96. Cortassa SC, Aon MA, O'Rourke B, Jacques R, Tseng HJ, Marban E, Winslow RL (2006) A computational model integrating electrophysiology, contraction and mitochondrial bioenergetics in the ventricular myocyte. Biophys J 91:1564–1589
97. Saks V, Dzeja P, Schlattner U, Vendelin M, Terzic A, Wallimann T (2006) Cardiac system bioenergetics: metabolic basis of the Frank-Starling law. J Physiol 571:253–273

Books and Reviews

Bak P (1996) How nature works: the science of self-organized criticality. Copernicus, New York
Barabasi AL (2003) Linked. Plume, New York
Capra F (1996) The web of life. Anchor books Doubleday, New York
Dewey GT (1997) Fractals in molecular biophysics. Oxford University Press, New York
Cortassa S, Aon MA, Iglesias AA, Lloyd D (2002) An Introduction to Metabolic and Cellular Engineering. World Scientific, Singapore
Feder J (1988) Fractals. Plenum Press, New York
Glass L, Mackey MC (1988) From clocks to chaos. The rhythms of life. Princeton University Press, Princeton
Haken H (1978) Synergetics. Springer, Heidelberg
Jantsch E (1989) The self-organizing universe. Scientific and human implications of the emerging paradigm of evolution. Pergamon Press, Oxford
Kauffman SA (1993) Origins of order: Self-organization and selection in evolution. Oxford University Press, New York
Lane N (2005) Power, sex, suicide. Mitochondria and the meaning of life. Oxford University Press, Oxford
Noble D (2006) The music of life. Oxford University Press, New York
Sornette D (2003) Why stock markets crash. Critical events in complex financial systems. Princeton University Press, New Jersey
Varela FJ (1989) Autonomie et connaissance. Seuil, Paris
West BJ, Deering B (1995) The Lure of Modern Science. Fractal Thinking, vol 3. World Scientific, Singapore
Yates EF (1987) Self-organizing systems. The emergence of order. Plenum Press, New York

Biological Complexity and Biochemical Information

Christoph Adami
Keck Graduate Institute of Applied Life Sciences, State University of New York, Claremont, USA

Article Outline

Glossary

C-value The haploid genome size of an organism, measured either in picograms (pg) or base pairs (bp).

Degree distribution The probability distribution $P(d)$ to find a node with d edges in a network.

Entropic profile A graph of the per-site entropy along the sites of a biomolecular sequence, such as a DNA, RNA, or protein sequence.

Epistasis Generally, an interaction between genes, where the fitness effect of the modification of one gene influences the fitness effect of the modification of another gene. More specifically, an interaction between *mutations* that can be either positive (reinforcing or *synergistic*), or negative (mitigating or *antagonistic*).

Erdös–Rényi network A random graph with a binomial degree distribution.

Fitness A numerical measure predicting the long-term success of a lineage.

Jensen–Shannon divergence In probability and statistics, a measure for the similarity of probability distributions, given by the symmetrized relative entropy of the distributions.

Module In network theory, a group of nodes that is closely associated in connections or function, but only weakly associated to other such groups.

Motif In network theory, a subgraph of small size.

Network diameter For networks, the average geodesic distance between nodes, defined as $D = 1/m \sum_{i=1}^{n} \sum_{j=1}^{n} d(i, j)$, where m is the number of edes of the graph, n is the number of nodes, and $d(i, j)$ is the shortest path distance between nodes i and j.

Phylogenetic depth A measure of the genetic distance between a genome and its ancestor on the same line of descent, given by the number of genetically different genomes on the line between the genomes plus one.

Random variable In probability and statistics, a mathematical object with discrete or continuous states that the object takes on with probabilities drawn from a probability distribution associated to the random variable.

Source entropy The entropy of a sequence generated by a process that generates symbols with a given probability distribution.

Wright–Fisher process In population genetics, a stochastic process that describes how genes are transmitted from one generation to the next.

Turing machine In mathematics, an abstract automaton that manipulates symbols on a tape directed by a finite set of rules.

Watson-Crick pairing In biochemistry, the pairing be-

tween nucleotides adenine and thymine (A-T), and guanine and cytosine (G-C).

Zipf's law A relationship between the frequency f and the rank k of words in a text, of the form $f(k) \sim k^s$, where s is the exponent of the distribution.

Definition of the Subject

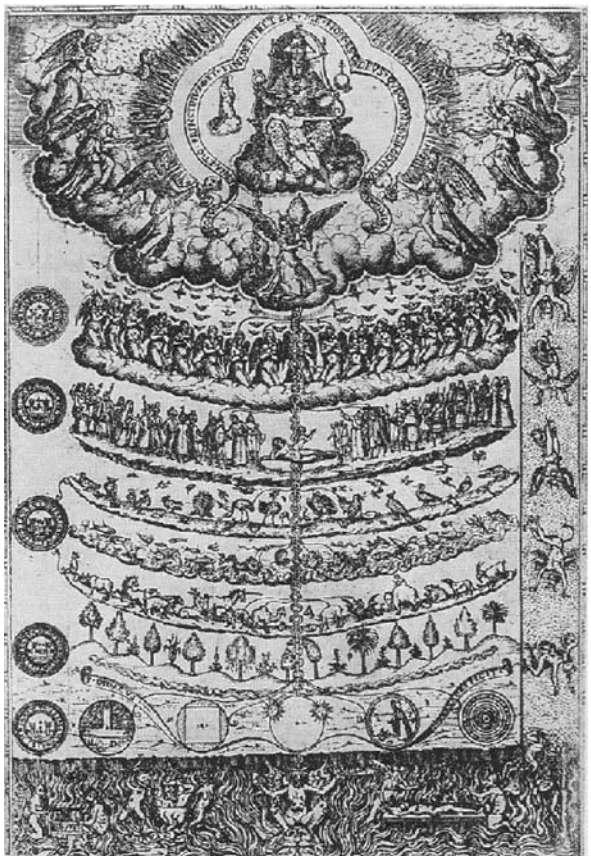
Biological complexity refers to a measure of the intricateness, or complication, of a biological organism that is directly related to that organism's ability to successfully function in a complex environment. Because organismal complexity is difficult to define, several different measures of complexity are often used as proxies for biological complexity, such as structural, functional, or sequence complexity. While the complexity of single proteins can be estimated using tools from information theory, a whole organism's biological complexity is reflected in its set of expressed proteins and its interactions, whereas the complexity of an ecosystem is summarized by the network of interacting species and their interaction with the environment.

Introduction

Mankind's need to classify the world around him is perhaps nowhere more apparent than in our zeal to attach to each and every living organism a tag that reveals its relationship to ourselves. The idea that all forms of life (and even inanimate matter) belong to a "Great Chain of Being" goes back to medieval times [1] and usually depicts rocks at the bottom of the chain, with plants next, followed by fish, birds, mammals, humans, angels, and ultimately god (see Fig. 1).

In more modern times, biologists and lay people alike have, sometimes unconsciously, given in to the same need by classifying organisms according to their perceived complexity. So, for example, viruses are often perceived as the least complex organisms, followed by bacteria, unicellular eukaryotes, fungi, plants, invertebrates, vertebrates, mammals, and, ultimately, ourselves. Such a hierarchical view of biological complexity with *Homo sapiens* at its top has repeatedly been decried (see, e. g., [2] and recently [3]). However, in the absence of a quantitative measure of biological complexity, neither a position in favor nor against a hierarchical ordering of species can authoritatively be maintained. Indeed, even though it has become less and less accepted to view *Homo* as the "crown jewel" of evolution, there is no reason a priori why this cannot in fact be true.

Many different measures for biological complexity have been introduced in the past literature, some of them



Biological Complexity and Biochemical Information, Figure 1
A depiction of the Great Chain of Being by Fra Diego de Valdes, in *Rhetorica Christiana, from 1579*

with obvious biological roots (such as number of cells, number of different tissue types, genome length, etc.), some of them inspired by the physics of dynamical systems. Neither of these measures has managed to convince a majority that it represents the best proxy for a biological complexity measure, and most have very obvious shortcomings. As we learn more about the genome and proteome of different organisms, it seems as if we are well on our way to accumulate a sufficient amount of detailed knowledge concerning the function of organisms that a universal measure of organism complexity is at last imaginable. For example, the set of all expressed proteins and their interactions (including their regulation and post-translational modification) could conceivably enter into an equation whose result is a number: the biological (functional) complexity of an organism that allows for comparisons across the tree of life.

Such a fantasy, however, is almost surely misguided unless we can enumerate along with the list of proteins and their interactions the possible environments within which

the organism can make a living. After all, the complexity of an organism must depend crucially on the environment within which it functions. Consider as an example a 150 pound rock and an average human, transplanted to the surface of the moon. Functionally, they are very comparable there. Likewise, a human is much less complex than even an algal bloom, if forced under water for an extended period of time. Given that we are unable to characterize even a single environment for most bacteria (environmental microbiologists estimate that fewer than 2% of all bacteria can be cultured in the laboratory) it is highly unlikely that the full functional complexity of most organisms can be ascertained.

Still, for organisms that make a living in comparable environments, a classification in terms of their biological complexity given their bioinformatic pedigree is not unreasonable, and the concepts and issues discussed below are steps toward such a goal. In the following section, I review standard types of complexity measures, how they compare to each other and how they fail to capture the essence of biological complexity. I then discuss a type of sequence complexity that can be shown to reduce to the information encoded into the biochemistry of DNA and proteins, and discuss its advantage over other measures, as well as its practical problems. I discuss complexity measures for biological and biochemical networks in the following section, and close with some thoughts on future directions.

Measures of Biological Complexity

The different measures put forward in the literature to quantify an organism's complexity can be grouped into three main categories: structural, functional, and sequence complexities. A fourth category—network complexity—is very recent, and I will consider that last.

Structural Complexity

Perhaps the most obvious way to define complexity is with reference to the structural complication of an organism. Clearly, structural complexity is not limited to biological organisms alone, and a good measure of this kind might allow us to compare biological organisms to human feats of engineering. Two problems are immediately apparent for any measure of structural complexity. First, to generate a scalar (that is, a single number) that will rank all possible structures appears to be a formidable task. Second, there is no guarantee that structural complexity is always a good predictor for an organism's success in the biosphere. On the one hand, something that looks like a complicated contraption could, in principle, be an evolutionary artifact: a non-adaptive feature that is either a necessity or a consequence of another feature (called a "spandrel" by Gould and Lewontin [4]). On the other hand, a complicated device could conceivably be functional only in a very different environment, perhaps one in which the organism can no longer survive. In other words, complex structure is not necessarily predictive of complex function, although we expect this to be true in the majority of cases.

The second difficulty, however, pales compared to the first. Commonly, a measure of structural complexity attempts to count the number of parts and their connection. Several such measures are reviewed by McShea [5]. A typical example for a structural complexity measure is the number of *different* cell types within an organism [6], perhaps normalized by the total number of cells in order to counter the bias of organism size. Bell and Mooers analyzed such a measure [7], and found that it robustly classifies animals as more complex than plants, and plants more complex than algae. However, the measure remains a very crude estimate of complexity, unable to shed light on finer gradations of the tree of life.

Hierarchical measures of complexity represent a different type of structural complexity. Generally speaking, a measure of hierarchical complexity seeks to quantify the number of levels needed to build a biological system, for example as the minimum amount of hierarchical structuring needed to build an understanding of the system [8]. The problem with a hierarchical scale of complexity for biological systems is that there are only four clear hierarchies: the prokaryote cell, the eukaryotic cell viewed as a symbiotic assembly of prokaryotic cells, the multicellular organism, and colonial individuals or integrated societies [9]. However, it is possible to introduce a higher-resolution scale by decomposing each hierarchy into levels and sublevels, for example by differentiating a monomorphic aggregate of elements of the lower hierarchy from a differentiated aggregate, and an aggregate of nested differentiated elements [9]. Even though a hierarchical measure of complexity necessarily represents a fairly coarse scale, it is one of only few measures of structural complexity that shows an unambiguous increase in complexity throughout the fossil record.

Functional Complexity

Ideally, any measure of biological complexity should be functional, that is, reflecting how the organism functions in a complex world. Because of the obvious difficulty in relating either form to function, or sequence to function, function-based measures are currently even less well-formed than structural ones. Because function

is understood to be a product of natural selection, it is often implicitly assumed that a measure of functional complexity should increase in evolution. However, arguments advanced earlier clearly destroy this notion: it is certainly possible that following a drastic change in the environment the functional complexity of an organism decreases, for example because a preferred metabolic substrate is eliminated from the environment, abolishing the metabolic pathway—and thus the function associated with it—for the unlucky species. A functional complexity measure for single molecules that is based on information theory has recently been introduced by Szostak [10]. I will discuss this measure in more detail in Sect. "Biochemical Information and Functional Complexity".

