Algorithmic complexity of motifs clusters superfamilies of networks

Hector Zenil, Narsis A. Kiani and Jesper Tegnér
Unit of Computational Medicine, Center for Molecular Medicine,
Karolinska Institutet, Stockholm, Sweden
Email: hector.zenil,narsis.kiani,jesper.tegner@ki.se

***Abstract*—Representing biological systems as networks has proved to be very powerful. For example, local graph analysis of substructures such as subgraph overrepresentation (or motifs) has elucidated different sub-types of networks. Here we report that using numerical approximations of Kolmogorov complexity, by means of algorithmic probability, clusters different classes of networks. For this, we numerically estimate the algorithmic probability of the sub-matrices from the adjacency matrix of the original network (hence including motifs). We conclude that algorithmic information theory is a powerful tool supplementing other network analysis techniques.**

***Keywords*-information theory; complex networks; network motifs; Kolmogorov complexity; algorithmic probability; information content; network typology.**

I. Network biology

A graph or network G, consists of a set of vertices V (also called nodes) and a set of edges E (also called links). Two vertices, i and j, form an edge of the graph if 2 vertices in E are connected. A useful representation of a graph (see Fig. 1) is what is called an adjacency matrix. The adjacency matrix of G, which we can denote by $Adj(G)$, is a $n \times n$ binary matrix with entries $a_{i,j} = 1$ if $(i, j) \in E$ and 0 otherwise. The adjacency matrix $Adj(G)$ fully determines the links E connecting all elements in V and therefore constitutes a full description of a graph. A subgraph G' of a graph G is a graph such that $V(G') \subseteq V(G)$ and $E(G') \subseteq E(G)$ satisfying the property that for every $\epsilon \in E(G')$, where ϵ has endpoints $u, v \in V(G)$ in the graph G, then $u, v \in V(G')$ and ϵ has endpoints u, v in G', i.e. the edge relation in G' is the same as in G. Directed graphs are graphs with links that have a direction (ingoing or outgoing) relative to a node, that is, if u and v are linked nodes, (u, v) is different from (v, u).

A. Network Motifs

One important development in network biology is the concept of network motifs [1], [10], that is recurrent and statistically significant sub-graphs found in a network, as compared to a uniform distribution in a random network. A thorough review can be found in [1] and [17]. As is to be expected, biological networks are not random networks because biological networks carry information necessary for an organism to develop. Motifs (see Fig. 2) are believed to be of signal importance largely because they may reflect functional properties.

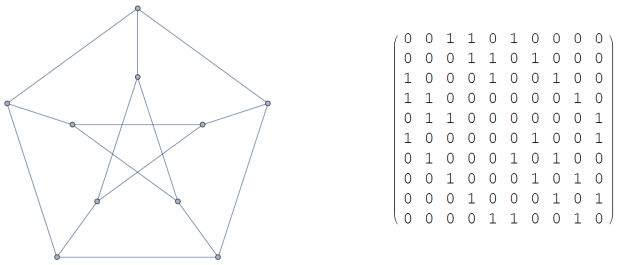

Figure 1. The so-called Petersen graph followed by its adjacency matrix. Regular graphs are Kolmogorov simple (see Section II) because every node requires the same amount of information to be specified.

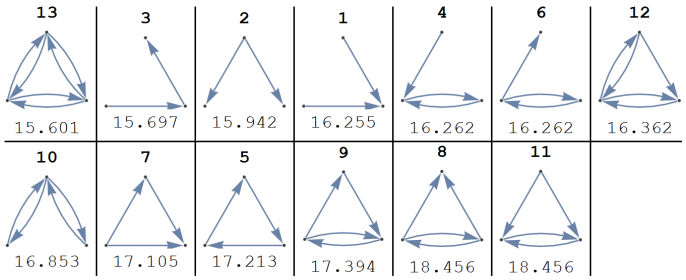

Figure 2. Algorithmic randomness estimation (by BDM) of the adjacency matrices of the 13 directed subgraphs of size 3 that can occur in a network as motifs (sorted from smaller to larger Kolmogorov complexity estimations).

II. Kolmogorov network motif analysis

Biological networks carry information, transfer information from one region to another and implement functions represented by the network interactions. In a biological network nodes usually represent proteins, metabolites, genes or transcription factors. A link represents the interactions between the nodes in a network that can correspond to protein-protein binding interactions, metabolic coupling or regulation. The degree of a node in a biological network of protein interactions represents the number of proteins with which it interacts. Connections among elements in a biological network are therefore unlikely to be like a regular graph or a random one.

Information theory [12] is often used as a basis for assessing the information content and randomness of an object. Given a finite sequence s of N symbols repeatedly chosen from a set of n elements, the average symbol information, also called *information entropy*, is given by $H(s) = \sum_i p_i \log_2(p_i)$, where p_i is the relative frequency of an element i in the finite sequence. This entropy is maximal when the relative frequencies are all equal disregarding the element internal structure, when it amounts to $H_{max} = \log_2(n)$.

