计算机科学 2005Vol. 32№. 11

求解蛋白质结构预测问题的局部搜索算法*)

吕志鹏　黄文奇

(华中科技大学计算机科学与技术学院　武汉 430074)

摘　要　蛋白质结构预测问题是计算生物学领域的核心问题之一。通过理论计算的方法根据蛋白质氨基酸序列直接预测其空间结构是解决这一问题的有效途径。构造了新的邻域结构,采用了部分随机跳坑策略,对此问题提出了新的局部搜索算法。计算结果表明,该算法计算效率要优于传统的遗传算法和 Monte Carlo 方法。对于链长为 50 的算例还找到了文献中所没有的全新的最低能量构形。

关键词　蛋白质结构预测,格点模型,局部搜索,跳坑

Local Search Algorithm for Solving Protein Structure Prediction Problem

LU Zhi-Peng　HUANG Wen-Qi

(School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

Abstract　Protein structure prediction has proven to be one of the central problems in the field of computational biology. It is a feasible approach to predict theoretically the three-dimensional structure of proteins based only on amino acid sequence information. Using a new neighborhood structure and partly randomized off-trap strategy, a novel local search algorithm for protein structure prediction is proposed. Computational results demonstrate that our algorithm not only is more efficient than conventional genetic and Monte Carlo algorithms, but also find new configurations of lowest energy states missed in previous papers for the sequence of length $N=50$.

Keywords　Protein structure prediction, Lattice model, Local search, Off-trap

1　前言

蛋白质结构预测问题是通过蛋白质一级结构的氨基酸序列来预测其空间结构。对其求解是后基因时代蛋白质工程的主要任务之一。研究表明,蛋白质的生物学功能由它的空间结构决定。因此,预测蛋白质的空间结构在生物学领域具有重要的理论意义和现实意义。

迄今为止,对蛋白质结构预测问题已提出了一些简化模型。其中 Dill 等[1]提出的 HP 格点模型(HP Lattice Model),已得到学术界的广泛认同。尽管组成蛋白质的氨基酸种类有二十几种,但根据疏水作用和亲水作用可将其分为疏水氨基酸(Hydrophobic,用 H 表示)和亲水氨基酸(Polar,用 P 表示)。因此,蛋白质氨基酸序列可表示为一 H、P 字符串,如 HPHPPHHPHPPHPH(为方便起见,下文称 H 为黑球,P 为白球,球的半径为 1/2)。HP 格点模型中,要求将任一给定的氨基酸序列摆放在二维平面上,且满足下面的条件:

1. 所有球必须摆放在格点上,即球的横纵坐标值均为整数。

2. 任意两个不同的球不能重叠。

3. 任意两个在链上相邻的球摆放后在空间也必须相邻,即球心的欧氏距离为 1。

称满足上述条件的空间摆放为一个合法的构形。蛋白质结构预测问题就是要找出任意给定氨基酸序列满足上面 3 个条件的最低能量构形。能量定义只考虑疏水氨基酸之间的相互作用力,即在序列中不相邻但在二维平面中相邻的两个疏水氨基酸之间的能量定义为 -1,其它情况为 0。构形能量 E 的形式化定义如:

$$E=\sum_{i,j=1,i<j-1}^{N}-\sigma_{ij}$$

其中 $\sigma_{ij}=\begin{cases}1, i,j \text{ 均为 } H \text{ 且 } d(i,j)=1;\\ 0, \text{否则。}\end{cases}$

尽管 HP 格点模型是一个简化模型,但对相应折叠问题的求解仍然是 NP 难度的[2]。国内外学者已提出了一些近似求解算法,如遗传算法[3,4],Monte Carlo[3]方法等。本文设计了新的邻域结构,给出了一种新的局部搜索算法(Local search algorithm, 简称 LS),并对国际文献中公认的一组典型算例进行了实算,最后给出计算结果和比较。