Attempts to count the number of different functions that an organism can perform are (and perhaps will remain) hopeless. This problem is directly related to our inability to characterize the necessary, or even sufficient, elements of an organism's environment. McShea proposed to catalogue the different behaviors of an organism as proxy for its different functions [11], as behaviors are more amenable to observation than functions. In turn, he suggested that the number of *behavioral parts* of an organism could be used as a proxy for the number of behaviors, if it is true that parts usually play a role only in one function or behavior. He gives the example of the Swiss Army knife as a device more complex than a screwdriver, on account of the different parts that represent different functions. McShea's analysis of functional parts leads him to consider networks of interacting parts that display varying degrees of modularity. I will return to such network-based measures further below.

Sequence Complexity

Given that all forms of life on Earth contain a genetic code that is responsible for generating their form and function, we might naively assume that the amount of haploid DNA (measured either in picograms (pg) as was done before the advent of whole genome sequencing or in millions of base pairs (mbp) as is more common today) would reflect—even if only roughly—the complexity of the organism. This hope was quashed relatively early on: Britten and Davidson showed conclusively [12] that no correlation between genome size and perceived complexity exists. This disconnect has been termed the "C-value paradox" [13] (reviewed in [14]) and is perhaps best exemplified by the giant free living amoeba *Amoeba dubia*, whose total DNA content was estimated at 700 pg, which would correspond to about 675,000 mpb if it was all haploid. This would correspond to about 200 times the C-value of humans. However, the haploidy of the *A. dubia* genome is now in doubt [15]. The variation in genome size and the absence of a correlation to a complexity scale such as that given by the classical chain of being is depicted in Fig. 2.

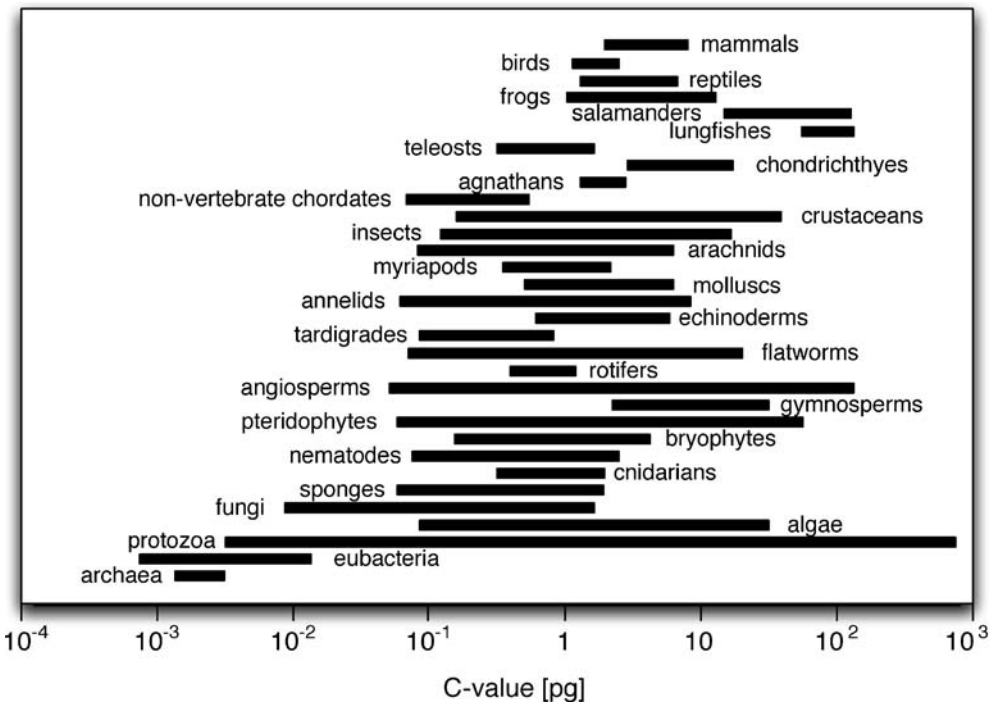
Because sequence length cannot be used, the focus for measures of sequence complexity instead has been on mathematical measures. The literature on mathematical sequence complexity, or more specifically, the complexity of symbolic strings, is far richer than that concerning functional or structural and morphological complexity. Among sequence complexities there are many different types, such as Kolmogorov, compositional, or information-theoretic ones. I shall give brief expositions of examples of each of these types here, without attempting to be even nearly exhaustive. For a good review of sequence complexities used in physics, see Ref. [16].

The most well-known sequence complexity is that introduced by the Russian mathematician Andrey Kolmogorov. He proposed to assign to each symbolic sequence a scalar that represents the *regularity* of the sequence [17]. So, for example, a string consisting of the repetition of a symbol or pattern is classified as regular, whereas a string with no discernable pattern would be irregular, and therefore complex. Note however that this algorithm does not classify a string as complex just because no pattern is readily identifiable. The Kolmogorov measure assumes that *all possible* computer programs (of a universal Turing machine) are tested so as to find the shortest one that produces the sequence in question. Mathematically, the Kolmogorov complexity of a string s is given by the length of the shortest program p (denoted as $|p|$) that produces s when executed on a universal Turing machine T:

$$K(s) = \min\{|p| : s = C_{\mathrm{T}}(p)\} \ , \tag{1}$$

where $C_{\mathrm{T}}(p)$ denotes the result of running program p on Turing machine T. So, for example, the binary equivalent of the irrational number π is random prima facie, however a concise algorithm (a short program p_π) exists to produce it, leading to a fairly low complexity for s_π. There is a vast amount of mathematical and information-theoretical literature concerning the Kolmogorov complexity (see, e. g., [18]), but when applied to biological sequences, this measure has two obvious flaws.

First, the procedure to generate the Kolmogorov complexity is uncomputable because the search for the smallest program may never end (the computation may not halt). Second, truly random strings, that is, those that cannot be generated by any computation, are assigned maximal complexity. But in physics (and in particular in biology!) truly random sequences are meaningless, and there-



Biological Complexity and Biochemical Information, Figure 2
Ranges in haploid genome size measured in picograms (the C-value) in different organisms [14]

fore should not be assigned a large complexity. In fact, the Kolmogorov complexity is even logically inconsistent. Remember that the procedure to calculate the Kolmogorov complexity is one where an automaton returns the smallest program that produces the sequence in question. For a random sequence, the procedure is to return the sequence itself (and thus the complexity estimate of a random sequence is the length of that sequence, the largest possible result). But, logically, a random sequence can never be the result of a computation, because a computation is a deterministic process. These flaws can be partly fixed by focusing on conditional and mutual Kolmogorov complexities, as I outline further below. In summary, while the Kolmogorov complexity is a mathematically sound measure of *mathematical* sequence regularity, it is not a measure of the complexity of a sequence that describes a physical object.

The key concept to account for objects in the physical—rather than mathematical—world is the *conditional Kolmogorov complexity* (first introduced in [17]) of a string s given another string t, defined as the length of the smallest program $|p|$ that produces s using that program and an external string t:

$$K(s|t) = \min\{|p| : s = C_{\mathrm{T}}(p|t)\}\ , \qquad (2)$$

with the notation $(p|t)$ pronounced as "p given t". Now, t can be a sequence from the physical world, or describing an object in the physical world. The program p is small if the sequence s can be obtained from t using a simple computation. Using this construction, the conditional complexity of a random sequence r can be defined rigorously, because the shortest program to produce a random sequence r involves both the sequence r as input to the machine, and the vanishingly small (in the limit of infinite strings) program p="print". In other words, the conditional Kolmogorov complexity of random strings *vanishes* in the limit of long sequences, rather than being maximal.

To define a physical complexity of a symbolic string (physical, because it refers to a physical world, rather than to abstract mathematics), we can imagine a sequence e that represents everything that can be measured in that world. We now ask of our Turing machine to compute string s given everything that is knowable about the physical world, that is, given e. In that case, the conditional complexity $K(s|e)$ represents everything that *cannot* be obtained from knowing the physical world. In other words, it represents the "remaining randomness": the unmeasurable. Naturally, the physical complexity of the string is then just the length of the sequence *minus* the remaining

randomness [19]:

$$C_P(s) = K_0(s) - K(s|e)\,, \tag{3}$$

where $K_0(s)$, the Kolmogorov complexity in the absence of both a "world" string e and the rules of mathematics, is just the length of s: $K_0(s) = |s|$. This notation is chosen in anticipation of results from information theory introduced in the following section. In particular, it allows us to see that $C_P(s)$ is simply a *mutual Kolmogorov complexity* [17], between string s and the world e. Put in another way, the physical complexity of a sequence is that part of the sequence that can be obtained from the world string e using a concise (and therefore short in the limit of very long sequences) computation using program p. Only *those* sequences can be obtained by computation from the world e that *mean* something in world e, or refer to something there. Note that by its construction, the physical complexity represents a special case of the "effective complexity" measure introduced earlier by Gell-Mann and Lloyd [20].

As an example, consider a world of machines, whose blueprints are stored in such a way that they can be represented as sequences of symbols. In this world, e represents all the blueprints of all the possible machines that exist there. A sequence s is complex in this world if a part—or all—of the sequence can be obtained by manipulating, or translating, the world tape e. (It is conceivable that s contains part of a blueprint from e in encrypted form, in which case the program p must try to compare e to encrypted forms of s.) Of course, it would be an accident bordering on the unreasonable to find that a string that is complex in e is also complex mathematically. Instead, from a mathematical point of view, such a sequence most likely would be classified as random, or rather, the search for a shortest program would not halt. Similarly, it is extremely unlikely that sequence s would be classified as complex in a world in which e represents, say, all the literature produced on Earth (unless there are a few books on the blueprints of certain machines!). Thus, the complexity of a sequence s, by this construction, is never absolute (like in mathematics), but always conditional with respect to the world within which the sequence is to be *interpreted*.

This is precisely what we need in order to quantify the complexity of biological sequences, because it is immediately clear that a biological sequence only means something in a very specific environment, given very specific rules of chemistry. So, according to this argument, the sequence describing a particular organism is complex only with respect to the environment within which that organism "makes its living", that is, its niche. Take the organism out of its niche, and it is unlikely to function as well as in its native niche; some of its structural complexity may turn into a useless appendage or, worse, a liability. In the following section, we will see under what circumstances this physical complexity can be understood in terms of information theory.

Biochemical Information

The inception of information theory [21] created the widely accepted expectation that biology would ultimately become a subdiscipline of cybernetics, information theory, and control theory [22]. That this expectation has not come to pass lies partly in a gross underestimate of the complexity of molecular and cellular biology by engineers and physicists, and partly in a fundamentally misguided application of information theory to genomes and molecular sequences in general. Most of the early applications of information theory to biomolecular sequences focused on an estimation of the entropy of sequences [23], rather than the information content. For example, a significant amount of work was expended on estimating the compressibility of DNA and protein sequences by studying long range correlations [24,25,26], all the while mistaking the average per-site entropy for information content [27]. These studies concluded, for the most part, that coding DNA sequences and proteins are *essentially random*, or uncompressible (non-coding DNA was found to be less random, due to repeats). From the point of view of coding theory, such a finding should not have been surprising, as Shannon's coding theorem [28] implies that the length of a message conveyed by a symbolic sequence is limited by the per-site (source) entropy [29]. In other words, we expect evolution to try to maximize the source entropy. However, source entropy does not equal information content. Let us briefly review the basics of information theory as applied to molecular sequences [30].

Entropy

For a random variable X that can take on states $x_1, \ldots, x_D$ with probabilities $p(x_i)$, the entropy of X is given by

$$H(X) = -\sum_i^D p(x_i) \log p(x_i)\,, \tag{4}$$

where the logarithm is taken to a convenient base, and determines the units of the entropy. If the base is chosen to be 2, for example, the entropy is given in bits. Often, it is convenient to take the number of possible states D as the basis. In that case, the entropy is bounded from above by 1, and a sequence of N such random variables (a polymer) has a maximal entropy of N.

Suppose we are given a DNA sequence

AGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGA
TCCCGCATAGCTCCACCA (5)

We can either take the stance that this is one of 4^{56} possible DNA sequences of length $N = 56$, or we can imagine that this is the record of 56 consecutive independent trials of a single DNA random variable that can take on the states $x = \mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}$ only. If we take the latter view, then we can try to estimate the *source entropy* by using the results of the 56 measurements to calculate the *plug-in probabilities*:

$$p_{\mathtt{A}} \approx \frac{10}{56} \approx 0.18\,, \quad p_{\mathtt{T}} \approx \frac{11}{56} \approx 0.20\,, \\ p_{\mathtt{G}} \approx \frac{16}{56} \approx 0.29\,, \quad p_{\mathtt{C}} \approx \frac{19}{56} \approx 0.34\,. \tag{6}$$

These probabilities are quite uncertain due to the lack of statistics, and only give us limited knowledge about the nature of the gene. The deviation of the calculated probabilities from the unbiased prediction $p_i = 1/4$ (an independent identically distributed, or i.i.d, random variable) is not significant. Indeed, using the four probabilities above, we obtain as an estimate of the source entropy

$$H(X) = -\sum_{i=\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}} p(x_i) \log_4 p(x_i) \approx 0.978\,, \tag{7}$$

which is almost maximal and confirms that, whatever message is encoded in the sequence, it is encoded in a nearly maximally compressed manner. For longer sequences, an analysis of the usage probabilities can reveal important information such as a bias in nucleotide usage, which will reduce the source entropy.

