Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power

Trey Ideker,[1,2,*] Janusz Dutkowski,[1] and Leroy Hood[3]
[1]Departments of Medicine and Bioengineering
[2]Institute for Genomic Medicine
University of California, San Diego, La Jolla, CA 92093, USA
[3]Institute for Systems Biology, Seattle, WA 98103, USA
*Correspondence: tideker@ucsd.edu

DOI 10.1016/j.cell.2011.03.007

A major difficulty in the analysis of complex biological systems is dealing with the low signal-to-noise inherent to nearly all large biological datasets. We discuss powerful bioinformatic concepts for boosting signal-to-noise through external knowledge incorporated in processing units we call filters and integrators. These concepts are illustrated in four landmark studies that have provided model implementations of filters, integrators, or both.

Introduction

Complexity is the grand challenge for science and engineering in the 21st century. Complex systems—by definition—have many parts in an intricate arrangement that gives rise to seemingly inexplicable or emergent behaviors. For example, a radio captures an electromagnetic signal and converts it through electronic circuitry into sound that we hear. To most, the radio is a black box with an input (electromagnetic waves) and an output (sound waves). However, understanding the inner workings of this box requires going head-to-head with the challenges of complexity. What are the component parts of the system and how are these parts interconnected? How do these connections influence functions and dynamic system outputs? In biology, ultimately one would like to create models that predict the emergent behaviors of complex entities—and even re-engineer these behaviors to humankind's benefit.

To decipher complexity, biologists have developed an impressive array of technologies—next-generation sequencing, tandem mass spectrometry, cell-based screening, and so on—that are capable of generating millions of molecular measurements in a single run. This enormous amount of data, however, is typically accompanied by a fundamental problem—an incredibly low rate of *signal-to-noise*. For example, the millions of single-nucleotide variants (SNVs) found in a typical genome-wide association study or by the International Cancer Genome Consortium (Hudson et al., 2010) make it extremely difficult to identify which particular SNVs are the true causes of disease. Due to the overwhelming number of measurements, such analyses either lack power to detect the true signal or must admit an unacceptable amount of noise.

Fortunately, biologists have two major weapons with which signal-to-noise may be improved. First is what we know about complexity, which can and should be used as strong prior assumptions when analyzing biological data. Known principles of complexity such as modularity, hierarchical organization, evolution, and inheritance (Hartwell et al., 1999) all provide important insights into how biological systems are constructed and how they function. Second is the availability of data in many complementary layers—including the genome, transcriptome, proteome, metabolome, and interactome. A recent wave of new bioinformatic methods has demonstrated how both weapons—strong prior assumptions related to complexity and systematic accumulation of complementary data—can be used together or separately to exact substantial increases in signal-to-noise.

In what follows, we summarize these developments within a general paradigm for signal detection in biology. Central to this paradigm are processing units we call filters and integrators, which draw on prior biological assumptions and complementary data to reduce noise and to boost statistical power. To illustrate these ideas in context, we review four landmark studies that have provided model implementations of filters and integrators.

The Signal Detection Paradigm

Imagine a biological dataset as a stream of information flowing into a hypothetical signal detection device (Figure 1A). The information flow is quantized into atomic units or events, representing measurements for entities such as genes or proteins, protein interactions, SNVs, pathways, cells, or individuals. Each event contains a certain amount of information, ranging from a single measurement (e.g., strength of protein interaction) to thousands (e.g., an SNV state or gene expression value over a population of patients). Some events represent true biological *signals*, with the definition of "signal" depending exquisitely on the type of results the experimentalist is looking for (e.g., an SNV causing disease or a true protein interaction; many examples are given later). The remaining events are *noise*, which can be due to errors that are technical in nature (uncontrollable variation in different instrument readings collected from the same sample) or biological in nature (uncontrollable variation in different samples collected from the same biological condition). An event may also be considered part of noise even if it is biological and

