三角网格模型的快速树搜索算法及可设计性分析*)

李小妹　王能超

(华中科技大学计算机科学与技术学院　武汉 430074)

摘　要　在蛋白质折叠格子模型的可设计性特征研究中，为了克服以往方格模型具有奇偶问题这一缺点，本文利用三角网格模型来进行穷举搜索。在简化的网格模型中，序列折叠为某一结构的能量值为在结构心部疏水氨基酸的个数取负值。在蛋白质折叠模型的二维 4+5+6+5+4 三角网格中穷举了所有的序列和致密结构。其中序列由两类氨基酸(疏水氨基酸和亲水氨基酸)组成，排除正反对称序列共 $2^{12}+2^{23}=8392704$ 种不同序列。在由 24 个格点组成的三角网格模型中共得到 219093 种简化结构串。在穷尽搜索算法中，为实现快速搜索，通过树结构将相似的结构串尽量聚类，通过计算各树结点的目标能量值以减少搜索算法中所需的计算量。经并行实验验证，利用该树结构可使快速搜索算法达到指数级加速比。最后对计算所得结果进行了统计分析。

关键词　三角网格模型，蛋白质折叠，聚类树，可设计性

Fast Tree Search for Triangular Lattice Model and Designability Analysis

LI Xiao-Mei　WANG Neng-Chao

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

Abstract　Using a triangular lattice model to research designability of protein folding, we overcame parity problem of previous cubic lattice model. This paper enumerated all the sequences and compact conformations on a two-dimension triangular lattice simple model of size 4+5+6+5+4. The energy of a sequence folded into a structure is minus the number of hydrophobic amino acids in the "core" of the structures. We used two types of amino acids-hydrophobic (H) and polar (P)-to make up the sequence, so there are $2^{24}+2^{12}$ different sequences excluding the reverse symmetry sequences. The total simple solvation string number of distinct structures was 219093 excluding reflection symmetry in the self-avoiding path of length 24 triangular lattice model. Based on triangular lattice model, a fast search algorithm is presented in order to search the protein conformation space by constructing a fast search tree. The algorithm decreased the computation by computing the objective energy of tree nodes. The parallel experiments proved that the fast tree search algorithm yielded an exponential speed-up in models of size 4+5+6+5+4. To understand the search result we perform statistical analysis in designability.

Keywords　Triangular lattice model, Protein folding, Fast search tree, Designability

蛋白质结构的问题[1]多年来一直吸引着各个领域的研究者，氨基酸序列和蛋白质的三维结构之间的关系不仅是生物学上的重要问题，同样也是科学研究中的基础问题。通常天然蛋白质构象空间庞大，但其折叠速度却相当快，这就是著名的 Levinthal's Paradox：蛋白质的折叠态不是通过空间的随机搜索得到的[2]。尽管四十年来研究者做出了不懈的努力，但是这一问题仍然没有得到根本解决。通过蛋白质的格子模型，可以从高可设计度结构出发探索其中的奥秘。

蛋白质是通过共价键将各种氨基酸的基本原子连接在一起的大分子。蛋白质的结构决定了其性能，因此理解蛋白质的结构和性能的关系至关重要。1963 年 Anfinsen 通过试验得出蛋白质的氨基酸序列在失性后可自发恢复其天然构象[3]，并通过该试验得到两个结论：(1)对大多数单域蛋白质而言，编码蛋白质的氨基酸序列即可以决定它的空间构象；(2)蛋白质的天然构象选择的是能量最低的结构。

基于这一理论基础，人们提出了用全原子模型来进行蛋白质折叠模拟[4]，由于该模拟涉及到大量原子以及复杂的原子力场模型，所需的计算量庞大，目前的计算能力无法达到。20 世纪 80 年代末 Dill 等人引入了格子模型[5,6]，这是一种极为简化的蛋白质结构模型。在格子模型中，一个蛋白质的结构由多个结点组成的链在二维或三维的正方格子空间的自回避行走所得的路径来表示，一条路径也就对应一种格子模型的致密结构。格子模型必须满足两个限制条件：(1)氨基酸序列中的共价键不能打断；(2)每个氨基酸占据一个格点，但一个格点不能被两个氨基酸同时占用。

