11. G. Salton, Ed., *The SMART Retrieval System: Experiments in Automatic Document Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1971).
12. M. E. Williams, *Science* **228**, 445 (1985).
13. F. W. Lancaster and E. G. Fayen, *Information Retrieval On-Line* (Melville, Los Angeles, 1973).
14. R. K. Summit, in *Proceedings of the 22nd National Conference of the Association on Computing Machinery* (Thompson, 1967), pp. 51–56.
15. C. T. Meadow, *Database* (October 1988), p. 23.
16. D. B. McCarn and J. Leiter, *Science* **181**, 318 (1973).
17. G. Salton and M. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, 1983).
18. R. S. Marcus and J. F. Reintjes, *IEEE Trans. Syst. Man Cybern.* **12**, 116 (March–April 1982).
19. M. E. Williams, *J. Am. Soc. Inf. Sci.* **37**, 204 (1986).
20. B. R. Schatz, in *Proceedings of IEEE Globecom †87* (IEEE, New York, November 1987), pp. 1181–1186.
21. For example, as a member of the Internet Research Task Force, I was one of the few members of the generation after the pioneers invited to speak at the 20th Anniversary Symposium for the ARPANET at the University of California at Los Angeles in 1989. My talk, "Telesophy: Towards World-Wide Information Spaces," although full of technical details and projections, seemed grand and futuristic at that point (August 1989).
22. B. R. Schatz, "Telesophy," *Bellcore TM-ARH-002487* (August 1984).
23. ———, in *Proceedings of the 5th IEEE International Conference on Data Engineering* (IEEE, New York, 1989), pp. 188–197.
24. D. C. Engelbart and W. K. English, in *Proceedings of the Fall Joint Computer Conference* (AFIPS Press, New York, 1968), vol. 33, part 1, pp. 395–410.
25. C. F. Herot, *Assoc. Comput. Mach. Trans. Database Syst.* **5**, 493 (1980).
26. An inspiration for knowledge regions was T. Nelson, who designed a grand system called Xanadu to handle all the world's knowledge as a single hyperliterature across multiple collections. His unimplemented treatise, *Literary Machines* (1981), contained many suggestions for building new documents by annotating and linking parts of old.
27. A. Goldberg and D. Robson, *Smalltalk-80: The Language and Its Implementation* (Addison-Wesley, Reading, MA, 1983).
28. B. Kahle *et al.*, *Electron. Networking* **2**, 59 (spring 1992).
29. B. Kahle, personal communication. Kahle developed the WAIS software at Thinking Machines with funding from Apple Computer and later started WAIS Inc., which was purchased by America Online.
30. B. R. Schatz and J. B. Hardin, *Science* **265**, 895 (1994).
31. The two predominant Web browsers are derived from Mosaic: Netscape Navigator was built by the original developers after they left NCSA, and Microsoft's Internet Explorer has at its core a licensed version of Enhanced Mosaic which is produced by Spyglass as the official commercial distributor of NCSA Mosaic. Historically, Telesophy played a role in Mosaic as well, because I have been the scientific advisor for information systems at NCSA since 1989, and Mosaic was one of several attempts at NCSA to reproduce the functionality of Telesophy for the general scientific community.
32. Lycos is a spin-off company from digital library projects at Carnegie-Mellon University. See http://www.lycos.com/
33. Alta Vista was a project, now a service, from Digital Equipment Corporation's Research Laboratories. See http://altavista.digital.com/
34. B. Schatz and H. Chen, *IEEE Comput.* **29**, 22 (May 1996).
35. The May 1996 special issue of *IEEE Computer* contains overview articles from all six DLI projects. See http://www.computer.org/pubs/computer/dli/
36. B. Schatz *et al.*, *Computer* **29**, 28 (May 1996).
37. E. van Herwinjnen, *Practical SGML* (Kluwer, Boston, 1994).
38. F. W. Lancaster, *Vocabulary Control for Information Retrieval* (Information Resources Press, Arlington, VA, 1986).
39. C. Lynch and H. Molina-Garcia, Eds., "Interoperability, Scaling, and the Digital Libraries Research Agenda," 22 August 1995. Available at http://www.hppc.gov/reports/report-nco/reports/iita-dlw/main.html. The Information Infrastructure Technology and Applications (IITA) group is the highest level technical committee of the Federal NII Program.
40. S. Nadis, *Science* **272**, 1419 (1996).
41. H. Chen *et al.*, *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 771 (1996).
42. H. Chen, J. Martinez, T. Ng, B. Schatz, *J. Am. Soc. Inf. Sci.* **48**, 17 (1997).
43. P. B. Kantor, *Annu. Rev. Inf. Sci. Technol.* **29**, 53 (1994).
44. R. T. Niehoff, *J. Am. Soc. Inf. Sci.* **27**, 3 (1976).
45. The vocabulary switching computation used bibliographic abstracts from Compendex (engineering and science) and Inspec (electrical engineering and computer science). Compendex has 40 broad subject classes (for example, computer science) and 600 class codes total. Inspec is narrower and deeper than Compendex, and the computation included about 150 classes at its highest level, the same as the lowest level of Compendex. Because Inspec has roughly 2500 classes all together, the collection spanned in total about (600/150)2500 = 10,000 community repositories across all of science and engineering. This size is similar to that calculated by Licklider, who stated 100 fields and 1000 subfields, because communities are the next deeper level (for example, Smalltalk is a community within the subfield of programming languages, within the field of computer science). A typical community repository in this computation or in the previous molecular biology computations has 5000 documents, at 20 kilobytes per document for full text. The size of a subfield literature is thus 10 times this, 1 gigabyte, just as computed by Licklider. The vocabulary switching computation thus spanned a representative set of all scientific literature (it used abstracts, not documents, and a sample of communities, so it did not compute the complete literature in toto).
46. B. R. Schatz, "Information Analysis in the Net: The Interspace of the Twenty-First Century", a CIC Forum White Paper for *America in the Age of Information: A Forum*, Committee on Information and Communications (CIC) of the National Science and Technology Council, July 1995. Available at http://www.hpcc.gov/cic/forum/CIC_Cover.html. The CIC is one of nine committees reporting directly to the Science Adviser to the President of the United States.
47. B. R. Schatz, "Building the Interspace," http://csl.ncsa.uiuc.edu/interspace.html
48. J. M. Nyce and P. Kahn, *From Memex to Hypertext: Vannevar Bush and the Mind's Machine* (Academic Press, San Diego, CA, 1991).
49. I thank the members of the DLI project at the University of Illinois in general and the Interspace project in particular, especially H. Chen, K. Powell, and C. Herring. C. Bourne, who was a pioneer in the early days of online information retrieval, carefully reviewed the historical details and suggested many corrections. L. Smith and P. Cochrane also kindly helped with the periods that predated my direct experiences. K. Powell helped with preparation of the figures. Support was provided through NSF-ARPA-NASA DLI grant IRI-94-11318COOP and my NSF Young Investigator award IRI-9257252 in science information systems.

Mathematical and Computational Challenges in Population Biology and Ecosystems Science

Simon A. Levin,* Bryan Grenfell, Alan Hastings, Alan S. Perelson

Mathematical and computational approaches provide powerful tools in the study of problems in population biology and ecosystems science. The subject has a rich history intertwined with the development of statistics and dynamical systems theory, but recent analytical advances, coupled with the enhanced potential of high-speed computation, have opened up new vistas and presented new challenges. Key challenges involve ways to deal with the collective dynamics of heterogeneous ensembles of individuals, and to scale from small spatial regions to large ones. The central issues—understanding how detail at one scale makes its signature felt at other scales, and how to relate phenomena across scales—cut across scientific disciplines and go to the heart of algorithmic development of approaches to high-speed computation. Examples are given from ecology, genetics, epidemiology, and immunology.

Mathematical and computational approaches to biological questions, a marginal activity a short time ago, are now recognized as providing some of the most powerful tools in learning about nature; such approaches guide empirical work and provide a framework for synthesis and analysis (*1*, *2*). In some areas of biology, such as molecular biology, the advent has been recent but rapid—for example, as an adjunct to the analysis of nucleic acid sequences or the structural analysis of macromolecules. In population biology, in contrast, the marriage between mathematical and empirical approaches has a century-long history, rich in tradition and in the insights it has provided. Statistics and stochastic processes, for example, derive their origins from biological questions, as in Galton's invention of the method of genetic correlations and Fisher's creation of the analysis of variance to study problems in agriculture (*1*). Branching processes were developed to describe genealogical histories, and even such

classical subjects as dynamical systems theory have been enriched by contact with problems in population biology [(see *3*, *4*)].

In recent years, the nature of the game has changed, primarily because of the availability of high-speed computation. Classical approaches to population biology—like classical approaches to other problems in biology—emphasized deterministic systems of low dimensionality, and thereby swept as much stochasticity and heterogeneity as possible under the rug. New techniques and the availability of more powerful computers have led to the development of highly detailed models in which a wide variety of components and mechanisms can be incorporated. In a model of animal grouping, every animal can be tracked; in a forest model, every tree; in an epidemiological model, every individual in the population.

Because models of this sort may provide an unjustified sense of verisimilitude, it is important to recognize them for what they are: imitations of reality that represent at best individual realizations of complex processes in which stochasticity, contingency, and nonlinearity underlie a diversity of possible outcomes. Individual simulations cannot be taken as more than representative of this diversity, but repeated simulations can provide statistical ensembles that contain robust kernels of truth. The problem becomes one of the central problems in science: determining what is signal and what is noise by understanding what detail at the level of individual units is essential to understanding more macroscopic regularities.

