

蛋白质氨基酸网络研究进展*

常 珊^{1△} 焦 雄² 王美华¹ 田绪红¹

(1 华南农业大学信息学院 广东 广州 510642 2 太原理工大学应用力学与生物医学工程研究所 山西 太原 030024)

摘要 氨基酸网络是运用复杂网络工具对蛋白质结构-功能关系研究的新方法。本文回顾了氨基酸网络中常用网络参量的计算方法,如 度分布,聚集系数,平均最短路径等。结合本研究小组的工作,介绍了常用的网络构建和分析方法,并总结了氨基酸网络在蛋白质折叠以及蛋白质分子对接问题中的应用。最后,分析了氨基酸网络研究目前存在的主要问题,并对未来的工作进行了展望。

关键词 蛋白质;复杂网络;氨基酸网络

中图分类号 Q591, Q615 **文献标识码** A **文章编号** 1673-6273 (2011) 01-190-04

Progress in Amino Acid Networks of Proteins *

CHANG Shan^{1△}, JIAO Xiong², WANG Mei-hua¹, TIAN Xu-hong¹

(1 College of Informatics, South China Agricultural University, Guangzhou, Guangdong 510642, China ;

2 Institute of Applied Mechanics and Biomedical Engineering, Taiyuan University of technology, Taiyuan, Shanxi 030024, China)

ABSTRACT: Amino acid network is a new approach on the relationship between protein structure and function by using the technology of complex network. In this article, we reviewed the computational methods of mainly network parameters in amino acid network, such as the degree distribution, the clustering coefficient and the average minimal distance. Combining the research groups works, we introduced the commonly used constructing and analyzing methods of amino acid network. Meanwhile, we concluded the application on protein folding and protein docking. Finally, the existing major questions of amino acid network were analyzed and the prospects of the future study were made.

Key words: protein; complex network; amino acid network

Chinese Library Classification: Q591, Q615 **Document Code:** A

Article ID: 1673-6273(2011)01-190-04

前言

随着人类基因组测序计划的完成,生命科学的重心开始转移到对基因的表达产物蛋白质的研究上来,蛋白质组学成为后基因组时代的研究前沿和热点领域。在后基因组时代,生物科学的中心任务是阐明基因组所表达的真正执行生命活动的全部蛋白质的表达规律和生物功能。蛋白质结构与功能之间的关系是后基因组时代研究的核心内容。一方面,由于基因组测序、蛋白质组学的快速发展,生物学积累了大量的数据,如何挖掘出大量实验数据所蕴含的生物基本规律成为生命科学研究关注的焦点;另一方面,研究生物学系统的方法从局部的定性描述向整体和系统的定量描述转变。这些进展使生物学家逐步地意识到生命科学研究需要数学、物理学、化学、信息科学等定量学科的参与。

近年来,复杂网络的研究迅速成为一个研究热点,同时相关的基础和应用研究已经渗透到物理学、生物学、计算机科学等许多学科之中,并具有重要的理论与实际应用价值^[1,2]。复杂网络能够有效地描述复杂系统的拓扑结构和整体性质,为挖掘生物学数据提供了很好的工具。复杂网络在生物学方面的应用

非常广泛,如蛋白质相互作用网络,代谢网络等等。本文将针对蛋白质氨基酸网络研究进行综述,氨基酸网络可以直接分析蛋白质的结构特征,对蛋白质结构-功能关系的研究有着重要的价值。

1 复杂网络的基本概念

网络不依赖于节点的具体位置和边的具体形态所表现出来的性质叫做网络的拓扑性质,相应的结构叫做网络的拓扑结构。在最初对网络的拓扑结构研究时,以一些规则的结构来近似代替真实系统中各因素之间的关系,从而系统被简化为规则网络,并以二维平面上规则网格来表示。随着研究的推进,提出了一种新的构造网络的方法,即以一定的概率确定两个节点之间是否连边,由此生成的网络被称为随机网络^[3,4]。此后的研究中,在描述真实系统时,随机网络得到广泛的应用。最近几年,由于计算能力的提高,通过大量的计算与研究,发现许多真实网络既不是规则网络,也不是随机网络,而是具有与两者皆不同的统计特征,将具有这些与规则网络或随机网络不同统计特征的网络命名为复杂网络 (Complex Networks)^[5]。

