Opinion

Whither genomics?

Andrew W Murray

Address: Department of Physiology, University of California San Francisco, San Francisco, CA 94143, USA.
From 15 July, Molecular and Cellular Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA.
E-mail: amurray@mcb.harvard.edu

Published: 9 June 2000

Genome ***Biology*** 2000, **1(1)**:comment003.1–003.6

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/1/comment/003

Abstract

The flood of data from genome-wide analysis is transforming biology. We need to develop new, interdisciplinary approaches to convert these data into information about the components and structures of individual biological pathways and to use the resulting information to yield knowledge about general principles that explain the functions and evolution of life.

What is genomics, where is it heading and how will it affect biology? Like the genetics it sprang from, genomics is both a science and a widely applicable tool. Just as genetics is both the science of how cells express and transmit the information in their DNA and a tool for studying any biological process, genomics is both the science of understanding the structure and evolution of genomes and a tool for learning about the functions of the genes therein. Genetics and genomics differ in scale and focus. Genetics uses mutants to identify and study the few genes that control a particular phenotype, whereas genomics collects data on all the genes in an organism. I have taken genomics to include all methods that collect and analyze comprehensive data about genes, including the sequence and abundance of nucleic acids and the properties of the proteins they encode (often called proteomics). In this article, I consider how genomics differs from previous approaches to biology, the type of information that current approaches provide, how it may influence biology, and the prospects for technical and intellectual improvements in gathering and exploiting genome-wide data.

Biology's industrial revolution

At the end of the 18th century, water power mechanized spinning and weaving in Britain, moving the textile industry from cottages to Blake's 'satanic mills'. Inventing rapid methods to clone and sequence DNA at the end of the 20th century has revolutionized biology. To begin with, these inventions simply sped up biology's traditional approach, the isolated study of a particular problem. But genomics moved sequencing from the cottage to factories, by determining the complete sequence of an organism's genome and developing sequence-based tools to follow the behavior of genes and ultimately the proteins they encode. Like everything new, genomics elates some and enrages others. One of its happiest consequences should be encouraging a shift to 'modular' biology, as I have advocated with Hartwell *et al.* [1]. Functional modules are biological pathways: collections of molecules, both large and small, that co-operate to perform a given function, such as protein synthesis, signal transduction, or the biosynthesis of small molecules. We want to know what the properties of a module are (such as the detailed, quantitative correlation between its inputs and outputs), how its parts are chemically and structurally connected to each other, how these connections explain the properties, and how different modules are connected to or insulated from each other. An important step in this analysis is listing the module's parts, a requirement that is currently satisfied by a laborious cottage industry that produces many dutiful, but intellectually dull papers that are entitled 'Protein X is involved in process Y' but shed little light into the biological or chemical role of X. If genomics can rapidly list the proteins in a given module, it will have done biology a major service by returning our attention to understanding

how biological pathways work. We should rejoice that the prodigious flow of data from genomics must draw into our circle mathematicians, physicists, computer scientists and others who have thought hard about how to analyze enormous quantities of sparsely connected data. Just as the refugees from physics gave molecular biology much of its intellectual excitement and altered the way we thought about biology, this new influx should do the same for genomics.

Genomics will accelerate the migration of biologists to the 'superb six': humans, mice, fruitflies, worms, yeast, and *Arabidopsis*. This shift towards a group of currently fashionable organisms is a self-reinforcing process, because the creatures with sequenced genomes and community-wide resources will tear biologists away from less fashionable organisms. The resulting concentration will be good and bad. Working on a few organisms means faster progress on many conserved features of biology, as witnessed by budding yeast's remarkable contribution to understanding fundamental aspects of cell biology over the last 15 years. But the rush to model eukaryotes threatens to extinguish work on a variety of organisms that have made important historical contributions to biology (the marine invertebrates), are rich reservoirs of interesting but poorly studied biological phenomena (epigenetic phenomena in protozoa), or are marvelous examples of phenomena that are harder to study in more complex creatures (cell-type specification in *Volvox*). The larger its genome and the fewer its students, the more likely work on an organism is to die.

In contrast to its winnowing effect on eukaryotes, genomics has stimulated work on the diversity of prokaryotes. Their small genomes make them easy to sequence, and the organization and diversity this endeavor has revealed has had a salutary effect on those who thought bacteria unsophisticated cells that were all alike. Together, bacterial genomics, the rise of bacterial cell biology, and a renewed interest in infectious diseases should lead to a renaissance in bacteriology. If sequencing and sequence-based resources get cheap enough, zoology and botany might enjoy the same sort of resurgence.

