PERSPECTIVES: MOLECULAR BIOLOGY AND EVOLUTION

Can Genes Explain Biological Complexity?

Eörs Szathmáry, Ferenc Jordán, Csaba Pál

Although natural selection does not guarantee that organisms will increase in complexity as they evolve, it is apparent that the complexity of certain lineages, such as our own, has increased during evolution. Although we have an intuitive appreciation of biological complexity—often thinking in terms of morphological or behavioral complexity, or the variety of cell types in an organism—the term itself is notoriously hard to define. One could resort to algorithmic complexity, where the number of steps in the shortest possible algorithm that solves a given task has proven to be a convenient measure (*1*). In this case, complexity could be defined as the number of steps in the developmental program out of which the embryo is "computed." The snag here is that evolution is not an engineer but a tinkerer, so that there is no reason to expect that, for example, elephants have developed according to a minimalist program (*2*).

Is the number of genes in an organism's genome an appropriate measure of biological complexity? It has been assumed that eukaryotes have more genes than bacteria, animals have more genes than plants, and vertebrates have more genes than invertebrates (*2*, *3*)—which nicely fits with the traditional notion of a *scala naturae*. The recent flurry of completed genome sequences, including our own, suggests that this is not necessarily the case (*4*–*6*). Rather surprisingly, it turns out that the worm *Caenorhabditis elegans* has 18,424 genes in its genome, the fruit fly *Drosophila melanogaster* 13,601, the plant *Arabidopsis* about 25,498, and humans about 35,000. This suggests that there must be other, more sensible genomic measures of complexity than the mere number of genes.

Transcription factors are DNA binding proteins that switch target genes on and off. For all transcription factor families, their members increase in number in the order yeast, nematode, fruit fly, human (*7*). The diversity of cell types in these organisms also increases in this order (*5*). This makes sense, given that maintaining the differentiated state of increasingly diverse cell types requires the presence of more and more molecular switches (*6*). In commenting on the human genome sequence, Claverie has suggested that we define biological complexity in terms of the number of transcriptome states (a transcriptome being the complete set of RNA transcripts) that the genome of an organism can achieve (*6*). Following this line of thought, how, then, can one obtain a measure of true biological complexity?

We propose that biological complexity might be better explained by considering networks of transcription factors and the genes they regulate, rather than by simply counting the number of genes or the number of interactions among genes. One could borrow indices from other fields that have an older tradition of quantifying networks. For instance, when trying to obtain a measure of ecosystem complexity, ecologists consider not only the number of species but also the types and numbers of interactions among them. For example, the complexity of interactions within a food web can be defined by the connectivity (C): $C = 2\,L/[N(N-1)]$, where the number of actual trophic links (L) is divided by the number of all possible links, with N as the number of species. It would be intriguing to know whether gene-regulation networks in bacteria or eukaryotic cells can also be defined in terms of their connectivity (see the table). A global analysis of transcriptional regulation in the bacterium *Escherichia coli* reveals that on average each transcription factor regulates three genes, and that each gene is under the control of two transcription factors (*8*). Certainly the connectivity of gene-regulation networks in eukaryotes is likely to be greater than that in bacteria, but for now we lack a way to measure the magnitude of this difference.

There are other indices derived from analyses of food web complexity that might be useful for analyzing the connectivity of gene-regulation networks (see the table). For instance, the clustering coefficient could be used to define relatively autonomous groups of developmental genes. The number of these groups (developmental modules) could then in turn provide a measure of developmental complexity.

