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): \( C = 2L/[N(N-1)] \), 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

### Indices describing interactions within networks

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

The authors are in the Collegium Budapest (Institute for Advanced Study), 2 Szentháromság u., H-1014 Budapest, Hungary. E-mail: szathmary@colbud.hu
nervous system contains similar Darwinian el-
though less well characterized the vertebrate
mammals, weeds them out, and destroys them. Al-
men (12). In the case of the vertebrate immune
munoglobulin genes) and internal selective fil-
thers comprehensible. Given their multitude
of transcription factors, should plants be
so we need to distinguish between two forms of genomic complexity: one measured
the number of genes and the other by the
complexity, achieved by genetically encoded in-
teractions with other regulatory proteins.
How does the genomic complexity of
plants compare with that of animals? Plants
have a surprisingly large number of tran-
scription 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 20000 (7). The wide variety of
plant transcription factors could be ex-
plained 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, al-
though 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 fac-
tors 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, verte-
brates manage to code for two highly complex
subsystems that are specialized for informa-
tion accumulation, storage, and retrieval:
hotel, 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 anti-
gens, weeds them out, and destroys them. Al-
though less well characterized, the vertebrate
nervous system contains similar Darwinian el-
ments. During development, a large surplus
of nerve cells and their myriad connections
are produced, from which only those that best
nerve 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. Where-
as 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
genome regulation and synthesis of new sec-
ondary metabolites to generate diversity.
So, we need to know that two forms of genomic complexity: one measured
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 com-
plexity, achieved by genetically encoded in-
formation-processing systems such as the
nervous and immune systems of vertebrates,
adds another dimension to biological com-
plexity. With the availability of more and
more completed genome sequences, bioin-
formatics is sure to yield additional measures
of complexity. We will then be able to devise
new ways to quantify these measures of bio-
complexity.

References

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 repre-
sent 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)
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 nonexponen-
tially, 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 recog-
nized (9). Locally structured networks often have many intermediates in paths be-
tween any given pair of individuals. They can also exhibit clique behavior, with pairs of
connected individuals sharing many common neighbors, reducing the opportu-
nities for secondary infection events. As a result, diseases may spread more slowly
when contact is mainly local, compared with well-mixed situations. Conversely,
early studies showed that even infrequent long-distance infection events can enhance
disease spread substantially (9). This fore-
shadowed some aspects of recent work on

PERSPECTIVES: EPIDEMIOLOGY

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