PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2014, 41(7): 712~718
www.pibb.ac.cn

一种在格点模型上模拟蛋白质折叠结构的优化算法 *

刘景发[1, 2)**] 宋蓓蓓[2, 3)] 刘朝霞[1)] 孙媛媛[2, 3)] 黄维波[2, 3)]

([1)] 南京信息工程大学网络信息中心，南京 210044；[2)] 南京信息工程大学计算机与软件学院，南京 210044；
[3)] 南京信息工程大学江苏省网络监控中心，南京 210044)

摘要 蛋白质折叠问题是生物信息学中一个经典的多项式复杂程度的非确定性(non-deterministic polynomial，NP)难度问题．势能曲面变平法(ELP)是一种启发式的全局优化算法．通过对 ELP 方法中的直方图函数提出一种新的更新机制，并将基于贪心策略的初始构象的产生，基于牵引移动的邻域搜索策略与 ELP 方法相结合，为面心立方体(FCC)格点模型的蛋白质折叠问题提出一种改进的势能曲面变平(ELP+)算法．采用文献中 9 条常用序列作为测试集．对于每条序列，ELP+算法均能找到与文献中的算法所得到的最低能量相等或更低的能量．实验结果表明，ELP+算法是求解 FCC 格点模型的蛋白质折叠问题的一种有效算法．

关键词 蛋白质折叠问题，势能曲面变平法，牵引移动，FCC 格点模型
学科分类号 180.1440，180.1450，520.1040 **DOI**: 10.3724/SP.J.1206.2013.00304

研究蛋白质的折叠机理并有效地预测其结构是生物信息学的核心问题之一．目前有一些实验方法来预测蛋白质的天然结构，如：核磁共振、X 射线晶体衍射法等．然而，这些方法不仅耗资耗时，而且能测的结构也很受限制，因此从理论上预测蛋白质的结构势在必行．根据 Anfinsen 的热力学假说[1]，蛋白质的自由能最低态及三级结构可以直接根据其氨基酸序列来进行预测．然而，即使对于最简单的疏水 - 亲水(HP)格点模型，这也是一个多项式复杂程度的非确定性(non-deterministic polynomial，NP)难度问题[2]．因此，寻找一个可以很好地反映氨基酸之间相互作用的模型和一个可以高效地搜索蛋白质最低能量构象的方法非常必要．

在 HP 格点模型中，二维正方形[3]和三维立方体[4]模型使用得最多，但这两个模型都存在奇偶性问题，即在序列中距离是偶数的氨基酸在折叠后的构象中不会拓扑相邻，进而也就不会产生相互作用力．为此，Hart 等[5]提出了一个面心立方体(FCC)格点模型．本文研究 FCC 格点模型的蛋白质折叠问题．一方面，该模型不存在奇偶性问题，同时，FCC 格点模型产生的结构也比正方形和立方体格点模型产生的结构更相似于真实蛋白质分子的结构．

目前，国内外应用于 FCC 格点模型进行蛋白质折叠研究的方法主要是一些启发式算法，如，简单遗传算法[6-7]及其变异算法(包括结合了牵引移动[8]与改进的交叉和变异操作的混合遗传算法[6]、带双胞胎移除策略的混合遗传算法[7, 9-10]、基于精英繁殖策略的遗传算法[11]和基于爬山策略的混合遗传算法[11])与带牵引移动策略的禁忌搜索算法[10]等．本文提出一种改进的势能曲面变平法(ELP+)．在 ELP+中，对势能曲面变平法(ELP)[12]提出了新的直方图更新机制，另外将形成初始构象的贪心策略及产生邻域解的牵引移动法融入到原始的 ELP 中．

ELP 算法最初由 Hansmann 和 Wille[12]提出，用来模拟全原子蛋白质折叠，后被一些学者引入到非格点模型[13]的蛋白质折叠问题和圆形 packing 问题[14]中．目前很少有研究者将 ELP 算法应用到离散优化问题中．为了验证 ELP 算法在离散空间的高效

