PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Discussion

**Cite this article:** Koonin EV. 2016 The meaning of biological information. *Phil. Trans. R. Soc. A* **374**: 20150065.
http://dx.doi.org/10.1098/rsta.2015.0065

Accepted: 27 July 2015

One contribution of 21 to a theme issue 'DNA as information'.

**Subject Areas:**
biophysics, complexity

**Keywords:**
information, meaning, evolution, selfish elements

**Author for correspondence:**
Eugene V. Koonin
e-mail: koonin@ncbi.nlm.nih.gov

THE ROYAL SOCIETY PUBLISHING

# The meaning of biological information

Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Biological information encoded in genomes is fundamentally different from and effectively orthogonal to Shannon entropy. The biologically relevant concept of information has to do with 'meaning', i.e. encoding various biological functions with various degree of evolutionary conservation. Apart from direct experimentation, the meaning, or biological information content, can be extracted and quantified from alignments of homologous nucleotide or amino acid sequences but generally not from a single sequence, using appropriately modified information theoretical formulae. For short, information encoded in genomes is defined vertically but not horizontally. Informally but substantially, biological information density seems to be equivalent to 'meaning' of genomic sequences that spans the entire range from sharply defined, universal meaning to effective meaninglessness. Large fractions of genomes, up to 90% in some plants, belong within the domain of fuzzy meaning. The sequences with fuzzy meaning can be recruited for various functions, with the meaning subsequently fixed, and also could perform generic functional roles that do not require sequence conservation. Biological meaning is continuously transferred between the genomes of selfish elements and hosts in the process of their coevolution. Thus, in order to adequately describe genome function and evolution, the concepts of information theory have to be adapted to incorporate the notion of meaning that is central to biology.

## 1. Entropy, information, meaning and genome evolution

One of the most common, textbook concepts in biology is that the genome encodes information on the

organism—or synonymously, the genotype encodes information about the phenotype. Genomes are ultimately strings of symbols, and this digital organization is naturally interpretable within the framework of standard information theory [1,2]. The classical Shannon formula for the mean entropy (often interpreted as information content) per position of a nucleotide (or amino acid) sequence of length $L$ be written as

$$H(L) = -\sum_{j=1}^{\alpha} f_{jL} \log_{\alpha} f_{jL}, \tag{1.1}$$

where $f_j$ is the frequency of the base $j$ ($j = \mathrm{A, T, G, C}$) in the given sequence, and $\alpha$ is the size of the alphabet (four in the case of nucleotide sequences and 20 for amino acid sequences). Applied this way, entropy only tells us how far the count of each base in the given sequence deviates from the random expectation $L/\alpha$ which does not convey any meaningful message on the genome in question let alone about the phenotype of the organism it is supposed to encode. Clearly, the message encoded in the genome is of a different nature. Within the classical information theory, the quantity we are interested in is not entropy but rather information (more precisely, information gain) that is obtained about a sequence $L$ as a result of some procedure that we will call measurement:

$$I(L) = H(L)_0 - H(L)_{\mathrm{m}}, \tag{1.2}$$

where $H(L)_0$ is the entropy of the sequence before the measurement and $H(L)_{\mathrm{m}}$ is the entropy of the same sequence after the measurement. An insightful discussion of the crucial distinction between information and entropy is presented by Adami in this theme issue of the *Philosophical Transactions* [3].

In qualitative terms, biological information is perhaps best described as the 'meaning' of a sequence. A nucleotide sequence assumes meaning only when it is either transcribed into a RNA molecule that directly carries out a biological function, or transcribed into a mRNA that is then translated into a functional protein, or else the DNA itself interacts with proteins or RNA molecules resulting in a functional (often, regulatory) effect.

What kind of measurements can yield biological information and allow one to quantify the meaning of genomic sequences? Obviously, one of the means to this end is direct experimentation. However, exhaustive characterization of the biological roles of each nucleotide in the genome is unrealistic even for the smallest model genomes such as those of viruses, let alone the expansive genomes of complex organisms such as animals and plants. Moreover, quantitative comparison of the 'meaningfulness' of different sites in the genome requires a whole other level of experimentation whereby the phenotypic (fitness) effects of changes in each site are measured in competition experiments. This type of experiment is central to experimental evolution research [4,5] but the complexity of bringing it to the genome scale far exceeds any imaginable laboratory capabilities.

Hence the alternative approach to information measurement involves extracting meaning from sequences themselves. A single genomic sequence is largely meaningless. The meaning of certain short nucleotide signals has been known for many years from multiple, definitive experiments. The most prominent signals of this type are the translation start and stop codons that mark protein-coding genes. A simple estimate shows that the presence of a long open reading frame between a start and a stop signal is highly unlikely, and therefore, at least for intronless genomes, protein-coding regions can be predicted reliably [6]. This does tell us something important about the meaning of the genome sequence, by delineating regions that most likely are used to produce proteins.

