一种新的蛋白质结构预测多模态优化算法

程正华 张贵军 邓勇跃 金媚媚
(浙江工业大学信息工程学院 杭州 310023)

摘 要 针对现阶段药物设计中对于蛋白质结构多模态的需求,提出了一种基于排挤差分进化策略的多模态优化算法。为了降低蛋白质构象空间求解的复杂度,算法采用能量极小化过程,有效缩小了可行域的搜索空间;同时,为了有效地平衡多模态优化问题的局部收敛性和模态多样性,在排挤差分进化算法的框架下,在保证算法收敛速度的前提下,算法采用空间局部性原理,同时随机选取不同交叉策略的集结思想又有效改善了种群的多样性。以脑啡肽为例,算法不仅得到了其全局最稳定结构,还获得了一系列局部最优结构。

关键词 差分进化算法,多模态优化,空间局部原理,集结过程,能量极小化

中图法分类号 TP301.6 **文献标识码** A

Novel Multimodal Optimization Algorithm for Protein Structure Prediction

CHENG Zheng-hua ZHANG Gui-jun DENG Yong-yue JIN Mei-mei
(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract Aiming at the multimodal demand of protein structure for the drug design, a multimodal optimization algorithm based on differential evolution was proposed. In order to reduce the computation complexity of the protein conformational space, energy minimization is applied to narrow the search space of feasible region. For balance local minima convergence and modal diversity of a multimodal optimization, under the framework of crowding differential evolution algorithm, under the premise of ensuring the convergence rate, the algorithm uses the principle of spatial locality and builds up procedures which randomly select a crossing strategy to increase the diversity of the population individual. Taking Met-enkephalin as benchmark, the new algorithm finds not only the global minimum energy conformation, but also many other distinct local minima.

Keywords Differential evolution, Multimodal optimization, Spatial locality, Build up procedures, Energy minimization

1 引言

在现代的工业生产中,由于系统工作者和决策者所掌握的信息是有差异的,许多优化问题在实际条件下不仅需要一个全局最优解,还需要其它高质量的局部极值解因此系统工作者应该向决策者提供尽可能多的解决方案,由决策者根据问题的实际情况进行抉择[1]。例如使用计算机等技术进行蛋白质药物设计时,由于蛋白质结构预测所选取的力场模型的复杂性和实验的误差性,可能使得算法预测所得的全局稳定结构和实测靶点的结构并不能很好地吻合[2],这就需要设计一种多模态优化算法能够给出蛋白质其它的高质量的局部稳定结构。多模态优化实质上就是设计一种能够求出问题所有全局最优解和尽可能多的高质量局部极值解的全局优化算法,其对蛋白质药物设计具有重要的实际应用意义。

人类基因组计划的实施和完成,极大地促进了一批与生命相关学科的发展和融合,使得计算机辅助药物设计已不再局限于药物化学一个概念,而是逐渐发展成为一门由数学、化学、药学、分子力学、蛋白质组学和计算机科学等学科综合交叉的新兴学科。计算机辅助药物设计就是依据受体、酶以及核酸等潜在的药物设计靶点,利用计算机等相关知识设计药物分子。但是现在许多疾病涉及多基因、多靶点通路的问题,传统的、针对单一靶点的单模态研究方法已难以适用相关治疗药物的设计[3],使得多模态蛋白质结构预测成为当前生物信息学中非常重要的一个研究热点。

近几十年来,许多随机全局优化算法陆续被提出来解决多模态优化问题,如遗传算法(Genetic Algorithms, GA)[4]、差分进化(Differential Evolution, DE)[5]等算法,但是面对多模态问题优化时,大多数智能优化算法极易陷入局部极值解,必须与拥挤因子、适应度共享机制等小生境模型结合使用,才能找到尽可能多的极值解。2012年,K-C Wong教授提出了一种基于空间局部性原理的排挤差分进化算法(Crowding-DE-SL)[6],它能较好地解决多模态优化问题。针对蛋白质结构多模态预测这个研究热点,近年出现了一系列多模态优化算法,例如1997年Lee提出了构象空间退火(Conformational Space Annealing, CSA)[7]算法;2003年Klepies等人综合了αBB和CSA算法[8],得到了脑啡肽在ECEPP/3模型下一系

