem. Phys.* **138** 174106
[72] McGibbon R T, Schwantes C R, Pande V S 2014 *J. Phys. Chem. B* **118** 6475
[73] Nuske F, Keller B G, Pérez-Hernández G, Mey A S J, Noé F 2014 *J. Chem. Theory Comput.* **10** 1739
[74] Sheong F K, Silva D A, Meng L, Zhao Y, Huang X H 2015 *J. Chem. Theory Comput.* **11** 17
[75] Zhu L Z, Sheong F K, Zeng X, Huang X H 2016 *Phys. Chem. Chem. Phys.* **18** 30228
[76] Wang W, Cao S, Zhu L Z, Huang X H 2018 *WIREs Comput. Mol. Sci.* **8** e1343
[77] Husic B E, Pande V S 2018 *J. Am. Chem. Soc.* **140** 2386
[78] Konovalov K A, Unarta I C, Cao S, Goonetilleke E C, Huang X H 2021 *J. Am. Chem. Soc. Au.* **1** 1330
[79] E W, Zhou X 2011 *Nonlinearity* **24** 1831
[80] Samanta A, Chen M, Yu T Q, Tuckerman M E 2014 *J. Chem. Phys.* **140** 164109
[81] Chen M, Yu T Q, Tuckerman M E 2015 *Proc. Natl. Acad. Sci. USA* **112** 3235
[82] E W, Ren W, Vanden-Eijnden E 2002 *Phys. Rev. B* **66** 052301
[83] E W, Ren W, Vanden-Eijnden E 2005 *J. Phys. Chem. B* **109** 6688
[84] Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G 2006 *J. Chem. Phys.* **125** 024106
[85] Ren W, Vanden-Eijnden E 2007 *J. Chem. Phys.* **126** 164103
[86] Maragliano L, Vanden-Eijnden E 2007 *Chem. Phys. Lett.* **446** 182
[87] Pan A C, Sezer D, Roux B 2008 *J. Phys. Chem. B* **112** 3432
[88] Chen C, Huang Y, Xiao Y 2012 *Phys. Rev. E* **86** 031901
[89] Chen C, Huang Y, Ji X, Xiao Y 2013 *J. Chem. Phys.* **138** 164122
[90] Chen C J, Huang Y Z, Jiang X W, Xiao Y 2014 *J. Chem. Phys.* **141** 154109
[91] Zhu L Z, Sheong F K, Cao S, Liu S, Unarta I C, Huang X H 2019 *J. Chem. Phys.* **150** 124105
[92] Xi K, Hu Z, Wu Q, Wei M, Qian R, Zhu L Z 2021 *J. Chem. Theory Comput.* **17** 5301
[93] Wang L, Xi K, Zhu L Z, Da L T 2022 *J. Chem. Inf. Model.* **62** 3213
[94] Xi K, Zhu L Z 2022 *Int. J. Mol. Sci.* **23** 14628
[95] Xi K, Zhu L Z 2023 *A Practical Guide to Recent Advances in Multiscale Modeling and Simulation of Biomolecules* (AIP Publishing) pp9-1–9-24
[96] Vanden-Eijnden E 2006 *Computer Simulations in Condensed Matter: From Materials to Chemical Biology* (Berlin: Springer) pp453–493
[97] Vanden-Eijnden E 2010 *Annu. Rev. Phys. Chem.* **61** 391
[98] Dellago C, Bolhuis P G, Csajka F S, Chandler D 1998 *J. Chem. Phys.* **108** 1964
[99] Bolhuis P G, Dellago C, Chandler D 1998 *Faraday Discuss.* **110** 421
[100] Dellago C, Bolhuis P G, Chandler D 1999 *J. Chem. Phys.* **110** 6617
[101] Dellago C, Bolhuis P G, Geissler P L 2002 *Adv. Chem. Phys.* **123** 1
[102] Noé F, Tkatchenko A, Müller K R, Clementi C 2020 *Annu. Rev. Phys. Chem.* **71** 361
[103] AlQuraishi M, Sorger P K 2021 *Nat. Methods* **18** 1169
[104] Karniadakis G E, Kevrekidis I G, Lu L, Perdikaris P, Wang S, Yang L 2021 *Nat. Rev. Phys.* **3** 422
[105] Ju F, Zhu J, Shao B, Kong L, Liu T Y, Zheng W M, Bu D 2021 *Nat. Commum.* **12** 2535
[106] Huang B, Xu Y, Hu X H, Liu Y R, Liao S H, Zhang J H, Huang C D, Hong J J, Chen Q, Liu H Y 2022 *Nature* **602** 523
[107] Du Z, Su H, Wang W, Ye L, Wei H, Peng Z, Anishchenko I, Baker D, Yang J 2021 *Nat. Prot.* **16** 5634
[108] Dai M, Dong Z, Xu K, Zhang Q C 2023 *J. Mol. Biol.* **435** 168059
[109] Yuan Q M, Chen J W, Zhao H Y, Zhou Y Q, Yang Y D 2022 *Bioinformatics* **38** 125
[110] Su M Y, Feng G, Liu Z, Li Y, Wang R 2020 *J. Chem. Inf. Model.* 60 1122
[111] Zeng C, Jian Y, Vosoughi S, Zeng C, Zhao Y 2023 *Nat. Commum.* **14** 1060
[112] Noé F, Olsson S, Köhler J, Wu H 2019 *Science* **365** 6457
[113] Jung H, Covino R, Arjun A, Leitold C, Dellago C, Bolhuis P G, Hummer G 2023 *Nat. Comput. Sci.* **3** 334
[114] Weiss D R, Levitt M 2009 *J. Mol. Biol.* **385** 665
[115] Israelewitz B, Gao M, Schulten K 2001 *Curr. Opin. Struct. Biol.* **11** 224
[116] Schlitter J, Engels M, Krüger P 1994 *J. Mol. Graph.* **12** 84
[117] Torrie G M, Valleau J P 1977 *J. Comput. Phys.* **23** 187
[118] Ryckaert J P, Ciccotti G, Berendsen H J 1977 *J. Comput. Phys.* **23** 327
[119] Babin V, Roland C, Sagui C 2008 *J. Chem. Phys.* **128** 134101
[120] Branduardi D, Gervasio F L, Parrinello M 2007 *J. Chem. Phys.* **126** 054103
[121] Applegate D L, Bixby R E, Chvátal V, Cook W J 2011 *The Traveling Salesman Problem: A Computational Study* (Princeton University Press) pp1–58
[122] Cox M A A, Cox T F 2008 *Handbook of Data Visualization* (Berlin: Springer) pp315–347
[123] Barducci A, Bussi G, Parrinello M 2008 *Phys. Rev. Lett.* **100** 020603
[124] Fischmann T O, Smith C K, Mayhood T W, Myers J E, Reichert J P, Mannarino A, Carr D, Zhu H, Wong J, Yang R S, Le H V, Madison V S 2009 *Biochemistry* **48** 2661
[125] Hanrahan A J, Sylvester B E, Chang M T, et al. 2020 *Cancer Res.* **80** 4233
[126] Schmidt M, Lipson H 2009 *Science* **324** 81
[127] Jung H, Okazaki K, Hummer G 2017 *J. Chem. Phys.* **147** 152716
[128] Swenson D W H, Prinz J H, Noe F, Chodera J D, Bolhuis P G 2019 *J. Chem. Theory Comput.* **15** 813
[129] Swenson D W H, Prinz J H, Noe F, Chodera J D, Bolhuis P G 2019 *J. Chem. Theory Comput.* **15** 837
[130] Glielmo A, Husic B E, Rodriguez A, Clementi C, Noé F, Laio A 2021 *Chem. Rev.* **121** 9722