该模型的生物学基础就是，蛋白质折叠的主要驱动力是由于氨基酸间的疏水作用[7,8]，使得疏水氨基酸尽量压缩到蛋白质结构的内部，而亲水氨基酸则处于蛋白质的表面，因此其体系的能量来自于疏水残基间的相互作用。构建该模型的目的是找到每条序列对应的能量最小且唯一的致密结构，也即该序列对应的基态结构。若一条序列对应的能量最小结构不止一个，那么该序列很有可能就是一种随机序列。因为任何具有生物学功能的序列，其天然构象是唯一的。对应这种序列我们在研究中不予考虑。一个致密结构可能对应多种序列，我们就把一种致密结构对应的序列数定义为该结构的可设计性。

为了通过该简化模型来研究蛋白质的热动力学性能和其它特点，需要枚举所有的序列和结构。由于天然球蛋白的结构几乎都是十分紧密的，因此在穷举搜索研究中只需计算格子模型的致密结构。近年来，对二维和三维的格子模型已进行了一些研究，比如 Li. H 等人于 1996 年研究了 3×3×3 以及 6×6 的 HP 网格模型[9]，其研究仅仅针对两种不同的氨基酸，即疏水氨基酸(Hydrophobic)和亲水氨基酸(Polar)。发现与高可设计性结构相联系的序列在热动力学上更稳定，比其它的随机序列折叠更快，而且这些高可设计性结构具有规则的二级结构以及整体的对称性。在 2002 年 Li. H 等人又

*)基金来源：国家自然科学基金，编号 70271069。李小妹　博士研究生，主要研究方向：生物信息学、复杂系统的演化分析。

对20种不同疏水特性的氨基酸研究了其格子模型，得到了相似的结果[10]。同年，Henry Cejtin 等人对 4×3×3 的网格模型进行了穷举搜索[11]。可以看出以前的研究模型都是集中在方形格子模型中，而这种方格模型在致密结构中存在着奇偶问题，同时天然蛋白质的结构表面是尽量光滑过渡的，而格子模型中存在一些尖角，为了改进该模型的这些问题，我们提出了用 4+5+6+5+4 的二维三角网格模型结构来研究蛋白质的可设计性问题，见图 1。

同时，我们知道蛋白质的枚举搜索需穷尽所有的序列和致密结构，如对 3×3×3 的立方网格模型而言，其可能的序列总数是 2^{27}，而致密结构总数为 103346 种，则所需计算的结构能量数约为 1.39×10^{13}，对三角网格模型 3+4+5+4+3 所需计算量为 $2^{19}\times20486=1.07\times10^{10}$，而对规模为 4+5+6+5+4 的三角网格模型，经统计发现其致密结构共 1475782 种，则其计算量达 $2^{24}\times1475782=2.48\times10^{13}$。因此寻找快速算法对穷举搜索而言是关键。本文在穷尽搜索算法中，对所有简化结构串构建了快速搜索树，将相似的结构尽量聚类，通过计算其目标能量值，以减少搜索算法中所需的计算量。经并行试验分析验证，该快速搜索算法可达到的加速因子为 $\Theta(1.486^{\log_2 M-\log_2 Max_{op}})$，其中 M 为模型的唯一结构串总数，$Max_{op}$ 为叶结点允许的最优结构串数。通过试验发现对 4+5+6+5+4 的三角网格模型穷举搜索加速比可达到 189。

