维普资讯 http://www.cqvip.com

基于并行遗传算法的蛋白质空间结构预测*)

王 敞 陈增强 袁著祉
(南开大学信息技术科学学院 天津300071)

Protein Conformation Prediction Based on Parallel Genetic Algorithm

WANG Chang CHEN Zeng-Qiang YUAN Zhu-Zhi
(College of Information Technology Science, Nankai University, Tianjing 300071)

Abstract For the prediction of protein structure, we propose the solution based on parallel genetic algorithm. This solution has modified the existing algorithm by omitting the process of simulated annealing, while compensating this modification by parallel programming to maintain the results quality. Experiments have demonstrated that the new solution has higher processing speed and same results quality. This solution also underlines the necessity of introducing the idea of parallel programming into the study of bioinformatics.

Keywords Protein conformation, Parallel genetic algorithm, Bioinformatics, Data mining, PVM

1 前言

蛋白质是生命的基础。从细胞分裂到细胞运动,从代谢到免疫,已知的生物功能没有一个能够离开蛋白质而实现。一个典型的细胞可能含有的蛋白质种类就有大约10万种,而且每种蛋白质的作用都不尽相同。更进一步,基因是生物细胞中的遗传物质,它也必须以蛋白质形式表达出来,才能显示出生物的各种遗传性状。因此,对于蛋白质的研究在生命科学中是最重要的课题之一。

蛋白质的功能很大程度上决定于它的空间结构。因此,研究蛋白质的空间结构就成了目前研究工作的热点。实验室方法(如核磁共振法等)确定这种空间结构花费高,速度慢,远远满足不了人们的要求。利用蛋白质的已知信息来对它的空间结构进行预测(也称蛋白质折叠、蛋白质构象)也就因此逐渐成为了生物信息学研究中一个最富挑战性的方向。

我们的研究主要是基于 Ron Unger, John Moult 等人的工作[1],他们在1993年使用了一种简化的蛋白质模型,而后利用遗传算法在这种简化的模型上进行了构象研究。后来,Arnold L. Patton 等人[2]又对原有方法进行了改进,通过在类似的模型上进行实验,获得了更好的效果。我们的工作也是基于文[1]使用的简化蛋白质模型的,我们同样使用了遗传算法的思想。但是,我们使用的是另外一套遗传算法的实现策略。我们的方法省略了原算法中每步遗传操作中的模拟退火过程,而用并行化的程序来弥补这种省略带来的解质量的下降,大大降低了程序运行的时间。

2 蛋白质构象研究

构象(Conformation)是指组成一个分子的各个原子和基团间的相对位置。因为分子中单键的旋转,使分子中的一些原子和基团相对位置有所不同,导致一个分子可以有多种构象体。所谓蛋白质的构象就是蛋白质二级结构中规则的肽段和其他无规则的肽段一起构成的完整的立体结构(蛋白质三级结构)。目前,对于蛋白质而言,立体结构、三级结构、空间结构和构象都可以认为是等同的概念。用计算机进行构象研究实际上就是对蛋白质整个构象空间进行搜索,然后找出能量最低的构象作为最后的预测构象,这种预测方式在科学上是有一定根据的[5]。

我们使用的蛋白质简化模型描述如下:①蛋白质一级序列,即氨基酸序列中只包含疏水型和亲水型2种氨基酸。②序列被认为是平面方格上面的二维结构。③序列中的每个点可以左转90度、右转90度和继续向前。④该模型的能量函数也相对简单,即初始能量为0,如果在原序列中不相连的两个疏水型氨基酸在平面方格上直接相连了,那么能量值就减1。

使用这种能量函数作为判断尺度,那些以疏水型氨基酸为核心的结构就具有相对较低的能量,因为那些疏水型氨基酸有更多的可能相互接触。这种简化的蛋白质结构模型具有真实蛋白质所能表现出来的特点,是很好的研究用模型,而且它大大简化了对于能量的计算。

