

Concepts

A Non-Adaptationist Perspective on Evolution of Genomic Complexity or the Continued Dethroning of Man

Eugene V. Koonin

National Center for Biotechnology Information; National Library of Medicine; National Institutes of Health; Bethesda, Maryland 20894 USA

Correspondence to: Eugene V. Koonin; National Center for Biotechnology Information; National Library of Medicine; National Institutes of Health; 8600 Rockville Pike; Bldg. 38A; Bethesda, Maryland 20894 USA; Tel.: 301.435.5913; Fax: 301.480.9241; Email: koonin@ncbi.nlm.nih.gov

Received 01/05/04; Accepted 01/06/04

Previously published online as a *Cell Cycle* E-publication:
<http://www.landesbioscience.com/journals/cc/abstract.php?id=745>

KEY WORDS

genomic complexity, entropy, genome evolution, introns, neutral evolution, population size reduction

ACKNOWLEDGMENTS

I appreciate constructive discussions with Michael Lynch and Alex Kondrashov. This article is an opinion piece, rather than a regular review, and I made no attempt, except in the part on gain and loss of genes and introns, to cite all the relevant references.

ABSTRACT

A new, non-adaptationist theory of evolution of genomic complexity was recently proposed by Lynch and Conery.¹ This concept holds that increase in complexity seen in eukaryotic genomes is a 'syndrome' caused by increase in genome entropy, which is inevitably triggered by reduction of population size. Here, I discuss the definitions of genomic entropy and complexity and the evidence supporting the entropic theory of genome complexity evolution, including new observations on concordant gain and loss of genes and introns in eukaryotic genomes. I further consider the far-reaching biological and philosophical implications of this theory.

ENTROPY AND COMPLEXITY OF GENOMES: PROBLEMS AND DEFINITIONS

Complexity is, arguably, one of the most popular and most misused buzzwords in today's science, most especially, biology. No one knows how to define complexity optimally but everyone agrees that it is an extremely important and interesting property of biological systems (e.g.).² It probably would not be much of an exaggeration to say that understanding why or, more precisely, how did it so happen that our planet is inhabited by creatures spanning a staggering range of organizational complexity is one of the prime goals of today's biology. Indeed, at the organism level, multicellular eukaryotes, such as animals and plants, obviously are substantially more complex than unicellular eukaryotes, let alone prokaryotes—suffice it to mention the elaborate differentiation of cell types, tissues, and organs in animals and plants. Is this increase in phenotypic complexity instantiated through a correspondingly more complex genome organization? At this level, the issue becomes somewhat murky. Eukaryotic genome organization certainly displays many features that point to an increase in complexity parallel to that of the organism itself. Comparison of the sequenced eukaryotic and prokaryotic genomes makes it obvious that eukaryotes have more genes, on average, larger proteins, longer and more elaborate regulatory regions, and unique modes of gene expression, such as alternative splicing, which could lead to a dramatic increase in the complexity of the proteome.

However, with this increase in genomic complexity seen in eukaryotes, comes a major puzzle. Along with the (at least potentially) 'useful' complex features, genomes of multicellular eukaryotes have accumulated numerous 'selfish' elements, which have no function in the 'host' organism.³⁻⁵ Questions of paramount importance and interest are why are selfish elements so abundant in the genomes of complex organisms but not in those of simpler ones, and what is the connection (if any) between the propagation of selfish elements and emergence of 'useful' aspects of complexity. A recently published theory suggests an astonishingly simple, general, and plausible solution.¹