* 国家自然科学基金(61373016), 江苏省自然科学基金(BK2010570), 江苏省“六大人才高峰”项目(DZXX-041), 中国博士后科学基金(201104572)和江苏省博士后科学基金(1001030B)资助.
** 通讯联系人.
Tel: 025-58695637, E-mail: jfliu@nuist.edu.cn
收稿日期：2013-07-01，接受日期：2013-09-06

性，我们将其应用于 FCC 格点模型，对给定的氨基酸序列进行结构预测．实验结果表明，ELP 算法是解决 FCC 格点模型蛋白质折叠问题的一个高效算法．

1　FCC 格点模型

在 FCC 格点模型中，氨基酸被简化为一组 H(疏水性或非极性)和 P(亲水性或极性)氨基酸的集合．FCC 格点模型的蛋白质折叠就是寻找氨基酸序列的自回避路径，即将氨基酸序列折叠到 FCC 格点平面或空间中，使得序列中相邻的氨基酸占据相邻的格点，并且一个格点只能被一个氨基酸占据．事实上，2D FCC 格点模型是一个无限图 $G=(V, L)$，其中顶点集 $V=(\sqrt{3}\cdot Z\times Z)\cup((\sqrt{3}\cdot Z+\sqrt{3/2})\times(Z+1/2))$，边集 $L=\{(x, x')|x, x'\in V, \|x-x'\|=1\}$，这里 Z 表示整数集，$\|x-x'\|$表示 x 和 x'之间的欧氏距离．3D FCC 格点可以看作是 2D FCC 格点的一个堆叠．在 2D FCC 格点模型中，一个氨基酸最多有 6 个邻点(图 1a)，而在 3D FCC 格点模型中，一个氨基酸最多有 12 个邻点(图 1b)．给定一个构象 c，其能量 $E(c)$被定义为拓扑相邻(排除序列中相邻的情况)的疏水氨基酸对的数量的负值．例如，在图 2 中，一条长度为 25 的氨基酸序列其二维构象的能量为–12．

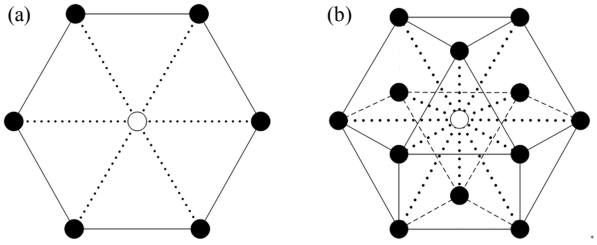

Fig. 1　FCC HP lattice model

(a) A unit of the 2D FCC HP lattice model. (b) A unit of the 3D FCC HP lattice model.

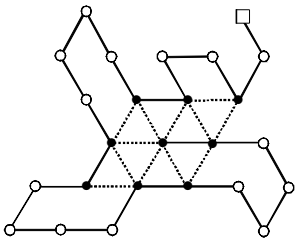

Fig. 2　A conformation of an amino acid sequence with length 25

"● " and "○ " ("□ ") indicate the hydrophobic and hydrophilic amino acids, respectively. "——" denotes the binding edge and "……" is the topological neighboring contact edge. "□ " is the first amino acid of the sequence.

蛋白质折叠问题的数学形式化提法为：给定一个氨基酸序列 $s=s_1s_2\cdots s_n$(s_i 为 H 或 P)，寻找 s 的最低能量构象，即找到 $c^*\in C(s)$，使得 $E(c^*)=\min\{E(c)|c\in C(s)\}$，其中 $C(s)$是 s 的所有合法构象的集合．

