研究论文

甲型 H1N1 流感病毒三维空间结构预测

靳佩轩，　高　洁*

（江南大学 理学院，江苏 无锡 214122）

摘要：采用从头计算思想，用改良的遗传算法对甲型 H1N1 流感病毒的蛋白质空间结构进行预测研究。基于蛋白质空间结构的 HP 模型，构建甲型 H1N1 流感病毒蛋白质空间结构的 3DHP 模型，并利用改良的遗传算法找到最小自由能结构，从而预测得到甲型 H1N1 流感病毒蛋白质三维空间结构。利用蛋白质空间结构数据建立距离矩阵，通过相关性分析和显著性检验，表明预测结构与已知结构存在高度的一致性。该模型提供了一种快速预测甲型 H1N1 流感病毒结构的方法。

关键词：3DHP 模型；改良的遗传算法；距离矩阵；相关性分析

中图分类号：Q 71；O 29　**文献标志码**：A　**文章编号**：1673—1689(2014)05—0492—06

Three-Dimensional Structure Prediction of Influenza A (H1N1) Virus

JIN Peixuan，　*GAO Jie**

（School of Science，Jiangnan University，Wuxi 214122，China）

Abstract：Basing on the HP model of protein spatial structure，this paper searches the protein three-dimensional structure of the influenza A （H1N1） virus according to the method of ab initio. This paper establishes the 3DHP model of the Influenza A （H1N1） virus protein spatial structure and uses the improved genetic algorithm to find the structure with minimum free energy，aiming at predicting the influenza A （H1N1） virus protein three-dimensional structure. Before correlation analysis and significance test，protein spatial structure data establishes distance matrices. The significance is to illustrate the highly consistent relationship between prediction structure and known structure. The model provides a quick method to predict the spatial structure of the influenza a （H1N1） virus.

Keywords：3DHP model，optimization of genetic algorithm，distance matrix，correlation analysis

流感病毒具有多样的变异性，历史上每一次流感大流行多是由流感病毒新亚型和以往出现过的亚型的再次出现，人类绝大多数对其缺乏相应的免疫力，并且现有的流感疫苗起不到有效作用，从而造成流感病毒在人群中快速广泛的传播，最终出现流感大流行[1-2]。面对流感病毒表面抗原蛋白结构复

收稿日期：2013-08-04

基金项目：国家自然科学基金项目(11271163)；中央高校基本科研业务费专项资金项目。

* **通信作者**：高　洁(1972—)，女，江苏无锡人，工学博士，副教授，硕士研究生导师，主要从事生物信息学方面的研究。

E-mail：ezhun6669@sina.com

杂多变的性状和变异的突发性,研究流感病毒蛋白的三维空间结构显得特别重要。

近年来,关于蛋白质三维空间结构的研究已有不少。1989 年,Dill and Lau[3]建立蛋白质空间结构的 HP 模型,忽视侧链的影响,将氨基酸分为亲水性(P)和疏水性(H)两类,从而使氨基酸序列抽象成一个二进制的序列,来构建一个存在疏水核的二维或三维蛋白质结构模型。2006 年,陈凤飞[4]又衍生出关于 HP 模型的修正模型,即三角化的 HP 格点模型。2004 年,Custodio 等[5]提出用遗传算法去优化三维蛋白质 HP 格点模型,得到结构最优的三维结构。2010 年,Vincent 等[6]用从头计算的思想,利用分子动力学方法深入研究蛋白质三维空间结构折叠的原理,为从头计算建立三维空间模型进行相关因素分析。2011 年,Ivan Dotu 等[7]提到将格点模型和非格点模型用 LNS(Large Neighborhood Search)去找到 HP 模型的自然态。2011 年,Islam[8]对蛋白质结构的三维 HP 格点模型用 MA(memetic algorithm)搜索算法进行优化。2013 年,张玲等[9]运用 ARFIMA 模型对甲型 H1N1 流感病毒的 HA 蛋白质的序列进行预测分析。2009 年,Manabu Igarashi 等[10]运用已经很成熟的比较建模的方法对 2009 年甲型 H1N1 流感病毒的结构进行了模建分析。