A typical application of sequence entropy is the calculation of "n-gram" entropies for blocks of sequences of length n:

$$H_n = -\sum_{i}^{4^n} P_i^{(n)} \log P_i^{(n)}\,, \tag{8}$$

where $P_i^{(n)}$ represents the probability to find the ith of the 4^n blocks of length n in the sequence. H_n then measures the average uncertainty within an average n-gram. We can write down the difference [31]

$$h_n = H_{n+1} - H_n\,, \tag{9}$$

which is sometimes called the amount of information necessary to predict the next symbol on a sequence given the n previous symbols. For example, for the sequence Eq. (5), $H_2 = 1.877$ (based on 55 2-grams) so that $h_1 = H_2 - H_1 = 0.899$, using the per-site entropy Eq. (7) for H_1. Note, however, that the difference Eq. (9) is, strictly speaking, not a measure of information but rather the conditional entropy (see below) to find a symbol A given the sequence S_n of length n: $h_n = H(A|S_n)$.

Using the entropy per letter $h = \lim_{n\to\infty} h_n$, we can calculate Grassberger's *effective measure complexity* [31]

$$EMC = \sum_{n=0}^{\infty} (h_n - h)\,. \tag{10}$$

This measure sums up the "memory" effects within the string, that is, it sums up the correlations at all scales. This measure vanishes if all sites are independent (because then $H_n = nh$ and $h_n = h$) and is maximal if there are strong long-range correlations within a sequence. The utility of this measure to capture the complexity of a gene is much in doubt, however, because the ability to predict the next symbol on a sequence, as discussed before, is unlikely to shed light on the function and utility of the gene if the information contained in the sequence is *encoded*.

The *compositional complexity* is another usage of n-gram entropies to capture sequence complexity, but this measure attempts to find the optimal segmentation of the sequence into m partitions based on a significance criterion. The more heterogeneous a sequence is, the higher the compositional complexity C_{comp} of a sequence s with length L [32]

$$C_{\text{comp}} = \max_{\pi} J(s_m)\,, \tag{11}$$

which is the maximum over all possible partitions π of the Jensen–Shannon divergence

$$J(s_m) = H(s) - \sum_{i=1}^{m} \frac{l_i}{L} H(s_i)\,, \tag{12}$$

where $H(s)$ is again the source entropy, and $H(s_i)$ is the entropy of the ith segment of length l_i. As with most of the sequence measures discussed up to now, this measure addresses coding-style more than function, and is unlikely to capture the essence of functional information.

Conditional Entropy

Before we can talk about information, we have to introduce the concept of *conditional probabilities*, that is, the probability $p(x_i|y_j)$ that a random variable X takes on one of its states x_i given that another variable Y is in one of its

states y_j. Then, the conditional entropy of X given that Y is in state y_j can be written as

$$H(X|Y = y_j) = -\sum_i^D p(x_i|y_j) \log p(x_i|y_j) \,. \qquad (13)$$

This concept allows us to discover the *relative state* of two different random variables, or in this case two different nucleotide positions. Suppose we are given 32 more sequences like the one in Eq. (5), and suppose we are told that these sequences represent the *same information*, that is, same gene, but from different samples. After aligning these 33 sequences, the result looks like Table 1.

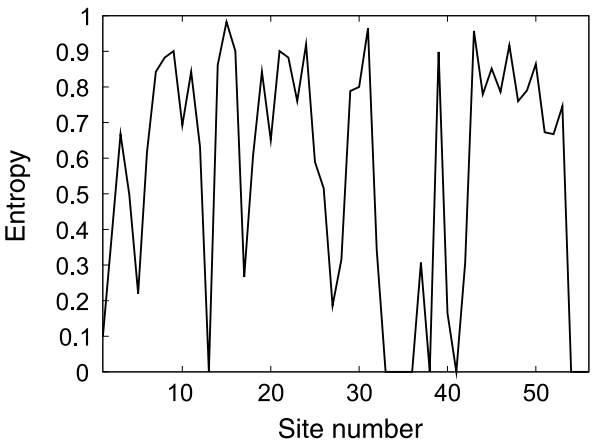
The extra symbols '–' and '.' in this alignment need further explanation. The symbol '–' denotes a "missing" symbol, that is, a deletion in that sequence with respect to the other sequences. The symbol '.', on the other hand, signals that the identity of the nucleotide at that position is not known. Here, we will treat each of these symbols as if there was no data for the random variable at that position, reducing the statistics at that site. Clearly, we now are viewing the set of sequences in Table 1 as representing independent samples of a joint random variable

$$Z = X_1 X_2 \cdots X_{56} \qquad (14)$$

Using the conditional entropy, we can now study *correlations* between nucleotide positions. For example, we can study position 8, or more precisely, the random variable X_8. We can gather a probability estimate for this position by gathering statistics *vertically*, rather than horizontally as before. Here, we find

$$\begin{aligned} p_{\mathrm{A}}(8) &= 5/33 \,, \quad p_{\mathrm{C}}(8) = 17/33 \,, \\ p_{\mathrm{G}}(8) &= 5/33 \,, \quad p_{\mathrm{T}}(8) = 6/33 \,. \end{aligned} \qquad (15)$$

which, even though 33 sequences does not represent a large sample, are significantly different from the random assumption $p_i = 1/4$. Even more important deviations from the i.i.d case can be found for X_{33} and X_{36}: only one particular nucleotide can be found at these positions, all others are absent. The reasons for this *abnormal* distribution is immediately clear when we reveal the origin of these sequences: they are the last 56 nucleotides of the tRNA gene of the *E. coli* bacterium [33], and the function of this molecule requires positions 33 and 36 to have that particular nucleotide. A mutation leading to an alternative nucleotide at these positions implies the death of the organism carrying the mutant, which explains why we do not find it in a sample of living organisms. We can now check how uneven the probabilities per-site are, by plotting the per-site entropy (normalized to lie between zero and one by choosing base $D = 4$ for the logarithm) against site-number in Fig. 3.



Biological Complexity and Biochemical Information, Figure 3
Per-site entropy of the last 56 positions of *E. coli* tRNA from the alignment in Table 1

We notice that there are sites that are almost random, and some sites that are absolutely conserved. While this pattern, conserved by evolution over billions of years, clearly reflects the function of the organism, it does not tell us whether the states of any two sites are correlated.

The relative state of two random variables will tell us whether knowing the state of one variable will reveal to us information about another variable. For example, we can ask whether knowing, say, X_8, allows us to say something about other sites that we would not already know in the absence of that knowledge. Let us see whether the probabilities characterizing X_{21}, say, depend on knowing X_8. Collecting frequencies for X_{21} gives approximately

$$\begin{aligned} p_{21}(\mathrm{A}) &= 0.24 \,, \; p_{21}(\mathrm{C}) = 0.46 \,, \\ p_{21}(\mathrm{G}) &= 0.21 \,, \; p_{21}(\mathrm{T}) = 0.09 \,, \end{aligned} \qquad (16)$$

while $p(X_{21}|X_8) =$

$$\begin{pmatrix} p(\mathrm{A}|\mathrm{A}) & p(\mathrm{A}|\mathrm{C}) & p(\mathrm{A}|\mathrm{G}) & p(\mathrm{A}|\mathrm{T}) \\ p(\mathrm{C}|\mathrm{A}) & p(\mathrm{C}|\mathrm{C}) & p(\mathrm{C}|\mathrm{G}) & p(\mathrm{C}|\mathrm{T}) \\ p(\mathrm{G}|\mathrm{A}) & p(\mathrm{G}|\mathrm{C}) & p(\mathrm{G}|\mathrm{G}) & p(\mathrm{G}|\mathrm{T}) \\ p(\mathrm{T}|\mathrm{A}) & p(\mathrm{T}|\mathrm{C}) & p(\mathrm{T}|\mathrm{G}) & p(\mathrm{T}|\mathrm{T}) \end{pmatrix}$$
$$= \begin{pmatrix} 0.2 & 0.235 & 0 & 0.5 \\ 0 & 0.706 & 0.2 & 0.333 \\ 0.8 & 0 & 0.4 & 0.167 \\ 0 & 0.059 & 0.4 & 0 \end{pmatrix} . \qquad (17)$$

Biological Complexity and Biochemical Information, Table 1
Alignment of the last 56 nucleotides of sequences of tRNA genes of *E. coli*

```
AGAGCGCCTGCTTTGCACGCAGGAGGTCTGCGGTTCGATCCCGCATAGCTCCACCA
AGAGCGCTTGCATGGCATGCAAGAGGTCAGCGGTTCGATCCCGCTTAGCTCCACCA
TATGTAGCGGATTGCAAATCCGTCTA-GTCCGGTTCGACTCCGGAACGCGCCTCCA
AGAATACCTGCCTGTCACGCAGGGGGTCGCGGGTTCGAGTCCCGTCCGTTCCGCCA
AGGACACCGCCCTTTCACGGCGGTAA-CAGGGGTTCGAATCCCCTAGGGGACGCCA
AGAGCAGGGGATTGAAAATCCCCGTGTCCTTGGTTCGATTCCGAGTCCGGGCACCA
ATTACCTCAGCCTTCCAAGCTGATGA-TGCGGGTTCGATTCCCGCTGCCCGCTCCA
AGAGCACGACCTTGCCAAGGTCGGGGTCGCGAGTTCGAGTCTCGTTTCCCGCTCCA
AGAACGAGAGCTTCCCAAGCTCTATA-CGAGGGTTCGATTCCCTTCGCCCGCTCCA
AGAGCCCTGGATTGTGATTCCAGTTGTCGTGGGTTCGAATCCCATTAGCCACCCCA
AGAGCGCACCCCTGATAAGGGTGAGGTCGGTGGTTCAAGTCCACTCAGGCCTACCA
AGAGCAGGCGACTCATAATCGCTTGGTCGCTGGTTCAAGTCCAGCAGGGGCCACCA
AGAGCAGTTGACTTTTAATCAATTGGTCGCAGGTTCGAATCCTGCACGACCCACCA
AGAGCACATCACTCATAATGATGGGGTCACAGGTTCGAATCCCGTCGTAGCCACCA
AGAACGGCGGACTGTTAATCCGTATGTCACTGGTTCGAGTCCAGTCAGAGGAGCCA
AGCGCAACTGGTTTGGGACCAGTGGGTCGGAGGTTCGAATCCTCTCTCGCCGACCA
AGCGCACTTCGTTCGGGACGAAGGGGTCGGAGGTTCGAATCCTCTATCACCGACCA
AGCGCACCGTCATGGGGTGTCGGGGGTCGGAGGTTCAAATCCTCTCGTGCCGACCA
AAGGCACCGGTTTTTGATACCGGCATTCCCTGGTTCGAATCCAGGTACCCCAGCCA
AAGGCACCGGATTCTGATTCCGGCATTCCGAGGTTCGAATCCTCGTACCCCAGCCA
AGAGCGCTGCCCTCCGGAGGCAGAGGTCTCAGGTTCGAATCCTGTCGGGCGCGCCA
AGAGCAACGACCTTCTAAGTCGTGGGCCGCAGGTTCGAATCCTGCAGGGCGCGCCA
AGAGCAACGACCTTCTAAGTCGTGGGCCGCAGGTTCGAATCCTGCAGGGCGCGCCA
AGAGTACTCGGCTACGAACCGAGCGGTCGGAGGTTCGAATCCTCCCGGATGCACCA
ATAACGAGCCCCTCCTAAGGGCTAAT-TGCAGGTTCGATTCCTGCAGGGGACACCA
AGAGCGCACCCTTGGTAGGGGTGGGGTCCCCAGTTCGACTCTGGGTATCAGCACCA
AGAGCGCACCCTTGGTAAGGGTGAGGTCGGCAGTTCGAATCTGCCTATCAGCACCA
AGAGCAACTGACTTGTAATCAGTAGGTCACCAGTTCGATTCCGGTA.TCGGCACCA
AGAGCAGCGCATTCGTAATGCGAAGGTCGTAGGTTCGACTCCTATTATCGGCACCA
AGAGCGCACCCTTGGTAAGGGTGAGGTCCCCAGTTCGACTCTGGGTATCAGCACCA
AGAGCACCGGTCTCCAAAACCGGGTGTTGGGAGTTCGAGTCTCTCCGCCCCTGCCA
AGCTCGTCGGGCTCATAACCCGAAGATCGTCGGTTCAAATCCGGCCCCCGCAACCA
AGCTCGTCGGGCTCATAACCCGAAGGTCGTCGGTTCAAATCCGGCCCCCGCAACCA
```