2　折叠结构模拟方法

2.1　势能曲面变平法

势能曲面变平(ELP)法[12]是一种启发式全局优化算法，也是一种 Monte Carlo(MC)方法．ELP 方法通过定义新的能量函数使搜索避开已访问过的区域．也就是说，如果一个构象 c 被采样，则其能量 $E(c, t)$将被$\widetilde{E}(c, t)=E(c, t)+f(H(q, t))$替换，其中惩罚项 $f(H(q, t))$是直方图 $H(q, t)$的函数，q 是有序参数，t 是 MC 迭代步数．本文中，令 $q=E, f(H(q,t))=kH(E(c, t), t)$，其中 k 是常数．如果能量 $E(c, t)$落入直方图的某个区间(每个区间的大小均设置为 1)，则该区间相应的频数就加 1．构象的抽样权重定义为 $\omega(\widetilde{E}(c, t))=\exp[-\widetilde{E}(c, t)/k_BT]$，其中 k_BT 是在低温 T 处的热能，k_B 是玻尔兹曼常数．

在 ELP 迭代中，随着局部极小点访问次数的增多，惩罚项将增加，局部极小点的抽样权重将减少而不再被优先采样，此时，计算将跳出该局部极小值点，并搜索权重更高的区域．然而，搜索又将很快落入一个新的局部极小点．这种过程不断重复，直到原始的势能曲面逐渐变平．然而，ELP 存在一个技术缺陷，即当前构象 c_1 通过邻域搜索(如 **2.3** 节中的牵引移动)产生新的构象 c_2 之后，只有在满足 random(0,1)<$\exp\{[\widetilde{E}(c_1, t)-\widetilde{E}(c_2, t)]/k_BT\}$(random(0, 1)表示 0～1 之间的随机数)的条件下才接受 c_2．这样，ELP 模拟可能会错过 c_1 周围的更低能量的构象．为此，文献[14]给出了另一个版本的 ELP：c_2 的接受与否由 $E(c_1, t)$和 $E(c_2, t)$的比较来决定，有两种情况：(a)$E(c_2, t)<E(c_1, t)$；(b)$E(c_2, t)\geqslant E(c_1, t)$．在情况(a)下，$c_2$ 被无条件接受；在情况(b)下，如果 c_2 满足 random(0, 1)<$\exp\{[\widetilde{E}(c_1, t)-\widetilde{E}(c_2, t)]/k_BT\}$，则 c_2 也被接受，否则 c_2 不被接受．然而，在原始的 ELP 方法[12]和 Liu 等提出的改进版[14]中，还存在一个缺陷，即不管利用邻域搜索产生的新构象是否被接受，直方图函数都会更新．这种更新机制会导致搜索中处于极小点周围未被接受的构象，可能在接下来的模拟中，由于极小点周围势能壁垒的增高更不会被接受，以至于模拟无法跳出局部极小点．

为此本文进一步改进了 ELP 方法，并且提出一种新的直方图函数更新机制，即只有当新产生的构象 c_2 被接受时才更新直方图.

2.2　初始构象的产生

在原始的 ELP 方法[12]及 Liu 等[14]提出的改进版中，算法可能从一个无效的初始构象开始迭代，而本文提出的改进的 ELP 方法是从一个有效的初始构象开始. 具体步骤如下：对于二维模型，首先将第一个和第二个氨基酸分别放在(0, 0)和(1, 0)两个格点上；而对于三维模型，则将其放在(0, 0, 0)和(0, 1, 1)上. 接着将第 i ($3\leqslant i\leqslant n$)个氨基酸试放在与第 i–1 个氨基酸相邻的空位格点上，这里试放是指第 i 个氨基酸是暂时放置的，在计算部分构象(指前 i–1 个氨基酸和第 i 个氨基酸构成的构象)的能量后将其移除. 计算将第 i 个氨基酸试放在每个可能的位置上产生的部分构象的能量，最后将第 i个氨基酸正式放在使部分构象能量最低的格点位置上. 如果不存在与第 i–1 个氨基酸相邻的空位格点，则移除第 i–1 个氨基酸，从第 i–2 个氨基酸开始继续生成构象. 重复上述过程直到产生具有 n 个氨基酸的构象.

2.3　邻域搜索策略

本文采用牵引移动法[8]作为邻域搜索策略来更新构象. 牵引移动最初由 Lesh 等[8]针对正方形和立方体格点模型提出. 牵引移动集对于正方形、立方体和 FCC 格点模型均具备完整性和可逆性[8, 10]，这就保证了其更新构象的高效性和全局最小值的可达性.

