

不同类型氨基酸网络参量与蛋白质折叠的关系*

严立成^{1)**} 苏计国^{1,2)**} 陈慰祖¹⁾ 王存新^{1)***}

(¹⁾北京工业大学生命科学与生物工程学院, 北京 100124; ²⁾燕山大学理学院, 秦皇岛 066004)

摘要 理论和实验研究表明, 蛋白质天然拓扑结构对其折叠过程具有重要的影响. 采用复杂网络的方法分析蛋白质天然结构的拓扑特征, 并探索蛋白质结构特征与折叠速率之间的内在联系. 分别构建了蛋白质氨基酸网络、疏水网、亲水网、亲水-疏水网以及相应的长程网络, 研究了这些网络的匹配系数 assortativity coefficient 和聚集系数 clustering coefficient 的统计特性. 结果表明, 除了亲水-疏水网, 上述各网络的匹配系数均为正值, 并且氨基酸网和疏水网的匹配系数与折叠速率表现出明显的线性正相关, 揭示了疏水残基间相互作用的协同性有助于蛋白质的快速折叠. 同时, 研究发现疏水网的聚集系数与折叠速率有明显的线性负相关关系, 这表明疏水残基间三角结构(triangle construction)的形成不利于蛋白质快速折叠. 还进一步构建了相应的长程网络, 发现序列上间距较远的残基接触对的形成将使蛋白质折叠进程变慢.

关键词 氨基酸网络, 疏水作用, 折叠速率, 匹配系数, 聚集系数

学科分类号 Q6

DOI: 10.3724/SP.J.1206.2009.00760

蛋白质是生命系统中重要的生物大分子, 它在生物体的繁衍、新陈代谢、生长发育等一切生命活动中发挥着重要的生物学功能. 而蛋白质特定的三维结构是行使其相应生物功能的物质基础. 研究蛋白质结构的形成机制具有重要的理论和现实意义. 随着疯牛病等蛋白质折叠疾病的出现, 蛋白质折叠的研究更加具有紧迫性与现实意义. 自由能曲面理论为蛋白质折叠机理的研究提供了很好的理论框架, 被广泛应用于蛋白质折叠热力学和动力学机制的研究^[1]. 然而, 能量函数和能量曲面的构建又存在一些困难. 大量的实验和理论研究表明, 蛋白质的折叠速率和机制在很大程度上是由天然态的拓扑结构所决定的^[2], 而不是完全取决于原子之间的精细的相互作用^[3-4]. 因而, 探究蛋白质拓扑结构和折叠机制之间的内在联系, 通过蛋白质拓扑特征的分析获取折叠动力学信息, 是一个很有意义的研究内容. 基于此, 本文引进复杂网络的理论和方法构建氨基酸网络, 来研究蛋白质折叠问题, 为解决上述问题提供了一个全新的思路.

复杂网络在分析蛋白质结构的拓扑特征方面, 无疑是一个非常好的工具. 目前常用的做法是, 构建蛋白质的氨基酸网络, 蛋白质的每个氨基酸作为

网络的一个节点, 氨基酸之间的相互作用作为节点之间的连接. 通过对网络特征量的分析可以获取蛋白质折叠速率、折叠核等重要信息. 利用蛋白质氨基酸网络拓扑特征的分析来发掘蛋白质折叠的相关信息, 已经取得了一些研究成果, 揭示了蛋白质拓扑结构对折叠的决定性作用. 比如: 折叠过渡态氨基酸网络中介数(Betweenness)大的节点对应于折叠核心^[5]; 蛋白质天然结构中全局近邻中心节点(globally full-closeness centrals)对应于折叠核^[6]; 过渡态的拓扑结构直接决定了蛋白质是否能够折叠到天然态^[7]; 疏水和亲水氨基酸网络、长程和短程氨基酸网络具有不同的拓扑特征, 在蛋白质折叠中起着不同的作用^[8-9]; 氨基酸网络的匹配性质(Assortative mixing)与蛋白质折叠速率存在很高的关联性^[10]等.

* 国家自然科学基金(10974008, 10947014)和教育部博士点基金(200800050003)资助项目.

