Commentary

THE EMBO JOURNAL

Informational limits of biological organisms

Jussi Taipale[1,2,3]

Genomic analyses have revealed that free-living biological organisms carry between 10^7 and 10^{11} bits of information in their genomes. In large organisms with relatively small population sizes, such as humans, only in the order of 1% of the genomic information is shaped by the environment via natural selection. A much larger amount of information than this is routinely being generated by biomedical researchers, and the rapidly accumulating data is often interpreted to mean that biological systems are extremely complex. However, as the genome is finite in length, it cannot define precisely optimal values for the quantitative parameters of the experimentally identified molecular phenotypes. Furthermore, because the genomic sequences orchestrate a biochemical system that is much more information-rich than the genome, the vast majority of the measured molecular phenotypes must represent "molecular spandrels", that is phenotypes that are not independent of each other, and instead co-determined by the same genomic sequences. These considerations are important in interpreting the results of individual experiments. In addition, they indicate that full understanding of biological systems requires a genome-centric model that does not abstract away the information contained in the genome, and instead explicitly maps all phenotypic data back to specific genomic sequences.

The EMBO Journal (2018) 37: e96114

Introduction

In physics, systems that appear extremely complex can be described by simple laws that themselves contain very little information. For example, the positions and velocities of atoms in a gas can be abstracted away, and the features of the macroscopic system described with high accuracy using a simple equation containing just two parameters, temperature and pressure. Biological systems have resisted such abstractions, not because they are inherently more complex than a gas, but because their structure and behavior (response to stimuli) are described by a genome, which is itself a store of information. The sequence of the genome cannot be abstracted away from a particular biological system without losing essential information about the system. To do so would be analogous to abstracting away meanings of words from a human language, and claiming that all that one needs to know is high-level concepts such as the notion that language is composed of words. We can interpret the genome to some extent, for example, determining which triplet codons encode particular amino acids. However, we cannot mostly figure out how DNA sequence determines gene expression, or where in DNA abstract concepts such "fear of snakes" are written. Nevertheless, information at different levels of abstraction is written in a molecule inside our cells, using an alphabet that we can read, but in a language that is alien to us.

Compared to the genome, the biochemical system that it directs is much more information-rich. According to some physical models, the space encompassing a single cell can contain more than 10^{30} bits of information (see Table 1 for terminology used), an amount far larger than the information contained in the sequence of the four letters in any genome. Despite this imbalance, the genome can constrain the trajectory of the biochemical system in such a way that both the genome and the biochemical system are copied. In theory, lifeforms can contain far more inheritable information than that contained in the genome, as the genome does not directly define the identities, conformations, rotational orientations, or spatial positions of the molecules of the biochemical system. However, much of the molecular structure of cells arises by a process of self-assembly, where the structures and activities of individual proteins determine the structure of their higher order complexes (reviewed in Cartwright, 2016). In addition, the genome has co-evolved with the biochemical system. For example, replacing leucine with norleucine in all proteins would potentially support life and result in a new lifeform, but this would require that this substitution would also induce leucine tRNA synthetases to incorporate norleucine to leu-tRNAs. In addition, many existing proteins would have to function when all their leucines are replaced by norleucine. This is highly unlikely, and thus, only a vanishingly small subset of all possible biochemical systems can support life using a particular genomic sequence. Given these considerations, it is likely that almost all information that is both mutable and inheritable is encoded directly by the genome.

Much early work on molecular mechanisms in biology used the tools of biochemistry and genetics. In biochemistry, an activity of an enzyme was purified from a mixture of proteins, identifying the specific protein that had the activity of interest. In genetics, in turn, a gene causing a specific phenotype was identified by linkage or association. Both of these classical approaches were set out to identify activities that caused a molecular or organismal phenotype in a highly specific manner. The fact that such

1 Department of Biochemistry, University of Cambridge, Cambridge, UK
2 Genome-Scale Biology Program, University of Helsinki, Helsinki, Finland
3 Division of Functional Genomics and Systems Biology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden.
E-mail: ajt208@cam.ac.uk
DOI 10.15252/embj.201696114 | Published online 18 April 2018

Table 1. Terminology.