A powerful measure of information content and randomness is the measure provided by Kolmogorov complexity [6] and [3]—that we will denote by K. This is because K is an intrinsic and universal measure of complexity guaranteed to asymptotically spot any possible computable regularity [13] of an object independent of any ensemble. Formally, the Kolmogorov complexity of a string s is $K(s) = \min\{|p| : U(p) = s\}$, that is, the length (in bits) of the shortest program p that when running on a universal Turing machine U outputs s upon halting. By the *invariance theorem* (see [9], [2], [4]), K_U only depends on U up to a constant, so as is conventional, the U subscript can be dropped.

A. Algorithmic Probability

Due to its great power, K comes with a technical inconvenience (called semi-computability) and it is proven that no effective algorithm exists which takes a string s as input and produces the exact integer $K(s)$ as output [6], [3], and this is related to a common problem in computer science known as the undecidability of the Halting problem [9], [4]. However, K has proven to be useful in graph theory [14]. This has been achieved by way of a seminal concept in the theory of algorithmic information, namely the notion of algorithmic probability. The algorithmic probability of a string s is a measure that describes the probability that a valid random program p produces the string s when run on a universal (prefix-free[1]) Turing machine U. In equation form this can be rendered as $m(s) = \sum_{p:U(p)=s} 1/2^{|p|}$ (Eq. 1). That is, the sum over all the programs p for which U outputs s and halts.

The algorithmic probability measure $m(s)$ is related to Kolmogorov complexity $K(s)$ in that $m(s)$ is at least the maximum term in the summation of programs, given that the shortest program carries the greatest weight in the sum. The Coding Theorem [8] further establishes the connection between $m(s)$ and $K(s)$ as follows: $|-\log_2 m(s) - K(s)| < c$, where c is a fixed constant, independent of s. The Coding Theorem implies that [5] one can estimate the Kolmogorov complexity of an object by calculating its frequency of appearance [5], [15].

[1]The group of valid programs forms a prefix-free set (no element is a prefix of any other, a property necessary to keep $0 < m(s) < 1$.) For details see [9].

The approach to determining the algorithmic complexity of network motifs thus involves considering how often the adjacency matrix of a subgraph is generated by a random Turing machine on a 2-dimensional array, also sometimes called a *termite* or Langton's *ant* Turing machine [7]. This same technique has been applied to image classification and classification of space-time diagrams of abstract computing machines [16].

We thus approximate a graph G motif complexity $K(G)$ (and denoted by BDM for Block decomposition method) using the following formula:

$$K(G) = \sum_{(r_u,n_u)\in Adj(G)_d} K(r_u) + \log_2(n_u) \qquad (1)$$

where $Adj(G)_{d\times d}$ represents the set with elements (r_u, n_u), obtained when decomposing the adjacency matrix of G into all subgraphs contained in G of size d. In each (r_u, n_u) pair, r_u is one such submatrix of the adjacency matrix and n_u its multiplicity (number of occurrences). As it can be seen from the formula, repeated subgraphs only contribute to the complexity value with the subgraph BDM complexity value once plus a logarithmic term as a function of the number of occurrences. This is because information content of subgraphs is only sub-additive, as one would expect from the growth of its description length ("n times a subgraph").

Worth to notice is that there is an important difference between calculating the complexity of motifs and calculating the complexity of submatrices of the adjacency matrix of a graph or network. Not all submatrices represent a motif or a subgraph. However, all motifs are a submatrix of the original adjacency matrix. So calculating the complexity of matrices does provide approximations to the network motifs but the rationale for reconstructing the adjacency matrix comes from the idea that several small Turing machines working on a grid can fully reconstruct all the submatrices that constitute of the original adjacency matrix. Yet to know and investigate is the removal rate of calculating the complexity of all the submatrices (including motifs but also non-proper subgraphs) against calculating the complexity of only the proper subgraphs (including motifs) of the original network.

Eq. 1, shows that motif diversity accounts for most of the Kolmogorov complexity of a network given that repetition is penalised only having a limited additive logarithmic effect. How much information about G is lost by estimating $K(G_d)$? The invariance theorem in the theory guarantees convergence only if K is measured over all d up to $d = |G|$ (the size of the network), yet it does not tell whether the most important information is captured for a small d, perhaps even suggesting a more effective way to discriminate graphs and networks by way of their local regularities as suggested in Fig. 3 (for example, in the case of biological networks, specific functions carried out and represented by the motifs).

III. Results

In [11], researchers found that complex biological, technological, and sociological networks of very different sizes and topologies can be clustered in superfamilies of networks when performing a network motif analysis.

In agreement with these results and over the same data (except for some networks unavailable from the original dataset [11] that authors did not provide), we show (Fig. 3) that looking at local regularities (graphs and motifs of size 4) of the same complex networks, a classification of networks by network family arises from the numerical approximation of the complexity of the subgraphs of size 4 (including motifs) by means of algorithmic probability as measured by the Block Decomposition Method (BDM), hence reinforcing the idea that key in these networks information-processing structure is the occurrence of subgraphs of certain specific configuration and their frequency of appearance.