到稿日期:2012-11-16 返修日期:2013-05-09 本文受国家自然科学基金(61075062),浙江工业大学重中之重学科开放基金(20120811)资助。

程正华(1986—),男,硕士生,主要研究方向为现代智能优化算法、生物信息学,E-mail:zgj@zjut.edu.cn;张贵军(1974—),男,博士,教授,主要研究方向为生物信息学、智能信息处理、全局优化理论及其算法实现。

列稳定结构；2010 年，K-C Wong 等人采用 CGA-mixed 方法[9]针对 HP 晶格模型得到了蛋白质众多稳定结构；2011 年，Lee 等人将 DFA 思想引入到 CSA 中，得到了 CASP8 中 14 个目标蛋白质的结构[10]。

尽管蛋白质结构预测取得很多成果，但是要稳定、有效地搜索到蛋白质的稳定结构，同时还保证得到良好的模态分布，蛋白质结构预测仍然是一项艰巨的任务。其主要原因有 3 个[11]：1)要有一个合理的势函数来将蛋白质结构抽象转化成一个数学模型；2)蛋白质能量模型是一个高维的非凸函数，要保证算法在有效的计算时间找到势能函数的全局最优；4)在蛋白质分子设计过程中，可能算法预测所得的全局稳定结构并不满足实际的需求，那么只能采用其稳定性较好的局部稳态结构。因此，新的算法不仅要求更快地得到蛋白质的全局稳定结构，还要尽可能地找到一系列高质量的局部最优结构。

针对现阶段蛋白质结构预测所遇到的难题，本文提出一种新的混合算法 BECDE-SL(Build-up＋Energy-minimization ＋CrowdingDE-SL)。该算法设计的主要思想是在 DE 算法的基础上进行了 3 方面的改进：首先，改善的 CrowdingDE-SL 算法采用全局搜索和内在并行的搜索方式来快速地定位优化问题的全局最优解和局部极值解；其次，针对蛋白质预测问题中高维构象空间极其复杂的瓶颈，算法采用能量极小化过程对种群进行处理，大大降低蛋白质结构预测的搜索空间；最后，采用集结过程的思想直接继承种子个体中局部较好片段，避免算法运行时蛋白质结构中较好片段被算法破坏，同时不同的交叉策略又保证了种群具有好的模态分布。本文以脑啡肽(Try^1-Gly^2-Gly^3-Phe^4-Met^5)为实例，并同其它算法进行比较。

2 能量模型

根据 Anfinsen 的理论，蛋白质的天然结构对应于其自由能最小时的构象[12]。在数学上，其自由能可以建模为蛋白质分子内部的原子和基团之间相互作用的总函数。ECEPP 力场模型是依赖于原子三维坐标的经验势能函数，由于其忽略了电子的相互作用，使得分子力场模型结构相对简单，同时达到很高的精度，能够用于生物大分子的结构预测。本文采用 ECEPP/3 力场模型能量函数的表示形式如下[13]：

$$\begin{aligned} f_1(x^1,x^2,\cdots,x^N) &= E_{bond}+E_{angle}+E_{electrostatic}+E_{torsion}+E_{udw}+E_{hydrogen}+E_{other} \\ &= \sum_{b\in BOND}\frac{k_b}{2}(b-b_0)^2+\sum_{a\in ANGLE}\frac{k_a}{2}(a-a_0)^2+\sum_{i,j\in ES}\frac{q_iq_j}{\varepsilon\cdot r_{ij}} \\ &\quad+\sum_{\tau\in TOR}\sum_{m\in MUL}V_{m,\tau}[1+\cos(m\tau-\gamma_{m,\tau})]+\sum_{i,j\in VDW}\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}-\sum_{i,j\in HB}\frac{C_{ij}}{r_{ij}^{12}}-\frac{D_{ij}}{r_{ij}^{10}}+E_{other} \end{aligned} \tag{1}$$

式中，N 表示肽链中原子个数，x^i 为第 i 个原子的坐标(x_1^i，x_2^i，x_3^i)，$i=1,2,\cdots,N$，k_b、b_0、k_a、a_0、$V_{m,\tau}$、$r_{m,\tau}$、ε、q_i、q_j、A_{ij}、B_{ij}、C_{ij}、D_{ij} 等参数设置参考文献[13]。