Knowing X_8 would help us in predicting the state of X_{21} if any of the numbers in any of the columns of Eq. (17) is significantly larger than 0.25, which is the probability to guess X_{21} right by random chance. In fact, however, only a very few of the probabilities are even above 0.5. Clearly, the knowledge of position 8 does not help much in predicting position 21. Given $X_8 = \mathtt{G}$, for example, the probabilities to correctly predict $\mathtt{A}, \mathtt{C}, \mathtt{G}$ or $\mathtt{T}$ at position 21 are given by the third column of Eq. (17), and none of those probabilities exceed 0.5. The only case where knowing X_8 helps us significantly in predicting X_{21} is if $X_8 = \mathtt{A}$, since then $X_{21} = \mathtt{G}$ with probability 0.8. But let us look instead at another pair of sites, this time X_8 and X_{22}. We can obtain the *unconditional* probabilities for X_{22} from the alignment (as before, we are rounding the probabilities)

$$p_{22}(\mathtt{A}) = 0.18\,,\ p_{22}(\mathtt{C}) = 0.15\,,$$
$$p_{22}(\mathtt{G}) = 0.52\,,\ p_{22}(\mathtt{T}) = 0.15\,, \tag{18}$$

but the conditional probability matrix is very different from Eq. (17):

$$p(X_{22}|X_8) = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}. \tag{19}$$

This matrix immediately tells us that, based on the alignment in Table 1, if we see a $\mathtt{T}$ at position 8, then we can be certain ($p = 1$) to find $\mathtt{A}$ at position 22, while if we see a $\mathtt{G}$ at position 8, we are sure to encounter $\mathtt{C}$ at 22, and so forth. Obviously, these are the associations implied by Watson-Crick pairing, so this conditional probability matrix suggests to us that position 8 and 22 are in fact *paired* within the molecule: this is how biochemistry stores information in molecules.

Calculating the conditional entropy $H(X_{22}|X_8 = \mathtt{T})$ makes the correlation apparent in another way. As the entropy of a random variable measures the amount of uncertainty that we have about it, we immediately see that

knowing X_8 tells us everything about the state of X_{22} that there is to know: for every one of X_8's possible states, $H(X_{22}|X_8 = \mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}) = 0$. We can now introduce the *average conditional entropy*:

$$H(X|Y) = \sum_{j=1}^{D} p(y_j) H(X|Y = y_j)\,, \tag{20}$$

which gives us the average value of the entropy of X given any of the states of Y. For the pair (X_{22}, X_8) this gives of course $H(X_{22}|X_8) = 0$, while for the pair (X_{21}, X_8) we find

$$H(X_{21}|X_8) = 0.58\,. \tag{21}$$

We saw earlier that knowing site 8 does not significantly affect our probability to predict site 21, but still the conditional entropy Eq. (21) is significantly smaller than the unconditional one, $H(X_{21}) = 0.9$.

Information

From the example we just discussed, the definition of biochemical information is clear: it is just the reduction in entropy (or uncertainty) of one variable using the knowledge of the state of another. So, if Y is again a variable whose entropy we seek to decrease using our knowledge of X, then the information X conveys about Y is just

$$I(X:Y) = H(Y) - H(Y|X)\,. \tag{22}$$

This definition of information (also sometimes called "shared" or "mutual" entropy) is symmetric: $I(X:Y) = I(Y:X)$, and we can therefore equivalently write

$$I(X:Y) = H(X) - H(X|Y)\,. \tag{23}$$

Thus, "what X knows about Y, Y also knows about X". This is particularly clear for the example of biochemical information through Watson-Crick pairing that we looked at above. We can calculate the information position 8 has about position 22 as

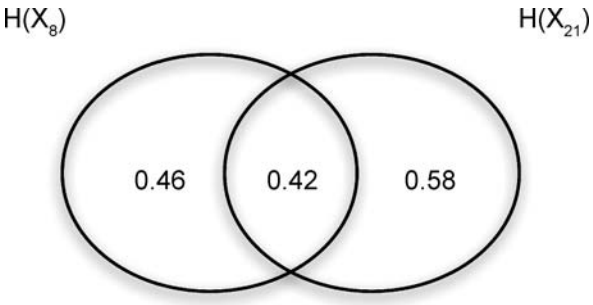
$$I(X_{22}:X_8) = H(X_{22}) - H(X_{22}|X_8) = H(X_{22}) \tag{24}$$

because $H(X_{22}|X_8) = 0$ as we found above. At the same time,

$$I(X_8:X_{22}) = H(X_8) - H(X_8|X_{22}) = H(X_8) \tag{25}$$

because $H(X_8|X_{22})$ also vanishes. Then, $H(X_8)$ and $H(X_{22})$ have to be equal. We can repeat this analysis for the other pair we looked at:

$$I(X_8:X_{21}) = H(X_{21}) - H(X_{21}|X_8) = 0.9 - 0.58 = 0.42\,. \tag{26}$$



Biological Complexity and Biochemical Information, Figure 4
Venn diagram of entropies between sites 8 and 21 of the sequence alignment (1)

A simple diagram (Fig. 4) helps to see how entropies are distributed among shared and conditional entropies.

As we discussed earlier, if two sites share most of their entropy, we can conclude that they bind to each other in a Watson-Crick pair. Because this binding is responsible for the structure of the molecule, information theory can help us to determine the molecule's *secondary structure*, that is, how the molecule is arranged as a chain in two dimensions [34].

Molecular Complexity

While it is important to understand how biochemistry encodes information in chemical bonds, calculating information-theoretical correlations between nucleotides (or between residues in proteins [35,36,37,38]) does not reveal how much information a molecule stores about the environment within which it functions. We can achieve this by imagining that the physical and chemical world has a *description* in terms of a symbolic sequence. Then, given this particular sequence (and thus, given the identity of the world within which we study the entropy of biomolecular sequences), the set of information-rich sequences can be determined, and the information content calculated. The mathematical form of the measure of a molecule's information content will be shown to be closely related to the physical complexity measure discussed earlier in Eq. (3).

We start by considering a molecule that is guaranteed *not* to encode information about its environment: a random molecule. Let us imagine a polypeptide P of L residues, written as a joint random variable $P = P_1 P_2 \cdots P_L$. If P is truly random, then each of its 20^L states are equally likely, and the entropy of P is

$$H(P) = \log 20^L\,. \tag{27}$$

If we choose to take 20 as the base of the logarithm, we see that the entropy of the random L-mer is $H(P) = L$, that is, one unit of entropy for each random monomer. Of course, functional proteins are nowhere near random. Instead, the probabilities to find a particular residue x at position n, $p_n(x)$, are strongly constrained by evolution. The opposite extreme of a protein would be the case where one and only one particular residue is allowed at each position in a protein that has a specific function. Then, only one state of the polymer is consistent with that function, and $H(P) = 0$. In general, the entropy of P is given by

$$H(P) = -\sum_{i}^{20^L} p_i \log p_i \,, \tag{28}$$

where p_i is the probability to find any of the possible 20^L states of P in an infinitely large ensemble of polymers of length L. In practice, however, ensembles large enough to estimate these p_i cannot exist (note that 20^{100} is approximately 10^{130}), so we need to find approximations to calculate $H(P)$ if we would like to measure a polymer's entropy. A common approximation is to neglect interactions between the different sites n, so that the entropy of P can be written in terms of a sum over the entropies of each monomer:

$$H(P) \approx \sum_{n=1}^{L} H(P_n) \,. \tag{29}$$

In practice this is not a bad approximation [39], but we should keep in mind that interactions between residues, known as *epistasis*, are extremely important in evolution and are often studied in detail [40,41,42]. Typically, while many pairs of residues interact epistatically, some do so positively and some negatively, so that on average the approximation Eq. (29) often holds.

If we can consider each residue separately, we can now focus on the probabilities $p_n(x)$ introduced above. Typically, we expect those sites that are very important for the function of the protein to be strongly conserved (all the $p_n(x)$ are either one or zero at position n), leading to a vanishing entropy at that site, while those that are less important are less constrained, leading to a larger entropy at those sites. We can now define the amount of information a sequence P stores about its environment as the mutual entropy (information) between the random variable describing protein P and a random variable E describing all possible environments (we imagine, as before, that all possible environments can be described in terms of sequences, and listed)

$$I(P:E) = H(P) - H(P|E) \,, \tag{30}$$

which is the same formula as Eq. (46). The first term on the right hand side of Eq. (30) is the *unconditional* entropy of the polymer P. An entropy that is not conditional on any environment is an entropy that is unspecified. This is the same as the entropy of a random protein, and thus $H(P) = H_{\max}(P) = L$. The second term in that equation is the average conditional entropy of the protein, averaged over all the possible environments described by the random variable E. Our world, of course, is just one particular such environment $E = e$, and therefore the amount of information stored in a polymer P about environment $E = e$ is given by

$$I(P:e) = L - H(P|e) \,, \tag{31}$$

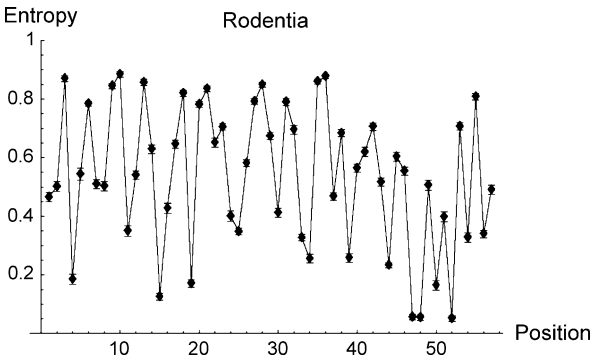
where

$$H(P|e) = \sum_{i=1}^{L} H(P_i|e) = -\sum_{n=1}^{L}\sum_{x=1}^{20} p_n(x) \log_{20} p_n(x) \,, \tag{32}$$

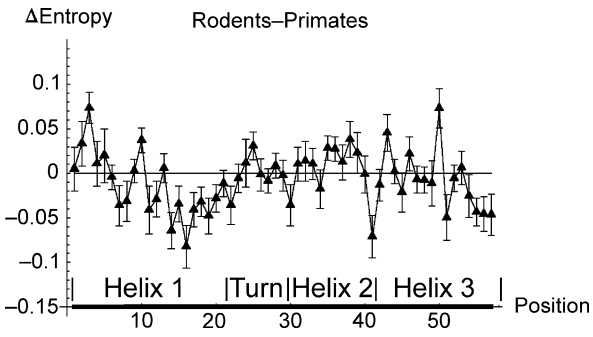
and the sum over x goes over the possible 20 amino acids at each site.

The probabilities $p_n(x)$ can be obtained by a sequence alignment of structurally identical proteins of different species, such as those listed in the *Pfam* database [43]. As an example, we can align the sequences of the homeobox proteins of the rodent family *Rodentia*. Aligning 703 of the sequences of 57 residues in the database (as of July 2006) allows us to estimate the $p_n(x)$ necessary for calculating the rodent homeodomain information content. Note that for this analysis to be significant, we have to ensure that the sequences of all the aligned proteins code for a protein with the *same* functionality. Indeed, any time two different residues are allowed at a particular position, we must be able to imply that these are neutral substitutions in the protein. Adaptive changes that influence protein function should not appear in the alignment. The average sequence identity of the set of 703 proteins is about 39%, which gives confidence that the set is composed of sequences coding for proteins that are at least structurally identical. (Sequences with more than 30% identity have more than a 90% chance of coding for structurally identical proteins [44].)

When estimating entropies from finite ensembles, care must be taken to correct the estimates for a bias that arises when the ensemble is small. The method for correcting this bias is well-known [45,46], and applied to all data shown here. We start by calculating the per-site entropy of all 57 aligned sites, giving the *entropy profile* of the sequence shown in Fig. 5. This view reveals a curious alterna-



Biological Complexity and Biochemical Information, Figure 5
Entropic profile of the 57 amino acid rodent homeodomain, obtained from 703 sequences in Pfam (St. Louis mirror, accessed July 20, 2006). Per-site entropy between zero and one by taking logs to base 20



Biological Complexity and Biochemical Information, Figure 6
Difference between entropic profile of the homeobox protein of rodents and primates (the latter from 504 sequences in Pfam St. Louis, accessed July 20, 2006)

tion between high and low entropy sites in the homeobox protein.