** 共同第一作者.

*** 通讯联系人.

Tel: 010-67392724, E-mail: cxwangbjut@gmail.com

收稿日期: 2009-12-23, 接受日期: 2010-03-23

构成蛋白质的 20 种氨基酸按其物理化学性质可以大致分为两类: 疏水残基和亲水残基. 蛋白质在生理环境下, 疏水残基会相互靠近并被包埋在蛋白质内部, 而亲水残基则倾向于分布在蛋白质表面. 研究发现, 疏水作用是蛋白质折叠的主要驱动力. 在本工作中, 基于残基的差异性, 分别构建了氨基酸网、疏水网、亲水网、亲水-疏水网, 进而计算了各网络的匹配系数和聚集系数, 并考察了这两个参量与折叠速率之间的关系, 发现疏水网的匹配系数与折叠速率之间有着非常明显的线性正相关关系. 另外, 还构建了相应的长程网络以观察各网络中长程残基间的相互作用对蛋白质折叠过程的影响.

1 研究体系和方法

1.1 蛋白质结构数据

从蛋白质数据库(PDB)下载了 155 个蛋白质, 包括了全 α 、全 β 、 α/β 和 $\alpha+\beta$ 4 种主要的蛋白质结构类型. 这些蛋白质的分辨率至少达到 1.8 Å, 序列同源性均低于 20%, 残基数目 (n_r) 范围为 $83 < n_r < 1011$, 其中包括有折叠速率实验数据的 23 个单体二态球状蛋白. 这 23 个蛋白质的 PDB 代码分别为 1LMB、1HRC、2ABD、1YCC、1CSP、1TEN、1SHF、2AIT、3MEF、1MJC、1AEY、1SHG、1SRL、1PKS、1UBQ、1CIS、1PCA、2PTL、1HDN、1APS、1URN、1FKB、1VIK. 它们的残基数 (n_r) 范围为: $56 < n_r < 402$, 折叠速率的自然对数值变化范围为: $-1.48 < \ln k_F < 9.62$.

1.2 氨基酸网络的构建

将 20 种氨基酸分为两类: 疏水残基(C, M, F, I, L, V, W, Y)和亲水残基(A, G, T, S, N, Q, D, E, H, R, K, P). 采取以下几种方法构建网络: a. 将每个残基用 C α 原子代替, 以统一的截断半径 8 Å 判断两个残基之间是否接触来确定它们之间是否存在相互作用, 从而建立残基之间的连接, 构建氨基酸网; b. 与氨基酸网类似, 但只取疏水残基为节点, 并将其按上述方法相连接, 构建疏水网; c. 类似地, 只取亲水残基为节点构建亲水网; d. 同时取亲水和疏水残基为节点, 但只考虑亲水残基和疏水残基之间的相互作用, 而忽略其他残基之间的相互作用, 构建亲水-疏水网; e. 进一步构建上述各网络的长程网络, 忽略短程的相互作用. 在这些网络中, 除了主链上近邻的残基, 只有主链上相隔至少 12 个残基且在截断半径

内的两个残基之间认为存在相互作用.

1.3 氨基酸网络参量的计算

本文分析了上述方法所构建的各种网络的拓扑结构特征, 包括匹配系数、聚集系数, 并考察了这些参量与蛋白质折叠机制之间的内在关联. 度是复杂网络理论的一个基本几何量, 一个节点的度是指与此节点连接的边的数量. 匹配模式是考察复杂网络度分布之间相关性的重要参量, 它的意义在于考察度值大的点倾向于与度值大的点相连接还是与度值小的点相连接. 网络匹配系数 r (Pearson correlation coefficient) 的计算方法如下^[10]:

$$r = \frac{M^{-1} \sum_i j_i k_i - [M^{-1} \sum_i \frac{1}{2} (j_i + k_i)]^2}{M^{-1} \sum_i \frac{1}{2} (j_i^2 + k_i^2) - [M^{-1} \sum_i \frac{1}{2} (j_i + k_i)]^2} \quad (1)$$

其中 j_i , k_i 分别为第 i 条边的两个节点的度值, M 是网络中边的个数, i 的取值范围为 $1 \leq i \leq M$.