However, the only general way to extract meaning from sequences involves comparative analysis of homologues. The premises are extremely simple, yet powerful. The great majority of the meaningful sites in nucleotide sequences, i.e. those sites that contribute to biological function, are subject to purifying selection, hence evolutionary conservation of meaningful sites. The stronger the selection, the more meaningful ('important') a site is. These simple considerations allow one to naturally quantify meaningful information contained in sequences [7–10].

For an alignment of orthologous sequences, the Shannon entropy formula (1.1) can be re-written as follows:

$$H(L)_n = \sum_{i=1}^{L} H_i = -\sum_{i=1}^{L} \sum_{j} f_{ij} \log f_{ij}, \tag{1.3}$$

where $H(L)_n$ is the total entropy of the alignment of $n$ sequences of length $L$; $H_i$ is the per site entropy; and $f_{ij}$ are the frequencies of each of the four nucleotides ($j = \mathrm{A, T, G, C}$) or each of the 20 amino acids in site $i$. Clearly, for a fully conserved site $H(i) = 0$, whereas for a completely random site $H(i) = 1$; accordingly, the values of $H(L)_n$ are between 0 and $L$. Note that equation (1.3) is equivalent to equation (1.1) except that instead of applying the Shannon formula 'horizontally', i.e. to a single sequence, we now apply it 'vertically', i.e. to an alignment of homologous sequences. This definition of entropy is consistent both with Boltzmann's famous statistical definition of entropy and with Shannon entropy (information content) and thus can be legitimately denoted 'evolutionary entropy' of a set of aligned sequences. In addition to being physically valid, evolutionary entropy seems to make perfect biological sense: low-entropy sites are most strongly conserved, and by inference, most functionally important (meaningful).

Then, using formula (1.3), 'biological (evolutionary) information' of a genome can be defined as

$$I(N) = N - \sum_{i=1}^{k} H(L_i), \tag{1.4}$$

and 'biological (evolutionary) information density' can be calculated as

$$D(N) = \frac{I(N)}{N} = \frac{N - \sum_{i=1}^{k} H(L_i)}{N} = 1 - \frac{\sum_{i=1}^{k} H(L_i)}{N}, \tag{1.5}$$

where $N$ is the total length (number of sites) in a genome; $L_i$ is the length of a genomic segment that is subject to measurable selection (such as a protein-coding or RNA-coding gene); $k$ is the number of such alignable segments in the genome; and $H(L_i)$ is the evolutionary entropy for the segment $L$ calculated using formula (1.2). Previously, the quantity defined by equation (1.4) has been denoted 'biological complexity' but at least for the purpose of the present discussion, 'biological (evolutionary) information' seems to be a more straightforward definition. The values of $I(N)$ are between 0 and $N$, and equation (1.4) is equivalent to equation (1.2), i.e. biological information is the information gain that can be extracted from an alignment of homologous sequences through the constraint on change in 'meaningful' positions. Indeed, biological information density is directly related to meaning: sites and sequences with the highest values of $D(N)$ are the most meaningful ones. Thus, Dobzhansky's famous dictum 'Nothing in biology makes sense except in light of evolution' [11] takes a literal, even technical interpretation: biological meaning (sense) effectively cannot be gleaned by any means other than direct evolutionary analysis.

To conclude this conceptual discussion, it seems pertinent to ask: what is biological information about? It has been persuasively argued that the genome stores information about the environment, allowing the organism to predict and exploit environmental changes [10,12]. Although environmental interactions certainly are an important part of the genomic information content, it seems prudent to indicate that another key part is about the (nearly) universal aspects of cellular and organismal design. A notable evidence of the universality of cellular design comes from the consistent observations on the universal scaling of different functional categories of genes with the total gene count in all cellular life forms [13,14]. The genes encoding universal components of cells, such as RNA and proteins that constitute the translation system, are endowed with by far the most pronounced meaning (i.e. evolutionary conservation) than genes that are involved in environmental interactions [15,16].

## 2. Informational and entropic genomes and evolution of organismal complexity

The exact values of $H$ are difficult to calculate for complete genomes because the distribution of evolutionary constraints is never known precisely [16]. Furthermore, there is always arbitrariness in the choice of orthologues to be included in the alignment for the calculation, and most important, the sequences of orthologous genes are actually not independent but rather are connected by an evolutionary tree. Thus, to produce accurate estimates of biological information density, an appropriate weighting scheme taking into account the evolutionary tree topology and branch lengths is required. However, these details are not essential if one is interested only in ballpark estimates. The fraction of sites under selection across the genome has been estimated with reasonable precision for some model organisms such as humans or *Drosophila* [16–18]. For others, particularly prokaryotes and unicellular eukaryotes, the fraction of coding nucleotides plus the estimated fraction of regulatory sites can be taken as a reasonable approximation; for sites under selection, $H_i = 0.5$ can be taken to approximate the mean entropy value.