The issues raised above cut across population biology and ecosystems science, from the immune system to the biosphere. At each level, dynamics can be observed to emerge from the collective behaviors of individual units. The challenge, then, is to develop mechanistic models that begin from what is understood (or hypothesized) about the interactions of the individual units, and to use computation and analysis to explain emergent behavior in terms of the statistical mechanics of ensembles of such units. In the following sections, this challenge is examined for a range of scientific problems. Many of the ideas are explicated in more detail in (*1*) and represent conclusions derived more recently in (*5*). The areas discussed range across a spectrum of problems in population biology, from the populations of B cells and T cells in the immune system, to the variety of genotypes within a population, to the diversity of populations in the biosphere. Though the nature of the biological problems differs, the similarity is what stands out: An individual organism is a biosphere in miniature—with competition, exploitation, mutualism, succession, and nutrient cycling—that provides the stage for evolutionary changes on the small scale, including selfish and cooperative behaviors. Although the subdisciplines that are highlighted have their individual cultures and dynamics, the commonality of the mathematical and computational challenges can foster positive feedbacks that would otherwise not occur.

Ecology

The characterization of ecological interactions provides one of the most venerable of venues for mathematical biology, dating back at least as far as Volterra's consideration of the fluctuations of the Adriatic fisheries. The challenges facing us today—for example, in the consideration of global change and the loss of biodiversity, and in achieving a sustainable future (*6*)—elevate the complexities to new levels.

General circulation models are providing detailed information on likely scenarios of climate change and the global fluxes of key elements such as carbon and nitrogen. Typically, the resolution of such models is at the scale of hundreds of kilometers; how then can we assess likely effects on natural and managed systems, where the scales of interest are typically on the order of meters or even centimeters? Even more difficult, how can we extrapolate from the level of effects on individual plants and animals to changes in the distribution of individuals over longer time scales and broader space scales, and hence in community-level patterns and the fluxes of nutrients?

Individual-based models, such as the forest growth simulators JABOWA (*7*), FORET (*8*), and SORTIE (*9*), provide a point of departure, but the amount of detail in such models cannot be supported in terms of what we can measure and parameterize. The result is that these models produce cartoons that may look like nature but represent no real systems. However, they do represent powerful experimental tools, which become more valuable when used to produce exhaustive simulations that allow exploration of parameter space and model structures; such models permit adequate representation of the full statistical ensemble of possible realizations associated with the many stochastic elements. The development of extensive sets of outputs from multiple runs forms the basis for extracting essential and more robust features that can be compared with data, and that can provide the foundation for simplification (*10*, *11*). Simplification techniques may include familiar tools such as renormalization or moment closure (*12*) in approximations that present more interpretable representations of pattern and dynamics. Computation is an essential adjunct to analysis in developing and testing these approximations.

SORTIE provides a case study in the range of computational problems that can arise with ecological data. Designed to simulate the growth of northeastern forests, SORTIE is a stochastic and mechanistic model that follows the fates of individual trees and their offspring. It uses species-specific information on growth rates, fecundity, mortality, and seed dispersal distances, as well as detailed, spatially explicit information about local light regimes, which change in response to changing distributional patterns of nine dominant or subdominant species. The outputs are dynamic maps of tree species distributions that look like real forests (Fig. 1) and match data observed in real forests at appropriate levels of spatial resolution. Models of this sort, if verified, obviously provide powerful tools for prediction under various hypothetical scenarios of future climate change; more reliably, they provide tools for exploring hypotheses regarding the mechanisms underlying the maintenance of biodiversity and ecosystem processes.

Yet it is fair and important to ask how seriously such predictions should be taken. Surely, such models should not be expected to predict where every tree will be at each point in time; only aggregate statistical properties can be reliably predicted, typically over broad spatial and temporal scales. The great detail regarding local light regimes may be important to the growth of individual trees, but forest dynamics can respond in predictable ways only to more general features of light regimes. To derive robust statements about these systems, it is essential to understand what detail at the local level affects the broader scale patterns, and what is noise.

One approach to this problem [for example, (*10*)] is to carry out extensive simulations in which different degrees of smoothing and aggregation are used, to determine how much information is lost by averaging, and to find out where error is compressed and where it is enlarged in the course of this process. SORTIE typically involves tens of thousands of trees, each having an associated light regime resolved into 216 pixels. The magnitude of the system requires high computational power even for individual simulations; the tasks

S. A. Levin is in the Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA. B. Grenfell is in the Zoology Department, Cambridge University, Downing Street, Cambridge CB2 3EJ, UK. A. Hastings is in the Division of Environmental Studies, Institute for Theoretical Dynamics, and Center for Population Biology, University of California, Davis, CA 95616, USA. A. S. Perelson is at Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

*To whom correspondence should be addressed.

described above magnify this challenge by requiring exploration of statistical ensembles through multiple runs and complex statistical analyses. Simulations carried out for heterogeneous environments require an interface between large dynamic simulations and geographic information systems, providing real-time feedbacks between the two. In some cases, these tasks simply involve known techniques and many cycles, and in other cases they involve the development of new algorithms. There are many outstanding theoretical challenges.

Simplification through extensive simulations is a powerful brute-force method, but the development of analytical approaches to simplification transforms art into science. Again, there is the need both for adapting existing methodologies and for developing new ones. SORTIE may provide the starting point, but abstracted analytical descriptions can potentially reproduce essential qualitative features, and thereby provide more robust and interpretable descriptions of vegetational dynamics. Evaluation of such simplifications requires the output from extensive simulations, the numerical solutions of coupled partial-differential integral equations (*11*, *13*), and the development of theoretical generalizations that may raise sophisticated mathematical challenges. The richness of mathematical and computational issues is matched only by the great potential for increasing our ability to understand and predict the dynamics of forests. Moreover, the creation of interfaces between the self-organizing dynamics implicit in these models and the imposed environmental regimes derived from geographical information systems, remote sensing, or the output of climate models allows exploration of the interplay between intrinsic and extrinsic factors in shaping vegetational patterns.

Global change and vegetational responses to it provide one set of challenges, but similar issues exist in the description of other ecological phenomena. Populations typically are made up of diverse and heterogeneous assemblages of individuals, each with unique characteristics. As such, they differ from the more uniform assemblages usually treated in statistical mechanics, but the challenges are similar. How do we represent the mean dynamics of such heterogeneous assemblages without retaining all of the detail, much of it irrelevant to the essential dynamics? How much information, beyond variances and covariances, do we need to retain in order to provide reasonable descriptions, and how can we close up those descriptions in terms of the dynamics of the higher moments? Similar questions exist classically not only in the physical sciences, but also in evolutionary biology (*14*). Evolution feeds off the variances and covariances within populations, and in return helps to shape that variance-covariance structure. The recognition of this phenomenon, and of ways to deal with it, has provided some of the most powerful approximations to the dynamics of quantitative inheritance.

The maintenance of biological diversity and approaches to sustainable use raise similar issues. The heterogeneous distribution of resources and exploiters is a fact of overwhelming importance to understanding dynamic interactions, as well as an ecological and evolutionary consequence of those interactions (*11*, *13*). Thus, the description of the dynamics of aggregations of fish, krill, birds, or foraging vertebrates requires an understanding of how factors at the level of individuals determine the cohesion, fusion, and fission of groups, and of the consequences of those processes and patterns for ecological interactions such as harvesting for food or predation. Again, a powerful starting point is the individual: Lagrangian descriptions of individual movements make attractive cartoons (*15*) and can provide a basis for analysis; and again, extensive simulations can provide the foundation for the exploration of robust cause-and-effect relations and for the extraction of statistical mechanical and Eulerian field descriptions that capture the essence of the dynamics. In the same manner as for the vegetational systems the interplay between extrinsic and intrinsic factors can also be explored through computation—for example, by imposing flow regimes derived from Navier-Stokes equations upon the dynamics of attraction and repulsion in marine systems (*16*).

Spatial heterogeneity is the most obvious of ways that nonuniform distributions may be important, but other dimensions provide even greater challenges. In epidemiology (see below), heterogeneous mixing among different risk groups can provide a fundamentally altered view of disease dynamics, especially for sexually transmitted diseases (STDs). Regarding biological diversity, although it is widely acknowledged that species are being lost at rates never before experienced, what is equally important is the loss of diversity at other scales—not only within species (genetic diversity, or even simply the loss of populations), but also within functional groups of species performing essential ecosystem functions. The most important consequences of the disappearance of biodiversity may be in the loss of such ecosystem services as the maintenance of fluxes of nutrients and pollutants, the mediation of climate and weather, and the stabilization of coastlines.

In developing priorities for the conservation of biodiversity, it becomes important to identify and understand the most fragile and critical components of ecological systems, in terms of their capability to sustain these services. Again, this means understanding the degree to which aggregate behavior is linked to the dynamics of higher moments representing distributional features. The approach is the same as discussed previously [for example, (*17*)]: extensive simulations of detailed models, comparison with aggregated models, and the development of rules for relating these models to one another and for providing simplified descriptions. In all of these problems, there are common mathematical and computational challenges that range from techniques for representing and accessing data sets, to algorithms for simulation of large-scale spatially stochastic systems, to the development and analysis of simplified descriptions. These themes will reappear below.