2　局部搜索算法

2.1　编码

对于一给定长度为 N 的氨基酸序列,其任一合法构形可用字母表{−1,0,1}上的长度为 $N-2$ 的字符串来表示(前两球固定)。字符串中第 i 个字母表示第 $i+2$ 个球对于第 i 和第 $i+1$ 个球的相对位置坐标,其中 -1 表示向左,0 表示向前,1 表示向右。如图 1,氨基酸序列 HPHPPHHPHPPHPH (N=14)的合法构形可用长度为 12 的字符串表示,即(−1,0,−1,−1,1,1,−1,0,−1,−1,1,−1)。

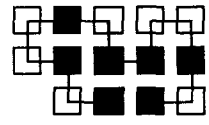

图 1　长度为 14 的一个氨基酸序列对应的一个合法构形

2.2　邻域结构

*)本工作为国家 973 计划(批准号:G1998030600)资助项目。吕志鹏　博士研究生,研究方向为人工智能,计算智能,求解 NP 问题的高效算法研究。黄文奇　教授,博士生导师,研究方向为求解 NP 问题的高效算法研究。

本文算法将基于局部搜索策略。因此，只要给出邻域结构的构造方法并描述搜索策略即可。

对任意一个给定的长度为 N 的蛋白质氨基酸序列，将它的某一合法构形的相对坐标编码 $\pi=(\sigma_1,\sigma_2,\cdots,\sigma_{N-2})$ 视为 $N-2$ 维空间中的一个点，其中 $\sigma_i\in\{-1,0,1\}(i=1,2,\cdots,N-2)$。考虑 π 的邻域，将长度为 k 的窗口覆盖于 π 的某段，并穷举被此窗口覆盖各球的所有可能的相对方向，总共有 3^k-1 种可能。同时将没有被该窗口覆盖的其它各球的相对方向保持不变。将由此产生的新方向序列集合中对应的合法构形 π' 看作 π 的邻点。新方向序列中不合法的构形将被抛弃。

例如一个 $N=7$ 的氨基酸序列，其一个合法构形表示为 $\pi=(0,-1,-1,1,-1)$，假设 $k=2$，且覆盖窗口的最左端位置为第 5 个球。则由 π 经过局部枚举产生的新方向序列为：$\pi_1=(0,-1,-1,-1,-1)$，$\pi_2=(0,-1,-1,0,-1)$，$\pi_3=(0,-1,0,-1,-1)$，$\pi_4=(0,-1,0,0,-1)$，$\pi_5=(0,-1,0,1,-1)$，$\pi_6=(0,-1,1,-1,-1)$，$\pi_7=(0,-1,1,0,-1)$和 $\pi_8=(0,-1,1,1,-1)$。其中 π_1 和 π_2 为非法构形。因此只有 $\pi_3,\pi_4,\pi_5,\pi_6,\pi_7$ 和 π_8 为 π 的邻点。

窗口的位置可通过滑动的方式进行选取，即将窗口置于方向序列最左端，找出此时对应原构形的所有邻点，再将窗口向右滑动一个单位并找出其对应所有邻点，重复这种操作直到窗口不能再向右滑动为止。即长度为 k 的窗口从最左端出发可向右滑动 $(N-2-k)$ 次。通过这种方式得到的所有的合法构形称为原构形的邻点。

2.3 算法描述

有了上面的编码方法及邻域结构的构造方法，算法可描述如下：

步骤 1：随机产生合法初始解 π，计算其能量 $E(\pi)$。设定最大跳坑数 T_{max}，令跳坑计数器 $T=0$；

步骤 2：领域搜索。构造 π 的领域结构 $N(\pi)\{\pi_1,\pi_2,\cdots,\pi_m\}$，计算并找出邻域中能量最低的构形，记作 $E(\pi')=\min\{E(\pi_i),i=1,2,\cdots,m\}$。
若 $E(\pi')<E(\pi)$，
　　则 $\pi=\pi'$，转到步骤 2，
否则，转到步骤 3；

步骤 3：跳坑策略。采取随机的方式进行跳坑。即在方向序列中随机选取一个位置，将对应球的相对方向随机地改变为其它方向，如果改变后构形仍合法则继续，如果不合法则不进行这种改变。重复这种操作直到已进行了 C_0 次成功改变为止。$T=T+1$，如果 $T=T_{max}$，则停机，否则转到步骤 2。

