(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 103729577 B
(45)授权公告日 2017.08.22

(21)申请号 201310683507.7

(22)申请日 2013.12.12

(65)同一申请的已公布的文献号
申请公布号 CN 103729577 A

(43)申请公布日 2014.04.16

(73)专利权人 深圳先进技术研究院
地址 518055 广东省深圳市南山区西丽大学城学苑大道1068号

(72)发明人 彭丰斌 魏彦杰 张慧玲 弓英瑛

(74)专利代理机构 深圳市科进知识产权代理事务所(普通合伙) 44316
代理人 沈祖锋 郝明琴

(51)Int.Cl.
G06F 19/18(2011.01)

(56)对比文件
CN 103761452 A,2014.04.30,
CN 102779239 A,2012.11.14,
US 2013/0018594 A1,2013.01.17,
R. Ghulghazaryan, et al..Efficient Combination of Wang-Landau and Transition Matrix Monte Carlo Methods for Protein Simulations.《Journal of Computational Chemistry》.2006,第28卷(第3期),715-726.
R. Ghulghazaryan, et al..Efficient Combination of Wang-Landau and Transition Matrix Monte Carlo Methods for Protein Simulations.《Journal of Computational Chemistry》.2006,第28卷(第3期),715-726.
彭丰斌 等.基于并行群体模拟退火算法的蛋白质折叠研究.《科研信息化技术与应用》.2013,第4卷(第5期),26-34.
T. Vogel, et al..Generic, Hierarchical Framework for Massively Parallel Wang-Landau Sampling.《PHYSICAL REVIEW LETTERS》.2013,第110卷(第21期),
M. Eisenbach.Future Proofing WL-LSMS: Preparing for First Principles Thermodynamics Calculations on Accelerator and Multicore Architectures.《CUG 2011 Proceedings》.2011,

审查员 王硕

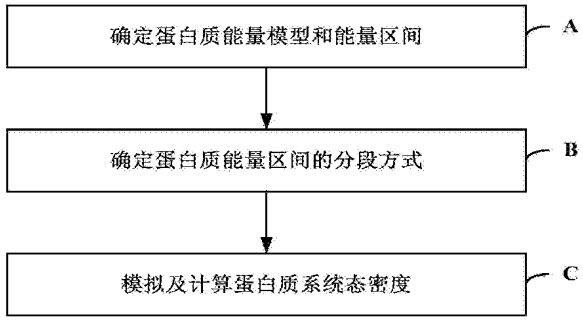
权利要求书3页 说明书7页 附图2页

(54)发明名称

基于混合并行方式的蛋白质热力学分析高效随机模拟方法

(57)摘要

本发明涉及生物信息分析技术领域,提供了一种基于混合并行方式的蛋白质热力学分析高效随机模拟方法,包括:步骤A:确定蛋白质能量模型和能量区间;步骤B:确定所述蛋白质能量区间的分段方式;步骤C:模拟及计算蛋白质系统态密度。采用本发明提供的方法,可以高效地分析和研究蛋白质折叠的整个热力学过程,进而对蛋白质折叠过程进行探索和研究。



CN 103729577 B

1.一种基于混合并行方式的蛋白质热力学分析高效随机模拟方法，其特征在于，包括：

步骤A：确定蛋白质能量模型和能量区间；

步骤B：确定所述蛋白质能量区间的分段方式；

步骤C：模拟及计算蛋白质系统态密度；

所述步骤A进一步包括：

采用ECEPP蛋白质能量模型，ECEPP能量力场的表达形式为：

$E_{ECEPP}=E_C+E_{LJ}+E_{HB}+E_{Tor}$

其中，$E_C=\sum_{(i,j)}\frac{332q_iq_j}{\varepsilon r_{ij}}$是两电荷之间的库伦作用力，$r_{ij}$表示原子i和j之间的距离；$E_{LJ}=\sum_{(i,j)}(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}})$是两原子之间的兰纳-琼斯作用力；$E_{HB}=\sum_{(i,j)}(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}})$是氢键作用力；$E_{Tor}=\Sigma_1U_1(1\pm\cos(n_1\xi_1))$是两面角旋转作用力，$\xi_1$是第1个两面角；所述步骤A进一步包括：

对所使用的蛋白质能量区间进行离散化处理，若取k个能量bin区间值，则对$[E_{min},E_{max}]$平均划分k个bin区间，用每个bin区间中间的一个能量值代表能量区间值；所述步骤B进一步包括：

步骤B1：对能量区间平均分为M段，设相邻子能量区间之间的重合度等于Δ个bin区间，则每一段含有$\left[\frac{k}{M}+\Delta\right]$个bin区间；

步骤B2：依照当前计算得到的蛋白质系统态密度函数的对数S(E)分布特点，自适应地对能量区间分段，若某个子能量区间为$[E_{begin},E_{end}]$，则$\nabla S(E)=S(E_{end})-S(E_{begin})$；

