Biochemical networks display universal structure across projections and levels of organization

Harrison B. Smith[1], Hyunju Kim[2] and Sara I. Walker[2,3,4,5]

[1]Earth-Life Science Institute, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan
[2]Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ, USA
[3]School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA
[4]ASU-SFI Center for Biosocial Complex Systems, Arizona State University, Tempe, AZ, USA
[5]Blue Marble Space Institute of Science, Seattle, WA, USA.
hbs@elsi.jp

Abstract

Biochemical reactions underlie all living processes. Like many systems, their web of interactions is difficult to fully capture and quantify with simple mathematical objects. Nonetheless, a huge volume of research has suggested many real-world systems–including biochemical systems–can be described simply as 'scale-free' networks, characterized by power-law degree distributions. More recently, rigorous statistical analyses upended this view, suggesting truly scale-free networks may be rare. We provide a first application of these newer methods across two distinct levels of biological organization: analyzing an ensemble of biochemical reaction networks generated from 785 ecosystem-level metagenomes and 1082 individual-level genomes (representing all domains of life). Our results confirm only a few percent of biochemical networks meet the criteria necessary to be more than super-weakly scale-free. We perform distinguishability tests across individual and ecosystem-level biochemical networks and find there is no sharp transition in the organization of biochemistry across distinct levels of the biological hierarchy–a result that holds across network projections. This suggests the existence of common organizing principles operating across different levels of biology, which can best be elucidated by analyzing all possible coarse-grained projections of biochemistry in tandem across scales.

Introduction

Broido and Clauset recently developed a methodology to compare the degree distributions of network projections of different complexities, classifying the degree to which they are scale-free on a scale from "Not scale-free" all the way to "strongest" [1]. This provides a framework for statistically analyzing many projections of a given system to determine how well scale-free structure describes the real underlying system when projected onto its different coarse-grained representations.

Herein, we build from the work of Broido and Clauset with specific application to the problem of characterizing biochemical systems. A novelty in our approach is recognizing that in order to really understand the structure of real-world biological systems, the relevant scale(s) for performing such analysis must also be considered. In particular, many biological systems are hierarchical, with networks describing interactions across multiple levels. For example, one may study the biochemistry of individual species, but ultimately the function of an individual in a natural system depends on a complex interplay of interactions among the many species comprising its host ecosystem. In this way, biochemistry is hierarchically organized into individuals and ecosystems. Indeed, much discussion about universal properties of life has shifted focus from individuals to ecosystems as the relevant scale best capturing the regularities of biological organization [2, 3]. It is unclear at present whether analysis of biochemical networks at the level of individuals or ecosystems will best uncover their structure and permit identifying generative mechanisms for biology, or whether all levels must be considered simultaneously.

Results

Utilizing the framework developed by Broido and Clauset [1], we perform statistical analysis of an ensemble of biochemical systems generated from 785 ecosystem-level metagenomes and 1082 individual-level genomes (representing all three domains of life). We use the full set of biochemical reactions encoded in each (meta)genome to construct eight distinct network representations of each respective biochemical system. This resulted in 8656 network projections for the 1082 individual-level biochemical datasets, and 6280 network projections for the 785 ecosystem-level biochemical datasets. We determine whether or not these datasets are scale-free, and analyze the aspects of them, and their diverse projections, that tend to lend themselves to be more or less scale-free.

Our results indicate most biochemistry at the individual and ecosystem-level is characterized by networks that are "super-weakly" scale-free (**Fig. 1**). That is, while the power-law is better than other models for fitting the shape of their degree distributions, the power-law is not itself a good model. When doing a goodness-of-fit test, we find that the majority of network representations of each ge-

