Information content of colored motifs in complex networks

Christoph Adami[1,2,4,*] Jifeng Qian[1] Matthew Rupp[3,4]
Arend Hintze[1,3,4]

[1]*Keck Graduate Institute of Applied Life Sciences, 535 Watson Drive, Claremont, CA 91711*
[2]*Department of Microbiology and Molecular Genetics*
[3]*Computer Science and Engineering*
[4]*BEACON Center for the Study of Evolution in Action*
Michigan State University, East Lansing, MI 48824

April 4, 2011

Abstract

We study complex networks in which the nodes of the network are tagged with different colors depending on the functionality of the nodes (colored graphs), using information theory applied to the distribution of motifs in such networks. We find that colored motifs can be viewed as the building blocks of the networks (much more so than the uncolored structural motifs can be) and that the relative frequency with which these motifs appear in the network can be used to define the information content of the network. This information is defined in such a way that a network with random coloration (but keeping the relative number of nodes with different colors the same) has zero color information content. Thus, colored motif information captures the exceptionality of coloring in the motifs that is maintained via selection. We study the motif information content of the *C. elegans* brain as well as the evolution of colored motif information in networks that reflect the interaction between instructions in genomes of digital life organisms. While we find that colored motif information appears to capture essential functionality in the *C. elegans* brain (where the color assignment of nodes is straightforward) it is not obvious whether the colored motif information content always increases during evolution, as would be expected from a measure that captures network complexity. For a single choice of color assignment of instructions in the digital life form Avida, we find rather that colored motif information content increases or decreases during evolution, depending on how the genomes are organized, and therefore could be an interesting tool to dissect genomic rearrangements.

Keywords Network complexity, network motifs, colored motifs, Caenorhabditis elegans, information theory, digital evolution, Avida platform

*Contact author

1 Introduction

One of the most common strategies to understand a complex system is to analyze it in a hierarchical manner. For example, in biology, we attempt to unravel a cell's function by finding all of its parts and understanding how they relate to each other. Because the rules of interactions between cellular components are very complicated, the cell is much more than the sum of its parts: the parts and their interactions form a network, whose properties we can analyze. On the next level, an organism is much more than the sum of its cells, and a society of organisms, in turn, much more than the sum of its members. Thus, networks can be used to study social interactions between individuals also, allowing us to understand the dynamics of groups from the perspective of the mathematics of networks.

While the "science of networks" [10, 38, 37] has developed tremendously in the last ten years, a comparison of networks across different disciplines, or even of networks within one discipline (such as the protein-protein interaction networks of different organisms) has not really been possible except on the level of the connectivity patterns alone. The complexity of a network–or even perhaps its capacity to perform particular functions–is difficult to quantify, simply because complexity is a multi-faceted concept that as yet does not have an empirical basis. Many different approaches to quantifying complexity exist [3] (a non-exhaustive list is presented in Ref. [31]), ranging from assessing the complexity of a system's structure [32, 41, 42, 59, 58, 54, 9], or the complexity of the sequence giving rise to that structure [26, 27, 18, 30, 19, 6], to quantifying the *function* of the sequence or system [33, 57, 21]. What the measures of complexity have in common is that they all attempt to capture "that which increases when self-organizing systems organize themselves" [12].

If a network is a succinct description of any complex system, shouldn't a measure of network complexity be the concept that unifies attempts to attach a number to our intuitive understanding of complication? Unsurprisingly perhaps, a network's complexity appears to be as difficult to quantify as any other complex system. Several attempts exist in the literature [62, 34, 15, 63], reviewed in [25].

Here we develop and study a measure that attaches a number to a network so that it can be ranked and compared to other networks, and that allows us to track network evolution. This complexity measure is based on the theory of information, and is closely related to a measure that has been proposed to study the complexity of genes. Without a network complexity measure it is not possible to correlate complexity with function. Armed with such a measure, however, we should be able to understand for example how different types of networks react to damage, something that is important for molecular networks as for neural networks, ecological networks, or our cyber infrastructure.

Information is perhaps the central commodity of a technologically advanced society. We use information to order the world around us, make predictions about the world that allow us to function within it, and to encode our knowledge so that it can be passed on to future generations. But while information is an intuitive notion in our day-to-day life, it also has a precise mathematical formulation that meshes perfectly with our intuitive understanding. The theory of information due to Shannon [49] allows us to quantify the amount of information in a book, say, or on a CD or on the hard drive of a computer. It also allows us to study information transmission as well as ways

to protect information from noise. Because the theory of information is mathematical in nature, it applies to any information anywhere, in particular to the information stored in our genes. And while this information is not written in the ones and zeros of computers, or the letters of an alphabet, it is written in the language of biochemistry: the nucleotides A, C, G, and T or the twenty amino acids that proteins are made of.

We have recently described how the information content of genes can be measured from biological sequence information alone [6, 3, 4]. This information content is measured as the deviation from the expected sequence of a random gene, by recording the frequency with which each symbol appears at any particular position within the sequence. Thus, a highly conserved nucleotide at a particular sequence position indicates strong selection for function there (and thus high information content), while a position where each nucleotide appears with equal probability–the random expectation–stores little or no information about the function of that gene given the particular environment. Because the probability distribution of symbols in the sequence is shaped by Darwinian selection within the environment in which the organism that harbors that sequence lives, it is immediately clear that this information is necessarily *functional*, that is, useful to the organism. Qualitatively speaking, this information is used by the organism in order to make predictions about its environment that are better than chance [4]. In other words, we expect the information content of genes to correlate with fitness, which has been shown to be the case in at least two different computational systems [7, 24, 39, 22] and one biochemical one [14].

A network, if it describes a functioning entity (such as a cell, a brain, the internet, or a group of friends), can be seen as an information-rich structure. Clearly, the nodes and edges carry meaning in such a network, because a rearrangement of the nodes and edges would describe an entirely different system, or at least one with severely impaired function. This meaning, of course, is relative to the environment in which the network functions, just as the meaning of genes is context-dependent. How then can we measure the information that is stored in networks? Previous approaches have studied the information contained in degree-degree correlations (the assortativity of the network) to study how functional constraints affect network structure, for undirected [52] as well as directed [44] networks. Another information-theoretic approach focused on the entropy of randomized ensembles of networks constrained by degree distribution, degree correlation, and community structure [13]. Here, we take a different approach and instead of considering the degree distribution as the "degree of freedom" that provides entropy to a network, we study the *subnetworks* (sometimes called subgraphs, or motifs)[51, 36, 48] of a network that are obtained when we break up a network into its components, just as we break up a gene into its nucleotide alphabet.