The information content is obtained by summing up the per-site entropies and subtracting this value from the length of the sequence, as implied by Eq. (31). If we call the unit of information obtained by using logarithms to the base of the size of the alphabet the "mer" (such that an L-mer has a maximum information content of L mers) then the information content of rodent homeobox proteins is

$$I(P_{\text{rodents}}) = 26.83 \pm 0.13 \,\text{mers}\,, \tag{33}$$

where the quoted error reflects the statistical nature of the probability estimate, quantified in [45,46]. Note that for amino acids, 1 mer equals about 4.32 bits, while for nucleic acids 1 mer equals 2 bits.

From the point of view of evolution, we may ask whether this information content has changed as species evolved. For example, we might ask whether animals that, according to our intuitive understanding of the "great chain", are considered "higher" than rodents show a different information content for this protein. The homeobox protein sequences of many other families of organisms are available in *Pfam* to examine this. As an example, the alignment of 504 sequences of primate homeobox proteins results in an entropic profile remarkably similar to that of Fig. 5. The information content for these proteins can be calculated to be

$$I(P_{\text{primates}}) = 26.71 \pm 0.14 \,\text{mers}\,, \tag{34}$$

in other words, identical to that of the rodents within statistical error. But just because the total information content is the same does not imply that information is coded in a similar manner. We can study this by subtracting one entropic profile form the other, to see if the entropy of some sites has increased, and some others decreased in the evolution from rodents to primates. We can see this difference plot in Fig. 6, which suggest that some recoding did indeed take place, in particular in the first helix of the protein, but the significance of this result is not strong.

Thus, information theory can track some aspects of evolutionary changes between proteins, even if the total amount of information is unchanged.

Biochemical Information and Functional Complexity

It is tempting to speculate that the information content of a biomolecule is related to the functional complexity of an organism. After all, possessing information about an ensemble enables the *prediction* of the possible states of an ensemble with accuracy better than random, something that is highly valuable for a biological organism whose ensemble is an uncertain environment. Let us recall the measure of sequence complexity ("physical complexity") introduced in the previous section:

$$C_P(s) = K_0(s) - K(s|e)\,, \tag{35}$$

for any sequence s, given an environment sequence e. If we take the average of this quantity over an infinite ensemble of sequences s_i drawn from an ensemble S, we obtain

$$\langle C_P(s)\rangle_S = \sum_{s_i} p(s_i)\,(K_0(s_i) - K(s_i|e))\,, \quad s_i \in S\,. \tag{36}$$

It is easy to prove the following inequality between average Kolmogorov complexities and the Shannon entropy [47]:

$$\sum_{s_i} p(s_i)K(s_i) \geq \sum_{s_i} p(s_i)\log\frac{1}{p(s_i)} = H(S)\,. \tag{37}$$

The inequality in Eq. (37) reflects the possibility that the complexities $K(s_i)$, which can be viewed as compressed *encodings* of the sequences s_i, do not necessarily form a *perfect code* that saturates the Kraft inequality [48]. However, because the $K(s_i)$ do represent the *smallest* program encoding s_i, it is reasonable to assume that the average Kolmogorov complexity is given by the Shannon entropy of the ensemble up to an additive constant, which represents the length of a program that tells one computer how to simulate another (see, e. g., [48]):

$$\langle K(s_i)\rangle_S \approx H(S) + c\,. \tag{38}$$

In that case (and assuming that the overall constant cancels from the difference), the average physical complexity becomes

$$\langle C_P(s)\rangle_S \approx H(S) - H(S|e)\,, \tag{39}$$

where $H(S)$ is the unconditional Shannon entropy of the ensemble of sequences, and $H(S|e)$ is the conditional entropy. If the ensemble S consists of sequences of fixed length L, then the unconditional entropy is $H(S) = L$, and $H(S|e)$ is the conditional entropy of the sequences as in Eq. (31). (Note that, technically, the Kolmogorov complexity for fixed length sequences $K(s|L)$ is related to the arbitrary length complexity $K(s)$ via $K(s) \leq K(s|L) + 2\log L + c$, where c is again the "simulation constant".) To summarize, the average physical complexity is (assuming perfect coding) equal to the Shannon information that the ensemble has about the environment, that is, a sequence's information content.

This interpretation of complexity is particularly satisfying from an evolutionary point of view. The value of information lies in the ability of the observer who is in possession of it to make *predictions* about the system that the information is *about*. Organisms, armed with the functionality bestowed upon them by their genetic code, do precisely that to survive. An organism's metabolism is a chemical machine making predictions about the availability and concentrations of the surrounding chemicals. A cell's surface proteins make predictions about the type of cells it might interact with, and so on. Viewed in this way, informational complexity should be a near perfect proxy for functional complexity, because information must be used for function: if it is not so used, a sequence represents entropy, not information. An investigation of the informational complexity of evolving computer programs (an instance of "digital life" [49,50]) has shown that the complexity increases in evolution [51] and correlates well with the functional complexity of the programs [52]. A good example for how the equivalence of function and information is achieved in biochemistry is the evolution of functionality in ribozymes by in-vitro evolution, achieved by Jack Szostak's group at Massachussets General Hospital.

This group evolved short GTP-binding RNAs (aptamers) in vitro, and found eleven distinct structures with different binding affinities [53]. By measuring the information content as outlined here (within each pool of sequences that evolved for each structure, there were sufficient mutants that could be aligned in order to determine the substitution probabilities $p_n(x)$), it was possible to show that increased functional activity went hand-in-hand with increased information content, so much so that the group was able to derive a simple law that predicts, within this GTP-binding set of ribozymes, that a ten-fold higher binding affinity is achieved by about 10 bits of extra information. In other words, the informational complexity is linearly proportional to the functional complexity. Even more, the structural complexity, as measured by the number of different stems (ladders) within the secondary structure of the enzyme, also seemed to increase with functional activity.

Based on this type of observation, Szostak has proposed a new measure of functional complexity [10,54] that is based both on function and on information. For a particular function x, let E_x represent the degree of that function achieved by a system. Then the *functional information* is defined as [54]

$$I(E_x) = -\log(F(E_x))\,, \tag{40}$$

where $F(E_x)$ is the fraction of all possible configurations of the system that possess a degree larger or equal to E_x. For sequences, the function could represent a binding affinity, or the number of ATPs produced by a pathway within which the enzyme is the bottleneck factor, or any other real-valued attribute that characterizes the performance of the sequence. This measure introduces a clear link between information and function, but fundamentally turns out to be a coarse-grained version of the information content Eq. (31), as can be seen as follows.

Suppose we are interested in measuring the information content of a sequence s that performs function x to the degree E_x. We can obtain the functional information of s by creating all possible mutants of s, and measuring the fraction $F(E_x)$ of sequences that have the same function as s, given by $\nu(s)/N$, where $\nu(s)$ is the number of neutral mutants of s within S, and N is the total number of possible sequences. Thus,

$$I(E_x) = \log N - \log \nu(s)\,. \tag{41}$$

The conditional probability to find a sequence s_i given environment e in an evolving populations of sequences of the type s is given by $p(s_i|e)$. If e specifies the function x at level E_x for s_i, then $p(s_i|e) = 1$ if s_i performs the function at the required level, and zero otherwise (coarse-graining of the entropy). There are $\nu(s)$ such sequences in the ensemble, and thus

$$H(S|e) = -\sum_i p(s_i|e) \log p(s_i|e)$$
$$= -\sum_{\nu(s)} \frac{1}{\nu(s)} \log \frac{1}{\nu(s)} = \log \nu(s) \,. \qquad (42)$$

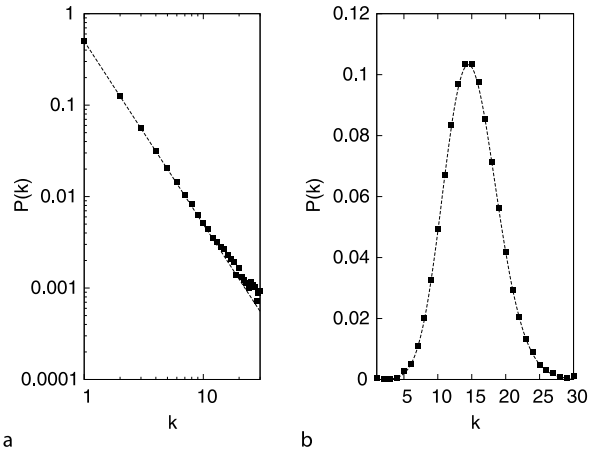
As $\log N = \log D^L = L$, Eq. (41) recovers Eq. (31), that is, functional information is a coarse-grained version of the Shannon information content of a sequence.

Network Complexity

If we are given the informational complexity of every single protein in an organism, is the sum of these complexities equal to the complexity of the cell, or the organism? Certainly not, because much of the complexity of an organism lies in how the proteins interact, and in particular in the complicated temporal sequence of events dictated by the regulation of the expression of proteins as a function of environmental influences.

It is well-known that functional biological networks have properties that distinguish them from a randomly connected set of nodes (a random graph, or Erdös–Rényi network). In fact, many biological networks have degree distributions that are approximately scale-free (Fig. 7a), in stark contrast to the binomial distribution of the random graph (Fig. 7b). Also, the diameter of biological networks, defined as the average of all internode distances $d(i, j)$, is small ("small-world-network", see [55]) and depends only weakly on the number of nodes in the network, again unlike what we find for random graphs.

Assessing the complexity of a network usually entails measuring the structural complexity of the network, as compared to a random graph. Often, the *modularity* of a network is used as a proxy for the structural or the functional complexity, even though the concept of a module for networks is not universally defined [56,57]. Usually, a module is defined as a discrete entity whose function is separable from those of other modules, but in biology modules can have significant overlap. Within protein-protein interaction networks, putative modules can be obtained by clustering, so that modules are sets of proteins that are strongly interconnected, but only weakly connected to other such sets. In this section I discuss



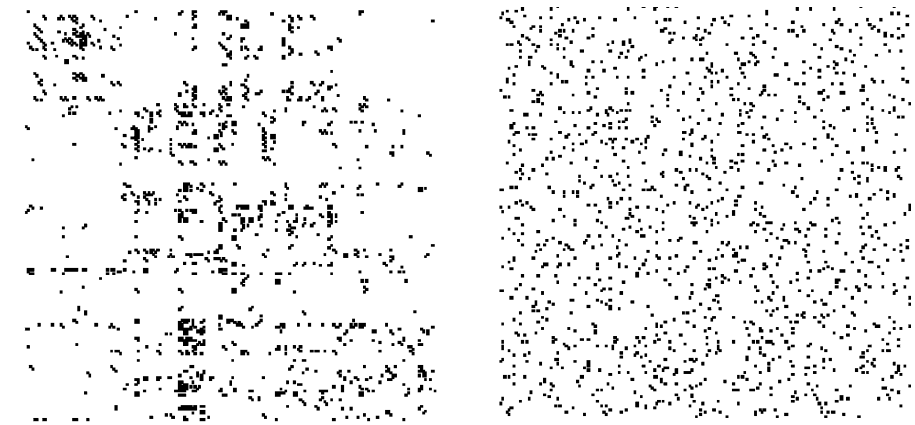
Biological Complexity and Biochemical Information, Figure 7
a Degree distribution $P(k)$ to find a node with k edges in a scale free network (squares) with degree distribution $P(k) \sim 1/k^2$ (*dashed line*), b degree distribution for a random graph with 1,000 nodes and a connectivity $p = 0.015$ (*squares*) and a Poisson distribution with mean $\langle k \rangle = 15$ (*line*)

several ways to understand network complexity, starting with an estimate of the information content of a functional network by assessing the information content of the genome that gave rise to it, using the methods discussed in Sect. "Molecular Complexity". I then review an information-theoretic approach to network modularity, followed by an exposition of a method to assess network structure and information by measuring subnetwork (subgraph) abundances.

Evolution of Information in Networks

The difficulty in assessing network complexity from network topology is clear to anyone who has studied the multitudes of networks arising in engineering and biology. Biological networks usually have thousands of nodes and several thousand edges, and often appear to be unstructured. For example, the network summarizing the connectivity of neurons in the brain of the nematode *C. elegans* shows very little modularity or structure at first sight, but is markedly different from a random network [58] (see Fig. 8).

Because the functionality of a network is not necessarily reflected in its topological structure, the best hope for assessing the complexity of a network is to measure the complexity of the set of rules used to construct it. In biology, this set of rules is encoded in the genome, so a first-order estimate of the complexity of a network should be given by the complexity of the genome that produced it. Of course, this is difficult for all the reasons given in the



Biological Complexity and Biochemical Information, Figure 8
Adjacency matrix of 179 of the 302-neuron neural network of a *C. elegans* brain (*left*), and a random network of the same size and connectivity (*right*)

previous section, but even more difficult in this case because a network of proteins, for example, is specified not just by the open reading frames coding for the proteins, but also all the untranslated regulatory regions as well as the transcription factors affecting them.

