(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.

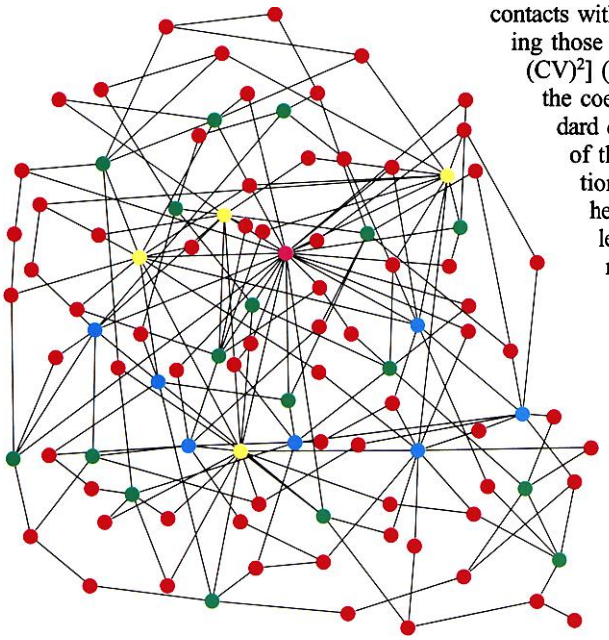
In a recent paper in *Physical Review Letters*, Pastor-Satorras and Vespignani (*6*) explore a dynamical model for the spread of viruses in networks of the kind found in the Internet and WWW (*7*, *8*). In striking contrast with the usual models for the spread of infection in human and other populations, 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

"small world" networks (*10*) and on the recent spread of foot and mouth disease in the UK (*11*).

In contrast to such results, which derive from the spatial structure of networks, Pastor-Satorras and Vespignani's results largely derive from the scale-free character of the internet and WWW (*6*). Scale-free networks (see the figure) can arise when a network grows through new nodes being linked preferentially to the most highly connected existing nodes (*12*). The probability for a node to be connected to k other nodes obeys a power law distribution, $P(k) \sim k^{-\gamma}$. In the case of the Internet and WWW, the observed exponent γ lies between 2 and 3 (*7*, *8*).



No matter of scale. Example of a scale-free network, consisting of 100 nodes, generated with the algorithm of Barabási and Albert (*12*). In order of increasing connectivity, the nodes are colored red, green, blue, and yellow, with the most highly connected nodes colored magenta. Note the small number of highly connected nodes; the majority of nodes have few connections.

Pastor-Satorras and Vespignani simulate the spread of computer viruses with a "susceptible-infected-susceptible" (SIS) model, in which susceptible individuals acquire infection at a rate β upon contact with an infected node and subsequently recover from the infected to the susceptible state after an average time D. In their scale-free network, γ equals 3 (*12*) and the least connected nodes have m connections. The average connectivity, $\langle k \rangle$, is then $2m$. The authors show that the results obtained with this model agree with observed patterns of viral spread and persistence. The system eventually settles to a steady state, in which the fraction of infected nodes is $\gamma \approx 2 \exp(-2/\rho_0)$, where $\rho_0 = \beta D \langle k \rangle$. Epidemiologists would call ρ_0 the "basic reproductive number" for the disease—the average number of infections produced by an infected individual in a wholly susceptible population—assuming a homogeneous network (that is, all nodes are assumed to interact with the same number of other nodes, namely the average, $2m$).

However, spurred largely by the need to understand the spread of human immunodeficiency virus (HIV) within complex networks of sexual partnerships, traditional epidemiology has advanced well beyond homogeneous models. The basic reproductive number, R_0, for HIV and other infections spread by binary contacts within complex networks, including those studied in (*6*), is $R_0 = \rho_0[1 + (CV)^2]$ (*5*, *13*, *14*), where CV denotes the coefficient of variation (the standard deviation divided by the mean) of the node-connectivity distribution. This expression shows that heterogeneity within the network leads to an increase in the basic reproductive number. The reason for the absence of a threshold for the spread of infection in Pastor-Satorras and Vespignani's study is now clear: Their scale-free distribution has infinite variance, and hence R_0 always exceeds unity, no matter how small the homogeneously approximated quantity ρ_0 may be.

