格子模型中蛋白质折叠的有效运动集

李小妹
（广东工业大学计算机学院　广州 510006）

摘　要　改进的遗传算法应用格子模型的一种新的运动集来优化蛋白质折叠过程。该新的运动集包含所有对称构象以及传统的运动模式，使其更适用于蛋白质折叠模拟。通过在格子模型中的实验我们发现，利用新的运动集明显优于传统使用的运动集。改进的遗传算法是一种进化算法，该算法适合搜索任何 HP 序列。

关键词　蛋白质折叠，格子模型，运动集

A New Move Set for Protein Folding in Lattice Model

LI Xiao-Mei
(Faculty of Computer, Guangdong University of Technology, Guangzhou 510006)

Abstract　An improved genetic algorithm applies the optimization procedures of protein folding through defining a new move set in lattice model. The new move set including rotation and mirror reflection is suitable for use in protein folding simulation. Through folding simulation on a simple lattice model it is found that the new move set is dramatically superior to that of the commonly used move set.

Keywords　Protein folding, Lattice model, Move set

1　引言

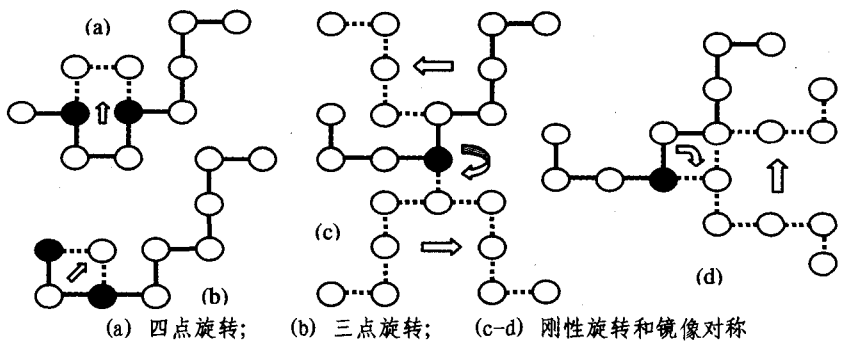
格子模型既简单，又具有真实蛋白质的很多特性，对于蛋白质的折叠问题，即使是这种最简化的格子模型却已证明是一种 NP 完备问题，因此只有通过启发式和近似算法来解决该问题。预测的最大困难在于庞大的搜索空间以及复杂的能量表面，其中包含很多的局部能量最小点和极少的全局能量最小点。

为了解决蛋白质折叠[1,2]的优化算法，目前的文献中有两种算法：进化算法和增长算法。前者包括模拟退火[3]、蒙特卡罗[4]、遗传算法[5]、改进的蒙特卡罗算法[6]以及遗传算法和禁忌算法的混合方法[7]。这种方法很容易陷入局部最小能量值，因此为搜索到最低能量构象，研究者们提出了各种改进策略来提高搜索性能。后者为串的增长算法，包括疏水内核引导的串增长算法[8]和基于权值的有偏串增长算法[9,10]。利用各种结构相关信息来提高算法的性能。通过前面一节的分析我们知道目前增长算法要比进化算法所需时间更短，得到的结果更优，但这种算法不适合搜索最低能量构象存在远程疏水残基间形成拓扑接触对的 HP 序列，因此本节通过提出一种新的运动集并结合蒙特卡罗和遗传算法来探讨蛋白质折叠的优化问题。

蛋白质的构象空间十分庞大，因此其折叠一定遵循着一定的途径使其从非折叠态到具有生物学功能的能量最小构象。从本节中可以看出，利用新的运动集的遗传算法也许能够在一定程度上用来模拟真实蛋白质的折叠途径。

2　运动集

格子模型的运动集可定义为一个单位时间步内构象可能发生的变化。若一个构象通过运动集中的一个运动转变为另一种构象，我们就称这两种构象相邻。也就是说，相邻的构象是构象空间的相邻点。多肽串的折叠与所选择的运动集有着十分密切的关系。虽然在目前的文献中提出了各种运动集[11~13]，但现已有的运动集并不完备，传统的运动集包括以下三大类：三点旋转，四点旋转和刚性旋转。这种运动集似乎简化了蛋白质折叠的运动途径，因此不能用来完整地模拟真实蛋白质的折叠。



