Journal of Theoretical Biology 355 (2014) 95–104

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Amino acid contact energy networks impact protein structure and evolution

Wenying Yan [a,1], Maomin Sun [a,b,1], Guang Hu [a], Jianhong Zhou [a], Wenyu Zhang [a], Jiajia Chen [a,c], Biao Chen [a], Bairong Shen [a,*]

[a] Center for Systems Biology, Soochow University, No. 1, Shizi Street, Suzhou, Jiangsu 215006, China
[b] Laboratory Animal Research Center, School of Medical, Soochow University, China
[c] Department of Chemistry and Biological Engineering, Suzhou University of Science and Technology, Jiangsu, Suzhou 215011, China

HIGHLIGHTS

- Protein structures are modeled as amino acid contact energy networks.
- Topology of an amino acid network is found correlated with its protein secondary structure density.
- Diameters of the amino acid networks show a negative correlation with evolutionary rates.

ARTICLE INFO

Article history:
Received 8 February 2014
Accepted 21 March 2014
Available online 1 April 2014

Keywords:
Clustering coefficient
Long-range link percentage
Network diameter
Protein evolutionary rate
Protein secondary structure density

ABSTRACT

One of the most challenging tasks in structural proteomics is to understand the relationship between protein structure, biological function, and evolution. An understanding of amino acid networks based on protein topology has an important role in the study of this relationship; however, the relationship between network parameters underlying protein topology with structural properties or evolutionary rate is still unknown. To investigate this further, we modeled the three dimensional structure of proteins as amino acid contact energy networks (AACENs) with nodes represented as amino acid residues and edges established according to environment-dependent residue–residue contact energies. Five other types of networks were also constructed to investigate their topological parameters and compare their effect on protein structure and evolution: (1) a random contact network (RCN), (2) a rewiring network with the same degree of distribution as AACEN (RNDD), (3) long-range contact energy networks with and without the backbone connectivity (LCEN_BBs and LCENs), and (4) short range contact energy networks (SCENs). The results indicated that the long-range link percentage and the network clustering coefficient showed a significantly positive and negative correlation, respectively, with protein secondary structure density. In addition, the long-range link percentage and network diameter had a significantly positive and negative correlation, respectively, with evolutionary rate. According to our knowledge, this is the first study to identify the potential role of long-range links and network diameter in protein evolution.

1. Introduction

Identifying the reasons underlying variation in protein evolutionary rate is an important objective in many fields, including molecular evolution, comparative genomics, and structural proteomics (Pal et al., 2006). The rate of protein evolution arguably provides one of the most powerful tools for quantification of the relative importance of selection and genetic drift, and for the identification of selective forces from genomic data. Moreover, studies on protein evolutionary rate aid in the identification of functionally important sites, which can be used to predict how mutations might contribute to disease (Sauna and Kimchi-Sarfaty, 2011). Systematic surveys have indicated numerous factors that correlate with evolutionary rate, including protein length (Ingvarsson, 2007), protein dispensability (Wall et al., 2005), number of a protein interaction partners (Fraser et al., 2002), protein function, level of gene expression (Pagan et al., 2012;

* Corresponding author. Tel./fax: +86 512 651 109 51.
E-mail address: bairong.shen@suda.edu.cn (B. Shen).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jtbi.2014.03.032
0022-5193/

Wall et al., 2005), and even protein position in a pathway (Ramsay et al., 2009).

In addition, a number of studies have addressed the relationship between a protein's native structure and evolutionary rate. Protein folding or topologies, such as contact density (i.e., the average number of contacts per residue (England and Shakhnovich, 2003)), secondary structure, and solvent exposure, play important roles in the evolutionary rate of a protein's corresponding genetic sequence (Bloom et al., 2006; Bustamante et al., 2000; Marsh and Griffiths, 2005; Rorick and Wagner, 2011; Zhou et al., 2008).

Graph-theoretical models of protein structure can be constructed to characterize their topological properties, study their possible role in protein secondary structure, and examine the molecular evolution of proteins. The protein structure can be modeled as an undirected network made of its constituent amino acids and their interactions, termed an amino acid network (AAN). Studying proteins from this network perspective permits topological investigations and captures the global connectivity in a protein molecule (Hu et al., 2013, 2014; Zhou et al., 2014a, 2014b). Additionally, it permits investigation into the role of each individual amino acid within the complex interacting network (del Sol and O'Meara, 2005; Vendruscolo et al., 2002).