对模型(1)进行平移及旋转变换后，优化问题维数为 $3N-6$，考虑到肽链的键长、键角均固定在平衡状态，设置 $E_{bond}=E_{angle}=E_{other}=0$，可将维数降至 $3N-E_{bond}-E_{angle}-6$。同时，为了消除转换过程中高阶的非线性等式约束条件，设 $r_{ij}=\zeta(\tau_1,\tau_2,\cdots,\tau_N)$，$i,j=(1,2,\cdots,N)$，$i\neq j$ 带入式(1)可得到(具体转换过程参考文献[14])：

$$\begin{aligned} \min\ & f_2(\tau)\equiv f_2(\phi_i,\psi_i,\omega_i,\chi_i^k) \\ &= f_1(\zeta(\tau_1,\tau_2,\cdots,\tau_N))\,|_{Ebond=Eangle=Eother=0} \\ \text{subject to}\ & -\pi\leqslant\phi_i\leqslant\pi,i=1,\cdots,N_{RES} \\ & -\pi\leqslant\psi_i\leqslant\pi,i=1,\cdots,N_{RES} \\ & -\pi\leqslant\omega_i\leqslant\pi,i=1,\cdots,N_{RES} \\ & -\pi\leqslant\chi_i^k\leqslant\pi,i=1,\cdots,N_{RES},k=0,1,\cdots,K_i \end{aligned} \tag{2}$$

式中，$\tau=(\tau_1,\tau_2,\cdots,\tau_N)\equiv\{\phi_i,\psi_i,\omega_i,\chi_i^k\,|\,i=1,\cdots,N_{RES},k=0,1,\cdots,K_i\}$为肽链二面角向量；$N$ 为肽链中二面角的自由度，N_{RES} 表示肽链长度(或残基)个数，K_i 为第 i 个残基侧链二面角的个数，且满足 $3N_{RES}+\sum_{i=1}^{N_{RES}}K_i=N$。

3 算法描述

3.1 基于空间局部原理的排挤差分进化算法

差分进化算法(DE)是 1995 年 Storn 等人提出的一种全局优化算法[15]，该算法采用基于种群的全局搜索策略和竞争生存策略对多个非劣解并行搜索，容易被改造成多模态混合优化算法。基本的 DE 算法是以变异的方式产生新个体，其数学表达式为：

$$v_{i,j}^{t+1}=x_{a,j}^t+mutatorFactor\cdot(x_{b,j}^t-x_{c,j}^t) \tag{3}$$

用差分进化这种群体优化算法处理多模态优化问题时，由于使用了全局选择因子，算法只能收敛到全局最优解，而忽略了众多局部极值解；其次，模型的复杂性造成这些算法极易陷入某个局优解；同时差分这种随机算法因缺乏全局收敛理论依据及解的不确定性而进一步限制了它们在实际问题中的应用。因此，Thomsen 2004 年将排挤因子融合到差分进化算法中，得到排挤差分进化算法(CrowdingDE)[16]。此算法通过替换相似亲代的方法更新种群，对于每一子代，根据欧式距离定义选择最近的一个种群个体，若子代适应值优于该亲代个体，则替换亲代个体；否则维持种群不变。

局部性原理借鉴了计算机高速缓存的设计策略，其主要思想是利用相邻或临近的个体预测和优化当前个体。2012 年，KC Wong 等人将空间局部性原理加入 CrowdingDE 中，我们称这种算法为 CrowdingDE-SL 算法。算法首先计算非亲代个体和亲代之间的距离，并使用转化函数将距离转换成轮盘形式，在轮盘中，那些和亲代个体比较近的非亲代个体占较大的份额。本文对 CrowdingDE-SL 算法做一点改进：首先，选取和亲代个体最相近的个体作为基准矢量 $x_{a,j}^t$；其次用轮盘形式随机选取产生矢量差的个体 $x_{b,j}^t$ 和 $x_{c,j}^t$；最后产生新个体 $v_{i,j}^{t+1}$。CrowdingDE-SL 算法对近亲个体进行各种操作，从而能够得到更好的后代，既保证算法快速收敛至极值点，又维持了种群的多样性。实验表明，此算法不仅能得到全局最优解，还能得到很多局部最优解。