(a) 四点旋转；　(b) 三点旋转；　(c-d) 刚性旋转和镜像对称

图 1　运动集 (a) 四点旋转；(b) 三点旋转；(c-d) 刚性旋转和镜像对称

李小妹　博士，主要研究方向：生物信息学和复杂系统的演化分析。

本文扩展了以前的运动集，将镜像对称包含进运动集，该运动集包含三类不同的运动(见图 1)，第一种为四点旋转(见图 1(a))，第二种为三点旋转(见图 1(b))，第三种为刚性旋转与镜像(见图 1(c-d))。违反自回避特性的运动应排除。图中的黑色球表示旋转关键点。

进一步对镜像对称结构进行描述，在二维网格模型中，我们将旋转和镜像关键点的坐标定义为(x,y)，将模型中的自回避路径沿着该关键点实施旋转和镜像，则坐标值存在两种置换，每种置换有四种符号组合，也就是$(\pm x, \pm y)$ $(\pm y, \pm x)$共 8 种对称结构，其中有两种肯定违反自回避特性，排除该构象本身，一种构象在二维格子模型中存在最多 5 种不同的对称结构(见图 1)。

尽管图 1 只描述了二维格子模型的对称结构，但该运动集很容易扩展到三维格子模型。基于同样的原理，我们定义对称关键点坐标为(x,y,z)，将模型中的自回避路径沿着该关键点实施旋转和镜像，则坐标值存在六种置换，每种置换有八种符号组合，也就是$(\pm x, \pm y, \pm z)$ $(\pm x, \pm z, \pm y)$ $(\pm y, \pm x, \pm z)$ $(\pm y, \pm z, \pm x)$ $(\pm z, \pm x, \pm y)$ $(\pm z, \pm y, \pm x)$共 48 种对称结构，其中有八种肯定违反自回避特性，排除该构象本身，在三维网格模型种，一种构象存在最多 39 种不同的对称结构。

3 算法

我们利用三维的格子模型对改进的蒙特卡罗算法(IMC)进行研究。该算法的步骤为：首先生成一条合法的随机串，记其构象的能量值为 E_1。接着随机选择一个旋转点，对该串实施运动集中所有不违反自回避特性的运动。在所有运动中，选择能量最低的构象作为当前构象的邻居构象，记其构象的能量值为 E_2；若所有构象均不合法，则重新选择旋转点。若邻居构象较当前构象能量更低，则选择邻居构象作为当前构象；若邻居构象能量高于当前构象，则根据蒙特卡罗算法的 Metropolis 规则以一定的概率($Ran<\exp[(E_1-E_2)/T]$，其中 T 为温度参数，该温度以一定的冷却速度下降)接受该邻居构象，该过程持续下去，直到找到最低能量构象或结束终止条件满足为止。

遗传算法是一种带指导性的随机搜索方法，通过选择合适的参数为问题找到一个理想的解。本质上是一种导向性猜测算法。这种导向性来自于候选解的适应值。猜测性来自于试图通过对适应值较高的候选解的组合和突变以期望获得更为理想的解。

蛋白质折叠改进的遗传算法 (IGA)的核心思想为进化，算法的具体描述如下：

(1) 随机生成构象种群，每个构象不能违反自回避特性，并计算每个构象的能量值或适应值。

(2) 突变操作：对种群中的每个构象实施一定步数的 IMC 循环。

(3) 交叉操作：根据适应值选择两个候选解进行交叉操作(在本节中选择交叉概率为 0.5)。选择一个随机的交叉点，将两个候选解在交叉点断开，重新组合成新解，在二维网格中有最多 6 种组合方式，在三维网格中，有最多 40 种组合方式。从中选出能量最低的构象作为候选构象的交叉子代。