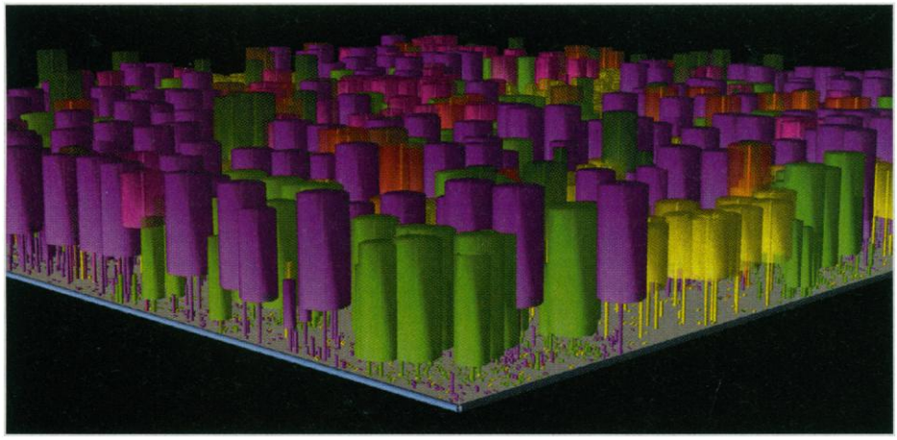


Fig. 1. Visualization of a 9-hectare SORTIE forest, 500 years into the simulation. Each cylinder represents an individual, where height and cylinder diameter are based on species-specific parameters (*96*). Green, Eastern hemlock; purple, beech; yellow, yellow birch.

Genetics and Evolution

The heritage of mathematics in evolutionary and genetic studies has been extraordinary, beginning with the work of the three giants—Fisher, Haldane, and Wright—and continuing to this day. Although much of the basic framework of population genetics thus has roots deep in the history of the subject, contemporary questions ranging from the very basic (*18*) to the applied [for example, conservation biology (*19*) and the use of transgenic organisms] are raising new and important mathematical challenges. Despite the relative simplicity of the underlying genetic models, complexities ranging from multiple loci to spatial factors to the role of frequency dependence in evolution (*20*) lead to problems that require sophisticated computational approaches. The considerations underlying the management and analysis of genetic sequence data are well known; hence, the following discussion focuses on other facets of evolution and genetics that lead to deep computational and mathematical challenges, especially regarding dynamics.

Although the dynamics of alleles at single loci were well understood in the 1920s, the inclusion of just one more locus leads to models whose dynamics are still not completely understood, even in the deterministic case (*21*). A full understanding of the behavior of these two-locus models has required the use of a variety of computational approaches, from straightforward simulation [for example, (*22*)] to more complex analyses based on optimization (*23*) or the use of computer algebra systems. The consideration of as few as three loci leads to models whose behavior can only be understood by means of numerical approaches, except for some very special cases (*21, 24*); yet the number of loci exhibiting genetic variation in populations of higher organisms is well into the thousands. Including all this complexity leads to the consideration of populations in which the number of possible genotypes could be much larger than the population. Thus, stochastic effects become paramount, and even the simulation of such populations (*25*) leads to problems of substantial computational difficulty (*26*).

Faced with the impossibility of constructing a theory of evolution of characters controlled at many loci by detailed consideration of what is going on at each locus, evolutionary biologists have turned to more macroscopic representations at the level of the phenotype, an attractive option because of the ease of observation and description. The simplest such approaches involve quantitative traits, such as height or weight, or other traits of ecological interest that represent the sum of multiple small effects. Recently, there have been substantial efforts (*14*) to integrate the long tradition of using statistical approaches to model the dynamics of quantitative traits with the more mechanistic genetic approaches, and hence to provide a rigorous basis for treating quantitative traits. The problem of closure arises again, and even under simplifying assumptions concerning the relation between genotype and phenotype, further approximations are required to obtain a closed system of equations (*14, 27*). Confirmation of the appropriateness of these approximations ultimately rests on comparisons with both natural and artificial populations as well as on the results of computer simulations.

The study of complex adaptations can lead to questions about the evolution of evolvability itself (*28, 29*). How does selection act to modify the capability of organisms to adapt to changing environments? This can become an extraordinarily complex question; one intriguing avenue to identifying the kinds of questions that arise has been to create "artificial life" through computer simulations [for example, (*30*)], and hence to explore how the rules that govern evolution develop and become modified. Often, the resulting simulations are so seductive that the boundary between truth and fiction becomes blurred, but the potential for developing novel insights cannot be denied. Needless to say, the computational problems that arise are substantial and are leading to new innovations in programming.

The flow between computation and biology is not one-way; as in the example of artificial life, computation can draw inspiration from biology. A case in point involves the invocation of evolutionary processes that use a variety of distinct approaches (*29, 31, 32*), all of which have at least some of the formal structure of genetic systems, to solve very complex optimization problems by identifying strategies with computer "genotypes." For various reasons, the solutions found by such approaches may bear little similarity to how natural selection solves similar problems (*32*). Historically, the search for optimization principles to apply to natural evolutionary systems has had limited success, largely because of frequency dependence (the dependence of relative fitnesses on the frequencies of types in the populations); that is, evolution is best understood as a problem in game theory rather than optimization theory.

To address problems of frequency dependence, which arise naturally in the consideration of most interesting ecological problems, Maynard Smith introduced the notion of an evolutionarily stable strategy (ESS) (*33*), which has been used extensively to understand the evolution of behavior, especially altruistic behaviors. An elegant theory developed by Hamilton (*34*) based on inclusive fitnesses can explain why individuals might forego their own fitnesses to help relatives, but the evolution of altruism between unrelated individuals is much more difficult to explain.

The central issue in the evolution of altruism is to determine how cooperation can evolve through individual selection. A simple model system is provided by the familiar game of prisoner's dilemma, for which the game theoretic solution (for a single encounter) is noncooperation (Fig. 2). Evolutionary biologists have been able to explain the evolution of altruism by focusing on multiple repetitions of the games and on correlations that arise in time or space; such correlations affect realized payoffs because they affect who plays with whom.

In particular, when the game is played repeatedly, as in iterated prisoner's dilemma (*35*), it can be shown that tit-for-tat, which consists of beginning with cooperation and then using the strategy used by the other "player" in the previous interaction, is better than the pure defecting strategy [and that no pure strategy is an ESS (*36*)]. Sophisticated simulations (*37*) allow exploration of more complex ESSs in which individuals remember past interactions, and the result is a greater ease of evolving cooperative strategies. Spatial localization of interactions further increases the probability that the same partners will play the game repeatedly and facilitate evolution of cooperatives.

In general, the introduction of explicit

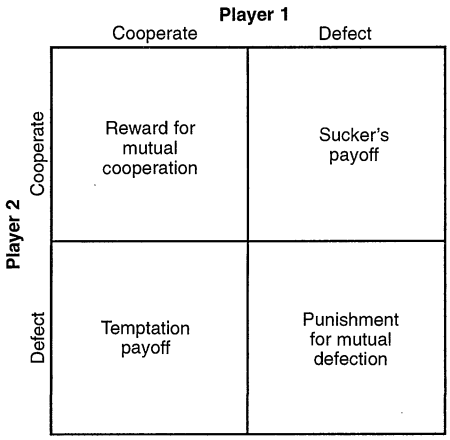
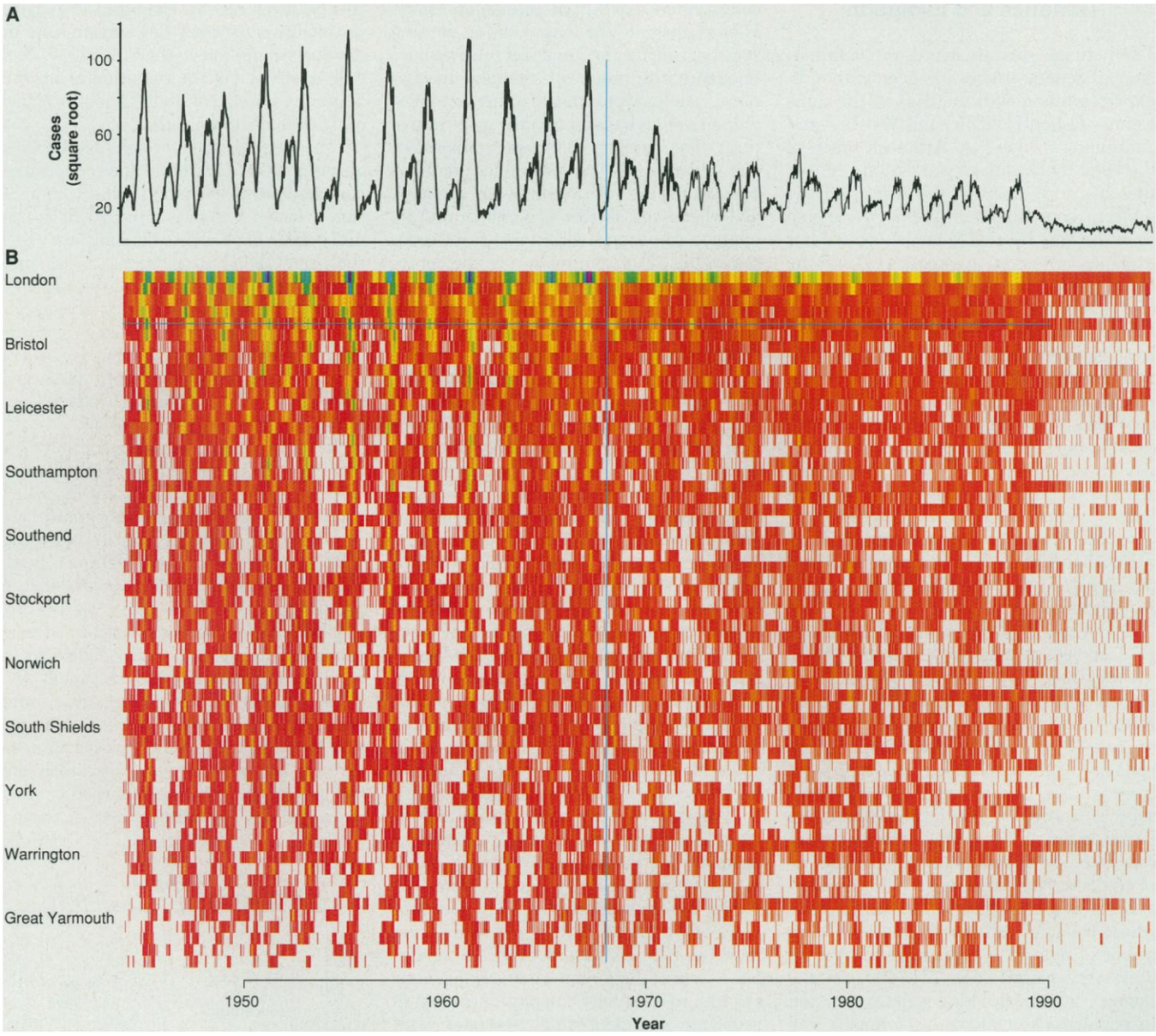