2D FCC 格点模型中牵引移动的主要思想如下：首先从当前构象中随机选择一个氨基酸 v，若在格点平面上至少存在一个空位格点 A，使得 v 及其前一个(或后一个)氨基酸都与 A 相邻，则将 v 移到 A(图 3a). 这可能会破坏构象，需要通过牵引移动来修复，即将 v 的前一个(或后一个)氨基酸移动到 v 原来的位置，若没有形成一个有效构象，则继续修复，直到得到一个有效的构象(图 3b).

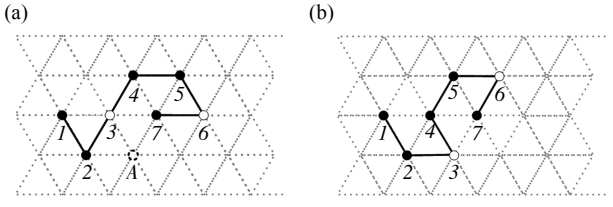

Fig. 3　An example of pull-move

"● " and "○ " indicate the hydrophobic and hydrophilic amino acids, respectively. As shown in (a), if position *A* is free, then amino acid 3 can be placed at *A*, and amino acid *4* can be moved to the position of amino acid *3*, *5* to *4*, *6* to *5*. Thus a valid conformation (see (b)) is obtained.

3D FCC 格点模型中的牵引移动也是类似的. 不同的是，二维模型中，每个氨基酸可以移动的相邻位置最多只有 6 个，而三维模型中最多有 12 个.

2.4　算法描述

通过对 ELP 方法中的直方图函数提出一种新的更新机制，并将基于贪心策略的初始构象的产生，基于牵引移动的邻域搜索策略与 ELP 方法相结合，为 FCC 格点模型的蛋白质折叠问题提出一种新的蛋白质折叠结构模拟方法——改进的势能曲面变平(ELP+)算法. 算法具体步骤如下：

(1)采用贪心策略产生一个有效的初始构象 c. 令 $\bar{c}=c$，$c_{\min}=c$. 初始化 $k=0.1$，$T=5$，$t=1$. 计算 $E(c, t)$，令其所在区间的直方图函数 $H(E(c, t), t)$的值为 1，其余区间的值为 0. 令$\tilde{E}(c, t)=E(c, t)+kH(E(c, t), t)$.

(2)从集合 $N=\{1, 2, \cdots, n\}$中随机选择一个整数 i.

(3)对于当前构象 c 的第 i 个氨基酸，如果存在合法的牵引移动空位格点，则对该空位格点进行牵引移动，并计算出所有成功执行牵引移动后所得到的构象的能量，最后挑出能量最低的构象作为 c 的更新构象，记为 c'，转(4)；否则转(7).

(4)计算 $E(c', t)$. 如果 $E(c', t)<E(c_{\min}, t)$，则令 $c_{\min}=c'$，$E(c_{\min}, t)=E(c', t)$.

(5)如果 $E(c', t)<E(c, t)$，则接受 c'，并令 $\bar{c}=c$，$c=c'$，更新 $H(E(c', t), t)$，令$\tilde{E}(c', t)=E(c', t)+kH(E(c', t), t)$，转(8)；否则转(6).

(6)如果 random(0, 1)<exp {$[\tilde{E}(c, t)-\tilde{E}(c', t)]/k_BT$}，则接受 c'，并令 $\bar{c}=c$, $c=c'$, 更新 $H(E(c', t), t)$, 令$\tilde{E}(c', t)=E(c', t)+kH(E(c', t), t)$，转(8)；否则转(7).

(7)如果 1 到 n 间的整数都被选过，则令 $c=\bar{c}$，转(8)；否则令 $N=N-\{i\}$，从 N 中选择另一个整数 j，令 $i=j$，转(3).

(8)如果 $t>2\times10^6$，则输出最低能量构象 $c_{\min}$，迭代结束；否则，令 $t=t+1$，转(2).