(4) 将当前代的能量最低构象直接复制进入下一代。判断终止条件是否满足，若不满足返回第(2)步。其中的终止条件为已找到能量最低构象或达到一定的循环代数。

在蛋白质折叠的优化过程中，每个候选解先经历一定步数的突变，具体的突变操作利用 IMC 算法，只是选用不同的温度参数和冷却速度，突变的步数为蛋白质串的长度。

用作交叉操作的候选解的选取方式根据适应值概率选取，适应值的大小等于对应构象的能量且符号相反。适应值越大，能量越小的构象被选取的概率也就越大，因此选取的算法为轮盘选取法。

和突变步一样，交叉操作也可以产生新的候选解，在串中随机找一个旋转点，通过交叉操作产生两个新的候选解，尝试所有的连接方式，将能量最低的构象作为交叉子代；若所有连接构象均不合法，则重新选取交叉点。计算子代构象能量，并与两个父代能量的平均值进行比较，根据蒙特卡罗算法的 Metropolis 规则决定是否接受该子代构象。

4 实验结果

我们用表 1 给出的三维 HP 序列样例做了测试，利用不同的运动集测试了蒙特卡罗方法和遗传算法。

表 1 三维格子模型中蛋白质折叠问题的 HP 序列及其最优或已知的最低的能量值

No	Length	E	Protein sequence
			3D HP sequence
1	48	−32	$HPH_2P_2H_4PH_3P_2H_2P_2HPH_3PHPH_2P_2H_2P_3HP_8H_2$
2	48	−32	$H_4PH_2PH_5P_2HP_2H_2P_2HP_6HP_2HP_3HP_2H_2P_2H_3PH$
3	48	−32	$PHPH_2PH_6P_2HPHP_2HPH_2(PH)_2P_3H(P_2H_2)_2P_2HPHP_2HP$
4	48	−32	$PHPH_2P_2HPH_3P_2H_2PH_2P_3H_5P_2HPH_2(PH)_2P_4HP_2(HP)_2$
5	48	−32	$P_2HP_3HPH_4P_2H_4PH_2PH_3P_2(HP)_2HP_2HP_6H_2PH_2PH$
6	48	−32	$H_3P_3H_2PH(PH_2)_3PHP_7HPHP_2HP_3HP_2H_6PH$
7	48	−32	$PHP_4HPH_3PHPH_4PH_2PH_2P_3HPHP_3H_3(P_2H_2)_2P_3H$
8	48	−32	$PH_2PH_3PH_4P_2H_3P_6HPH_2P_2H_2PHP_3H_2(PH)_2PH_2P_3$
9	48	−32	$(PH)_2P_4(HP)_2HP_2HPH_6P_2H_3PHP_2HPH_2P_2HPH_3P_4H$
10	48	−32	$PH_2P_6H_2P_3H_3PHP_2HPH_2(P_2H)_2P_2H_2P_2H_7P_2H_2$

表 2 和表 3 给出了 10 条三维序列应用传统运动集和新运动集的蒙特卡罗方法和遗传算法的结果，表中的黑体表示三次循环的最好结果。表 2 给出了算法得到的最低能量值和得到该能量值时执行的 MC 步数，表 3 给出了算法得到的最低能量值和得到该能量值时执行的突变和交叉的总步数。对这 10 条测试样例，所用的蒙特卡罗方法，选定开始温度参数为 2，每 10,000 步 MC 步后温度降低到原来的 0.98 倍；所用的遗传算法选定种群大小为 200，循环 300 代，遗传算法中的

突变步开始温度参数为2，每循环5代温度参数降低到原来的0.97倍，遗传算法中的交叉步开始的温度参数为0.3，每循环5代温度参数降低到原来的0.99倍，对所有序列在相同参数和相同条件下运行3次。