1 模型

以前在蛋白质折叠可设计性的研究中，所使用的均为正方网格模型，该模型自由度少、操作方便且可以用来对全部可能的序列和致密结构进行全搜索。但该模型存在着奇偶问题，也就是说，链中只有处于奇数位和偶数位的氨基酸间能形成拓扑接触对，同时位于奇数位和偶数位的氨基酸不可能形成这种结构关系。而在三角网格模型中不存在这种问题，而且该模型的表面过渡比方形网格模型光滑，更接近天然蛋白质的结构。

在 4+5+6+5+4 三角网格模型的致密结构搜索中，共找到 5903128 种自回避路径，其中包括一种旋转对称，二种轴线对称，也即每种有向构象存在四种同构结构，排除这些对称结构剩下 1475782 种不同的有向自回避路径，其中正向结构和反向结构对称的构象共 738189 对，逆向对称的构象 596 种（见图 1），因此排除正反对称的结构后还剩 738189+596=738785 种不同的致密结构。

我们知道为了形成更低的能量构象，疏水氨基酸尽量占据结构的心部。在该模型中，我们定义氨基酸序列为 σ_i，在氨基酸序列中，我们只选用两种，疏水氨基酸（$\sigma_i=1$）和亲水氨基酸（$\sigma_i=0$）。而结构序列则是格子模型中的致密自回避路径。氨基酸序列折叠为某一结构的能量计算公式可考虑为每个氨基酸的埋藏程度。具体的表达公式为：

$$E=-\sum_{i=1}^{N}\sigma_i s_i$$

其中 s_i 表示为氨基酸链中第 i 个位点在结构中的埋藏程度。我们把结构中位于心部位点的埋藏度指派为“1”（$s_i=1$），而面部位点的埋藏度指派为“0”（$s_i=0$），在 4+5+6+5+4 的二维三角网格模型中的心部位点（用黑色表示）和面部位点（用灰色表示）见图 1。利用该简化结构模型，有些致密结构具有相同的结构串。1475782 种自回避路径共得到具有唯一结构串的构象 219093 种，其中逆向对称结构 219 种，为节省存储空间将 109497 对正反对称结构用一个结构串来表示，因此需存储的结构串 109656 种，其中 25825 种致密结构得到的结构串是唯一的，而其余 83831 种简化结构串对应两种或多种致密结构路径。

2 树搜索算法

从图 1 中我们知道对于 4+5+6+5+4 三角网格模型，一个结构串应有 10 个心部位点，14 个面部位点。而我们的搜索计算目的是找出待测序列$\{\sigma_i\}$对应的能量值最小且唯一的目标结构$\{s_j\}$。简而言之就是找到目标序列与简化结构点积最大且唯一的结构。很显然，目标结构与待测序列应有较好的相似性。我们可以考虑将待测序列定位到对相似结构串的搜索，以避免大量的无用搜索以达到加快搜索的目的。为了实现这种快速搜索，可以将相似的结构尽量聚类，通过目标能量值的计算来尽快定位目标结构，以减少所需搜索的结构串。

实现该算法的最佳方案就是利用树结构。首先考虑树中每个结点应包含的信息（见图 2），其中长度为 N 的 01 串向量 K 表示在该结点下所有结构串的已知信息，若该结点下的所有结构串在某一位点值为“1”，则向量 K 中的该位点值为“1”，否则为“0”，我们称之为已知位点（known-ones）；长度为 N 的 01 串向量 U 表示在该结点下所有结构串的未知信息，若所有结构串在某一位点的值不确定，则向量 U 中的该位点值为“1”，否则为“0”，我们称之为未知位点（unknown-ones）；整数 m 表示在未确定的位点中还应包含的“1”的个数，我们可以称之为缺失位点个数（Missing ones）。