3 并行遗传算法

并行遗传算法是在遗传算法的实施中加入自然生物种群中的空间结构因素,其目的是可以在大规模并行机上运行。即使是序列式实施并行遗传算法,也可克服经典遗传算法性能上的不足,从而提高算法的效率。并行遗传算法大致分为两种:一种是孤岛模型(Island Modeling),称为粗粒度并行遗传算法;另一种是细胞模型(Cellular Model),称为细粒度并行遗传算法。孤岛模型的实施方法是在几个独立按着经典遗传算法方式运行的种群中加入迁徙因素;而细胞模型的实施则是在一个单一种群中,通过对经典整体性选择与交叉机制的局部化来达到并行化目的。孤岛模型考虑的是种群的空间结构,而细胞模型考虑的是一种群中个体的空间结构。

4 PVM 概述

PVM (Parallel Virtual Machine)是一种基于局域网络的异构编程环境,由 Emory 大学与 Oak Ridge 国家实验室联合开发。PVM 具有通用性强及系统规模小的特点,即适合

*)本文得到国家自然科学基金(60174021)资助。王 敞 硕士研究生,主要研究领域为数据挖掘、生物信息学、人工智能。陈增强 教授,博士生导师,主要研究领域为智能控制。袁著祉 教授,博士生导师,主要研究领域为智能控制。

维普资讯 http://www.cqvip.com

TCP/IP 环境，又适合 MPP 大型并行系统。目前，PVM 已成为最成功的并行开发环境，所有的并行机公司都宣布支持 PVM。一些标准软件已经或正在移植到 PVM 平台。

PVM 能把一组由使用 UNIX/Linux 系统的计算机组成的异构网络环境虚拟成单个的并行计算机，可以让一群计算机去合力解决一个大型计算问题。一般用虚拟机这个术语指代这个分布的并行计算机。PVM 提供在该虚拟机上创建任务以及支持任务间通信、同步的手段。

PVM 系统包括两个部分：监控程序部分叫 PVMD，必须运行在构成虚拟机的所有宿主机上。所有合法注册的用户都可把 PVMD 装在他的机器上。当用户要运行 PVM 应用程序时，必须先在一个机器上执行 PVMD，这个 PVMD 将根据配置文件自动启动虚拟机所有成员上的 PVMD，然后在任一台宿主机上，用户可以在 UNIX 提示符下执行他的应用程序。PVM 的第二部分就是 PVM 的接口例程库(PVMLIB)，用户可以利用这些例程去完成消息传递、创建子进程、任务同步等功能。

PVM 提供两种编程模式：一种是主从式，由主程序激活一组从程序运行，并调度这些从程序；另一种是 SPMD 模式，在这种模式下，所有进程地位平等，没有所谓的主进程。PVM 应用程序可以用 C 或 Fortran 编写，并使用 PVM 提供的消息传递机制实现并行计算。通过发送和接受消息，应用程序的多个任务可以相互协作，并发地解决一个问题。PVM 支持在应用、机器、网络三个层次上的异构。

5 基于并行遗传算法的蛋白质空间结构预测

5.1 编码

对目标蛋白质，我们对其一级结构直接编码。每个氨基酸都用一个两元组表征，元组1表示其属性：疏水型或是亲水型；元组2表示与之相连的下一个氨基酸的移动方向(直接前方、左方、右方)。

5.2 初始群体生成

对于待处理蛋白质，因为其一级结构已知，所以其氨基酸序列所有元素的属性(亲水或疏水)都是确定的。多次实验表明，种群越大，含有的个体种类越丰富，程序最后获得解的效果就越好。为保证初始群体解的多样性，我们对于表征下一元素移动方向的信息采取了完全随机生成的方法。这里需要注意一点，由于将氨基酸序列在平面空间展开，可能导致一些非法个体(序列在平面上发生了重叠)的产生。在程序初始化的时候，要判断产生的个体是否合法，若不合法则需要重新产生。

5.3 适应度计算

假设序列长度为 L，我们需要构造一个(2L－1)*(2L－1)的矩阵。我们将初始点置于矩阵的中心，然后通过顺次扫描序列元素，将整个序列的二维平面图"画"在矩阵上。设初始能量值为0。我们首先判断序列是否达到矩阵的四个边。如果达到，那么该序列肯定排列成一条直线，能量值仍为0；如果没达到四边，那么对于(2行，2列)，(2行，2L－2列)，(2L－2行，2L－2列)，(2L－2行，2列)四点顺次连线构成的矩阵的元素从左上到右下，开始逐个扫描，判断每个点是否构成了使能量减小的氨基酸对。并累加这种减少值，算出总能量。