聚集系数是反映网络聚集程度的几何量, 即考察一个节点的近邻之间的连接情况^[12]. 可以这样定义聚集系数, 对于一个节点 V , 找到其近邻集合 N_v , 记 $n = |N_v|$, N_v 中存在的边的数量为 M , 若以 C_v 表示该点的聚集系数, 则

$$C_v = \frac{M}{C_n^2} \quad (2)$$

其中 C_n^2 为该近邻集合中所有可能存在的边的数目. 于是可以得到所有节点的集聚程度, 它的统计分布是刻画网络的一个重要几何量, 其平均值称为平均集聚程度 C . 平均聚集系数的意义在于考察网络的整体紧密程度, 如果一个网络的平均聚集系数大, 将对网络中的信息渗透和传递以及网络的结构稳定性都有很重要的意义.

2 研究结果

2.1 氨基酸网络的拓扑特征

目前, 复杂网络的方法被应用于很多领域的研究, 如社会网络、技术网络、生物网络等. 研究发现, 在很多现实的网络中都存在着一种连接偏好性, 即在网络中, 任意一条边的两个顶点的度值具有某种规律, 这种规律可以用网络的匹配性来刻画. 如果一个点倾向于跟自身度值相似的点相连接, 称之为正匹配, 反之, 称之为负匹配. 目前研究表明, 社会网络是正匹配的, 如科学家合作网络、演员合作关系网络、疾病传播网络等. 而在一些技术网络和生物网络中则表现出负匹配的特征,

如万维网、神经网络、海洋食物网络等。人们发现,正匹配网络的健壮性和渗透性要明显高于负匹配网络,因此在疾病传播网络中要切断疾病传播路径以达到控制目的就相对较难,而技术网络则脆弱得多,容易受攻击而导致整个网络瘫痪^[1]。蛋白质的结构、折叠机制和功能有着密不可分的关系,因此本文构建了各种蛋白质氨基酸网络,并定量计算了各网络的匹配系数,考察其中是否存在匹配性,并研究这种特征与折叠机制有什么样的关系。

由图1中可以看出,在155个蛋白质所构建的网络中,氨基酸网络的匹配系数全部为正,平均值达到0.360,标准方差为0.079,显示出很高的正匹配性,这与Bagler等^[10]的统计结果一致。氨基酸网络的正匹配性质有利于提高蛋白质的渗透性,使得蛋白质的抗攻击能力增强,同时生物信息的传递也更加便捷,对于其发挥生物学功能有着重要的意义;疏水网的网络匹配系数全部为正,平均值达到0.400,标准方差为0.094,说明在疏水基团内部,度值大的点倾向于与度值大的点相连接;亲水网的匹配系数基本都大于0(只有一个为负,其值为-0.039),平均值达到0.361,标准方差为0.122;亲水-疏水网的匹配系数都接近于0(其中小于零的有95个,大于零的有60个),平均值为-0.026,标准方差0.082,基本没有匹配性。

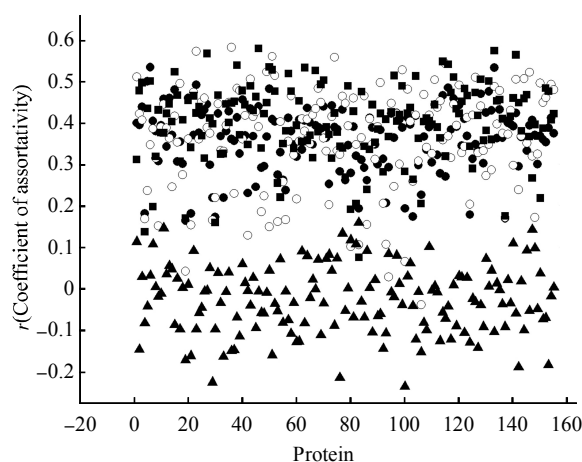


Fig. 1 Coefficient of assortativity for the amino acid network of the 155 proteins

Amino acid network(●), Hydrophobic-Hydrophilic amino acid network(▲), Hydrophilic amino acid network(■), Hydrophobic amino acid network(○).

