Physica A 441 (2016) 199–214

Contents lists available at ScienceDirect

Physica A

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

Analysis of heat kernel highlights the strongly modular and heat-preserving structure of proteins

Lorenzo Livi [a,*], Enrico Maiorino [b], Andrea Pinna [c], Alireza Sadeghian [a], Antonello Rizzi [b], Alessandro Giuliani [d]

[a] Department of Computer Science, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada

[b] Department of Information Engineering, Electronics, and Telecommunications, SAPIENZA University of Rome, Via Eudossiana 18, 00184 Rome, Italy

[c] CRS4 Bioinformatica, Parco Tecnologico della Sardegna, Edificio 1, 09010 Pula (CA), Italy

[d] Department of Environment and Health, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy

HIGHLIGHTS

- We study the structure and dynamical properties of protein contact networks.
- Results are compared with respect to other biological networks and various models.
- Networks are represented as numerical vectors according to two methodologies.
- Protein contact networks cannot be assimilated to any of the considered models.
- Heat trace decay shows subdiffusion, a peculiar property of protein molecules.

ARTICLE INFO

Article history:
Received 17 March 2015
Received in revised form 11 May 2015
Available online 9 September 2015

Keywords:
Protein contact networks
Heat kernel
Subdiffusion
Network analysis

ABSTRACT

In this paper, we study the structure and dynamical properties of protein contact networks with respect to other biological networks, together with simulated archetypal models acting as probes. We consider both classical topological descriptors, such as modularity and statistics of the shortest paths, and different interpretations in terms of diffusion provided by the discrete heat kernel, which is elaborated from the normalized graph Laplacians. A principal component analysis shows high discrimination among the network types, by considering both the topological and heat kernel based vector characterizations. Furthermore, a canonical correlation analysis demonstrates the strong agreement among those two characterizations, providing thus an important justification in terms of interpretability for the heat kernel. Finally, and most importantly, the focused analysis of the heat kernel provides a way to yield insights on the fact that proteins have to satisfy specific structural design constraints that the other considered networks do not need to obey. Notably, the heat trace decay of an ensemble of varying-size proteins denotes subdiffusion, a peculiar property of proteins.

* Corresponding author.
E-mail addresses: llivi@scs.ryerson.ca (L. Livi), enrico.maiorino@uniroma1.it (E. Maiorino), andreapinna@gmail.com (A. Pinna), asadeghi@ryerson.ca (A. Sadeghian), antonello.rizzi@uniroma1.it (A. Rizzi), alessandro.giuliani@iss.it (A. Giuliani).
URLs: https://sites.google.com/site/lorenzlivi/ (L. Livi), http://infocom.uniroma1.it/~rizzi/ (A. Rizzi).

http://dx.doi.org/10.1016/j.physa.2015.08.059
0378-4371/

1. Introduction

As aptly pointed out by Nicosia et al. [1] "Networks are the fabric of complex systems". Concepts like organized complexity (or middle way [2], mesoscopic systems [3]) revolve around the existence of features shared by systems made of many interacting parts, which can be suitably described in terms of networks. In 1948 Weaver [4] defined the notions of simplicity and complexity in Science by a three categories classification: (1) *Organized Simplicity*. The paradigm of this kind of problems is exemplified by Newton's law of universal gravitation; (2) *Disorganized Complexity*. In these cases, systems are made by a very large number of interacting elements. Even if these systems cannot be studied at the single-element level, nevertheless they allow for a very efficient statistical treatment (e.g., in terms of statistical mechanics). There is a third kind of complexity that cannot be studied this way and that Weaver identified with biological systems: *Organized Complexity*.

Both the arguments and terminology used by Weaver were largely coincident with Laughlin et al., identifying "The Middle Way" [2] as the real frontier of basic Science in the XXI century. The natural entities in which the organized complexity approach was more fruitful were protein molecules. Such molecules can be collocated at the boundary between simple and complex physics [5], allowing also for both very accurate experiments and for a very refined structural description (e.g., via X-ray crystallography). The situation nowadays continues to be unchanged with respect to Laughlin et al. [2]: protein science is still the most explored field in the realm of organized complexity [6–12].