Fig. 2. Payoff matrix in the prisoner's dilemma game, where each box lists the payoff to player 2 when players 1 and 2 play the pair of strategies indicated [redrawn from (*97*)]. The game is a prisoner's dilemma if the reward for cooperation is greater than the average of the sucker's payoff and the temptation payoff, and the payoffs are ordered so that temptation payoff > reward for cooperation > punishment > sucker's payoff. In an evolutionary sense, the problem is to explain how strategies involving cooperation among nonrelated individuals evolve.



space produces further complications, leading to results that depend fundamentally on population structure and movement rules. The underlying principle is that the evolution of traits for which fitnesses are frequency-dependent requires knowledge of which individuals are interacting; thus, for large populations, simulations (*38*, *39*) are needed to understand dynamics in spatially structured populations. Prisoner's dilemma is a caricature, and more biologically relevant studies are beginning to show the importance of the spatial localization of interactions in the evolution of both cooperative and antagonistic behaviors (*38*, *40*). Substantial questions remain to be explored, including the evolution of more complex behaviors [for example, (*41*)] and coevolutionary questions. For parasite-host systems, the problem has been well studied [for example, (*42*)], but more diffuse interactions involving many species introduce challenges similar to those that arise in going from two loci to many loci. Fundamental challenges exist in understanding how community properties emerge from the evolution of component species, an issue that is at the core of research into biodiversity.

Infectious Diseases

The mathematical theory of the population biology of infectious diseases dates back at least as far as Daniel Bernoulli's mathematical analysis of smallpox control in 1760. The main impetus for this highly successful field has been the great impact of disease on human health and agriculture, both historically and in facing the threat of acquired immunodeficiency syndrome (AIDS) and other emerging diseases. However, parasite ecology—which effectively links ecological and immunological dynamics—also presents a number of fundamental questions for mathematical and computational research. Simple models have been remarkably successful in capturing many features of host-parasite dynamics and control (*43*, *44*). However, as with ecology, the interaction between spatial and genetic heterogeneity, nonlinearity, and stochasticity can complicate this picture.

• A major preoccupation for epidemiological modeling is how transmission varies

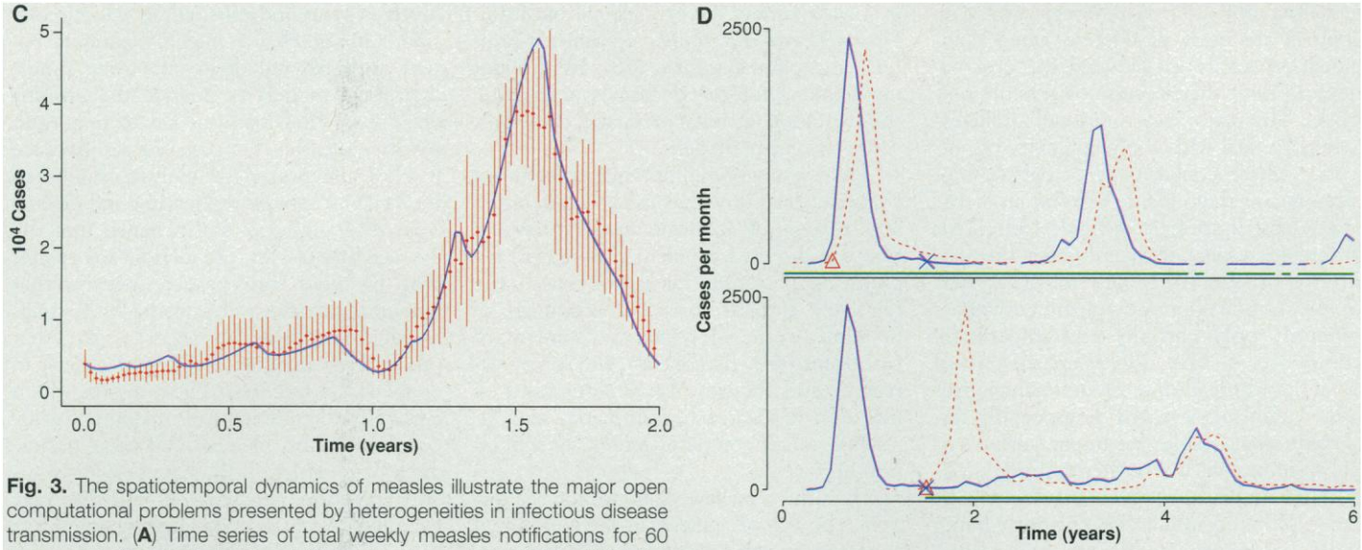


Fig. 3. The spatiotemporal dynamics of measles illustrate the major open computational problems presented by heterogeneities in infectious disease transmission. (**A**) Time series of total weekly measles notifications for 60 towns and cities in England and Wales, for the period 1944 to 1994; the vertical blue line represents the onset of mass vaccination around 1968. (**B**) An image plot, showing the breakdown of cases for individual centers, ranked by population size; red indicates zero notifications, and other colors represent cases on a topological scale, from blue (small) to green and brown (large). The large-scale prevaccination dynamics are well represented by age-structured deterministic models (*46, 47, 54, 98–101*). (**C** and **D**) Pattern comparisons. The average observed biennial pattern (± SE) is compared in (C) with the limit cycles of the best-fit deterministic model (solid line) [see (*47, 98*) for more details]. By contrast with homogeneous models, which tend to predict large-amplitude chaotic dynamics (*50*), this age-structured formulation indicates that stochastically perturbed coexisting limit cycles may be the norm (*102*). A horizontal line in (B) marks the population threshold—the critical community size (CCS) (*103*)—above which measles persisted endemically, without local extinction of infection, in the prevaccination era. Recent developments of stochastic models can begin to capture this threshold (*54*), though much more needs to be done in explaining fully the complex spatiotemporal structure summarized in (B). The framing of explicit spatial structure—as "patch" models (*46, 99*), pair approximations to individual-level interactions (*104*), and power-law approaches to irregular epidemics in small populations (*105*)—shows promise for exploring the persistence of measles metapopulations (*101*). However, modeling anything approaching the full hierarchical spatial dynamics will require refinements to computational and analytical approaches and to nonlinear statistical analyses of the balance between deterministic and stochastic dynamics (*58*). One of the most interesting questions for such models is to explore the emergent spatial effects of vaccination. The CCS in (B) remained remarkably constant during most of the vaccine era (*53, 106*), and there was considerable persistence of measles even in the 1990s when high vaccine uptake greatly reduced its incidence. Preliminary analyses (*53*) indicate that this may be the result of a "rescue effect" arising from the observed decorrelation of epidemics caused by vaccination (*107*). This is illustrated in (D), which shows how simulated epidemics in two coupled centers (center 1, solid blue line; center 2, dashed red line) show global extinctions of infection, when the epidemics are in phase (top panel). Moving the epidemics out of phase (bottom panel) eliminates fadeouts attributable to cross-infection between the centers; details are given in (*53*). Global fadeouts of infection in the two centers are denoted by breaks in the green line on the time axis. The triangle and X in each panel illustrate the phase shift; these points, which are a year apart in the top panel, are brought together in the bottom panel by shifting center 2's dynamics forward in time. Long-term changes in the availability of susceptibles, as a result of birth rate trends, can also affect the spatiotemporal dynamics of infection in complex ways (*108*).

with social or geographical space (*44, 45*). A key theoretical issue here is how, and in what detail, to represent spatial variations in the intrinsically nonlinear contact process underlying transmission. One of the best illustrations of this process is provided by the highly dynamic spatiotemporal epidemic pattern of measles (Fig. 3) (*46–49*). An important set of analyses of simple, homogeneous models predicted the possibility of chaotic dynamics (*50*); however, the resulting large-amplitude epidemics generate unrealistically low persistence of infection in small communities (*51*). Adding successive layers of social and geographical space—and moving from deterministic to stochastic models—improves spatial realism and may reduce the propensity for chaos (*46, 47, 52–54*).