网络的平均聚集系数反映网络节点聚合成小聚

集体的能力。研究表明,疏水网和亲水网都会形成一些小聚集体和孤立的点,并且亲水网中小集体和孤立点的数目要多于疏水网^[10]。由图2可见,氨基酸网的平均聚集系数最大,平均值达到0.543,标准方差为0.021,疏水网的平均聚集系数平均值为0.396,标准方差为0.043,亲水网的平均聚集系数平均值为0.528,标准方差为0.024,亲水-疏水网的平均聚集系数平均值最小,其值为0.342,标准方差为0.025。

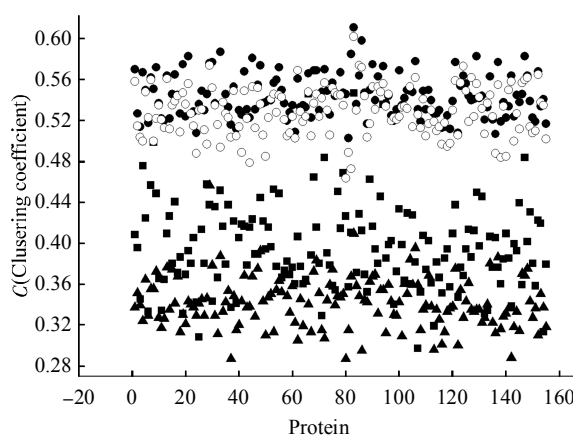


Fig. 2 Clustering coefficients of the amino acid network of the 155 proteins for 4 types of networks

Amino acid network(●), Hydrophobic-Hydrophilic amino acid network(▲), Hydrophilic amino acid network(■), Hydrophobic amino acid network(○).

2.2 网络拓扑参量与蛋白质折叠之间的关联

2.2.1 网络匹配系数与蛋白质折叠速率之间的关联

大量的理论和实验研究表明,在长时间的生物进化过程中,蛋白质序列得到了优化,使其具有“极小的能量阻挫”特征,即蛋白质的折叠过程在很大程度上由天然拓扑结构所决定。探索蛋白质折叠速率与蛋白质拓扑结构之间的内在联系是一个很有意义的内容。通过前面的网络参量分析发现,蛋白质氨基酸网络具有正匹配性,即度值大的节点倾向于与度值大的节点相连,这种匹配性反映了天然结构中氨基酸之间相互作用的普遍特征。这里有两个问题:为什么蛋白质天然结构会选择这样的相互作用方式?这种相互作用方式是否有利于蛋白质折叠?

为了回答上面两个问题, 需要进一步探索氨基酸网络的匹配性与蛋白质折叠之间的关系. 我们搜集了有折叠速率实验数据的 23 个蛋白质, 根据式 (1), 计算了每一个蛋白质的匹配系数, 并与实验上获得的折叠速率进行了对比. 图 3a 反映了二者之间的统计关系, 由图 1 可以看出, 氨基酸网络的匹配系数与折叠速率之间存在明显的线性正相关, 即网络的匹配系数越大, 其折叠速率也越大. 二者的

关联系数为 0.600 ($P < 0.003$), 这一结果与 Bagler 等^[10]的统计结果一致. 上述结果表明, 网络的正匹配有助于蛋白质快速折叠. 蛋白质的快速折叠需要氨基酸之间具有高度的协同性, Newman 等^[11]的研究表明, 相对于负匹配网络, 正匹配网络有助于节点之间信息的快速传递. 在蛋白质折叠过程中, 这种快速的信息交流, 可以提高氨基酸之间的协同性.