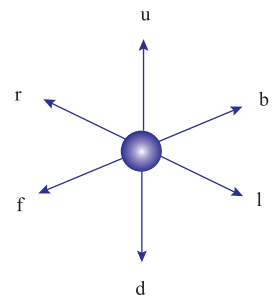
作者基于蛋白质空间结构的 HP 模型,将蛋白质结构中氨基酸种类分成两类,构建了甲型 H1N1 流感病毒蛋白质空间结构的 3DHP 模型,利用改良的遗传算法找到最小自由能结构,从而预测了甲型 H1N1 流感病毒蛋白质三维空间结构,并在 PDB 数据库中取了 8 条已测得的甲型 H1N1 流感病毒的蛋白质空间结构中心碳原子坐标数据,将甲型流感病毒蛋白的空间坐标转换为距离矩阵量化表示,对预测结构与实际测得的结构进行比较,通过相关性分析和显著性检验,表明预测结构与已知结构存在高度的一致性。

1 材料与方法

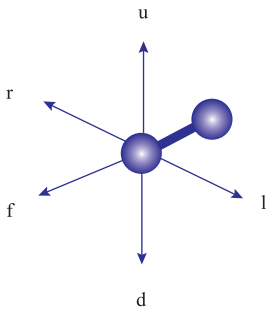
1.1 基于蛋白质空间结构的 3DHP 模型

将氨基酸按亲疏水性分类,即将 20 种氨基酸分两类:疏水性 $H=\{A,I,L,M,F,P,W,V\}$,亲水性 $P=\{N,C,Q,G,S,T,Y,D,E,R,H,K\}$,又令:$H=1$ 和 $P=0$,蛋白质的氨基酸序列即可转换为由 0 和 1 组成的序列。由蛋白质三维空间结构产生的主要驱动力是氨基酸的疏水效应[11],则在蛋白质结构中疏水残基间相互作用在蛋白质的中心形成一个疏水核,亲水残基包围在这个核的外面形成了一个稳定的蛋白质空间结构。

HP 格点模型在三维空间中的折叠简称 3DHP 模型,将氨基酸看作一个节点,定义在模型中各节点间最小距离为单位 1,在折叠过程中各节点位置不能重叠,每个节点的折叠方向在空间中有六个,即在立体方格中每个氨基酸节点可分别进行 90°的向上(u)、下(d)、左(l)、右(r)、前(f)和后(b)六个方向的折叠(见图 1a)。由于每个节点不能重复,确定前一个节点的位置后,下一个节点最多有 5 个折叠方向(见图 1b)。此模型中任意两个节点不重合,并忽略折叠中侧链的影响,虽将蛋白质的空间结构简化,但是整体的蛋白质空间结构骨架符合真实蛋白质结构的基本特征,能很好地模拟真实蛋白的折叠行为,且计算简单,有利于对比不同折叠搜索算法。



(a) 每一个节点有六个折叠方向



(b) 确定前一个节点后,下一个节点有五个折叠方向

图 1 3DHP 结构中氨基酸节点的折叠方向

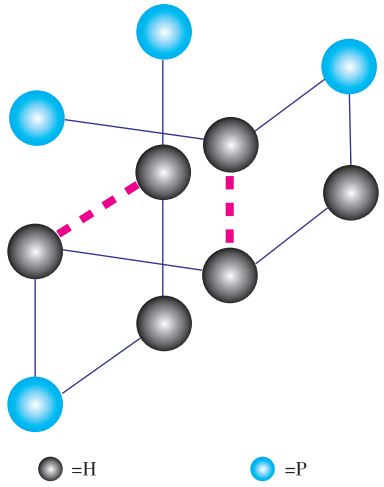
Fig. 1 Folding direction about amino acid nodes in the 3DHP structure