3.2 集结过程

原始集结过程首先将分子分成小的片段，对各个小片段采用系统搜索的方法寻找其最低能量构象，然后把两个最低能量构象的片段连接起来组成新的片段，选择其中低能量的片段，迭代进行这种连接操作，最后得到整个分子的稳定构象。本文采用一种改进的集结过程，其主要思想是：采用不同的方法将蛋白质分子分成不同的片段组合，并对这些个体片段进行能量极小化处理。在种群交叉操作时，随机选取一种

组合方法，将种子中相应的片段直接复制给测试个体中对应的片段。这样算法继承了种子个体中能量较低的片段，既尽可能地保持种群的模态分布性，又能防止个体中的局部较好片段被算法破坏，加快了算法的收敛速度。

3.3 能量极小化过程

理想的蛋白质结构预测方法必须是基于能量极小化的理论计算方法，因为蛋白质是一个大分子的复杂体系，其能量势能面存在大量的局部极小值，如何避免陷入局部极小的陷阱、找到全局能量最小的稳定构象，成为解决蛋白质结构预测问题的关键。常见的能量极小化方法有分子动力学方法、最陡下降法、牛顿方法、共轭梯度法、随机搜索方法等。

差分进化算法具有很强的全局搜索能力，但是其局部搜索的能力较弱。本文采用拟牛顿法，提高了差分进化算法的局部搜索能力，大大降低了算法的搜索空间。特别是当群体进入全局最小的区域时，采用拟牛顿法可快速找到最优解。

3.4 BECDE-SL 算法描述

针对蛋白质结构预测这个多模态优化问题，在 CrowdingDE-SL算法的框架下，针对 ECEPP/3 势能模型，CrowdingDE-SL算法保证了不仅得到蛋白质的最低能量结构，而且尽可能地得到更多高质量的局部稳定结构；集结过程直接继承种子个体中较好片段，同时随机选取不同的交叉策略又保证了种群的多样性，使算法能够快速收敛到能量最低状态；能量极小化过程使算法避免早熟，降低搜索空间，提高了差分进化算法的局部搜索能力。

BECDE-SL 算法流程描述如下(CrowdingDE-SL 算法具体操作参照文献[10])：

Step1 算法初始化：设置种群规模 popSize，变异因子 mutatorFactor，交叉因子 crossFactor，初始种群 $POP=x^1,x^2,\cdots,x^{popSize}\mid x^i=(x_1^i,x_2^i,\cdots,x_N^i)i\in I$，计算 $f_2(x^i),i\in I$，并设 $f_2^*=\min_{i\in I}f_2(x^i)$，$I\in(1,2,\cdots,popSize)$。

Step2 对种群进行能量极小化处理，并对其排序，选取种群前 M 个个体作为种子。

Step3 对于每个目标 x^i 做以下处理：

3a)设置 i=1；

3b)计算种群中非亲代和亲代的距离，并由大到小排序，再通过高斯转换函数将距离转化为轮盘形式。

3c)首先选取距亲代最近的个体作为变异的基准矢量 x^a；然后在轮盘中随机选取个体 x^b，x^c，并保证 a≠b≠c；最后对{x^a，x^b，x^c}执行变异操作。