Before discussing this theory and its implications, it makes sense to present the contrasting definitions of complexity used in mathematics, information theory, and biology in a somewhat more formal manner. Understandably, it is prohibitively hard to develop a general mathematical definition of 'functional complexity' such that it could be estimated from genomic sequence. The closest attempt I am aware of was made by Christoph Adami who proposed the notion of 'physical complexity'.⁶ Physical complexity is defined as the amount of information on the organism's environment (this could be generalized to include both the external and the internal physical milieu) stored in the genome sequence. In practice, this is approximated by sequence conservation: the greater the conservation the more information about the environment is contained in the sequence. Obviously, this definition of complexity is quite different from and, in a sense, is the opposite of the more traditional mathematical notion of Kolmogorov complexity. Kolmogorov complexity is,

essentially, equivalent to Shannon entropy and can be thought of as the length of the shortest message in which the given sequence can be encoded. When defined in this way, complexity is minimal for a homopolymer (the length of the message is just one letter) and maximal for a random sequence, in which case complexity is simply equal to the sequence length. Clearly, this does not correspond to our intuitive notion of what is meaningfully complex in biology; to emphasize this and for brevity, I will use the term ‘entropy’ for this quantity. Physical complexity (again, I think ‘biological complexity’ makes more sense in this context; from here on, I will use this term or simply ‘complexity’) is linked to entropy through a simple relationship:

$$C = L - \sum_{j=1}^L H(j)$$

where L is the sequence length and H(j) is per site entropy determined from an alignment of the analyzed sequence with its homologs. Entropy is calculated using the classic Shannon formula (given here for a nucleotide sequence):

$$H(j) = -\sum_{i=A,T,G,C} p(j) \log p(j)$$

such that H(j) = 1 for a completely random, nonconserved site, whereas, for a fully conserved site, H(j) = 0. Thus, for a completely conserved sequence, biological complexity reaches the maximum and is equal to entropy, i.e., C = L, whereas, for a sequence that is not evolutionarily conserved at all, C = 0. In other words, such a non-conserved sequence carries no biological information, which is quite compatible with our intuition. Essentially, this can be thought of as the definition of the notorious junk DNA. Biological complexity, as defined above, is the portion of the entropy of the genome, to which biological meaning (function) has been assigned by evolution.

One of the most remarkable features of eukaryotic evolution is that, although biological complexity of the genomes of multicellular eukaryotes is greater than that of prokaryotic genomes, the difference in entropy is much more dramatic. Indeed, in the large genomes of most complex eukaryotes, such as mammals, a substantial majority of the nucleotides appears not to be under any selective pressure, i.e., these sequences have C = 0. Although the actual fraction of eukaryotic DNA that is subject to selection is intensely debated,⁷ the amount of DNA that is indisputable junk is huge. Remarkably, numerous junk sequences have invaded genes themselves: in mammalian genomes, introns, in which << 10% of the nucleotides are subject to selection, comprise ~30% of the genome, whereas only ~1.5% is allotted to protein-coding sequences. With the increase of genome size, the biological complexity to entropy ratio, which we may call ‘effective complexity’ of the genome,

$$C = (L - \sum_{j=1}^L H(j))/L = 1 - \sum_{j=1}^L H(j)/L$$

notably drops. We know too little about noncoding functional elements of eukaryotic genomes to produce confident estimates of E, and such estimates are further complicated by the need to take into account various types of repeats present in the genome. However, back of the envelope calculations suggest values of ~0.8 to 0.9 for prokaryotes (these are ‘wall-to-wall’ genomes that consists largely of protein-coding sequences), but < 0.1 for mammalian genomes.

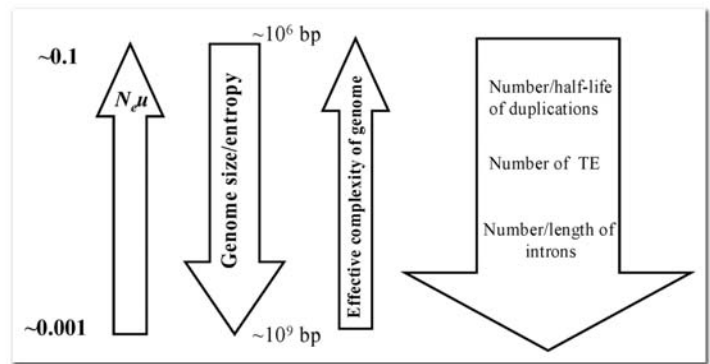


Figure 1. Effective population size and genomic characteristics: positive and negative correlations. Arrows indicate the direction of increase of the respective quantity(ies). The $N_e u$ estimates are from.¹ TE, transposable (mobile) elements; bp, base pairs.