1.2 能量函数

1973 年 Anfinsen 提出了蛋白质的天然构象对应其能量最低的结构这一热力学假说。为搜索到

3DHP 模型构建的能量最小的蛋白质三维空间结构,在此定义其能量函数。

本研究中,采用改良的遗传算法对蛋白质三维空间结构的选择优化的过程中所依据的能量函数函数就是其适应度函数。则能量函数计算如下:在空间折叠结构中,对序列上不相邻而空间上相邻的两个氨基酸节点(见图 2)赋予它们一个相互作用的能量值。由此可以得到的蛋白质空间结构模型的总能量值 E:$E=\sum_{i=1,j=i+1}^{i<N,j<N}\lambda E_{ij}$ 即,其中 N 为氨基酸序列的长度,如果 i 与在空间中相邻而在序列中并不相邻,则 λ 等于 1,否则 λ 等于 0。E_{ij} 表示在蛋白质空间结构中空间相邻的第 i 个氨基酸与第 j 个氨基酸之间的能量值(见图 3)。



空间相邻的两个节点由粉色虚线相连,序列相邻的两个节点由蓝色实线相连

图 2 序列 PHPHHHPHHP 的 3DPH 模型

Fig. 2 Model of PHPHHHPHHP

	0	1
0	0	0
1	0	-1

图 3 空间结构中空间相邻的节点间的能量表

Fig. 3 Energy value table about the space of the adjacent nodes in the structure

由图 3 可知,蛋白质空间结构中空间相邻的三种情况 1-1、0-1、0-0 的能量值分别为 E_{11}=-1,E_{01}=0,E_{00}=0,则具有最小自由能的蛋白质空间结构就是搜索得到空间相邻 1-1 个数最多的空间结构。

1.3 改良的遗传算法

改良的遗传(GA)算法主要引入局部优化策略,将 GA 算法和模拟退火(SA)算法结合,既有效地克服了 SA 算法求最优解耗时大的缺点,又有效地避免了 GA 算法因为其早熟收敛而得到非全局最优解的问题。

算法的基本步骤设计如下:1)由 3DHP 模型随机产生 100 个合法的蛋白质空间结构,计算其个体适应度,即每个空间结构的能量值;2)选择过程,采用随机选择思想在 100 个结构中随机选择两个结构作为算法优化目标,并计算各自能量值;3)单点交叉过程,针对步骤 2 中选出的两个结构,由前面可知这两个结构为同一个序列得到的不同的空间结构,则在这两个结构中分别随机选择其对应氨基酸序列上的同一个位置的氨基酸作为交叉位点,将两个结构中位于这个交叉位点后的结构进行互换得到两个新的空间结构,并读取两个新结构的节点坐标;4)变异过程,在进行交叉过程后形成的两个新的个体会发生结构中节点坐标重叠的现象,对重叠的节点进行变异操作,即依次改变交叉点后出现重叠的节点的折叠方向来保证新的个体为合法的蛋白质空间结构,并计算其各自的自由能;5)对得到的四个结构进行筛选,即从这四个结构中先选出一个自由能最低的结构作为下一个优化的目标,同时剔除自由能最高的一个结构,对剩下的两个结构根据其自由能大小,按照 SA 算法进行概率筛选;6)经过以上过程对蛋白质空间结构进行优化,重复步骤 3~5 直到产生自由能最低的空间结构,即视作自然状态下稳定的蛋白质空间结构。

2 甲型 H1N1 流感病毒的蛋白质空间结构

2.1 预测蛋白质空间结构

基于 3DHP 模型运用优化的遗传算法,分别以长度为 13 的 HPPHPPHPHPPHP 序列、长度为 17 的 HHHHPPHHHHHHHPPPH 序列、长度为 20 的 PHHHHHHPHHHHPHHPHHPP 序列和长度为 21 的 PHPHPPHPHPPHPPHPHPPHP 序列为例,并与文献[12]所得最低能量进行比较,比较结果见表 1。图 4~5 给出长度分别为 13 和 17 的蛋白质序列的空间结构。