We can test the evolution of network complexity in computational models where a genome represents the functionality of a cellular network, as was done recently in Ref. [59]. In this work, an artificial chemistry and genetics was encoded in a simple linear (circular) code based on the monomers `0,1,2,3`, where enzymatic proteins with variable specificity act on 53 precursor molecules to form up to 555 metabolites. The metabolic reactions involving transport and enzymatic proteins are obtained by a translation of the genetic code into reactions, and implementing chemostat physics and reaction kinetics. Evolution proceeded from a simple ancestral genome with only 3 genes to large and complex metabolic networks of thousands of nodes and several thousand edges, in a completely asexual Wright–Fisher process acting on two chromosomes.

In order to be considered fit, an artificial cell has to import precursor molecules that are available outside the cell walls and convert them into metabolites. The fitness of an organism was determined by calculating the produced biomass of metabolites (see [59]). Evolution was carried out in three different environments that differ in their predictability. In the simplest environment, the location of precursor sources and their abundance is constant during evolution (the "static" environment), while in the quasistatic environment one randomly selected precursor source location is changed per update. In the dynamic environment, the source location of all precursors is changed randomly, and 25% of all precursors are made unavailable, giving rise to a highly unpredictable environment.

The information content of the genomes was measured as outlined above, that is

$$I = L - H(s)\,, \tag{43}$$

where L is the total length of the sequence and $H(s)$ is the sum of per-site entropies

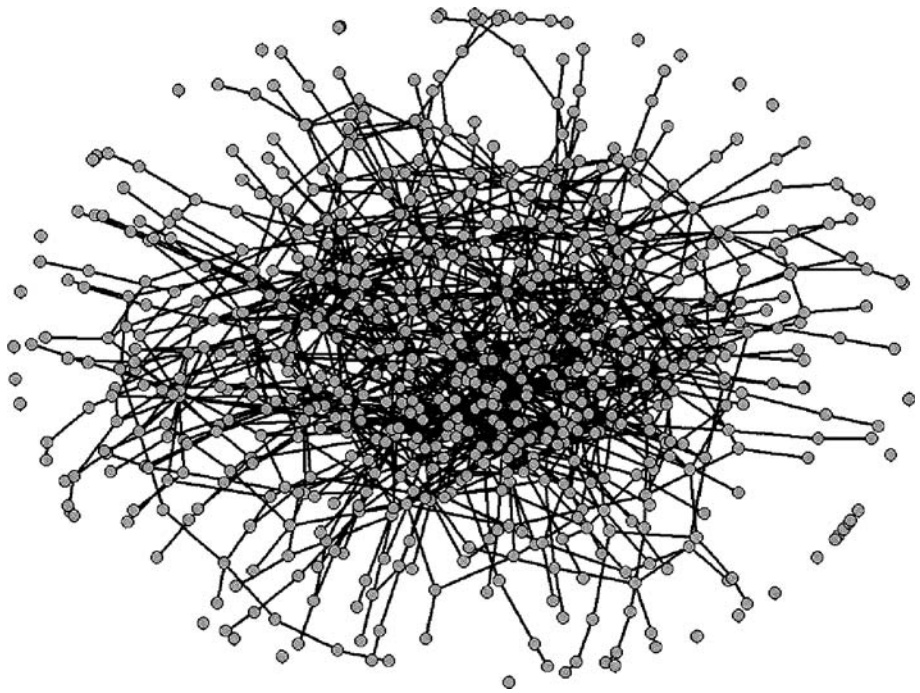
$$H(s) = \sum_{x=1}^{L} H(x) \tag{44}$$

and the per-site entropy $H(x)$ is obtained by summing over the substitution probabilities p_i at that site:

$$H(x) = -\sum_{i=0}^{3} p_i \log_4 p_i\,. \tag{45}$$

Because we only have 4 possible symbols per site, taking the logarithm to base 4 again ensures that the per-site entropy lies between 0 and 1.

As the evolutionary mechanism allows for insertions and deletion of entire genes or genetic regions along with point mutations, genomes can change length during evolution. As a consequence, an alignment of genomes in a population to ascertain substitution probabilities is problematic. Instead, an approach can be used that determines the substitution probabilities p_i from the fitness effect of the substitution on organism fitness, along with an application of population genetics theory. If a substitution of allele i has fitness effect w_i, then the probability to find this allele in an equilibrated population evolving at mutation



Biological Complexity and Biochemical Information, Figure 9
Evolved metabolic network with 969 nodes and 1,698 edges, rendered with PAJEK [60]

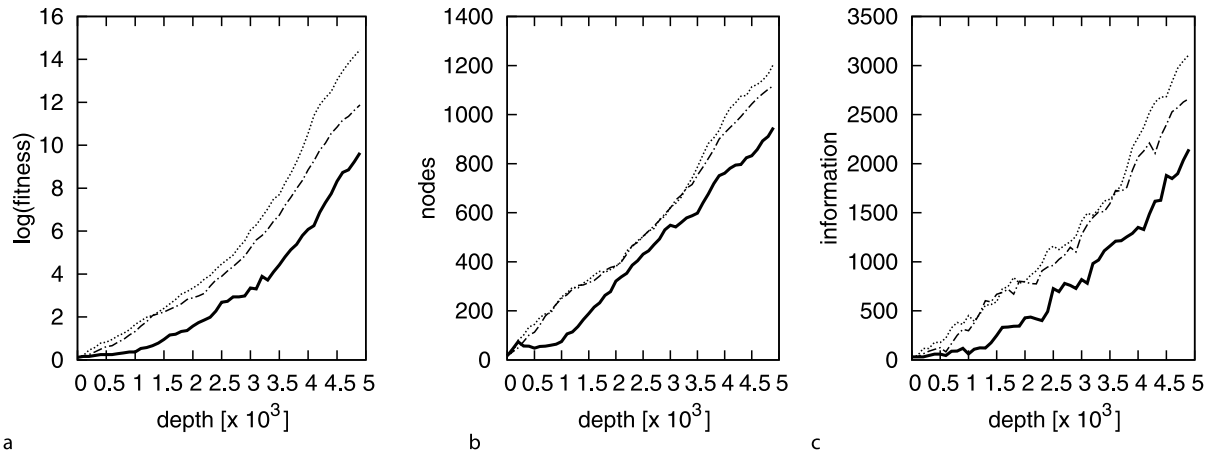
rate μ is given by [61]

$$p_i = \frac{p_i w_i}{\bar{w}}(1-\mu) + \frac{\mu}{4}\sum_{j=0}^{3}\frac{p_j w_j}{\bar{w}}\,, \tag{46}$$

where $\bar{w} = \sum_{i=0}^{3} p_i w_i$ is the mean fitness of the 4 possible alleles at that position.

Figure 10 shows the evolution of network fitness, size and complexity (measured in units of monomer entropy or "mer", where one mer equals 2 bits) for the three different environments discussed above), as a function of the phylogenetic depth of the organism. The phylogenetic depth of an organism is its position on the line of descent of the particular evolutionary run: The organism with the highest fitness at the end of the run is used to reconstruct the line of descent by following its direct ancestry and increasing the depth counter whenever an organism's genome differs from that of its direct parent. When arriving at the initial organism, this counter is set to zero, so that the phylogenetic depth counter increases up until the last organism on the line. Because on average we find about one new organism on the line of descent per generation in these runs, the phylogenetic depth is a good proxy for evolutionary time, even though in principle many generations could pass without an advance on the line of descent.

Because the fitness of an organism is multiplicative in the biomass (discovering how to produce a new metabolite multiplies the previous fitness by a number greater than one), the log of the fitness grows about linearly for all three environments (Fig. 10a). The fitness grows fastest for the static environment that is the easiest to predict (dotted line), while it takes longer for complexity to emerge in the dynamic environment (solid line). The same trend is reflected in the growth of the number of nodes and edges in these environments (Fig. 10b). Finally, the information content as calculated by Eq. (43) using the substitution probabilities Eq. (46) follows the same trend: the informational complexity grows the fastest for the static and quasi-static environments, and lags behind for evolution in a dynamic environment. The reason for the slower growth of complexity for networks evolving in dynamic environments is clear: because the availability of precursors necessary for the production of complex metabolites cannot be relied upon in such environments, the cells end up manufacturing the precursor molecules within the cells (rather than importing them from the outside). This machinery is complex in itself, but takes time to evolve. Ultimately, we expect networks evolving in dynamic environments to be more complex than those evolving in static environments because of the added flexibility of producing precursor molecules within the cells. However, such net-



Biological Complexity and Biochemical Information, Figure 10
Evolution of complexity in artificial metabolic networks. a log fitness for networks evolving in a static (*dotted line*), quasistatic (dash-dotted), and dynamic environment (*solid line*). b Evolution of the number of nodes (number of edges follows a similar trend). c Evolution of informational complexity, lines as in a

works lag behind slightly during the time this complexity is generated.

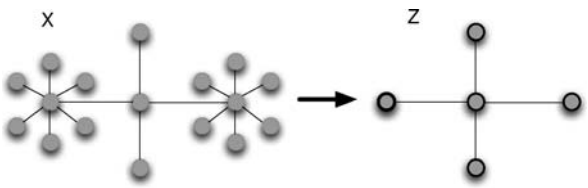
Note that the informational complexity of the networks used to seed these evolutionary experiments is rather low: the network with three genes (two for importing precursors and one for metabolizing those) is specified with a genome of informational complexity of just 36 mers (72 bits), even though the starting genome has 1,000 "nucleotide" positions in each of the two chromosomes. The non-coding part of these initial genomes thus does not contribute to the informational complexity, because changing any of these positions to any other allele cannot change the fitness of the organism (we do not take beneficial mutations into account in the fitness tests). For these non-coding nucleotides, the p_i calculated by Eq. (46) all are exactly equal ($p_i = 1/4$), guaranteeing that they do not contribute to I, as is easily checked. However, the informational complexity grows rather quickly once more and more metabolic reactions are discovered and optimized, at a pace of about 0.5 mers (1 bit) per depth step (roughly one bit per generation).

Modules from Information Theory

The path towards an understanding of the functional organization of a network in the absence of genomic information usually involves the *decomplexification* of the network, either by clustering nodes that are related in function [62], removing those nodes that are immaterial to (or redundant in) function, or analyzing the subgraph decomposition [63] as discussed further below. A particularly insightful method to decompose networks into modules—both overlapping and non-overlapping—uses information theory to estimate how much information about the original network is present in the abstract—that is, decomplexified— version, while maximizing a variable that measures the *relevance* of the abstraction. This method, sometimes called the "information-bottleneck" approach [64] was applied to biological and engineering networks by Ziv et al. [65].

The *Network Information Bottleneck* (NIB) approach attempts to replace a complex network by a simpler one while still retaining the essential aspects of the network. For example, a highly connected star topology could be replaced by a single node that represents the modular function of the star, as in Fig. 11.

The main idea of the method is that while there are many different ways in which one can collapse a topology, the optimal mapping is one where the new topology retains as much information as possible about the original



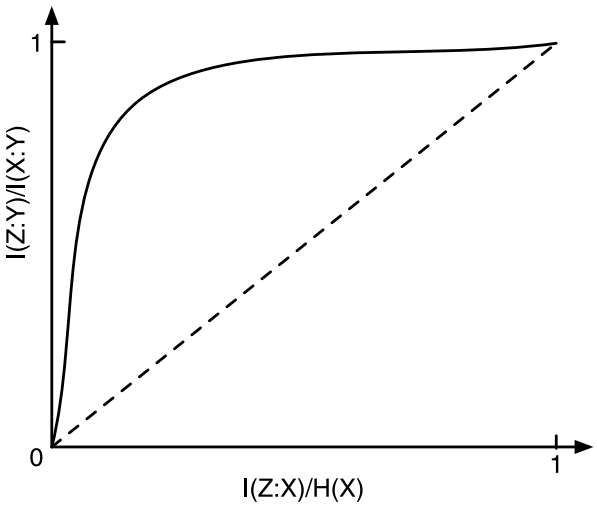
Biological Complexity and Biochemical Information, Figure 11
Collapse of a topology described by the nodes *X* to a more succinct one described by *Z* in which clusters are replaced by a cluster assignment variable *Z*

one, while maintaining as much *relevance* of the description as possible. Say, for example, that a random variable X stands for nodes $x \in X$ in a network that occur with probability $p(x)$, and total number of states $|X| = N$, where N is the size of the network. A *model* of this network can then be made by a random variable Z with *fewer* states $|Z| < N$, and where a cluster assignment is given by a set of probabilities $p(z|x)$: the probability that z is assigned to a particular cluster given the input node x. Ideally, we would like to maximize the mutual entropy (information) between the random variables X and Z, but we shall do this with a constraint given by a relevance variable mentioned earlier. This relevance variable will distinguish different ways in which clusters are assigned. In this application of the NIB to networks, the relevance variable involves diffusion on the network: those nodes that are close to each other are preferentially visited by a diffusion algorithm, and are more likely to be clustered together. In this algorithm, the relevance is represented by a random variable Y that is defined such that a diffusive process determines the probability to arrive at node y. The relation to the network variable X is given by the joint probability $p(x, y) = p(y|x)p(x)$, the probability to arrive at node y via a diffusive process given the process started at node x, times the probability that we started at node x. The latter probability is always assumed to be uniform, that is, $p(x) = 1/N$.