A NON-ADAPTATIONIST, ENTROPIC CONCEPT OF EVOLUTION OF GENOMIC COMPLEXITY

These observations prompt two alternative interpretations of the evolution of biological complexity and entropy of genomes:

1. the increase of biological complexity is an adaptation driven by positive selection, but the mechanisms that lead to its emergence are imperfect and the entropy increase is the price to pay
2. increase of biological complexity of genomes is a by-product of the entropy growth, which in itself is a nonadaptive, neutral process.

In a recent landmark work, Michael Lynch and John Conery provide a remarkably simple theory that implies the latter, nonadaptationist interpretation of the evolution of genome complexity.¹ According to this theory, the common denominator in genome evolution is the characteristic effective size of an evolving population. It is well known from classic evolutionary genetics that the intensity of selection acting on a population is proportional to $N_e u$, where N_e is effective population size and u is the mutation rate. Given that the u value is more or less the same (at least within an order of magnitude) throughout the entire spectrum of life, effective population size becomes critical. Purifying selection is intense in large populations, essentially precluding fixation of significantly deleterious mutations. In contrast, in small populations, even mutations with a substantial deleterious effect can be and often are fixed by random drift. Any increase in genome size initially leads to an increase of genomic entropy genome, but not of biological complexity, and therefore effectively represents a (slightly) deleterious mutation. Such increase occurs via duplication of genome segments large or, more frequently, small, and via amplification of mobile elements. The deleterious effect of these events is inevitable because, even if the new entity does not interfere with any function, the emerging junk DNA becomes part of the genome and needs to be replicated, and that consumes time and energy. Thus, evolutionary theory predicts that population size controls the genome entropy increase, and there might be a population size threshold above which proliferation of any genomic elements becomes negligible. Lynch and Conery examined the connections between the $N_e u$ estimates, the genome size, and the most notable features thought of as hallmarks of eukaryotic genome complexity (more precisely, entropy, in most of these cases), such as number and length of introns, number of mobile elements, and half-life of duplicate genes. The trend is

clear, consistent, and highly statistically significant: a strong negative correlation between $N_e\mu$ and genome size, and equally strong positive correlation between genome size and each of the above quantities (Fig. 1). Particularly impressive is the existence of the predicted sharp threshold (defined in terms of either $N_e\mu$ or genome size) for proliferation of introns and mobile elements.

The above connections certainly do not prove that evolution of biological complexity is driven by the entropy increase triggered by reduced population size, but they appear to be fully compatible with this hypothesis. This means that, simply by the power of the Okkam razor, we must accept this nonadaptationist explanation as the current null hypothesis on the origin of biological complexity. Because, under this theory, the increase in genomic complexity is driven by the entropy increase, I will take the liberty of coining the phrase entropic theory of complexity evolution, at least for the rest of this article. The situation here is very similar, conceptually, to the change in evolutionary thinking brought about by the neutral theory of molecular evolution.⁸ For 100 years between Darwin and Kimura, evolutionary biologists perceived positive selection as the main, if not the only mechanism of fixation of mutations. The neutral theory changed the null hypothesis by showing that most of the fixed mutations were actually neutral (or slightly deleterious) such that cases of positive selection had to be proved.

To put it succinctly once again, the null hypothesis on evolution of biological complexity is that complexity is but a by-product of entropy increase which, in turn, is a 'genomic syndrome' caused by reduced population size. Of course, this is not at all to say that proliferating genomic elements have no functions. The role of duplications in evolution of new functions is well recognized,⁹⁻¹¹ there is accumulating information on the importance of introns in nucleocytoplasmic transport and quality surveillance of mRNAs,¹² and mobile elements have been recruited for regulatory and, possibly, structural functions.^{13,14} However, under the 'entropic' theory of the origin of biological complexity, these functions are by-products of the entropy increase: as long as there are no efficient mechanisms for purging proliferating genomic elements, some of them are eventually adopted for roles that are useful for the host organism. This line of thinking can be extended by conjecturing that an important aspect of selection in species with high-entropy genomes was the ability to recruit proliferated genomic elements for physiological functions. Thus, the survivors have an impressive array of functional adaptations of initially useless (and hence slightly deleterious) sequences. Because species with small populations do not have effective means to purge the proliferating elements (that is, statistically—the life expectancy of an individual entropic element is likely to be low), the motto of eukaryotic evolution seems to be not so much 'use it or lose it' as 'use it or die'.