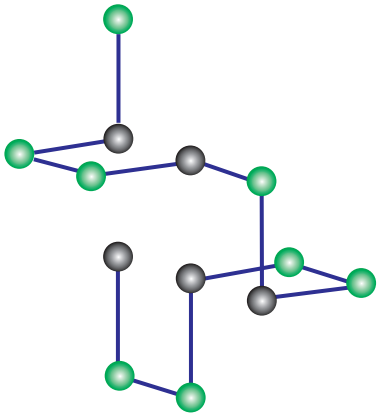


图 4　长度为 13 的氨基酸序列 3DHP 模型(其能量 E=–5)

Fig. 4　Sequence model about the length of the size for 13 (the energy value E=–5)

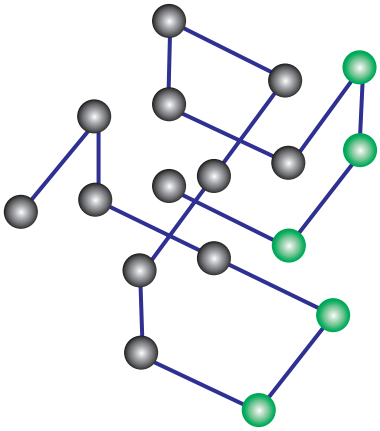


图 5　长度为 17 的氨基酸序列 3DHP 模型(其能量 E=–9)

Fig. 5　Sequence model about the length of the size for 17 (the energy value E=–5)

表 1　3DHP 预测模型最小能量比较

Table 1　Minimum energy comparison on the 3DHP

序号	长度	序列	文献[11、12]	本算法
1	13	HPPHPPHPHPPHP	–5	–5
2	17	HHHHPPHHHHHHHPPPH	–9	–9
3	20	PHHHHHHPHHHHPHHPHHPP	–13	–13
4	21	PHPHPPHPHPPHPPHPHPPHP	–10	–10

2.2　预测甲型 H1N1 流感病毒蛋白质空间结构

选取 8 条甲型 H1N1 流感病毒 HA 蛋白质序列,编号分别为:1RUZ、1RUY、1RU7、3HTO、2WRH、3SM5、3UYW、4B7M (序列来源于 pdb 网站, 网址:http://www.rcsb.org/pdb/home)。运用 3DHP 模型分别对 1RUZ 的 H 链、1RUY 的 H 链、1RU7 的 H 链、3HTO 的 A 链、2WRH 的 H 链、3SM5 的 A 链、3UYW 的 A 链、4B7M 的 A 链进行空间结构预测,获得每个氨基酸节点的空间坐标,并从 PDB 数据库中取相应的已知空间结构的中心碳原子的坐标,将其视为所在氨基酸的空间坐标,由此得到实际测得的蛋白质空间结构骨架,与预测所得的结构进行比较。

2.2.1　构建距离矩阵　对结构的比较采用将蛋白质的空间坐标转换为距离矩阵形式量化表示的方法。对长为 n 的序列,其中氨基酸节点的坐标分别为$(x_1,y_1,z_1),\cdots,(x_n,y_n,z_n)$,i=1,2,$\cdots$,$n$。

定义距离矩阵 $\boldsymbol{A}$:

$$\alpha_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}\ ,$$

$$i,j=1,2,\cdots n。$$

其中α_{ij} 为在矩阵 $\boldsymbol{A}$ 的第 i 行 j 列的元素。

由此建立关于两个蛋白质空间结构的距离矩阵 A 中第 i 行、第 j 列的元素α_{ij} 表示空间结构中第 i 个氨基酸节点到第 j 个氨基酸节点间的欧氏距离。即得到由模型预测所得结构和实验室测得结构的两个距离矩阵。

2.2.2　相关性分析　定义两个距离矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的相关系数:

$$r=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(p_{ij}-\bar{A})(q_{ij}-\bar{B})}{\sqrt{\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(p_{ij}-\bar{A})^2\right)}\sqrt{\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(p_{ij}-\bar{B})^2\right)}}$$