Information	Information here refers to the number of bits required to describe the DNA sequence using a four-letter alphabet without a compression algorithm. The abstract string describing the sequence of the genome can be transmitted with less information than this by using many simple compression algorithms. However, as the genome is actually transmitted to daughter cells in an uncompressed form as a molecule that contains each of the bases, for purposes of this commentary, I will not use such formalism. For readers interested in that area, please see, for example, reviews by Fabris (2009), Adami (2012) and Koonin (2016a). Because sequences that arise by mutation without selection can result in phenotypes that are of biomedical importance (e.g., disease), information is also used here without regard to a specific "meaning" or "purpose", simply as a measure of the memory capacity (coding capacity) of a polymer
Compression	Decreasing the amount of memory capacity needed to store a set of data. Lossless compression (e.g., gzip algorithm) allows complete regeneration of original data, whereas lossy compression (e.g., jpeg algorithm) results in loss of some data and introduction of noise. Because human genome has many low complexity sequences, it can be compressed. This means that the genome does not "use" its entire coding capacity. More technically, the Shannon entropy of the genome sequence is lower than the memory capacity of the DNA polymer itself
Phenotype	What is meant by "phenotype" in this review is a feature that can be defined and measured, and whose measurement results in generation of information about the system. For example, height and Km of an enzyme are examples of organismal and molecular phenotypes, respectively
Optimal biochemical system	A system that optimally uses energy and molecular building blocks to reproduce, without production of side products or unnecessary waste
Spandrel	An evolutionary concept that describes an observable phenotype that itself has no effect on fitness, but has emerged and persists as a consequence of another phenotype that is selected
Exaptation	The utilization of a feature for a purpose other than that for which it originally emerged. In the case of a spandrel, the feature originally evolved for no purpose of its own
Purifying selection	Selective removal of deleterious alleles

activities were identified led some scientists to think that biological organisms are efficient, specific, and orderly. However, actually encoding such a biological system would require far more information than what is available in the genome.

In contrast to the classical approaches, genomics and functional genomics have not analyzed specific functional features, but instead determined genome-wide all activities found in cells. These studies have revealed that biological systems are in fact imprecise and wasteful. I will argue below that this is a necessary consequence of the fact that biological organisms are described by a finite amount of genomic information that orchestrates a biochemical system that is far more information-rich, and that this imbalance imposes limits on the capacity of biological systems for precision and efficiency.

A genome cannot specify optimal biochemical parameters

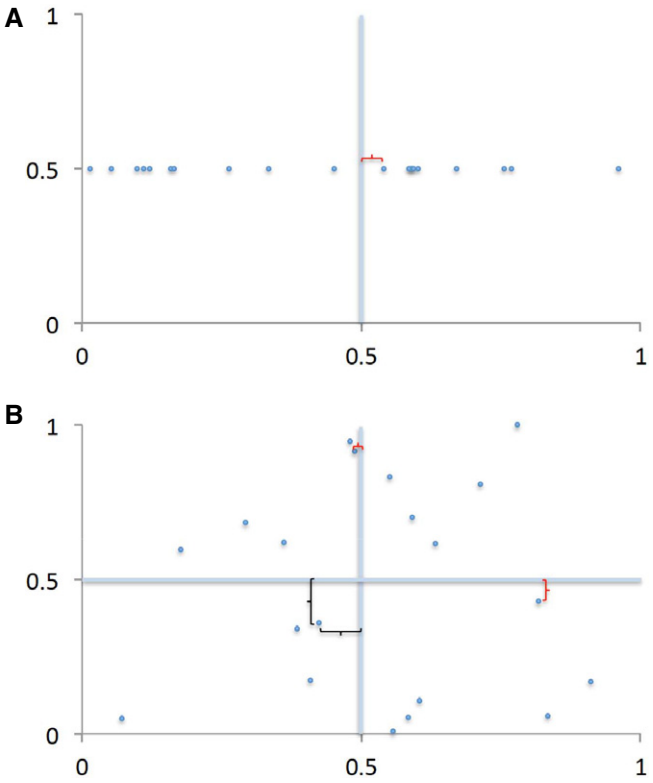
Even relatively small biomedical laboratories can now generate more information about a biological system than what is contained in the genome of the corresponding organism. For example, our laboratory has determined binding specificities of pairs of human transcription factors (TFs) to DNA (Jolma *et al*, 2015). This dataset, when completed, will contain more than 40 million parameters. The value of each parameter is clearly determined by the genome, through the protein sequences of the TFs. If we measure these parameters even at a precision that only specifies whether a particular pair of TFs prefers a particular spacing and orientation, the dataset would contain 40 megabits, an amount of information that is two orders of magnitude larger than that contained by the entire coding sequences of the TF protein constructs themselves (~400 kilobits). In practice, we can measure the parameters at a higher precision, leading to a difference of approximately four orders of magnitude between the amount of information in the measured values and in the part of the genome that determines them. What does this mean?