5.4 选择、交叉与变异

为了让遗传算法能够得到全局最优解，我们采用了完全随机的选择方式。交叉采用了一点交叉。对于本问题，变异的作用尤为明显，恰当的变异率可以部分代替模拟退火的作用，因此它是算法能否成功的关键。我们的变异对象是染色体中表示方向的信息。这里我们的交叉和变异概率都是自适应变化的，而且变异概率相对较大。

5.5 遗传算法的并行化实现

本程序采取了粗粒度并行化的遗传算法实现方式，也就是平时所说的孤岛模型。每个孤岛内部都有一个遗传算法在进行计算。三个孤岛两两相连，构成一个三角形。每隔若干代，每个孤岛都将自己最优的个体以消息传递的方式传给其右侧的节点，实现了最优个体在种群间的迁徙。每个孤岛上负责蛋白质结构预测部分的程序都是完全相同的，初始数据也由各孤岛内部独立产生。

从 PVM 的实现角度，我们选用了3台普通 PC 组建了并行开发环境，每台都运行 PVM 程序。我们采用了主从方式的编程模式。其中一台作为主机，由它启动整个运算，并输出结果，并要负责1/3的运算量。另外两台从机，各自负责运算量的1/3，并负责将其最后运算结果返回给主机输出。

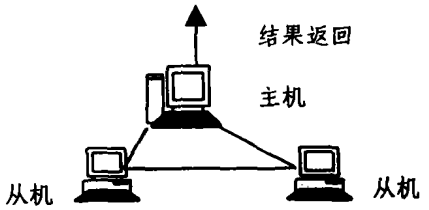

图1 本研究中 PVM 的组织结构

6 仿真实验

我们在 Linux 平台下的 PVM3.4环境上用 C 语言实现了我们的算法。实验数据选用由20，24，25，36，48个氨基酸组成的5种蛋白质。实验表明，本程序在前三种情况下可以获得最优解。随着问题空间的增大，则只能获得近似最优结果。与文[1]的方法相比较而言，在每种情况下，程序的运行速度都要比文[1]方法快得多。对于问题空间不太大的情况，新方法与文[1]中的方法都可以获得最优结果(序列长为20：最低能量为－9；序列长为24：最低能量为－9；序列长为25：最低能量为－8)。对于问题空间较大的情况，新方法获得的结果质量则不

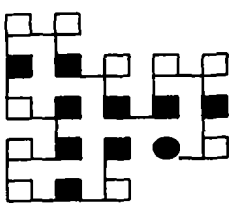

图2

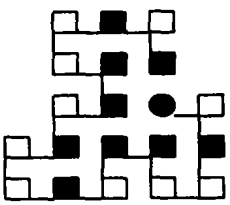

图3

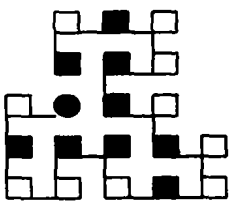

图4

注：●代表序列开始 ■代表疏水型氨基酸 □代表亲水型氨基酸

(下转第171页)

维普资讯 http://www.cqvip.com

定理1 问题(1)式中第 j 个源信号 $s_j(t)$ 理论上能被提取的必要条件是矩阵 A 的第 $2j-1$ 列不能被矩阵 A 的其它列线性表示。

证明：在假设(1)、(2)的条件下，由于卷积混叠盲分离问题(1)式能够转化为线性混叠盲分离问题：

$$x_k(t)=\sum_{i=0}^{P_{K1}}a_{k1}(i)s_1(t)-\tau\sum_{i=0}^{P_{K1}}ia_{k1}(i)\dot{s}_1(t)+\cdots+\sum_{i=0}^{P_{Kn}}a_{kn}(i)s_n(t)-\tau\sum_{i=0}^{P_{Kn}}ia_{kn}(i)\dot{s}_n(t),k=1,2,\cdots,m \quad (4)$$