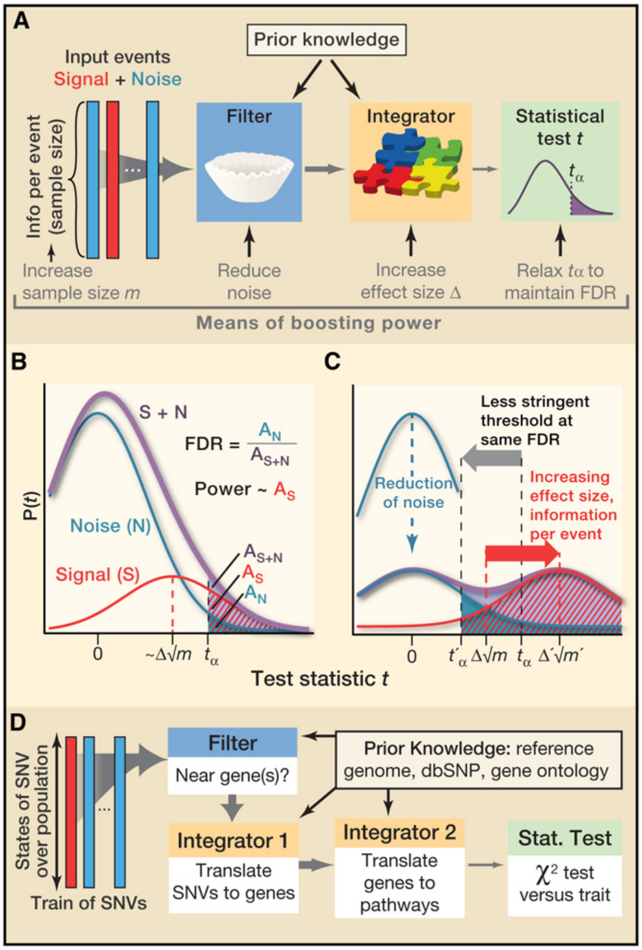

Figure 1. Boosting Signal-to-Noise in Biological Data using Prior Knowledge

(A) Signal detection paradigm in which an input data stream is routed through a series of filtering and integration units, ending in a statistical test that makes accept or reject decisions. Symbols: *m*, information per event or sample size; Δ, effect size; t_α, decision threshold; FDR, false discovery rate.

(B) Probability distribution P(*t*) of the test statistic *t* over the entire data stream of signal plus noise (purple). This distribution is factored into a red signal and a blue noise component. FDR and power are visualized in terms of the areas under these curves to the right of t_α.

(C) Effect of varying parameters on the signal, noise, and signal plus noise probability distributions. The power is increased by more than 6-fold compared to (B), at an identical FDR. Colors are shown as in (B).

(D) MAGENTA, a specific implementation of the signal detection paradigm for pathway-based disease gene mapping as described in Segrè et al. (2010).

reproducible, simply because it encodes aspects of phenotype irrelevant to the current studies.

To make a decision on which events are signal, the device scores each event and accepts those for which the score exceeds a statistically defined decision threshold (Figure 1A). It is precisely this decision that becomes problematic in many large-scale biological studies, in which one either mistakenly rejects a large proportion of the true signal (low *statistical power*) or must tolerate a high proportion of accepted events that are noise (high *false discovery rate* or FDR).

Boosting Signal with Filters and Integrators

To increase signal-to-noise, a pivotal trend in bioinformatics has been to augment the signal detection process with complementary datasets and with prior knowledge about the nature of signal. The vast majority of these approaches fall into either of two categories that we call *filters* and *integrators* (Table S1 available online). Filters attempt to cull some events from the information flow immediately and reject them as noise. For example, a detection system for differential expression might reject certain genes immediately if their expression levels fail to exceed a background value in any condition. Integrators, on the other hand, transform the information flow by aggregating individual events into larger units to yield a fundamentally new type of information, or by integrating together different types of information (Hwang et al., 2009). For example, genes might be aggregated into clusters of similar expression or of related function, in which the median levels of the clusters—not their individual genes—are propagated as the ‘‘events’’ on which final accept/reject decisions are made (Park et al., 2007). Importantly, the combining of filters or integrators results in a new device that itself can be recombined with other signal detection systems in a modular fashion.