Firstly, it means that measuring biochemical parameters at very high precision is futile, as the genome cannot specify a value for any given parameter that is arbitrarily close to a theoretical optimal value (Fig 1A). This is clear if one considers that each possible sequence for a particular protein selects one available affinity value for its interaction with another protein. The distribution from where the values are selected cannot be infinitely complex, as it is constrained by physical laws that have a relatively simple mathematical form. Therefore, it is not possible to select a sequence that sets a precisely optimal value for any single parameter.

Secondly, specifying multiple parameter values with the same sequence imposes a further limit on the combined accuracy to which the optimal values can be set, as the same amino acids must determine the values for multiple parameters (Fig 1B). This means that we cannot reasonably infer that the value of a particular experimentally determined parameter is important even if it is conserved between multiple species. That is because any given parameter is most likely correlated with a large number of others, only one of which could be the true cause of the phenotype that is under evolutionary pressure. Thus, most of the values are molecular variants of the spandrel (Gould & Lewontin, 1979), an evolutionary concept that describes an observable phenotype that itself has no effect on fitness, but has emerged and persists as a consequence of another phenotype that is selected. More formally, the spandrel concept is an approximation, as many co-determined phenotypes can together be important for fitness.

Interactions are often spandrels

Considering bimolecular and higher order interactions leads to exponentially increasing numbers of molecular phenotypes that can far exceed the number of base pairs in the genome, leading to a massive increase in the fraction of co-determined phenotypes.