3　模拟结果与分析

我们在 1.6 GHz 的 Intel Core 2 Duo 的 PC机(内存为 2G)上用 java 语言实现了 ELP+算法(源程序可发邮件至 sbb_67348@126.com 索取). 本文所测试

的 9 条序列(表 1)来自于文献[10]和[11]. 对于 2D FCC 格点模型，每条序列分别独立运行 20 次，将每条序列运行所得到的最低能量和平均能量分别列于表 2 中，并与文献中的简单遗传算法(SGA)[6]、混合遗传算法(HGA)[6]、带双胞胎移除策略的混合遗传算法(HGA+TR)[10]、基于精英繁殖策略的遗传算法(ERS-GA)[11]、基于爬山策略的混合遗传算法(HHGA)[11]、禁忌搜索算法(TS)[10]以及基于牵引移动的不带新的直方图函数更新机制的 ELP 算法(简记为 ELP)的计算结果形成对比.

Table 1 Nine sequences for FCC HP lattice model

No.	Length	Sequence
1	20	$HPHP_2H_2PHP_2HPH_2P_2HPH$
2	24	$H_2P_2(HP_2)_6H_2$
3	25	$P_2HP_2(H_2P_4)_3H_2$
4	36	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$
5	48	$P_2H(PH_3)_2P_5H_{10}P_6(H_2P_2)_2HP_2H_5$
6	50	$H_2(PH)_3PH_4P(HP_3)_3P(HP_3)_2HPH_4(PH)_4H$
7	54	$H_2(PH)_3PH_4P(HP_3)_4P(HP_3)_2HPH_4(PH)_4H$
8	60	$P(PH_3)_2H_5P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$
9	64	$H_{12}(PH)_2((P_2H_2)_2P_2H)_3(PH)_2H_{11}$

从表 2 中可以看出，对于 2D FCC 格点模型，ELP+很容易找到文献中 4 条短序列 *1*～*4* 所给出的最低能量，而对于长度分别为 48、50、54、60 和 64 的长序列，ELP+找到的最低能量均优于文献中所给出的最低能量. 实际上，对于长度为 20 的序列，除了 SGA 外，其余 7 种算法都能找到最低能量. 对于序列 2，ELP+得到的最低能量与 HHGA、TS 和 ELP 得到的相同，并且比 SGA、HGA、HGA+TR 和 ERS-GA 得到的更低. 对于序列 3，ERS-GA、HHGA、TS、ELP 和 ELP+均得到了最低能量为 –12 的基态构象，而 SGA、HGA 和 HGA+TR 只找到了能量为–10 的次优解. 8 种方法中只有 3 种(TS、ELP 和 ELP+)找到了序列 *4* 的最低能量–24. 值得注意的是，对于序列 *5*～*9*，ELP+找到了新的最低能量，分别为 –44，–41，–42，–71 和–75. 表 2 中括号中的数字表示 ELP+方法运行 20 次所得到的平均能量值. 表 2 中最后一列中的 Time 和 MC step 分别表示 ELP+对于每条序列在 20 次运行中所能找到最优解的次数和找到最优解所需的平均迭代步数. 表 2 中最后一列显示，对于序列 *1*～*4*，ELP+每次都能快速地找到最优解，而对于稍长的序列，ELP+需要更长的时间来找到最优解，并且不能保证每次都能找到最优解. 这是合理的，因为随着氨基酸序列长度的增加，解空间将增大，能量曲面也会变得更为复杂. 为了验证新的直方图更新机制的效果，表 2 中还列出了原始的 ELP 所得到的最低能量和平均最低能量. 为了让比较结果更有说服力，对于所有的算例，ELP 和 ELP+所使用的参数是相同的. 从表 2 可以看出，除了 4 条短序列 *1*～*4*，两种方法得到的最低能量相同外，对于其余的 5 条序列，ELP+都能找到更低的能量. 同时，对于所有的序列，ELP+所得到的平均最低能量都比 ELP 得到的平均最低能量要低得多. 由 ELP+得到的序列 *7*、*8* 和 *9* 的最低能量的典型构象如图 4 所示. 很显然，每个构象都有一个很紧凑的疏水核结构.