Both filters and integrators influence statistical power and FDR, but by fundamentally different means. Filters reduce the fraction of noise passing through the system and, as a consequence, the FDR. Alternatively, as filters are added, FDR can be held constant by relaxing the decision threshold, resulting in higher statistical power (Figures 1B and 1C). By comparison, integrators combine a train of weak signals into fewer stronger events, leading to an increase in ‘‘effect size’’ and thus a direct increase in statistical power. These methods complement the more classical means of boosting power by increasing the amount of information per event (also called the sample size) (Figure 1A).

In each of the following four examples, boosting power with a combination of filters and integrators has been critical to the success of a landmark genome-scale analysis project.

Example 1: Pathway-Level Integration of Genome-wide Association Studies

Genome-wide association studies (GWAS) seek to identify polymorphisms, such as SNVs, that cause a disease or other phenotypic trait of interest. Despite the success of this strategy in mapping SNVs underlying many diseases, the identified loci typically explain only a small proportion of the heritable variation. For such diseases, one likely explanation is that the genetic contribution is distributed over many functionally related loci with large collective impact but with only modest individual effects that do not reach genome-wide significance in single-SNV tests (Wang et al., 2010; Yang et al., 2010).

Based on this hypothesis, Segrè et al. (2010) investigated the collective impact of mitochondrial gene variation in type II diabetes. They described a method called MAGENTA that performs a meta-analysis of many different GWAS to achieve larger sample sizes than any single study, thereby increasing statistical power. MAGENTA also includes both filtering and integration steps (Figure 1D). First, a filter is applied so that SNVs that fall far from genes are removed. Next an integrator is applied to transform SNVs to genes, such that each gene is assigned a score equal to the most significant p value of association among its SNVs. Gene scores are further corrected for confounding factors such as gene size, number of SNVs per kilobase, and genetic linkage. Finally, a second integrator combines the scores across sets of genes assigned to the same biochemical function or pathway, resulting in a single pathway-level p value of association.

Simulation studies using MAGENTA suggest a potentially large boost in power to detect disease associations (Figure S1A). For example, the method has 50% power to detect enrichment for a pathway containing 100 genes of which 10 genes have weak association to the trait of interest. This performance is compared to only 10% power to detect any of the 10 genes at the single-SNV level. At this increased power, MAGENTA did not identify any mitochondrial pathways as functionally associated with type II diabetes, suggesting that mitochondria have overall low genetic contribution to diabetes susceptibility—a surprise given the conventional wisdom about the disease. On the other hand, in an independent analysis of genes influencing cholesterol, MAGENTA identified pathways related to fatty acid metabolism that had been missed by classical GWAS.

Example 2: Mapping Disease Genes in Complete Genomes

Sequencing and analysis of individual human genomes is one of the most exciting emerging areas of biology, made possible by the rapid advances in next-generation sequencing (Metzker, 2010). As complete genome sequencing becomes pervasive, one of the most important challenges will be to determine how such sequences should best be analyzed to map disease genes. The signal filtering and integration paradigm provides an excellent framework for developing methods in this arena. As a landmark example, Roach et al. (2010) described a filtering methodology for disease genes based on the complete genomic sequences of a nuclear family of four. This approach was used to identify just three candidate mutant genes, one of which encoded the Miller syndrome, a rare recessive Mendelian disorder for which both offspring, but neither parent, were affected.