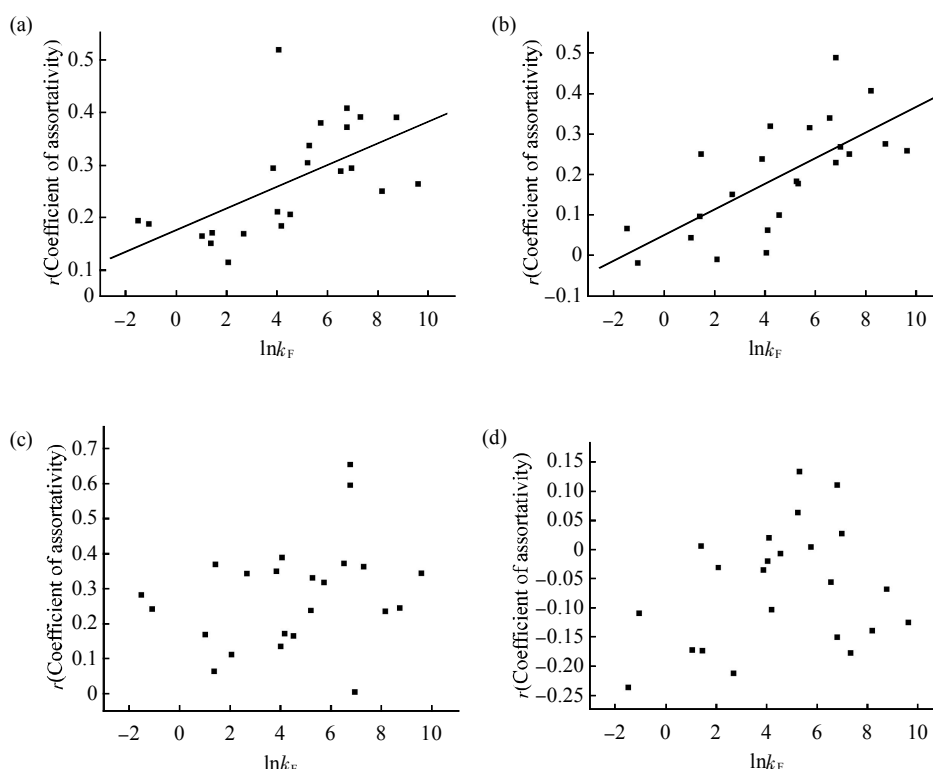


Fig. 3 Correlation between the coefficient of assortativity of the amino acid network and the experimental folding rate for 4 types of networks

(a) Amino acid networks. (b) Hydrophobic amino acid networks. (c) Hydrophilic amino acid networks. (d) Hydrophobic-Hydrophilic amino acid networks.

根据不同的物理化学性质, 可以把氨基酸分为疏水残基(C, M, F, I, L, V, W, Y)和亲水残基(A, G, T, S, N, Q, D, E, H, R, K, P). 理论和实验研究表明, 疏水相互作用是蛋白质折叠的主要驱动力. 为了更进一步考查不同类型残基之间相互作用对蛋白质折叠的影响, 我们分别统计了疏水氨基酸网络、亲水氨基酸网络和疏水-亲水氨基酸网络的匹配系数与蛋白质折叠速率之间的关系, 如图 3b~d 所示. 结果表明, 疏水氨基酸网络的匹配系数与折叠速率之间有明显的线性正相关(图 3b), 相关系数达到 0.694 ($P < 0.000\ 23$). 而亲

水网络和疏水-亲水网络的匹配系数与折叠速率之间没有明显的线性相关关系. 这一结果揭示了疏水残基之间相互作用对蛋白质折叠的重要影响, 疏水网络的匹配性与折叠速率的正相关是导致氨基酸网络匹配性与折叠速率正相关的主要原因. 疏水相互作用是蛋白质折叠的主要驱动力, 疏水网络的正匹配有助于提高疏水残基之间相互作用的协同性, 进而促使蛋白质的快速折叠. 亲水网络的匹配系数与蛋白质折叠速率之间没有线性关联, 表明其正匹配性并不能促使蛋白质的快速折叠. 疏水-亲水网络本身并不具有正匹配性, 其匹配系数对蛋白质折叠

也没有明显的关联性。

2.2.2 网络聚集系数与折叠速率之间的关联。

根据式(2)，计算了有折叠速率实验数据的 23 个蛋白质的聚集系数，并考查了聚集系数与折叠速率之间的关系(图 4)。由图 4a 可以看出，氨基酸网络的聚集系数与折叠速率之间没有明显的相关性，该结果与目前的文献报道一致。进而，我们计算了疏水氨基酸网络、亲水氨基酸网络以及疏水-亲水氨基酸网络的聚集系数，并研究了这些网络的聚集系数与折叠速率之间的关系，如图 4b~d 所示。结果显示，疏水网络的聚集系数与折叠速率之间存在明显的负相关，即聚集系数越大，折叠速率越小，