There is some freedom in defining the "network alphabet": we can use subgraphs of two, three, four nodes, or more. Naturally, subgraphs with more nodes give rise to a network alphabet with more letters (motifs). But once we settle on an alphabet, we can obtain the frequency of each motif in the network, for example as in Fig. 1, where we illustrate the procedure of motif counting for a simple graph of six nodes only. Using the frequency of motifs we can estimate *motif probabilities* just as we can estimate the probability to find words in an English sentence using the frequency of words in a text. For the latter case, it is possible to estimate the information content of English text as compared to random sequences of words, an exercise that

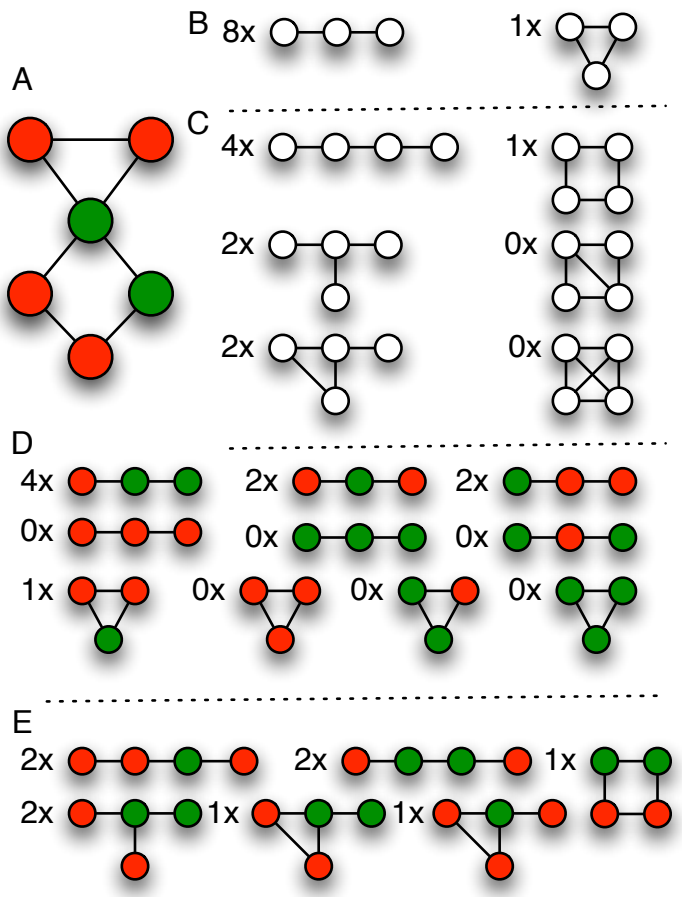

Figure 1: Motif counting. A: A six-node undirected example graph with two colors can be seen to be made from various motifs. B: For the size-three alphabet (motifs made from three nodes), we find two different structural motifs. C: For a size-four alphabet we see that two of the six possible structural motifs do not occur in the example graph. D: The colored motif frequencies (three node, two colors) for the network in A. E: Colored motif frequencies in the six-node graph shown in (A) for two colors and four nodes.

Shannon already conducted in 1951 [50]. If we assume that a random sequence of words contains no information, then the deviation from the uniform distribution (probability of each word of a given length appearing with equal probability) could be used to distinguish and perhaps classify functional (that is, meaningful) text from gibberish. Indeed, such a test was recently used to distinguish living from non-living matter both in biochemistry and in ALife [17]. In the same vein, it is possible that functional networks differ significantly from random networks in the subgraph utilization, and we can study this difference by estimating the *motif information content*. However, it is also not surprising that some of the differences in motif utilization across networks that have been noted previously [36, 35] could be due to constraints imposed by the degree distribution, or other constraints imposed by the growth-process of the network [23]. In other words, it is possible that the non-random "expression" of structural motifs could be a "spandrel" of cellular complexity [53].

But in fact, it is not difficult to see that networks contain more information than their topology (that is, the local patterns of connections) alone. Imagine, for example, a network of friends that know each other from high school, say, together with their friends. While we can learn a lot about common interests by looking at the clusters and the type of subgraphs that occur often, this approach assumes that all the nodes (and all the edges, for that matter) are qualitatively the same. However, more information can be gleaned from the network if we attach tags to each node or edge to classify the nodes or edges. For example, we can assign the tags male and female to each node, or we can assign the tag high-school and after-high-school to the edges that define the relationship between the nodes (referring to the time the two nodes became friends). If we color the graph according to these tags, the subgraph alphabet suddenly becomes much larger because each motif now comes in a variety of colorations. Here, we limit ourselves to colored nodes (leaving edges uncolored), and define an alphabet of colored motifs that we can use to calculate network information content. We show an example of colored motif counting for a six-node graph in Figure 1C and D, using only two colors.

2 Motif entropy and information

Entropy in information theory [16] is a measure of the uncertainty about the identity of objects in an ensemble. Let X be a random variable describing the structural (or topological) motifs of a network, given the size of motifs (different motif sizes define a different set of possible topological motifs). X can then take on the states $x_1, ..., x_N$, where N is the number of possible motifs of the given kind. Note that even when the number of nodes is fixed, the number of possible topological motifs still depends on the kind of edges that are allowed in the network (directed or undirected), and whether "self-edges" are allowed. If q_i are the probabilities to find motifs x_i, we can define the topological motif entropy as

$$H_{\rm top}(X) = -\sum_{i=1}^{N} q_i \log_2 q_i \ . \tag{1}$$

Each network has a particular topological motif entropy $H_{\rm top}(X)$ that reflects the

motif "utilization". We can determine whether the distribution of motifs in a network is functionally constrained, by randomizing the edges in the network (while keeping the edge distribution, for example, unchanged). Each randomization will create an instance $H^R_{\rm top}(X)$. The topological information content of the network would then be

$$I_{\rm top}(X) = \langle H^R_{\rm top}(X)\rangle - H_{\rm top}(X) \tag{2}$$

with $H_{\rm top}(X)$ from Eq. (1), and where $\langle H^R_{\rm top}(X)\rangle$ is the topological motif entropy averaged over different edge-randomizations of the network. This definition is formally the equivalent of the definition of the information content at a single nucleotide or residue site X [4].

If nodes can carry colors, they add an element of uncertainty even if the structure of the motif is given, because each particular topological motif can, given the possible colors that nodes can take on, appear in different colorations. Many of these colorations may be meaningless or downright detrimental for a functioning organism. We can quantify the functional constraints that affect colored motifs by studying the *color entropy* of a particular structural (topological) motif. If a particular structural motif x_i is now interpreted as a random variable Y_i that can take on the states $y^{(i)}_1, ..., y^{(i)}_M$ (its possible colorations) with probabilities $p^{(i)}_1, ..., p^{(i)}_M$, we can define the *color entropy* $H_{\rm color}(Y_i)$ of this motif by measuring how many times each of the colorations $y^{(i)}_j$ appears in the network:

$$H_{\rm color}(Y_i) = -\sum_{j=1}^{M} p^{(i)}_j \log_2 p^{(i)}_j \ . \tag{3}$$

The average color entropy of motifs in the network is then

$$H_{\rm color} = \sum_i q_i H_{\rm color}(Y_i) \ . \tag{4}$$

The total entropy of motifs, obtained by counting all possible colored motifs (within each class of motif sizes) is simply given by the sum of the topological and color entropy by virtue of the grouping axiom of information theory [11], i.e.,

$$H_{\rm total} = H_{\rm color} + H_{\rm top} \ . \tag{5}$$

However, this decomposition does not allow us to determine whether more information is stored in the topology or the functional assignment of nodes, because the baseline (unselected) distribution of motifs depends strongly on the method of randomization used. Furthermore, given a color assignment, an edge randomization automatically implies a color randomization, that is, color information and topological information cannot strictly be separated.