Comparison of the estimates of $H(N)$, $I(N)$ and $D(N)$ for genomes of different life forms reveals a paradox. The total biological information $I(N)$ (arguably, the measure of biological complexity) monotonically increases with the genome size, in particular, in multicellular eukaryotes compared to prokaryotes, but the entropy $H(N)$ increases dramatically faster, and as the result, the evolutionary information density $D(N)$ sharply drops (figure 1). Thus, the genomes of organisms that are usually perceived as the most complex, such as animals and plants, indeed have the highest total information content but also are 'entropic' genomes with a low biological information density. By contrast, organisms that we traditionally think of as primitive, such as bacteria, have 'informational' genomes with high information density [9]. To rephrase the same statement more provocatively, the genomes of unicellular organisms and viruses appear to be incomparably 'better designed' than the genomes of plants or particularly animals. Certainly, this conundrum is already apparent in a simple comparison of the genome architectures of multicellular and unicellular organisms, with the former being dominated by non-coding sequences (introns and intergenic regions), whereas the latter are 'wall to wall' genomes that are almost completely comprised of genes [19]. Nevertheless, the formal approach to biological information outlined above allows one to emphasize and quantify the differences between the informational landscapes of different life forms.

The primary cause behind the low information density of the genomes of the complex life forms seems to follow directly from straightforward population genetic theory [20–22]. In populations with a small effective size that are characteristic of complex multicellular organisms, the weak purifying selection and the high intensity of genetic drift preclude efficient purging of meaningless sequences and conversely allow proliferation of such sequences, in particular, various mobile elements. Evolutionary and functional plasticity is the other side of the same coin [16]. This plasticity is manifested in the numerous demonstrated cases of recruitment of mobile element sequences and other originally 'meaningless' sequences for biological functions. What matters for the evolution of phenotypic (organismal) complexity appears to be the total biological information content of the genome rather than information density ('design'). I discuss these aspects of biological information in the following sections.

## 3. Junk DNA or sequences with fuzzy meaning?

The now well-established phenomenon of pervasive transcription [23–25], that has triggered the (in)famous debate around the results of the ENCODE project [26–30], when pitted against the formal considerations outlined above, suggests a radical line of thinking on the nature of biological information and meaning. The indisputable findings that (nearly) all sequences in complex genomes, such as human, are transcribed at some level (at least in some cell types and at some life stages) most likely fit the same population genetic perspective [20–22]. Conceivably, transcription is pervasive because selection against spurious promoters and enhancers is not sustainable in small populations subject to drift. However, in the context of the above