A CASE IN POINT: CONGRUENT GAIN AND LOSS OF GENES AND INTRONS IN EUKARYOTES

I believe that evidence from various avenues of comparative genomics comes together in an impressive coherence resulting in a picture that is compatible with the entropic concept of evolution of genomic complexity. To emphasize this, I present here in some detail a juxtaposition of two major aspects of eukaryotic genome evolution: gain and loss of genes and introns. Lineage-specific gene loss and emergence of new genes, probably in large part via duplication with subsequent radical divergence, are the most prominent mechanisms of eukaryotic evolution. This has been originally demonstrated by

comparison of fungal genomes^{15,16} and, more recently, by reconstructing the evolutionary scenarios for animals, fungi and plants.^{17,18} Thus, yeast *Saccharomyces cerevisiae* apparently lost ~300 of the genes that were present in its common ancestor with *Schizosaccharomyces pombe* and were retained by the latter; the fruit fly *Drosophila melanogaster* was estimated to have lost ~500 genes since its divergence from the common ancestor with humans. This massive gene loss is counterbalanced by emergence of new genes. The distribution of gene loss and gene emergence events among the branches of the eukaryotic phylogenetic tree is strongly nonuniform. Thus, in the chordate lineage, emergence of new genes appears to be much more common than gene loss, whereas the opposite is true of fungal evolution.¹⁸

In a parallel line of study, we recently examined the conservation of intron positions in a large set of genes that are represented in all sequenced eukaryotic genomes.¹⁹ This analysis, in agreement with other studies,²⁰ showed remarkable conservation of intron positions throughout eukaryotic evolution, but also revealed numerous gains and losses of introns, which were nonuniformly distributed across the phylogenetic tree. Figure 2A shows the currently preferred topology of the phylogenetic tree of the eukaryotic crown group,^{21,22} with gains/losses of genes and gains/losses of introns in conserved genes assigned to each branch according to the previously derived most parsimonious evolutionary scenarios.^{18,19} Figures 2B C show scatterplots of the number of gene gains and losses against the number of intron gains and losses for all tree branches. There are strong, statistically significant, positive correlations between the rates of gene and intron gain (Fig. 1B), and gene and intron loss (Fig. 1C) in eukaryotic lineages. Thus, lineages that "invent" more new genes than they lose old ones also have a tendency to insert new introns (Fig. 1B), and lineages that mostly lose genes also tend to lose many introns from the conserved genes (Fig. 1C). The lineages of *D. melanogaster* and *S. cerevisiae* might be considered exceptions because, in each of these, more genes were gained than lost, but loss of introns dominated over gain (Fig. 1A). Even in these cases, however, gene loss was more pronounced than it was in those lineages that showed the tendency to insert numerous new introns, e.g., the human lineage (Fig. 1A).

It should be emphasized that the evolutionary dynamics of genes and introns analyzed here are, a priori, completely independent because evolution of introns was examined in genes that are conserved throughout the eukaryotic crown group. Thus, the observed correlated gain (loss) of genes and introns are likely to be caused by the same, distinctive features of evolution of the respective lineages. The concordance of gain and loss of genes and introns is fully compatible with the predictions of the entropic theory of complexity evolution. Indeed, according to this theory, lineages with characteristic small population sizes are much more likely to fix propagating 'junk' DNA, such as new introns, and also to retain gene duplications long enough for them to be fixed via the subfunctionalization or neofunctionalization mechanisms²³ than lineages with large populations, which experience intensive purifying selection. Hence the entropic theory predicts that gain and loss of genes and introns should occur, at least roughly, in parallel; the present analysis shows that this is, indeed, the case.

