Seminars in Cell & Developmental Biology 51 (2016) 32–43

Contents lists available at ScienceDirect

Seminars in Cell & Developmental Biology

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

Review

Methods of information theory and algorithmic complexity for network biology☆

Hector Zenil*, Narsis A. Kiani, Jesper Tegnér

Unit of Computational Medicine, Department of Medicine, Karolinska Institute & Center for Molecular Medicine, Karolinska University Hospital, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 16 August 2015
Accepted 7 January 2016
Available online 21 January 2016

Keywords:
Information theory
Complex networks
Kolmogorov complexity
Algorithmic randomness
Algorithmic probability
Biological networks

ABSTRACT

We survey and introduce concepts and tools located at the intersection of information theory and network biology. We show that Shannon's information entropy, compressibility and algorithmic complexity quantify different local and global aspects of synthetic and biological data. We show examples such as the emergence of *giant components* in Erdös-Rényi random graphs, and the recovery of topological properties from numerical kinetic properties simulating gene expression data. We provide exact theoretical calculations, numerical approximations and error estimations of entropy, algorithmic probability and Kolmogorov complexity for different types of graphs, characterizing their variant and invariant properties. We introduce formal definitions of complexity for both labeled and unlabeled graphs and prove that the Kolmogorov complexity of a labeled graph is a good approximation of its unlabeled Kolmogorov complexity and thus a robust definition of graph complexity.

Contents

1. Introduction

Over the last decade network theory has become a unifying language in biology, giving rise to whole new areas of research in computational systems biology. Gene networks are conceptual models of genetic regulation where each gene is considered to be

☆ Dr. Ali Masoudi-Nejad, Guest Editor for this paper.

* Corresponding author.
E-mail address: hector.zenil@algorithmicnaturelab.org (H. Zenil).

http://dx.doi.org/10.1016/j.semcdb.2016.01.011
1084-9521/

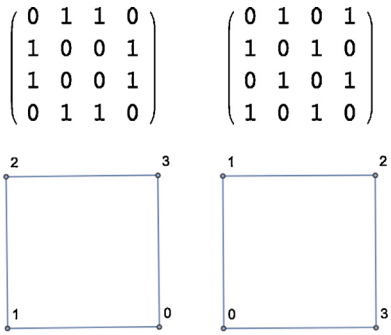

Fig. 1. Isomorphic graphs with two different adjacency matrix representations, illustrating that the adjacency matrix is not an invariant of a graph under relabelings. However, similar graphs have adjacency matrices with similar algorithmic information content.

directly affected by a number of other genes, and are usually represented by directed graphs.

Classical information theory has for some time been applied to networks, but Shannon entropy, like any other computable measure (i.e. one that is a total function, returning an output in finite time for every input), is not invariant to changes of object description [45].

More recently, algorithmic information theory has been introduced as a tool for use in network theory, and some interesting properties have been found [40,42,46]. For example, in [40] correlations were reported among algebraic and topological properties of synthetic and biological networks by means of algorithmic complexity, and an application to classify networks by type was developed in [42].

We review and explore further these information content approaches for characterizing biological networks and networks in general. We provide theoretical estimations of the error of approximations to the Kolmogorov complexity of graphs and complex networks, offering both exact and numerical approximations. Together with [40] and [42], the methods introduced here represent a novel view and constitute a formal approach to graph complexity, while providing a new set of tools for the analysis of the local and global structure of networks.

2. Graph notation and complex networks

A graph G is *labeled* when the vertices are distinguished by names such as $u_1, u_2, \ldots u_n$ with $n = |V(G)|$. Graphs G and H are said to be *isomorphic* if there is a bijection between the vertex sets of G and H, $\lambda : V(G) \to V(H)$ such that any two vertices u and $v \in G$ are adjacent in G if and only if $\lambda(u)$ and $\lambda(v)$ are adjacent in H. When G and H are the same graph, the bijection is referred to as an *automorphism* of G. The adjacency matrix of a graph is not an invariant under *graph relabelings*. Fig. 1 shows two adjacency matrices for isomorphic graphs. A *canonical form* of G is a labeled graph $Canon(G)$ that is isomorphic to G, such that every graph that is isomorphic to G has the same canonical form as G. An advantage of $Canon(G)$ is that unlike $A(G)$, $A(Canon(G))$ is a graph invariant of $Canon(G)$ [1].