其中:$\bar{A}=\frac{\sum_{i=1,j=1}^{n}p_{ij}}{n^2}$,$\bar{B}=\frac{\sum_{i=1,j=1}^{n}p_{ij}}{n^2}$

相关系数值越接近 1,检验效果越显著,表示两距离矩阵越相似,即对应的两结构越相近,以此评价建立模型的建模效果。

取 1RUZ 的 H 链、1RUY 的 H 链、1RU7 的 H 链、3HTO 的 A 链、2WRH 的 H 链、3SM5 的 A 链、3UYW 的 A 链、4B7M 的 A 链的前 20 个氨基酸为例进行分析,运用 3DHP 模型分别对 8 条氨基酸链进行空间结构预测获得氨基酸节点空间坐标,并在 PDB 数据库中获取相应结构的坐标。

以 1RUZ 的 H 链前 20 个氨基酸 ATNADTICIGYHANNSTDTV 为例, 其在 PDB 数据库中的坐标见表 2。运用 3DHP 模型分别对 1RUZ 的H 链前 20 个氨基酸序列进行空间结构预测得到其空间结构,见图 6,对应的氨基酸节点坐标见表 3。

表 2　1RUZ 的 H 链在 PDB 中中心碳原子 C_α 的坐标(Å)

Table 2　Coordinates of center carbon atoms about 1RUZ-H chain in the PDB

氨基酸	X 坐标	Y 坐标	Z 坐标
1	−16.873	136.138	16.934
2	−13.548	134.679	18.110
3	−11.935	131.538	19.555
4	−8.813	131.095	21.675
5	−6.370	128.26	22.442
6	−5.051	128.027	25.993
7	−3.937	125.844	28.881
8	−5.148	124.911	32.356
9	−3.931	126.697	35.489
10	−4.739	126.933	39.200
11	−3.943	127.835	42.798
12	−1.432	124.995	43.283
13	1.742	126.056	45.132
14	3.848	122.917	44.637
15	7.295	123.550	43.127
16	9.770	121.315	41.274
17	13.302	121.878	40.059
18	14.847	121.624	36.614
19	18.373	121.805	35.289
20	17.706	125.274	33.826

表 3　1RUZ 的 H 链模型氨基酸节点坐标

Table 3　Node coordinates of amino acids on the model of 1RUZ-H chain

氨基酸	X 坐标	Y 坐标	Z 坐标
1	0	0	0
2	1	0	0
3	1	−1	0
4	0	−1	0
5	0	−1	−1
6	0	−2	−1
7	0	−2	0
8	0	−2	1
9	0	−1	1
10	0	−1	2
11	−1	−1	2
12	−1	−1	1
13	−1	−1	0
14	−2	−1	0
15	−2	−1	1
16	−2	−1	2
17	−2	−2	2
18	−1	−2	2
19	−1	−2	1
20	−1	−2	0

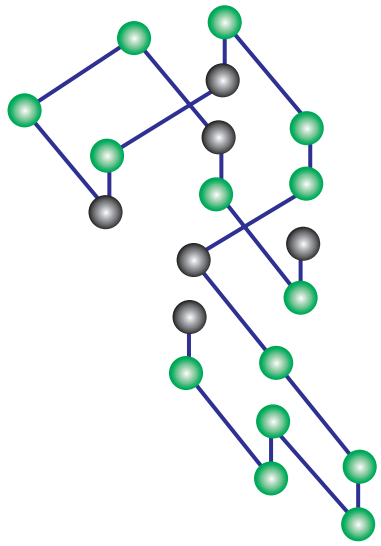


图 6　1RUZ 的 H 链结构预测模型

Fig. 6　Prediction model about 1RUZ-H chain structure

根据距离矩阵定义，对 1RUZ 的 H 链通过 3DHP 模型所得结构和 PDB 数据库提供结构分别构造距离矩阵进行相关性分析。求得两个距离矩阵的相关系数,P 值为，可知这两个距离矩阵高度相关,即预测结构与已知结构存在高度的一致性。对另外 7 条:1RUY 的 H 链、1RU7 的 H 链、3HTO 的 A 链、2WRH 的 H 链、3SM5 的 A 链、3UYW 的 A 链、4B7M 的 A 链进行结构模建后空间结构坐标的距离矩阵相关性分析见表 4，可知预测结构与已知结构也存在高度的一致性。

