Published in final edited form as:

Science. 2008 April 18; 320(5874): 325–327. doi:10.1126/science.1158343.

The Ultimate Model Organism

Atul J. Butte

Stanford Center for Biomedical Informatics Research, Department of Medicine and Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305 USA, and Lucile Packard Children's Hospital, Palo Alto, CA 94304, USA. E-mail: abutte@stanford.edu

Abstract

A deeper understanding of disease requires a database of human traits and disease states that is integrated with molecular information.

This month, the scientific community celebrates the 25th anniversary of GenBank, the open access database of DNA sequences and the molecules they encode. Heralded as one of the earliest bioinformatics community projects, it has fueled our need to understand how this information can be linked to physiology and disease. Since then, biocomputational, informatics, and statistical methods have been used to relate sequences and molecules to diseases. But as highlighted in meetings such as last month's Summit on Translational Bioinformatics (1), the same high-bandwidth measurement style that has accelerated the molecular and genetic study of disease must be practiced in physiology if we are to gain a deeper understanding of normal and impaired health.

Within the last 5 years, systematic studies on the commonalities (2) and differences (3) across diseases have shown that particular biological signaling pathways and modules share similar properties. Other studies have shown that diseases that resemble each other can share genes with variants (4,5) or share genes coding for proteins that interact with each other (6). So many diseases have now been studied that publicly available data can be used to find genes with common changes in expression for each condition (7).

The difficulty with interpreting such analyses lies with how diseases are defined. The definition of a disease is often specified by a particular knowledge base and is thus subject to limitations and biases. For example, a network built from a knowledge base of monogenic diseases (those associated with a single gene) may not be generalizable to more common diseases caused by multiple genes (5). More recently described diseases, such as sudden infant death syndrome, may be less well characterized, so searches for gene variants by matching syndromes through clinical descriptions could yield false-negative predictions (6). In some studies, gene expression or genotyping samples are studied based on clinical disease labeling by physicians, and thus, genes found to be “associated” with a condition do not yet fully explain the observed traits present in a disease. Moreover, there is a growing movement toward direct-to-consumer testing, with the promise that consumer-provided DNA samples could be used for genome-wide association studies. It remains to be seen how such samples can be analyzed when the “assignment” of phenotype and disease is provided by a consumer.

On the other hand, parallel measurements of physiological variables have been successfully linked to genetic markers in animal models of diseases, such as hypertension (8). These efforts have driven the organization of efforts such as the Physiome Project (9), an international collaboration to model the human body through computational methods that integrate biochemical, biophysical, and anatomical information about cells, tissues, and organs. There have also been calls for a Human Phenome Project (10), whose goal is to establish databases

of phenotypes that are associated with physiology, to determine their relation with genes and proteins. Data for some complex human physiological traits are already publicly available for analysis at resources such as PhysioNet (11).

Yet, the current approach for defining phenotypes for molecular discovery is not adequate, and therefore, doesn't optimize the use of physiological data. Phenotypes—traits ranging from height and weight to glucose metabolism, predisposition to disease, and behavioral characteristics— and their differences between individuals can be due to environmental influences and/or genetic variation. One solution for defining richer phenotypes is to take advantage of clinical measurements, which are born from physiological measurements. Enormous numbers of clinical tests are performed each year, and are increasingly being captured in electronic health records, along with patient interventions (medications or procedures). These kinds of data could be used to answer basic biological questions (12,13). Mathematical arrays of such data have already been assembled from hospital-based clinical measurements or epidemiological information and have successfully identified biomarkers for human maturation and aging (14,15). Connections between clinical findings and molecular measurements can also now be tested across a large set of findings and molecules. For example, gene expression profiles of individual liver cancer samples have been predicted by prior radiological findings on abdominal computed tomography scan (16).

How can we take advantage of the petabytes of clinical measurements on patients for whom genetic or genomic measurements may not yet have been obtained? The same broad consideration across diseases used successfully in molecular studies could also be applied to clinical measurements. For example, suppose three diseases are separately considered by a quantitative clinical laboratory test measurement (obtained from an electronic health record), and a gene expression measurement, (from a public repository of gene expression data) (see the figure). Within a disease, the distribution of gene and clinical measurements can be shown, but whether the clinical and gene measurements correlate cannot be determined, as the measurements were not taken from the same patients. But trends might be observed across the three diseases. For instance, as a disease shows more or less of a clinical measurement in patients, microarray samples of the disease may show more or less expression of a particular gene. Thus, associations could be discovered between molecular and clinical measurements, even when these measurements are not made using the same samples or patients. Instead of studying samples or patients as data points in the traditional reductionist manner, one could study and plot diseases.

