PERSPECTIVES

ESSAY

Back to the future: education for systems-level biologists

Ned Wingreen and David Botstein

Abstract | We describe a graduate course in quantitative biology that is based on original path-breaking papers in diverse areas of biology; each of these papers depends on quantitative reasoning and theory as well as experiment. Close reading and discussion of these papers allows students with backgrounds in physics, computational sciences or biology to learn essential ideas and to communicate in the languages of disciplines other than their own.

The genome has presented biologists with an opportunity to study genetic processes on a genomic scale, and to achieve quantitative understanding, not just of individual molecular mechanisms but also of their interactions and regulation at the systems level. The transformation of biology into a fully quantitative, theory-rich science now seems inevitable, if not yet quite within reach.

These developments have produced a challenge to the educational system. It is clear that the future of biology will require combinations of skills that are rarely found in individual scientists today. The existing educational system teaches biologists very few mathematical or computational skills, and gives scientists with backgrounds in physics and informatics comparably limited exposure to even the most basic biological phenomena and principles. The problem begins early in undergraduate education, and by the doctoral level there are severe interdisciplinary communication difficulties that are encountered by even the most motivated of collaborators[1].

An integrated science curriculum

At Princeton, we have begun to address the educational challenges at both the undergraduate and the graduate levels. Our goal, at each level, is to equip students to succeed across scientific disciplines using the language and mathematics appropriate to each. To this end, we are teaching courses that integrate subjects that were traditionally taught separately, such as mathematics, physics, chemistry and biology. We are finding that, at the undergraduate level, integration works best when it is initiated in the first year of college. An integrated science curriculum, taught at the level of the most challenging physics, computation, chemistry, and biology courses, serves to bring the students to a level of sophistication that allows them to major in any of the sciences.

At the graduate level the problem is more difficult as the students have already differentiated to some extent and view themselves as, for example, physicists, chemists or biologists. We find that first-year graduate students already have some lacunae in their education: the biologists have had limited education in physics, computation and mathematics, and the physical and computational scientists have little or no biology at their command. The temptation is to prescribe remedial undergraduate courses; however, this approach does not appeal to graduate students. Our alternative is to begin with an integrated graduate course in quantitative biology.

We teach a single course to a population that consists of approximately equal numbers of graduate students in biology and physics, mostly in their first or second years. These students are interested in the interface between biology and quantitative science (FIG. 1). We meet with the students together, in a seminar format, having assigned two papers — often classics — that use sophisticated quantitative methods and concepts to study biological problems. We discuss the papers in detail, but not necessarily in the original order of presentation, for almost three hours. We use, with some