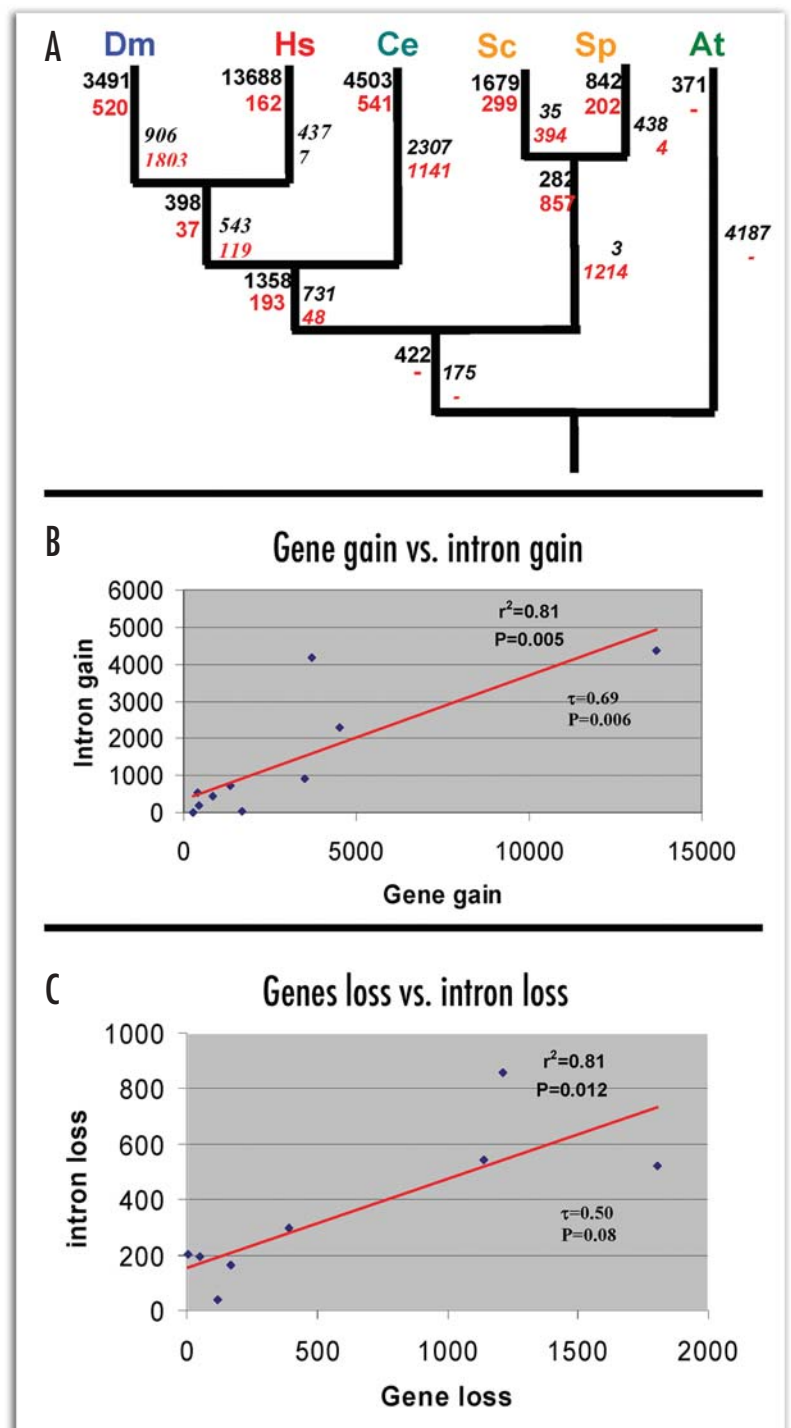
IMPLICATIONS FOR BIOLOGICAL EVOLUTION: CHERCHEZ LA CATASTROPHE

The implications of the entropic concept of complexity evolution are many, but perhaps the most obvious and dramatic one is a