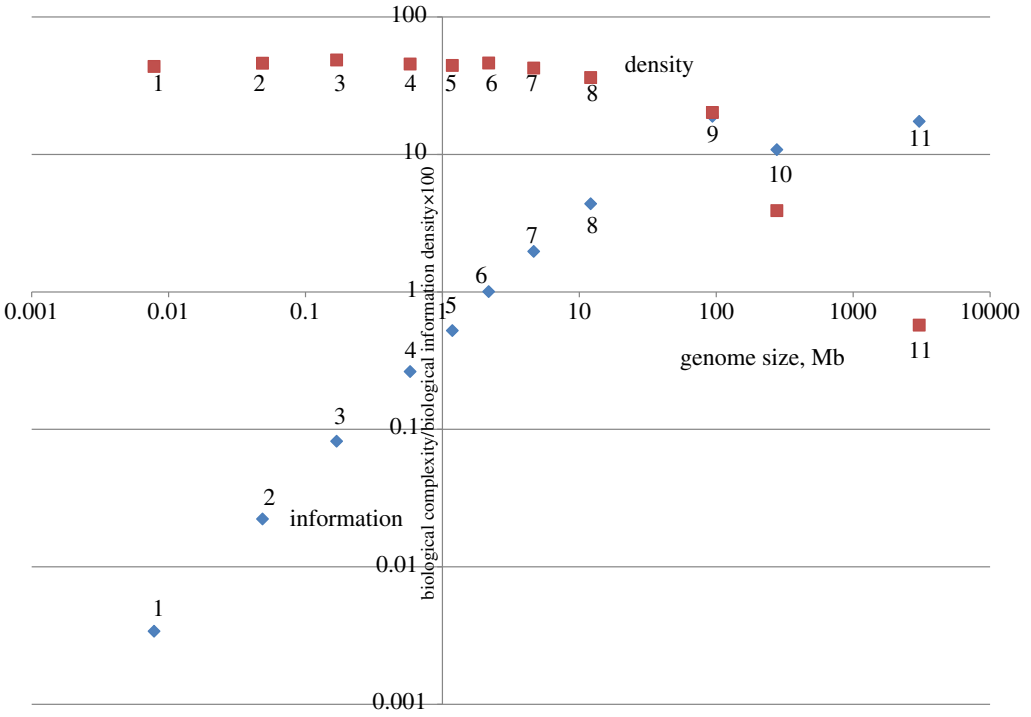


**Figure 1.** Biological information and information density depending on genome size: viruses, prokaryotes and eukaryotes. The biological information and density values were calculated using equations (1.4) and (1.5), respectively, and the data on genomes were from Genbank. The plot is on a double logarithmic scale. 1, encephalomyocarditis virus (RNA virus); 2, lambda phage; 3, T4 phage; 4, *Mycoplasma genitalium* (parasitic bacterium); 5, acanthamoeba polyphaga mimivirus (giant virus); 6, *Archaeoglobus fulgidus* (free-living archaeon); 7, *Escherichia coli* (free-living bacterium); 8, *Saccharomyces cerevisiae*; 9, *Arabidopsis thaliana*; 10, *Drosophila melanogaster*; 11, *Homo sapiens*. (Online version in colour.)

formalization of biological information, would it be appropriate to view most of the sequences in complex genomes, with (extremely) low biological information density, as being endowed with 'fuzzy meaning' (figure 2)? Operationally, sequences with fuzzy meaning can be defined as those that cannot be aligned between genomes that have diverged beyond the threshold of sequence conservation that is due to common ancestry alone, e.g. after the synonymous sites in protein-coding genes have reached saturation. This is a conservative definition because it assumes neutrality of the synonymous sites that in actuality are subject to selection albeit substantially weaker than that affecting non-synonymous sites [31–33].

The exact sequences of genomic regions with fuzzy meaning are (almost) inconsequential but their expression has meaning that can be rationalized at least at two levels. First, the sequences with fuzzy meaning form the material basis of plasticity from which functional molecules, primarily but not exclusively, regulators of various processes, are continuously recruited to assume better defined meaning, a process that can be denoted 'gain of meaning'. Second, although numerous sequences might not encompass any specific meaning whatsoever, their transcription itself could be meaningful, in particular, for maintaining particular chromatin states that in turn regulate transcription of regions with specific meaning (genes) [24]. The information flow between the domains of defined meaning and fuzzy meaning certainly is a two-way street: loss of meaning continuously occurs, e.g. in the process of pseudogenization.

The boundary between the genomic sequences with well-defined meaning and those with fuzzy meaning is not necessarily sharp. The long non-coding (lnc) RNAs that recently have been identified in abundance in mammals [34–36] appear to bridge the islands of highly meaningful

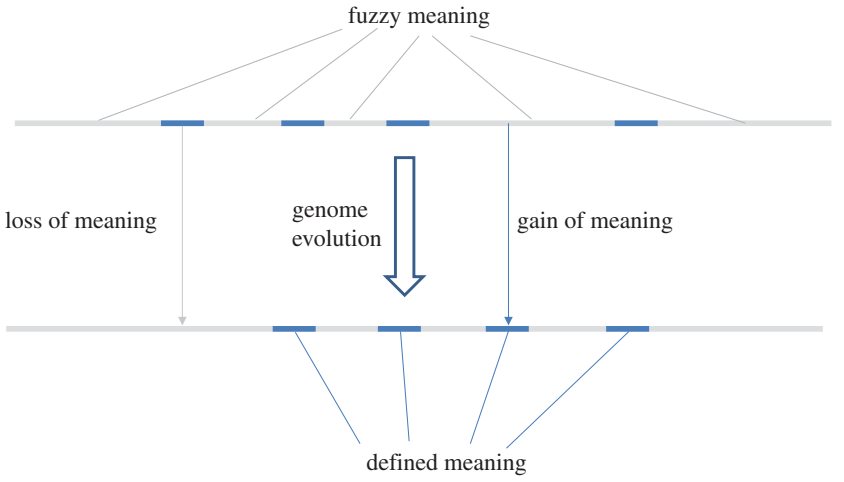


**Figure 2.** The fuzzy meaning concept and gain and loss of meaning. The cartoon schematically shows a fragment of a genome of a complex multicellular organism (animal or plant) that consists mostly of sequences with fuzzy meaning, interspersed with 'islands' of defined meaning such as genes (exons) encoding structural RNAs and proteins as well as evolutionarily conserved regulatory elements. (Online version in colour.)

protein-coding regions (and those that encode structural RNAs) with the sea of fuzzy meaning sequences (figure 2). In the expansive pool of lncRNAs, there are many that are represented by orthologues even in distant organisms, such as primates and rodents, although sequence conservation (biological information density) is low [35,37–40]. However, numerous lncRNAs, even within well-defined, relatively highly expressed sets [39], are lineage-specific and hence belong in the fuzzy meaning domain.