The NIB algorithm allows optimal solutions where several different nodes z could be assigned to any given input node x. This flexibility allows for the possibility of *overlapping* clusters or modules (soft clustering), but in the following we will follow only the algorithm for hard clustering, so that for any choice of x and z, $p(z|x)$ is either one or zero.

One version of the algorithm (agglomerative clustering) begins with a random variable Z that has exactly one fewer nodes than X, and attempts to find the optimal pair of x nodes to join to produce a model of the network with one fewer nodes. For each possible cluster assignment $p(z|x)$ we can execute a diffusion process to determine the matrix $p(y|z)$, that is, the probability to arrive at node y given a node z as starting point. We choose to merge those nodes that maximize $I(Y : Z)$, and the algorithm repeats with a set Z smaller by one node until all nodes have been merged. At each step, we can calculate the normalized variables $0 < I(Z : X)/H(X) < 1$ and $0 < I(Z : Y)/I(X : Y) < 1$ and plot them against each other, giving rise to the *information curve* [65], as in Fig. 12.

A completely random network gives rise to the diagonal in Fig. 12, and represents the least modular network. We can define a modularity score, the *network modularity*,



Biological Complexity and Biochemical Information, Figure 12
The information curve for a modular network (*solid line*), obtained by starting with a model network Z of the same size as X (maximal information $I(Z : X)/H(X) = 1$ and $I(Z : Y)/I(X : Y) = 1$, *upper right corner*), and merging nodes while maximizing $I(Z : Y)$. This process generates the information curve from the *upper right corner* all the way down to the *lower left corner*, where $|Z| = 1$ and the mutual entropy vanishes. The dashed line represents the information curve for a random network. The modularity score is given by the area under the information curve

as the area under the information curve. Perfectly modular networks then have a maximal modularity score of 1, whereas random networks have a score of 1/2. Several different networks have been analyzed using this modularity measure in Ref. [65], such as the network of coauthors for papers presented at a meeting of the American Physical Society. This network has 5,604 nodes and 19,761 edges, and yielded a modularity core of 0.9775. The regulatory network of the bacterium *E. coli* (328 nodes and 456 edges), also analyzed by these authors, yielded a score of 0.9709. Thus, both of these networks are highly modular according to this measure.

Interestingly, the network of connections of the *C. elegans* brain (see Fig. 8) has a network modularity score 0.9027, whereas a randomized version retaining the same number of nodes and edges scores 0.4984 on average, as predicted for a random network. The evolved metabolic networks discussed in Sect. "Evolution of Information in Networks" also score high on this scale. For the largest connected component of a 453 node network, we find a modularity score of 0.8486 for the 5,000*th* organism on the line of descent (about one organism per generation). While a score for small networks is fairly meaningless, the score increases slightly as the networks become more complex.

Information in Motifs

We have seen that networks that are functional are built from modules, or at least can be understood in terms of strongly connected sets of nodes that are only weakly connected to other such clusters. This clustering—carried out using information theory in the previous section—can also be performed on the basis of topology alone. For example, every network can be analyzed in terms of its *subgraph composition* [63], that is, the frequency with which particular subgraphs or motifs appear within the entire network. The degree with which certain motifs are overutilized—and some others underutilized—compared to a uniform distribution or one obtained from a random network, reflects the *local structure* of the network and can be used to classify networks from very different realms into similar categories [66].

Subgraph abundances can also be used to study the modular composition of networks, in analogy to the modular composition of sentences in written language. Meaning can be conveyed in text only because the utilization frequency of letters in words, and words in sentences, is different from uniform. For example, the letters e,t,a,i,o, and n appear in decreasing frequency in an average text written in English, while the rank-abundance distribution of the words follows a scale-free distribution (Zipf's Law [67,68]). If we assume that a random sequence of letters contains no information, then the deviation from the uniform distribution could be used to distinguish and perhaps classify functional (that is, meaningful) text from gibberish. 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 *subgraph information content* as follows.

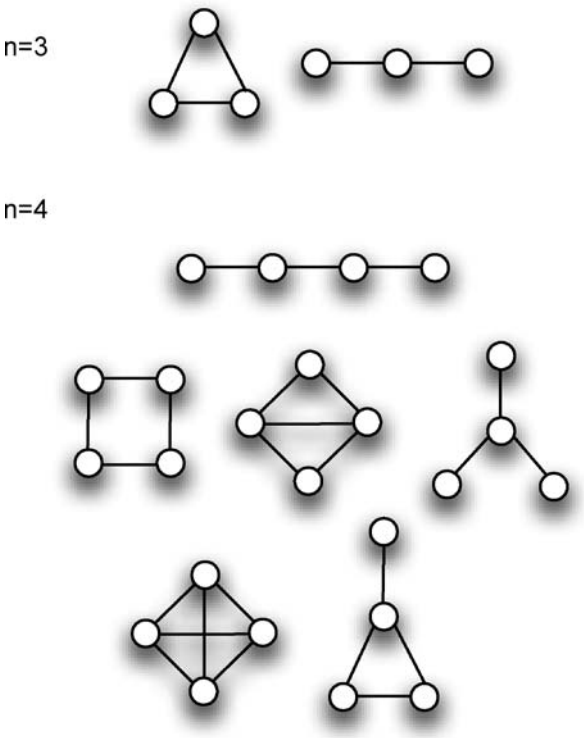
Suppose we compute the probability to find any of the two possible motifs that can be made of three nodes (see Fig. 13). For simplicity, we are considering here only undirected graphs, and do not allow self-interactions, that is, nodes that link to themselves. We can then compare these empirical probabilities to the probabilities with which these subgraphs appear in a random network.

A priori, we might think that any of the two motifs of size $n = 3$ should appear with equal probability in a random network, giving rise to a motif entropy that is maximal:

$$H_3(\tfrac{1}{2}, \tfrac{1}{2}) = -\tfrac{1}{2}\log_2 \tfrac{1}{2} - \tfrac{1}{2}\log \tfrac{1}{2} = 1\,. \tag{47}$$

Here, we defined the size-n motif entropy

$$H_n(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log_m p_i\,, \tag{48}$$



Biological Complexity and Biochemical Information, Figure 13
Undirected motifs of size $n = 3$ and $n = 4$, without self-interaction

where m is the number of possible connected motifs of size n, and the p_i are the probabilities to find the ith motif in the network. (Because the base of the logarithm is also m, this entropy is normalized to lie between zero and one.) The information stored within $n = 3$-motifs would then be (the superscript (u) refers to the uniform baseline distribution)

$$I_3^{(\mathrm{u})} = H_3(\tfrac{1}{2}, \tfrac{1}{2}) - H_3(p_1, p_2) = 1 - H_3(p_1, p_2)\,, \tag{49}$$

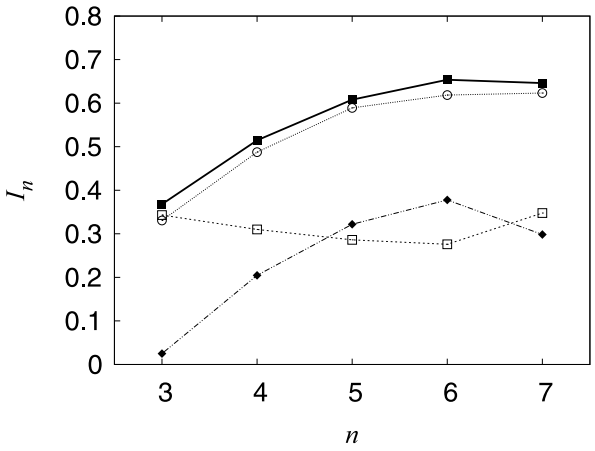
while the information stored in n-motifs is naturally

$$\begin{aligned} I_n^{(\mathrm{u})} &= H_n\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) - H_n(p_1, \ldots, p_m) \\ &= 1 - H_n(p_1, \ldots, p_m)\,. \end{aligned} \tag{50}$$

However, even random networks do not have a uniform distribution of motifs, and we can instead consider the information stored in motifs compared to a random network as baseline, as (for $n = 3$)

$$I_3^{(\mathrm{r})} = H_3(p_1^{(\mathrm{r})}, p_2^{(\mathrm{r})}) - H_3(p_1, p_2)\,. \tag{51}$$

where the $p_i^{(\mathrm{r})}$ refer to the probability of finding motif i in a random network.



Biological Complexity and Biochemical Information, Figure 14
Motif information I_n for motifs of size n for evolved networks with a uniform distribution as a baseline (*solid line, filled squares*), an Erdös–Rényi network of the same size and number of edges and connectivity $p \approx 0.005$ (*dashed line, open squares*), and a randomized network of the same size with a scale-free edge distribution (*dotted line, open circles*). The motif information using the probability distribution of a random network as the baseline Eq. ($I_n^{(r)}$) is the difference between the solid and dashed lines (*dash-dotted line, filled diamonds*)

In Fig. 14, we see the information $I_n^{(u)}$ stored in motifs of size n for $n = 3 - 7$ (filled squares, solid line), for a single, evolved, functional, simulated metabolic network discussed in Sect. "Evolution of Information in Networks", of 598 nodes. The network information increases as more complex motifs (larger n) are used for encoding, but appears to stabilize. This behavior mirrors the statistics of n-gram entropies in English text, as noted early on by Shannon [68]. Note that because the shortest path between any two nodes is on average of the order of 4-5 in these networks [59], motifs of size 7 or larger are not well-sampled.

We can study whether the network information is dictated by functionality, edge distribution, or both, by constructing analogous networks that have the functionality removed by randomizing connections but keeping the scale-free edge distribution, and by randomizing the network but destroying also the edge distribution. If, for example, we randomize the connections in our functional evolved network while keeping the scale-free edge distribution, we find that the network information is only slightly lowered (open circles and dotted line in Fig. 14). On the other hand, if we randomize the network in such a way that the degree distribution is that of a random graph (but still keeping the same number of nodes and edges), the dependence of the network information as a function of the subgraph size is markedly different (open squares, dashed line in Fig. 14). This suggests that the network information is significantly dictated by the biological scale-free distribution of edges per nodes, and only weakly by the actual function of the network.

Because random graphs do not utilize subgraphs with equal probability (as witnessed by the non-zero network information in Erdös–Rényi networks), it is more appropriate to use the random network probabilities as the baseline to calculate the network information. In this case, we see the network information $I_n^{(r)}$ increase from small values at $n = 3$ up to $n = 6$ (dash-dotted line, filled diamonds in Fig. 14). The decrease noted for $n = 7$ is due to incomplete sampling of size $n = 7$ networks for this small graph (598 nodes), and is not significant.

In summary, network information as measured by subgraph "n-gram" entropies behaves similar to the dependence of n-gram entropies in written language, and can be used to distinguish functional networks from random ones. However, the network information appears to be controlled mostly by the form of the edge distribution. Insight into the modular structure of networks will likely depend on understanding how the subgraphs of networks are assembled into modules, in analogy to how letters are assembled into words in written text.

Future Directions

Several billions of years of evolution have shaped our biosphere to become the complicated, interdependent, hierarchical complex system we witness today. Among the parts and members of this system, there are certainly differences in complexity—some obvious to any observer—some less so. Structural complexity, the most intuitive of all measures, is notoriously difficult to define because there is no universal system to rank all possible physical structures. We have seen that automata theory and information theory allow us to quantify the complexity of a sequence in terms of its information content about the environment within which it has evolved, and that this information is a good proxy for the functional complexity of the organism precisely because the information is used by the organism to function in a complex environment.