1.1 网络的表示

* 基金项目 国家自然科学基金 (31070829), 广东省科技计划项目 (2010B080701070), 华南农业大学校长基金 (5600-k09325)

作者简介 常珊 (1982), 男, 博士, 讲师, 主要研究方向: 计算机辅助药物设计, 分子模拟与生物信息学

△通讯作者 常珊, 电话 020-85280320, E-mail: schang@scau.edu.cn

(收稿日期 2010-10-06 接受日期 2010-10-30)

用图论的语言来描述,一个网络可以抽象为一个由节点的集合和边集组成的图 $G=(V,E)$ 。节点数记为 $N=|V|$,边数记为 $M=|E|$ 。E 中每条边都有中一对点与之相对应。如果图中的任意节点对 (i,j) 和 (j,i) 对应同一条边,即网络中的边没有固有的方向,这种网络称为无向网络,否则称为有向网络。网络中的边还可以赋予各种权重,这种网络称为加权网络。反之,如果把网络中所有的边都同等看待,则称为无权网络。

1.2 度分布

节点的度定义为与该节点直接连接的其它节点的数目,和节点连接的这些节点称为的邻居。网络中所有节点度的平均值称为网络的平均度,即

$$\langle k \rangle = \frac{1}{N} \sum_i k_i \quad (1)$$

网络中节点度的分布情况可用分布函数 $P(k)$ 来描述,它表示随机选取一个节点的度刚好是 k 的几率。

1.3 聚集系数

如果一个网络节点有几个邻居,那么这几个邻居之间可能彼此也有边相连,聚集系数可以用来描述网络连接的这种聚集性质^[9]。网络中节点 i 的所有邻居所构成的子图称为节点的邻域 Γ_i , $|E(\Gamma_i)|$ 是 i 的邻域中边的数目。则节点 i 的聚集系数定义为

$$C_i = \frac{|E(\Gamma_i)|}{\binom{k_i}{2}} \quad (2)$$

其中 k_i 为节点 i 的度, Γ_i 中可能存在的最大的边数是

$$\binom{k_i}{2} = \frac{1}{2} k_i (k_i - 1) \quad (3)$$

1.4 平均路径长度

网络中两个节点 i 和 j 之间的距离 d_{ij} 定义为连接这两个节点的最短路径上的边数。网络中任意两个节点之间距离的最大值称为网络的直径。网络的平均路径长度 定义为任意两个节点之间距离的平均值,即

$$L = \frac{1}{\frac{1}{2} N(N-1)} \sum_{i>j} d_{ij} \quad (4)$$

其中 N 为网络节点数。网络的平均路径长度也称为网络的特征路径长度。然而,式(4)的定义不适用于网络非全连通的情形,即网络中存在两个节点之间没有任何路径连接的情况。原因是,如果把两个无路径连接的节点间的距离定义为无穷大的话,将导致 L 也变为无穷大。为了解决这个问题,一种方法是将无路径连接的节点间的距离定义为网络最大的那个最短路径长度。另一种方法是采用调和平均来定义网络的平均路径长度,即

$$L^{-1} = \frac{1}{\frac{1}{2} N(N-1)} \sum_{i>j} d_{ij}^{-1} \quad (5)$$

这种定义自然消除了无穷大的问题。

1.5 其它参数

表征网络拓扑结构性质的参数还有很多,如介数^[7,8]、同配

性系数^[9]、分形维数^[10]等等。网络中通过某个节点或某条边的最短路径的数目称为这个节点或这条边的介数。它表征了一个节点或一条边在网络交通中的重要性程度。同配性系数则刻画相互连接的节点度大小的关联性,如果网络中度大的节点倾向于与度大的节点连接,则同配性系数大于 0,如果度大的节点倾向于连接到度小的节点上,则同配性系数小于 0。分形维数是表征网络分形或自相似性质的参数。