即 $X(t)=AS(t)$，其中 $X(t)=(x_1(t),x_2(t),\cdots,x_m(t))^T$，$A$ 为由(3)式定义的混叠矩阵，$S=(s_1(t),\dot{s}_1(t),\cdots,s_n(t),\dot{s}_n(t))^T$。

当信号 $s_1(t),\dot{s}_1(t),s_2(t),\dot{s}_2(t),\cdots,s_n(t),\dot{s}_n(t)$ 相互独立时，必有该组信号不相关；若第 j 个源信号理论上能被分离，则存在某个向量 $W_j=(w_{j1},w_{j2},\cdots,w_{jm})^T$ 和 $\lambda_j(\lambda_j\neq0)$ 使得

$$W_j^TX(t)=W_j^TAS(t)=\lambda_js_j(t)。$$

设 $W_j^TA=(b_{11},b_{12},\cdots,b_{(j-1)2},b_{j1},\cdots,b_{n2})$，显然

$$W_j^TAS(t)=b_{11}s_1(t)+b_{12}\dot{s}_1(t)+\cdots+b_{(j-1)2}\dot{s}_{j-1}(t)+b_{j1}s_j(t)+b_{j2}\dot{s}_j(t)+\cdots+b_{2n}s_{2n}(t)=\lambda_js_j(t)。$$

由于信号 $s_1(t),\dot{s}_1(t),s_2(t),\dot{s}_2(t),\cdots,s_n(t),\dot{s}_n(t)$ 不相关，因此可推出

$$b_{11}=0,b_{12}=0,\cdots,b_{(j-1)2}=0,b_{j1}=\lambda_j,b_{j2}=0,\cdots,b_{n2}=0。$$

即

$$W_j^TA=(b_{11},b_{12},\cdots,b_{(j-1)2},b_{j1},b_{j2},\cdots,b_{n2})=(0,0,\cdots,0,\lambda_j,0,\cdots,0)。$$

因为 $W_j^TA=(0,0,\cdots,0,\lambda_j,0,\cdots,0)$ 等价于方程组

$$\begin{cases}w_{j1}\sum_{i=0}^{P_{11}}a_{11}(i)+w_{j2}\sum_{i=0}^{P_{21}}a_{21}(i)+\cdots+w_{jm}\sum_{i=0}^{P_{m1}}a_{m1}(i)=0\\ \cdots\cdots\\ w_{j1}\sum_{i=0}^{P_{1j}}a_{1j}(i)+w_{j2}\sum_{i=0}^{P_{2j}}a_{2j}(i)+\cdots+w_{jm}\sum_{i=0}^{P_{mj}}a_{mj}(i)=\lambda_j\\ \cdots\cdots\\ -\tau w_{j1}\sum_{i=0}^{P_{1n}}ia_{1n}(i)-\tau w_{j2}\sum_{i=0}^{P_{2n}}ia_{2n}(i)-\cdots-\tau w_{jm}\sum_{i=0}^{P_{mn}}ia_{mn}(i)=0\end{cases}$$

假设 A 的第 j 列能由 A 的其它列线性表示，即存在一组数 $c_1,c_2,\cdots,c_{j-1},c_{j+1},\cdots,c_n$ 使得

$$(\sum_{i=0}^{P_{1j}}a_{1j}(i),\sum_{i=0}^{P_{2j}}a_{2j}(i),\cdots,\sum_{i=0}^{P_{mj}}a_{mj}(i))=c_1(\sum_{i=0}^{P_{11}}a_{11}(i),\sum_{i=0}^{P_{21}}a_{21}(i),\cdots,\sum_{i=0}^{P_{m1}}a_{m1}(i))+\cdots+c_{j-1}(\sum_{i=0}^{P_{1j-1}}a_{1j-1}(i),\sum_{i=0}^{P_{2j-1}}a_{2j-1}(i),\cdots,\sum_{i=0}^{P_{mj-1}}a_{mj-1}(i))+c_{j+1}(\sum_{i=0}^{P_{1j+1}}a_{1j+1}(i),\sum_{i=0}^{P_{2j+1}}a_{2j+1}(i),\cdots,\sum_{i=0}^{P_{mj+1}}a_{mj+1}(i))+\cdots+c_n(\sum_{i=0}^{P_{1n}}a_{1n}(i),\sum_{i=0}^{P_{2n}}a_{2n}(i),\cdots,\sum_{i=0}^{P_{mn}}a_{mn}(i))。$$

