CT 1.3.5
Information-theoretic measure of network complexity

A. Hintze, C. Adami

Keck Graduate Institute, 535 Watson Dr., Claremont, California, 91711, USA

Introduction

Consider any biological or engineering network (a protein-protein interaction network, the neuronal connectivity network of the nematode *C. elegans*, or a social network of friends). In the standard approach to network theory, we can analyze the network's structure and topology: observe a small-world property, study the degree distribution, or simply note that the network appears to be quite complex. But the human notion of complexity cannot distinguish a functional network, a random graph, or a randomized version of the functional network, where edges have been reconnected such that the degree distribution of the network is conserved. In what way then is a random graph different from a functional graph? Here we present a new method to quantify the complexity of a network by using motif frequencies in conjunction with information theory. We apply this measure to modular and bipartite artificially grown networks to demonstrate the potential of the measure, and then apply it to a number of well-known biological networks.

Results

Let us consider motifs of size n with a measured abundance distribution that translates into a probability distribution of motifs $p_i^{(n)}$ which is the probability to observe motif i among the M_n possible motifs of size n. We can then define an n-gram motif entropy

$$H^{(n)} = -\sum_{i=1}^{M_n} p_i^{(n)} \log_2 p_i^{(n)} . \qquad (1)$$

The amount of information per n-motif is then given by the difference between the maximal entropy $H_{\max}^{(n)} = \log_2 M_n$ and Eq. (1):

$$I^{(n)} = \log_2 M_n - H^{(n)} \qquad (2)$$

However, the maximal entropy $\log_2 M_n$ is dependent on the degree distribution, it is also clear that a good part of the information content is determined by the degree distribution itself. To correct for this, we can calculate the entropy of motifs under a randomization procedure that keeps the degree distribution intact. Counting the motif abundances for such a random network gives us an entropy $H_R^{(n)}$ and respectively the information content $I_R^{(n)} = \log_2 M_n - H_R^{(n)}$. However, because some of the structural information content is likely used for function, the purely functional information content remains ambiguous. As an alternative we can calculate the *relative entropy* (or Kullback-Leibler distance):

$$D^{(n)}(p \,\|\, q) = \sum_i p_i^{(n)} \log \frac{p_i^{(n)}}{q_i^{(n)}} , \qquad (3)$$

using the abundances of motifs $q_i^{(n)}$ obtained by randomizing the network while conserving the degree distribution. This measure of functional network complexity is positive, and vanishes if the motif distribution of the network studied is unchanged under the randomization procedure. To test this measure, we simulate a process that

creates complexity in a random graph, by growing scale-free modular networks using two types of nodes [1]. During growth, we allow edges to connect nodes of the same type with probability 1-*p* and nodes of different types with *p* (so that a value *p=0.5* creates completely random scale-free networks). At the extreme points we find networks very different from random, that is, structurally complex. For *p=1* nodes of the same type are always connected while for *p=0* we have no connection between nodes of the same type. If nodes of the same type form a module, then networks with *p=1* are perfectly modular while graphs with *p=0* are bipartite, or *anti-modular*. For the simplest case of motifs with depth *n=3* and undirected edges (and no node can connect to itself) we have two possible motifs (triangle and line). In highly modular (*p=1*) networks we find more triangles than lines, in contrast to bi-partite graphs where triangles are absent. Figure 1 shows a calculation of the network complexity in terms of the relative entropy as a function of *p*, demonstrating that only the modular and anti-modular networks are deemed complex using this measure. We also see that larger motifs (n>3) have a higher resolution of complexity.

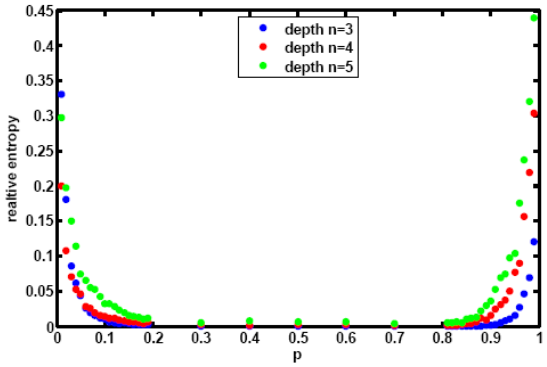

Figure 1: Relative complexity for various motif sizes. Networks with two different types of nodes were grown with different values of *p*.

Discussion

We introduce a new measure of complexity that utilizes motifs as symbols in an information-theoretic way to assess how much information is stored in a network. The information content is used as a proxy for complexity in analogy to information-theoretic measures of sequence complexity [2]. The general idea behind this approach is that any process (such as evolution) that influences a random distribution generates information. If this information reflects function, the information content is predictive of function, and hence complexity. The measure works well in computational test cases, but also in biological networks and their randomized versions.

Acknowledgements

We thank Alpan Raval and Jifeng Qian for discussions. This work was supported by the NSF's Frontiers in Integrative Research Program, Grant No. FIBR-0527023.

References

1. Hintze A, Adami C (2008) Universal growth model for random and modular networks with any degree distribution. Preprint: Keck Graduate Institute.
2. Adami C, Cerf NJ (2000) Physical complexity of symbolic sequences. Physica D 137: 62-69