3 结果和比较

利用本文算法在 CY 333MHz CPU 计算机上用 C 语言编制程序，对当前国际文献中公认的一组典型算例进行了计算。对文中所有算例，算法中各参数设定如下：窗口宽度 $k=5$，$T_{max}=1000$。

对于这组算例中除链长为 64 的算例外，其余均在较短的计算时间内找到了最低能量构形。表 1 列出了算例以及与文[3]中的遗传算法和 Monte Carlo 算法、文[4]中的并行遗传算法的结果比较。图 2 至图 7 为本文算法找到的典型的最低能量构形，图中黑色矩形表示为 H，白色矩形表示 P。对链长为 50 的算例，本文还找到了文献中所没有的全新的最低能量构形，如图 6 所示。

表 1　算例及结果比较

No	N	HP 链①	E_{min}①	LS②	GA③	MC④	PGA⑤
1	20	HPHPPHHPHPPHPHHPPHPH	−9	−9	−9	−9	−9
2	24	HHPPHPPHPPHPPHPPHPPHPPHH	−9	−9	−9	−9	−9
3	25	PPHPPHHPPPPHHPPPPHHPPPPHH	−8	−8	−8	−8	−8
4	36	PPPHHPPHHPPPPPHHHHHHHPPHHPPPPHHPPHPP	−14	−14	−14	−13	>−14
5	50	HHPHPHPHPHHHHPHPPPHPPPHPPPPHP-PPHPPPHPHHHHPHPHPHPHH	−21	−21	−21	−21	–
6	64	HHHHHHHHHHHHPHPHPPHHPPHHPPHP-PHHPPHHPPHPPHHPPHHPPHP HPHHHHHHHHHHHH	−42	−38	−37	−35	–

HP 链①：从文[2]中获得；LS②：本文局部搜索算法找到的最低能量；GA③：文[2]中遗传算法找到的最低能量；MC④：文[2]中模拟退火算法找到的最低能量；PGA⑤：文[4]中并行遗传算法找到的最低能量，–表示未报道。

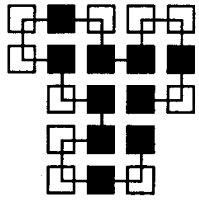

图 2　N=20，$E_{min}=-9$ 的一个最低能量构形

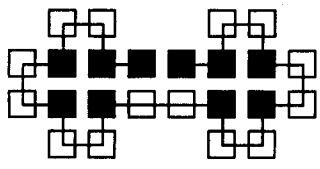

图 3　N=24，$E_{min}=-9$ 的一个最低能量构形

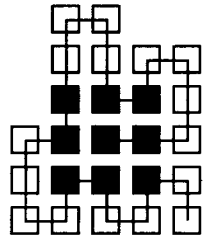

图 4　N=25，$E_{min}=-8$ 的一个最低能量构形

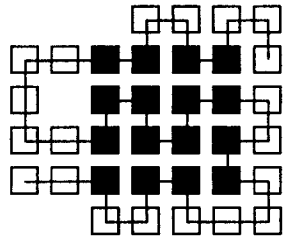

图 5　N=36，$E_{min}=-14$ 的一个最低能量构形

（下转第 151 页）

则

(1)$\underline{B}(d)=\underline{A}(d)\Leftrightarrow M_B^{(1)}(x)=M_A^{(1)}(x)\ \forall x\in U$

(2)$\underline{B}(d)=\underline{A}(d)\Leftrightarrow M_B^{(2)}(x)=M_A^{(2)}(x)\ \forall x\in U$

证:(1)因 $x\in M_B^{(1)}(x)\Leftrightarrow x\in\overline{R}_B^*(D_j)$,$x\in M_A^{(1)}(x)\Leftrightarrow x\in\overline{R}_A^*(D_j)$而$\overline{B}(d)=\overline{A}(d)$,即$\overline{R}_B^*(D_j)=\overline{R}_A^*(D_j)$即证。

(2)类似于(1)可证。

定理 2 设$(U,A\cup\{d\})$是目标信息系统,$B\subseteq A$,

则

(1)$\overline{B}(d)=\overline{A}(d)\Leftrightarrow$对$\forall x,y\in U$,当$M_A^{(1)}(x)\neq M_A^{(1)}(y)$时,$[x]_B\cap[y]_B=\phi$