The nonexponential nature of the initial spread of infection has also been noted in heterogeneous epidemiological models for HIV (*13*). The initial exponential epidemic phase is rapidly curtailed because the highly active classes quickly saturate with infection, giving way to a more gradual increase, with new infections largely coming from the slower dissemination of infection to less active classes.

In SIS models, the fraction infected at any one time comes almost entirely from continual reinfection of the most highly connected nodes. In reality, these are exactly the sophisticated nodes most likely to avoid this fate. Moreover, for many computer viruses, infected nodes are likely to recover to an immune, rather than a susceptible, state (by using antiviral software or simply losing susceptibility to "I LOVE YOU" enticements). In this case, the somewhat more complicated class of "susceptible-infected-recovered" (SIR) epidemiological models is more appropriate.

In SIS situations, we can observe endemic levels of infection in a closed population, whereas in SIR models, the epidemic waxes and then wanes as the progressing epidemic reduces the number of susceptible nodes. Again, analytic and simulation-based results on the spread of sexually transmitted diseases within heterogeneously connected networks are informative here. For instance, Anderson and May (*5*, *13*) have derived formulas for the fraction of the population, I, ever infected in an SIR epidemic. Interestingly, for Pastor-Satorras and Vespignani's scale-free distribution, this proportion is of much the same form as the asymptotic fraction infected in the SIS model: $I \approx C \exp(-2/\rho_0)$ (a detailed calculation shows that the constant $C \approx 3.05$). Note that in those circumstances where ρ_0 is small, so that R_0 exceeds unity by virtue of the infinite variance in the contact distribution, the fraction infected (both in the steady state for SIS and in total as the epidemic sweeps through for SIR) will be very small.

At first sight, it might seem as if the extreme heterogeneity exhibited by the scale-free networks of Pastor-Satorras and Vespignani makes them poor models for human interactions. Complicated networks of social interactions cannot be treated as if they were homogeneous (*5*, *14*), but heterogeneity is often low in networks describing friendships between individuals (*15*), which might be appropriate models for diseases passed by casual social contact (or computer viruses that use e-mail address lists found on infected machines). Pastor-Satorras and Vespignani's results may be less appropriate for diseases passed by social contact.

On the other hand, sexual partnership networks are often extremely heterogeneous because a few individuals (such as prostitutes) have very high numbers of partners. Pastor-Satorras and Vespignani's results may be of relevance in this context. The study highlights the potential importance of studies on communication and other networks, especially those with scale-free and small world properties, for those seeking to manage epidemics within human and other animal populations.

References

1. F. B. Cohen, *A Short Course on Computer Viruses* (Wiley, New York, 1994).
2. J. O. Kephart, G. B. Sorkin, D. M. Chess, S. R. White, *Sci. Am.* **277** (no. 5), 88 (1997).
3. R. Albert, H. Jeong, A.-L. Barabási, *Nature* **406**, 378 (2000).
4. N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases* (Griffin, London, 1975).
5. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).
6. R. Pastor-Satorras, A. Vespignani, *Phys. Rev. Lett.* **86**, 3200 (2001).
7. M. Faloutsos, P. Faloutsos, C. Faloutsos, *Comput. Commun. Rev.* **29**, 251 (1999).
8. R. Albert, H. Jeong, A.-L. Barabási, *Nature* **401**, 130 (1999).
9. D. Mollison, *J. R. Stat. Soc. B* **39**, 283 (1977).
10. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).
11. N. M. Ferguson, C. A. Donnelly, R. M. Anderson, *Science* **292**, 1155 (2001).
12. A.-L. Barabási, R. Albert, *Science* **286**, 509 (1999).
13. R. M. May, R. M. Anderson, *Philos. Trans. R. Soc. London B* **321**, 565 (1988).
14. R. M. May, S. Gupta, A. R. McLean, *Philos. Trans. R. Soc. London B*, in press.
15. L. A. N. Amaral, A. Scala, M. Barthelemy, H. E. Stanley, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11149 (2000).