The resurgent interest in graph-theoretical methods for the analysis of complex systems [13–17,10,18–29] set forth by the works by Barabasi, Strogatz and other pioneers in the first decade of XXI century, allowed for a rich set of measures providing a multifaceted description of complex networks [30–32,25,14,33–35,33,36–38]. It is sufficient that a given problem is formalized by means of a vertex-and-edge representation – with vertices being the parts and edges their pairwise relations – to allow facing the problem according with the terms of organized complexity [15]. Protein molecules are at the forefront of complex network way to organized complexity. In fact, it is possible to represent a protein molecule in terms of protein contact network (i.e., a network having amino acid residues as vertices and edges representing physical contacts between them) by considering its native three-dimensional structure [9,39]. It is worth noting that graph-based representations are at the core of chemical sciences: the structural formulas are in fact graphs [40].

In this work, we face a data-driven quest for mesoscopic organization principles of complex biological systems by analyzing different complex networks: protein contact networks, metabolic networks, and genetic networks, together with simulated archetypal networks whose wiring scheme is generated by mathematical rules. All considered networks are characterized in terms of two collections of numerical features. The first collection is based on classical topological descriptors, such as the modularity and statistics of the shortest paths. The second one exploits the discrete heat kernel (HK), elaborated using the eigendecomposition of the (normalized) graph Laplacian [41]. With a first preliminary analysis, we show that the different classes of networks are discriminated by a suitable embedding of such numeric features. This is reasonably expected given the substantially different natures of the analyzed networks, but by no means can be considered as a trivial result. As a matter of fact, the distinction in terms of metabolic, genetic, and protein contact networks is based on network functions with no necessary on-to-one relation with corresponding structure—identical network structures can perform very different functions. The demonstration of a link between functional and structural properties of the corresponding graph representation is a prerequisite for the soundness of the proposed strategy of analysis. An important result is that the two network characterizations resulted to be strongly correlated with each other, so giving a proof-of-concept of the reliability and interpretability of the adopted network descriptions. From this first analysis, it also emerged that protein contact networks display unique properties that do not allow for a straightforward classification in any of the considered classical archetypal networks, stressing the need for the search of new, specialized generative models for protein contact networks. The second, and more important, claim of this paper is that an ensemble HK analysis allows us to derive results in agreement with known chemico-physical properties of protein molecules. Our analysis demonstrates that a (simulated) diffusion process in protein contact networks proceeds slower than normal diffusion, i.e., we observe subdiffusion. Notably, a two-regime diffusion emerged from the analysis of the heat trace decay: a fast and a slow regime. The fast regime is driven by shortcuts putting in contact amino acid residues far-away along the sequence. Subdiffusion is ubiquitous in Nature [42] and it is a well-studied peculiarity describing energy flow [43–46] and vibration dynamics [47–51] in protein molecules. Interestingly, here we are able to observe the same phenomenon by exploiting only algebraic properties of the graph representations of an ensemble of varying-size proteins. The demonstration of the ability of such a network formalization to explain a unique chemico-physical property of protein molecules, as heat diffusion, is a proof-of-concept of the usefulness of graph theory in chemical system modeling.

There is sufficient agreement on the fact that proteins, considering their native structure, are highly modular and fractal networks [52,53,14,54–56]; yet they are characterized also by short paths connecting distant regions of the molecules responsible for the fast-track transport of energy and protein allosteric properties [43]. The modular character of protein contact networks has a direct impact on diffusion processes that are simulated via graph-theoretical methods. In fact, well-defined modular organizations slow down diffusion processes in networks [57]. Additionally, theoretical and experimental results regarding diffusion on porous and fractal media predict fractional scaling exponents for the mean squared displacement (a well-known measure of diffusion), which give rise to the so-called anomalous diffusion [58,59]. To conclude, another interesting fact deduced from our analysis is that, at odds with the other networks, modularity of protein graphs increases with the size of the network. This observation is fairly interesting since path efficiency and modular properties are two conflicting features in networks [60,61], which however seem to be suitably optimized in proteins.