Table 2 Comparison of computational results by different methods on 2D FCC HP lattice model

No.	Length	SGA[6]	HGA[6]	HGA+TR[10]	ERS-GA[11]	HHGA[11]	TS[10]	ELP	ELP+	Time/MC step[b)]
1	20	–11	**–15**	**–15**	**–15**	**–15**	**–15**	**–15**(–15)	**–15**(–15)[a)]	20/732
2	24	–10	–13	–13	–13	**–17**	**–17**	**–17**(–17)	**–17**(–17)	20/4321
3	25	–10	–10	–10	**–12**	**–12**	**–12**	**–12**(–12)	**–12**(–12)	20/2763
4	36	–16	–19	–19	–20	–23	**–24**	**–24**(–24)	**–24**(–24)	20/32034
5	48	NA[c)]	NA	–32	–32	–41	–40	–43(–43)	**–44**(–43.32)	9/253384
6	50	NA	NA	NA	–30	–38	NA	–36(–33.42)	**–41**(–40.14)	5/583347
7	54	–21	–23	–23	NA	NA	–31	–41(–38.43)	**–42**(–41.5)	9/168842
8	60	–40	–46	–46	–55	–66	–70	–69(–66.81)	**–71**(–70.08)	6/1203846
9	64	–33	–46	–46	–47	–63	–50	–70(–65.32)	**–75**(–74.08)	4/368623

[a)] Numbers in bold indicate the lowest energies so far and the numbers in parentheses denote the average energies. [b)] Time and MC step indicate the times of the runs which can find the lowest energies and the average number of energy evaluation before the lowest energy is found by ELP+. [c)] NA means data not available.

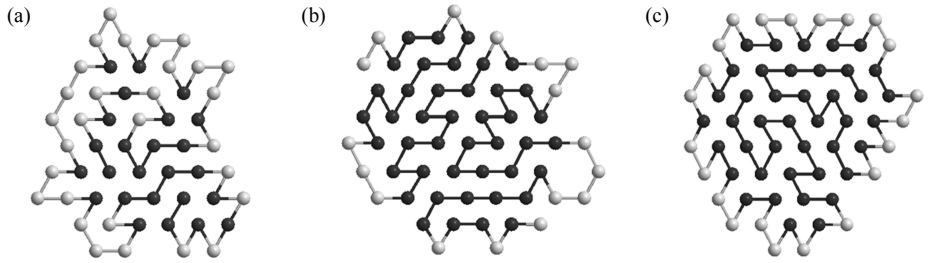

Fig. 4　Conformations with the lowest energies found by the ELP+ algorithm for sequences 7～9 on 2D FCC HP lattice model

(a) Typical conformation with E=–42 of instance 7. (b) Typical conformation with E=–71 of instance 8. (c) Typical conformation with E=–75 of instance 9. "● " and " ○ " indicate the hydrophobic and hydrophilic amino acids, respectively.

为了进一步验证 ELP+算法的高效性，我们又将该算法应用于 3D FCC 格点模型来模拟蛋白质的折叠结构．对于表 1 中的每条序列，分别独立运行 ELP+算法 20 次，并将 ELP+算法与 SGA[7]、HGA[7]、HGA+TR[7, 9]和 TS[10]所得到的最低能量列于表 3 中(对于序列 *1*～*4*，由于 ELP+ 算法每次均能很容易地得到文献中所给出的最低能量，所以表 3 中就不再列出)．从表 3 的计算结果可以看出，除序列 *6* 外(TS 算法没报道结果)，TS 和 ELP+算法所得到的最低能量均优于 SGA、HGA 和 HGA+TR 得到的最低能量．这说明 ELP+算法在 3D FCC 格点模型的蛋白质折叠结构预测中也是高效的．图 5 列出了 ELP+算法所模拟的序列 *7*～*9* 的 3D FCC 格点模型的最低能量构象．