Nevertheless, we can calculate the information content of motif coloration by randomizing the colors in each network, while keeping the relative numbers of colors unchanged. In this way, we introduce $\langle H^R_{\rm color}\rangle$, which is calculated just as Eq. (3) but using a color-randomized version of the network, and averaged over a sufficient number of such randomizations. The color information content of the entire network is then simply

$$I_{\rm color} = \langle H^R_{\rm color}\rangle - H_{\rm color} \ . \tag{6}$$

2.1 Motifs in the *C. elegans* brain

To test these measures, we can analyze motifs in the network of synaptic and gap-junction connections of the neuronal network of the nematode *C. elegans*. This network controls one of the most well-understood complex biological systems to date, and most of the network architecture of the 302 neurons of the hermaphrodite worm is known from experimental work [61, 20] as well as recent reconstructions [60]. The most up-to-date wiring information covers 279 neurons of the somatic nervous system, excluding 20 neurons of the pharyngeal system and three neurons that appear to be unconnected from the rest [60]. There are 3,606 edges between these nodes, of which some (the synaptic connections) are directed, while gap-junctions are undirected. In our analysis of this network, we describe an undirected edge as a "bi-directional" edge, and also place bi-directional edges between nodes if synaptic connections run in both directions between the nodes.

2.1.1 Two-node colored motifs

In previous work that analyzed structural motifs only [56, 47, 55], the uni-directional two-node motif was found to be unremarkable (in the sense that the probability with which it was observed in the actual *C. elegans* network was not significantly different from the frequency observed in an edge-randomized version), while the bi-directional motif was deemed over-represented [47, 55]. We can look at both of those motifs in terms of the exceptionality of their colorations, by coloring neurons according to three possible functional tags, such as motorneuron (blue), sensor neuron (green), or interneuron (red).

We can study the functional constraints imposed on motifs by node function (color) by analyzing the constraints separately for each of the color realizations of the two motifs. In Fig. 2 we show the measured counts of each of the color realizations of the directed (Fig. 2A) and bi-directional (Fig. 2B) motifs. These distributions show that the observed functional constraints make intuitive sense. For example, the "S→I" (green→red) as well as "I→I" (red→red) motifs appear significantly more often then expected by chance, while the motif "M→S" (blue→green) is significantly suppressed: we do not expect muscles to relay information to sensory neurons in a functioning worm (even though some of these connections are indeed observed). So, while the uni-directional motif was unremarkable compared to an edge-randomized control, the motifs with "sensible" colorations such as sensor-→inter-neuron and inter-→inter-neuron are in fact highly significant, while non-sense pairs such as motor-→sensor-neuron are highly unlikely. Such an analysis can reveal motifs in the *C. elegans* brain that are used much more frequently than would be expected by chance, which can allow us to dissect the computational building blocks of the network [45].

In Fig. 3 we compare the color entropy $H_{\text{color}}(X)$ of the two structural motifs with two nodes to the distribution of $H^R_{\text{color}}(X)$ of 1,000 independent color randomizations of the same network (in color randomizations, the relative count of colors in the network is kept constant). We find that the color entropy of the *C. elegans* motifs of two nodes are significantly smaller than their randomized counterparts, a result that is particularly strong for the directed link motif in Fig. 3A. Thus, in terms of significant colorations, the uni-directional motif is more remarkable than the bi-directional motif. From those

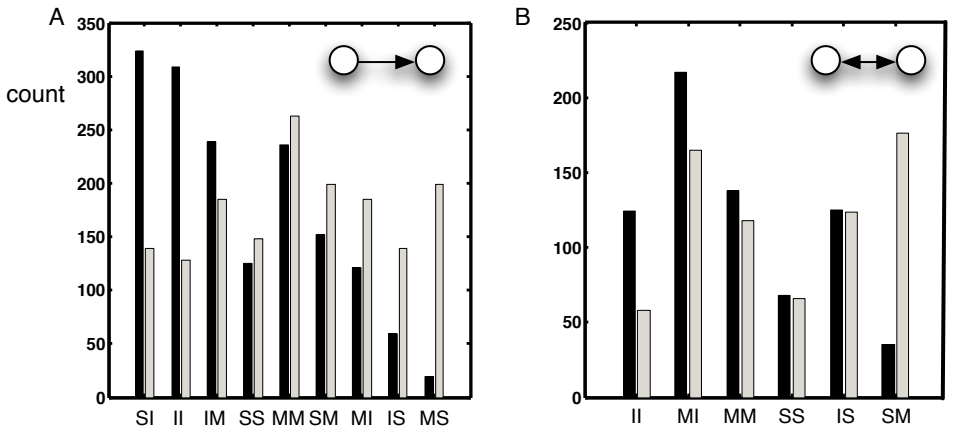

Figure 2: Histogram of abundances of directed structural motifs with particular coloration in *C. elegans* (black) compared to the average abundance in 1,000 color randomizations (grey). S: sensory neuron, I: interneuron, M: motor neuron. A: directed pairs (the direction of information flow is left-to-right: SI means S→I and so forth). B: bi-directional pairs.

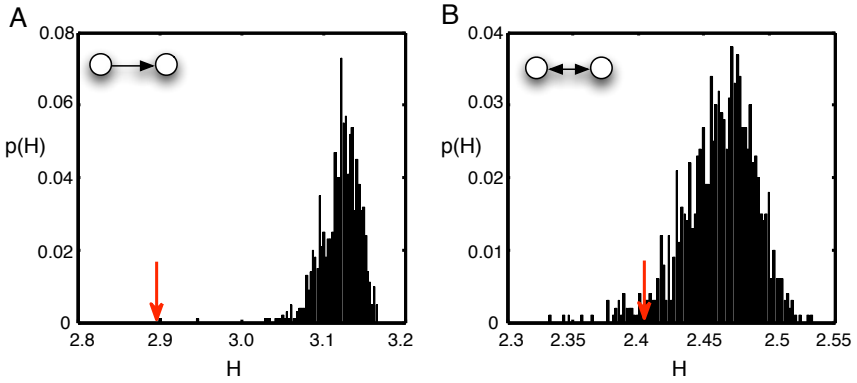

Figure 3: Distribution of color entropy for the two directed structural motifs with two nodes, obtained from 1,000 color randomizations of the *C. elegans* neuronal network. The color entropy of the actual *C. elegans* network $H_{\mathrm{col}}(X)$ is indicated by the arrow. A: unidirectional two-node motif, B: bi-directional two-node motif.

graphs, we can also estimate the color information content for each motif based on Eq. (6). We find for the information content of the uni-directional motif with two nodes $I_2^{\rm uni} = 3.15 - 2.9 = 0.35$ bits per symbol, while the color-information content of the bi-directional motif is significantly less: $I_2^{\rm bi} = 2.47 - 2.41 = 0.06$ bits per symbol.