To begin the analysis, the four genome sequences were processed to identify approximately 3.7 million SNVs across the family. SNVs were then directed through a series of filters (Figure S2A). In the first, SNVs were rejected if they were unlikely to influence a gene-coding region annotated in the human genome reference map (http://genome.ucsc.edu/), leaving approximately 1% of SNVs that led to missense or nonsense mutations or fell precisely onto splice junctions. A second filter removed SNVs that were common in the human population and thus were unlikely to cause a rare Mendelian disorder. Like the first one, this filter yielded an approximate 100-fold decrease in the number of candidates. A third filter was designed to check inheritance patterns, which can be gleaned only from a family of related genomes. SNVs were removed that had a non-Mendelian pattern of inheritance (result of DNA sequencing errors) or did not segregate as expected for a recessive disease gene, in which each affected child must inherit recessive alleles from both parents. This filter yielded another 4- to 5-fold decrease in candidate SNVs versus using only a single parental genome. Finally, an integrator was used to translate all remaining SNVs into their corresponding genes.

Using the entire system of filters and integrators under a compound heterozygote recessive model, a total of three genes were identified as candidates. One of these (DHODH) was concurrently shown to be the cause of Miller syndrome. In this way, the family genome sequencing approach used the principles of Mendelian genetics (prior knowledge) to correct approximately 70% of the sequencing errors, directly identify rare variants (those present in two or more family members), and reduce enormously the search space for disease traits (corresponding to an increase in statistical power from 0.15% to 33%) (Figure S1B).

Example 3: Assembly of Global Protein Signaling Networks

Another area in which filtering and integration are turning out to be key is assembly of protein networks. An excellent example of network assembly is provided by the recent work of Breitkreutz et al. (2010), in which mass spectrometric analysis was used to report a high-quality network of 1844 interactions centered on yeast kinases and phosphatases. Central to the task of network assembly was a signal detection system for quality control and interpretation of the raw data. The data consisted of a stream of more than 38,000 proteins that had been coimmunoprecipitated with a different kinase or phosphatase used as bait. Bait proteins can interact both specifically and nonspecifically with a wide variety of peptides, and the nonspecific interactions comprise a major source of noise. To remove nonspecific interactions, the authors introduced a method called *s*ignificance *a*nalysis of *int*eractome (SAINT), in which each putative interacting protein is assigned a likelihood of true interaction based on its number of peptide identifications (representing the amount of information per event or sample size) (Figure S2B). After filtering, the remaining protein interactors are funneled to an integrator stage in which they are clustered into modules based on their overall pattern of interactions (Table S1).

The resulting modular interaction network reveals an unprecedented level of crosstalk between kinase and phosphatase units during cell signaling. In this network, kinases and phosphatases are not mere cascades of proteins ordered in a linear fashion. Rather, they are more akin to the neurons of a vast neural network, in which each kinase integrates signals from myriad others, enabling the network to sense cell states, compute functions of these states, and drive an appropriate cellular response. It is likely that evolution tunes this network, such that some interactions dominate and others are minimized in a species-specific fashion. This might help explain two paradoxical effects seen pervasively in both signaling and regulation: (1) the same network across species can be used to control very different phenotypes (McGary et al., 2010); and (2) very different networks across species can be used to execute near identical responses (Erwin and Davidson, 2009).

Example 4: Filtering Gene Regulatory Networks using Prior Knowledge

One of the grand challenges of biology is to decipher the networks of transcription factors and other regulatory components that drive gene expression, phenotypic traits, and complex behaviors (Bonneau et al., 2007). Toward this goal, probabilistic frameworks such as Bayesian networks have been extensively applied to learn gene regulatory relationships from mRNA expression data gathered over multiple time points and/or experimental conditions (Friedman, 2004). However, due to a limited sample size, large space of possible networks, and probabilistic equivalence of many alternative models, these approaches are often unable to find the underlying causal gene relationships.