Table 3　Comparison of computational results by different methods on 3D FCC HP lattice model

No.	Length	SGA[7]	HGA[7]	HGA+TR[7, 9]	TS[10]	ELP+
5	48	NA[a)]	NA	–69	**–74**	**–74**(–74)[b)]
6	50	–55	–59	–59	NA	**–73**(–72.2)
7	54	NA	NA	NA	**–77**	**–77**(–76.33)
8	60	–97	–114	–117	**–130**	**–130**(–130)
9	64	–81	–98	–103	**–132**	**–132**(–132)

[a)] NA means data not available. [b)] Numbers in bold indicate the lowest energies so far and the numbers in parentheses denote the average energies.

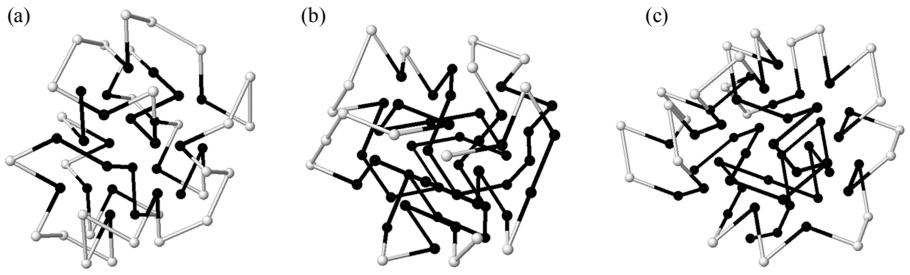

Fig. 5　Conformations with the lowest energies found by the ELP+ algorithm for sequences *7*～*9* on 3D FCC HP lattice model

(a) Typical conformation with E=–77 of instance *7*. (b) Typical conformation with E=–130 of instance *8*. (c) Typical conformation with E=–132 of instance *9*. "● " and " ○ " indicate the hydrophobic and hydrophilic amino acids, respectively.

4 结　论

庞大的搜索空间以及粗糙的势能曲面使得寻找蛋白质天然结构的过程异常复杂，搜索过程极易陷入局部极小点．势能曲面变平法(ELP)是一种新的全局优化算法．该算法中的 Metropolis 抽样准则和累积的直方图函数可以帮助搜索逃离局部极小点．目前，ELP 算法已成功地应用于连续空间，例如，非格点模型蛋白质折叠问题和圆形 Packing 问题．然而，很少有研究人员将其应用到离散空间．为了验证 ELP 算法在离散空间同样具有高效性，本文针对 FCC 格点模型蛋白质折叠问题提出了一种改进的 ELP 算法(ELP+)．该算法提出了新的直方图更新机制，同时采用了基于贪心策略的初始构象产生机制和基于牵引移动的邻域搜索策略．对于文献中的每条序列，无论是二维模型还是三维模型，ELP+ 算法均能找到与文献中的算法所得到的最低能量相等或更低的能量．实验结果表明，ELP 算法是一种搜索 FCC 格点模型蛋白质折叠低能构象的有效算法．今后，我们将进一步改进 ELP+ 算法，并将其应用于全原子模型来模拟真实蛋白质的结构．

参 考 文 献

[1] Anfinsen C B. Principles that govern the folding of protein chains. Science, 1973, **181**(4096): 223−230

[2] Unger R, Moult J. Finding the lowest free energy conformation of a protein is an NP-hard problem: proof and implications. Bull Math Biol, 1993, **55**(6): 1183−1198

[3] Chira C, Horvath D, Dumitrescu D. Hill-climbing search and diversification within an evolutionary approach to protein structure prediction. BioData Mining, 2011, **4**(23): 1−17

[4] Lin C J, Su S C. Protein 3D HP model folding simulation using a hybrid of genetic algorithm and particle swarm optimization. Int J Fuzzy Syst, 2011, **13**(2): 140−147

[5] Hart W E, Istrail S. Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than 86% of optimal. J Comp Biol, 1997, **4**(3): 241−259