2 蛋白质氨基酸网络(残基网络)的研究

蛋白质分子是由氨基酸组成且通过氨基酸之间各种相互作用共同维系的复杂系统,因此,蛋白质分子可视为由氨基酸残基及其相互作用构成的复杂网络。建立蛋白质的氨基酸网络时,通常将残基简化为一个点作为网络节点,并依据节点之间的距离判定节点间的连接情况。通常认为氨基酸网络具有小世界性,即相比于随机网络具有大的聚集系数和小的平均最短路径。

2.1 网络构建与分析

目前常用的建网方法都是以氨基酸残基作为节点,不同在于有的以原子截断半径建立连接^[11,12],而有的则采用原子间接触作为截断半径建立连接^[13,14]。除了连接的不同,网络往往还依据不同的氨基酸类型进行分类。不同类型的氨基酸网络往往能够分析得到特别的结论。Kundu 等人将氨基酸分成疏水、亲水、带电的不同类型,分析了不同氨基酸网络的静态参量特征^[15,16]。网络加权也是一种特殊的分析方法,能量以及连接的原子数量等都可以作为权重定义的方式^[17]。国内也有小组进行了类似的研究,严立成等人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和聚集系数的统计特性^[18]。

我们也根据氨基酸的特征建立并研究了蛋白质保守氨基酸的无向网络^[19]。蛋白质结构被视为一个无向网络,其中保守性氨基酸残基为网络节点,残基之间的接触视为连接。我们的研究发现,保守性氨基酸网络具有介于规则网络和随机网络之间的聚集系数和特征路径长度,属于小世界类型的网络。蛋白质复合物界面的残基比表面残基通常具有大的度值和低的聚集系数。此外,还发现了保守残基的空间聚集是一个普遍现象。保守性氨基酸网络的性质将能够给蛋白质-蛋白质界面预测提供一些新的参量。

2.2 蛋白质折叠研究

蛋白质如何由氨基酸序列折叠成三维结构,是一个仍未解决的问题,称为“蛋白质折叠”^[20]。蛋白质折叠是一个涉及分子生物学、物理学、化学的交叉研究领域^[21-24]。对单个蛋白质的网络结构分析往往能够揭示折叠的一些信息^[25],如 Vendruscolo 和 Karplus 等人^[12]通过对 PDB 数据库中 978 个具有代表性的蛋白质进行结构分析,得到氨基酸之间的成对几率,以该几率的倒数作为对应氨基酸之间的距离。对一个特定的蛋白质结构,通过截断半径 (0.85 nm) 判定残基之间是否接触,并对边赋予该对残基之间的权重,通过研究蛋白质网络的拓扑性质,计算网络特征量介数来寻找蛋白质折叠中的关键残基 (Key Residues) Dokholyan 等人^[26]通过研究 C12 和 C- SrcSH3 两个小

蛋白的折叠过渡态的氨基酸网络性质,来揭示蛋白质结构的拓扑特性对蛋白质折叠动力学的影响;Atilgan 等人^[11]研究了氨基酸残基位置变化对蛋白质网络性质的影响,从而揭示了蛋白质结构和动力学之间的联系;汪炜等人的采用蛋白质天然结构中全局近邻中心节点 (globally full-closeness centrals) 来预测蛋白质的折叠核残基^[27]。

我们采用加权网络的方法研究了蛋白质折叠过程中的参量变化和折叠机制^[28]。通过对 197 个非同源蛋白质的加权与非加权氨基酸网进行统计,比较分析了不同加权方式的网络参量,以及权重对网络参量的影响,并对比不同残基类型的网络参量以及亲疏水残基之间的连接情况,证实了加权网较非加权网包含了更多的蛋白质结构信息。同时,分析了加权网的网络特征量在胰凝乳蛋白酶抑制剂 2 (Chymotrypsin Inhibitor 2, CI2) 高温去折叠路径上的变化,发现结构解聚主要表现为疏水核心的破坏;网络集聚系数对二级结构的变化不敏感;平均最短路径随着结构的解聚而增大;折叠核心的介数值为局部极大,且加权网介数对折叠核的区分能力强于非加权网。