One of the most basic properties of graphs is the number of links per node. When all nodes have the same number of links, the graph is said to be *regular*. The *degree* of a node v, denoted by $d(v)$, is the number of (incoming and outgoing) links to other nodes. We will also say that a graph is *planar* if it can be drawn in a plane without its edges crossing. Planarity is an interesting property because only planar graphs have *duals*. A *dual graph* of a planar graph G is a graph that has a vertex corresponding to each face of G, and an edge joining two neighboring faces for each edge in G.

A popular type of graph that has been studied is the so-called *Erdös-Rényi* [12,14] (*ER*) graph, in which vertices are randomly and independently connected by links with a fixed probability (also called *edge density*) (see Fig. 2 for a comparison between a regular and a random graph of the same size). The probability of vertices being connected is called the *edge probability*. The main characteristic of random graphs is that all nodes have roughly the same number of links, equal to the average number of links per node. A *ER* graph $G(n, p)$ is a graph of size n constructed by connecting nodes randomly with probability p independent from every other edge. Usually *ER* graphs are assumed to be non-recursive (i.e. truly random), but *ER* graphs can be constructed recursively with, for example, pseudo-random algorithms. Here we will assume that *ER* graphs are non-recursive, as theoretical comparisons and bounds hold only in the non-recursive case. For numerical estimations, however, we use a pseudo-random edge connection algorithm, in keeping with common practice.

ER random graphs have some interesting properties, but biological networks are not random. They carry information, connections between certain elements in a biological graph are favored or avoided, and not all vertices have the same probability of being connected to other vertices. The two most popular complex network models consist of two algorithms that reproduce certain characteristics found in empirical networks. Indeed, the field has been driven largely by the observation of properties that depart from properties modeled by regular and random graphs. Specifically, there are two topological properties of many complex networks that have been a focus of interest. A *simple graph* is a graph with no self-loops and no multi-edges. Throughout this paper we will only consider simple graphs.

A network is considered a *small-world* graph G (e.g. see Fig. 3) if the average graph distance D grows no faster than the log of the number of nodes: $D \sim \log V(G)$. Many networks are *scale-free*, meaning that their degrees are size independent, in the sense that the empirical degree distribution is independent of the size of the graph up to a logarithmic term. That is, the proportion of vertices with degree k is proportional to γk^{τ} for some $\tau > 1$ and constant γ. In other words, many empirical networks display a power-law degree distribution.

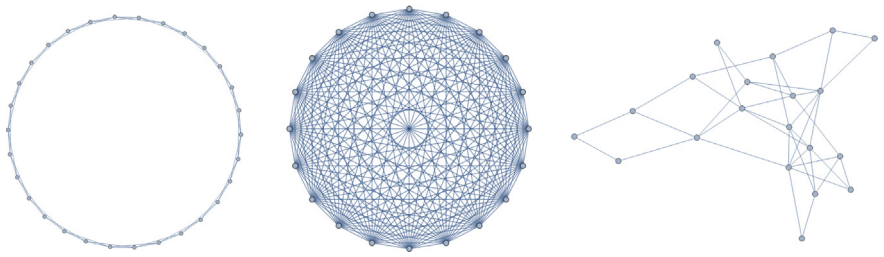

Fig. 2. Examples of two regular graphs (left and middle) are a $2n$ circular graph with 20 nodes and a complete graph with 20 nodes, both of whose descriptions are very short, hence $K(G) \sim \log |V(G)| \sim 4.32$ bits. In contrast, a random graph (right) with the same number of nodes and number of links requires more information to be specified, because there is no simple rule connecting the nodes and therefore $K(G) \sim |E(G)| = 30$ bits.