为了使结构串尽量聚类，各树结点的分支点选取是关键。我们利用熵值最小的原理来确定分支位点。首先计算当前结点中所有结构串在 N 个位点上“0”和“1”出现的概率值并计算各位点的熵值，其中熵值最小的位点作为该结点的分支点。熵值的计算公式为：

$$S=\mathrm{Min}\{-(p_i\log p_i+q_i\log q_i)\}$$

其中 p_i 为位点为“1”的概率，q_i 为位点为“0”的概率。

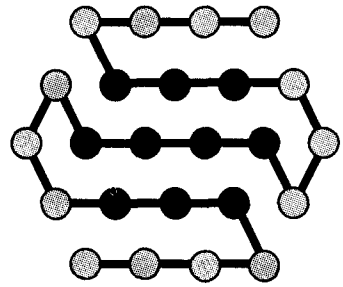

图 1 4+5+6+5+4 模型心部位点（黑色）和面部位点（灰尘色）

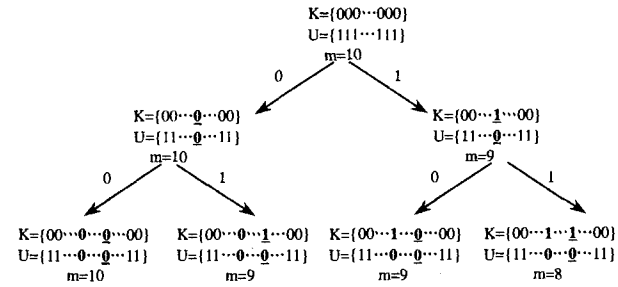

图 2 结构的聚类树加粗带下划线字体表示当前的 UK 更新信息，加粗不带下划线字体表示已更新的 UK 信息，K 中的省略号表示“0”，而 U 中的省略号表示“1”。

具体搜索树的构造结构如下：根结点包含所有的结构串，然后根据熵值最小原理，确定当前结点的分支位点，若当前树

结点在分支位点上的位点值为"0"则生成当前结点的左子结点，该左子结点应包含当前结点在该位点为"0"的所有结构，若当前位点值为"1"则生成当前结点的右子结点，该右子结点应包含当前结点在该位点为"1"的所有结构，并更新各个子结点的已知信息、未知信息以及缺失位点个数。若子结点包含的序列数目大于预定的叶结点最大结构串数 Max，则该过程继续，否则停止转向下一个子结点，直到所有的子结点包含的结构串数小于 Max 则停止。

为了加快树的搜索，我们必须知道各树结点的目标能量值，很显然对于某个待测序列$\{\sigma_i\}$，各树结点对应的该结点下所有结构串点积的上限值(也即目标值)可表示为如下公式：

$$B_{\perp}=\mathrm{Min}\{\sigma_i*(U+K),\sigma_i*K+m\}$$

在对聚类树进行搜索时，首先根据上面的公式确定各树结点的目标能量值，再根据该目标能量值来确定是否搜索该结点。若达不到该目标值，则不用搜索进行回溯。若能达到当前目标值则判断当前结点是否为叶结点，若当前搜索的结点为叶结点，则搜索该叶结点的每个结构串；若为非叶结点，先检查与待测序列在该位点匹配的子结点(经试验验证仅此一步即可提高加速因子 1.782 倍)。若整个聚类树已搜索完毕仍未找到目标结构串，则将目标能量值减 1，再进行树搜索。若找到目标序列且唯一则读出结果且转向下一条待测序列。若不唯一则直接转向下一条待测序列。

3 试验

对于 4+5+6+5+4 三角网格模型，其唯一结构串为 219093 种，为存储方便将其中排除反向构象的 109656 种结构拿来建树，其中含反向构象的序列需另做标记，计算时应考虑反向结构串。首先我们利用前面介绍的算法对唯一结构串构建聚类树，并对全部待测序列进行有目标的搜索，记录下具有唯一能量最小的所有序列，及其对应的结构串和最低能量值，以便于进一步做统计分析。

很明显该算法有很好的并行性能，我们在联想深腾 1800 高性能机群系统的服务器上进行计算，该机群的结点机采用基于 Intel 构架的商用服务器，通过高速通讯网络实现结点间的互连，对外提供单一系统映射。机群包含 24 个结点，每个结点均为 2.8GHZ 的微处理器。