In Fig. 3, networks running along the x-axis are distributed as follow: 1 and 2 are developmental genetic networks (in yellow), networks 3 to 5 are power grid networks (red), 6 and 7 (green) are protein signalling networks and 8 and 9 are social networks (pink). BZIP2 values did not present such a clustering capability but within each family, the Kolmogorov complexity approximation was preserved, that is, if networks G_i, G_j are of the same type or belong to the same network family (e.g. genetic, social, electric or protein), and and G' is the matrix decomposition of the adjacency matrix of G, then if $BDM(G'_i) < BDM(G'_j)$ then $BZIP2(G_i) < BZIP2(G_j)$. All networks come from the supplemental material of [11], where a similar result, but by different means, was derived.

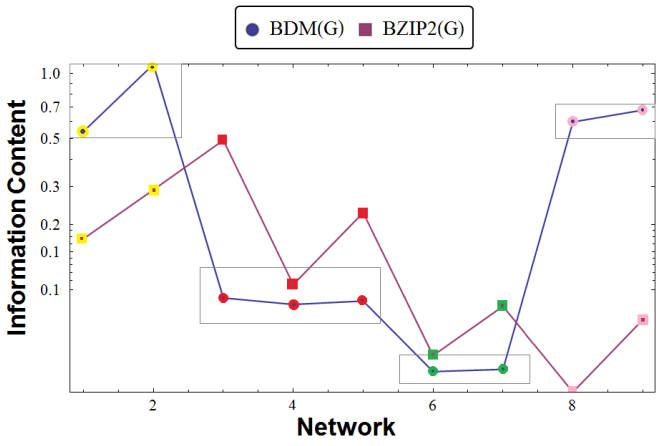

Figure 3. Kolmogorov complexity, approximated by BDM and normalised by a constant, of the motifs and subgraphs of size 4 of the nine networks from [10]. Network BDM and BZIP2 values (plotted in log scale) were coloured by networks family (only visible online and colour version of the paper). Joined data points are for illustration only (of the relative positions of the networks with respect to the others). Average distance between network families is 1 944 bits, that represents a 25% variance between BDM mean values by network family, hence apart enough from each other.

IV. Concluding remarks

We have praised for an encompassing information-theoretic study of biological networks at different scales only to find out that motif complexity suggests to be a powerful tool to be further studied and exploited. The method here advanced demonstrates that local scale analysis for algorithmic information content approximations deliver results suggesting that, as reported in [11], by looking at a very small scale one can characterize important features of the original networks. These partial results show that a local complexity approach retrieves enough information about the networks to be distinguished from other type of networks.

References

[1] U. Alon, Network Motifs: theory and experimental approaches, *Nature*, 450, vol. 8, June 2007.

[2] C.S. Calude, *Information and Randomness: An Algorithmic Perspective*, EATCS Series, 2nd. edition, 2010, Springer.

[3] G.J. Chaitin. On the length of programs for computing finite binary sequences *Journal of the ACM*, 13(4):547–569, 1966.

[4] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, 2nd Edition, Wiley-Blackwell, 2009.

[5] J.-P. Delahaye and H. Zenil, Numerical Evaluation of the Complexity of Short Strings: A Glance Into the Innermost Structure of Algorithmic Randomness, *Applied Mathematics and Computation* 219, pp. 63-77, 2012.

[6] A. N. Kolmogorov. Three approaches to the quantitative definition of information, *Problems of Information and Transmission*, 1(1):1–7, 1965.

[7] C.G. Langton, Studying artificial life with cellular automata, *Physica D: Nonlinear Phenomena 22* (1–3): 120–149, 1986.

[8] L. A. Levin. Laws of information conservation (non-growth) and aspects of the foundation of probability theory, *Problems of Information Transmission*, 10(3):206–210, 1974.

[9] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed., Springer, 2009.

[10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks, *Science* 298, no. 5594: 824-827, 2002.

[11] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, V. Ayzenshtat, M. Sheffer, U. Alon, Superfamilies of designed and evolved networks, *Science* 303, 1538–1542, 2004.

[12] C.E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, IL, 1949.

[13] R.J. Solomonoff, A formal theory of inductive inference: Parts 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[14] H. Zenil, F. Soler-Toscano, K. Dingle and A. Louis, Correlation of Automorphism Group Size and Topological Properties with Program-size Complexity Evaluations of Graphs and Complex Networks, (submitted).

[15] H. Zenil and J-P. Delahaye, On the Algorithmic Nature of the World. In G. Dodig-Crnkovic and M. Burgin (eds), *Information and Computation,* World Scientific Publishing Company, 2010.

[16] H. Zenil, F. Soler-Toscano, J.-P. Delahaye and N. Gauvrit, Two-Dimensional Kolmogorov Complexity and Validation of the Coding Theorem Method by Compressibility, (submitted).

[17] H. Zenil, and J. Tegnér, Methods of information theory and algorithmic complexity for network biology. In M. Elloumi, A.Y. Zomaya (eds.), *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data* (Wiley Series in Bioinformatics), IEEE Computer Society and Wiley, 2014 (forthcoming)