Genomics as an integrative science

Genomics should help reunite the reductionist and integrative branches of biology. By looking at all the genes or proteins at once, genomics encourages its practitioners to step back and consider the overall physiology of cells and organisms and reminds us that cell biology is in essence molecular physiology. For example, analyzing mRNA levels in mammalian cells treated with fibroblast-derived growth factor, normally used as a paradigm for studying the control of cell proliferation, also revealed a variety of responses that are connected with wound healing rather than proliferation [2]. The intimate links between genomics and evolutionary biology should strengthen connections between those who study biology at very different timescales and organizational levels. One can even imagine doing ecological experiments by using organism arrays, in which each organism was represented by a unique DNA sequence and the relative abundance of different organisms would be determined by harvesting samples, preparing RNA and hybridizing it to the array.

How will the post-sequencing phase of genomics affect the balance between small and big science? Small science will adapt and prosper. Bright students in small labs are already seizing on genomics as a way to devise clever, new approaches to old and difficult problems. I believe the challenge lies in picking and executing the right big projects to produce comprehensive genome-wide datasets. How should we decide which post-sequencing projects are most important, who should carry them out, and how do we make sure that the data become rapidly and freely available? For example, in budding yeast the technology exists to determine the effect of every viable gene deletion on the level of the mRNA expressed by every gene [3], and to determine all the pairs of gene deletions that are synthetically lethal with each other [4]. Both projects would produce tens of millions of data points, cost less than $5,000,000, and will have enormous value both for those working on specific modules and for those concerned with general issues about how modules are connected to and insulated from each other.

To make sure that we use people and money well, biologists and their patrons need to vigorously debate where genomics is going. What sorts of data should we collect, and how accurate should we be? Should we focus on building the most sophisticated tools and databases for a few organisms, or spread our resources more broadly? Should our primary motivation be intellectual curiosity, profit, or prolonging human lives, and how honest should we be about the answer? Will genomics in academia and industry co-exist by encouraging collaborations that produce publicly accessible data or by lowering the cone of secrecy over corners of academic labs? How do we deal with the competition between academic and commercial labs to generate genome-based information that is of equal interest to shareholders and academics (assuming that these are separable entities!)? Unless we participate in the discussion, we are likely to find ourselves unhappy with the answers that evolve.

Hypothesis-driven versus exploratory research

Just as the industrial revolution drastically lowered the price of textiles, so the genomic one is lowering the price of data, changing biology from a data-limited to an analysis-limited science. The raw product of genomics is data. At the lowest level, these data are like the contents of a phone book: they allow us to look up details about an individual gene or protein. The next level of abstraction is turning data into useful information, for example by using patterns of mRNA abundance to infer that a group of genes are involved in a common process [5]. Finally there is knowledge, the refining

of information into an understanding of biological processes that allows us to make useful predictions, and more importantly satisfies our curiosity about how cells and organisms work and where they came from.

When data were hard to get, collecting the right data was critical. Thus funding decisions (especially those of the National Institutes of Health in the USA) have traditionally favored hypothesis-driven (I think I know what will happen if...) over exploratory (what happens if...?) research. By making data so easy to collect, genomics encourages exploratory research, but increases the need for rigor in the design, execution, and interpretation of experiments. As microarray analysis becomes widely available, more and more researchers are testing what their favorite perturbation does to gene expression. Seeing the biological equivalent of a Rorschach blot - collections of genes that go up or down - they ask themselves and others what it all means. In many cases, especially in human cells, the design of the experiment precludes a clear answer, but all too often authors argue that the changes they see are the cause rather than the effect of the phenomena they study, and a paper appears because the technology is novel and genomics is fashionable.

An alternative form of exploration is finding all the genes that are linked by some form of genomic information and studying the effect of perturbing them. If genomic information argues that the genes function in a common pathway, this work is no more or less hypothesis-driven than a geneticist's decision to isolate and study all the genes that can be mutated to yield a given phenotype, and the two approaches have pleasingly complementary strengths and weaknesses. An approach that has no parallel with genetics is studying all the members of a class of proteins, such as motors that move along microtubules, or in the extreme case, every gene in a genome. The utility of such approaches will depend on the quality and detail of the analysis, and, particularly for whole-genome investigations, on developing methods that turn data into information and information into knowledge.