When considering gene-regulation networks, we think in terms of the transcriptome. But even greater complexity is conferred on organisms by the proteome (that is, all possible proteins that an organism can make). Alternative splicing and posttranscriptional modification of RNA transcripts

GENETIC NETWORKS AND BIOCOMPLEXITY

Index	Scale	Relevance
Number of nodes, N	Global	Number of relevant genes in a genetic network
Number of links, L	Global	Number of gene interactions
Connectivity, $C = 2\,L/[N(N-1)]$	Global	Realized fraction of possible gene interactions
In-degree, D_{in}	Local	Number of genes affecting a particular gene
Out-degree, D_{out}	Local	Number of genes affected by a particular gene
Degree, D	Local	The number of genes directly interacting with a particular gene
Average degree, D_{av}	Global	Average number of gene interactions per gene
Heterogeneity (the standard deviation of degrees)	Global	Evenness of link distribution among genes
Clustering coefficient, (the average connectivity of subnetworks containing each nodes's neighbors), CC	Global	Appearance of tightly connected regulatory subnetworks
Average distance, $D_{av} = [\Sigma d_{ij}]/[N(N-1)]$	Global	Number of communication steps between two randomly chosen genes
Arc connectivity	Global	Minimal number of gene interactions whose deletion results in a disconnected network
Node connectivity	Global	Minimal number of genes whose deletion results in a disconnected network

Indices describing interactions within networks. Such indices include those used by ecologists to determine complex interactions within food web networks. These indices can be applied to the measurement of interactions within and between gene-regulation networks (*13*, *14*).

The authors are in the Collegium Budapest (Institute for Advanced Study), 2 Szentháromság u., H-1014 Budapest, Hungary. E-mail: szathmary@colbud.hu

(RNA editing) can generate many more proteins than the number encoded by genes (*9*). In *Drosophila*, alternative splicing and RNA editing theoretically could generate 1,032,192 mRNA transcripts (each encoding a slightly different protein) from the single *para* gene, which encodes a sodium channel. In yeast, only three genes are known to be alternatively spliced whereas in the human, at least 35% of the gene transcripts undergo alternative splicing. Unfortunately, little is known about the proteins that regulate alternative splicing, although splicing is known to be location- and time-specific (*9*). This suggests that the protein complex carrying out the splicing (the spliceosome) may itself be under strict regulation, perhaps through its interactions with other regulatory proteins.

How does the genomic complexity of plants compare with that of animals? Plants have a surprisingly large number of transcription factors—more than 1500 genes (5% of the genome) encode transcription factors, and half of these are plant-specific (*10*). For comparison, the worm genome has 500 transcription factor genes, the fly genome about 700, and the human genome more than 2000 (*7*). The wide variety of plant transcription factors could be explained by a unique feature of plants: their complex secondary metabolism. As many as 25% of all plant genes are associated with a unique array of secondary metabolites not found in animals (the total number of plant secondary metabolites is close to 50,000, although each plant species produces only a fraction of these). The expression of genes associated with secondary metabolism is both tissue- and time-specific (*11*), which makes the large number of transcription factors comprehensible. Given their multitude of transcription factors, should plants be considered more complex than vertebrates? Obviously, the answer is no, but the reason why requires a closer look at the complexity of vertebrate organ systems.

With a limited number of genes, vertebrates manage to code for two highly complex subsystems that are specialized for information accumulation, storage, and retrieval: namely, the immune system and the nervous system. Both systems operate on a generative basis, that is, they can store huge amounts of information based on a fixed set of rules. These rules reside in variation-generating mechanisms (such as the reshuffling of immunoglobulin genes) and internal selective filters (*12*). In the case of the vertebrate immune system, reshuffling of immunoglobulin genes produces an enormous variety of antibodies. An internal selective filter then recognizes cells producing antibodies against self antigens, weeds them out, and destroys them. Although less well characterized, the vertebrate nervous system contains similar Darwinian elements. During development, a large surplus of nerve cells and their myriad connections are produced, from which only those that best innervate a given territory are retained (*12*). The immune and nervous systems might yield clues as to how an extremely complex and highly connected system could develop from a limited number of genetic instructions. Whereas vertebrates have delegated a large part of their complexity to their immune and nervous systems, plants seem to compensate for their lack of generative systems by depending on gene regulation and synthesis of new secondary metabolites to generate diversity.