相关系数达到 $-0.645(P < 0.0009)$ 。在复杂网络中，若同时存在两个节点与一个节点相互作用，并且这两个节点之间也存在相互作用，就会形成一种三角结构(triangle construction)。聚集系数描述了与同一节点相互作用的各个节点之间相互作用的程度。聚焦系数越大，天然态中形成的三角结构就越多，上述结果说明疏水残基之间三角结构的形成可能会拖慢折叠进程的进行，降低蛋白质折叠速率。而亲水网络和疏水-亲水网络的聚集系数与折叠速率之间不存在这种负相关。上述结果进一步表明疏水网络的拓扑结构对蛋白质折叠具有重要的影响。

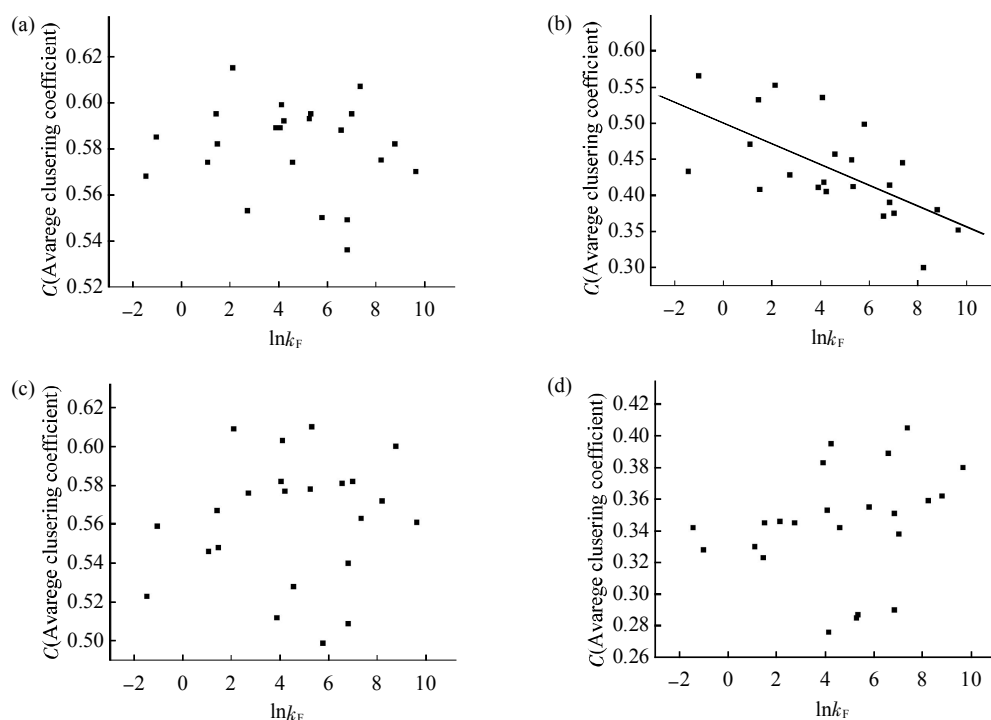


Fig. 4 Comparison of the average clustering coefficient of the amino acid network with the experimental folding rate for 4 types of networks

(a) Amino acid networks. (b) Hydrophobic amino acid networks. (c) Hydrophilic amino acid networks. (d) Hydrophobic-Hydrophilic amino acid networks.

我们进一步构建了以上 4 种类型氨基酸网络所对应的长程网络，构建方法是保持原有网络的节点数不变，只将在截断半径之内并且主链相隔 12 或更多个残基的两个节点相连接。根据式(2)计算了这 4 种长程网络的聚集系数，并探索了聚集系数与蛋白质折叠速率之间的统计关系，结果如图 5 所示。由图 5 可以看出，这 4 种长程网络的聚集系数

都与折叠速率线性负相关，相关系数分别为 $-0.636(P < 0.002)$ (长程氨基酸网络)、 $-0.674(P < 0.001)$ (长程疏水氨基酸网络)、 $-0.487(P < 0.019)$ (长程亲水氨基酸网络)、 $-0.529(P < 0.010)$ (长程疏水-亲水氨基酸网络)。长程氨基酸网络中大的聚集系数表明很多序列上相隔较远的残基聚集到一起形成相互作用，聚集系数越大，折叠速率越小，这对于快速折

叠是不利的, 所以会拖慢折叠进程^[10]. 前人的研究结果证实, 蛋白质天然结构的接触序(contact order)

与蛋白质折叠速率之间存在明显的负相关^[10], 这与我们的结果是一致的.