Finally, there are analyses that explore a specific hypothesis. A beautiful example is Futcher's analysis of genes transcribed early in mitosis of the budding yeast cell cycle (Bruce Futcher, personal communication). Because the factor stimulating their transcription had not been found by genetic analysis, he argued that there were probably two related proteins and that since either could activate mitotic transcription, a mutation in only one gene would have little effect. Because the mRNAs encoding the transcription factors that drove other groups of cell-cycle-regulated genes rose and fell with the genes they controlled, he argued that the missing pair of transcription factors must rise and fall with the other early mitotic genes. Applying these two filters to cell-cycle-regulated gene expression, he identified a single candidate pair of genes. Gratifyingly, inactivating both genes did indeed eliminate the cell-cycle fluctuation of the early mitotic genes. This example illustrates the synergy between exploratory (collecting data on cell-cycle-regulated gene expression) and hypothesis-driven research, and the fallacy of arguing too strongly for one or the other.

'Hard data is good to find'

The value of any genomic approach depends on the quality of the data collected and the intelligence used to analyze them. Such an obvious statement is required because the novelty of being able to look at all genes simultaneously appears to have led to a certain amount of amnesia about the value of repeated measurements, error bars, and other forms of statistical analysis. For example, if you look at the ratio of gene expression between two different repetitions of the same experiment in an organism with 10,000 genes, and the error in the measurements is normally distributed, 500 of the ratios will be more than two standard deviations from the mean value (hopefully, 1). Thus, if the standard deviation is 0.5, many genes will show more than two fold changes in gene expression. A good example of what careful analysis can provide is the clustering of genes into groups whose expression changes in correlated ways in response to a wide range of genetic, physiological, and pharmacological perturbations, and the converse clustering of perturbations that have similar effects on gene expression. A striking instance is the recognition, in some experiments, of genes whose response was correlated with their location on the same chromosome, revealing that the studied strain carried an unsuspected extra copy of that chromosome [6].

Genomics increases the chance that biology will experience a split like the one in physics, between those who collect and those who analyze data. This will challenge the majority of biologists who believe that modeling, simulation, and theory have little to contribute to biology. This prejudice rests on insecurity engendered by most biologists' weakness in mathematics (including my own) and previous efforts to model systems using more variables than there were data points. If we keep clinging to this prejudice, we will drown in a sea of data.

Genomics as a classification tool

Even in budding yeast, the paragon of genetic analysis, we know surprisingly little about the function of most genes. How can genomics help us? One way is to ask which genes help cells survive and proliferate under various conditions. By using integrated oligonucleotides (bar codes) to uniquely mark each strain that carries a deleted gene and microarrays to follow the abundance of the different bar codes as cells proliferate, experiments can be performed on pools of strains, so that the role of all the non-essential genes can be tested at the same time [4]. In principle, this approach can be extended to essential genes by using the combination of inducible promoters and inducible protein degradation [7,8] to control levels of gene expression. Extending this approach

beyond micro-organisms will be difficult and expensive. An attractive alternative is to use computational tools to group genes that function in particular pathways. Two tools depend solely on the sequence of multiple genomes. One groups genes that appear and disappear together during evolution [9], whereas the other groups proteins that are encoded by separate genes in one organism, but fused into a single gene in another [10,11]. Genes whose expression patterns cluster with each other in transcriptional profiles also have a high probability of functioning in the same module [5].

Combining all these approaches produces robust clues that gene X functions in process Y [12]. Can genomics go further and tell us what gene X is doing in that process? So far, no. But if this limitation can be overcome, genomics will truly revolutionize biology. We would be much better off if multiple genome sequences produced accurate predictions of protein structure (see the Comment by Petsko [13] in this issue on the subject of what accuracy means in this context). The sequences of the same protein in different organisms must encode information about which parts of the primary sequence touch each other when the protein folds up, much as phylogenetic comparisons of RNA sequences reveal information about secondary structure. Although we know how to find conserved base-pairs in nucleic acids, we do not know how to decode the information encoded in multiple related protein sequences. If we did, combining this constraint with the restriction that all the related protein sequences must have very similar tertiary structures should vastly improve the accuracy of structural predictions. Since protein structure and function are more strongly conserved than amino-acid sequence, we could detect evolutionary and functional relationships that had been obscured by sequence changes during evolution. If genome-based structure prediction were successful, it could be extended to predict which proteins interacted with each other and to reveal the structure of the interfaces. Exciting as this possibility is, we would still need other information to define a protein's function in the module that contains it.

The future: nucleomics versus proteomics

The base-pairing of nucleic acids makes genomics as well as life possible. Sequence information is sufficient for what might be called nucleomics, generating and using the nucleic acids needed for expression profiling, constructing a wide variety of genetically manipulated strains, and amplifying vanishingly small amounts of material. Corresponding general mechanisms for analyzing and perturbing protein function do not exist, meaning that genome-wide analysis of protein behavior lags far behind that of RNA.