The concept of fuzzy meaning seems to reconcile two fundamental, undeniable but apparently contradictory lines of evidence: (i) in complex, large genomes, the substantial majority of the sequences is subjected to extremely weak or effectively no purifying selection and (ii) most of these apparently meaningless sequences are at least occasionally transcribed, i.e. have a distinct phenotypic manifestation. Rather than dismissing most of the genome as junk DNA [41–44], the fuzzy meaning concept seems to offer a more adequate description of this vast pool of sequences.

The relevance of fuzzy meaning for evolution, and in particular evolutionary innovations, does not appear to be limited to non-coding DNA or to large, complex genomes. It has been observed that most of the novel eukaryotic proteins adopt $\alpha$-helical folds and seem to have evolved from generic, repetitive coiled coil proteins [45]. Even more strikingly, evidence has been presented that in various eukaryote organisms, numerous short genes evolve from non-coding sequences through a stage of 'pre-proteins' [46–52]. This route of evolution appears to be a clear manifestation of fuzzy meaning. Upon acquiring a specific function, the evolution of these proteins substantially slowed down: their meaning was sharply defined as they left the fuzzy domain (figure 2).

## 4. The agency of biological meaning: parasite–host interaction, arms races and exaptation

Much like in human affairs and unlike in standard information theory, the meaning of genomic sequences can only be meaningfully defined if the beneficiary of the message is identified ('meaning for whom?'). The sequences of the innumerable selfish, mobile genetic elements that are integrated into the genomes of cellular lifeforms [53–57]—and represent the majority of the genome sequences in many animals and plants [58,59]—are generally meaningless for the host

organisms. For most of these elements, orthologous relationships cannot be established between any distant host species, and hence biological information density cannot be estimated from the host genome comparison. Yet, 'from the selfish element's point of view', i.e. when biological information density is estimated from an alignment of homologous sequences of elements in the same family, these sequences are densely packed with meaning.

Continuous transfer of meaning between selfish elements and hosts is a major evolutionary trend that comes in several guises. First, genes from selfish elements are often recruited by host organisms such that the specific activity of the encoded protein is modified and appropriated for host functions. Examples include the essential eukaryotic enzyme telomerase that is required for linear chromosome replication that was recruited from a bacterial retroelement (group II self-splicing intron) for its reverse transcriptase activity [60,61]; hedgehogs, key regulators of animal development, that have been derived from an intein and employ the autoprotease activity of the latter [62–64]; and syncytins, placental receptors derived from retrovirus genes [65,66]. A remarkable, common phenomenon involves what can be described as a change of biological meaning to its opposite. Under this trend, 'offensive weapons' of selfish elements are captured by the hosts and turned into means of defence [67]. Striking examples include the parallel recruitment of integrases from unrelated selfish elements for adaptive immunity systems in prokaryotes (CRISPR-Cas) and in animals [68] as well as the system of DNA elimination in the ciliate macronucleus [69]. Conversely, selfish elements, particularly viruses with comparatively large genomes, consistently capture host genes involved in defence and adopt them for counter-defence, e.g. as dominant-negative inhibitors [70–72]. Thus, sequences that have been meaningful for selfish elements become meaningful for the host and vice versa.

Another common trend in the coevolution of selfish elements and hosts is the erosion of meaning that accompanies integration of genomes. This phenomenon includes inactivation and deterioration of all kinds of mobile elements that occurs on a limited scale in bacteria and archaea but is a massive contribution to the genomes of animals and plants. Including splicesosomal introns which appear to be descendants of bacterial retrotransposons (group II self-splicing introns) [61], the majority of the DNA in animal and plant genomes is derived from mobile elements. Thus, these elements are the principal source of fuzzy meaning discussed in the preceding section.

## 5. Conclusion

The biologically relevant information is more akin to meaning than to entropy. This type of information can be quantified by applying theoretical informational concepts to aligned sequences of orthologous genes or proteins: biological information density (meaning) is defined vertically, i.e. across an alignment of homologous sequences, rather than horizontally, i.e. along a single genome. Sites with the lowest entropy have the highest biological information density or in other words, are the most meaningful ones. The meaningfulness of genomic sequences spans the entire range from sharply defined, universal meaning to effective meaninglessness. Sequences with low biological information density can be assigned to the domain of fuzzy meaning which encompasses most of the genomic sequence in animals and plants. The sequences with fuzzy meaning serve as a pool for recruitment for diverse functions and could also be involved in generic functional roles that require little if any sequence conservation. The concept of fuzzy meaning seems to capture better the status of non-conserved genomic sequences than the more rigid notion of junk DNA. The evolution of life involves the perennial arms race between parasites and hosts that involves continuous transformation of the agency of biological meaning. Sequences that are meaningful for selfish elements are appropriated by the hosts to assume meaning, particularly as means of defence, and vice versa, host genes involved in antivirus defence and other processes are recruited by selfish elements, with their meaning changed in the process. An even more common phenomenon is the erosion of meaning of selfish element genes upon integration with the host genomes. These sequences replenish the fuzzy meaning domain. In summary, meaningful analysis of genomes from an informational theoretical standpoint requires