3d)为保证种群多样性，算法以不同的概率(m、n、p)从下列 3 种策略中选取一种执行交叉操作。

①以概率 m 执行基本的 DE 交叉策略。

②以概率 n 随机选取一个小组执行交叉操作。

③以概率 p 随机选取一个集合组执行交叉操作。

3e)对所得的测试个体进行能量极小化处理。

3f)对所得的测试个体执行基本 DE 的选择操作，如果测试个体比亲代好，则替换亲代个体，否则保持种群不变。

Step4 如果 i<popSize，i=i+1。

Step5 判断是否满足迭代终止条件(迭代 400 次)，如不满足，转至 Step2 继续执行操作。

注：

1)在执行变异操作前，首先对种群进行能量极小化处理，保证种子有能量较低的局部片段。

2)Step2 中的高斯函数形式及其距离转换具体过程参见文献[6]。

3)执行交叉操作前，随机从 M 个种子中选取一个做交叉操作中的种子个体。

4)在交叉操作过程中，若选取的是小组或集合组，将种子中和选取的集合组或小组相对应的局部片段直接复制给测试个体。

4 案例研究

脑啡肽(Try^1-Gly^2-Gly^3-Phe^4-Met^5)是由 5 个氨基酸组成的，它作为一种理想的评估新的蛋白质预测算法效率的基准，现在被广泛应用在各种蛋白质结构预测算法中。脑啡肽分子由 75 个原子组成，可用 24 个独立的主-侧链二面角描述，公认的脑啡肽稳定结构能量值为 −11.7073kcal/mol，其势能面上至少包含 10^{11} 个以上的局部最优解，优化难度极大，为了说明该优化问题的复杂性，我们使用多起点局部搜索方法对模型(2)进行优化：首先，随机产生 100000 个初始点；然后以这些起点进行能量极小化处理得到各模态的局部最优解；最后将得到的 100000 个局部最优解与公认的标准解进行结构对比，并计算 C_α-RMSD。从图 1 中可以清楚地看到，100000 个随机结构的局部优化都没有收敛到公认的脑啡肽稳定结构解。

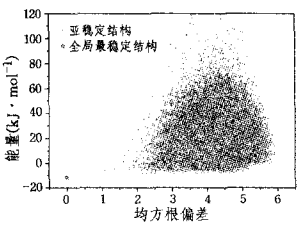

图 1 100000 局部稳定结构与全局稳定结构对比结果

算法以脑啡肽为例，将脑啡肽的二面角搜索范围定在 −180°到 180°之间，并将其对应的 24 个二面角分为 8 个小组，如图 2 所示，小组中的 φ、ψ、ω 代表脑啡肽主链中的二面角，χ^i 代表脑啡肽侧链中的二面角。在算法中，我们进一步将 8 个小组分为 7 个集合组，如表 1 所列。小组和集合组中的成员分别对应 24 个二面角中的某些片段，这些小组和集合组类似于原始集结过程中两种不同的片段。

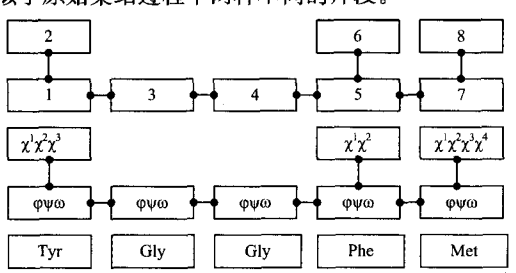

图 2 将 24 个二面角分为 1−8 个小组

表 1 将 8 个小组划分为 7 个集合组

小组	集合组
1	1,2,3
2	1,2,4
3	3,4,5
4	4,5,6
5	5,6,7
6	5,6,8
7	6,7,8

设计一种蛋白质结构多模态预测的优化算法必须要解决的难点是：由于蛋白质势能曲面上拥有大量的局部极小值，虽然算法得到大量的蛋白质结构，但是其中大部分是蛋白质的重叠结构，或者不是蛋白质稳定的模态结构。针对实验中所得的重叠结构，本文设定两个标准（式(4)和式(5)）来区分不同的模态：1）两个蛋白质结构解中二面角的 d_{ij} 不大于 24°（d_{ij} 中参数参考文献[7]）；2）两个蛋白质结构解中任意二面角 θ 相差不大于 5°。如果满足 1）、2）中任意一个，则认为这两个蛋白质结构为同一个模态。针对蛋白质的伪模态结构，算法同时采用能量极小化对这些解进行处理，使其尽可能地逼近离它们最近的模态。

$$d_{ij}=\sum_{k=1}^{n}\min[\mathrm{mod}\{(\theta_k-\theta_k'),sym(k)\},\{sym(k)-\mathrm{mod}\{(\theta_k-\theta_k'),sym(k)\}\}]\leqslant 24^\circ \tag{4}$$