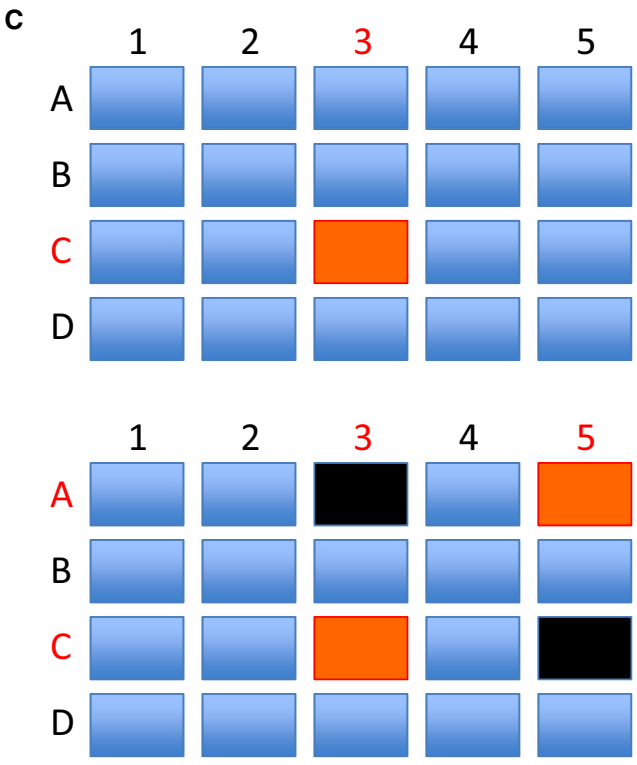


Figure 1. A genome cannot set precisely optimal parameter values or independently determine a larger number of phenotypic states than it has itself.
(A) An attempt to optimize a single phenotypic parameter value to the optimal value of 0.5 using variation in a single amino acid. The 20 different values to be selected from are indicated by the blue dots. Note that the best approximation (red bracket) is not exactly at 0.5. (B) Optimization becomes more difficult when the same amino acid affects two phenotypes. Note that the requirement for optimizing multiple parameters using the same sequence led to less optimal values for both parameters (compare black and red brackets). (C) A system can determine more states than it has itself, but cannot determine them independently of each other. As an example, a gene has four possible start sites (A to C) and five termination sites (1–5). Top: By choosing the start and the termination site, any one of the 20 possible products can be expressed, without expressing any other product. For example, C3 product (orange) can be selected by starting from C and terminating at 3. Bottom: It is not possible to specifically express all combinations of two products. For example, selecting C3 and A5 (orange) leads to expression of A3 and C5 (black, spandrels) as well. It is possible to change the rules, but no single rule system will allow independent selection of all possible 20 consequent states (20 bits), as only nine causative states (nine bits, four in A–C and five in 1–5) exist.

For example, an average protein, with ~500 amino acids (~2 kb of information), cannot independently determine its affinities to all other proteins in the genome, as even describing the ranks of the affinities requires ~500 kb. It is thus clear that each node in a dense protein–protein interaction network cannot be important for the fitness of an organism. It is intuitive that binding to 10 specific proteins requires more information than binding to a single one. However, for scientists not familiar with information theory, it appears counterintuitive that binding to 10 proteins, nine of which are random and one of which is biologically important, requires less information than specifically binding to just the important one.

Specificity is informationally expensive and will not be achieved if it is not required for function (see Fig 1C). Thus, the number of edges in a protein–protein interaction network is likely to exceed the number that is important for fitness, and even apparently conserved interactions could be spandrels. Similar considerations about the informational limits of the systems studied should be extended to all functional genomic studies, such as those analyzing genetic interactions and studies analyzing the sequences of carbohydrate chains covalently linked to proteins.

Consequences for repair and maintenance

The high informational requirements for specifying interactions also have consequences for repair and maintenance of biological systems. Damage or disease often involves pathological interactions between macromolecules, including new genetic regulatory interactions between transcription factors and oncogenes (Sur & Taipale, 2016). As the number of potentially damaging interactions is extremely high, the genome cannot contain instructions to prevent all such interactions, and it is also too small to specify repair pathways for all possible molecular contingencies. Instead of relying heavily on repair, the genome thus must control its biochemical environment by reasserting its orders. This is effected by recycling macromolecules, by specifying that defective cells die by apoptosis, and by continually generating new cells and new whole organisms according to the genomic instruction set.

Information in the regulatory genome

As described above, the affinities of all TF–TF interactions cannot be independently determined by the genome. However, the specific interacting pairs can be recruited to specific positions in the genome by encoding their recognition site at those particular positions, thus "exapting the spandrel in *trans*". Much of conserved genomic information is located in such regions outside of protein-coding genes, and substantial fraction of these regions are involved in protein–DNA interactions.

Functionally, hundreds of thousands of alternatively spliced RNA transcripts with millions of start and end positions, together yielding hundreds of millions of distinct molecular species, have been characterized (see, e.g., Barbosa-Morais *et al*, 2012; Pelechano *et al*, 2013; The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014; Gerstein *et al*, 2014). A scientist familiar with the genetic or biochemical paradigm might think that each of these molecules is important, and has specific roles in particular biological processes. This could be effected by the large swathes of information-rich non-coding DNA that could potentially have evolved to encode very specific instructions on where and when to express each specific version of these transcripts. Recent functional genomic experiments have indeed revealed that a large part of the genome participates in biochemical reactions, with "biochemical activity" assigned to 80% of the human genome (ENCODE Project Consortium, 2012). However, comparative sequence analyses have yielded estimates that only 2–12% of the genome is under "purifying selection" (Ponting & Hardison, 2011; Ward & Kellis, 2012), suggesting that not all of the biochemical activities contribute to organismal fitness. Biochemically, all base pairs contribute directly or indirectly to all organismal phenotypes,[1] albeit often very, very little. Thus, in a formal sense, both types of estimates depend on thresholds that are set for "biochemical activity" and "selection", respectively. A more rigorous approach to understanding the informational limit set by evolution starts with the concept of information, and the realization that selection (survival of the fittest) is based on survival and reproduction of an individual organism, and thus does not directly act on base pairs or genes, but on whole genomes.