表 4　其余 7 条序列的 r 值、p 值

Table 4　r value and p value of the rest of the 7 amino acids sequence

序列	r 值	p 值
1RUY 的 H 链	0.725 19	1.735 4e−066
1RU7 的 H 链	0.645 19	1.845 9e−048
3HTO 的 A 链	0.745 92	3.024 7e−072
2WRH 的 H 链	0.694 17	8.559 1e−059
3SM5 的 A 链	0.650 9	1.447 6e−049
3UYW 的 A 链	0.624 04	1.468 3e−044
4B7M 的 A 链	0.714 06	1.305 5e−063

3 结　语

目前来看,对流感病毒的研究大部分都是针对其一级结构的分析,基于理论对流感病毒蛋白质空间结构的研究还很少,虽 Manabu Igarashi 等运用比

较建模的方法对 2009 年甲型 H1N1 流感病毒的结构进行了模建分析[9],但还很少有运用从头计算来对流感病毒蛋白质空间结构进行模建预测。作者采用 3DHP 模型和改良的遗传算法获取最小能量的空间构象,预测甲型 H1N1 流感病毒蛋白质空间结构,直接由一级结构预测甲型 H1N1 流感病毒的蛋白质三维空间结构,得到其空间结构数据,与 PDB 数据库中的实际结构进行比较分析得到了很好的检验结果,这就提供了一种快速预测甲型 H1N1 流感病毒结构的方法。

参考文献:

[1] Layne S P,Monto A S,Taubenberger J K. Pandemic influenza :an inconvenient mutation[J]. **Science**,2009,323(5921):1560.

[2] Tokuriki N,Tawfik D S. Stability effects of mutations and protein evolvability [J]. **Current Opinion in Structural Biology**,2009,19(5):596-604.

[3] Lau K F,Dill K A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins [J]. **Macromolecules**,1989,22(10):3986-3997.

[4] 陈凤飞. 蛋白质结构预测的三角化模型和算法[D]. 武汉:华中科技大学,2006.

[5] Custódio F L,Barbosa H J C,Dardenne L E. Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm[J]. **Genetics and Molecular Biology**,2004,27(4):611-615.

[6] Voelz V A,Bowman G R,Beauchamp K,et al. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39)[J]. **Journal of the American Chemical Society**,2010,132(5):1526-1528.

[7] Dotu I,Cebrian M,Van Hentenryck P,et al. On lattice protein structure prediction revisited [J]. **Computational Biology and Bioinformatics**,2011,8(6):1620-1632.

[8] Islam M,Chetty M. Clustered memetic algorithm with local heuristics for ab initio protein structure prediction[J]. 2013,17(4):558-576.

[9] 张玲,高洁. 甲型流感病毒 HA 蛋白质序列的预测[J]. 食品与生物技术学报,2013,32(8):828-831.
ZHANG Ling,GAO Jie. Prediction for base of influenza virus A/HA protein sequence [J]. **Journal of Food Science Journal of Health Laboratory Technology**,2013,32(8):828-831.(in Chinese)

[10] Igarashi M,Ito K,Yoshida R,et al. Predicting the antigenic structure of the pandemic(H1N1) 2009 influenza virus hemagglutinin [J]. **PLoS One**,2010,5(1):e8553.

[11] Dill K A,Bromberg S,Yue K,et al. Principles of protein folding—a perspective from simple exact models [J]. **Protein Science**,1995,4(4):561-602.

[12] ZHOU Changjun,HOU Caixia,ZHANG Qiang,et al. Enhanced hybrid search algorithm for protein structure prediction using the 3D-HP lattice model[J]. **Journal of Molecular Modeling**,2013,19(9):3883-3891.