在聚类树的搜索阶段，需要对全部待测序列数据实施穷举搜索，而每个序列数据的搜索是独立的，因此，只要对全部序列数据进行正确分割，每个序列就可以独立执行搜索操作，可以并行处理。我们的做法是：把本阶段全体数据分为 $m+1$ 组，分别分派给 $m+1$ 个处理机执行，又因为序列数据与结构串间的相关性不同，树搜索的操作计算量难以估算，为保持负载平衡，采用工作池的动态任务分派策略进行并行处理。

并行算法描述如下：

因待处理的序列对应某一种简化结构，则其反向序列对应该简化结构的反向结构，因此待处理的序列可排除反向序列，则待测序列总数为 $2^{12}+2^{23}$ 种，我们将这些待测序列数据分成 $n=129$ 组 $f_i(0\leqslant i<129)$，前 128 组每组 2^{16} 个序列，最后一组包含 2^{12} 种序列。设参数 num 表示主结点机当前准备发送的数据组标号，初始为 0；设参数 count 表示已处理完毕的数据组，初始为 0。主结点机 P_0 作如下操作：(1)取出 m 个数据组 $f_0, f_1 \cdots\cdots f_{m-1}$，分别发送给结点机 P_1，$P_2 \cdots\cdots P_m$；num=m；count=0；(2)对 f_{num} 组数据实施树搜索，执行完毕时 count 增 1；num 增 1；(3)接收各结点机 $P_i(1\leqslant i\leqslant m)$ 的处理结果；count 增 1；(4)如果 num$<n-1$，则将第 f_{num} 组数据发送给结点机 P_i，否则通知结点机 P_i 终止执行；num 增 1；(5)如果 num$<n-1$，则转(2)执行；如果 count 等于 n，则转(6)；(6)集合各结点机的处理结果并终止程序执行。其他各结点机 $P_i(1\leqslant i\leqslant m)$ 循环作以下操作直至收到终止信息：接收来自主结点机 P_0 的数据组；实施树搜索；发送处理结果给 P_0。

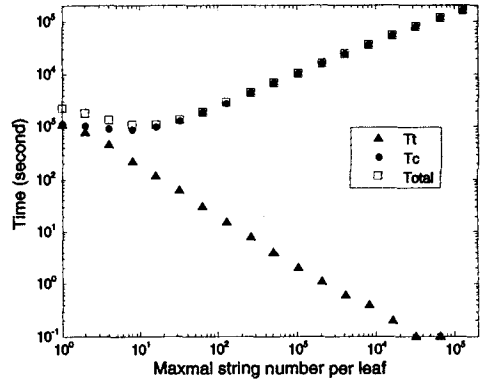

图 3 在不同 Max 值时所需的建树和计算时间

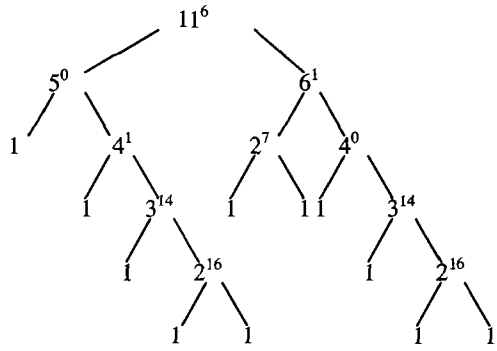

图 4 三角网络模型 4+5+6+5+4 的 109656 种简化结构当 Max=1 时构建的部分树结构。其中数据表示该结点包含的结构串数，上标为该结点的分支位点。