Figure 2. Concordant gain and loss of genes and introns in eukaryotic evolution. (A) Phylogenetic tree of the eukaryotic crown group with the number of gene and intron gains and losses assigned to branches. The intron data are to the right of each branch and rendered in italics. For both genes and introns, the upper number indicates gain and the lower number (in red) indicates loss. For Arabidopsis and for the internal branch leading to the animal-fungal clade, the number of losses could not be determined as indicated by a dash. Species abbreviations: At, *Arabidopsis thaliana*, Ce, *Caenorhabditis elegans*, Dm, *Drosophila melanogaster*, Hs, *Homo sapiens*, Sc, *Saccharomyces cerevisiae*, Sp, *Schizosaccharomyces pombe*. (B) Scatter-plot of gene gains against intron gains. (C) Scatter-plot of gene losses against intron losses. The plotted data are for all branches from (Fig. 1)A (both individual species and internal branches). The linear regression lines are shown. For each plot, Pearson correlation coefficient (r^2), Kendall rank correlation coefficient (τ), and the probability that the correlation of each type was due to chance (P) are indicated. The P-values were calculated using the *t*-test for the Pearson correlation coefficient and using a previously described method²⁸ for the Kendall correlation coefficient.

revival, on a new turn of the scientific spiral, of the catastrophist view of evolution. In the early 19th century, the great paleontologist George Cuvier and some other scientists held the interesting view that the existence of distinct repertoires of fossils in successive geological strata was due to catastrophes that from time to time wipe out the entire biota, after which a new one is created. With the creationist part discrediting it, the catastrophist theory decidedly fell out of vogue, with gradualism of Lyell and Darwin dominating biology and geology as well. The later years have seen a creeping, partial comeback of catastrophism brought about by several striking observations, such as the ascent of mammals immediately after the demise of dinosaurs ~65 mln years ago, apparently triggered by the famed meteorite hit. It is notable that this catastrophe also coincides with the dramatic proliferation of mobile elements in mammalian genomes. The potential pivotal role of environmental catastrophes in major transitions in the history of life is most strikingly emphasized by the latest incarnation of catastrophism in geology, the Snowball Earth theory. Geological evidence indicates that the two most prominent total glaciations of the planet (Snowball Earth epochs) occurred ~2.4 billion years ago and ~700 million years ago.²⁴ These dates are an uncanny fit with two of the principal jumps of organizational complexity in the evolution of eukaryotes, the origin of eukaryotic cell itself and the caembrian explosion, the rapid diversification of the animal phyla. An evolutionary-theoretical counterpart of catastrophism was provided by the punctuated equilibrium theory of Gould and Eldredge, which is based on fossil record analysis and states that brief bursts of dramatic morphological change punctuate long epochs of stasis.²⁵

The entropic theory of complexity evolution makes a qualitatively new step by establishing a direct connection between environmental catastrophes, which result in an abrupt drop of population size of many species, and genome evolution. It is tempting to think that there is a general chain of causation between global or local environmental catastrophes and the increase in organismal complexity, through drastic reduction in population size and the inevitably



ensuing increase in genome entropy, which creates the opportunity for an increase in biological complexity of the genome (Fig. 3). When looking for the ultimate causes of evolutionary transitions, ‘Cherchez la catastrophe!’ could be a powerful guiding principle. Of course, ‘catastrophe’ here needs not to be an apocalyptic event, just any condition that results in a reduction in population size of one or more species (the greater the number of affected species the higher the likelihood of new developments).