$$|\theta_{1i}-\theta_{2j}|\leqslant 5^\circ(i\neq j=1,2,\cdots,24) \tag{5}$$

5 算法测试及其结果分析

为了同 BECDE-SL 算法进行对比，文中给出了其它 3 种同类的算法。算法 1 为 DE 算法＋能量极小化，算法 2 为 DE 算法＋集结过程＋能量极小化，算法 3 为 CrowdingDE 算法＋集结过程＋能量极小化，算法 4 为 BECDE-SL 算法。算法 1—4 各自针对模型(2)进行优化运算，并对实验结果进行对比。为了保证实验结果对比的公平性，算法 1 和 2、3、4 的实验参数保持一致，设定如下：popSize＝100，mutatorFactor＝0.9，crossFactor＝0.1，迭代次数为 400 次，每种算法都独立运行 50 次，实验中选取的种子个数 M＝10，随机概率 m＝0.5，n＝0.2，p＝0.3。针对 ECEPP/3 势能模型，当脑啡肽的能量值收敛至－11.7070 kcal/mol 时，我们认定其达到最稳态结构。同时在对数据进行处理时，当迭代至 400 次仍然没有达到稳态值时，我们认定算法在 400 次迭代时达到稳态值。实验以文献[13]中所得到的脑啡肽全局稳定结构解为公认稳定结构，实验结果以此解为标准进行对比。

表 2　4 种算法各方面性能的对比

算法	可靠性(%)	达到稳态的平均迭代次数	每次所得模态的平均个数	50 次所得模态总数(＜－10kcal/mol)	最低稳定结构(kcal/mol)
1	72	265.66	48.16	30	－11.7073
2	74	215.45	1.12	5	－11.7073
3	88	181.06	95.28	67	－11.7073
4	92	188.68	96.64	84	－11.7073

从表 2 运行结果中可以看出，算法 1—4 虽然都能搜索到脑啡肽能量最小值－11.7073kcal/mol 对应的结构，但是加入了集结过程和排挤因子的算法 3、4 运行结果明显较好。和算法 3 比较，空间局部性原理的加入虽然增加了算法 4 的复杂度，但是算法 4 在单个模态上的收敛速度（达到稳态的平均迭代次数/50 次所得到模态总数）比算法 3 要好（即表 2 的数据 188.68/84＜181.06/67）；同时在保证收敛速度的前提下，算法 4 在 50 次运行中有 46 次能够搜索到脑啡肽的全局最稳态结构，可靠性相对最好。如图 3 所示的是 4 种算法 50 次运行的种群平均能量分布图，算法 4 其种群的平均能量曲线整体较为平缓，并且误差波动小，表明算法 4 拥有更好的性能，得到了更多的脑啡肽模态结构。

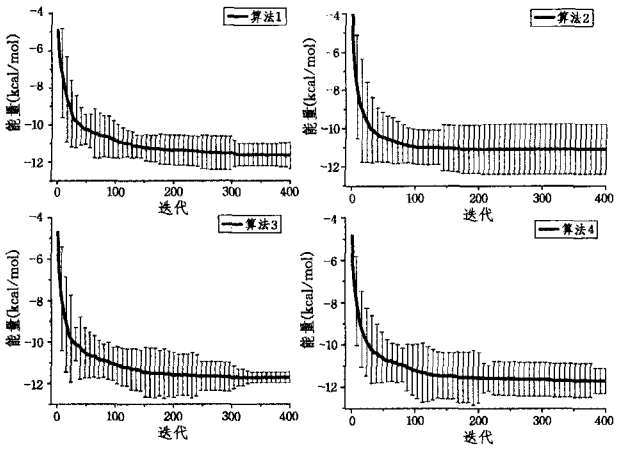

图 3　种群运行 50 次的平均能量图

由于蛋白质结构的多重对称性和实验的误差，多模态优化算法得到了大量的蛋白质结构，但其中有蛋白质的重叠结构和不稳定的伪模态结构。图 4 所示的是算法 1、2、3、4 在 50 次运行后得到的所有能量值小于－10 kcal/mol 的种群个体的聚类曲线，从图中可以直观地看出，算法 1 是 DE＋能量极小化，导致种群收敛特性不是很好，400 次迭代后，种群中也出现了一些脑啡肽的结构。从图 4 中看出，尽管算法 3 和算法 4 都是多模态优化算法，但是算法 4 种群个体的聚类曲线明显更加曲折，表明其有更好的种群多样性。算法 4 加入了空间局部原理，虽然增加小部分收敛速度，但是其平均每次运行就能得到 96.64 个不同的蛋白质结构，50 次运行总共得到了 679 个能量值小于－10kcal/mol 的脑啡肽结构，经过 d_{ij} 等限制条件的筛选，得到了 84 个独立的高质量脑啡肽稳定结构，明显比算法 3 所得到的运算结果要好。