所述步骤C进一步包括：

通过MPI的主从进程模式和OpenMP的多线程并行模式，模拟及计算蛋白质系统态密度；

在所述主从进程模式的N个分进程中，分进程1为主进程，其余分进程均为从进程；

所述主进程包括如下步骤：

步骤S11：初始化蛋白质系统态密度函数的对数S(E)=lng(E)=0，直方图H(E)=0，$E_{min}\leqslant E\leqslant E_{max}$，修正因子df=In(f)，定义因子f，则修正因子df定义为f的自然对数，并初始化df=1；

步骤S12：s=1；

步骤S13：依照所确定的蛋白质能量区间的分段方式将能量区间$E_{min}\leqslant E\leqslant E_{max}$分成M段，并分配到M个分线程中，t=1；

步骤S14：在每个分线程中，对原来的构型限制在相应的子能量区间里进行随机变动，产生新的构型，计算能量E_{new}，根据Metropolis准则确定新构型被接受的概率，t=t+1；

所述步骤S14循环tmax次；

步骤S15：所有线程间相互通信，综合得到整个区间的S(E)和H(E)，s=s+1；

所述步骤S14和S15循环smax次；

步骤S16：所有进程间相互通信，主进程收集所有从进程的$S_{tmp}(E)$和$H_{tmp}(E)$并累加计算

出全局的S(E)和H(E)，即全局的S(E)＝S(E)+所有从进程的$S_{tmp}(E)$，全局的H(E)＝H(E)+所有从进程的$H_{tmp}(E)$，将全局的S(E)和H(E)广播给所有从进程，判断直方图平缓条件：

$$\frac{\max(H(E))-\min(H(E))}{\max(H(E))+\min(H(E))}<\phi,\quad 0<\phi<1$$

若不满足则返回执行步骤S12继续迭代；若满足则执行步骤S17；

步骤S17：改变修正因子df，再返回执行步骤S12继续迭代，直到满足进程终止条件$df<\varphi$，其中$0<\varphi<1$；求得S(E)，得到蛋白质系统相对的态密度$g(E)=e^{S(E)}$。

2.如权利要求1所述的方法，其特征在于，在所述步骤S14中，根据Metropolis准则确定新构型被接受的概率进一步包括：

$$P(old\rightarrow new)=\min(1,e^{-[S(E_{new})-S(E_{old})]})$$

若接受新构型，则：

$S(E_{new})=S(E_{new})+df$，$H(E_{new})=H(E_{new})+1$；

否则：

$S(E_{old})=S(E_{old})+df$，$H(E_{old})=H(E_{old})+1$。

3.如权利要求1所述的方法，其特征在于，所述从进程包括如下步骤：

步骤S21：初始化蛋白质系统态密度函数的对数S(E)＝lng(E)＝0，$S_{tmp}(E)=lng_{tmp}(E)$＝0，直方图H(E)＝0，$H_{tmp}(E)$＝0，$E_{min}\leqslant E\leqslant E_{max}$，修正因子df＝In(f)，定义因子f，则修正因子df定义为f的自然对数，并初始化df＝1；

步骤S22：s＝1；

步骤S23：依照所确定的蛋白质能量区间的分段方式将能量区间$E_{min}\leqslant E\leqslant E_{max}$分成M段，并分配到M个分线程中，t＝1；

步骤S24：在每个分线程中，对原来的构型限制在相应的子能量区间里进行随机变动，产生新的构型，计算能量E_{new}，根据Metropolis准则确定新的构型被接受的概率，t＝t+1；

所述步骤S24循环tmax次；

步骤S25：所有线程间相互通信，综合得到整个区间的$S_{tmp}(E)$和$H_{tmp}(E)$，s＝s+1；

所述步骤S25循环smax次；

步骤S26：所有进程间相互通信，从进程将$S_{tmp}(E)$和$H_{tmp}(E)$发送给主进程，然后接收经主进程计算得出的全局的S(E)和H(E)更新原来的S(E)和H(E)，将$S_{tmp}(E)$和$H_{tmp}(E)$初始化为0，判断直方图平缓条件：