表2 三维序列实施MC和IMC算法运行3次得到的结果

串号	应用传统运动集的MC算法			应用新运动集的IMC算法		
	1	2	3	1	2	3
1	−30(798,033)	−31(740,591)	−29(568,249)	−31(756,071)	−30(433,864)	−30(381,756)
2	−29(666,309)	−29(737,238)	−28(614,426)	−30(649,730)	−31(891,592)	−30(335,756)
3	−29(558,133)	−32(862,199)	−29(433,650)	−31(629,120)	−33(624,943)	−32(521,275)
4	−29(713,170)	−30(688,774)	−30(620,299)	−31(545,625)	−31(553,864)	−30(298,953)
5	−29(731,325)	−29(769,564)	−29(792,887)	−30(760,971)	−30(300,803)	−30(270,693)
6	−27(553,963)	−29(716,030)	−28(617,068)	−30(1,239,628)	−28(45,751)	−29(430,756)
7	−27(545,895)	−29(612,065)	−28(548,871)	−29(115,316)	−30(634,579)	−28(184,917)
8	−29(785,572)	−28(770,749)	−28(932,209)	−30(506,683)	−28(295,546)	−29(408,858)
9	−29(690,905)	−31(786,993)	−31(573,488)	−32(749,143)	−31(134,247)	−32(450,903)
10	−30(728,773)	−31(845,823)	−32(1,021,128)	−30(315,036)	−32(421,176)	−31(568,081)

从表2可看出除了三维序列的1号串和10号串利用两种不同运动集的蒙特卡罗方法得到了相同能量构象，其他8条序列利用新的运动集均得到了较传统运动集能量更低的结果。从表7可知有9条串利用新运动集的遗传算法较传统运动集能得到能量更低的构象，只有6号串利用两种运动集得到了相同的能量构象。表4给出了这10条三维HP序列的最低能量构象序列。利用新运动集找到了10条串中9条串的最低能量构象。

表3 三维序列实施GA和IGA算法运行3次得到的结果

串号	应用传统运动集的GA算法			应用新运动集的IGA算法		
	1	2	3	1	2	3
1	−31(1,503,394)	−30(1,494,369)	−31(1,451,041)	−31(482,447)	−32(685,353)	−32(750,145)
2	−31(1,241,536)	−30(1,213,086)	−30(960,432)	−34(829,914)	−32(863,255)	−32(1,192,364)
3	−32(906,550)	−31(1,304,485)	−32(1,251,45)	−33(434,301)	−34(535,197)	−34(1,076,581)
4	−31(1,714,537)	−31(1,185,200	−30(834,220)	−32(714,175)	−31(659,754)	−33(1,212,444)
5	−31(1,196,030)	−30(950,295)	−31(1,500,842)	−31(559,732)	−32(831,479)	−31(1,101,393)
6	−31(2,175,648)	−30(872,871)	−30(1,326,206)	−31(984,386)	−30(365,083)	−31(958,401)
7	−30(1,537,923)	−30(1,260,322)	−29(1,474,434)	−32(1,090,611)	−30(457,963)	−31(715,189)
8	−29(1,204,737)	−29(1,314,223)	−30(1,300,165)	−30(1,080,855)	−31(912,924)	−31(756,491)
9	−31(1,391,625)	−32(2,311,717)	−31(1,255,551)	−33(656,289)	−34(711,443)	−33(841,466)
10	−31(1,736,623)	−32(1,345,063)	−31(953,195)	−32(743,005)	−33(991,126)	−33(327,344)

表4 10条三维HP序列的最低能量构象序列

串号	结构序列
1	DLLLURRFRDLLULFRDDRURULULBUBDBRFFUBRDRDRFUULLDF
2	DLUURBLDDRURRULUFDDRDLDDBRUULDBUBUUFRDLLBDFLUUR
3	DBDDDFUULDDDLUBRBUFLFUBRUFUFRDFDBDDLUUFUBLDLDRD
4	DRDFFLBLBRBBLFDDRRBUUFDLFRRDLFULDBLULUUBRUFDFRUR
5	DDRBDLDFUFDFLBUBDLUUURDFFRBULLDFDBLBDBRRULURRUL
6	Global minimum was not reached.
7	DLBDRFLFRFFURBLBLURFLDDDFUUURUBRDBBBDFDDFUURFDL
8	DLLDDRRULFUULDDDLDDRRUURUURFLDFLUUBDLFDDBRULLBL
9	DRDRDBULULBDBDFRFFLUBDDLUUUBDBDFLFULFRURDDDLLUR
10	DLDDDRUUBDDDBULFLURBRULFURBRDFRULFRDLDFDBBURBLB