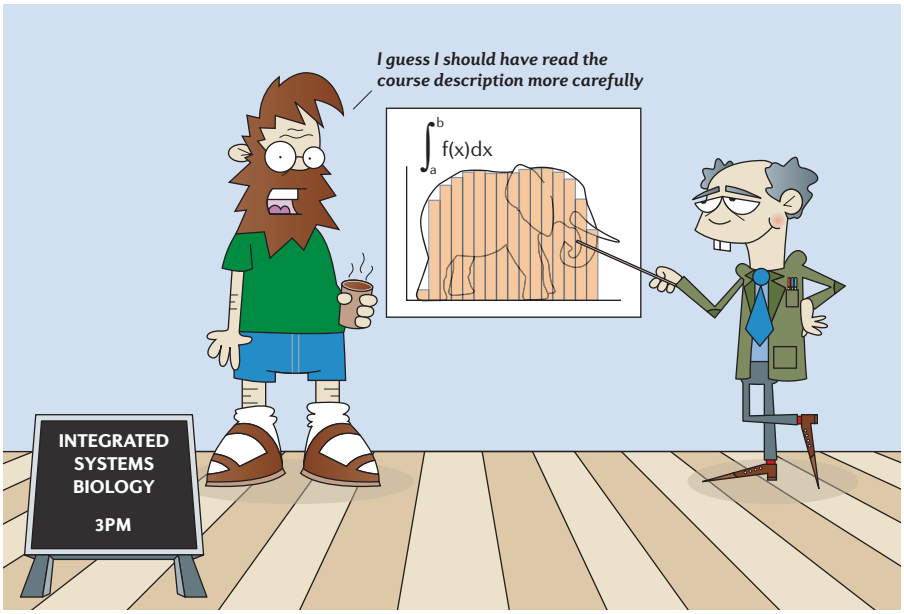

Figure 1 | **Taking integrated systems biology a step too far.**

Box 1 | ***Diauxie* — understanding a paradigm in gene regulation**

This figure, from Jacques Monod's Ph.D. dissertation (*La Croissance des Cultures Bactériennes*, 1942), illustrates a basic phenomenon that led directly to the discovery of gene regulation by repressors in bacteria, and to a Nobel Prize. This figure (or something similar to it, usually with lactose instead of galactose as the second sugar) is in virtually every textbook of genetics and molecular or cell biology.

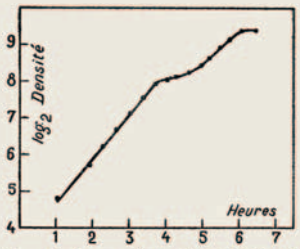

FIG. 46. — Croissance d'une culture de *B. coli* dans un mélange glucose-galactose. (Coordonnées semi-logarithmiques).

However, understanding of this phenomenon is not possible without a quantitative analysis, which is introduced in one of the papers read in our course[7]. The persistent failure of the galactose (or lactose) operon to be induced in the presence of inducer depends crucially on the absence of permease molecules. Novick and Weiner[7] clearly showed that *diauxie* works because induction is a positive-feedback system in which the inducer is taken up from the medium by a permease molecule that is itself induced. It is this feature that makes the relatively weak phenomenon of catabolite repression into an absolute barrier to induction. Because this phenomenon illustrates so nicely the importance of feedback and memory, it is still studied today by systems-level biologists (see REF 14).

success, the skills of one group of students to help the other group; when biological issues arise, the biologists tend to speak, and when quantitative issues are on the table, the physicists tend to participate most actively. Each group develops increasing respect for the value of listening to the other, and we are optimistic that the students become motivated to help each other.

Course materials

The nature and quality of the papers is crucial to the success of this course. The papers need to be sufficiently sophisticated and important to repay detailed study. The papers also have to function as vehicles for teaching both biology and quantitative analysis. We have found, somewhat to our surprise, that some of the best papers are quite old, possibly harking back to a time when biologists' and physical scientists' educations were less different than they are today. We have also found that the lessons in these old papers are being re-emphasized in the current systems biology literature (BOX 1).

In this section, we present nearly a dozen of the pedagogically most successful papers. We also provide a few words on the lessons we find embodied in them.

Random processes and distributions. Luria, S.E. and Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance (1943)[2].

This classic paper, which was published more than 50 years ago, first presented the appropriate mathematics for analysing random mutations as they arise in a population. This mathematical analysis can be applied to a class of problems in biology that, in subsequent years, has been shown to also involve statistical physics and chemistry. The analysis begins with an understanding of Poisson processes and the Poisson distribution (see Glossary).

> We have found, somewhat to our surprise, that some of the best papers are quite old, possibly harking back to a time when biologists' and physical scientists' educations were less different than they are today.

Luria and Delbrück addressed the specific problem of mutations arising in a growing colony of *Escherichia coli* cells derived from a single phage-sensitive ancestor. To analyse their data they had to consider the distribution of cells that have mutated to resistance to bacteriophage attack. The Luria–Delbrück distribution has a remarkably long 'tail', due to the occasional 'jackpot' in which a resistance-endowing mutation occurred at an early generation in the history of the clone. It was, indeed, the use of a slot machine that stimulated Luria to think about mutation in this way, and the entirely appropriate term jackpot was used by him.

As well as teaching the students to think about random processes in populations (for homework, see BOX 2), this paper introduces a considerable amount of basic microbiology. A good example lies in the primary motivation for the study. Does resistance to phage arise from random mutations, or from the development of immunity after exposure? We ask the students to discuss questions such as: what if, instead of studying the lytic phage T1, Luria and Delbrück had studied the lysogenic phage λ, which can integrate into the genome and provide immunity?

Individuality. Elowitz, M.B. *et al.* Stochastic gene expression in a single cell (2002)[3].

