

Geometricity of Residue Interaction Graphs

Tijana Milenković^{1,+}, Ioannis Filippis^{2,+}, Michael Lappe², Nataša Pržulj^{1,*}

1. Department of Computer Science, University of California, Irvine, CA 92697-3435, USA

2. Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, D-14195 Berlin, Germany

⁺These authors contributed equally to this work

* Corresponding author (E-mail: natasha@ics.uci.edu)

Abstract

Finding a well fitting null model for biological networks is important in many research areas. A good model should generate graphs that resemble real data as closely as possible across a wide range of statistical measures. Degree-preserving randomized models have been widely used for this purpose in biomolecular networks. However, such a single summary global statistic of a network may not be detailed enough to capture the complex topological characteristics of a network.

Here, we consider residue interaction graphs (RIGs) as network representations of protein structures with residues as nodes and inter-residue interactions as edges. The RIGs observed in this study are derived from a structurally diverse data set covering nine proteins. For each protein, in addition to a series of distance cut-offs, we examine three different “contact types”: we denote by “BB” (“SC”) the RIGs that contain as edges only the residue pairs that have heavy backbone (side-chain) atoms within the given distance cut-off; we denote by “ALL” the most commonly used RIG model, in which all heavy atoms of every residue are taken into account when determining residue interactions.

In order to find a well-fitting network model for RIGs, we evaluate the fit of RIGs to five random graph models: Erdős-Rényi random graphs (“ER”) [1], random graphs with same degree distribution as the RIGs (“ER-DD”), 3-dimensional geometric random graphs constructed using Euclidean boxes and Euclidean distance norm (“GEO-3D”) [2], Barabási-Albert type scale free networks (“SF-BA”) [3], and stickiness-index based networks (“STICKY”) [4]. Each of the generated model networks that corresponds to a RIG has the same number of nodes and the number of edges within 1% of those in the RIG.

Exact comparisons of large networks are computationally infeasible due to NP-completeness of the underlying subgraph isomorphism problem [5]. Thus, to evaluate the fit of the data to the model networks, we compare the RIGs to the model networks with respect to a set of *network properties*. To overcome the limitations introduced by using a single network property (such as the degree distribution), we perform a fine-grained analysis of RIGs that is based on a variety of *local* and *global* network properties. The local properties used in this study are based on *graphlets*, small connected non-isomorphic induced subgraphs of large networks [6]. The two local properties that we use are *relative graphlet frequency distance (RGF-distance)* [6] and *graphlet degree distribution agreement (GDD-agreement)* [7]. Additionally, we use four global network properties: the *degree distribution*, the *clustering coefficient*, the *clustering spectrum*, and the *average network diameter*.

We show that 3-dimensional geometric random graphs provide the best fit to these RIGs for all reasonable and practically used distance cut-offs. All analyzed local and global network properties offer support to superiority of the GEO-3D model. Illustrations showing the fit of one of the analyzed proteins to the five network models according to GDD-agreements and RGF-distances are presented in Figure 1. For all distance cut-offs and all contact types, RGF-distances and GDD-agreements between the RIGs and the model networks strongly favor geometric random graphs. The similar trends follow for other network properties and other proteins.

To summarize, we address the important issue of finding a well fitting null model for protein structure networks. We show the superiority of the fit of geometric random graphs over four other random graph models to RIGs that correspond to nine structurally different proteins and are constructed using three different contact types and a series of residue distance cut-off values. This superiority of the fit is demonstrated by examining two highly constraining measures of network local structure, as well as four standard measures of global network structure. Our geometric random graph null model may facilitate further graph-based studies of protein conformation space and the discovery of significant structural motifs. This analysis may also have important implications for protein structure comparison and prediction.

References

- [1] Erdős, P. and Rényi, A. (1959) On random graphs. *Publicationes Mathematicae*, **6**, 290–297.
- [2] Penrose, M. (2003) *Geometric Random Graphs*, Oxford University Press.
- [3] Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.
- [4] Pržulj, N. and Higham, D. (2006) Modelling proteinprotein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, **3**(10), 711–716.
- [5] Cook, S. (1971) In *Proc. 3rd Ann. ACM Symp. on Theory of Computing* Association for Computing Machinery pp. 151–158.
- [6] Pržulj, N., Corneil, D. G., and Jurisica, I. (2004) Modeling interactome: Scale-free or geometric?. *Bioinformatics*, **20**(18), 3508–3515.
- [7] Pržulj, N. (2006) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**, e177–e183.