This concept has an obvious and extremely general corollary. If major transitions in biological evolution are triggered by unique environmental catastrophes, a hypothetical rerun (or computer simulation) of life’s evolution could not possibly reproduce the history

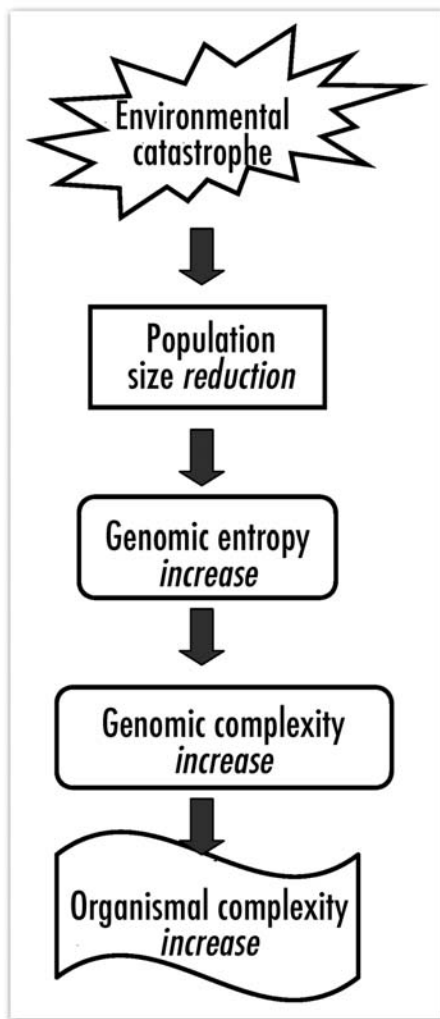


Figure 3. Proposed chain of causation connecting environmental catastrophes with emergence of organismal complexity.

of life as we know it. However, the same concept implies that evolution of some life forms with complex genomes and, accordingly, complex organization is as inevitable as catastrophes resulting in population size reduction. Unpredictability of the course of evolution has been interpreted within the framework of deterministic chaos, particularly by Daniel Dennett.²⁶ The entropic perspective is quite compatible with this view but seems to provide a simpler and more straightforward (even if, perhaps, more mundane) explanation.

PHILOSOPHICAL AND CULTURAL IMPLICATIONS

Beyond the purely scientific implications of the entropic concept of genome complexity evolution, it is hard to refrain from mentioning some philosophical and historical associations. Stephen Jay Gould and others noticed that the history of modern Western science (or, more broadly, culture) is punctuated with successive breakthroughs that led to demotion of man from his traditional (at least in the Judeo-Christian culture) position at the centre of the Universe to a progressively humbler status.²⁷ The two greatest marks in this historical line are, arguably, the Copernican and Darwinian revolutions. Copernicus proved that Earth was not the center of the Universe. In

the domain of physical sciences, this was followed by the developments in astronomy and cosmology, which showed that the Solar System itself is but a tiny part of a garden variety galaxy. Darwin demonstrated that man emerged not by a special act of creation in God's image, but as a regular result of biological evolution, his ancestors being decidedly nondivine creatures. More importantly, Darwin's theory once and for all buried the 'ennobling' idea of innate progress, which was the guiding principle of earlier evolutionists, such as Geoffroy Saint-Hilaire and Lamarck. Darwin argued that natural selection for advantageous traits, which emerge by chance as a result of random mutations, was the principal evolutionary force (with remarkable prescience, Darwin did emphasize the importance of neutral changes as evolutionary markers, but it took over a century for this idea to take off).

Darwin's concept of natural selection at least allowed for some notion of progress, albeit driven by random factors. The developments in evolutionary biology of the 20th century, although not comparable to Darwinism in their cultural shock value, continued the trend of stripping man of the remains of his unique status and evolution of 'purpose' and 'sense'. The neutral theory, which was developed primarily by Kimura and Ohta in the 1970s, showed that the majority of mutations that are fixed during evolution are actually (nearly) neutral. Consequently, the main form of selection is not Darwinian positive, diversifying selection, but rather purifying selection, which acts by eliminating mutations with substantial deleterious effect. The contribution of positive selection, which is, traditionally, linked to the origin of 'progressive' innovations and increase of complexity during evolution, is quantitatively small. The selfish (junk) DNA concept delivered yet another blow. Although the debate on the exact amount of functional sequences in our genomes is far from being settled, it is quite clear that the majority of our DNA is there just because it is capable of propagation and the species has no adequate way to purge it. As a result, genomes of complex animals have a much lower 'effective complexity' than simpler genomes (so much for 'progress'). The entropic concept of complexity evolution makes another important step by arguing that not only is our genome a haphazard jumble of a few functional and many more useless pieces of DNA, but its complexity itself—and hence the ensuing organismal complexity—evolved not as adaptations but as a 'pathological' reaction to reduced population size. What is more, if accidental reductions in population size comprise 'the' (or even 'a') major driving force of evolution, there is no doubt that any particular complex life form (*Homo sapiens* obviously included) emerged by chance.