$$\frac{\max(H(E))-\min(H(E))}{\max(H(E))+\min(H(E))}<\phi,\quad 0<\phi<1$$

若不满足则返回执行步骤S22继续迭代；若满足则执行步骤S27；

步骤S27：改变修正因子df，再返回执行步骤S22继续迭代，直到满足进程终止条件$df<\varphi$，其中$0<\varphi<1$。

4.如权利要求3所述的方法，其特征在于，在所述步骤S24中，根据Metropolis准则确定新的构型被接受的概率进一步包括：

$$P(old\rightarrow new)=\min(1,e^{-[S(E_{new})-S(E_{old})]})$$

若接受新构型，则：

$S(E_{new})=S(E_{new})+df$，$H(E_{new})=H(E_{new})+1$，

$S_{tmp}(E_{new})=S_{tmp}(E_{new})+df$，$H_{tmp}(E_{new})=H_{tmp}(E_{new})+1$；

否则：

$S(E_{old})=S(E_{old})+df$，$H(E_{old})=H(E_{old})+1$，

$S_{tmp}(E_{old})=S_{tmp}(E_{old})+df$，$H_{tmp}(E_{old})=H_{tmp}(E_{old})+1$。

5.如权利要求1所述的方法，其特征在于，在所述步骤S17中，改变修正因子df的方式为：

先连续进行N次迭代的，$f=f^{\alpha}$，$0<\alpha<1$，再进行1次迭代的$f=f^{1/\alpha^{N-1}}$；反复重复上述方式。

6.如权利要求3所述的方法，其特征在于，在所述步骤S27中，改变修正因子df的方式为：

先连续进行N次迭代的，$f=f^{\alpha}$，$0<\alpha<1$，再进行1次迭代的$f=f^{1/\alpha^{N-1}}$；反复重复上述方式。

基于混合并行方式的蛋白质热力学分析高效随机模拟方法

【技术领域】

[0001] 本发明涉及生物信息分析技术领域，特别是涉及一种基于混合并行方式的蛋白质热力学分析高效随机模拟方法。

【背景技术】

[0002] 蛋白质折叠主要研究蛋白质如何在短时间内从一维多肽链折叠为天然三维结构，形成具有生命功能的大分子。生物体的遗传信息(DNA)通过RNA转录和翻译过程传递给蛋白质(即中心法则)，因此蛋白质折叠也被称为第二遗传密码，它的研究可以帮助揭示生命遗传信息的表达和功能传递的奥秘。在从一维多肽链到天然三维结构的折叠过程中，蛋白质可发生误折叠或聚集，其结构和功能因此受到破坏，从而引起‘折叠病’，比如老年痴呆症等。因此蛋白质折叠研究对探索多种‘折叠病’机理意义重大。

[0003] 目前，研究蛋白质折叠的算法大多数都在分子动力学模拟和随机模拟中实现。一般而言，分子动力学模拟常用于研究蛋白质系统的动力学过程；而随机模拟则可以研究蛋白质系统的整个热动力学过程。针对使用高精确度的全原子蛋白质模型的模拟，需要计算成千上万个原子之间的多种相互作用力，对于分子动力学模拟只能模拟纳秒级的蛋白质折叠过程，故其在微秒到毫秒时间内的蛋白质折叠研究中具有很大的局限性；此外，分子动力学模拟也受一个初始实验构型的影响。而随机模拟不但能用于微秒到毫秒时间内的蛋白质折叠研究，而且不依赖于一个具体的初始构型，可以更广泛地搜索构型空间。

[0004] 经典的WangLandau算法就是随机模拟领域最吸引人最有发展情景的新算法，它能解决生物信息学、统计物理学等多个领域的很多复杂问题。比如在蛋白质折叠研究中，该算法有两个最显著的优点：第一，蛋白质模拟不会局限在局部最小能量状态，因而能较好地在整个能量区间进行自由行走；第二，通过该算法可模拟和计算出蛋白质系统态密度，因而就能进一步求解得到宽广温度范围内的很多热动力学量如比热等，这样就能高效地分析和研究蛋白质折叠的整个热力学过程。但WangLandau算法在计算精度和速度上还有待于进一步提升。

[0005] 鉴于此，克服该现有技术所存在的缺陷是本技术领域亟待解决的问题。

【发明内容】

[0006] 本发明要解决的技术问题是提供一种基于混合并行方式的蛋白质热力学分析高效随机模拟方法。

[0007] 本发明采用如下技术方案：

[0008] 一种基于混合并行方式的蛋白质热力学分析高效随机模拟方法，包括：

[0009] 步骤A：确定蛋白质能量模型和能量区间；

[0010] 步骤B：确定所述蛋白质能量区间的分段方式；

[0011] 步骤C：模拟及计算蛋白质系统态密度。

[0012] 进一步地，所述步骤A进一步包括：

[0013] 采用ECEPP蛋白质能量模型，ECEPP能量力场的表达形式为：

[0014] $E_{ECEPP}=E_C+E_{LJ}+E_{HB}+E_{Tor}$

[0015] 其中，$E_C=\sum_{(i,j)}\frac{332q_iq_j}{\varepsilon r_{ij}}$ 是两电荷之间的库伦作用力，r_{ij}表示原子i和j之间的距离；$E_{LJ}=\sum_{(i,j)}(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}})$ 是两原子之间的兰纳-琼斯作用力；$E_{HB}=\sum_{(i,j)}(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}})$ 是氢键作用力；$E_{Tor}=\Sigma_l U_l(1\pm\cos(n_l\xi_l))$ 是两面角旋转作用力，ξ_l是第l个两面角。

[0016] 进一步地，所述步骤A进一步包括：

[0017] 对所使用的蛋白质能量区间进行离散化处理，若取k个能量bin区间值，则对[E_{min}，E_{max}]平均划分k个bin区间，用每个bin区间中间的一个能量值代表能量区间值。

[0018] 进一步地，所述步骤B进一步包括：

[0019] 步骤B1：对能量区间平均分为M段，设相邻子能量区间之间的重合度等于Δ个bin区间，则每一段含有$\left[\frac{k}{M}+\Delta\right]$个bin区间；

[0020] 步骤B2：依照当前计算得到的蛋白质系统态密度函数的对数S(E)分布特点，自适应地对能量区间分段，若某个子能量区间为[E_{begin}，E_{end}]，则$\nabla S(E)=S(E_{end})-S(E_{begin})$。