Informational limit set by evolution

Not all genomic information in higher organisms has been set by evolutionary selection. This was suggested in the 1960s by the neutral and nearly neutral theories of molecular evolution (Kimura, 1968; Ohta, 1973) and more recently by the elegant synthesis of genomics, evolutionary theory, and population genetics (Lynch, 2006, 2007). These studies suggested that the effect of selection for most variants in relatively small populations is far smaller than the effect of genetic drift. In the information-theoretic sense, this is caused by the fact that a real population contains a discrete number of individuals, and this number limits the precision at which frequencies of alleles can be maintained under selection.

In addition to random changes in frequencies of preexisting variants caused by drift, the amount of information that the genome has acquired about its environment is also limited by the balance of the information contents of selection and mutational load. In higher organisms, this balance is heavily tilted toward mutation. Mutation generates far larger number of variants in each generation than what can be selected for or against. As a thought experiment, consider an attempt to breed a new perfect[2] cow, given two perfect cows of opposite sex whose whole genomes are perfect, and only differ in sequence of their sex chromosomes. Their offspring would have on average ~70 mutations, and selection of a new perfect cow with no mutations would require selecting one out of $>10^{30}$ cows, which would involve breeding them for far longer than the expected lifetime of the universe. This is because the information content of selection itself is much lower than the aggregate information content of all the mutations that occur in a single generation.

As another example, human mutation rate is approximately 75 base pairs per generation (Kong *et al*, 2012), indicating that mutation writes 150 bits of information to the genomic memory at (almost) random positions. Selection generates only around three bits per generation, as in the order of one in eight embryos survive, develop, and reproduce ($-\log_2(1/8) = 3$). The way that the genome learns the features of the environment is through selection. In the information-theoretic (Shannon, 1948) sense, the environment represents the source of the information and the genome its destination, whereas selection represents the channel through which the signal is sent, with its capacity limited by the magnitude of the change in the number of genomic copies due to death and reproduction.

Selection can act on different phenotypes in different lineages, and the three bits of information can be shared via recombination and sexual reproduction, substantially increasing the rate by which

[1] In the sense that their effect can theoretically be measured and are "real" in the physical sense. In practice, most effects are not specific to the measured phenotype, and/or vanishingly small, and are routinely ignored in for example the "near neutral" model of evolution (Ohta, 1973).

[2] A cow cannot be perfect as its phenotype is described by a discrete system (the genome), thus always resulting in a cow that steps either under or over the optimal position in the fitness landscape. What "perfect" means here is "best of all possible cows", with a mutation rate that is similar to that observed in another large mammal, human. Interestingly, being only best of all possible cows has evolutionary benefits over being perfect. Because of low information content of the genome, and the large number of co-determined phenotypes, the phenotype of a real cow cannot very specifically and narrowly fit the optimal phenotype, and thus, unlike the perfect cow, can shift to the next local maxima due to exaptation of the "unintended" associated phenotypes when conditions change.

the system adapts to features of the environment. However, as the mutations occur in all individuals and at random positions, the overall mutational load cannot be effectively cleared by this mechanism. This indicates that in a steady state, 98% of the memory capacity of the genome is filled with information about the specific molecular mutational processes and 2% contains descriptions of other aspects of the microscopic and macroscopic environment. At this ratio, most mutations will miss selected bases, and the ones that hit them can still be removed by selection. However, as individual-level selection is not perfect and often acts on events only partially determined by the genome, the 2% is most likely an overestimate. Thus, most of the information in the genome is solely written by random mutational processes and cannot direct gene expression of millions of specific transcripts independently of each other, in such a way that all have increased the fitness of the organism (see also Lynch, 2012; and Koonin, 2016b). Instead, large biological organisms must function using the small fraction of information that has accumulated into the genomic memory during the hundreds of millions of years of Darwinian evolution. The remaining 98% of information can and will have phenotypic effects, but those phenotypes, by definition, have not been under selection.