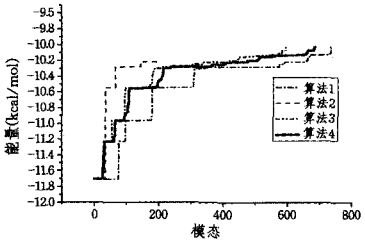

图 4　各算法 50 次运行所得的模态

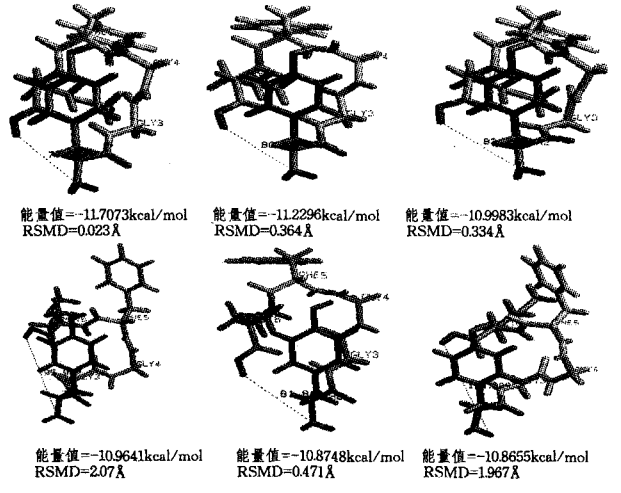

图 5　算法 4 所得到的 6 个高质量稳定结构的三维图

（下转第 229 页）

[15] Chattopadhyay R. A study of test functions for optimization algorithms[J]. Opt. Theory Appl. ,1971,8:231-236
[16] Schoen F. A wide class of test functions for global optimization [J]. Global Optimization,1993,3:133-137
[17] Shang Y W,Qiu Y H. A note on the extended rosenrbock function[J]. Evolutionary Computation,2006,14:119-126
[18] Shilane D,Martikainen J,Dudoit S,et al. A general framework for statistical performance comparison of evolutionary computation algorithms[J]. Information Sciences:an Int. Journal,2008,178:2870-2879
[19] Deep K,Bansal J C. Mean particle swarm optimisation for function optimisation[J]. Int. J. Comput. Intel. Studies,2009,1:72-92
[20] Eberhart R C,Shi Y H. Comparing inertia weights and constriction factors in particle swarm optimization[C]// Proceedings of the Congress on Evolutionary Computing. La Jolla,2000:84-89

(上接第 215 页)

表 3 是算法 4 所得到的全局最稳态结构的二面角解。表 4 给出的是算法 4 同 CSA 等算法所得的全局最稳态结构的精度对比。尽管这几种算法都能得到脑啡肽的全局最稳态结构,但是算法 4 所求的最优解同公认稳定结构之间的标准差为 0.606°,最接近公认的脑啡肽稳定结构。图 5 所示的是算法 4 所得到的脑啡肽全局稳定结构和其它 5 个高质量的局部稳定结构的三维图。从图中可以直观地看出,尽管能量值相差很小,但 6 个三维结构图有很大的差别。图 5 中还给出了这 6 种高质量脑啡肽稳定结构与公认稳定结构相比较的 C_α-RMSD。如图 5 中所示,能量值为 −10.8748kcal/mol 所对应的脑啡肽结构尽管能量值相比较大,但是其结构和公认稳定结构的误差为 0.471Å。由此表明,优化算法所得到的脑啡肽结构虽然其能量值较小,但是其三维结构和脑啡肽的全局最稳定结构可能有较大的误差。