[6] Hoque M T, Chetty M, Dooley L S. A hybrid genetic algorithm for 2D FCC hydrophobic-hydrophilic lattice model to predict protein folding // Sattar A, Kang B H. Proceedings of the 19th Australian Joint Conference on Artificial Intelligence, Hobart, 2006. Berlin: Springer, 2006: 867−876

[7] Hoque M T, Chetty M, Sattar A. Protein folding prediction in 3D FCC HP lattice model using genetic algorithm // Sriuivasan D, WANG Li-po. Proceedings of IEEE Congress on Evolutionary Computation, Singapore, 2007. Washington DC: IEEE, 2007: 4138−4145

[8] Lesh N, Mitzenmacher M, Whitesides S. A complete and effective move set for simplified protein folding // Vingron M, Istrail S, Pevzner P, *et al*. Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, Berlin, 2003. New York: ACM Press, 2003: 188−195

[9] Hoque M T, Chetty M, Lewis A, *et al*. Twin removal in genetic algorithm for protein structure prediction using low-resolution model. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, **8**(1): 234−245

[10] Bökenhauer H J, Ullah A D, Kapsokalivas L, *et al*. A local move set for protein folding in triangular lattice models // Keith C, Jens L. Proceedings of the 8th International Workshop on Algorithms in Bioinformatics, Karlsruhe, 2008. Berlin: Springer, 2008: 369−381.

[11] Su S C, Lin C J, Ting C K. An effective hybrid of hill climbing and genetic algorithm for 2D triangular protein structure prediction. Proteome Science, 2011, **9**(Suppl 1): S19

[12] Hansmann U H E, Wille L T. Global optimization by energy landscape paving. Phys Rev Lett, 2002, **88**(6): 068105

[13] Liu J F, Huang W Q. Studies of finding low energy configuration in off-lattice protein models. J Theo Comp Chem, 2006, **5**(3): 587−594

[14] Liu J F, Li G. Basin filling algorithm for the circular packing problem with equilibrium behavioral constraints. Science China: Information Sciences, 2010, **53**(5): 885−895

An Optimization Algorithm for Simulating Protein Folding Structures in Lattice Models*

LIU Jing-Fa[1,2)**], SONG Bei-Bei[2,3)], LIU Zhao-Xia[1)], SUN Yuan-Yuan[2,3)], HUANG Wei-Bo[2,3)]

([1)] *Network Information Center, Nanjing University of Information Science & Technology, Nanjing* 210044, China;

[2)] *School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing* 210044, China;

[3)] *Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing* 210044, China)

Abstract Protein folding problem is a classical non-deterministic polynomial (NP) hard problem in bioinformatics. The energy landscape paving (ELP) method is a class of heuristic global optimization algorithm. This paper applies the ELP method to simulate protein folding conformations for the hydrophobic-polar (HP) model on the face-centered-cube (FCC) lattice. By putting forward a new update mechanism of the histogram function in ELP and incorporating the generation of initial conformation based on the greedy strategy and the neighborhood search strategy based on pull-moves into ELP, an improved energy landscape paving (ELP+) method is put forward for the protein folding problem on the FCC lattice model. We test the method on nine benchmark sequences. The lowest energies by ELP+ are as good as or better than those of other methods in the literature for all instances. Computational results show that ELP+ is an effective method for protein folding problem on FCC lattice model.

Key words protein folding problem, energy landscape paving method, pull-moves, FCC lattice model

DOI: 10.3724/SP.J.1206.2013.00304

*This work was supported by grants from The National Natural Science Foundation of China (61373016), The Natural Science Foundation of Jiangsu Province (BK2010570), The "Six Talent Peaks" of Jiangsu Province (DZXX-041), Special Foundation of China Postdoctoral Science Foundation (201104572), and Jiangsu Planned Projects for Postdoctoral Research Funds (1001030B).

**Corresponding author.

Tel: 86-25-58695637, E-mail: jfliu@nuist.edu.cn

Received: July 1, 2013 Accepted: September 6, 2013