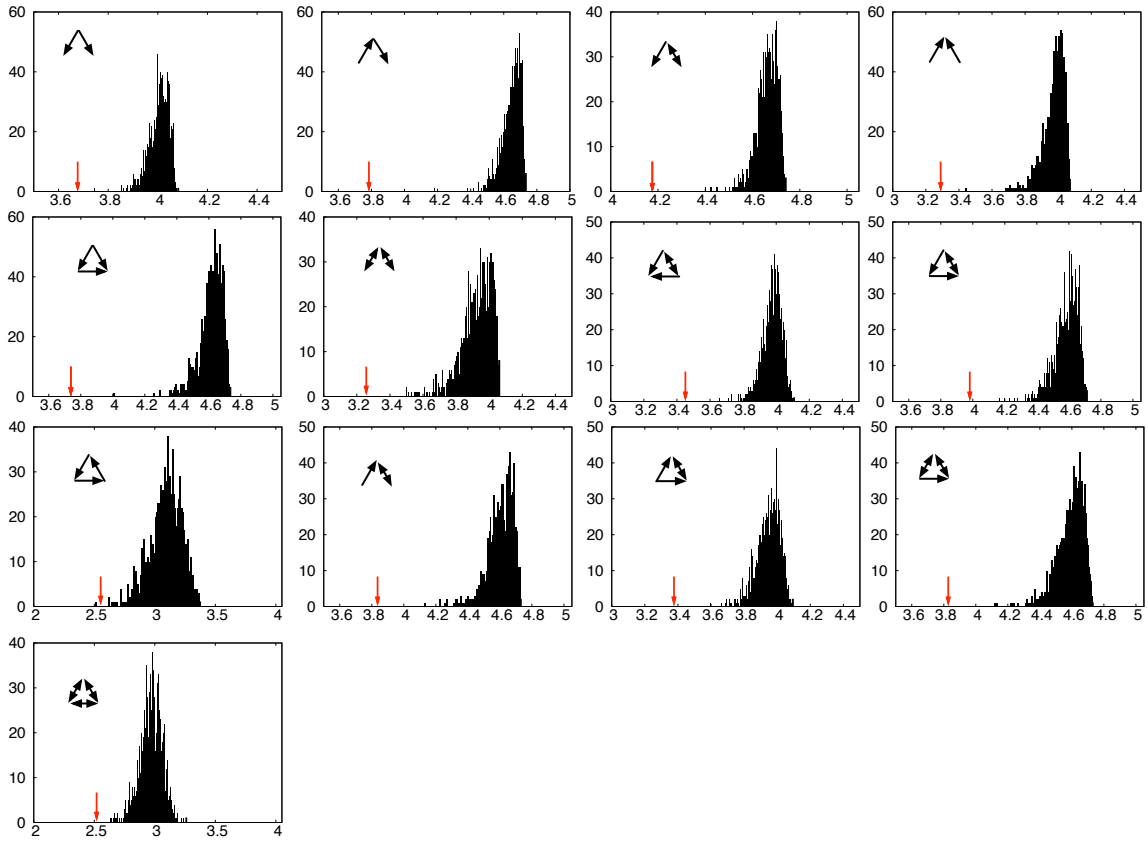

Figure 4: Distribution of color entropies for the 13 topologically different motifs of directed graphs of three nodes. The motif itself is identified in the upper left corner of each of the 13 histograms, which show the entropy on the x-axis and the count of how many times this entropy was observed in 1,000 randomizations of the network on the y-axis. The arrow indicates the color entropy of the actual *C. elegans* version of this motif, which in all cases is significantly lower than any of the entropies of the randomized networks, but the level of significance varies.

2.1.2 Three-node motifs

We can repeat the same analysis for motifs of size 3 (see Fig. 4). There are thirteen different structural motifs, whose color entropy can be measured for *C. elegans* and

compared to randomized-color controls. Fig. 4 shows that *all* three-node motifs in *C. elegans* have exceptional color combinations that reflect strong selective pressures on which motifs make sense within a functioning worm.

2.1.3 Entropy and information trends

Figures 3 and 4 indicate that each topological motif has a color entropy that is significantly lower than the average color entropy of that motif in a randomized network. But what is the average color entropy per motif, as a function of motif size? The average color entropy is about 2.8 bits for two-node motifs (averaged over the two types studied in Fig. 3), and increases slowly as the number of nodes increases (see Fig. 5, dash-dotted line). At the same time, the color entropy for a randomized graph starts at about 2.9 bits per symbol, but increases more quickly, indicating that the amount of functional (that is, color) information per symbol increases from about 0.1 bits per symbol (two-node motifs) to 1.2 bits (four-node motifs).

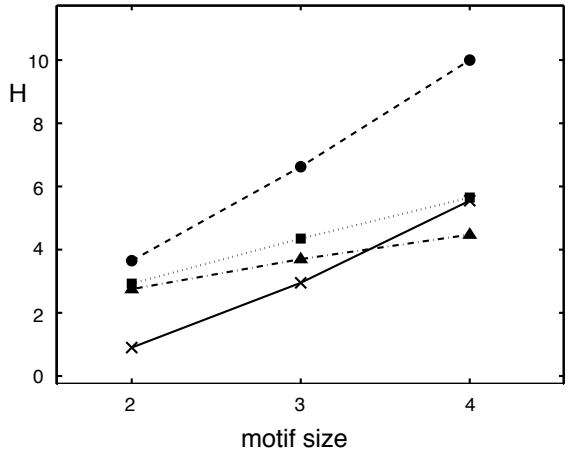

Figure 5: Motif entropies as a function of motif size. Average color entropy (▲, dash-dotted line), average randomized color entropy (■, dotted), topological entropy (✕, solid), and total entropy (●, dashed).

The analysis of information content in structural and colored motifs shows that information is stored in both topology and function, and that the information content depends on the size of the alphabet that is used. At the same time, it is clear that how we assign colors to nodes will also significantly affect information content. For example, other classifications of neuronal functions in the worm exist (such as into ten different morphological classes [1]). However, using more than a handful of colors can quickly make a computational analysis of colored motifs unwieldy because of the explosion in the number of motifs, and we do not expect to see dramatic changes in the information content once a meaningful set of colors is found for a network.

To study how color information changes as a network evolves, we have to use a different example than the worm brain, as it represents only one snapshot in time. To

study motif information evolution, we turn instead to Artificial Life.