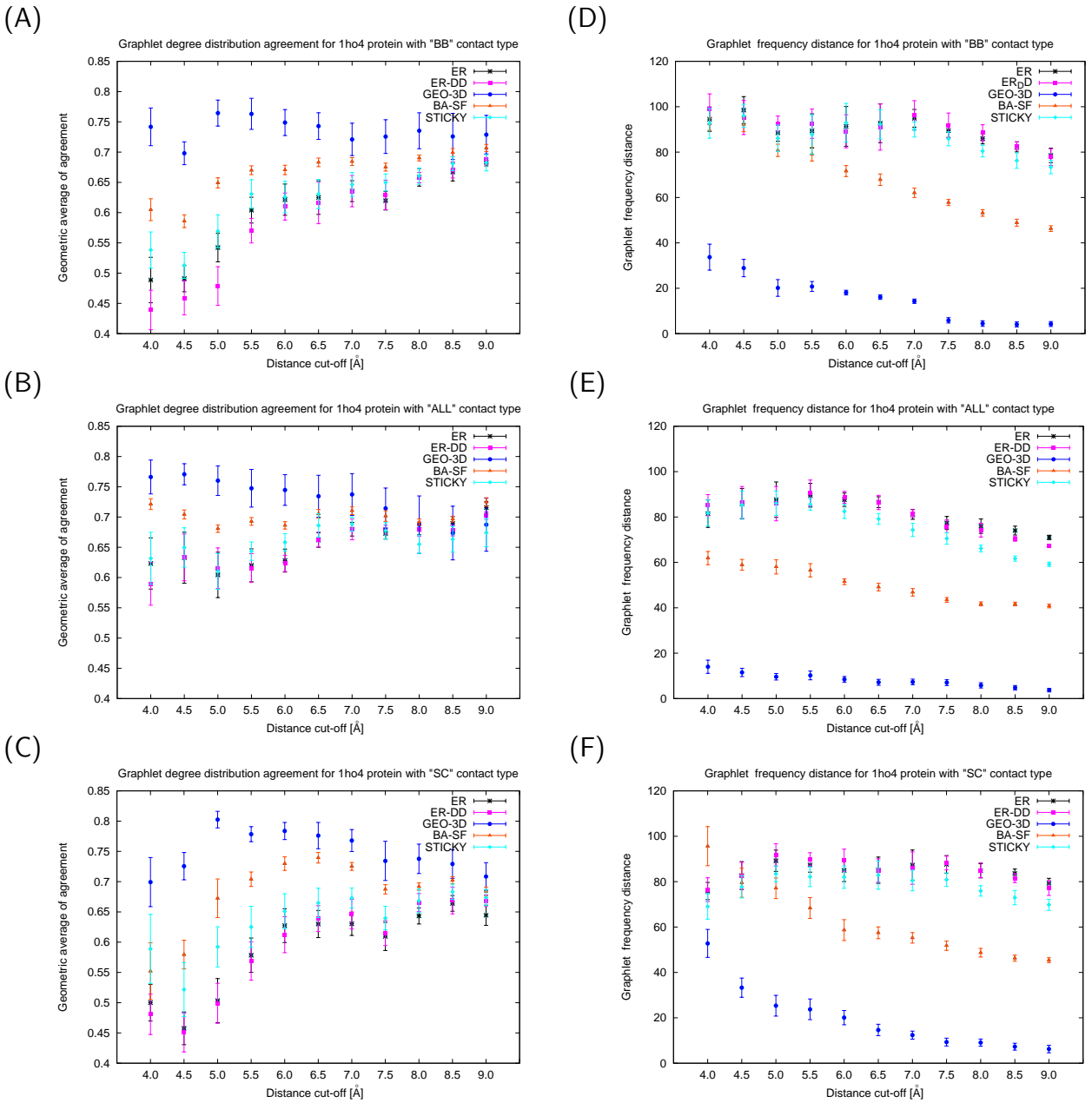


Figure 1: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1ho4 protein that are constructed for each of the three contact types ("BB", "ALL", and "SC") and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for "BB" contact type. **B.** GDD-agreement for "ALL" contact type. **C.** GDD-agreement for "SC" contact type. **D.** RGF-distance for "BB" contact type. **E.** RGF-distance for "ALL" contact type. **F.** RGF-distance for "SC" contact type. The larger the GDD-agreement in panels A-C, the better the fit. The smaller the RGF-distance in panels D-F, the better the fit.