S1

Computational Approaches to Biological Inference: Baby Steps to Computing the Organism

Junhyong Kim

Department of Biology, Penn Center for Bioinformatics, University of Pennsylvania, USA

In the last 10 years, biological data gathered by high-throughput systematic methods such as those employed in the Human Genome project has been accumulating at an exponential rate. At the same time, actual experimental investigation of this data has been increasing only at a constant rate creating a large gap between the collected data and extracted biological knowledge. It is now well recognized that this gap can only be closed by the employment of large-scale computational analyses and high-throughput experimentation *a' la* functional genomics. The need for computational analysis prompted a rapid growth in the field of computational biology and bioinformatics and led to their extreme popularization. More recently, somewhat analogous to the Internet *boom-and-bust* that paralleled the growth of bioinformatics, there has been a reflexive disappointment at the perceived failure of the nascent field to deliver on the high expectations of computational approaches. Yet, the gap in the progression from data to knowledge will not go away and high-throughput experimental techniques such as functional genomics exacerbate the need for computational analysis rather than reducing its need. In this talk I first broadly examine the structure of computational problems in biology. I categorize the computational activities into those involving "Support and Infrastructure", "Technology", and "Theory". I relate these problems to the traditional Information Technology classification of "Data", "Information", and "Knowledge". I next show short examples of work in my lab from each of these categories. Under "Support and Infrastructure", I show work we are doing under NSF ITR program to establish a National Center for Computational Phylogenomics. In particular, I outline the problem of optimally extracting comparative genomics data from existing databases. Under "Technology", I describe the work done to develop a new algorithm for searching for multi-transmembrane proteins that led to a collaborative work resulting in the first cloning of olfactory receptor genes in insects. The main technical problem here was to generate a new data model in which primacy was placed on statistical characterization of protein structure rather than homology of individual amino-acid strings. In the main part of my talk, I concentrate on Computational Biology as a scientific endeavor rather than a technical enterprise. All natural sciences in their development experience a transition from asking, "What is it?" to asking, "What are its organizational principles?" Biology in the 20th century has slowly experienced this transformation, which is exemplified by such instances as the re-establishment of the transmission principles of Mendel and the discovery of the remarkable organization of the genetic code. The exponential growth of data-especially at the system-wide level has particularly focused our attention on the need to ask questions of organizing principles. Problems and answers of organizational principles usually require data abstraction, data modeling, and quantitative calculation. Thus, computational and mathematical approaches are not only the new tools for aiding traditional biology but in fact the backbone towards developing a new theory of biological organization. We wish to know what are the principles by which the organism computes itself and how has this organization resulted in the evolution of myriad of organisms in nature. We are just starting to make baby steps towards asking such questions. Here, I first present data we collected on whole-genome gene expression evolution in different species of Drosophila and mutation accumulation lines. I show that even closely related strains of Drosophila show a surprising amount of gene expression change. I also show that the evolution of gene expression can be used to infer functional importance of transcription regulation of a particular gene. This analysis is analogous to that employed in sequence analysis where conservation of a trait is used as a signature for stabilizing selection for function. One of the fundamental organizing principles in nature is that complex entities are built up from modular building blocks (e.g., molecules fromatoms). Thus one of the first steps in understanding the organization of a complex system like whole genome transcriptional regulation is to ask whether it can be decomposed into modules of semi-independent units. I describe our attempt at developing an algorithm that uses whole-genome expression data to identify modular units. I show that modules identified in this manner correspond well with biochemical and physiological functions. Furthermore, under selective response to noble environments, the genome seems to respond in these same modular units. Modular organization can be selected for under evolutionary process. A surprising kind of modular organization can be found in protein 3D structures. Most natural proteins display a high degree of self-similarity where geometrically similar substructures are found in the same protein. Such self-similar units may correspond to modules in the folding process of the protein. There is reasonable evidence that natural proteins may be selected to "fold fast"--i.e., reach their minimal free energy configuration with minimal number of transition states. We developed a new algorithm for measuring the degree of self-similarity in protein structures and found evidence that common protein structures tend to be more self-similar consistent with the theory that nature tends to select for modular organizations. The transition in biology from data gathering to theorizing--especially theorizing at a global scale is already taking place. Some people are beginning to call some of this activity "Systems Biology." However, all natural sciences have in fact undergone such transformation where theory and prediction are part of the mainstream activities of the field. In particular, the key to this transition lies not only in more data but also in asking new questions. Just as physicists discovered Boyle's laws of gas without computing the movement of individual molecules, even without a complete molecular characterization of the organism, there will be new biological laws to be discovered if proper abstractions areemployed and new questions are asked.