Recently, Zhu et al. (2008) showed that supplementing gene expression profiles with complementary information on genotypes may help to overcome some of these problems (Figure S2C). These authors sought to assemble a gene regulatory network for the yeast *Saccharomyces cerevisiae* using previously published mRNA expression profiles gathered for 112 yeast segregants. Rather than assemble a Bayesian network from expression data alone, the data were first supplemented with the genotypes of each segregant. The combined dataset was then analyzed to identify *expression quantitative trait loci* (eQTL)—genetic loci for which different mutant alleles associate with differences in expression for genes at the same locus (*cis*-eQTL) or for genes located elsewhere in the genome (*trans*-eQTL). The eQTLs were used as a filter to prioritize some gene relations and demote others. Any candidate cause-effect relations in which the effect gene is near an eQTL were removed, as the *cis*-eQTL already explains the gene expression changes at that locus. Conversely, cause-effect relations that were supported by trans-eQTLs and passed a formal causality test were prioritized. Supplementing gene expression profiles with genetic information significantly enhanced the power to identify bona fide causal gene relationships. Further improvement was achieved by introducing a second filter that prioritized cause-effect relations that correspond to measured physical interactions, including data from the many genome-wide chromatin immunoprecipitation experiments published for yeast that document physical interactions between transcription factors and gene promoters.

Summary

Biology is expanding enormously in its ability to decipher complex systems. This ability derives from the expanded power to incorporate diverse and complementary data types and to inject prior understanding of biological principles. Signal detection systems such as those discussed here—along with their filters, integrators, and other components—are leading to fundamental new biological discoveries and models, some of which will ultimately transform our understanding of disease and therapeutics. It is also likely that many of the strategies, technologies, and computational tools developed for healthcare can be applied to problems of complexity inherent in other scientific domains, including energy, agriculture, and the environment. Healthcare and energy will demand significant societal resources moving forward—and hence offer unique opportunities to push the development and application of approaches for attacking complexity.

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and one table and can be found with this article online at doi:10.1016/j.cell.2011.03.007.

ACKNOWLEDGMENTS

We gratefully acknowledge G. Hannum, S. Choi, I. Shmulevich, D. Galas, J. Roach, and N. Price for helpful comments and feedback. This work was funded by grants from the National Center for Research Resources (RR031228, T.I., J.D.), the National Institute of General Medical Sciences (GM076547, L.H.; GM070743 and GM085764, T.I.), the Department of Defense (W911SR-07-C-0101, L.H.), and the Luxembourg strategic partnership (L.H.).

REFERENCES

Bonneau, R., Facciotti, M.T., Reiss, D.J., Schmid, A.K., Pan, M., Kaur, A., Thorsson, V., Shannon, P., Johnson, M.H., Bare, J.C., et al. (2007). Cell *131*, 1354–1365.

Breitkreutz, A., Choi, H., Sharom, J.R., Boucher, L., Neduva, V., Larsen, B., Lin, Z.Y., Breitkreutz, B.J., Stark, C., Liu, G., et al. (2010). Science *328*, 1043–1046.

Erwin, D.H., and Davidson, E.H. (2009). Nat. Rev. Genet. *10*, 141–148.

Friedman, N. (2004). Science *303*, 799–805.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). Nature *402*, C47–C52.

Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bell, C., Bernabe, R.R., Bhan, M.K., Calvo, F., Eerola, I., Gerhard, D.S., et al. (2010). Nature *464*, 993–998.

Hwang, D., Lee, I.Y., Yoo, H., Gehlenborg, N., Cho, J.H., Petritis, B., Baxter, D., Pitstick, R., Young, R., Spicer, D., et al. (2009). Mol. Syst. Biol. *5*, 252.

McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., and Marcotte, E.M. (2010). Proc. Natl. Acad. Sci. USA *107*, 6544–6549.

Metzker, M.L. (2010). Nat. Rev. Genet. *11*, 31–46.

Park, M.Y., Hastie, T., and Tibshirani, R. (2007). Biostatistics *8*, 212–227.

Roach, J.C., Glusman, G., Smit, A.F., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., et al. (2010). Science *328*, 636–639.

Segrè, A.V., Groop, L., Mootha, V.K., Daly, M.J., and Altshuler, D. (2010). PLoS Genet. *6*, e1001058.

Wang, K., Li, M., and Hakonarson, H. (2010). Nat. Rev. Genet. *11*, 843–854.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Nat. Genet. *42*, 565–569.

Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E., and Schadt, E.E. (2008). Nat. Genet. *40*, 854–861.