SPECIAL TOPIC—Machine learning in biomolecular simulations

Transition state searching for complex biomolecules: Algorithms and machine learning*

Yang Jian-Yu # Xi Kun # Zhu Li-Zhe †

(*Warshel Institute for Computational Biology, School of Medicine, The Chinese University of Hong Kong, Shenzhen* 518172, *China*)

(Received 13 August 2023; revised manuscript received 9 September 2023)

Abstract

Transition state is a key concept for chemists to understand and fine-tune the conformational changes of large biomolecules. Due to its short residence time, it is difficult to capture a transition state via experimental techniques. Characterizing transition states for a conformational change therefore is only achievable via physics-driven molecular dynamics simulations. However, unlike chemical reactions which involve only a small number of atoms, conformational changes of biomolecules depend on numerous atoms and therefore the number of their coordinates in our 3D space. The searching for their transition states will inevitably encounter the curse of dimensionality, i.e. the reaction coordinate problem, which invokes the invention of various algorithms for solution. Recent years, new machine learning techniques and the incorporation of some of them into the transition state searching methods emerged. Here, we first review the design principle of representative transition state searching algorithms, including the collective-variable (CV)-dependent gentlest ascent dynamics, finite temperature string, fast tomographic, travelling-salesman based automated path searching, and the CV-independent transition path sampling. Then, we focus on the new version of TPS that incorporates reinforcement learning for efficient sampling, and we also clarify the suitable situation for its application. Finally, we propose a new paradigm for transition state searching, a new dimensionality reduction technique that preserves transition state information and combines gentlest ascent dynamics.

Keywords: transition state, gentlest ascent dynamics, path methods, reinforcement learning, generative models

PACS: 87.10.Tf, 87.15.A–, 87.15.H–, 87.15.hp **DOI:** 10.7498/aps.72.20231319

* Project supported by the National Natural Science Foundation of China (Grant No. 31971179) and the Science Technology and Innovation Commission of Shenzhen Municipality, China (Grant Nos. JCYJ20200109150003938, RCYX20200714114645019).

# These authors contributed equally.

† Corresponding author. E-mail: zhulizhe@cuhk.edu.cn

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

生物大分子过渡态搜索算法及其中的机器学习

杨建宇　席昆　竺立哲

Transition state searching for complex biomolecules: Algorithms and machine learning

Yang Jian-Yu　Xi Kun　Zhu Li-Zhe

引用信息 Citation: Acta Physica Sinica, 72, 248701 (2023)　DOI: 10.7498/aps.72.20231319
在线阅读 View online: https://doi.org/10.7498/aps.72.20231319
当期内容 View table of contents: http://wulixb.iphy.ac.cn