This gap should not obscure the importance of proteins for biology. RNA levels are easy to measure, but cells depend on an intricate web of carefully regulated interactions between proteins. The activities of proteins transform a fertilized frog egg from a single cell to a hollow ball of more than a thousand cells; there is no new transcription, and changes in protein stability, rather than in rates of translation, drive the oscillation of the few proteins whose abundance fluctuates during the cell cycle [14,15]. Thus, understanding much of biology will require the branch of genomics often called proteomics: the development and use of techniques to measure the post-translational modification state, localization, binding partners, and ideally the catalytic activity of every protein in the cell.

How close are we to achieving this goal? Two-dimensional gels can in principle report on protein abundance and many protein modifications, but suffer from the disadvantage that the identities of most proteins on the gels are not known, and low abundance proteins are hard to detect. A more expensive alternative is proteolysis followed by peptide fractionation and identification by mass spectrometry. For both methods, tricks that label proteins *in vitro* make it possible to directly compare protein abundance and modification between two samples [16,17]. Protein localization [18] and protein-protein interactions [19] are being probed in yeast in a systematic way, but although both approaches will produce a great deal of useful data, various technical limitations will keep them from providing information about every protein. I believe that this area is the one most in need of new technical developments, such as arrays which display every protein in its native conformation, new chemical tools for rapidly and specifically perturbing protein function, and new computational approaches to biology. One important advance is the ability to use the properties of puromycin to couple proteins to the nucleic acids that encode them [20], which could in principle identify the genes that encode all the proteins that bind to particular molecule.

The discussion above ignores small organic molecules in the cell, which clearly play key roles in all aspects of biology. Can we develop methods to measure these comprehensively in the same way that transcript arrays measure RNA levels?

Can genomics guide reverse engineering?

Genomics can tell us about the components of biological modules, but can it tell us how these modules work? Even if we can use genomics to identify all the proteins that function in a particular module, detect or predict their interactions with each other, perturb their levels, and follow the levels of mRNAs, proteins, and protein modifications, can we uncover the pathway's function and mechanism? This is the challenge of 'reverse engineering', deducing the circuitry and properties of a biological pathway from its components and changes in their levels as the input to the pathway varies. So far, most modelers have concentrated on validating general schemes that biologists have proposed, and trying to distinguish between different variants of a single general scheme. The more challenging task of trying to develop models for pathways solely from whole-genome or whole-pathway data has not been attempted. Perhaps we should mimic the structural

biologists, who hold contests that compare predicted structures to experimentally determined but unpublished ones. The appropriate analog for module prediction might be to construct modules *de novo* with known connectivity and then generate the data on inputs, outputs, and other properties and supply them to reverse engineers. Recent advances in constructing artificial biological networks [21] suggest that this approach will soon become a possibility. A weaker alternative is to ask reverse-engineering programs to predict the structure and behavior of well known modules without using any information (such as biochemical data or detailed patterns of genetic interactions) that would be unavailable for a novel module in a novel organism.

I suspect that machine-based reverse engineering, based solely on data from a single pathway, will be difficult or impossible, mostly because we lack general methods to measure the activity of protein molecules or the rates of chemical reactions they catalyze in living cells. A less satisfying but more feasible aim would be to develop computational tools that recognize similarities between the performance and components of different pathways and thus suggest that they share common design principles. Success in this endeavor would allow very detailed information about a small number of well understood pathways to inform the study of all cellular and molecular biology. In a very ambitious future, we might be able to deduce chemical reactions by using existing data to train programs that would use predicted structures, systematic measurements of the levels and fluxes of protein modifications and small molecules, and docking programs that predict interactions between proteins and small molecules.

The interplay between genomics and evolution

If life had not evolved from a single origin, genomics would lack much of its power. The conservation of sequences allows the function of proteins to be implied from their sequence and critical regulatory sites to be identified in DNA and RNA sequences. Just as thinking about evolution increases the knowledge we can derive from genome sequences, genomics is revolutionizing the study of evolution. The first level of this revolution comes from sequence information, which has already revealed that horizontal gene transfer within prokaryotes has been so extensive that we must think of family mazes as opposed to family trees [22], and the unexpected finding that there have been duplications of the entire genome during the evolution of vertebrates [23]. The second is likely to come from comparing information on gene expression, protein-protein interactions, and genetic interactions between organisms. Such studies will reveal how some pathways have diverged in function while keeping a similar overall structure, while others have evolved from very different structures to have a similar function. For example, only 40% of the predicted proteins in the nematode genome have clear homology to other eukaryotic proteins. Are the other 60% specific to nematodes, or has their primary sequence, but not their function, diverged beyond our ability to detect it? The third, and most important level will be the ability of genomics to enhance experiments on evolution in the laboratory and in the field. For example, it will be possible to determine how selection for a given phenotype has influenced gene expression, and to use genome-wide methods of detecting sequence polymorphisms to map the mutations that produce the phenotype.