[0021] 进一步地，所述步骤C进一步包括：

[0022] 通过MPI的主从进程模式和OpenMP的多线程并行模式，模拟及计算蛋白质系统态密度。

[0023] 进一步地，在所述主从进程模式的N个分进程中，分进程1为主进程，其余分进程均为子进程。

[0024] 进一步地，所述主进程包括如下步骤：

[0025] 步骤S11：初始化蛋白质系统态密度函数的对数S(E)=lng(E)=0，直方图H(E)=0($E_{min}\leqslant E\leqslant E_{max}$)，修正因子df=1(=lnf=lne)；

[0026] 步骤S12：s=1；

[0027] 步骤S13：依照所确定的蛋白质能量区间的分段方式将能量区间($E_{min}\leqslant E\leqslant E_{max}$)分成M段，并分配到M个分线程中，t=1；

[0028] 步骤S14：在每个分线程中，对原来的构型限制在相应的子能量区间里进行随机变动，产生新的构型，计算能量E_{new}，根据Metropolis准则确定新构型被接受的概率，t=t+1；

[0029] 所述步骤S14循环tmax次；

[0030] 步骤S15：所有线程间相互通信，综合得到整个区间的S(E)和H(E)，s=s+1；

[0031] 所述步骤S14和S15循环smax次；

[0032] 步骤S16：所有进程间相互通信，主进程收集所有从进程的$S_{tmp}(E)$和$H_{tmp}(E)$并累加计算出全局的S(E)和H(E)，即全局的S(E)=S(E)+所有从进程的$S_{tmp}(E)$，全局的H(E)=H(E)+所有从进程的$H_{tmp}(E)$，将全局的S(E)和H(E)的广播给所有从进程，判断直方图平缓条件：

[0033] $$\frac{\max(H(E))-\min(H(E))}{\max(H(E))+\min(H(E))}<\phi \quad (0<\phi<1)$$

[0034] 若不满足则返回执行步骤S12继续迭代;若满足则执行步骤S17;

[0035] 步骤S17:改变修正因子df,再返回执行步骤S12继续迭代,直到满足进程终止条件 $df<\varphi$,其中 $0<\varphi<1$;求得S(E),得到蛋白质系统相对的态密度g(E)=$e^{S(E)}$。

[0036] 进一步地,在所述步骤S14中,根据Metropolis准则确定新构型被接受的概率进一步包括:

[0037] $$P(old\rightarrow new)=\min(1,e^{-[S(E_{new})-S(E_{old})]})$$

[0038] 若接受新构型,则:

[0039] S(E_{new})=S(E_{new})+df,H(E_{new})=H(E_{new})+1;

[0040] 否则:

[0041] S(E_{old})=S(E_{old})+df,H(E_{old})=H(E_{old})+1。

[0042] 进一步地,所述从进程包括如下步骤:

[0043] 步骤S21:初始化蛋白质系统态密度函数的对数S(E)=lng(E)=0,S_{tmp}(E)=lng_{tmp}(E)=0,直方图H(E)=0,H_{tmp}(E)=0($E_{min}\leqslant E\leqslant E_{max}$),修正因子df=1(=lnf=lne);

[0044] 步骤S22:s=1;

[0045] 步骤S23:依照所确定的蛋白质能量区间的分段方式将能量区间($E_{min}\leqslant E\leqslant E_{max}$)分成M段,并分配到M个分线程中,t=1;

[0046] 步骤S24:在每个分线程中,对原来的构型限制在相应的子能量区间里进行随机变动,产生新的构型,计算能量E_{new},根据Metropolis准则确定新构型被接受的概率,t=t+1;

[0047] 所述步骤S24循环tmax次;

[0048] 步骤S25:所有线程间相互通信,综合得到整个区间的S_{tmp}(E)和H_{tmp}(E),s=s+1;

[0049] 所述步骤S24和S25循环smax次;

[0050] 步骤S26:所有进程间相互通信,从进程将S_{tmp}(E)和H_{tmp}(E)发送给主进程,然后接收经主进程计算得出的全局的S(E)和H(E)更新原来的S(E)和H(E),将S_{tmp}(E)和H_{tmp}(E)初始化为0,判断直方图平缓条件:

[0051] $$\frac{\max(H(E))-\min(H(E))}{\max(H(E))+\min(H(E))}<\phi \quad (0<\phi<1)$$

[0052] 若不满足则返回执行步骤S22继续迭代;若满足则执行步骤S27;

[0053] 步骤S27:改变修正因子df,再返回执行步骤S22继续迭代,直到满足进程终止条件 $df<\varphi$,其中 $0<\varphi<1$。

[0054] 进一步地,在所述步骤S24中,根据Metropolis准则确定新构型被接受的概率进一步包括:

[0055] $$P(old\rightarrow new)=\min(1,e^{-[S(E_{new})-S(E_{old})]})$$

[0056] 若接受新构型,则:

[0057] S(E_{new})=S(E_{new})+df,H(E_{new})=H(E_{new})+1,

[0058] S_{tmp}(E_{new})=S_{tmp}(E_{new})+df,H_{tmp}(E_{new})=H_{tmp}(E_{new})+1;

[0059] 否则：

[0060] $S(E_{old})=S(E_{old})+df$，$H(E_{old})=H(E_{old})+1$，

[0061] $S_{tmp}(E_{old})=S_{tmp}(E_{old})+df$，$H_{tmp}(E_{old})=H_{tmp}(E_{old})+1$。

[0062] 进一步地，在所述步骤S17和S27中，改变修正因子f的方式为：

[0063] 先连续进行N次迭代的$f=f^{\alpha}$ $(0<\alpha<1)$，再进行1次迭代的 $f=f^{1/\alpha^{N-1}}$；

[0064] 反复重复上述方式。

[0065] 与现有技术相比，本发明的有益效果在于：

[0066] 与经典的WangLandau算法相比，本发明使用基于退火机制的更新修正因子方式可提高计算精度和速度，利用一种灵活的能量区间分段方式可使分线程间负载均衡，采用MPI+OpenMP混合编程模型的混合并行方式可大大加快模拟和计算速度。采用本发明提供的方法，可以高效地分析和研究蛋白质折叠的整个热力学过程，进而对蛋白质折叠过程进行探索和研究。

【附图说明】

[0067] 图1是本发明实施例基于混合并行方式的蛋白质热力学分析高效随机模拟方法流程图；

[0068] 图2是本发明实施例中蛋白质能量区间的分段方式示意图；

[0069] 图3是模拟及计算蛋白质系统态密度的混合并行方法流程图。

【具体实施方式】

[0070] 为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

[0071] 此外，下面所描述的本发明各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

[0072] 本发明提供了一种基于混合并行方式的蛋白质热力学分析高效随机模拟方法，如图1所示，该方法包括：

[0073] 步骤A：确定蛋白质能量模型和能量区间；

[0074] 步骤B：确定蛋白质能量区间的分段方式；

[0075] 步骤C：模拟及计算蛋白质系统态密度。

[0076] 在步骤A中，可采用ECEPP蛋白质能量模型，ECEPP能量力场的表达形式为：

[0077] $E_{ECEPP}=E_C+E_{LJ}+E_{HB}+E_{Tor}$

[0078] 其中，$E_C=\sum_{(i,j)}\frac{332q_iq_j}{\varepsilon r_{ij}}$ 是两电荷之间的库伦作用力，r_{ij}表示原子i和j之间的距离；$E_{LJ}=\sum_{(i,j)}(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}})$ 是两原子之间的兰纳-琼斯作用力；$E_{HB}=\sum_{(i,j)}$

$(\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}})$是氢键作用力；$E_{Tor}=\sum_1 U_1(1\pm\cos(n_1\xi_1))$是两面角旋转作用力，$\xi_1$是第1个两面角。由于ECEPP能量力场采用的是角度坐标系，所以其计算效率高于基于笛卡尔坐标系的其他蛋白质模型。

[0079] 为了便于计算机模拟仿真，还可对所使用的蛋白质能量区间进行离散化处理，若取k个能量bin区间值，则对$[E_{min},E_{max}]$平均划分k个bin区间，用每个bin区间中间的一个能量值代表能量区间值。

[0080] 在步骤B中，为了使分线程间负载均衡，可采用如下具有自适应特点的一种灵活的能量区间分段方式：

[0081] 步骤B1：对能量区间平均分为M段，设相邻子能量区间之间的重合度等于Δ个bin区间，则每一段含有$\left[\frac{k}{M}+\Delta\right]$个bin区间；

[0082] 其中，设相邻子能量区间之间的重合度等于Δ个bin区间，是为了能综合得到整个区间的S(E)和H(E)（即综合所有子能量区间的S(E)和H(E)），需要相邻子能量区间有合适的重合度。

[0083] 步骤B2：依照当前计算得到的蛋白质系统态密度函数的对数S(E)分布特点，自适应地对能量区间分段，若某个子能量区间为$[E_{begin},E_{end}]$，则$\nabla S(E)=S(E_{end})-S(E_{begin})$。

[0084] 其中，一般地，蛋白质系统态密度函数的对数S(E)=lng(E)是单调递增的上凹函数，自适应地对能量区间分段需使得每一段的$\nabla S(E)$均衡，也即若某个子能量区间为$[E_{begin},E_{end}]$，则$\nabla S(E)=S(E_{end})-S(E_{begin})$，如图2所示。

[0085] 在步骤C中，通过MPI的主从进程模式和OpenMP的多线程并行模式，模拟及计算蛋白质系统态密度；在主从进程模式的N个分进程中，分进程1为主进程，其余分进程均为子进程。

[0086] 如图3所示，主进程包括如下步骤：

[0087] 步骤S11：初始化蛋白质系统态密度函数的对数S(E)=lng(E)=0，直方图H(E)=0($E_{min}\leqslant E\leqslant E_{max}$)，修正因子df=1(=lnf=lne)；