所以，

$$\lambda_j=w_{j1}(\sum_{i=0}^{P_{1j}}a_{1j}(i)+w_{j2}\sum_{i=0}^{P_{2j}}a_{2j}(i)+\cdots+w_{jm}(\sum_{i=0}^{P_{mj}}a_{mj}(i)=(w_{j1}\sum_{i=0}^{P_{11}}a_{11}(i)+w_{j2}\sum_{i=0}^{P_{21}}a_{21}(i)+\cdots+w_{jm}\sum_{i=0}^{P_{m1}}a_{m1}(i))+\cdots+(w_{j1}(\sum_{i=0}^{P_{1n}}a_{1n}(i)+w_{j2}\sum_{i=0}^{P_{2n}}a_{2n}(i)+\cdots+w_{jm}\sum_{i=0}^{P_{mn}}a_{mn}(i)=0$$

这与 $\lambda_j\neq0$ 矛盾。

由于矩阵 A 的行数 m 越大，矩阵 A 的某列不能被矩阵 A 的其它列线性表示的可能性越大，因此由上述定理可以推出，传感器的个数越多，越有利于分离出源信号。

结论 在卷积混叠盲源分离中，当时间是微小延迟的情况下，本文研究了传感器个数少于源信号个数的病态卷积混叠的可分性；给出了病态卷积混叠盲源分离的必要条件，该条件同样适合非病态微延迟盲反卷积问题。

参 考 文 献

1 Jutten C, Herault J. Blind separation of sources, Part I: An adaptive algorithm based on neuromimetic architecture [J]. Signal Processing, 1991, 24(1): 1～10

2 Yellin D, Wensten E. Criteria for multichannel signal separation [J]. IEEE Trans Signal Process, 1994, 42(8): 2158～2168

3 Thi H N, Jutten C. Blind source separation for convolutive mixtures [J]. Signal Processing, 1995, 45(2): 209～229

4 Lee T W, Bell A J, Lambert R H. Blind separation of delayed and convolved sources [A]. Advances In Neural Information Processing Systems [C]. Cambridge MA: MIT Press, 1997. 758～764

5 谭丽丽，韦岗. 多输入多输出盲解卷问题的最大熵解法[J]. 电子学报，2000，28(1)：114～116

6 刘琚，何振亚. 盲源分离和盲反卷积[J]. 电子学报，2002，30(4)：570～576

7 Jean B, Gill C. A compact sensor array for blind separation of sources [J]. IEEE Trans Circuits and systems-I: Fundamental theory and applications, 2002, 49(5): 565～574

(上接第148页)

如文[1]的方法。因此，新方法适用于问题空间不是很大、但对速度要求较高的情况。

对于由20个氨基酸组成的蛋白质的空间结构，研究成果最多。研究表明[1]，这个问题的解空间规模为8400万，其中最优解只有4个，能级为－9(由此可见，该类问题的复杂性)。前人的工作给出了两种最优构象(图2和图3)。我们的计算得到了另外一种最优构象(图4)。

结论 本文针对蛋白质的空间结构预测问题，在前人工作的基础上，提出了基于并行遗传算法的实现方案。实践证明，并行方案有很好的适用性、很快的运行速度，尤其适用于生物信息学领域中经常遇到的大计算量问题。进一步的研究可以侧重于以下两个部分：①使用更能体现蛋白质真实特性的简化模型。②分析并行程序运行中涉及到的负载平衡问题，切实提高并行机运行效率。

参 考 文 献

1 Unger R, Moult J. Genetic Algorithms for Protein Folding Simulations, J Mol Biol, 1993

2 Patton A L, Punch W F III, Goodman E D. A standard GA Approach to Native Protein Conformation Prediction, Proc Incl Conf on Genetic Algorithms

3 Holland J H. Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, Michigan, 1975

4 PVM 3.4 Manual

5 Levinthal C. Are there pathways for protein folding? J. Chem. Phys, 1968

6 阎隆飞，孙之荣. 蛋白质分子结构. 清华大学出版社，1999