Implications to modeling

Biologists generally prefer understanding of biological systems at a conceptual level, described using natural language. The main thrust of the biomedical research enterprise is still focused on identifying novel concepts and fundamental principles at a relatively high level of abstraction. However, searching for more useful concepts is becoming progressively more difficult, as many fundamental principles are already known. Testing which concepts are useful using hypothesis-driven research, which generally tests a small number of simple models and often results in low information feedback, is also becoming exponentially more difficult as more information accumulates. In addition, even the best possible high-level conceptual models will describe biological systems at a level of abstraction that is generally too high for effective prediction of the structure or behavior of the systems. For example, knowing that transcription factors exist and activate gene expression by binding to gene regulatory elements does not allow one to predict expression of genes from their regulatory sequences. This lack of knowledge of the gene regulatory code has severely hampered translation of the exponentially accumulating sequence data toward biotechnological applications and medical benefits. Thus, in abstracting away the genome, the high-level conceptual models that the human mind finds so fascinating fail to accomplish key objectives of biomedical research. Continuation of investment of most resources to projects aimed at identification of such appealing abstractions is likely to yield diminishing returns, ultimately leading to decreased public support for science itself.

In contrast to the relatively straightforward models of biologists, computational modeling of biological systems has developed extremely information-rich models for even relatively simple systems (see also Apostolico, 2001). However, if the whole system is less complex than the model, does this mean that simpler computational models could be used to model important features of biological systems? Ultimately, in modeling biological systems, the analogue of the ideal gas law of physics would be the genome sequence itself. A perfectly efficient model of a biological system would only use as much information as what is contained in the genome, implemented in an algorithm that starts from a set of initial conditions and models the physical world. A straightforward implementation of such a model is presently not computationally tractable. Thus, it will be necessary to build computable approximations. However, the genome may be 98% "junk", but the remaining useful information it contains is encoded with a highly efficient algorithm. In other words, the code used by the genome to determine important phenotypes is not bloated, as each bit is coded the hard way by selection. Thus, identifying computationally tractable approximations of the encoding that would allow similar levels of compression may be very challenging. However, we will never know this without testing different classes of simple models and assessing their ability to predict biological phenotypes. Such experiments could identify a region traversing the middle ground between the overly abstract "mind models" of biologists and fully realistic physical models, allowing development of models whose algorithmic complexity is within a few orders of magnitude from that of the genome that could predict the behavior of biological systems at a practically useful accuracy.

Perspective

The genome can be thought of as the product of a specific type of machine learning algorithm that learns environmental features using selection, which in turn acts on random variation introduced by mutation. From this perspective, it is clear why evolution has resulted in an increase in the length of the genome over time, as organisms with higher memory capacity can fit their environment better. More efficient evolvers that inherit acquired characteristics (Lamarckism) or even predict future environment could accumulate far more information to their genomes and would rapidly take over the biosphere, provided that they would not overfit to minor fluctuations in the environment. Although we have evolved epigenetic mechanisms and cellular systems such as the immune system and the brain that adapt and learn environmental features very effectively, such learned features are not written to the genome and cannot thus be genetically inherited. Therefore, present biological organisms are subject to two informational limits, one defined by the size of their genome and the other by the relationship between mutation rate and selection. The limit set by evolution is generally hit first. Evolution also limits what kind of biological systems can exist, as it requires that all previous generations have been able to reproduce. But even human-designed synthetic organisms would have to obey the limit set by the size of their genomes. The scale of biomedical experimentation allowed by modern technology can easily lead to generation of datasets that consume far larger amount of memory than the genome itself. The results can define much larger number of consequent states than what is available in the genome to cause them (see Fig 1C). Thus, the observed effects must be dependent of each other, and extreme care should be taken in interpretation of any individual data point.