值得一提的是,我们还根据蛋白质折叠的机理建立了氨基酸网络的演化模型,该模型能够揭示氨基酸网络的形成原因^[29]。由于蛋白质的氨基酸网络是在蛋白质折叠过程中形成的,通常的网络模型难以解释其演化的机制。基于蛋白质折叠的观点,我们提出了一个氨基酸网络的演化模型。在此模型中,演化从天然蛋白质的氨基酸序列开始,由两个基本假设进行引导,即近邻偏好性规则和能量偏好性规则。研究发现近邻偏好性规则主要决定通常的网络性质,而能量偏好性规则主要决定特殊的生物学结构特征。应用于天然蛋白质体系发现,该模型能够很好地模拟出氨基酸网络的性质。

2.3 蛋白质分子对接研究

蛋白质分子对接方法是研究蛋白质间相互作用与识别的理论手段,对细胞中蛋白质的分子生物学机理研究,计算机辅助药物设计和复合物结构预测具有重要的意义^[30-32]。蛋白质分子对接模拟是已知两个单体蛋白质分子的三维结构,通过计算机模拟来预测其复合物的结合模式和三维结构。因此,蛋白质复合物的氨基酸网络同样对蛋白质-配体相互作用有着重要的方法学意义,Amitai 等人^[33]通过研究蛋白质网络的特征量 closeness 来预测蛋白质活性位点;Jacobs 等人^[34]通过蛋白质网络来识别蛋白质柔性区和刚性区。del Sol 等人^[35-37]采用介数定义蛋白质-蛋白质相互作用的热点残基,Daily 等人^[38]分析了蛋白质-蛋白质结合过程中构象变化的机理。

我们根据氨基酸类型将蛋白质复合物的氨基酸网络分成两种类型,即疏水和亲水氨基酸网络。分析这两种不同类型的网络,发现他们都具有小世界的性质。通过分析网络参量发现,正确结合的复合物构象比错误结合的结构具有更高的界面度值和更低的网络特征路径长度。这些性质反映出正确结合的复合物结构有更好的几何和残基类型互补,同时正确的结合模式对于保证天然蛋白质复合物的特征路径长度起着重要作用。此外,建立了两个基于网络参量的打分项,它们能够很好地反映复合物整体形状和残基类型互补特性。将基于网络的打分项与其他打分函数项进行组合后,提出了一个新的多项打分函数

HPNCscore,它能够与 RosettaDock 组合打分函数的区分能力提高 12%^[39,40]。另外,基于加权氨基酸网,我们还构建了新的对接打分函数 S_n ^[41]。采用该打分函数对 42 个体系的 bound 与 unbound 对接采样结果进行打分排序,并与其他打分函数 (RP) 结果进行比较,发现基于氨基酸网络的打分函数 S_n 在一些指标上优于 RP 打分函数。通过分析发现, S_n 打分函数更加适合于考虑蛋白质柔性的对接采样。

3 存在的问题和展望

氨基酸网络是蛋白质结构-功能关系研究的新方法,目前仍然有很多问题没有解决^[42]。如:(1) 氨基酸网络的形成机制仍然不清楚。当前的网络演化模型无法对蛋白质的网络结构进行有效的解释,分析蛋白质结构网络的动力学和演化机制将帮助理解蛋白质结构的形成机理。(2) 系统的比较和分析合适的连接权重(变换距离阈值和非加权连接)以及网络建立规则是将来的任务之一。后续可以引入非成对相互作用^[43],也许能够更细致的分析蛋白质结构的细节。加权网络和向有连接的引入,以及系统的模块分析构象网络也许能够解决长期困扰的能量漏斗理论。(3) 更精确的对蛋白质结构进行模块分析,可能在定义蛋白质酶活性关键残基,变构调节,信号转导和蛋白质稳定性中发挥作用。网络的中心化度量应该考虑权重,并在拓扑结构的多层面上展开,可以定义分级的蛋白质关键残基(模块中心,模块间桥联结构和多个重叠区域的元素)。相信将来氨基酸网络的研究会取得更大的发展,成为蛋白质结构-功能关系研究的重要方向。