3 Motifs in digital genomes

In Digital Life [46, 2], populations of self-replicating computer programs are adapting to a user-defined landscape, using a short instruction set of between 20-30 instructions. Since the initial implementation by Tom Ray in the tierra software, most digital life research has been carried out using the Avida platform (see, e.g., the Artificial Life Journal Special Issue [8], and [5]). Here, we use digital genomes evolved with Avida 2.8.1 (available from SourceForge.net) to create networks of interacting instructions. The 26 instructions in this experiment can be assigned to four different classes of instructions, as shown in Table 1.

Reproductive (Black)	Computational (Green)	Flow Control (Blue)	No-Ops (Red)
h-divide	IO	set-flow	nop-A
h-copy	nand	if-less	nop-B
h-alloc	swap	if-label	nop-C
	shift-l	get-head	
	shift-r	mov-head	
	push	jump-head	
	swap-stk	mov-head	
	pop	if-n-equ	
	add		
	sub		
	inc		
	dec		

Table 1: Functional and color assignment of the 26 Avida instructions. The class of "reproductive" instructions are involved solely in the management of inheritance, while the "computational" instructions play the role of "metabolic" instructions, as they are involved in harnessing the energy that avidians need to reproduce. "Flow control" instructions manage the information flow in the network, while the No-Op instructions are themselves inert, but typically modify the instruction (or instructions) just preceding it.

We evolve genomes in the standard "logic task" landscape, which rewards the performance of all one-input and two-input logical tasks with bonus CPU time depending on the difficulty of the task (there are nine distinct such tasks, see, e.g., [5]). The experiment is started with a population of 3,600 ancestral genomes with a length of 50 instructions that are only capable of self-replication (the self-replicating sequence is padded with the `nop-C` instruction to arrive at the sequence length of 50). The population evolves for 100,000 updates (a measure of time within which each sequence in the population has 30 of its instructions executed), with a mutation rate of 0.0025 per

instruction per copy-event (no cross-over), and an instruction-insert and delete probability of 5% in mass-action mode (well-mixed chemostat). Because of the insert/delete probability, the sequence length is not constant, but instead increases slightly during evolution to 56 instructions. Fig. 6A shows the evolution of fitness as a function of updates, on the *line of descent* (LOD) of the population. The line of descent is created by picking a representative of the most fit genotype of the population at the end of the experiment, and tracing its lineage backwards in time via its direct ancestors, ending at the seed genotype. Because these populations evolve in a single niche, the LODs of all genotypes present in the population at the end of the experiment quickly coalesce, so that a single LOD characterizes the evolutionary dynamics of the experiment [29]. Analysis of evolutionary experiments in terms of the LOD (rather than population averages) has the advantage of recapitulating the salient events in evolutionary history, while disregarding any changes that did not leave a trace in the final product. In that manner, the LOD allows for a reconstruction of the path that evolution took to arrive at the adapted sequence.

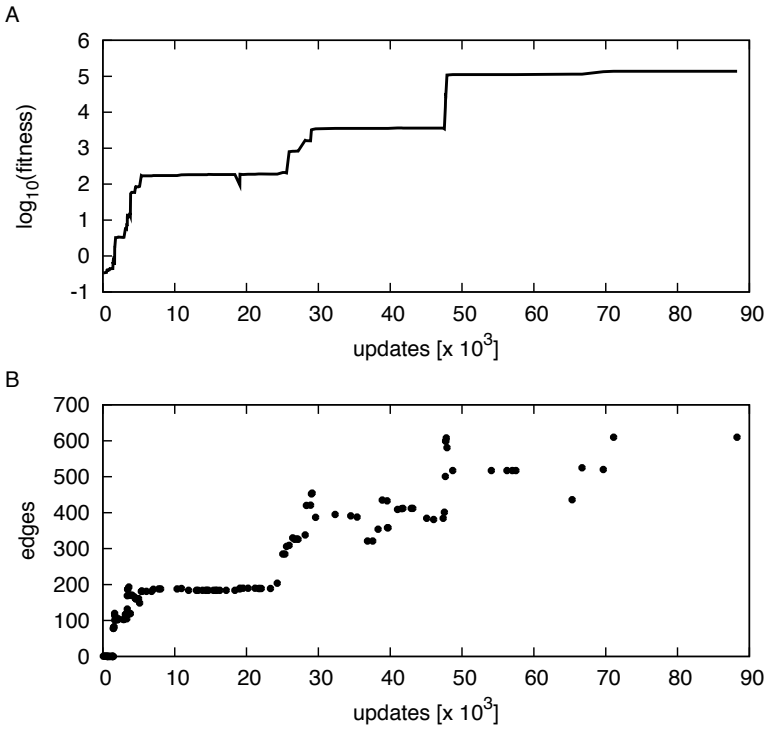

Figure 6: Fitness, and number of epistatic edges on the line of descent. A: Fitness for 138 genotypes on the LOD. B: Number of edges between interacting instructions with epistasis $|\epsilon_{ij}| > 2$, on the LOD.

3.1 Epistasis

We determine that two instructions within an avidian sequence interact if the fitness effects of knocking out these instructions depend on each other, that is, if the fitness contribution of one instruction is contingent on the identity of the other. Two instructions that are linked in such a way are called *epistatic* [64] (see also the review [43]). Epistasis is an important concept within evolutionary genetics, and is usually defined to quantify the interaction between genes [40] rather than between the set of monomers that code for the gene, but can easily be used the way we show here [28]. It has been shown earlier that complex genomes show more positive epistasis between deleterious mutations than simple ones [28] (a finding we corroborate here), and that epistasis between avidian instructions is crucial to understand the evolution of complex features [29].

Instruction knockouts are performed by replacing each instruction by an inert instruction `nop-X`, in order to prevent fitness effects that are due to a change in sequence length only rather than the identity of the instruction. For each sequence on the LOD, we can calculate the epistasis ϵ_{ij} for any pair of mutations at instruction sites i and j as follows. Let the unmutated (that is wild-type) fitness of the sequence be w_0. Here, fitness is measured as the rate at which an avidian produces offspring per generation, and is equivalent to the growth rate of more conventional organisms. The fitness effect of mutating instruction i then is w_i/w_0. On the LOD, we find many substitutions that are neutral or beneficial, however, most knockouts of arbitrary instructions are either neutral or deleterious. After creating the mutant with fitness w_i, mutate another instruction j to obtain the double-mutant with fitness w_{ij}. At the same time, revert mutation i on the double mutant to obtain a genome with only the single mutation j, with fitness w_j. This is sufficient to compare the two single-mutant effects w_i/w_0 and w_j/w_0 with the effect of the double mutant w_{ij}/w_0. The quantity

$$\epsilon_{ij} = \log_e \left(\frac{w_{ij} w_0}{w_i w_j} \right) \tag{7}$$

then measures the epistasis between the two instructions i and j (see, e.g., [40] and references therein). Positive epistasis between mutations implies that the fitness of the double mutant is higher than we would have expected from the effect of each single mutation, while negative epistasis signifies that the double mutation has made things worse than either of the single mutations would have led you to believe. A typical example of genetic (epistatic) interaction is a pair of redundant instructions, where each of the mutations by themselves does not affect organism fitness, while the mutation of both instructions creates a fitness deficit. In this case, the epistasis is clearly negative. This effect is called "synthetic lethality" (if the double mutant is non-viable) in the genetic literature. The opposite case can also occur, when the knockout of one gene compensates for the loss of function due to the knockout of another, but the effect is less common. In general, more interactions in avidians are of the positive sort [28], simply because a second mutation that affects the same functional block as the first has virtually no effect anymore. As a consequence, groups of epistatically connected instructions often outline functional blocks or modules.