But there are challenges to using clinical data as physiological measurements. Access issues to patients' private health information can dissuade basic researchers from using clinical measurements. Although patient data could be deidentified and patients approached for informed consent, much clinical data exists as documents that are difficult to deidentify and/or sift through using automated processes. Even as these challenges are addressed, purely numerical quantitative clinicians measurements could be used to start, as these are the easiest to deidentify and analyze.

Whereas there are public international repositories for many molecular measurements, we do not yet have an equivalent for deidentified clinical measurements. There are multiple reasons for this. The fear that personal medical information could be inappropriately released is a powerful disincentive for sharing. Clinical data may also be viewed by clinical and hospitals as a "trade secret," and only recently are data on performance and quality being published. This fear could be averted if health care networks pooled deidentified data sets, thus deidentifying the source of care as well. Clinical researchers are also justifiably protective of the resources they create and might fear missing a discovery within their own patient cohort. Availability agreements could address retention of rights, intellectual property, and publication

embargoes. Instead of viewing data availability as a disadvantage, clinical researchers and institutions should be encouraged to look at the success of resources such as GenBank to see how the public availability of deidentified data can yield many more discoveries when shared.

A population of well-supported and trained scientists and physicians must be nurtured to relate the enormity of physiological and clinical measurements to molecular measurements. The multiscale models of health they build will finally yield an understanding of disease that is more than just the sum of its parts.

References

1. www.amia.org/meetings/stb08/
2. Rhodes DR, et al. Proc. Natl. Acad. Sci. U.S.A 2004;101:9309. [PubMed: 15184677]
3. Segal E, Friedman N, Koller D, Regev A. Nat. Genet 2004;36:1090. [PubMed: 15448693]
4. The Wellcome Trust Case Control Consortium. Nature 2007;447:661. [PubMed: 17554300]
5. Goh KI, et al. Proc. Natl. Acad. Sci. U.S.A 2007;104:8685. [PubMed: 17502601]
6. Lage K, et al. Nat. Biotechnol 2007;25:309. [PubMed: 17344885]
7. Butte AJ, Kohane IS. Nat. Biotechnol 2006;24:55. [PubMed: 16404398]
8. Stoll M, et al. Science 2001;294:1723. [PubMed: 11721057]
9. Hunter PJ, Borg TK. Nat. Rev. Mol. Cell Biol 2003;4:237. [PubMed: 12612642]
10. Freimer N, Sabatti C. Nat. Genet 2003;34:15. [PubMed: 12721547]
11. <http://www.physionet.org/>
12. Sung NS, et al. JAMA 2003;289:1278. [PubMed: 12633190]
13. Payne PR, et al. J. Investig. Med 2005;53:192.
14. Chen DP, et al. Pac. Symp. Biocomput 2008;2008:243. [PubMed: 18229690]
15. Fliss, A.; Ragolsky, M.; Rubin, E. Summit on Translational Bioinformatics Proceedings; San Francisco, CA. 10 to 12 March 2008; Bethesda, MD: AMIA; 2008. p. 111SCB, La Jolla, CA
16. Segal E, et al. Nat. Biotechnol 2007;25:675. [PubMed: 17515910]

1. . The critical intersection of information

Three diseases are separately considered by a quantitative clinical laboratory test measurement, obtained from an EHR, and by one gene's expression measurement, obtained from a public repository of microarrays. Associations can be discovered between molecular and clinical measurements, even when these measurements are not made using the same samples or patients. For example, Disease A, when studied across all patients and time points, shows a high average level of a clinical test (red line), and a low level of a gene (blue line). The distribution of gene and clinical measurements are shown by sampling from both independent data sets (colored regions). The trend across the three diseases shown is that as a disease shows less of a clinical measurement in patients, it shows more expression of a particular gene.