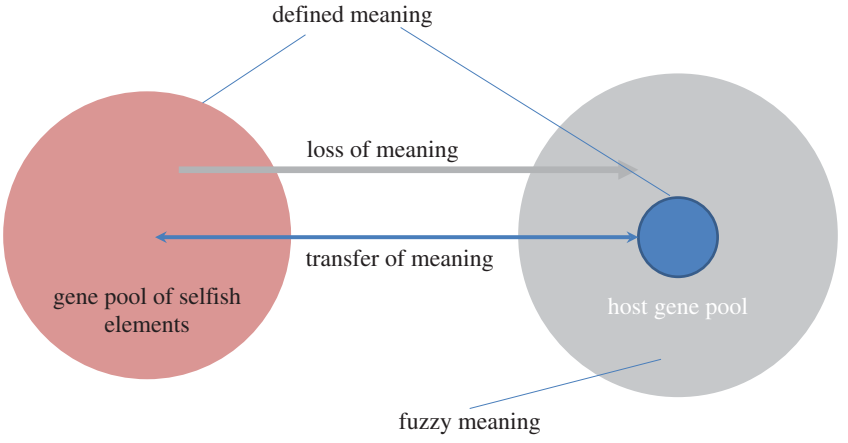


**Figure 3.** Flow of meaning between selfish elements and hosts. (Online version in colour.)

re-interpretation of the very notion of information as a concept of meaning that is specific to biology (figure 3).

Competing interests. I declare I have no competing interests.
Funding. The author's research is supported by intramural funds of the US Department of Health and Human Services (to National Library of Medicine).
Acknowledgements. I thank Yuri Wolf for help with the preparation of figure 1 and Chris Adami for invaluable discussions and critical reading of the first version of the manuscript.

## References

1. Shannon CE, Weaver W. 1949 *The mathematical theory of communication*. Chicago, IL: University of Illinois Press.
2. Pierce JR. 2012 *An introduction to information theory: symbols, signals and noise*. New York, NY: Dover Publications.
3. Adami C. 2016 What is information? *Phil. Trans. R. Soc. A* **374**, 20150230. (doi:10.1098/rsta.2015.0230)
4. Barrick JE, Lenski RE. 2013 Genome dynamics during experimental evolution. *Nat. Rev. Genet.* **14**, 827–839. (doi:10.1038/nrg3564)
5. Wiser MJ, Lenski RE. 2015 A comparison of methods to measure fitness in *Escherichia coli*. *PLoS ONE* **10**, e0126210. (doi:10.1371/journal.pone.0126210)
6. Nielsen P, Krogh A. 2005 Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21**, 4322–4329. (doi:10.1093/bioinformatics/bti701)
7. Adami C. 2002 What is complexity? *Bioessays* **24**, 1085–1094. (doi:10.1002/bies.10192)
8. Koonin EV. 2004 A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. *Cell Cycle* **3**, 280–285. (doi:10.4161/cc.3.3.745)
9. Koonin EV. 2011 *The logic of chance: the nature and origin of biological evolution.* Upper Saddle River, NJ: FT Press.
10. Adami C. 2012 The use of information theory in evolutionary biology. *Ann. NY Acad. Sci.* **1256**, 49–65. (doi:10.1111/j.1749-6632.2011.06422.x)
11. Dobzhansky T. 1973 Nothing in biology makes sense except in the light of evolution. *Amer. Biol. Teacher* **35**, 125–129. (doi:10.2307/4444260)
12. Adami C, Ofria C, Collier TC. 2000 Evolution of biological complexity. *Proc. Natl Acad. Sci. USA* **97**, 4463–4468. (doi:10.1073/pnas.97.9.4463)
13. van Nimwegen E. 2003 Scaling laws in the functional content of genomes. *Trends Genet.* **19**, 479–484. (doi:10.1016/S0168-9525(03)00203-8)
14. Koonin EV, Wolf YI. 2008 Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* **36**, 6688–6719. (doi:10.1093/nar/gkn668)