The major computational challenge in these highly nonlinear stochastic systems is to represent hierarchical spatial complexity and especially its impact on vaccination strategies. Depending on the problem, all scales—from the individual level to big cities—may be important, both in terms of social space [family and school infection dynamics (*55*)] and in terms of geographic spread and coherency (Fig. 3). As in ecology and evolution, a central question is: How spatially aggregated and parsimonious a model can provide useful results in a given context? This is particularly important in comparisons between directly transmitted human infections—where long-range movements may bring infection dynamics comparatively close to mean field behavior (in which every individual is assumed to have equal contact with every other individual, thus experiencing the mean or average field)—and the equivalent infections in natural populations, where more restricted movements and host population dynamics add extra complexities (*56*).

It is risky to model at a given level of detail without having data at the relevant spatial grain. Notifiable infectious diseases are unusually well provided here (Fig. 3), with large and often as yet uncomputerized spatiotemporal data sets. These data provide a huge potential testbed for developing methods for characterizing spatiotemporal dynamics in nonlinear, nonstationary stochastic systems. An encouraging development is that the current, generally nonparametric, approaches to characterizing chaos and other nonlinear behaviors are increasingly incorporating lessons from mechanistic epidemiological models (*49, 57, 58*).

The main focus for modeling social space (the space of social interactions) and disease is, of course, on AIDS and other sexually transmitted infections. Simple models illustrated clearly that heterogeneities in contact rates can substantially alter the predicted course of epidemics (*43*). This area has seen an explosion of research, both in data analysis of contact structures and in graph-theoretic and other approaches to

modeling (*43*, *59*, *60*). Models and data analysis are most productive when combined, especially in allowing the observations to limit the universe of possible networks. The major computational challenge is how to deal with the complexity of networks, where concurrency of partnerships often means that closure to a few moments of the distribution is difficult (*60*). This problem is especially acute given the sensitivity of obtaining data for STD networks, in that the nature of the network is generally only partially and imperfectly known (*61*). The use of mathematical models for human immunodeficiency virus (HIV) transmission will be especially important in assessing the impact of potential vaccines (*62*). Another major computational challenge—which developed with the AIDS epidemic and is currently being applied to another pathogen, the bovine spongiform encephalopathy agent (*63*)—is to estimate the parameters of transmission models from disease incidence and other demographic data.

One hope for the future for both of these areas is network information embedded in viral genomes. A body of recent work indicates exciting possibilities for estimating epidemiological parameters from the birth and death processes of pathogen evolutionary trees (*64*). More generally, new mathematical and computational techniques will be needed to understand the epidemiological implications of the rapidly accumulating data on pathogen sequences, especially in the context of parasite genetic diversity and the host immunological response to it (*65*).

The other major area of current epidemiological interest, the impact of host and parasite genetic heterogeneity and coevolution (*66*), has a distinguished history in population genetics and epidemiology. However, the revolution in both genome research and molecular epidemiology is now providing the foundations for much more detailed explorations of the dynamics of host and parasite strains. An important linked area here is the question of immunoepidemiology (*67*)—modeling the population-dynamic implications of the immunological processes described in the next section. These approaches come together, for example, in recent work on the strain dynamics of malaria (*68*), in which models of observed strain and immunological variation indicate a set of cocirculating strains rather than the traditional homogeneous picture of a single, highly transmissible entity.

The major computational question is again to represent hierarchical spatial dynamics, but with the added problem (and hence the added dimensionality) of complex within-host dynamics and host-parasite genetic diversity. The genetic dynamics of a wide variety of pathogens, from influenza (*69*) and HIV to macroparasitic worms (*70*) and plant parasites (*66*), have major implications for the dynamics of control, the evolution of resistance, and the emergence of new pathogens.

These issues present a range of technical computational problems in the assimilation and analysis of data and model construction. For instance, moment closure (*12*) is a promising possibility for approximating the relatively smooth stochastic dynamics of helminth worm infections and some plant pathogens (*66*). By contrast, the spikey dynamics and frequent local extinctions of infection in measles and influenza seem to require more computer-intensive simulation approaches.

Over the next few years, we foresee further major development in computational approaches to the complexities of host-parasite spatial and genetic dynamics. Two areas that are likely to be of particular interest are integrating dynamics at the epidemiological, genetic, and immunological levels and exploring the new dynamical properties of systems revealed by parasite control strategies (Fig. 3). In terms of impact on human welfare, research on the dynamics of infectious diseases in developing nations is an important priority.

Immunology and Virology

Historically, mathematical and computational methods have not played a large role in immunology and virology. This is now changing, and impressive advances have come from the use of simple models applied to the interpretation of quantitative data.

The best example is in AIDS research. As is well known, AIDS develops slowly; the average time from HIV infection to the development of full-blown AIDS is about 10 years. Modeling of the progression to AIDS has received considerable attention and has been able to capture much of the observed phenomenology (*71*, *72*). The suggestion that progression to AIDS involves a diversity threshold (*72*) has generated debate, new theory, and new experimentation (*73*). The role of the immune response in determining the pace of disease progression has yet to be clarified, but mathematical modeling has helped focus attention on the role of cytotoxic T cells (*74*, *75*). Other key areas in which modeling has played and will continue to play an important role are the understanding of how HIV evolves resistance to antiretroviral drugs and the design of treatment strategies (*76*).

Much of the 10-year period until AIDS develops has been characterized as a period of clinical latency, with low but constant levels of virus and infected cells in circulation. Giving HIV-1–infected patients potent antiretroviral drugs and using simple dynamical models to analyze the ensuing decline in viral load has led to important insights into the in vivo processes involved in HIV infection. This analysis established that HIV is rapidly replicating and cleared from the body (*77*) and revealed that the average rate of HIV production was greater than 10 billion virus particles per day, that free virus particles were cleared with a half-life that is probably 6 hours or less, and that productively infected T cells had a life-span of about 1.5 days (*78*). These results, which derive from mathematical modeling, firmly put to rest the view of AIDS as a slow disease in which little happens for years after infection, and replaced it with a new paradigm in which rapid viral dynamics was the centerpiece. Most important, uncovering the rapid replication of HIV led to a new understanding of the observed rapid evolution of the virus and the seemingly inevitable emergence of drug-resistant forms of HIV-1. In part as a result of this increased understanding, treatment protocols using a single drug are being replaced by protocols using combinations of antiretroviral drugs, which have a greater antiretroviral effect and which increase the number of mutations needed for resistance. The early clinical results of combination therapy, along with mathematical modeling, have now been used to obtain minimal estimates for how long therapy needs to be maintained until HIV is eliminated from the body (*79*).

The new finding that HIV uses two receptors for entry into target cells—a primary receptor (CD4) and a coreceptor [a chemokine receptor, either fusin (now renamed CXCR4) or CCR5] (*80*)—provides new challenges and opportunities for modeling. Using concepts from population genetics, researchers have argued that individuals who are homozygous for a 32-nucleotide deletion in the CCR5 gene are resistant to HIV-1 infection and otherwise show no drastic decrease in fitness as a result of this deletion (*81*). The homozygous defect is found in approximately 1% of Caucasians of Western European ancestry (*81*). Models of HIV-1 dynamics have assumed that infection is a single-step process. New models need to account for coreceptors and for the interesting finding that high-affinity binding of HIV-1 gp120 to the first HIV receptor, CD4, causes conformational changes in gp120 that lead to the creation of a new recognition site on gp120 for CCR5 (*82*). Lastly, CCR5 has been identified as the major coreceptor for macrophage-tropic HIV-1 strains. Although some mathematical models have considered macrophage in-

fection (79, 83), none yet have incorporated coreceptors.

Opportunities also exist for modeling to provide insights into the dynamics of other infectious diseases. Hepatitis, which currently infects more than 250 million people worldwide, is an important target for modeling, and work in this direction has begun (84). Models that incorporate immune responses and deal with the issue of drug resistance that can arise during treatment are of great importance and can yield insights into treatment strategies for tuberculosis, HIV, and other infectious agents (76, 85).

Spatial considerations, which play a large role in ecological and epidemiological modeling, also enter into virological and immunological problems. For example, in humans, detection of virus is most easily done in the blood, yet virus can be distributed throughout the body. Models and experiments now need to address the question of observability—that is, how well do measurements in blood reflect other compartments? New experiments and models are being designed that take into consideration bodily compartments where virus and T cells are found, for example, lymph nodes (86). Also, because drugs are transported through tissues, drug concentrations vary in space and time. Models need to be developed that allow for drug transport and differing concentrations at different locations, although some modeling has been initiated in other contexts (87). Such models are particularly relevant for agents such as monoclonal antibodies that can rapidly bind to cells as they move through tissue (88). The implication of spatial and temporal gradients for the generation and selection of drug-resistant organisms needs to be examined.