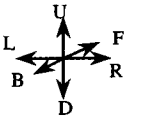


表5 三维格子模型中IMC和IGA与其他算法的比较

串号	IGA	IMC	MC	IGA	IMC	MC
1	−32(685,353)	−31(756,071)	30	−31(958,401)	−30(1,239,628)	30
2	−34(829,914)	−31(891,592)	30	−32(1,090,611)	−30(634,579)	31
3	−34(535,197)	−33(624,943)	31	−31(756,491)	−30(506,683)	31
4	−33(1,212,444)	−31(545,625)	30	−34(711,443)	−32(450,903)	30
5	−32(831,479)	−30(300,803)	30	−33(991,126)	−32(421,176)	30

三维格子模型利用IGA算法得到的结果与其他算法的结果比较见表5。表中IGA算法所示数据为表3利用IGA

每条序列运行3次得到的最好结果。IMC算法所示数据为表2利用IMC每条序列运行3次得到的最好结果。MC算法所示数据为文[17]报告的最小能量。通过利用新运动集和传统运动集在MC算法得到的结果比较可知，IMC方法在10条序列中有6条能找到较MC算法更低的能量构象，有2条串能得到相同的能量构象，只有两条序列所得结果要差于MC算法。

结论 利用方形网格模型，我们应用一种扩展的运动集，并与传统的运动集进行了比较，实验结果表明改进的遗传算法IGA能在最长的两条二维HP序列上比以前的算法找到更好的解。相比较而言，IGA方法在二维序列的7号串和8号串上均找到了目前已知的最优解，而前面的所有进化算法只是找到了次最优解。改进的遗传算法是目前格子模型中优化算法较为有效的一种算法。可以得出这样的结论，蛋白质折叠的优化过程与所选用的运动集有很大的关系，在折叠过程中，使用更为灵活的运动集能更快更好地找到最低能量构象。因此可以这么说，改进的遗传算法尤其适用于再现蛋白质折叠的折叠途径，同时我们也希望该运动集在处理蛋白质折叠的模拟过程中起到一定的作用。

参考文献

1 Berger B, Leighton T. Protein folding in the hydrophobic-hydrophilic model is NP complete, J. Comput. Biol, 1998(5):27～40
2 Paterson M, Przytycka T. On the complexity of string folding, Discrete. Appl. Math., 1996,71: 217～230
3 Kirkpatrick S, Gelatt Jr C D, Vecchi M P. Optimization by simulated annealing. Science, 1983,220:671
4 Unger R, Moult J. Genetic algorithms for protein folding simulations. J Mol Biol, 1993, 231:75
5 Konig R, Dandekar T. Improving genetic algorithms for protein folding simulations by systematic crossover. Biosystems, 1999, 50: 17～25
6 Liang Faming. Evolutionary Monte Carlo for protein folding simulations. J Chem Phys,2001, 115:3374
7 Jiang Tianzi. Protein folding simulations of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithms. J Chem Phys, 2003,119:4592
8 Beutler T C, Dill K A. A fast conformational search strategy for finding low energy structures of model proteins. Protein Science, 1996,5:2037～2043
9 Zhang J L, Liu J S. A new sequential importance sampling method and its application to the two-dimensional Hydrophobic- Hydrophilic model. J Chem Phys, 2002, 117(7): 3492～3498
10 Grassberger P. The pruned-enriched Rosenbluth method: simulations of Theta polymers of chain length up to 1000000. Phys Rev E, 1997, 56(3): 3682～3693
11 Shin J, Oh W S. Study of move set in cubic lattice model for protein folding. J Phys Chem, 1998, 102(33): 6405～6412
12 Nunes N L, Chen K, Hutchinson J S. A flexible lattice model to study protein folding. J Phys Chem, 1996, 100(24): 10443～10449
13 Yesylevskyy S O, Demchenko A P. Towards realistic description of collective motions in the lattice protein folding models. Biophysical Chemistry, 2004, 109(1): 17～40
14 Konig R, Dandekar T. Improving Genetic Algorithms for Protein Folding simulations by systematic crossover. Biosystems, 1999, 50(1): 17～25
15 Li H, Tang C, Wingreen N. Nature of driving force for protein folding: A result from analyzing the statistical potential. Physical review letters,1997, 79(4): 765～768
16 Błażewicz J, Łukasiak P. Application of tabu search strategy for finding low energy structure of protein. Artificial Intelligence in Medicine, 2005, 35(1-2):135～145
17 Yue K, Fiebig KM. A test of lattice protein folding algorithms. Proc Natl Acad Sci USA,1995, 92(1): 325～329