15. Koonin EV. 2003 Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136. (doi:10.1038/nrmicro751)
16. Koonin EV, Wolf YI. 2010 Constraints and plasticity in genome and molecular-phenome evolution. *Nat. Rev. Genet.* **11**, 487–498. (doi:10.1038/nrg2810)
17. Rands CM, Meader S, Ponting CP, Lunter G. 2014 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* **10**, e1004525. (doi:10.1371/journal.pgen.1004525)
18. Ponting CP, Hardison RC. 2011 What fraction of the human genome is functional? *Genome Res.* **21**, 1769–1776. (doi:10.1101/gr.116814.110)
19. Koonin EV. 2009 Evolution of genome architecture. *Int. J. Biochem. Cell Biol.* **41**, 298–306. (doi:10.1016/j.biocel.2008.09.015)
20. Lynch M. 2007 The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl Acad. Sci. USA* **104**(Suppl. 1), 8597–8604. (doi:10.1073/pnas.0702207104)
21. Lynch M. 2007 *The origins of genome architecture*. Sunderland, MA: Sinauer Associates.
22. Lynch M, Conery JS. 2003 The origins of genome complexity. *Science* **302**, 1401–1404. (doi:10.1126/science.1089370)
23. Amaral PP, Dinger ME, Mercer TR, Mattick JS. 2008 The eukaryotic genome as an RNA machine. *Science* **319**, 1787–1789. (doi:10.1126/science.1155472)
24. Berretta J, Morillon A. 2009 Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* **10**, 973–982. (doi:10.1038/embor.2009.181)
25. Jacquier A. 2009 The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.* **10**, 833–844. (doi:10.1038/nrg2683)
26. Dunham I *et al.* 2012 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74. (doi:10.1038/nature11247)
27. Graur D, Zheng Y, Price N, Azevedo RB, Zufall RA, Elhaik E. 2013 On the immortality of television sets: 'function' in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* **5**, 578–590. (doi:10.1093/gbe/evt028)
28. Doolittle WF. 2013 Is junk DNA bunk? A critique of ENCODE. *Proc. Natl Acad. Sci. USA* **110**, 5294–5300. (doi:10.1073/pnas.1221376110)
29. Doolittle WF, Brunet TD, Linquist S, Gregory TR. 2014 Distinguishing between 'function' and 'effect' in genome biology. *Genome Biol. Evol.* **6**, 1234–1237. (doi:10.1093/gbe/evu098)
30. Graur D, Zheng Y, Azevedo RB. 2015 An evolutionary classification of genomic function. *Genome Biol. Evol.* **7**, 642–645. (doi:10.1093/gbe/evv021)
31. Resch AM, Carmel L, Marino-Ramirez L, Ogurtsov AY, Shabalina SA, Rogozin IB, Koonin EV. 2007 Widespread positive selection in synonymous sites of mammalian genes. *Mol. Biol. Evol.* **24**, 1821–1831. (doi:10.1093/molbev/msm100)
32. Shabalina SA, Spiridonov NA, Kashina A. 2013 Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res.* **41**, 2073–2094. (doi:10.1093/nar/gks1205)
33. Hunt RC, Simhadri VL, Iandoli M, Sauna ZE, Kimchi-Sarfaty C. 2014 Exposing synonymous mutations. *Trends Genet.* **30**, 308–321. (doi:10.1016/j.tig.2014.04.006)
34. Louro R, Smirnova AS, Verjovski-Almeida S. 2009 Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics* **93**, 291–298. (doi:10.1016/j.ygeno.2008.11.009)
35. Marques AC, Ponting CP. 2009 Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol.* **10**, R124. (doi:10.1186/gb-2009-10-11-r124)
36. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. 2011 Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927. (doi:10.1101/gad.17446611)
37. Managadze D, Rogozin IB, Chernikova D, Shabalina SA, Koonin EV. 2011 Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.* **3**, 1390–1404. (doi:10.1093/gbe/evr116)
38. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC. 2012 Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* **8**, e1002841. (doi:10.1371/journal.pgen.1002841)
39. Managadze D, Lobkovsky AE, Wolf YI, Shabalina SA, Rogozin IB, Koonin EV. 2013 The vast, conserved mammalian lincRNome. *PLoS Comput. Biol.* **9**, e1002917. (doi:10.1371/journal.pcbi.1002917)