(2)$\underline{B}(d)=\underline{A}(d)\Leftrightarrow$对$\forall x,y\in U$,当$M_A^{(2)}(x)\neq M_A^{(2)}(y)$时,$[x]_B\cap[y]_B=\phi$

证:必要性:因$\overline{B}(d)=\overline{A}(d)$,由定理 1,得$M_B^{(1)}(x)=M_A^{(1)}(x)$,$M_B^{(1)}(y)=M_A^{(1)}(y)$

若对$\forall x,y\in U$,$[x]_B\cap[y]_B\neq\phi$,一定有$[x]_B=[y]_B$因此$M_B^{(1)}(x)=M_B^{(1)}(y)$,所以有$M_A^{(1)}(x)=M_A^{(1)}(y)$。

充分性:记$J([x]_B)=\{[y]_A:[y]_A\subseteq[x]_B\}$

由于$B\subseteq A$,因此$J([x]_B)$构成了$[x_B]$的一个分划。

因对$\forall x\in U$,当$[y]_A\subseteq[x]_B$时,有$[x]_B\cap[y]_B\neq\phi$

由已知有$M_A(x)=M_A(y)$,对$\forall j\leqslant r$,如果$x\in\overline{R}_B^*(D_j)$,则$[x]_B\in\overline{R}_B^*(D_j)$

由于$[x]_B=\cup\{[y]_A:[y]_A\in J([x]_B)\}$因此对任意$[y']_A\in J([x]_B)$,有$[y']_A\subseteq\overline{R}_B^*(D_j)$,从而$D_j\in M_A(y')$,$D_j\in M_A(x)$,所以$x\in\overline{R}_A^*(D_j)$,即$\overline{R}_B^*(D_j)\subseteq\overline{R}_A^*(D_j)$,反过来,若$x\in\overline{R}_A^*(D_j)$,则$D_j\in M_A(x)$

当$[y]_A\in J([x]_B)$时,$[y]_B\cap[x]_B\neq\phi$,因此$M_A(y)=M_A(y)$,从而$D_j\in M_A(y)$,即$P(D_j|[y]_A\geqslant P(D_j))$,因此,$P(D_j|[x]_B)=(\sum\{[y]_A\cap D_j:[y]_A\in J([x]_B)\}/|[x]_B|=\sum\{D(D_j/[y]_A)\frac{|[y]_A|}{|[x]_B|}:[y]_A\in J([x]_B)\}\geqslant P(D_j)\sum\{\frac{|[y]_A|}{|[x]_B|}:[y]_A\in J([x]_B)\}=P(D_j)$,所以$x\in\overline{R}_B^*(D_j)$,即$\overline{R}_A^*(D_j)\subseteq\overline{R}_B^*(D_j)$,因此$\overline{R}_B^*(D_j)=\overline{R}_A^*(D_j)$,对$\forall j\leqslant r$,所以$\overline{B}(d)=\overline{A}(d)$。

定义 3 设$(U,A\cup\{d\})$是目标信息系统,定义

$$D^{(i)}(x,y)=\begin{cases}\{a_l\in A:a_l(x)\neq a_l(y)\} & M_A^{(i)}(x)\neq M_A^{(i)}(y)\\ A & M_A^{(i)}(x)=M_A^{(i)}(y)\end{cases}$$

$(i=1,2)$

则$D^{(i)}(x,y)(i=1,2)$被分别称为对象x,y关于$M_A^{(i)}(x)(i=1,2)$的可辨识属性集,$D^{(i)}=\{D^{(i)}(x_i,y_j)i,j\leqslant r\}(i=1,2)$被分别称为目标信息系统关于$M_A^{(i)}(x)(i=1,2)$可辨识矩阵。

定理 3 设$(U,A\cup\{d\})$是目标信息系统,$B\subseteq A$,则

(1)$\overline{B}(d)=\overline{A}(d)\Leftrightarrow B\cap D^{(1)}(x,y)\neq\phi\quad\forall x,y\in U$