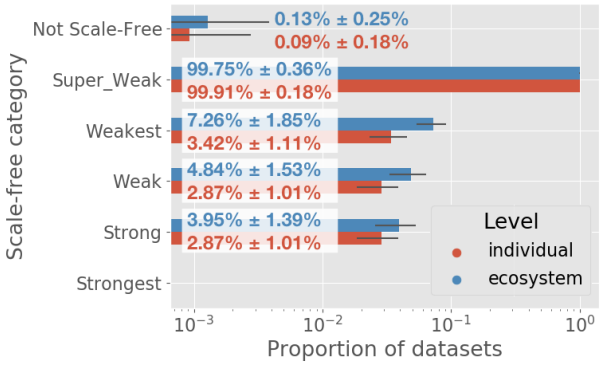

Figure 1: **The vast majority of individual and ecosystem level networks are not "scale-free".** Most datasets are super weak, indicating that when compared to other models, a power-law distribution is a better fit. However, the power-law distribution is not a "good" fit for most dataset network representations. Overlaid values show the percent of networks of each level which fall into each category ($\pm$2SD.).

nomic/metagenomic dataset have $p < 0.1$, indicating there is less than a 10% chance that our data is truly power-law distributed. This effectively rules out the possibility that our data is drawn from a power-law shaped degree distribution, despite the fact that, when compared to other distributions through log-likelihood ratios, 99% of all datasets do not favor alternative heavy tail distributions for the majority of their network-projections

While other literature [4, 5] has advocated for unipartite networks (with all compounds participating in a reaction connected), we find that these networks overestimate power-law goodness-of-fit p-values and α values compared to reaction and bipartite networks.

Furthermore, we demonstrate that using random forest distinguishability analyses on a combination of all the results of scale-free analyses completed in this work can predict, better than chance, whether the data comes from individuals or ecosystems.

Discussion

The fact that multiple levels and multiple projections of biochemistry reveal common structure suggests universal principles may be within reach if cast within an ensemble theory of biochemical network organization.

Whether or not the observed structure is truly a universal property of life's chemical systems is more difficult to conclude. Achieving such a theory requires recognizing that, unlike simple physical systems where statistics over individual components is sufficient to describe and predict their behavior, biological systems require additional information about the structure of interactions among their many components. Perhaps incorporating additional reaction data on the particular flows of atoms between compounds would further elucidate regularities of biological structure. But how to project this structure onto simple mathematical objects that can be quantifiably characterized and compared remains a central problem of complex systems science. In physics, the relevant coarse-graining procedure is well understood, but we are not so far in complexity science: the first hurdle we must traverse is to identify the proper coarse-grained network representations for analysis. Existing literature cautions against using unipartite network projections, as it is argued they can lead to "wrong" interpretations of system properties such as degree in biochemical networks [6, 7]. We find instead that whether or not this conclusion should be drawn is highly dependent on the particular characteristics of degree or the degree distribution under consideration.

The similarities and differences in the structure of different projections provides insight into the actual structure of the underlying system of interest. In physics we have become accustomed to one unique coarse-grained descriptor providing insight into the structure of a system. It may be that to really understand complex interacting systems, such as the systems of reactions underlying all life on Earth, we must forget the allure of simple, singular models. Instead, to characterize the regularities associated with living processes, we should perform statistical analyses over many (still relatively simple) coarse-grained projections.

Acknowledgements

We thank the Emergence@ASU team (especially Doug Moore, Cole Mathis, and Jake Hanson) for feedback through various stages of this work.

References

[1] Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1017, 2019.

[2] Eric Smith and Harold J Morowitz. *The origin and nature of life on Earth: the emergence of the fourth geosphere*. Cambridge University Press, 2016.

[3] Hyunju Kim, Harrison B Smith, Cole Mathis, Jason Raymond, and Sara I Walker. Universal scaling across biochemical networks on earth. *Science Advances*, 5(1):eaau0149, 2019.

[4] Petter Holme. Model validation of simple-graph representations of metabolism. *Journal of The Royal Society Interface*, 6 (40):1027–1034, 2009.

[5] Petter Holme and Mikael Huss. Substance graphs are optimal simple-graph representations of metabolism. *Chinese Science Bulletin*, 55(27-28):3161–3168, 2010.

[6] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and cellular networks. *PLoS computational biology*, 5(5):e1000385, 2009.

[7] Raul Montanez, Miguel Angel Medina, Ricard V Sole, and Carlos Rodríguez-Caso. When metabolism meets topology: Reconciling metabolite and reaction networks. *Bioessays*, 32 (3):246–256, 2010.