In basic immunology, issues related to mutation also have been the focus of mathematical modeling and intense experimentation (89, 90). During the course of an immune response, B lymphocytes within germinal centers can rapidly mutate the genes that code for antibody variable regions. The immune system thus provides an environment in which evolution occurs on a time scale of weeks. Among the large number of mutant B cells that are generated, selection chooses for survival those B cells that have increased binding affinity for the antigen that initiated the response. After 2 to 3 weeks, antibodies can have improved their equilibrium binding constant for antigen by one to two orders of magnitude, and may have sustained as many as 10 point mutations. How can the immune system generate and select variants with higher fitness this rapidly and this effectively? An optimal control model has suggested that mutation should be turned on and off episodically in order to allow new variants time to expand without being subjected to the generally deleterious effects of mutation (90). Time-varying mutation could be implemented by having cells recycle through one region of the germinal center, mutating while there, and proliferating in a different region of the germinal center (90). This suggestion has generated new experimental investigations of events that occur within germinal centers (91). Opportunities exist for a range of models that address basic questions about in vivo cell population dynamics and evolution, as well as more detailed questions involving the immunological mechanisms underlying affinity maturation.

Control of the immune response is another area ripe for modeling. What determines the intensity of a response? How is the response shut off when the antigen is eliminated? Feedback mechanisms may exist to control the response intensity, response length, and type of response (cellular or antibody). Some models of a basic feedback mechanism involving two types of helper T cells, T_H1 and T_H2, have been developed (92); others are needed. Regulatory mechanisms involve interactions among many cell populations that communicate by direct cell-cell contact and through the secretion of cytokines. Diagrams representing the elements of regulatory schemes commonly have scores of elements. Because of the complexities involved, theorists have an opportunity to lead experimentation by providing suggestions as to what needs to be measured and how such measurements can be used to provide an insightful view of possible control mechanisms.

A fundamental feature of the immune system is its diversity. Successful recognition of antigens appears to require a repertoire of at least 10^5 different lymphocyte clones. The diversity of the immune system has challenged experimentalists, and many recent advances have come from developing experimental models with limited immune diversity. However, models based on ecological concepts may provide insights into the control of clonal diversity (75, 93), and modern computational methods now make it practical to consider models with tens of thousands of clones. Thus, it is possible to develop models that start to approach the size of small immune systems. Simulations have suggested that from simple rules of cell response, emergent phenomena arise that may have immunological significance (94). The challenge in using computation is to develop models that address important questions, are realistic enough to capture the relevant immunology, and yet are simple enough to be revealing.

Conclusions

The problems discussed above are distinguished by their centrality to basic and applied biological research as well as by the mathematical and computational challenges they pose. In this regard, they are in a great tradition that reaches back to Galton and Fisher, to Lotka and Volterra, with such recent examples as the contribution of population biology to the development of the theory of chaos (1, 3, 5, 95). This is not surprising; the central issues—understanding how detail at one scale makes clear its signature on other scales, and how to relate phenomena across scales—cut across scientific disciplines, and indeed go to the heart of algorithmic development of approaches to high-speed computation.

Imaginative and efficient computational approaches are essential in dealing with the overwhelming complexity of biological systems. Such approaches should comprise the storage and retrieval of vast amounts of information as well as the development of simulation methods that must interact with those data structures and deal with complex hierarchical systems, taking advantage where possible of parallel structures and symmetries that allow simplification and efficient organization of computational steps. The potential for benefits to mathematics and computational sciences as well as to the applications of these methods will create a rich mutualism, in which the rate of advance is nonlinear. The face of the science of computational population biology and ecosystems science will change in the next decade. Key challenges involve ways to describe the dynamics of systems that are aggregates of heterogeneous units, representing the behavior of the means and lowest moments in closed form. Spatial heterogeneity and spatial localization of interactions introduce qualitatively new dynamics, and they present theoretical and computational issues that are similar across a range of biological levels.

REFERENCES AND NOTES

1. S. A. Levin, Ed., *Mathematics and Biology: The Interface* (Lawrence Berkeley Laboratory, University of California, Berkeley, CA, 1992).
2. J. D. Murray, *Mathematical Biology*, vol. 19 of *Biomathematics* (Springer-Verlag, Heidelberg, 1990); P. J. Hilts, "Eric Steven Lander: Love of Numbers Leads to Chromosome 17," *New York Times*, 10 September 1996, p. C1.
3. R. M. May, *Science* **186**, 645 (1974).
4. J. Roughgarden, R. M. May, S. A. Levin, Eds., *Perspectives in Ecological Theory* (Princeton Univ. Press, Princeton, NJ, 1989).
5. P. Kollman, Ed., *Modeling of Biological Systems: A Workshop at the National Science Foundation*, University of California, San Francisco, 14 and 15 March 1996 (technical report).
6. J. Lubchenco *et al.*, *Ecology* **72**, 371 (1991).
7. D. B. Botkin, J. F. Janak, J. R. Wallis, *J. Ecol.* **60**,