在计算中将叶结点允许的最大结构串数 Max 作为一个可变参数，通过试验得到所需的建树时间(T_t)树搜索时间或计算时间(T_c)以及总时间(Total=T_t+T_c)见图 3。图中的计数时间为所需的 CPU 时间，单位为秒。从图中可以很明显地看出在 Max 为 8 时所需计算时间和总时间均为最低。建树所需时间是随着叶结点允许的最大结构串数的增加呈指数递减，这是因为在建树时，每个结点均需对所有结构进行搜索，而随着 Max 的减小，树的结点呈指数增加，从而建树时间也就随之指数增加。而总的来说计算所需时间(T_c)则是随着 Max 的增加呈指数递增，很显然是由于随着 Max 的增加，叶结点数指数增加，通过计算目标能量值，排除了大量的无用搜索，能更快地定位到目标结构串，因此所需计算时间也就指数递减。但是当 Max 小于 8 时，计算所需时间比 max 为 8 时所需时间还要多，这是因为从图 4 中可以看出当 Max 较小时，各子结点包含的序列串较少，而其父结点的已知信息已足够描述这些序列，继续二分下去只会增加目标能量值的计算，使得所需计算时间不降反升。当然不同的模型其最优的 Max 值会不同，可以对部分序列进行计算以得到计算时间最优的 Max 信息。当 Max 为模型的结构串总数时，此时树中仅有一个根结点，它包含所有的结构串，对树的搜索也就是对

所有结构串进行全搜索，此时的建树时间则为0。对4+5+6+5+4三角网格模型得到的最优时间加速比为189，利用直线拟合算法得到图3中计算时间的斜率为0.57155($8 \leqslant Max$)，因此其加速因子的底数为$2^{0.57155}=1.486$。很显然是一个指数加速算法$\Theta(1.486^{\log_2 M-\log_2 Max_{op}})$。其中$M$为模型的唯一结构串总数，$Max_{op}$为叶结点允许的最优结构串数。

4 统计分析

由于一条HP序列若以某种结构串作为其唯一基态，则其反向HP序列则以该结构串的反向序列作为其唯一基态，为简化分析，我们排除反向结构和反向串做统计分析。通过上述计算，排除正反对称序列串将$2^{23}+2^{12}$种HP序列，对219093种结构进行枚举搜索，对其中109656种结构串进行统计发现至少对应一条序列的构象有25825种，占所有结构的23.55%。通过计算找到了每种结构对应的序列数，也即是可设计性Ns。每种结构对应的可设计性差别很大。可设计性最大的结构对应101种序列。对给定的Ns，其对应的序列数随着Ns的增加快速单调递减(见图5)。具有唯一基态的序列共181375种，因此每个致密结构的平均可设计度7.02。图5中位于曲线尾部的结构其可设计度均远远大于平均可设计度。我们知道蛋白质的空间结构较一级序列更稳定，更能承受一定的序列突变而不改变其三维构象，这与试验中得到的一些高可设计性结构能对应较多的序列相吻合。