In order to fully understand biological systems, it is necessary to develop models that are detailed and described in a formal, computational language. For this purpose, one needs to adopt an information-based

analysis of evolutionary and functional genomics. A biological system is determined by the information contained in the genome, and estimating the relative amounts of information used to encode distinct molecular phenotypes would ultimately allow development of efficient and plausible models of biological systems. As the genome is severely limited in its information content, each individual measurement should not be considered in isolation, but instead a model that takes into account the informational limit should be built as a framework that guides further experimentation. This would take the form of annotation of the genome itself, with information in each base pair accounted for either by mutation alone, or mutation followed by selection for a specific function or combination of functions (e.g., protein folding, interaction, enzymatic activity, sequence that binds to RNA- or DNA-binding protein). This ultimately would prevent over-interpretation of the phenotypic data, as any new information would have to be considered within the context of the existing knowledge and subject to the informational limits. The completion of such a model would necessitate factory-scale experimentation to systematically and precisely measure orthogonal cellular and molecular phenotypes and to globally analyze macromolecular interactions and determine residues required for different purposes in all proteins. The analysis and interpretation of the data should then respect the limitations imposed by the most profound aspect of the genome, its finite and digital nature, and place the genome at its rightful place at the center of our models of life.

Acknowledgements

I wish to thank Drs. Sampsa Hautaniemi, Teemu Kivioja, Sten Linnarsson, Ben Lehner, and Minna Taipale for critical reading of the manuscript. I also want to thank all the scientists who have taken the time to discuss with me the topics covered here. To avoid an excessively technical tone, I have attempted to use common English words that do not have more formal meaning in any particular subfield covered by the text; the treatment may therefore appear somewhat imprecise to specialists. Some level of imprecision is, however, unavoidable, as the text itself is relatively short and low in information content. Therefore, in the same way that a finite genome cannot describe an optimal organism, this text cannot perfectly describe the underlying subject matter. I also thank the SciLife Laboratory, where I learned that directing complex organizations with decisions that have low information content (e.g., yes or no, funding level) is difficult, leading to the development of some of the ideas described here.

Conflict of interest

The author declares that he has no conflict of interest.

References

Adami C (2012) The use of information theory in evolutionary biology. *Ann N Y Acad Sci* 1256: 49 – 65

Apostolico A (2001) Of maps bigger than the empire. Proceedings of Eighth Symposium on String Processing and Information Retrieval, 2–9

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338: 1587 – 1593

Cartwright JH (2016) Directed self-assembly, genomic assembly complexity and the formation of biological structure, or, what are the genes for nacre? *Philos Trans A Math Phys Eng Sci* 374: 20150449

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57 – 74

Fabris F (2009) Shannon information theory and molecular biology. *J Interdiscip Math* 12: 41 – 87

FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. *Nature* 507: 462 – 470

Gerstein MB, Rozowsky J, Yan KK, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME *et al* (2014) Comparative analysis of the transcriptome across distant species. *Nature* 512: 445 – 448

Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci* 205: 581 – 598

Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J (2015) DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* 527: 384 – 388

Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217: 624 – 626

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U *et al* (2012) Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* 488: 471 – 475

Koonin EV (2016a) The meaning of biological information. *Philos Trans R Soc Lond A* 374: 20150065

Koonin EV (2016b) Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. *BMC Biol* 14: 114

Lynch M (2006) The origins of eukaryotic gene structure. *Mol Biol Evol* 23: 450 – 468

Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA* 104: 8597 – 8604

Lynch M (2012) Evolutionary layering and the limits to cellular perfection. *Proc Natl Acad Sci USA* 109: 18851 – 18856

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246: 96 – 98

Pelechano V, Wei W, Steinmetz LM (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* 497: 127 – 131

Ponting CP, Hardison RC (2011) What fraction of the human genome is functional? *Genome Res* 21: 1769 – 1776

Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 28: 379 – 423

Sur I, Taipale J (2016) The role of enhancers in cancer. *Nat Rev Cancer* 16: 483 – 493

Ward LD, Kellis M (2012) Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 337: 1675 – 1678