[0088] 步骤S12：s=1；

[0089] 步骤S13：依照所确定的蛋白质能量区间的分段方式将能量区间($E_{min}\leqslant E\leqslant E_{max}$)分成M段，并分配到M个分线程中，t=1；

[0090] 步骤S14：在每个分线程中，对原来的构型限制在相应的子能量区间里进行随机变动，产生新的构型，计算能量E_{new}，根据Metropolis准则确定新构型被接受的概率：(简称为MCS步)

[0091] $$P(old\rightarrow new)=\min(1,e^{-[S(E_{new})-S(E_{old})]})$$

[0092] 若接受新构型，则：

[0093] $S(E_{new})=S(E_{new})+df$，$H(E_{new})=H(E_{new})+1$；

[0094] 否则：

[0095] $S(E_{old})=S(E_{old})+df$，$H(E_{old})=H(E_{old})+1$。

[0096] t=t+1，所述步骤S14循环tmax次（也即经过tmax次（如10次）MCS步）；

[0097] 步骤S15：所有线程间相互通信，综合得到整个区间的S(E)和H(E)，s=s+1；

[0098] 所述步骤S14和S15循环smax次（如100次）；

[0099] 步骤S16：所有进程间相互通信，主进程收集所有从进程的$S_{tmp}(E)$和$H_{tmp}(E)$并累加计算出全局的S(E)和H(E)，即全局的S(E)=S(E)+所有从进程的$S_{tmp}(E)$，全局的H(E)=H(E)+所有从进程的$H_{tmp}(E)$，将全局的S(E)和H(E)的广播给所有从进程，判断直方图平缓条件：

[0100] $$\frac{\max(H(E))-\min(H(E))}{\max(H(E))+\min(H(E))}<\phi \quad (0<\phi<1)$$

[0101] 若不满足则返回执行步骤S12继续迭代；若满足则执行步骤S17；

[0102] 步骤S17：改变修正因子df，再返回执行步骤S12继续迭代，直到满足进程终止条件$df<\varphi$，（也即$f<e^{\varphi}, 0<\varphi<1$，如$\varphi$可取0.0001）；求得S(E)，得到蛋白质系统相对的态密度$g(E)=e^{S(E)}$。改变修正因子f的方式为：

[0103] 先连续进行N次迭代的$f=f^{\alpha}$（$0<\alpha<1$），再进行1次迭代的$f=f^{1/\alpha^{N-1}}$；

[0104] 反复重复上述方式。

[0105] 从进程包括如下步骤：

[0106] 步骤S21：初始化蛋白质系统态密度函数的对数S(E)=lng(E)=0，$S_{tmp}(E)=lng_{tmp}(E)=0$，直方图H(E)=0，$H_{tmp}(E)=0$（$E_{min}\leq E\leq E_{max}$），修正因子df=1（=lnf=lne）；

[0107] 步骤S22：s=1；

[0108] 步骤S23：依照所确定的蛋白质能量区间的分段方式将能量区间（$E_{min}\leq E\leq E_{max}$）分成M段，并分配到M个分线程中，t=1；

[0109] 步骤S24：在每个分线程中，对原来的构型限制在相应的子能量区间里进行随机变动，产生新的构型，计算能量E_{new}，根据Metropolis准则确定新构型被接受的概率：（简称为MCS步）

[0110] $$P(old\rightarrow new)=\min(1,e^{-[S(E_{new})-S(E_{old})]})$$

[0111] 若接受新构型，则：

[0112] $S(E_{new})=S(E_{new})+df$，$H(E_{new})=H(E_{new})+1$，

[0113] $S_{tmp}(E_{new})=S_{tmp}(E_{new})+df$，$H_{tmp}(E_{new})=H_{tmp}(E_{new})+^{1}$；

[0114] 否则：

[0115] $S(E_{old})=S(E_{old})+df$，$H(E_{old})=H(E_{old})+1$，

[0116] $S_{tmp}(E_{old})=S_{tmp}(E_{old})+df$，$H_{tmp}(E_{old})=H_{tmp}(E_{old})+1$。

[0117] t=t+1，所述步骤S24循环tmax次（即经过tmax次（如10次）MCS步）；

[0118] 步骤S25：所有线程间相互通信，综合得到整个区间的$S_{tmp}(E)$和$H_{tmp}(E)$，s=s+1；

[0119] 所述步骤S24和S25循环smax次（如100次）；

[0120] 步骤S26：所有进程间相互通信，从进程将$S_{tmp}(E)$和$H_{tmp}(E)$发送给主进程，然后接收经主进程计算得出的全局的S(E)和H(E)更新原来的S(E)和H(E)，将$S_{tmp}(E)$和$H_{tmp}(E)$初始化为0，判断直方图平缓条件：

[0121] $$\frac{\max(H(E))-\min(H(E))}{\max(H(E))+\min(H(E))}<\phi \quad (0<\phi<1)$$

[0122] 若不满足则返回执行步骤S22继续迭代；若满足则执行步骤S27；

[0123] 步骤S27：改变修正因子df，再返回执行步骤S22继续迭代，直到满足进程终止条件 $df<\varphi$，（也即 $f<e^{\varphi}$，$0<\varphi<1$ 如 φ 可取0.0001）。改变修正因子f的方式为：