This is a very good paper for discussing the individuality of genetically identical cells, which is rapidly becoming the standard paradigm for studying stochastic gene expression in single cells. Simultaneous measurement of two 'identically' regulated fluorescent proteins in single cells distinguishes intrinsic, biochemical noise from extrinsic, cell-to-cell variation. Of course, what gets classed as extrinsic and intrinsic depends on how the two reporters are constructed; for example, if the reporters are on opposite sides of the chromosome from the origin of replication, what happens in strains in which replication can stall?

This paper is useful for discussing how noise propagates from DNA copies, to mRNA copies, to number of proteins. More generally, it teaches the importance of single-cell measurements; what does one fail to notice in population-averaged measurements?

Stable switching. Novick, A. and Wiener, M. Enzyme induction as an all-or-none phenomenon (1957)[4].

This is another classic paper, not nearly so well known today as the Luria and Delbrück paper, but equally influential in its time. It provides a complete and convincing analysis of how genetically identical cells can stably maintain two heritable states. The *lac* operon, as an experimental system, remains the premier model not only for classic gene-expression regulation, but also for systems-level analysis (BOX 1).

Over a range of lactose concentrations, the *lac* system supports two stable states: induced and uninduced. The paper's most interesting lesson is that the history of the cells matters. Another basic biology (as well as mathematical) lesson is that the presence or absence of permease molecules determines the nature of *diauxic* growth — that is, permease molecules are the agents of a cell's 'memory' in the *lac* system. The comparison of growth rates of induced and uninduced cells foreshadows the recent interest in quantifying fitness (see also REF. 5).

Robustness. Barkai, N. and Leibler, S. Robustness in simple biochemical networks (1997)[6].

An elegant example of how 'robustness' — preservation of function despite fluctuations in components — can be built-in at the network level. Individual *E. coli* cells adapt precisely to ambient levels of chemoattractants, in some cases over a range of five orders of magnitude, despite fluctuations of network components.

Barkai and Leibler proposed a model in which active and inactive receptors are distinguished by different net rates of methylation. To achieve steady state, the average methylation and demethylation rates must come into balance, which only happens when receptors reach a particular average activity, therefore ensuring the precise adaptation of receptor activity. The wealth of information about the chemotaxis network makes this system the model for understanding signal transduction at a quantitative level.

Ultrasensitivity. Goldbeter, A. and Koshland, D.E.Jr. An amplified sensitivity arising from covalent modification in biological systems (1981)[7].

Chemical modification of proteins by enzymatic cycles (for example, phosphorylation and dephosphorylation) lies at the heart of many cellular signalling systems. Goldbeter and Koshland analysed the input–output characteristics of such cycles. How do changes in enzyme levels affect the steady-state level of a modified protein? Increasing enzyme saturation results in increasing ultrasensitivity — that is, a sharper, graphically more-step-like response. The paper segues smoothly between mathematics and biology, and provides critical insights into our current understanding of developmental switches in cells.

Specificity. Hopfield, J.J. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity (1974)[8].

How do cells achieve specificity? In a classic paper that is more relevant today than ever, Hopfield showed that a simple, kinetic (energy-driven) mechanism can multiplicatively enhance the intrinsic specificity of enzyme–substrate reactions. Examples of kinetic proofreading are taken from transcription and translation, and the mechanism beautifully explains the seemingly wasteful, extra energy inputs to these processes. This paper teaches that when studying biological processes, one should ask; what might go wrong and at what rate would such errors be expected? When there is a mismatch between expectation and observation, as there was in the fidelity of protein synthesis in the decade before Hopfield's paper, then new biological processes might be waiting to be discovered.

> "When there is a mismatch between expectation and observation, ... then new biological processes might be waiting to be discovered."

Similarity. Smith, T.F. and Waterman, M.S. Identification of common molecular subsequences (1981)[9].

Smith and Waterman tackled the problem of local sequence alignment. This seminal paper provides students with a good introduction to the problem, and an appreciation for the importance of how different algorithms scale with problem size. How does the use of dynamic programming bypass the (exponentially difficult) task of comparing every possible alignment of two sequences? This study also provides a springboard for discussing what one means by similarity, and how homology (evolution from a common ancestor) influences our understanding of the meaning of statistically unlikely similarities between sequences.