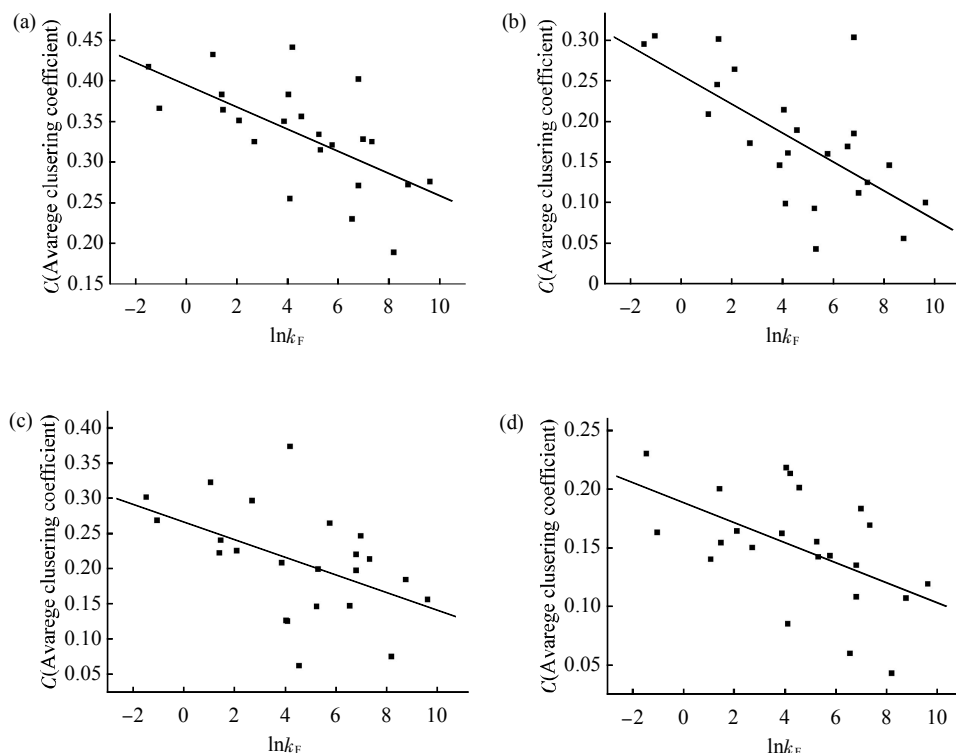


Fig. 5 Correlation between the average clustering coefficient of the long-range interaction networks and the experimental folding rate for 4 types of networks

(a) Amino acid networks. (b) Hydrophobic amino acid networks. (c) Hydrophilic amino acid networks. (d) Hydrophobic-Hydrophilic amino acid networks.

3 讨 论

本文将复杂网络的方法应用于蛋白质折叠机理的研究, 构建了氨基酸网、亲水网、疏水网、亲水-疏水网, 并计算出各网络的匹配系数和聚集系数. 进而探讨了这些拓扑参量与蛋白质折叠速率之间的关系. 研究结果表明, 除了亲水-疏水网外, 其他类型的氨基酸网络都表现出明显的正匹配性. 并且疏水网的匹配系数与折叠速率有明显的线性正相关关系, 揭示了疏水残基之间相互作用对蛋白质折叠的重要影响. 进一步, 通过对聚集系数的分析发现, 疏水网的平均聚集系数与折叠速率有明显的负相关, 表明疏水残基之间三角结构的形成不利于蛋白质快速折叠. 我们还构建了各种长程氨基酸网络, 研究发现, 长程网络的平均聚集系数与折叠速率具有明显的负相关, 说明序列上长距离残基之间接触的形成不利于快速折叠.