[0124] 先连续进行N次迭代的 $f=f^{\alpha}$ $(0<\alpha<1)$，再进行1次迭代的 $f=f^{1/\alpha^{N-1}}$；

[0125] 反复重复上述方式。

[0126] 本发明和经典的WangLandau算法相比，使用基于退火机制的更新修正因子方式可提高计算精度和速度，利用一种灵活的能量区间分段方式可使分线程间负载均衡，采用MPI+OpenMP混合编程模型的混合并行方式可大大加快模拟和计算速度。MPI+OpenMP混合编程模型可以充分利用这两种编程模式的优点，即MPI可以解决多处理器进程间的粗粒度通信，而OpenMP提供轻量级线程可以很好地解决每个多处理器计算机内部各处理器间的交互。

[0127] 本发明能有效模拟和计算蛋白质系统态密度，能进一步求解得到宽广温度范围内的很多热动力学量如比热等，因此能研究蛋白质折叠的整个热力学过程，进而对蛋白质折叠过程进行探索和研究。

[0128] 本领域普通技术人员可以理解实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：只读存储器（ROM，Read Only Memory）、随机存取存储器（RAM，Random Access Memory）、磁盘或光盘等。

[0129] 以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

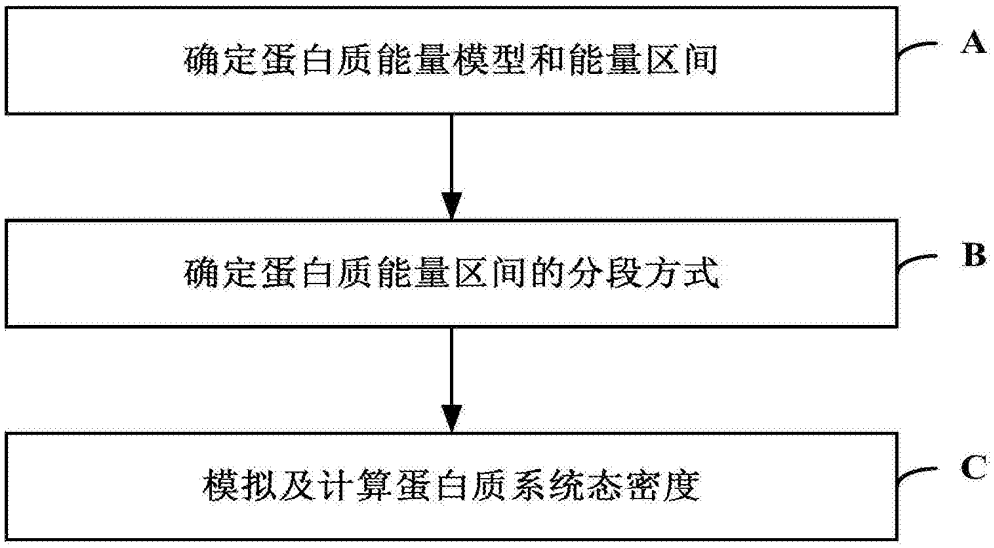

确定蛋白质能量模型和能量区间
A
确定蛋白质能量区间的分段方式
B
模拟及计算蛋白质系统态密度
C

图1

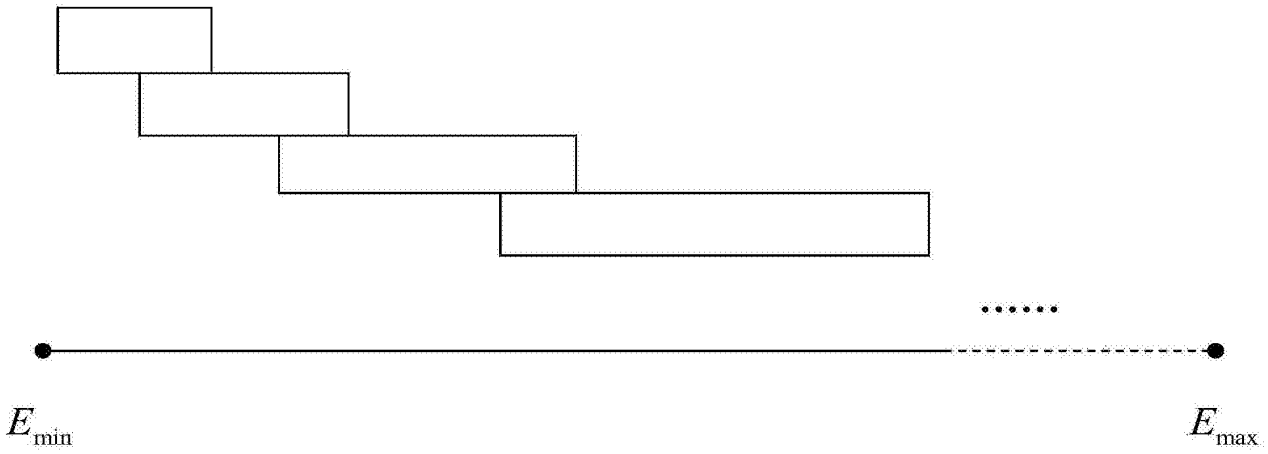

……
E_{min}
E_{max}

图2

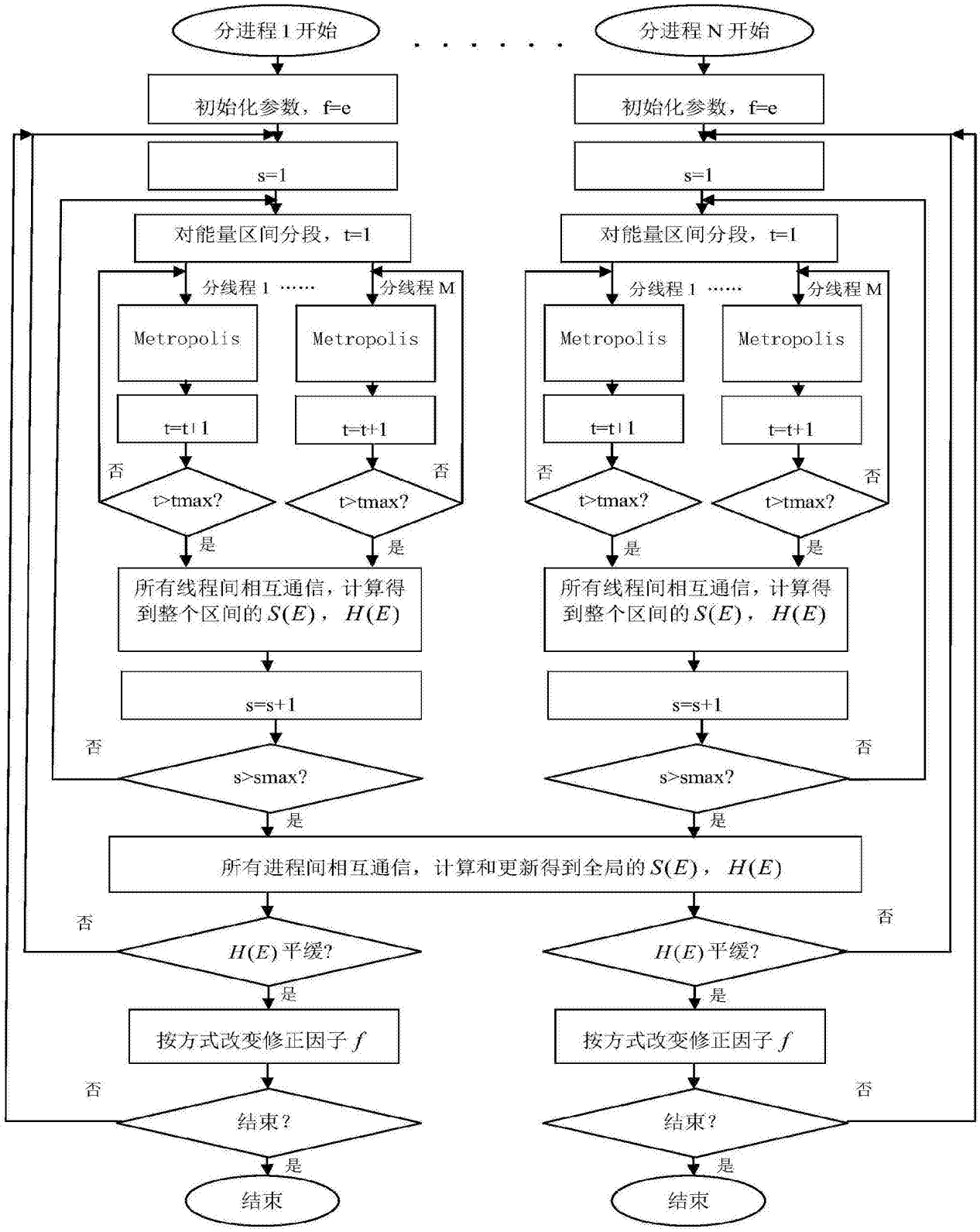

分进程 1 开始
分进程 N 开始
初始化参数，f=e
初始化参数，f=e
s=1
s=1
对能量区间分段，t=1
对能量区间分段，t=1
分线程 1 ……
分线程 M
分线程 1 ……
分线程 M
Metropolis
Metropolis
Metropolis
Metropolis
t=t+1
t=t+1
t=t+1
t=t+1
否
否
否
否
t>tmax?
t>tmax?
t>tmax?
t>tmax?
是
是
是
是
所有线程间相互通信，计算得到整个区间的 $S(E)$，$H(E)$
所有线程间相互通信，计算得到整个区间的 $S(E)$，$H(E)$
s=s+1
s=s+1
否
否
s>smax?
s>smax?
是
是
所有进程间相互通信，计算和更新得到全局的 $S(E)$，$H(E)$
否
否
$H(E)$ 平缓?
$H(E)$ 平缓?
是
是
按方式改变修正因子 f
按方式改变修正因子 f
否
否
结束?
结束?
是
是
结束
结束

图3