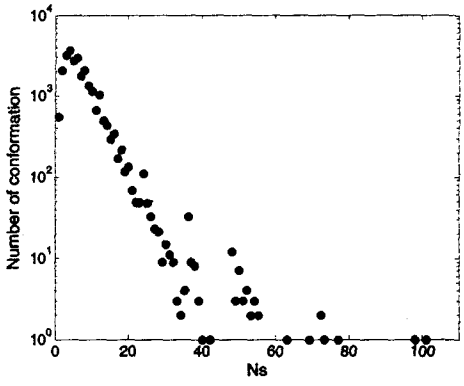

图5 可设计性与构象数间的关系

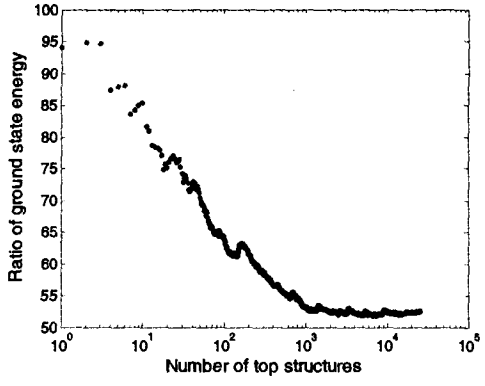

图6 基态能量−10与可设计性的关系

在不同可设计性结构的前提下，对各种基态能量所占的比率进行了统计，发现在计算得到的181375种非简并性序列中，基态能量为−10，−9，−8和−7所占的比率分别为62.64%，33.50%，3.83%和0.32%。其中基态能量在−8以上的序列占到99.7%。因此我们仅对基态能量为−10，−9和−8时基态能量与可设计性的关系进行了统计，统计结果如图6和图7所示，很明显可设计性越高，其对应的基态能量越低。从图中可以发现，可设计性最高的结构其对应的基态能量为−10的序列平均占到95%，而基态能量为−9和−8的序列仅占到5%和0%。而可设计性较低的部分结构其对应的基态能量为−10的序列仅占到52%，而基态能量为−9和−8的序列占到37%和11%。该试验表明高可设计性结构对应的序列的基态能量较一般结构低，因此在热动力学上要比普通结构更稳定。这就从可设计性的角度验证了高可设计性结构在热动力学上更稳定。

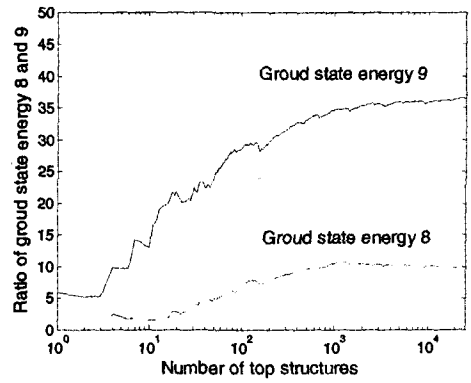

图7 基态能量−9，−8与可设计性的关系

结论 从算法出发，我们利用待测序列与目标结构串的关系，构建了快速聚类树，将相似的结构串尽量聚类，通过目标能量值的计算避免了大量结构的搜索，使得该算法达到了指数级的加速因子。从模型出发，我们提出了用三角网格模型来研究不同结构的可设计度问题，通过搜索所有可能的结构串和所有的HP序列，试验发现不同的致密结构其可设计度有很大的区别。高可设计度的结构所对应的序列基态能量更低，也就表明其结构的热动力学结构更稳定。而从结构上来看，三角网格模型不含奇偶问题，结构表面过渡光滑，与天然蛋白质更为接近。

参考文献

1 Anfinsen C B. (1973) Principles that govern the folding of protein chains. Science, 1973, 181: 223～230
2 Frederic M. Richards The protein folding problem. Scientific American, January 1991. 54～63
3 Levinthal C. Are there pathways for protein folding? J. Chim Phys, 1968, 65: 44～45
4 Dill K A. polymer principles and protein folding. Protein Science. 1999, 8: 1166～1180
5 Dill K A. Theory for the folding and stability of globular gproteins. Biochemistry, 1985, 24: 1501
6 Dill K A, Bromberg S, Yue K. Principles of protein folding: A perspective from simple exact models. Protein Science, 1995, 4: 561～602
7 Chan H S, Dill K A. Intrachain loops in polymers: Effect of excluded volume. J Chem Phys, 1989, 90: 492～509
8 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: 765～768
9 Li H, Helling R, Tang C, Wingreen N. Emergence of Preferred Structures in a Simple Model of Protein Folding. Science 273, 666 (1996)
10 Li H, Tang C, Wingreen N. The Designability of Protein Structures: A Lattice-Model Study using the Miyazawa-Jernigan Matri. Proteins 49, 403 (2002)
11 Cejtin H, Edler J. Fast tree search for enumeration of a lattice model of protein folding. J Chem Phys, 2002, 116(1): 352～358
12 Li H, Tang C, Wingereen N. Are protein folds atypical? Proc. Natl Acad Sci USA, 1998, 95: 4987～4990