In Fig. 7, we see avidian sequences at different time points in their evolution, with instructions colored according to the functional tags defined in Table 1, and edges

indicating epistatic interactions for pairs of instructions i and j if $|\epsilon_{ij}| > 2$. With this cutoff, there are no interactions between instructions in the ancestral genotype (Fig. 7A), but they start to emerge around update 1,450. Note that even though the fitness rises exponentially, epistasis is defined in terms of the relative effect of a knockout on fitness, and we should not expect a priori that the number of edges should increase as fitness increases. The cutoff $|\epsilon_{ij}| > 2$ is quite stringent: it implies that the double mutants fitness effect must be more than e^2 times the product of the fitness effects of the single mutations (for positive epistasis), or less than $\approx 13.5\%$ of the product (for negative epistasis).

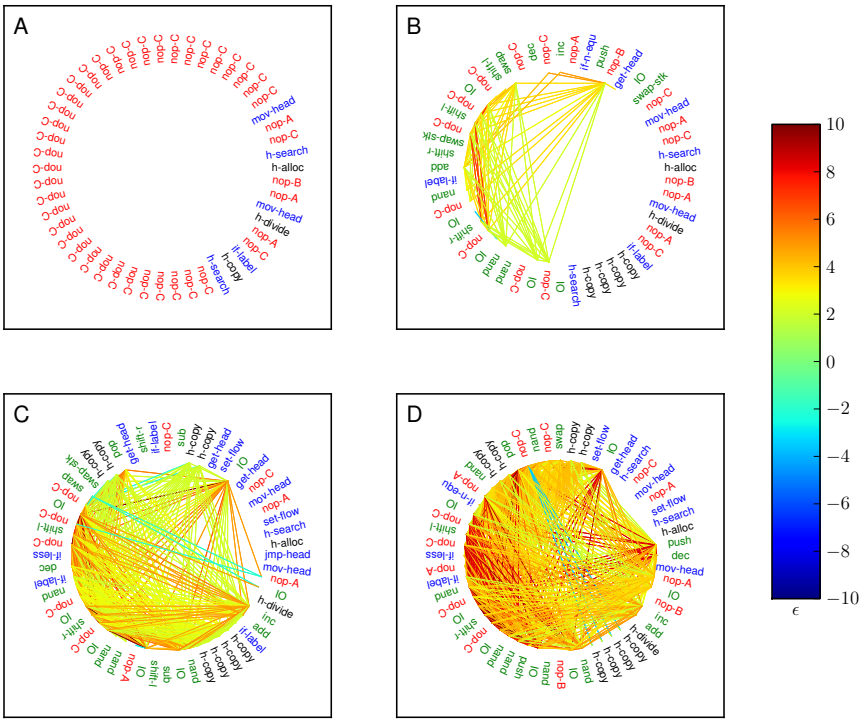

Figure 7: Epistatic interactions between instructions for four genomes on the LOD. Instructions are colored according to the scheme detailed in Table 1, while epistatic edges are colored according to their strength and direction, in a graded manner between blue ($\epsilon = -10$) over green (vanishing epistasis, not shown in these plots because of the threshold) to red ($\epsilon = 10$, see color bar). A: Ancestral genome (50 instructions). B: A genome early on the LOD (at update 3,742). C: Genotype on the LOD at update 29,035. D: Epistatic network for the last genotype on the LOD, at update 88,297, with 610 edges.

3.2 Motif entropy and information

For the epistatic colored networks of the type shown in Fig. 7, we can calculate colored motif entropies, structural motif entropies, and colored motif information content as described above. We focus here on motifs of size four, which are undirected. There are six structural motifs of size four (shown in Fig. 1C), which come in a total of 566 different colorations. Thus, the maximal total motif entropy is $H_{\rm tot}^{\rm max} \approx 9.14$ bits. We count the frequency of colored motifs in the network as described for the *C. elegans* network, and also calculate the mean frequency with which we observe that motif in a color-randomized network. In Fig. 8A, we show the topological entropy calculated using Eq. (1), the color entropy [Eq. (4)], as well as the total entropy [from Eq. (5)], along the LOD of the experiment. Generally, entropies are increasing as the network evolves, not because the network increases in size but because the number of edges is increasing (as seen in Fig. 6B), which leads to a greater diversity of colored motifs. However, the genetic changes that give rise to the fitness jump around 50,000 updates (see Fig. 6A) appear to change the genetic architecture in such a manner that color diversity *decreases* somewhat. This decrease is most apparent in the color entropy, less so in the structural motif entropy.

In order to calculate the information stored in the color assignment of instructions, we need to calculate the average color entropy for color randomized networks. We do this as described above for the *C. elegans* motifs, to obtain $\langle H_{\rm color}^R \rangle$. This entropy also increases with evolutionary time, and as a consequence the difference between the two is mostly constant, but also shows a decrease at least for some periods of time on the LOD (see Fig. 8). It is not immediately clear what kind of changes in the genetic architecture of the sequences is responsible for the drop or increase in motif information content. However, because the network is comparatively small, small changes in the genome can potentially give rise to large changes in the colored motif distribution. Note that the color entropy for 1000 color-randomized networks has an error in the mean that is much smaller than the changes seen on the LOD (between 0.02 for the earliest networks to 0.005 for the fittest ones), indicating that the fluctuations are not due to sampling error. We conclude that the color assignment (shown in Table 1) that we chose for the instructions shows that some information is stored within the colored motifs in the epistatic network, but that this information does not necessarily increase with an increase in fitness. In particular, it is possible that a different choice of color assignments captures *more* motif information, and correlates differently with fitness. Thus, while the genomic information content [6, 7, 4] correlates very well with fitness (see also [24]), the colored motif information content appears to be better suited to track changes in the genome architecture and organization.