(2)$\underline{B}(d)=\underline{A}(d)\Leftrightarrow B\cap D^{(2)}(x,y)\neq\phi\quad\forall x,y\in U$

证:(1)必要性

若$\overline{B}(d)=\overline{A}(d)$,对$\forall x,y\in U$,如果$M_A^{(1)}(x)=M_A^{(1)}(y)$,则$D^{(1)}(x,y)=A$,如果$M_A^{(1)}(x)\neq M_A^{(1)}(y)$,一定有$B\cap D^{(1)}(x,y)\neq\phi$。

由定理 2,$[x]_B\cap[y]_B=\Phi$,于是存在$a_k\in B$,使$a_k(x)\neq a_k(y)$,所以$a_k\in D^{(1)}(x,y)$,因此$B\cap D^{(1)}(x,y)\neq\phi$。

充分性:

若存在$x,y\in U$,使得$B\cap D^{(1)}(x,y)\neq\phi$,则$M_A^{(1)}(x)=M_A^{(1)}(y)$,对$\forall a_k\in B$,必有$a_k\notin D^{(1)}(x,y)$,所以$a_k(x)\neq a_k(y)$,因此$[x]_B=[y]_B$,由定理 2,$\overline{B}(d)\neq\overline{A}(d)$。(2)类似于(1)可证得。

结论 属性约简是粗糙理论的主要内容之一,即在保持信息不丢失的情况下约去冗余属性。本文利用上、下分布及可辨识属性矩阵,讨论了贝叶斯粗糙集模型的知识约简,这一方法比原来的约简方法在实际中具有可操作性,在理论及应用上都是有意义的。

参 考 文 献

1 Pawlak Z. Rough set [J]. International Journal of Computer and Information Science ,1982,11:341～356

2 Ziarko W. variable precision rough sets model [J]. Journal of computer and systems sciences, 1993,46(1):39～59

3 Slezak D, ziarko W. Attribute reduction in the Bayesian Version of Variable precision rough set model [J]. Electronic Notes in Theoretical Computer Science,2003(4):1～11

4 张文修,梁怡,吴伟志,等. 信息系统与知识发现[M]. 北京:科学出版社,2003

5 张文修,米据生,吴伟志. 不协调目标信息系统的知识约简[J]. 计算机学报,2003,26(1):12～18

(上接第 149 页)

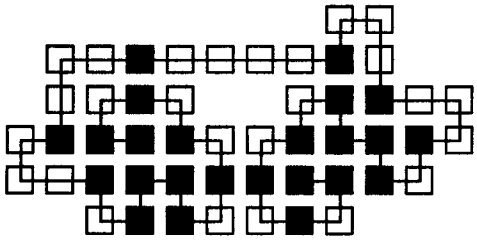

图 6 N=50, $E_{min}=-21$ 的一个典型最低能量构形

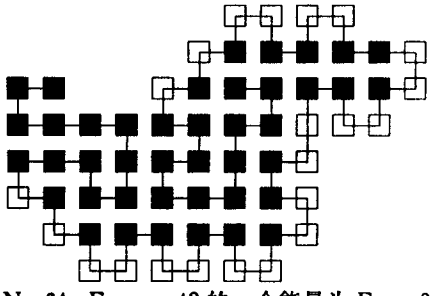

图 7 N=64, $E_{min}=-42$ 的一个能量为 E=－38 构形

通过上面的结果比较不难发现,本文算法对大多数算例的计算都达到了最优解,其计算效率要高于遗传算法和 Monte Carlo 方法。局部搜索算法具有局部寻优的特点,而跳坑策略则增加了构形的多样性,使算法在限人僵局的情况下跳出"陷阱",同时保持原有构形的信息以引导算法走向有希望的区域。二者结合有效地提高了算法的计算效率。

参 考 文 献

1 Dill K A, Bromberg S, Yue K, et al. Principles of protein folding: A perspective from simple exact models. Protein Sci. , 1995, 4:561～602

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 Unger R, Moult J. Genetic algorithms for protein folding simulations. J. Mol. Biol. , 1993, 231: 75～81

4 王敞,陈增强,袁著祉. 基于并行遗传算法的蛋白质空间结构预测. 计算机科学,2003,30(7): 147～150