Maximum likelihood. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach (1981)[10].

How do we distinguish between different models given a particular set of data? Felsenstein presents a simple and clearly written example using alternative unrooted phylogenetic trees as models and DNA sequences from extant species as data. Given a simple assumption for mutation rates, which tree is most likely to lead to the observed data? How confident can we be in the maximum likelihood solution? This article is a good lead-in to a discussion of Bayesian analysis and its wide application to large, and potentially mixed, data sets in this era of genomics and systems-level biology.

Evolutionary perspective. Eisen, J.A. A phylogenomic study of the MutS family of proteins (1998)[11].

Another important aspect of the course is to introduce the idea of phylogenomics — knowledge of genomes is vital to determining the evolutionary tree, and knowledge of the evolutionary tree is a powerful guide to interpreting genomes.

Box 2 | **A representative homework problem: 'Beyond grad school!'**

Problem. You're the technical advisor for a group of investors that are interested in buying a company that engineers bacteria to detoxify waste. An important issue is that the engineered bacteria are slightly less fit than the wild-type in the wild (where they will be used) but not in the laboratory (where they will be grown up and packaged for sale dispersal). The engineered bacteria revert (mutate) to wild-type in the laboratory at a rate μ — that is, at each cell division there is a probability μ that one daughter will become wild-type, and then that all of the daughter's progeny remain wild-type. The company has performed a set of 10 experiments — each starts with a single engineered cell, which grows for 20 generations, and then measures the number of wild-type (revertant) cells. In these 10 trials, they found 0, 0, 8, 14, 16, 24, 38, 44, 70 and 80 revertant cells. The company statisticians argue that the probability of getting 0 revertants is given by the probability of the 0 outcome in a Poisson process, namely $p(0) = e^{-\lambda}$, in which the mean is given by the expected number of mutation events in a colony, that is, $\lambda = \mu \times 2^{20}$. So, taking the probability of 0 revertants from the data to be 2/10, they estimate that $\lambda = 1.61$, which corresponds to $\mu = 1.5 \times 10^{-6}$.

The investors have asked you to verify this estimate for the reversion rate. They are only interested in buying the company if the rate is below 2×10^{-6}. Your intuition is that you can get a more reliable estimate for the rate using the entire data set rather than just the fraction of trials with 0 revertants.

Use the maximum likelihood approach to find the mutation rate μ for which the observed data has the highest likelihood (you will need to generate accurate distributions of revertants for different values of μ, calculate the likelihood of the observed data for each μ, and find the μ that maximizes this likelihood). Should the investors buy the company?

Authors note. The intention of this problem was to have the students do an exhaustive simulation, a technique not available to Delbrück in 1943. However, when this problem was given, several of the students realized that in addition to the possibility of simulating the full Luria–Delbrück distribution, it is possible to analytically calculate the distribution for the low numbers used here. We thought this was a great success.

Glossary

Croonian Lecture
The Croonian Lectures are prestigious lectureships given at the invitation of the Royal Society and the Royal College of Physicians.

***Diauxic* growth**
In a medium that contains glucose and a less preferred carbon source, bacteria exhaust the glucose before consuming the other carbon source. Monod called this behaviour *diauxie* ('double growth' in French).

Hodgkin–Huxley equations
Hodgkin and Huxley published a series of (now classic) papers in 1952 on electrical activity and transmembrane ion currents in the squid giant axon. In these papers, they derived the Hodgkin–Huxley equations, which accurately describe the action potential.

Jackpot
In a series of parallel cultures, the Luria–Delbrück 'jackpot' is the rare observation of a large clone of mutant cells that are derived from a single mutational event early in the growth of the culture. Though rare, these jackpots occur much more frequently than the probability $p(n)$ that is expected from simple Poisson statistics.

Poisson distribution
This distribution, $p(n) = \exp(-\lambda)\,\lambda^n/n!$, gives the probability of observing n rare random events in a very large population, for which λ is the average expected number of these rare events.

Eisen presents a particularly good example: the MutS family of DNA-repair proteins that is found in virtually all organisms. The function of MutS has changed in several ways along different lineages, which becomes readily apparent only when studying the evolutionary tree.