849 (1972); *IBM J. Res. Dev.* **16**, 101 (1972).
8. H. H. Shugart and D. C. West, *J. Environ. Manage.* **5**, 161 (1977).
9. M. A. Huston, *Tree Physiol.* **9**, 293 (1991); H. H. Shugart and I. C. Prentice, in *A Systems Analysis of the Global Boreal Forest*, H. H. Shugart, R. Leemans, G. B. Bonan, Eds. (Cambridge Univ. Press, Cambridge, 1992), pp. 313–333; H. H. Shugart and T. M. Smith, *Annu. Rev. Ecol. Syst.* **23**, 15 (1992); D. L. Urban and H. H. Shugart, in *Plant Succession: Theory and Prediction*, D. C. Glenn-Lewin, R. K. Peet, T. T. Veblen, Eds. (Chapman and Hall, London, 1992), pp. 249–292; S. W. Pacala, C. D. Canham, J. A. Silander, *Can. J. Forest Res.* **23**, 1980 (1993).
10. D. H. Deutschman, thesis, Cornell University (1996); S. W. Pacala and D. H. Deutschman, *Oikos* **74**, 357 (1995).
11. S. W. Pacala and S. A. Levin, in *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions*, D. Tilman and P. Kareiva, Eds. (Princeton Univ. Press, Princeton, NJ, in press).
12. L. R. Taylor, *Nature* **189**, 732 (1961); S. K. Ma, *Modern Theory of Critical Phenomena* (Benjamin, New York, 1976); K. Wilson, *Rev. Mod. Phys.* **55**, 583 (1983); G. I. Barenblatt, *Scaling Phenomena in Fluid Mechanics* (Cambridge Univ. Press, Cambridge, 1994).
13. S. A. Levin and S. W. Pacala, in (*11*).
14. N. H. Barton and M. Turelli, *Genetics* **138**, 913 (1994); E. S. Lander and N. J. Schork, *Science* **265**, 2037 (1994); N. Risch and K. Merikangas, *ibid.* **273**, 1516 (1996).
15. C. W. Reynolds, *Comput. Graphics* **21**, 25 (1987); S. Gueron and S. A. Levin, *J. Theor. Biol.* **165**, 541 (1993); S. Gueron, S. A. Levin, D. I. Rubenstein, *ibid.* **182**, 85 (1996); D. Grünbaum, *Evol. Ecol.*, in press.
16. G. Flierl, D. Grünbaum, S. A. Levin, D. B. Olson, unpublished manuscript.
17. B. M. Bolker, S. W. Pacala, F. A. Bazzaz, C. D. Canham, S. A. Levin, *Global Change Biol.* **1**, 373 (1995).
18. N. H. Barton, *Philos. Trans. R. Soc. London Ser. B* **351**, 785 (1996).
19. R. Lande, *Evolution* **48**, 1460 (1994).
20. S. Gavrilets and A. Hastings, *Proc. R. Soc. London Ser. B* **261**, 233 (1995).
21. T. Nagylaki, *Introduction to Theoretical Population Genetics*, vol. 21 of *Biomathematics* (Springer-Verlag, Berlin, 1992); A. Hastings, in *Some Mathematical Questions in Biology: Models in Population Biology*, A. Hastings, Ed., vol. 20 of *Lectures on Mathematics in the Life Sciences* (American Mathematical Society, Providence, RI, 1989), pp. 27–54.
22. M. W. Feldman, F. B. Christiansen, S. P. Otto, *Genetics* **129**, 297 (1991).
23. G. A. Fox and A. Hastings, *ibid.* **132**, 277 (1992).
24. As an example, the analytical results for a special case [W. P. Robinson, M. A. Asmussen, G. Thomson, *Genetics* **129**, 925 (1991)] were coupled with numerical results (W. P. Robinson, A. Cambonthomsen, N. Borot, W. Klitz, G. Thomson, *ibid.*, p. 931) to understand the evolution of a three-locus system.
25. R. Burger, G. P. Wagner, F. Stettinger, *Evolution* **43**, 1748 (1989).
26. S. J. E. Baird, *ibid.* **49**, 1038 (1995).
27. S. Gavrilets and A. Hastings, *Genet. Res.* **65**, 63 (1995).
28. S. A. Kauffman, *Physica D* **42**, 135 (1990).
29. G. P. Wagner and L. Altenberg, *Evolution* **50**, 967 (1996).
30. T. S. Ray, in *Artificial Life II*, C. G. Langton, C. Taylor, J. D. Farmer, S. Rasmussen, Eds. (Santa Fe Institute, Santa Fe, NM, 1992), pp. 371–408; C. G. Langton, Ed., *Artificial Life*, vol. 12 of *Santa Fe Institute Studies in the Sciences of Complexity* (Addison-Wesley, Redwood City, CA, 1989).
31. W. Banzhaf and F. H. Eckman, Eds., *Evolution As a Computational Process*, vol. 899 of *Lecture Notes in Computer Science* (Springer-Verlag, Heidelberg, 1995); J. H. Holland, *Adaptation in Natural and Artificial Systems* (MIT Press, Cambridge, MA, 1992); J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (MIT Press, Cambridge, MA, 1992); I. Rechenberg, *Evolutionsstrategie '94* (Frommann, Stuttgart, 1994).
32. S. Kauffman and S. A. Levin, *J. Theor. Biol.* **128**, 11 (1987); H. Muhlenbein and D. Schlierkamp-Voosen, in *Evolution as a Computational Process*, W. Banzhaf and F. H. Eckman, Eds., vol. 899 of *Lecture Notes in Computer Science* (Springer-Verlag, Heidelberg, 1995), pp. 142–168.
33. J. Maynard Smith, *Evolution and the Theory of Games* (Cambridge Univ. Press, Cambridge, 1982).
34. W. D. Hamilton, *J. Theor. Biol.* **7**, 1 (1964).
35. R. Axelrod and W. D. Hamilton, *Science* **211**, 1390 (1981).
36. R. Boyd and J. P. Lorberbaum, *Nature* **327**, 58 (1987).
37. J. H. Miller, *J. Econ. Behav. Organ.* **29**, 87 (1996).
38. M. A. Nowak and R. M. May, *Nature* **359**, 826 (1992); M. A. Nowak, S. Bonhoeffer, R. M. May, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 4877 (1994); R. Durrett and S. A. Levin, *J. Theor. Biol.*, in press.
39. C. Dytham and S. B. Shorrock, *Evol. Ecol.* **9**, 508 (1995).
40. A. P. Kinzig and J. Harte, in preparation; E. Klopfer, unpublished manuscript.
41. A. Bergman and M. W. Feldman, *Theor. Popul. Biol.* **48**, 251 (1995).
42. S. A. Levin, in *Coevolution*, M. Nitecki, Ed. (Univ. of Chicago Press, Chicago, 1983), pp. 21–65; in *Population Biology*, H. Freedman and C. Strobeck, Eds., vol. 52 of *Lecture Notes in Biomathematics* (Springer-Verlag, Berlin, 1983), pp. 328–334; R. M. May and R. M. Anderson, in *Coevolution*, D. J. Futuyma and M. Slatkin, Eds. (Sinauer, Sunderland, MA, 1983), pp. 186–206.
43. R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).
44. K. Dietz, *Statist. Methods Med. Res.* **2**, 23 (1993).
45. A. D. Cliff and P. Haggett, *Atlas of Disease Distributions: Analytic Approaches to Epidemiologic Data* (Blackwell, Oxford, 1988); D. Mollison and S. A. Levin, in (*56*), pp. 384–398.
46. N. M. Ferguson, R. M. Anderson, R. M. May, in (*11*).
47. B. M. Bolker and B. T. Grenfell, *Proc. R. Soc. London Ser. B* **251**, 75 (1993).
48. A. D. Cliff, P. Haggett, M. Smallman-Raynor, *Measles: An Historical Geography of a Major Human Viral Disease from Global Expansion to Local Retreat, 1840–1990* (Blackwell, Oxford, 1993).
49. B. T. Grenfell, in *Chaos from Real Data: The Analysis of Non-Linear Dynamics in Short Ecological Time Series*, J. N. Perry and R. Smith, Eds. (Academic Press, New York, in press).
50. L. F. Olsen and W. M. Schaffer, *Science* **249**, 499 (1990).
51. B. T. Grenfell, *J. R. Stat. Soc. Ser. B* **54**, 383 (1992).
52. A. L. Lloyd and R. M. May, *J. Theor. Biol.* **179**, 1 (1996).
53. B. M. Bolker and B. T. Grenfell, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 12648 (1996).
54. M. J. Keeling and B. T. Grenfell, *Science* **275**, 65 (1997).
55. N. G. Becker and K. Dietz, *Math. Biosci.* **127**, 207 (1995); N. G. Becker, A. Bahrampour, K. Dietz, *ibid.* **129**, 189 (1995).
56. B. T. Grenfell and A. P. Dobson, *Ecology of Infectious Diseases in Natural Populations* (Cambridge Univ. Press, Cambridge, 1995).
57. G. V. Bobashev, S. Ellner, D. W. Nychka, B. T. Grenfell, *Math. Biosci.*, in press.
58. S. Ellner, in (*49*).
59. P. Blanchard, G. F. Bolz, T. Kruger, *Mathematical Modelling on Random Graphs of the Spread of Sexually-Transmitted Diseases with Emphasis on HIV Infection* (Springer, Berlin, 1990).
60. M. Kretzschmar and M. Morris, *Math. Biosci.* **133**, 165 (1996); M. Morris and L. Dean, *Am. J. Epidemiol.* **140**, 217 (1994).
61. A. C. Ghani, J. Swinton, G. P. Garnett, *Sex. Transm. Dis.*, in press.
62. R. M. Anderson and G. P. Garnett, *Lancet* **348**, 1010 (1996).
63. R. M. Anderson *et al.*, *Nature* **382**, 779 (1996).
64. S. Nee, E. C. Holmes, R. M. May, P. H. Harvey, *Philos. Trans. R. Soc. London Ser. B* **344**, 77 (1994); P. H. Harvey, R. M. May, S. Nee, *Evolution* **48**, 523 (1994).
65. D. J. Austin and R. M. Anderson, *Parasitology* **113**, 157 (1996).
66. C. M. Lively and V. Apanius, in (*56*), pp. 421–449; C. A. Gilligan, *Phytopathology* **75**, 61 (1985).
67. B. T. Grenfell, K. Dietz, M. G. Roberts, in (*56*), pp. 362–383; B. T. Grenfell, K. Wilson, V. S. Isham, H. E. G. Boyd, K. Dietz, *Parasitology* **111**, S135 (1996).
68. S. Gupta *et al.*, *Nature Med.* **2**, 437 (1996).
69. C. Pease, *Theor. Popul. Biol.* **31**, 422 (1987); V. Andreasen, S. A. Levin, J. Lin, *Z. Angew. Math. Mech.* **76** (suppl. 2), 421 (1996); V. Andreasen, J. Lin, S. A. Levin, *J. Math. Biol.*, in press.
70. R. M. Anderson, R. M. May, S. Gupta, *Parasitology* **99** (suppl.), S59 (1989).
71. A. S. Perelson, in *Mathematical and Statistical Approaches to AIDS Epidemiology*, C. Castillo-Chavez, Ed., vol. 83 of *Lecture Notes in Biomathematics* (Springer-Verlag, New York, 1989), pp. 350–370; R. M. Anderson and R. M. May, in *Cell to Cell Signalling: From Experiments to Theoretical Models*, A. Goldbeter, Ed. (Academic Press, New York, 1989), pp. 335–349; A. R. McLean and M. A. Nowak, *J. Theor. Biol.* **155**, 69 (1992); A. S. Perelson, D. E. Kirschner, R. J. De Boer, *Math. Biosci.* **114**, 81 (1993); A. R. McLean, *Trends Microbiol.* **1**, 9 (1993); D. Schenzle, *Stat. Med.* **13**, 2067 (1994); P. Essunger and A. S. Perelson, *J. Theor. Biol.* **170**, 367 (1994); S. D. W. Frost and A. R. McLean, *J. Acquired Immune Defic. Syndr.* **7**, 236 (1994); H. J. Bremermann, *ibid.* **9**, 459 (1995); J. E. Mittler, B. R. Levin, R. Antia, *ibid.* **12**, 233 (1996).
72. M. A. Nowak, R. M. May, R. M. Anderson, *AIDS* **4**, 1095 (1990); M. A. Nowak *et al.*, *Science* **254**, 963 (1991).
73. S. M. Wolinsky *et al.*, *Science* **272**, 537 (1996); M. A. Nowak *et al.*, *ibid.* **274**, 1008 (1996); S. M. Wolinsky *et al.*, *ibid.*, p. 1010; R. J. De Boer and M. C. Boerlijst, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 544 (1994); N. I. Stilianakis, D. Schenzle, K. Dietz, *Math. Biosci.* **121**, 235 (1994).
74. M. A. Nowak *et al.*, *Nature* **375**, 606 (1995); M. A. Nowak and C. R. M. Bangham, *Science* **272**, 74 (1996); P. Klenerman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
75. M. A. Nowak, R. M. May, K. Sigmund, *J. Theor. Biol.* **175**, 325 (1995).
76. A. R. McLean and M. A. Nowak, *AIDS* **6**, 71 (1992); S. D. W. Frost and A. R. McLean, *ibid.* **8**, 323 (1994); A. R. McLean and S. D. W. Frost, *Rev. Med. Virol.* **5**, 141 (1995); D. Kirschner, *Notices Am. Math. Soc.* **43**, 191 (1996); D. Kirschner and G. F. Webb, *Bull. Math. Biol.* **58**, 367 (1996); M. D. de Jong *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5501 (1996); R. J. De Boer and C. A. B. Boucher, *Proc. R. Soc. London Ser. B* **263**, 899 (1996); M. A. Nowak *et al.*, *J. Theor. Biol.*, in press; N. I. Stilianakis *et al.*, *J. Virol.*, in press.
77. X. Wei *et al.*, *Nature* **373**, 117 (1995); D. D. Ho *et al.*, *ibid.*, p. 123.
78. A. S. Perelson *et al.*, *Science* **271**, 1582 (1996).
79. A. S. Perelson *et al.*, in preparation.
80. Y. Feng, C. C. Broder, P. E. Kennedy, E. A. Berger, *Science* **272**, 872 (1996); H. K. Deng *et al.*, *Nature* **381**, 661 (1996); T. Dragic *et al.*, *ibid.*, p. 667.
81. R. Liu *et al.*, *Cell* **86**, 367 (1996); M. Samson *et al.*, *Nature* **382**, 722 (1996); Y. Huang *et al.*, *Nature Med.* **2**, 1240 (1996).
82. L. Wu *et al.*, *Nature* **384**, 179 (1996); A. Trkola *et al.*, *ibid.*, p. 184.
83. D. E. Kirschner and A. S. Perelson, in *Mathematical Population Dynamics: Analysis of Heterogeneity and the Theory of Epidemics*, O. Arino, D. E. Axelrod, M. Kimmel, M. Langlais, Eds. (Wuerz, Winnipeg, Canada, 1995), pp. 295–310.
84. M. A. Nowak *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 4398 (1996); N. P. Lam *et al.*, *Hepatology*, in press.
85. B. R. Bloom, *Nature* **358**, 538 (1992); D. D. Richman, *Adv. Exp. Med. Biol.* **394**, 383 (1996); H. H.