Such studies should illuminate the fundamental question in evolution: how do organisms balance robustness, the ability to minimize the impact of short-term genetic variability on their phenotype in favorable environments, with evolvability, the need to respond to long-term exposure to unfavorable environments by producing adaptive phenotypic change? Possible mechanisms include the ability of environmental stress to increase mutation rates and the phenotypic change due to given mutations (reduced phenotypic buffering) [24], as well as the possibility that some mutations that would be deleterious in an optimal environment are advantageous in a difficult one. Using genomics to study how organisms evolve under clearly defined selective pressures should help us solve this riddle. I believe that work on evolution and functional modules will mutually stimulate each other. Understanding how evolutionary pressures constrain modules may help us to decide whether historical chance or the need to mingle robustness with evolvability is the main explanation of why such a small fraction of the pathways an engineer could design for a particular purpose are found in nature. If necessity, rather than chance, largely determines how modules function and interact with each other, experimentation and evolutionary analysis may discover design principles that can be exploited to turn genome-wide data into biological knowledge.

Acknowledgements

I thank Doug Melton for helpful comments. Work in my lab is supported by the National Institutes of Health and the Human Frontiers in Science Program.

References

1. Hartwell LH, Hopfield JJ, Leibler S, Murray A W: **From molecular to modular cell biology.** *Nature* 1999, **402**:C47-52.
2. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS, *et al.*: **The transcriptional program in the response of human fibroblasts to serum.** *Science* 1999, **283**:83-87.
3. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray.** *Science* 1995, **270**:467-470.
4. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, *et al.*: **Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis.** *Science* 1999, **285**:901-906.
5. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
6. Hughes TR, Roberts CJ, Dai H, Jones AR, Meyer MR, Slade D, Burchard J, Dow S, Ward TR, Kidd MJ, *et al.*: **Widespread aneuploidy revealed by DNA microarray expression profiling.** *Nature Genetics* 2000, in press.

7. Dohmen RJ, Wu P, Varshavsky A: **Heat-inducible degron: a method for constructing temperature-sensitive mutants.** *Science* 1994, **263**:1273-1276.
8. Tercero JA, Labib K, F X Diffley J: **DNA synthesis at individual replication forks requires the essential initiation factor Cdc45p.** *EMBO J* 2000, **19:**2082-2093.
9. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.
10. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86-90.
11. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences**. *Science* 1999, **285**:751-753.
12. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**:83-86.
13. Petsko GA: **The Grail problem.** *Genome Biology* 2000, **1**: comment002.
14. Murray AW, Kirschner MW: **Cyclin synthesis drives the early embryonic cell cycle.** *Nature* 1989, **339**:275-280.
15. Murray AW, Solomon MJ, Kirschner MW: **The role of cyclin synthesis and degradation in the control of maturation promoting factor activity.** *Nature* 1989, **339**:280-286.
16. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.** *Nat Biotechnol* 1999, **17:**994-999.
17. Unlu M, Morgan ME, Minden JS: **Difference gel electrophoresis: a single gel method for detecting changes in protein extracts.** *Electrophoresis* 1997, **18**:2071-2077.
18. Ross-Macdonald P, Coelho PS, Roemer T, Agarwal S, Kumar A, Jansen R, Cheung KH, Sheehan A, Symoniatis D, Umansky L, *et al.*: **Large-scale analysis of the yeast genome by transposon tagging and gene disruption.** *Nature* 1999, **402**:413-418.
19. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403**:623-627.
20. Roberts RW, Szostak JW: **RNA-peptide fusions for the *in vitro* selection of peptides and proteins.** *Proc Natl Acad Sci USA* 1997, **94**:12297-12302.
21. Elowitz MB, Leibler S: **A synthetic oscillatory network of transcriptional regulators.** *Nature* 2000, **403**:335-338.
22. Doolittle WF: **Phylogenetic classification and the universal tree.** *Science* 1999, **284**: 2124-2129.
23. Meyer A, Schartl M: **Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions.** *Curr Opin Cell Biol* 1999, **11**:699-704.
24. Rutherford SL, Lindquist S: **Hsp90 as a capacitor for morphological evolution.** *Nature* 1998, **396**:336-342.