COMMENTARY

Barriers to progress in systems biology

For the past half-century, biologists have been uncovering details of countless molecular events. Linking these data to dynamic models requires new software and data standards, argue **Marvin Cassman** and his colleagues.

The field of systems biology is lurching forwards, propelled by a mixture of faith, hope and even charity. But if it is to become a true discipline, several problems with core infrastructure (data and software) need to be addressed. In our view, they are too critical to be left to *ad hoc* developments by individual laboratories.

Systems biology has been defined in many ways, but has at its root the use of modelling and simulation, combined with experiment, to explore network behaviour in biological systems — in particular their dynamic nature. The need to integrate the profusion of molecular data into a systems approach has stimulated growth in this area over the past five or so years, as worldwide investments in the field have increased. However, this early enthusiasm will need to overcome several barriers to development.

A recent survey carried out by these authors — conducted by the World Technology Evaluation Center (WTEC) in Baltimore, Maryland, and funded by seven US agencies — compared the activities of system biologists in the United States, Europe and Japan[1]. The survey reveals that work on quantitative or predictive mathematical modelling that is truly integrated with experimentation is only just beginning. Progress is limited, therefore, and major contributions to biological understanding are few. The survey concludes that the absence of a suitable infrastructure for systems biology, particularly for data and software standardization, is a major impediment to further progress.

Come together

The WTEC survey confirmed that vital software is being developed at many locations worldwide. But these endeavours are highly localized, resulting in duplicated goals and approaches. Tellingly, one Japanese group called their software YAGNS, for 'yet another gene network simulator'. There are many reasons for this cottage industry: the need to accommodate local data; the requirements of collaborators to visualize data; and limited knowledge of what is already available. In general, however, it is a terrible waste of time, money and effort. Most software remains inaccessible to external users, even when the developers are willing to release it, because supporting documentation is so poor.

For software developers and skilled users these problems are not insurmountable. But sharing of the benefits of systems biology more widely will occur only when working biologists, who are not themselves trained to develop and modify such software, can manipulate and use these techniques. Unfortunately, the translation of systems biology into a broader approach is complicated by the innumeracy of many biologists. Some modicum of mathematical training will be required, reversing the trend of the past 30 years, during which biology has become a discipline for people who want to do science without learning mathematics.

> "During the past 30 years biology has become a discipline for people who want to do science without learning mathematics."

A reasonable set of expectations is that different pieces of shared software should work together seamlessly, be transparent to the user, and be sufficiently documented so that they can be modified to suit different circumstances. Funding agencies would be unwise to support software development without also investing in the infrastructure needed to preserve and enhance the results. One way to do this would be to create a central organization that would serve both as a software repository and as a mechanism for validating and documenting each program, including standardizing of the data input/output formats.

As with centralized databases, having a shared resource with appropriate software-engineering standards should encourage users to reconfigure the most useful tools for increasingly sophisticated analysis. A group sponsored by the US Defense Advanced Research Projects Agency, and involving one of us (M.C.), has developed a proposal for such a resource[2]. This repository would serve as a central coordinator to help develop uniform standards, to direct users to appropriate online resources, and to identify — through user feedback — problems with the software. The repository should be organized through consultation with the community, and will require the support of an international consortium of funding agencies.

Diverse data

The problems with software diversity are mirrored by the diversity of ways that data are collected, annotated and stored. Such issues are even worse than those faced by the DNA-sequencing community, because experimental data in systems biology is highly context dependent. For data to be useful outside the laboratory in which they were generated, they must be standardized, presented using a uniform and systematic vocabulary, and annotated so that the specific cell type, growing conditions and measurements made — from metabolite- and messenger-RNA-profiling to kinetics and thermodynamics — are reproducible.

Easy access to data and software is not a luxury, it is essential when results undergo peer review and publication. For the scientific community to evaluate the increasingly complex data types, the increasingly sophisticated analysis tools, and the increasingly incomplete papers (that cannot include all information because of the very complexity of the experiments and tools), it is vital that it has access to the source data and methods used.

Dealing with these complex infrastructure issues will require a focused effort by researchers and funding agencies. We propose that the annual International Conferences on Systems Biology would be an appropriate venue for initial discussions. Whatever the occasion, it must be done soon. ■

Marvin Cassman lives in San Francisco, California, USA.
Co-authors are Adam Arkin of the Bioengineering Department, University of California, Berkeley; Fumiaki Katagiri of the Department of Plant Biology, University of Minnesota, St Paul; Douglas Lauffenburger of the Biological Engineering Division, Massachusetts Institute of Technology, Cambridge; Frank J. Doyle III of the Department of Chemical Engineering, University of California, Santa Barbara; and Cynthia L. Stokes who is at Entelos, Foster City, California.

1. Cassman, M. *et al. Assessment of International Research and Development in Systems Biology* (Springer, in the press) www.wtec.org/sysbio
2. Cassman, M., Sztipanovits, J., Lincoln, P. & Shastry, S. S. *Proposal for a Software Infrastructure in Systems Biology* www.csl.sri.com/users/lincoln/SystemsBiology/SI.doc

©2005 Nature Publishing Group