4 Discussion

We have shown how an information-theoretic analysis of networks in which nodes are assigned a color based on their functionality allows us to determine the information content of the network motifs in a manner that significantly expands the purely topological treatment. The method is general and can be applied to any network where both structural information (connectivity) and functional annotation of the nodes is

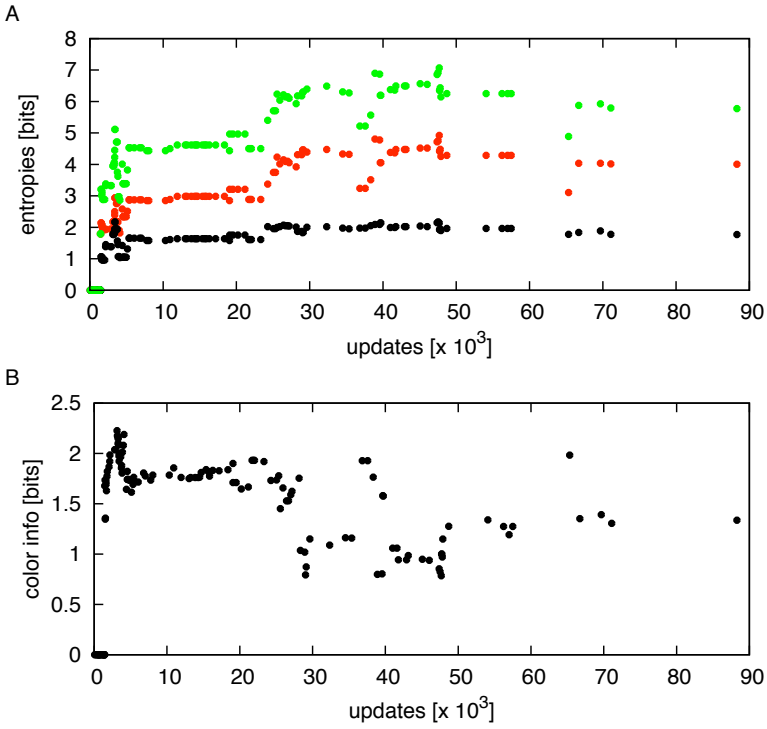

Figure 8: Motif entropies and colored motif information content. A: Motif entropies for the 138 genotypes on the LOD. Black dots: topological entropy of motifs of size 4 (maximal entropy for 6 motifs is 2.585 bits). Red dots: average color entropy of motifs of size 4. Green dots: Total motif entropy [Eq. (5)], given by the sum of topological and color entropy, according to the grouping axiom. B: Information content (per motif) in colored motifs of size 4, according to Eq. (6).

available. When considering the neuronal network of *C. elegans* as an example, we note that (depending on what size motifs we consider) more information is stored in the coloration of the motifs than in the structure. Indeed, an analysis of the *C. elegans* brain in terms of structural motifs has generated only limited insight [47, 55], while adding the color degree of freedom creates a wealth of information about what computations are performed by the worm's brain [45]. An analysis of the information stored in motif colorations shows that the information per symbol increases with the size of the motifs considered, but while it is clear that this information is a consequence of selection (because it is precisely the difference between a random color entropy and that selected by evolution) it is not clear how this information changes as the organism adapts. To address this question, we have analyzed the information content of colored motifs in networks created by the interaction between instructions of avidian genomes. By choosing a particular functional coloring of instructions (here 4 colors tagging instructions that have either a biological, a computational, a flow-control, or a modifying function), we discover that while information is stored in the colorations, this information neither has to increase nor decrease with adaptation. While the number of epistatic edges increase as the organism adapts to its environment, the colored motif distribution (while clearly constrained by the functionality of the sequence) can become more narrow or more broad, depending on the genetic architecture of the sequence that gives rise to them. We do see clear indications that the distribution changes at stages in which new functionality is evolved, which points to a relation between genomic architecture and colored motif distribution, but monitoring the information content alone is not sufficient to dissect what these changes are.

Of course, no general conclusions about the evolution of colored motif information in complex networks can be drawn from this single example, not because a single experiment would not be reflective of the average evolutionary trajectory (we believe it is in the present case) but rather because the assignment of colors to functions of the instructions reflects the investigator's intuition, but is not necessarily the assignment that maximizes the information content. Thus, while it is clear from the example we studied here that information content of colored motifs cannot be a universal measure of network complexity independently of what the color assignment is (or how edges are defined), it is nevertheless a promising tool for dissecting the functional complexity of a network. It is interesting to ask whether a search over possible colorations (using a limited number of colors, of course) looking for that coloration that maximizes the information content of motifs could generate insight into the functionality of instructions (and their dependence) that is not obvious from the outset.

5 Acknowledgements

We would like to thank Charles Ofria and Bjørn Østman for discussions. This work was supported in part by the National Science Foundation's BEACON Center for the Study of Evolution in Action, under contract No. DBI-0939454 and by the Agriculture and Food Research Initiative Competitive Grant no. 2010-65205-20361 from the USDA National Institute of Food and Agriculture. We wish to acknowledge the support of the Michigan State University High Performance Computing Center and the Institute

for Cyber Enabled Research.

References

[1] Achacoso, T. & Yamamoto, W. (1992). *AY's Neuroanatomy of C. elegans for computation*. Boca Raton: CRC Press.

[2] Adami, C. (1998). *Introduction to Artificial Life*. New York: Springer Verlag.

[3] Adami, C. (2002). What is complexity? *BioEssays*, *24*, 1085–94.

[4] Adami, C. (2004). Information theory in molecular biology. *Physics of Life Reviews*, *1*, 3–22.

[5] Adami, C. (2006). Digital genetics: unravelling the genetic basis of evolution. *Nature Reviews Genetics*, *7*, 109–118.

[6] Adami, C. & Cerf, N. J. (2000). Physical complexity of symbolic sequences. *Physica D*, *137*, 62–69.

[7] Adami, C., Ofria, C., & Collier, T. (1999). Evolution of biological complexity. *Proc. Natl. Acad. Sci. USA*, *97*, 4463–4468.

[8] Adami, C. & Wilke, C. (2004). Experiments in digital evolution (Editors' introduction to the special issue). *Artificial Life*, *10*, 117–122.

[9] Ahnert, S. E., Johnston, I. G., Fink, T. M. A., Doye, J. P. K., & Louis, A. A. (2010). Self-assembly, modularity, and physical complexity. *Physical Review E*, *82*, 026117.

[10] Albert, R. & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.

[11] Ash, R. B. (1965). *Information Theory*. New York, N.Y.: Dover Publications, Inc.

[12] Bennett, C. (1995). Universal computation and physical dynamics. *Physica D*, *86*, 268–273.

[13] Bianconi, G. (2008). The entropy of randomized network ensembles. *Europhysics Letters*, *81*, 28005.

[14] Carothers, J. M., Oestreich, S. C., Davis, J. H., & Szostak, J. W. (2004). Informational complexity and functional activity of RNA structures. *J. American Chem. Society*, *126*, 5130–5137.

[15] Claussen, J. C. (2007). Offdiagonal complexity: A computationally quick complexity measure for graphs and networks. *Physica A*, *375*, 365–373.

[16] Cover, T. M. & Thomas, J. A. (1991). *Elements of Information Theory*. New York, NY: John Wiley.