So, we need to distinguish between two forms of genomic complexity: one measured by the number of genes and the other by the connectivity of gene-regulation networks. The complexity of organisms (in terms of morphology and behavior) correlates better with the second definition. Delegated complexity, achieved by genetically encoded information-processing systems such as the nervous and immune systems of vertebrates, adds another dimension to biological complexity. With the availability of more and more completed genome sequences, bioinformatics is sure to yield additional measures of complexity. We will then be able to devise new ways to quantify these measures of biocomplexity.

References

1. H. Atlan, M. Koppel, *Bull. Math. Biol.* **52**, 335 (1990).
2. J. Maynard Smith, E. Szathmáry, *The Major Transitions in Evolution* (Oxford University Press, Oxford, 1995).
3. P. Bird, *Trends Genet.* **11**, 94 (1995).
4. P. Bork, R. Copley, *Nature* **409**, 818 (2001).
5. S. B. Carroll, *Nature* **409**, 1102 (2001).
6. J.-M. Claverie, *Science* **291**, 1255 (2001).
7. R. Tupler *et al.*, *Nature* **409**, 832 (2001).
8. D. Thieffry *et al.*, *BioEssays* **20**, 433 (1998).
9. B. R. Graveley, *Trends Genet.* **17**, 100 (2001).
10. J. L. Riechmann *et al.*, *Science* **290**, 2105 (2000).
11. E. Pichersky, D. R. Gang, *Trends Plant Sci.* **5**, 439 (2000).
12. J. Gerhart, M. Kirschner, *Cells, Embryos and Evolution* (Blackwell, Oxford, 1997).
13. F. Harary, *Graph Theory* (Addison Wesley, Cambridge, MA, 1969).
14. M. Higashi, T. P. Burns, Eds., *Theoretical Studies of Ecosystems—the Network Perspective* (Cambridge Univ. Press, Cambridge, 1991).

PERSPECTIVES: EPIDEMIOLOGY

How Viruses Spread Among Computers and People

Alun L. Lloyd and Robert M. May

The Internet and the world wide web (WWW) play an ever greater part in our lives. Only relatively recently, however, have researchers begun to study how the patterns of connectivity in these networks affect the spread of computer viruses within them (*1*, *2*) and their ability to handle perturbation or attack (*3*). Many models for communication can be formulated in terms of networks, in which nodes represent individuals (such as computers, web pages, people, or species) and edges represent possible contacts between individuals (network links, hyperlinks, social or sexual contact, and species interactions). The study of communication networks therefore has interesting parallels both with conventional epidemiology (*4*, *5*) and with the ability of ecosystems to handle disturbances.

In a recent paper in *Physical Review Letters*, Pastor-Satorras and Vespignani (*6*) explore a dynamical model for the spread of viruses in networks of the kind found in the Internet and WWW (*7*, *8*). In striking contrast with the usual models for the spread of infection in human and other populations, they find no threshold for epidemic spread: Within the observed topology of the internet and WWW, viruses can spread even when infection probabilities are vanishingly small. They also find that, in its early phase, the epidemic spreads relatively slowly and nonexponentially, again in contrast with the initial exponential behavior in conventional epidemics. These are notable findings, and the authors suggest they may be relevant to other types of social networks.

The importance of spatial structure for disease transmission has long been recognized (*9*). Locally structured networks often have many intermediates in paths between any given pair of individuals. They can also exhibit clique behavior, with pairs of connected individuals sharing many common neighbors, reducing the opportunities for secondary infection events. As a result, diseases may spread more slowly when contact is mainly local, compared with well-mixed situations. Conversely, earlier studies showed that even infrequent long-distance infection events can enhance disease spread substantially (*9*). This foreshadowed some aspects of recent work on

A. L. Lloyd is in the Program in Theoretical Biology, Institute for Advanced Study, Einstein Drive, Princeton, NJ 08540, USA. E-mail: alun@alunlloyd.com R. M. May is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. E-mail: robert.may@zoo.ox.ac.uk