GuestEditorial

BY DR. ANDREA CALIFANO

Taking the leap: from single genes to the molecular choreography of the cell

The Human Genome Project has spurred extraordinary developments in our ability to characterize cellular systems in high-throughput fashion. Polymorphism, methylation, gene expression, and proteomics profiles are just a few of the data modalities that we could not have hoped to measure without a genomic sequence anchor. Yet, despite this flood of new data, we are still struggling in our understanding of the complex molecular choreographies that ultimately determine physiological and pathological phenotypes. The reasons are multifold and define some of today's grand challenges of molecular biology and medicine. Two issues, however, perhaps most significantly contribute to our still limited understanding of normal and disease-related cellular processes.

First, while we have now characterized the majority of our gene repertoire and, to some extent, its byproducts (e.g., microRNA, proteins, etc.), cellular processes are not regulated by individual genes but rather by complex webs of molecular interactions that are precisely time- and space-coordinated and exquisitely phenotype specific. Unfortunately, we still know far too little about these interactions. Our estimate, is that only 1 to 20 percent of the full human interactome may be represented in the literature and in databases. Also, as reported at the 2007 Hinxton Interactome meeting in the United Kingdom, even for yeast, as many as 50 percent of the protein-protein interactions reported in individual articles may not be reproducible. Thus, our existing molecular interaction maps constitute a small, partially incorrect, and certainly largely incomplete view of the interactions that *could exist* rather than the ones that *are implemented* in a specific cellular context of interest.

Second, we are still mostly looking at individual genes as the determinants of both normal cellular processes and of their dysregulation in disease. For instance, most genome-wide disease association studies still test polymorphisms either in isolation or within small contiguous genomic regions, rather than in the context of interacting gene products. Even worse, individual factors such as copy number, chromatin methylation and acetylation, SNPs, expression levels, and post-translational modifications, are also being mostly studied in isolation, rather than integrated within pathways. Yet the same function may be regulated or dysregulated by any of these factors individually or in combination, each one contributing a very small fraction of the total penetrance. As a result, the total number of distinct molecular phenotypes may be as high as the product of all genetic/epigenetic variants across all interactions in a set of synergistic pathways: a very high number indeed!

To capture any unifying principles behind the heterogeneity of specific diseases, it will be crucial that we start integrating all available information over reasonably accurate, genome-wide interaction maps. To accomplish this goal, we may need once again to come together as a community of computational and experimental scientists to embark on an even greater challenge than the human genome project: the genome-wide, high-resolution mapping of molecular interactions, within highly specific cellular phenotypes. These maps will provide critical information on how several gene products work together to implement specific cellular functions. Additionally, they will provide an extraordinary resource to integrate disparate information on normal and disease-related tissue. Specifically, we are starting to see that distinct molecular determinants of the same cellular phenotype cluster within relatively compact regions of these molecular interaction maps, rather than being randomly distributed.

While promising approaches have been demonstrated to accomplish this goal both experimentally and computationally, the end-goal of this activity cannot be reached by using either approach alone. Indeed even once the experimental data becomes available, complex, multi-faceted computational challenges will have to be addressed. These range from an appropriate ontological classification of the cellular phenotypes, processes, and molecular species, to the 2D/3D visualization of the intricate interaction networks (possibly in a time- and space-dependent manner), to the development of reverse engineering approaches that leverage both functional and structural data, to the identification of relevant biomedical problems. Sound familiar? It should, because these are precisely the computational challenges that the seven National Centers for Biomedical Computing are currently trying to address. □

DETAILS

Dr. Andrea Califano is a professor of biomedical informatics at Columbia University, associate director of the Herbert Irving Comprehensive Cancer Center, and principal investigator for MAGNet—the National Center for Multiscale Analysis of Genomic and Cellular Networks (MAGNet) at Columbia University.