The paper provides a good introduction to the diversity of evolutionary events that must be considered, including gene duplications, gene deletions, horizontal transfer and, of course, speciation. More generally, it shows the value of taking an evolutionary perspective in interpreting biological function.

Microarray-data analysis. Eisen, M.B. *et al.* Cluster analysis and display of genome-wide expression patterns (1998)[12].

The analysis of microarray data using correlation coefficients between genes and hierarchical clustering has stood the test of time, compared to more complex approaches. This paper is also a good introduction to systems-level thinking about cellular dynamics. It is apparent from the data that genes don't turn on and off one at a time, rather, cohorts of genes turn on and off together. We stress to the students the importance of data visualization. The similarity of gene-expression patterns can be seen at a glance when data are presented as coloured patterns. We ask the students to imagine the same data presented as the table of numbers that underlie these coloured patterns. We also contrast this kind of analysis with methods that require removing, summarizing or otherwise changing the underlying data.

Biophysical modelling. Hodgkin, A.L. Croonian Lecture, ionic movements and electrical activity in giant nerve fibres (1958)[13].

An important theme in quantitative biology is biophysical modelling. This lecture presents a beautiful example of biophysical modelling that is the foundation of neurobiology. The story of Hodgkin and Huxley's study of giant nerve fibres runs step-by-step from the basic biology and physiology of axons and action potentials, through the crucial electrophysiology experiments, to the now famous Hodgkin–Huxley equations and their solution.

The students learn something about electrical engineering, ion channels, chemical potentials and differential equations, and a lot about reasoning from experiment. An alternative to this Croonian Lecture, which we have considered but not tried, might be to use one or more of the original Hodgkin and Huxley publications, as we think that papers presenting original data are probably more effective in such a course than reviews.

Conclusions

To summarize our experience, we have found that close reading of papers that made a significant contribution to basic biology through the use of quantitative reasoning and theory is useful in the setting of first-year graduate students with diverse educational backgrounds. The papers help those students with a biology background to focus on the mathematical and computational ideas and methods that are most relevant not only to current practice, but also to the students' areas of interest.

In a complementary way, the students with physics and computational backgrounds are exposed to many issues of significance in basic biology. They also become familiar with research areas to which they might well be able to contribute, if they continue to pursue their interests in biology. Last, we are hopeful that the emerging practice of teaching students with such diverse backgrounds together will facilitate communication and professional relationships that will serve these future members of the genomic and systems biology community well.

Ned Wingreen is at the Department of Molecular Biology, and David Botstein is at the Lewis-Sigler Institute and the Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA.
Correspondence to D.B.
e-mail: botstein@princeton.edu

doi:10.1038/nrm2023
Published online 20 September 2006

1. Bialek, W. & Botstein, D. Introductory science and mathematics education for 21st century biologists. *Science* **303**, 788 –790 (2004).
2. Luria, S. E. & Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491–511 (1943).
3. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
4. Novick, A. & Wiener, M. Enzyme Induction as an all-or-none phenomenon. *Proc. Natl Acad. Sci. USA* **43**, 553–566 (1957).
5. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–592 (2005).
6. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
7. Goldbeter, A. & Koshland, D. E. Jr. An amplified sensitivity arising from covalent modification in biological systems. *Proc. Natl Acad. Sci. USA* **78**, 6840–6844 (1981).
8. Hopfield, J. J. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl Acad. Sci. USA* **71**, 4135–4139 (1974).
9. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
10. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
11. Eisen, J. A. A phylogenomic study of the MutS family of proteins. *Nucleic Acids Res.* **26**, 4291–4300 (1998).
12. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
13. Hodgkin, A. L., Croonian Lecture, ionic movements and electrical activity in giant nerve fibres. *Proc. R. Soc. Lond. B. Biol. Sci.* **148**, 1–37 (1958).
14. Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I. & Van Oudenaarden A. Multistability in the lactose utilization network of *Escherichia coli*. *Nature* **427**, 737–740 (2004).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Ned Wingreens's homepage: http://www.molbio.princeton.edu/research_facultymember.php?id=65
David Botstein's homepage: http://www.molbio.princeton.edu/research_facultymember.php?id=60
Access to this links box is available online.