The concept of complexity evolution discussed here has such general implications that it might be tempting to brand it 'natural philosophy'. I think there is some truth to this view. Although the theory certainly yields falsifiable predictions (the 'Cherchez la catastrophe!' principle could be one of the strongest of these), the implications are so rich that it is hardly possible to falsify the entire concept. However, the greater significance of concepts like this is that they change the entire way we design and interpret our analyses aimed at understanding genome function and evolution. In that respect, the entropic concept of evolution of biological complexity is comparable to the neutral theory, the selfish DNA idea and, perhaps, in principle, if not in general importance, to Darwin's theory of evolution by natural selection, that greatest feat of natural philosophy in modern times.

References

1. Lynch M, Conery JS. The origins of genome complexity. *Science* 2003; 302:1401-4.
2. Gell-Mann M. *The Regular and the Random* Basic Books, New York: 2004.
3. Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 1980; 284:601-3.
4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* 2001; 409:860-921.
5. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002; 420:520-62.
6. Adami C. What is complexity? *Bioessays* 2002; 24:1085-94.
7. Koonin EV, Rogozin IB. Getting positive about selection. *Genome Biol* 2003; 4:331.
8. Kimura M. *The Neutral Theory of Molecular Evolution* Cambridge: Cambridge University Press, 1983.
9. Ohno S. *Evolution by gene duplication* Springer-Verlag, Berlin-Heidelberg-New York: 1970.
10. Lynch M, Force A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 2000; 154:459-73.
11. Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. Selection in the evolution of gene duplications. *Genome Biol* 2002; 3:Research0008.
12. Maniatis T, Reed R. An extensive network of coupling among gene expression machines. *Nature* 2002; 416:499-506.
13. Nekrutenko A, Li WH. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* 2001; 17:619-21.
14. Jordan IK, Rogozin IB, Glazko GV, Koonin EV. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* 2003; 19:68-72.
15. Aravind L, Watanabe H, Lipman DJ, Koonin EV. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc Natl Acad Sci USA* 2000; 97:11319-11324.
16. Braun EL, Halpern AL, Nelson MA, Natvig DO. Large-scale comparison of fungal sequence information: Mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae*. *Genome Res* 2000; 10:416-30.
17. Krylov DM, Wolf YI, Rogozin IB, Koonin EV. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* 2003; 13:2229-35.
18. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* 2003; in press.
19. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* 2003; 13:1512-7.
20. Fedorov A, Merican AF, Gilbert W. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc Natl Acad Sci USA* 2002; 99:16128-33.
21. Hedges SB. The origin and evolution of model organisms. *Nat Rev Genet* 2002; 3:838-49.
22. Wolf YI, Rogozin IB, Koonin EV. Coelomata and not Ecdysozoa: Evidence from genome-wide phylogenetic analysis. *Genome Res* 2004; 14:29-36.
23. Walsh B. Population-genetic models of the fates of duplicate genes. *Genetica* 2003; 118:279-94.
24. Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP. A neoproterozoic snowball earth. *Science* 1998; 281:1342-6.
25. Gould SJ, Eldredge N. Punctuated equilibrium comes of age. *Nature* 1993; 366:223-7.
26. Dennett D. *Darwin's Dangerous Idea*. New York: Simon & Schuster, 1995.
27. Gould SJ. *Full House: The Spread of Excellence from Plato to Darwin* New York: Three Rivers Press, 1997.
28. Babenko VN, Rogozin IB. Use of a rank correlation coefficient for comparing mutational spectra. *Biofizika* 1999; 44:632-8.