[17] Dorn, E. D., Nealson, K. H., & Adami, C. (2011). Monomer abundance patterns as a universal biosignature: Examples from terrestrial and artificial life. *Journal of Molecular Evolution*, *72*, 283–295.

[18] Ebeling, W. & Jimenez-Montano, M. (1980). On grammars, complexity, and information measures of biological macromolecules. *Mathematical Biosciences*, *52*, 53–71.

[19] Gell-Mann, M. & Lloyd, S. (1996). Information measures, effective complexity, and total information. *Complexity*, *2*, 44–52.

[20] Hall, D. & Russell, R. (1991). The posterior nervous system of the nematode caenorhabditis elegans: Serial reconstruction of identified neurons and complete pattern of synaptic interactions. *J. Neuroscience*, *11*, 1–22.

[21] Hazen, R. M., Griffin, P. L., Carothers, J. M., & Szostak, J. W. (2007). Functional information and the emergence of biocomplexity. *Proc Natl Acad Sci U S A*, *104 Suppl 1*, 8574–81.

[22] Hintze, A. & Adami, C. (2008). Evolution of complex modular biological networks. *PLoS Computational Biology*, *4*, e23.

[23] Hintze, A. & Adami, C. (2010). Modularity and anti-modularity in networks with arbitrary degree distribution. *Biol Direct*, *5*, 32.

[24] Huang, W., Ofria, C., & Torng, E. (2004). Measuring biological complexity in digital organisms. In J. Pollack, M. A. Bedau, P. Husbands, T. Ikegami, & R. Watson (Eds.) *Proceedings of Artificial Life IX*. Cambridge, MA: MIT Press, (pp. 315–321).

[25] Kim, J. & Wilhelm, T. (2008). What is a complex graph? *Physica A*, *387*, 2637–2652.

[26] Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, *1*, 4.

[27] Lempel, A. & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions On Information Theory*, *22*, 75–81.

[28] Lenski, R. E., Ofria, C., Collier, T. C., & Adami, C. (1999). Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, *400*, 661–664.

[29] Lenski, R. E., Ofria, C., Pennock, R. T., & Adami, C. (2003). The evolutionary origin of complex features. *Nature*, *423*, 139–44.

[30] Li, M. & Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. New York, NY: Springer Verlag.

[31] Lloyd, S. (2001). Measures of complexity: A nonexhaustive list. *IEEE Control Systems Magazine*, *21*, 7–8.

[32] Löfgren, L. (1977). Complexity of description of systems: A foundational study. *Int. J. Gen. Sys.*, *3*, 197–214.

[33] McShea, D. W. (2000). Functional complexity in organisms: Parts as proxies. *Biology and Philosophy*, *15*, 641–668.

[34] Meyer-Ortmanns, H. (2004). Functional complexity measure for networks. *Physica A*, *337*, 679–690.

[35] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., & Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, *303*, 1538–42.

[36] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, *298*, 824–7.

[37] Newman, M., Barabasi, A.-L., & Watts, D. (2006). *The Structure and Dynamics of Networks*. Princeton, N.J.: Princeton University Press.

[38] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, *45*, 167–256.

[39] Ofria, C., Huang, W., & Torng, E. (2008). On the gradual evolution of complexity and the sudden emergence of complex features. *Artif Life*, *14*, 255–63.

[40] Østman, B., Hintze, A., & Adami, C. (2011). Impact of epistasis and pleiotropy on evolutionary adaptation. Preprint arXiv:0909.3506 on arxiv.org.

[41] Papentin, F. (1980). On order and complexity I: General considerations. *J. theor. Biol.*, *87*, 421–456.

[42] Papentin, F. (1982). On order and complexity II: Application to chemical and biochemical structures. *J. theor. Biol.*, *95*, 225–245.

[43] Phillips, P. C. (2008). Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, *9*, 855–867.

[44] Piraveenan, M., Prokopenko, M., & Zomaya, A. (2010). Assortative mixing in directed biological networks. *IEEE/ACM Trans Comput Biol Bioinform.*

[45] Qian, J., Hintze, A., & Adami, C. (2011). Colored motifs reveal computational building blocks in the *C. elegans* brain. *PLoS ONE*, *6*, e17013.

[46] Ray, T. S. (1992). An approach to the synthesis of life. In C. G. Langton, J. D. Farmer, & S. Rasmussen (Eds.) *Artificial Life II*. Redwood City: Addison-Wesley, (pp. 371–408).

[47] Reigl, M., Alon, U., & Chklovskii, D. B. (2004). Search for computational modules in the C. elegans brain. *BMC Biol*, *2*, 25.

[48] Rice, J. J., Kershenbaum, A., & Stolovitzky, G. (2005). Lasting impressions: Motifs in protein-protein maps may provide footprints of evolutionary events. *Proc Natl Acad Sci U S A*, *102*, 3173–4.

[49] Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423,623–656.

[50] Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*, 50–64.

[51] Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, *31*, 64–68.

[52] Solé, R. & Valverde, S. (2004). Information theory of complex networks: On evolution and architectural constraints. *Lect. Notes Phys.*, *650*, 189–207.

[53] Solé, R. V. & Valverde, S. (2006). Are network motifs the spandrels of cellular complexity? *Trends Ecol Evol*, *21*, 419–22.

[54] Soloveichik, D. & Winfree, E. (2006). Complexity of self-assembled shapes. *SIAM Journal On Computing*, *36*, 1544–1569.

[55] Song, S., Sjöström, P. J., Reigl, M., Nelson, S., & Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol*, *3*, e68.

[56] Sporns, O. & Kötter, R. (2004). Motifs in brain networks. *PLoS Biol*, *2*, e369.

[57] Szostak, J. W. (2003). Functional information: Molecular messages. *Nature*, *423*, 689.

[58] Thomas, R. D., Shearman, R. M., & Stewart, G. W. (2000). Evolutionary exploitation of design options by the first animals with hard skeletons. *Science*, *288*, 1239–1242.

[59] Thomas, R. D. K. & Reif, W.-E. (1993). The skeleton space: A finite set of organic designs. *Evolution*, *47*, 341–360.

[60] Varshney, L. R., Chen, B. L., Paniagua, E., Halland, D. H., & Chklovskii, D. B. (2011). Structural properties of the *Caenorhabditis elegans* neuronal network. *PLoS Computational Biology*, *7*, e1001066.

[61] White, J., Southgate, E., Thomson, J., & Brenner, S. (1986). The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos Trans R Soc Lond Biol Sci*, *314*, 1–340.

[62] Wilhelm, T. (2003). An elementary dynamic model for non-binary food webs. *Ecological Modelling*, *168*, 145–152.

[63] Wilhelm, T. & Hollunder, J. (2007). Information theoretic description of networks. *Physica A*, *385*, 385–396.

[64] Wolf, J., Brodie, E., & Wade, M. (Eds.) (2000). *Epistasis and the Evolutionary Process*. Oxford: Oxford University Press.