参考文献(References)

- [1] Watts DJ. The 'new' science of networks [J]. Annu Rev Sociol, 2004, 30: 243-270
- [2] Strogatz SH. Exploring complex networks [J]. Nature, 2001, 410 (6825): 268-276
- [3] Karonski M. A review of random graphs [J]. J Graph Theory, 1982, 6 (4): 349-389
- [4] Erdős PR, Rényi A. On the evolution of random graphs [J]. Publ Math Inst Hung Acad Sci, 1960, 5: 17-61
- [5] Boccaletti S, Latora V, Moreno Y, et al. Complex networks: Structure and dynamics [J]. Phys Rep, 2006, 424(4-5): 175-308
- [6] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks [J]. Nature, 1998, 393(6684): 440-442
- [7] Newman MEJ. Mixing patterns in networks [J]. Phys Rev E, 2003, 67 (2): 026126
- [8] Freeman LC. Set of measures of centrality based on betweenness [J]. Sociometry, 1977, 40(1): 35-41
- [9] Newman MEJ. Assortative mixing in networks [J]. Phys Rev Lett, 2002, 89(20): 208701
- [10] Song CM, Havlin S, Makse HA. Self-similarity of complex networks [J]. Nature, 2005, 433(7024): 392-395
- [11] Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics [J]. Biophys J, 2004, 86(1): 85-91
- [12] Vendruscolo M, Dokholyan NV, Paci E, et al. Small-world view of the amino acids that play a key role in protein folding [J]. Phys Rev E,