(上接第167页)

如果本体O_1的用户要在本体O_2中查找与本体O_1中的某个概念相关的信息，那么需要在O_1和O_2之间进行映射。以Wordnet为参考，得到概念的一个共享近义词表，如表1所示。利用共享近义词汇表，获得近义词间的最初相似度，然后再进一步计算相似度。

表1 共享近义词表

course	undergraduate courses, graduate courses, postgraduate courses,
people	staff, faculty, academic staff, technical staff, teacher
name	frist -name, last-name
job	professor, assistant-professor, associated-professor, lecturer, senior lecturer,

表2 数据类型匹配

匹配值	实型	整型	字符型	日期型
实型	1	0.9	0.1	0.7
整型	0.9	1	0.1	0.8
字符型	0.1	0.1	1	0.1
日期型	0.7	0.8	0.1	1

在参考层中，设定一个数据类型匹配表，如表2所示。相似度的计算由映射模块来完成。最后，输出相似矩阵。设定一个阈值，当相似度大于该值则认为两个概念相似并更新相似矩阵，否则这两个概念不相似并且相似矩阵保持不变。

总结与展望 两个本体间的映射是1：1，而多个本体间的映射有1：n和m：n。对于m：n关系的映射，可以分解成m个1：n关系的映射。让这m个用户按一定的优先级排队等待并进行查询。每个用户进行查询时就可以按1：n关系进行映射和查询。m个用户的查询是串行进行的。排队的顺序可以是先进先出或其它的排队方法。多个本体间的1：n映射可以转换成n对两个本体间的映射。这n对两个本体间的映射可以同时进行。

随着计算机的发展，本体的应用领域越来越多，本体的数量也越来越多。总的来说，本体的研究和应用还处于起步阶段，许多问题还需要近一步的研究。另外，本体还没有统一的生命周期定义和标准，也没有统一的本体开发的方法学和技术。因此，创建系统的、全面的、完整的方法体系仍是本体未来的研究方向。

参考文献

1 Kivela A, Hyvonen E. Ontological theories for the Semantic Web [M], Helsinki: HIIT Publications,2002. 111～136
2 Gruber T. Towards principles for the design of ontologies used for knowledge sharing [J]. International Journal of Human-Computer Studies,1995,43(5-6): 907～928
3 Gruber T R. A translation approach to portable ontology specification [J]. Knowledge Acquisition, 1993, 5(2):199～220
4 邓志鸿，唐世渭，张铭，等. Ontology研究综述[J]. 北京大学学报(自然科学版)，2002,38(5):730～738
5 Maedche A, Motik B. Ontologies for Enterprise Knowledge Management [J]. IEEE Intelligent Systems,2003. 26～33
6 Ehrig M, Sure Y. Ontology Mapping - An Integrated Approach [J]. In: Proceedings of the 1st European Semantic Web Symposium, Heraklion,Greece, Springer, LNCS,2004. 10～12
7 Doan A, Madhavan J, Domingos P. Learning to Map between Ontologies on the Semantic Web [J]. In: Proc. World-Wide Web Conf. ACM Press, May 2002. 662～673
8 Wiederhold G. An algebra for ontology composition [D], U. S. Naval Postgraduate School, Monterey CA, 1994
9 Macedche A, Motik B. MAFRA——A Mapping Framework for Distributed Ontologies [J]. Web Intelligence and Agent System, 2003,1: 235～248
10 KalfoglouY, SchorlemmerM. Information-flow-based ontology mapping [J]. In: Proceedings of the 1st International Conference,Springer, 2002. 1132～1151
11 Mitra P, Noy N F, Jaiswal A R. OMEN: A Probabilistic Ontology Mapping Tool [J]. In: Workshop on Meaning coordination and negotiation at the Third International Conference on the Semantic Web (ISWC-2004), Hisroshima, Japan