But the measure is limited because so far it is only practical for short stretches of DNA or single proteins. Thus, a quantitative measure for the complexity of a whole genome using information theory will only be possible when a large number of complete genomes of closely related species is available. Another shortcoming of the informational complexity is that it refers to a particular niche only, and furthermore cannot quantify the complexity of genes that are adapted to varying environments. So, for example, so far we cannot use the informational complex-

ity to estimate the complexity of an ecosystem, nor of an organism that spends part of its lifecycle in one environment, and another part in a completely different one (simple forms of life such as arboviruses are notorious for such a cycle). However, a natural extension of the informational complexity exists that may cover multiple environments both in time and space. Recall that the informational complexity of an ensemble of sequences S refers to a single environment description $E = e$:

$$I(S : e) = L - H(S|e)\,, \tag{52}$$

where $H(S|e)$ is the ensemble entropy of the sequences. We can generalize this expression by promoting the environment to a true random variable E that can take on states e_i with probability $p(e_i)$. This formalism can describe environments that are composed of different (spatially separated or overlapping) niches e_i, as well as environments that take on the different states e_i periodically or even randomly, in time. The informational complexity then becomes

$$I(S : e) \to \sum_e p(e) I(S : e) = I(S : E) = L - H(S|E)\,. \tag{53}$$

Here, $H(S|E)$ is the average conditional entropy of the sequence ensemble S given the environment E. Whether this construction will turn out to be useful for characterizing the complexity of ecosystems or variable environments remains to be seen, as the practical obstacles are only amplified by having to measure the informational complexity in multiple environments. But at the very least this construction addresses one of the fundamental problems in assessing functional complexity that we encountered in the introduction, namely that for organisms that have adapted to be functional in a variety of environments, we can find genes that appear to show no phenotype upon knockout in the laboratory. Such genes, however, may very well show a phenotype in a particular environment that the organism encounters in the wild, and this functional capacity of an organism needs to be taken into account when assessing functional complexity. If the random variable E accounts for the multitude of environments with the correct probabilities $p(e)$, then the full functional complexity of the organism may be characterized using Eq. (53). But to apply this measure, more efforts need to be expended towards understanding the modes in which an organism functions in its native environment(s) (as opposed to the unnatural laboratory conditions that are the norm today). If such an effort is made, then as we can expect an exponential increase in sequence data in the coming years, the prospects for a general understanding of biological complexity in terms of sequence complexity are good.

Acknowledgments

I am grateful to Arend Hintze for the collaborative work in Sect. "Network Complexity", as well as numerous discussions. I am also indebted to Matthew Rupp for the analysis shown in Figs. 5 and 6. This work was supported in part by the National Science Foundations Frontiers in Integrative Biological Research grant FIBR-0527023, a Templeton Foundation research grant, and DARPA's FunBio initiative.

Bibliography

1. Lovejoy AO (1936) The great chain of being: A study of the history of the idea. Harvard University Press, Cambridge
2. Gould S (1996) Full house: The spread of excellence from Plato to Darwin. Harmony Books, New York
3. Nee S (2005) The great chain of being. Nature 435:429
4. Gould S, Lewontin R (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. Proc R Soc London B 205:581–598
5. McShea DW (1996) Metazoan complexity and evolution: Is there a trend? Evolution 50:477–492
6. Valentine J, Collins A, Meyer C (1994) Morphological complexity increase in metazoans. Paleobiology, 20:131–142
7. Bell G, Mooers A (1997) Size and complexity among multicellular organisms. Biol J Linnean Soc 60:345–363
8. Nehaniv CL, Rhodes JL (2000) The evolution and understanding of hierarchical complexity in biology from an algebraic perspective. Artif Life 6:45–67
9. McShea D (2001) The hierarchical structure of organisms: A scale and documentation of a trend in the maximum. Paleobiology 27:405–423
10. Szostak JW (2003) Functional information: Molecular messages. Nature 423:689
11. McShea DW (2000) Functional complexity in organisms: Parts as proxies. Biol Philosoph 15:641–668
12. Britten RJ, Davidson EH (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol 46:111–138
13. Cavalier-Smith T (1985) Eukaryotic gene numbers, non-coding DNA and genome size. In: Cavalier-Smith T (ed) The evolution of genome size. Wiley, New York, pp. 69–103
14. Gregory TR (2004) Macroevolution, hierarchy theory, and the c-value enigma. Paleobiology 30:179–202
15. Gregory TR (2005) Genome size evolution in animals. In: Gregory TR (ed) The evolution of the genome. Elsevier, San Diego, pp. 3–87
16. Badii R, Politi A (1997) Complexity: Hierarchical structures and scaling in physics, Cambridge Nonlinear Science Series, vol. 6. Cambridge University Press, Cambridge (UK)
17. Kolmogorov A (1965) Three approaches to the quantitative definition of information. Probl Inf Transm 1:4
18. Li M, Vitanyi P (1997) An introduction to Kolmogorov complexity and its applications. Springer, New York
19. Adami C, Cerf NJ (2000) Physical complexity of symbolic sequences. Physica D 137:62–69

20. Gell-Mann M, Lloyd S (1996) Information measures, effective complexity, and total information. Complexity 2:44–52
21. Shannon C, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana
22. Quastler H (ed) (1953) Information theory in biology. University of Illinois Press, Urbana
23. Gatlin L (1972) Information theory and the living system. Columbia University Press, New York
24. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, et al (1994) Linguistic features of noncoding DNA sequences. Phys Rev Lett, 73:3169–3172
25. Schmitt AO, Herzel H (1997) Estimating the entropy of DNA sequences. J Theor Biol 188:369–377
26. Weiss O, Jimenez-Montaño MA, Herzel H (2000) Information content of protein sequences. J theor Biol, 206:379–386
27. Herzel H, Ebeling W, Schmitt AO (1994) Entropy of biosequences: The role of repeats. Phys Rev E 50:5061–5071
28. Shannon C (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423, 623–656
29. MacKay DJC (2002) Information theory, inference and learning algorithms. Cambridge University Press, Cambridge
30. Adami C (2004) Information theory in molecular biology. Phys Life Rev 1:3–22
31. Grassberger P (1986) Toward a quantitative theory of self-generated complexity. Int J Theor Phys 25:907–938
32. Bernaola-Galvan P, Roman-Roldan R, Oliver J (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. Phys Rev E 53:5181–5189
33. Sprinzl M, Horn C, Brown M, Ioudovitch A, Steinberg S (1998) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res, 26:148–153
34. Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucl Acids Res, 22:2079–2088
35. Korber BT, Farber RM, Wolpert DH, Lapedes AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. Proc Natl Acad Sci USA 90:7176–7180
36. Clarke ND (1995) Covariation of residues in the homeodomain sequence family. Protein Sci, 4:2269–2278
37. Atchley WR, Wollenberg KR, Fitch WM, Terhalle W, Dress AW (2000) Correlations among amino acid sites in bhlh protein domains: an information theoretic analysis. Mol Biol Evol 17:164–178
38. Wang LY (2005) Covariation analysis of local amino acid sequences in recurrent protein local structures. J Bioinform Comput Biol 3:1391–1409
39. Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular systematic, 2nd edn, Sinauer, Sunderland, pp. 407–514
40. Wolf JB, Brodie III ED, Wade MJ (eds) (2000) Epistasis and the evolutionary process. Oxford University Press, Oxford
41. Bridgham JT, Carroll SM, Thornton JW (2006) Evolution of hormone-receptor complexity by molecular exploitation. Science 312:97–101
42. Cowperthwaite MC, Bull JJ, Ancel Meyers L (2006) From bad to good: Fitness reversals and the ascent of deleterious mutations. PLoS Comput Biol 2:e141
43. Finn RD et al (2006) Pfam: Clans, web tools and services. Nucleic Acids Res 34:D247–D251
44. Brenner SE, Chothia C, Hubbard TJP (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA, 95:6073–6078
45. Miller GA, Madow WG (1954) On the maximum likelihood estimate of the Shannon-Wiener measure of information. Technical Report 54–75, Air Force Cambridge Research Center, Bedford
46. Basharin GP (1959) On a statistical estimate for the entropy of a sequence of independent random variables. Theory Probab Appl 4:333–337
47. Zurek WH (1990) Algorithmic information content, Church-Turing thesis, physical entropy, and Maxwell's demon. In: Zurek WH (ed) Complexity, entropy, and the physics of information. SFI Studies in the Sciences of Complexity, vol. 8 Addison-Wesley. Redwood City pp. 73–89
48. Cover TM, Thomas JA (1991) Elements of Information Theory. John Wiley, New York
49. Adami C (1998) Introduction to Artificial Life. Springer, New York
50. Adami C (2006) Digital genetics: Unravelling the genetic basis of evolution. Nat Rev Genet 7:109–118
51. Adami C, Ofria C, Collier T (1999) Evolution of biological complexity. Proc Natl Acad Sci USA 97:4463–4468
52. Ofria C, Huang W, Torng E (2008) On the gradual evolution of complexity and the sudden emergence of complex features. Artif Life 14, to appear
53. Carothers JM, Oestreich SC, Davis JH, Szostak JW (2004) Informational complexity and functional activity of RNA structures. J Amer Chem Soc, 126:5130–5137
54. Hazen RM, Griffin PL, Carothers JM, Szostak JW (2007) Functional information and the emergence of biocomplexity. Proc Natl Acad Sci USA 104:8574–8581
55. Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101–113
56. Schlosser G, Wagner GP (eds) (2004) Modularity in development and evolution. University of Chicago Press, Chicago, IL
57. Callebaut W, Rasskin-Gutman D (eds) (2005) Modularity: Understanding the development and evolution of natural complex systems. MIT Press, Cambridge, Mass
58. Reigl M, Alon U, Chklovskii DB (2004) Search for computational modules in the C. elegans brain. BMC Biol 2:25
59. Hintze A, Adami C (2008) Evolution of complex modular biological networks. PLoS Comput Biol 4:e23
60. Batagelj V, Mrvar A (2003) Pajek: Analysis and visulaization of large networks. In: M Jünger PM (ed) Graph Drawing Software. Springer, Berlin, pp. 77–103
61. Huang W, Ofria C, Torng E (2004) Measuring biological complexity in digital organisms. In: Pollack J, Bedau MA, Husbands P, Ikegami T, Watson R (eds) Proceedings of Artificial Life IX, MIT Press, Cambridge, pp. 315–321
62. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442
63. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D et al. (2002) Network motifs: simple building blocks of complex networks. Science 298:824–827
64. Tishby N, Pereira F, Bialek W (1999) The information bottleneck method. In: Hajek B, Sreenivas RS (eds) Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing, University of Illinois Press, Champaign, IL, pp. 368–377
65. Ziv E, Middendorf M, Wiggins CH (2005) Information-theoretic approach to network modularity. Phys Rev E 71:046117

66. Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S et al. (2004) Superfamilies of evolved and designed networks. Science, 303:1538–1542
67. Zipf GK (1935) The psycho-biology of languages. Houghton-Mifflin, Boston
68. Shannon CE (1951) Prediction and entropy of printed English. Bell System Tech J 30:50–64

Biological Data Integration and Model Building

James A. Eddy[1], Nathan D. Price[1,2,3]
[1] Department of Bioengineering, University of Illinois, Urbana-Champaign, USA
[2] Department of Chemical and Biomolecular Engineering, University of Illinois, Urbana-Champaign, USA
[3] Institute for Genomic Biology, University of Illinois, Urbana-Champaign, USA

Article Outline

Glossary

Constraint-based analysis A modeling framework based on excluding infeasible network states via environmental, physicochemical, and regulatory constraints to improve predictions of achievable cellular states and behavior.

Data space Multidimensional space containing all possible states of a system; this space can be reduced using defined constraints.

Interaction network A graph where the nodes represent biomolecules (e. g. genes) and the edges represent defined interactions between the nodes, whether they be direct physical interactions (e. g. protein–protein binding, protein–DNA binding) or functional relationships (e. g. synthetic lethality).

Biochemical reaction network Collection of metabolic, signaling, or regulatory chemical reactions described in stoichiometric detail.

Statistical inferrence network A network model designed from statistical inference from large-scale biological data sets to be quantitatively predictive for novel perturbations and/or environmental conditions.

Genome The complete DNA nucleotide sequence in all chromosomes of an organism.

Transcriptome The complete set of RNA transcripts produced from an organism's genome under a particular set of conditions.

Proteome The complete set of expressed proteins produced by the genome.

Metabolome The complete set of small molecules which are the intermediates and products of an organism's metabolism.

Boolean network A set of N discrete-valued variables, $\sigma_1, \sigma_2, \ldots, \sigma_N$ where $\sigma_n \in \{0, 1\}$. To each node a set of k_n nodes, $\sigma_{n_1}, \sigma_{n_2}, \ldots, \sigma_{n_{k_n}}$ is assigned, which controls the value of σ_n through the equation $\sigma_n(t+1) = f_n(\sigma_{n_1}(t), \ldots, \sigma_{n_{k_n}}(t))$. In the case of Boolean networks, the functions f_n can be chosen from the ensemble of all possible Boolean functions.

Definition of the Subject

Data integration and model building have become essential activities in biological research as technological advancements continue to empower the measurement of biological data of increasing diversity and scale. High-throughput technologies provide a wealth of global data sets (e. g. genomics, transcriptomics, proteomics, metabolomics), and the challenge becomes how to integrate this data to maximize the amount of useful biological information that can be extracted. Integrating biological data is important and challenging because of the nature of biology. Biological systems have evolved over the course of billions of years, and in that time biological mechanisms have become very diverse, with molecular machines of intricate detail. Thus, while there are certainly great general scientific principles to be distilled – such as the foundational evolutionary theory – much of biology is found in the details of these evolved systems. This emphasis on the details of systems and the history by which they came into being (i. e. evolution) are distinct features of biology as a science, and influence the need for large-scale data integration. Also, biological systems are responsive to vary-