- 2002, 65(6): 061910
- [13] Brinda KV, Vishveshwara S. A network representation of protein structures: Implications for protein stability [J]. Biophys J, 2005, 89(6): 4159-4170
- [14] Greene LH, Higman VA. Uncovering network systems within protein structures [J]. J Mol Biol, 2003, 334(4): 781-791
- [15] Aftabuddin M, Kundu S. Hydrophobic, hydrophilic, and charged amino acid networks within protein [J]. Biophys J, 2007, 93(1): 225-231
- [16] Kundu S. Amino acid network within protein [J]. Physica A, 2005, 346(1-2): 104-109
- [17] Aftabuddin M, Kundu S. Weighted and unweighted network of amino acids within protein [J]. Physica A, 2006, 369(2): 895-904
- [18] 严立成, 苏计国, 陈慰祖, 等. 不同类型氨基酸网络参量与蛋白质折叠的关系 [J]. 生物化学与生物物理进展, 2010, 37(7): 762~768
Yan Li-Cheng, Su Ji-Guo, Chen Wei-Zu, et al. Study on the characters of different types of amino-acid networks and their relations with protein folding [J]. Prog Biochem Biophys, 2010, 37(7): 762~768 (In Chinese)
- [19] Chang S, Li CH, Gong XQ, et al. An analysis of protein conservation residues network [C]. ICBBE., 2007: 29-32
- [20] Lazaridis T, Karplus M. "New view" of protein folding reconciled with the old through multiple unfolding simulations [J]. Science, 1997, 278(5345): 1928-1931
- [21] Zhang Y. Protein structure prediction: when is it useful? [J]. Curr Opin Struct Biol, 2009, 19(2): 145-155
- [22] Zhang J, Li WF, Wang J, et al. Protein folding simulations: From coarse-grained model to all-atom model [J]. IUBMB Life, 2009, 61(6): 627-643
- [23] Chen YW, Ding F, Nie HF, et al. Protein folding: Then and now [J]. Arch Biochem Biophys, 2008, 469(1): 4-19
- [24] 王会峰. 分子伴侣及其在蛋白质折叠中的作用研究进展 [J]. 现代生物医学进展, 2009, 9(4): 746-748
Wang Hui-feng. Progress on molecular chaperone and its role in protein folding [J]. Progress in Modern Biomedicine 2009, 9(4): 746-748 (In Chinese)
- [25] 李海彦. 复杂网络在蛋白质折叠中的应用 [J]. 德州学院学报, 2008, 24(2): 1-4
Li Hai-yan. Application Of complex network in protein research [J]. J Dezhou Univ, 2008, 24(2): 1-4 (In Chinese)
- [26] Dokholyan NV, Li L, Ding F, et al. Topological determinants of protein folding [J]. Proc Natl Acad Sci U S A, 2002, 99(13): 8637-8641
- [27] Li J, Wang J, Wang W. Identifying folding nucleus based on residue contact networks of proteins [J]. Proteins, 2008, 71(4): 1899-1907
- [28] Jiao X, Chang S, Li CH, et al. Construction and application of the weighted amino acid network based on energy [J]. Phys Rev E, 2007, 75(5): 051903
- [29] Chang S, Jiao X, Gong XQ, et al. Evolving model of amino acid networks [J]. Phys Rev E, 2008, 77(6): 061920
- [30] Moreira IS, Fernandes PA, Ramos MJ. Protein-Protein Docking Dealing With the Unknown [J]. J Comput Chem, 2010, 31(2): 317-342
- [31] Ritchie DW. Recent progress and future directions in protein-protein docking [J]. Curr Protein Pept Sci, 2008, 9(1): 1-15
- [32] Gray JJ. High-resolution protein-protein docking [J]. Curr Opin Struct Biol, 2006, 16(2): 183-193
- [33] Amitai G, Shemesh A, Sitbon E, et al. Network analysis of protein structures identifies functional residues [J]. J Mol Biol, 2004, 344(4): 1135-1146
- [34] Jacobs DJ, Rader AJ, Kuhn LA, et al. Protein flexibility predictions using graph theory [J]. Proteins, 2001, 44(2): 150-165
- [35] del Sol A, Fujihashi H, Amoros D, et al. Residues crucial for maintaining short paths in network communication mediate signaling in proteins [J]. Mol Syst Biol, 2006, 2: 0019
- [36] del Sol AO'Meara P. Small-world network approach to identify key residues in protein-protein interaction [J]. Proteins, 2005, 58(3): 672-682
- [37] del Sol A, Fujihashi H, O'Meara P. Topology of small-world networks of protein-protein complex structures [J]. Bioinformatics, 2005, 21(8): 1311-1315
- [38] Daily MD, Upadhyaya TJ, Gray JJ. Contact rearrangements form coupled networks from local motions in allosteric proteins [J]. Proteins, 2008, 71(1): 455-466
- [39] Chang S, Jiao X, Li CH, et al. Amino acid network and its scoring application in protein-protein docking [J]. Biophys Chem, 2008, 134(3): 111-118
- [40] Chang S, Gong XQ, Jiao X, et al. Network analysis of protein-protein interaction [J]. Chin Sci Bull, 2010, 55(9): 814-822
- [41] 焦雄, 常珊, 安美文, 等. 加权氨基酸网络及其应用 [J]. 太原理工大学学报, 2009, 40(3): 323-325
Jiao Xiong, Chang Shan, An Mei-wen, et al. Weighted amino acid network model and its application [J]. J Taiyuan Univ Tech, 2009, 40(3): 323-325 (In Chinese)
- [42] Bode C, Kovacs IA, Szalay MS, et al. Network analysis of protein dynamics [J]. FEBS Lett, 2007, 581(15): 2776-2782
- [43] Finkelstein AV, Roytberg MA. Computation of biopolymers: A general approach to different problems [J]. Biosystems, 1993, 30(1-3): 1-19