van Es, E. Skamene, E. Schurr, *Clin. Invest. Med.* **16**, 285 (1993); M. L. Cohen, *Trends Microbiol.* **2**, 422 (1994).
86. A. T. Haase *et al.*, *Science* **274**, 985 (1996).
87. W. L. Walker and J. Cook, *Bull. Math. Biol.* **58**, 1047 (1996).
88. T. Saga *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8999 (1995).
89. T. B. Kepler and A. S. Perelson, *ibid.*, p. 8219; J. Jacob, J. Przylepa, C. Miller, K. J. Kelsoe, *J. Exp. Med.* **178**, 1293 (1993); P. E. Seiden and F. Celada, *Eur. J. Immunol.* **26**, 1350 (1996); Z. Agur, G. Mazor, I. Meilijson, *Proc. R. Soc. London Ser. B* **245**, 147 (1991).
90. T. B. Kepler and A. S. Perelson, *Immunol. Today* **14**, 412 (1993).
91. S. Han *et al.*, *J. Exp. Med.* **182**, 1635 (1995).
92. M. A. Fishman and A. S. Perelson, *J. Theor. Biol.* **170**, 25 (1994); B. F. Morel *et al.*, *Bull. Math. Biol.* **58**, 569 (1996); J. Caneiro, J. Stewart, A. Coutinho, *Int. Immunol.* **7**, 1265 (1995); A. Brass, R. K. Grencis, K. J. Else, *J. Theor. Biol.* **166**, 189 (1994).
93. M. Fishman and A. S. Perelson, *J. Theor. Biol.* **160**, 311 (1993); R. J. De Boer and A. S. Perelson, *ibid.* **169**, 375 (1994); S. J. Merrill, R. J. De Boer, A. S. Perelson, *Rocky Mountain J. Math.* **24**, 213 (1994).
94. R. J. De Boer and A. S. Perelson, *J. Theor. Biol.* **149**, 381 (1991); P. E. Seiden and F. Celada, *ibid.* **158**, 329 (1992).
95. J. Gleick, *Chaos: Making of a New Science* (Penguin, New York, 1987).
96. Figure 1 is a frame from the video, D. Deutschman and S. A. Levin, *SORTIE Simulations*, by Cornell and Princeton Universities; computation by L. Buttel; visualization by C. Devine, Cornell Theory Center, Cornell University.
97. L. A. Dugatkin and D. S. Wilson, *Am. Nat.* **138**, 687 (1991).
98. D. Schenzle, *IMA J. Math. Appl. Med. Biol.* **1**, 169 (1984).
99. B. Bolker, *ibid.* **10**, 83 (1993).
100. B. T. Grenfell, A. Kleczkowski, S. P. Ellner, B. M. Bolker, *Philos. Trans. R. Soc. London Ser. A* **348**, 515 (1994).
101. B. M. Bolker and B. T. Grenfell, *Philos. Trans. R. Soc. London Ser. B* **348**, 309 (1995); B. T. Grenfell, B. M. Bolker, A. Kleczkowski, *Proc. R. Soc. London Ser. B* **259**, 97 (1995).
102. N. M. Ferguson, D. J. Nokes, R. M. Anderson, *Math. Biosci.*, in press.
103. M. S. Bartlett, *J. R. Stat. Soc. Ser. A* **120**, 48 (1957); *ibid.* **123**, 37 (1960).
104. M. J. Keeling and D. A. Rand, unpublished data.
105. C. J. Rhodes and R. M. Anderson, *Nature* **381**, 600 (1996).
106. P. E. M. Fine and J. A. Clarkson, *Int. J. Epidemiol.* **12**, 332 (1983).
107. A. D. Cliff, P. Haggett, D. F. Stroup, E. Cheney, *Stat. Med.* **11**, 1409 (1992); B. T. Grenfell, A. Kleczkowski, C. A. Gilligan, B. M. Bolker, *Statistical Methods Med. Res.* **4**, 160 (1995).
108. B. T. Grenfell, A. Kleczkowski, S. P. Ellner, in *Forecasting and Chaos*, H. Tong, Ed. (World Scientific, Singapore, 1994), pp. 321–345.
109. We thank P. Kollman and the participants in the meeting, "Modeling of Biological Systems," which inspired this article, and NSF, which funded the workshop. Supported by NASA grant NAGW 4688 and the Andrew Mellon Foundation (S.L.), NSF grant DEB 9629236 (A.H.), the Wellcome Trust (B.G.), NIH grants RR06555 and AI28433 (A.S.P.), and the Jeanne M. Sullivan and Joseph P. Sullivan Foundation. D. Deutschman provided useful comments. Most of all, we thank A. Bordvik, who brought order to a chaotic sequence of drafts of this manuscript.

An Information-Intensive Approach to the Molecular Pharmacology of Cancer

John N. Weinstein,* Timothy G. Myers, Patrick M. O'Connor, Stephen H. Friend, Albert J. Fornace Jr., Kurt W. Kohn, Tito Fojo, Susan E. Bates, Lawrence V. Rubinstein, N. Leigh Anderson, John K. Buolamwini,† William W. van Osdol,‡ Anne P. Monks, Dominic A. Scudiero, Edward A. Sausville, Daniel W. Zaharevitz, Barry Bunow, Vellarkad N. Viswanadhan,§ George S. Johnson, Robert E. Wittes, Kenneth D. Paull

Since 1990, the National Cancer Institute (NCI) has screened more than 60,000 compounds against a panel of 60 human cancer cell lines. The 50-percent growth-inhibitory concentration (GI_{50}) for any single cell line is simply an index of cytotoxicity or cytostasis, but the patterns of 60 such GI_{50} values encode unexpectedly rich, detailed information on mechanisms of drug action and drug resistance. Each compound's pattern is like a fingerprint, essentially unique among the many billions of distinguishable possibilities. These activity patterns are being used in conjunction with molecular structural features of the tested agents to explore the NCI's database of more than 460,000 compounds, and they are providing insight into potential target molecules and modulators of activity in the 60 cell lines. For example, the information is being used to search for candidate anticancer drugs that are not dependent on intact p53 suppressor gene function for their activity. It remains to be seen how effective this information-intensive strategy will be at generating new clinically active agents.

Drug discovery is being transformed by new developments in molecular cell biology and the information sciences. A case in point is the drug discovery program conducted by the Developmental Therapeutics Program (DTP) of the NCI. Before 1985, the NCI used mice bearing murine leukemia P388 cells to screen new compounds for anticancer activity. That strategy identified agents active against leukemias but relatively few that were effective against solid tumors, including the most common human carcinomas. Hence, the NCI established a primary screen in which compounds are tested in vitro for their ability to inhibit growth of 60 different human cancer cell lines (*1*). Included are melanomas, leukemias, and cancers of breast, prostate, lung, colon, ovary, kidney, and central nervous system origin. A highly schematic view of this portion of the NCI drug discovery–development process is shown in Fig. 1. Compounds for testing have come principally from synthetic chemistry and natural product sources, but combinatorial libraries and products of biotechnology are also being screened.

This "disease-oriented" strategy for drug discovery was based on the hypothesis that selective activity in vitro against cancer cell lines from a particular organ would predict selective activity against corresponding tumors in humans. That concept is being tested as agents progress through clinical trials, and the answer is not yet clear. However, patterns of activity observed in the screen have proved predictive in an even more powerful way at the molecular level: They provide incisive information on the mechanisms of action of the compounds tested and on molecular targets and modulators of activity within the cancer cells. The cell lines are not fully representative of solid tumors in humans, but their patterns of pharmacological response are rich in information. We refer to this test system as a "screen," but it has also become a way to "profile" or "fingerprint" potential therapeutic agents.

The patterns of activity were first analyzed by the COMPARE algorithm (*2*). Given one compound as a "seed," COMPARE searches the database of screened agents for those most similar to the seed in their patterns of activity against the panel of 60 cell lines. Similarity in pattern often indicates similarity in mechanism of action, mode of resistance, and molecular structure (*2*). This form of analysis has been applied productively to topoisomerase II inhibitors (*3*), pyrimidine biosynthesis inhibitors (*4*), and tubulin-active compounds (*5*), among