表 3 算法 4 运行所得最稳定结构解

氨基酸	符号	二面角(°)	氨基酸	符号	二面角(°)
	ϕ_1	−83.499		ϕ_4	−136.847
	φ_1	155.793	苯丙	φ_4	19.097
酪氨酸	ω_1	−177.127	氨酸	ω_4	−174.09
Try^1	χ_1	−173.177	Phe^4	χ_1	58.857
	χ_2	79.384		χ_2	−85.476
	χ_3	−166.329		ϕ_5	−163.446
	ϕ_2	−154.261		φ_5	160.899
甘氨酸	φ_2	85.817	甲硫	ω_5	−179.79
Gly^2	ω_2	168.503	氨酸	χ_1	52.863
	ϕ_3	82.918	Met^5	χ_2	175.293
甘氨酸	φ_3	−75.029		χ_3	−179.868
Gly^3	ω_3	−169.97		χ_4	−58.589

表 4 各算法所得最优解的精度对比

算法	CSA	1	2	3	4
方差(Å)	0.672	0.625	0.618	0.628	0.606
标准差(Å)	0.812	0.791	0.786	0.793	0.779

结束语 本文针对现在药物设计对蛋白质结构多模态的需求,提出了一种新的多模态混合算法 BECDE-SL,新算法可较好地解决多模态问题,能优化存在的无法有效地平衡收敛性和多样性的问题,特别是针对蛋白质结构预测这种高维非凸函数的优化。算法在 CrowdingDE-SL 算法的框架下,对种群个体进行能量极小化处理,同时集结过程和空间局部原理大大加快了算法的收敛速度,不同的交叉策略又增加了种群的多样性。仿真结果表明,算法 BECDE-SL 同其它算法相比,不仅具有较好的收敛特性,而且保持了种群的多样性,能够得到众多高质量的脑啡肽稳定结构。

参 考 文 献

[1] 顾培亮. 系统分析与协调[M]. 天津:天津出版社,1998
[2] Bradley P,Misura K M S,Baker D. Toward high-resolution de novo structure prediction for small proteins[J]. Science,2005,309(5742):1868-1871
[3] 许忠能. 生物信息学[M]. 北京:清华大学出版社,2008:361-363
[4] Unger R,Moult J. Genetic algorithms for protein folding simulations[J]. Journal of Molecular Biology,1993,231(1):75-81
[5] Kalegari D H,Lopes H S. A differential evolution approach for protein structure optimization using a 2D off-lattice model[J]. International Journal of Bio-Inspired Computation,2010,2(3/4):242-250
[6] Wong K C,Wu C H,Peng C B,et al. Evolutionary multimodal optimization using the principle of locality[J]. Information Sciences,2012,194:138-170
[7] Jooyoung L,Harold A S,Shalom R. New optimization method for conformational energy calculations on polypeptides:conformational space annealing[J]. Journal of Computational Chemistry,1997,18(9):1222-1232
[8] Klepeis J L,Pieja M J,Floudas C A. A new class of hybrid global optimization algorithms for peptide structure prediction:integrated hybrids[J]. Computer Physics Communications,2003(151):121-140
[9] Wong K C,Leung K S,Wong M H. Protein structure prediction on a lattice model via multimodal optimization techniques[C]// GECCO'10 Genetic and Evolutionary Coputation Conference. Portland,USA:ACM,2010:155-162
[10] Juyong L,Jinhyuk L,Takeshi N S,et al. De novo protein structure prediction by dynamic fragment assembly and conformational space annealing[J]. PROTEINS:Structure,Function,and Bioinformatics,2011,79(8):2403-2417
[11] 来鲁华. 蛋白质结构预测与分子设计[M]. 北京:北京出版社,1993
[12] Anfinsen C B. Principles that govern the folding of protein[J]. Science,1973,181(96):223-230
[13] George N,Kenneth D G,Kathleen A P,et al. Energy parameters in polypeptides,10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm,with application to proline-containing peptides[J]. Journal of Physical and Chemical,1992,96(15):6472-6484
[14] Maranas C D,Floudas C A. Global minimum potential energy conformations of small molecules[J]. Journal of Global Optimization,1994,4(2):135-170
[15] Storn R,Price K. Differential evolution-a simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of Global Optimization,1997(11):341-359
[16] Thomsen R. Multimodal optimization using crowding-based differential evolution[C]// CEC2004,IEEE Congress on Evolutionary Computation. Portland Marriott Downtown,USA:IEEE,2004(2):1382-1389