40. Diederichs S. 2014 The four dimensions of noncoding RNA conservation. *Trends Genet.* **30**, 121–123. (doi:10.1016/j.tig.2014.01.004)
41. Ohno S. 1972 So much 'junk' DNA in our genome. *Brookhaven Symp. Biol.* **23**, 366–370.
42. Doolittle WF, Sapienza C. 1980 Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603. (doi:10.1038/284601a0)
43. Orgel LE, Crick FH. 1980 Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607. (doi:10.1038/284604a0)
44. Palazzo AF, Lee ES. 2015 Non-coding RNA: what is functional and what is junk? *Front. Genet.* **6**, 2. (doi:10.3389/fgene.2015.00002)
45. Aravind L, Iyer LM, Koonin EV. 2006 Comparative genomics and structural biology of the molecular innovations of eukaryotes. *Curr. Opin. Struct. Biol.* **16**, 409–419. (doi:10.1016/j.sbi.2006.04.006)
46. Carvunis AR *et al.* 2012 Proto-genes and de novo gene birth. *Nature* **487**, 370–374. (doi:10.1038/nature11184)
47. Tautz D, Domazet-Loso T. 2011 The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702. (doi:10.1038/nrg3053)
48. Neme R, Tautz D. 2013 Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**, 117. (doi:10.1186/1471-2164-14-117)
49. Neme R, Tautz D. 2014 Evolution: dynamics of de novo gene emergence. *Curr. Biol.* **24**, R238–R240. (doi:10.1016/j.cub.2014.02.016)
50. Zhao L, Saelao P, Jones CD, Begun DJ. 2014 Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343**, 769–772. (doi:10.1126/science.1248286)
51. Palmieri N, Kosiol C, Schlotterer C. 2014 The life cycle of *Drosophila* orphan genes. *Elife* **3**, e01311. (doi:10.7554/eLife.01311)
52. Schlotterer C. 2015 Genes from scratch—the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219. (doi:10.1016/j.tig.2015.02.007)
53. McClintock B. 1984 The significance of responses of the genome to challenge. *Science* **226**, 792–801. (doi:10.1126/science.15739260)
54. Kazazian Jr HH. 2004 Mobile elements: drivers of genome evolution. *Science* **303**, 1626–1632. (doi:10.1126/science.1089670)
55. Miller WJ, Capy P. 2004 Mobile genetic elements as natural tools for genome evolution. *Methods Mol. Biol.* **260**, 1–20. (doi:10.1385/1-59259-755-6:001)
56. Goodier JL, Kazazian Jr HH. 2008 Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23–35. (doi:10.1016/j.cell.2008.09.022)
57. Piegu B, Bire S, Arensburger P, Bigot Y. 2015 A survey of transposable element classification systems—a call for a fundamental update to meet the challenge of their diversity and complexity. *Mol. Phylogenet. Evol.* **86**, 90–109. (doi:10.1016/j.ympev.2015.03.009)
58. Jurka J, Kapitonov VV, Kohany O, Jurka MV. 2007 Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genomics Hum. Genet.* **8**, 241–259. (doi:10.1146/annurev.genom.8.080706.092416)
59. Chalopin D, Naville M, Plard F, Galiana D, Volff JN. 2015 Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol. Evol.* **7**, 567–580. (doi:10.1093/gbe/evv005)
60. Gladyshev EA, Arkhipova IR. 2011 A widespread class of reverse transcriptase-related cellular genes. *Proc. Natl Acad. Sci. USA* **108**, 20 311–20 316. (doi:10.1073/pnas.1100266108)
61. Lambowitz AM, Belfort M. 2015 Mobile bacterial group II introns at the crux of eukaryotic evolution. *Microbiol. Spectr.* **3**, MDNA3-0050-2014. (doi:10.1128/microbiolspec.MDNA3-0050-2014)
62. Hall TM, Porter JA, Young KE, Koonin EV, Beachy PA, Leahy DJ. 1997 Crystal structure of a Hedgehog autoprocessing domain: homology between Hedgehog and self-splicing proteins. *Cell* **91**, 85–97. (doi:10.1016/S0092-8674(01)80011-8)
63. Perler FB. 1998 Protein splicing of inteins and hedgehog autoproteolysis: structure, function, and evolution. *Cell* **92**, 1–4. (doi:10.1016/S0092-8674(00)80892-2)
64. Burglin TR. 2008 The Hedgehog protein family. *Genome Biol.* **9**, 241. (doi:10.1186/gb-2008-9-11-241)

65. Blaise S, de Parseval N, Heidmann T. 2005 Functional characterization of two newly identified Human Endogenous Retrovirus coding envelope genes. *Retrovirology* **2**, 19. (doi:10.1186/1742-4690-2-19)
66. Cornelis G *et al.* 2015 Retroviral envelope gene captures and syncytin exaptation for placentation in marsupials. *Proc. Natl Acad. Sci. USA* **112**, E487–E496. (doi:10.1073/pnas.1417000112)
67. Koonin EV, Krupovic M. 2015 A movable defense. *The Scientist*, 1 January 2015. See http://www.the-scientist.com/?articles.view/articleNo/41702/title/A-Movable-Defense/
68. Koonin EV, Krupovic M. 2015 Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* **16**, 184–192. (doi:10.1038/nrg3859)
69. Swart EC, Nowacki M. 2015 The eukaryotic way to defend and edit genomes by sRNA-targeted DNA deletion. *Ann. N Y Acad. Sci.* **1341**, 106–114. (doi:10.1111/nyas.12636)
70. Bugert JJ, Darai G. 2000 Poxvirus homologues of cellular genes. *Virus Genes* **21**, 111–133. (doi:10.1023/A:1008140615106)
71. Seet BT *et al.* 2003 Poxviruses and immune evasion. *Annu. Rev. Immunol.* **21**, 377–423. (doi:10.1146/annurev.immunol.21.120601.141049)
72. Hughes AL, Irausquin S, Friedman R. 2010 The evolutionary biology of poxviruses. *Infect. Genet. Evol.* **10**, 50–59. (doi:10.1016/j.meegid.2009.10.001)