参 考 文 献

- [1] Wolynes P G. Recent successes of the energy landscape theory of protein folding and function. *Q Rev Biophys*, 2005, **38**(4): 405-410
- [2] Alm E, Baker D. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc Natl Acad Sci USA*, 1999, **96**(20): 11305-11310
- [3] Riddle D S, Grantcharova V R, Santiago J V, *et al.* Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct Biol*, 1999, **6**(11): 1016-1024
- [4] Baker D. A surprising simplicity to protein folding. *Nature*, 2000, **405**(6782): 39-42
- [5] Vendruscolo M, Dokholyan N V, Paci E, *et al.* Small-world view of the amino acids that play a key role in protein folding. *Phys Rev E*, 2002, **65**(6): 061910.1-061910.4
- [6] Li J, Wang J, Wang W. Identifying folding nucleus based on residue contact networks of proteins. *Proteins*, 2008, **71**(4): 1899-1907
- [7] Dokholyan N V, Li L, Ding F, *et al.* Topological determinants of protein folding. *Proc Natl Acad Sci USA*, 2002, **99**(13): 8637-8641
- [8] Greene L H, Higman V A. Uncovering network systems within

- protein structures. *J Mol Biol*, 2003, **334**(4): 781–791
- [9] Kundu S. Amino acids network within protein. *Physica A*, 2005, **346**(1–2): 104–109
- [10] Bagler G, Sinha S. Assortative mixing in Protein Contact Networks and protein folding kinetics. *Bioinformatics*, 2007, **23**(14): 1760–1767
- [11] Newman M E J. Assortative mixing in networks. *Phys Rev Lett*, 2002, **89**(20): 208701–208705
- [12] 吴金闪, 狄增如. 从统计物理学看复杂网络研究. *物理学进展*, 2004, **24**(1): 18–46
- Wu J S, Di Z R. *Progress in Physics*, 2004, **24**(1): 18–46

Study on The Characters of Different Types of Amino-acid Networks and Their Relations With Protein Folding*

YAN Li-Cheng^{1)**}, SU Ji-Guo^{1,2)**}, CHEN Wei-Zu¹⁾, WANG Cun-Xin^{1)***}

⁽¹⁾ College of Life Science and Bioengineering, Beijing University of Technology, Beijing 100124, China;

⁽²⁾ College of Science, Yanshan University, Qinhuangdao 066004, China)

Abstract The theoretical and experimental studies have showed that the topology of the native structure of protein plays an important role in determining its folding process. The complex network approach was used to analyze the topological characters of the native structure of protein and then explore the relationship between these characters and the experimental folding rates. Several types of network were constructed, including all-amino-acids network, hydrophobic amino acid network, hydrophilic amino acid, hydrophobic-hydrophilic amino acid network and their corresponding long-range interaction network. The statistic characters of the assortativity coefficient and the clustering coefficient were studied. The results indicate that all types of network except for the hydrophobic-hydrophilic one are of the positive assortativity coefficient. Furthermore, there is an obvious linear positive correlation between the assortativity coefficient and the folding rate for the all amino acids network and the hydrophobic amino acid network, which implies that the cooperative interactions of the hydrophobic amino acids are important for proteins rapidly folding into their native states. Moreover, it is found that there is a clear linear negative correlation between the clustering coefficient of the hydrophobic amino acid networks and the experimental folding rates of the corresponding proteins, which indicates that the formation of the triangle construction among the amino acids reduces the folding rates. It is also found that in the long-range interaction networks, the formation of contact pairs linking two distant residues in sequence would slow down the process of protein folding.

Key words amino acid network, hydrophobic interaction, rate of folding, assortativity coefficient, clustering coefficient

DOI: 10.3724/SP.J.1206.2009.00760

*This work was supported by grants from The National Natural Science Foundation of China (10974008, 10947014) and The Research Fund for the Doctoral Program of Higher Education of China (200800050003).

**These authors contributed equally to this work.

***Corresponding author.

Tel: 86-10-67392724, E-mail: cxwangbjut@gmail.com

Received: December 23, 2009 Accepted: March 23, 2010