The amino acid network, based on distance between atoms, is an abstract description of protein structure that only considers amino acid contacts on a geometric level and not on the chemical properties of the protein. Therefore, another approach for the simulation of residue interactions in a protein has been proposed that uses the energy between amino acid residues and assigns energy as the weight of the edges. In one method, the energy is composed of two separate energy terms: the electrostatic interaction energy (Coulomb potential) and the van der Waals interaction energy (Lennard–Jones potential). Veloso et al. (2007) constructed the networks of 12 myoglobin proteins with atoms as the nodes and the energy between the atoms as edges. Vijayabaskar and Vishveshwara (2010) also modeled the protein structure as AANs by summing the two energy compositions at the residue level, averaged over the equilibrium ensemble.

Another method for estimating the interactions between amino acid residues according to the contact energy between two residues was suggested by Miyazawa and Jernigan (Jiao et al., 2007). This method was more efficient and easier for characterizing the energy between residues.

Giuliani et al. (2009) and Krishnan et al. (2008) summarized the applications of network models in protein science, especially amino acid networks based on alpha carbons. More recently, they suggested that the amino acid network may provide a novel paradigm for the high throughput investigation of structural proteomics (Di Paola et al., 2012). In our previous work, we also reviewed the recent advances of the network theory for exploring the topology and dynamics of protein–ligand and protein–nucleic acid complexes (Hu et al., 2013; Yan et al., 2014).

In the present study, we propose another method for the construction of amino acid contact networks (AACEN) based on a coarse-grained contact energy termed environment-dependent residues contact energy (ERCE) (Shen and Vihinen, 2003; Yang et al., 2013; Zhang and Kim, 2000; Zhou et al., 2014a). Differing from the normal contact energy, ERCE also takes into account the type of secondary structure for each residue in proteins. Hence, there are 60 types of residues (20 amino acids $\times$ 3 secondary structure states). This method outperformed many predictions based on residue contact energy.

In AACENs, each node stands for a residue unit in the protein and the edge between two nodes shows their contact energies (mainly reflecting hydrophobic interactions). To explore the role of long-range contact in AACENs, the long-range contact energy network with and without residue connectivity on the backbone (LCEN_BB and LCEN), and the short range contact energy network (SCEN) of each protein were also considered.

The effectiveness of AACEN was evaluated in our recent work and used to discriminate native structures from decoys (Zhou et al., 2014a). In this work, AACENs were further characterized. The topological characteristics of AACENs were analyzed and compared to random networks and rewiring networks with the same degree distribution. In addition, the influence of long-range links on the amino acid contact energy network was detected by comparing network parameters in the AACEN, LCEN_BB, LCEN, and SCEN. Furthermore, the AACENs were used to determine whether the network parameters uncover features about protein secondary structure and the relationship of protein structural properties with evolutionary rate.

2. Results

2.1. Degree and degree distribution

The relationship between average degree and node number in a network (protein size) has been extensively studied. Some researchers have suggested that the average degree of a protein structure network is independent of protein size (Aftabuddin and Kundu, 2006; Susan, 2011), while other researchers have found that the average degree increases at a slower rate as protein size increases (Greene and Higman, 2003). To detect whether the average degree of our amino acid contact energy network was dependent on protein size, Pearson's correlation test was performed between them in AACENs, SCENs, LCEN_BBs, and

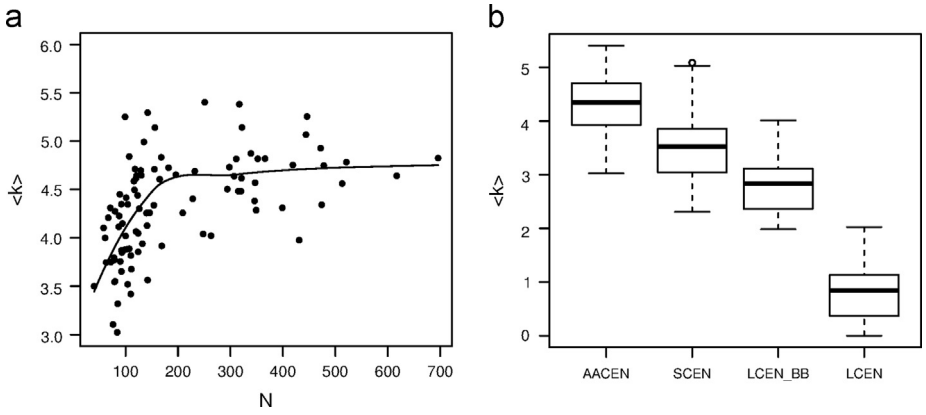

Fig. 1. Average degree distributions. (a) Average degree plotted versus node number in AACEN. The line is a smooth curve fitted by Loess. (b) Boxplots of the distributions of average degree for different types of contact energy network.