Task Group Summary 1

How would you design the acquisition and organization of the data required to completely model human biology?

Challenge Summary

In many fields that relate to complexity, datasets are still fragmentary and questionable in terms of their overall quality. This is particularly true in the field of biology. Small-scale empirical data have been described for decades in hundreds of thousands of papers published in thousands of journals. This information, although generally perceived as highly accurate, is extremely hard to extract in reliable ways. On the other hand, high-throughput systematic biological datasets tend to be widely accessible, but are currently perceived as lesser quality information. This represents a considerable challenge if one considers the fact that, relative to its widely accepted complexity, the molecular aspects of human biology have been described only superficially.

Key Questions

With the general assumption that we are given funding in the range of what was allocated to sequence the human genome between the late 1980s and the early 2000s (~$3,000,000,000), the following questions will be addressed:

• How would you design the acquisition of new data pertaining to human biology?
• How would you validate the inherent quality of such data?
• How would you organize this information into practical, usable datasets made available in databases ready to be used by the research community?
• How would you design the development of analytical tools to attempt to entirely model the molecular and physiological complexity of the human body?
• How would you relate this information with genetic and environmental factors that influence disease and good health?

Required Reading

Suggested Reading

TASK GROUP MEMBERS
• Ananth Annapragada, University of Texas Houston
• James Glazier, Indiana University
• Amy Herr, University of California, Berkeley
• Barbara Jasny, Science/AAAS
• Paul Laibinis, Vanderbilt University
• Suzanne Scarlata, Stony Brook University
• Gustavo Stolovitzky, IBM Research
• Eric Schwartz, Boston University

TASK GROUP SUMMARY

By Eric Schwartz, Graduate Science Writing Student, Boston University

The question of how to put together and organize the data needed to simulate human biology is large and complex. At the 2008 National Academies Keck Futures Initiative Conference on Complex Systems, a Task Group (1) of scientists from multiple disciplines met to contemplate the problem.

The goal is a complete, easily queried simulation that would be comprehensive and could synthesize different data to give useful answers to questions about human physiology in health and disease. This is obviously a monumental undertaking, especially when we realize the limitations of current state-of-the-art computers and technology, and our mental ability to conceptualize such problems. Nevertheless, the group developed an initial plan upon which many future directions can be based.

First, there is the challenge of obtaining information that scientists know would be essential. For instance, it is estimated that humans have approximately 25,000 genes, but the interactions of only about 10% of their products are known. Proteins made by genes, in multiple forms, interact with each other in different ways. A comprehensive simulation might require knowledge of protein production and behavior in space and time. (Not all proteins are active at all times or at all places.) Metabolic, signaling, and gene regulatory pathways, known and yet to be understood, would be part of the simulation, as would patterns of neuronal growth and decay, and whole organ anatomy and function, and nervous, endocrine, circulatory, respiratory physiology systems, etc. Altogether, simulating human biology is an immense problem not only of biological research but also bioinformatics, biomedical computation, epistemology, and computer power. Unless all of this information is combined in an understandable format, a lot of important and relevant medical data for a human simulation would be ignored.

The Initial Plan

The group considered many options for collating and organizing data. It was decided that one of the most important steps is to find out what empirical data compilations already exist and organize them according to some basic principle to avoid covering ground already covered on a scale ranging from the molecular, protein, cellular, organ, and full-organism scales. There may be more than 250 types of cells in the human body, each with their own unique functioning and
Therefore the group decided the interaction between different levels of biology is just as important as what occurs on those levels alone. Ultimately, correlation of the different types of data rests on good indexing, or a metastructure to define all of the phenomena. The group concluded that the best way to deal with the question of human biology as a whole is to break it up into different parts. Five basic databases were outlined by the group with the expectation that the databases could then interact with each other. The databases enumerated were:

1. Simulations of subsystems and connections. In this case meaning cellular, protein, and other systems and how they relate to each other.
2. Limitations: that is, the limitations posed by the lack of standardized datasets on human biology and the ability to relate these data to information about biological simulations.
3. Experimental data plus metadata for different cases and perturbations.
4. Templates of appropriate subsystem choices and connections for different categories of problems.
5. Sample complete simulations and outputs.

The group then broke down the databases into smaller subsets of knowledge. The database of parameters was for example further subdivided into spatial and temporal distribution, mechanical properties, cell behavior, and biomarkers. The group decided the most important issue facing them was the many gaps in their knowledge. The different databases currently in existence aren’t standardized and there is no consensus ontology or unified computational tools to deal with the data already compiled. Ontologies are logical structures which provide a formal description of concepts. An ontology is simply a hierarchy of terms with understood meanings and sets of subterms and modifiers which can be applied to each term.

In order to know what to do with the overwhelming complexity of the whole of human biology, the group decided that a pilot program to test their basic ideas is essential. If successful, the pilot simulation could then be the basis for simulations of other parts of human biology. The pilot would need to be something simple but at the same time useful for discussion. Although many options, from cancer to neurodegenerative diseases, were considered, the group settled on the effects of an injection of norepinephrine into the body. This chemical is used on people in anaphylactic shock caused by allergies or toxins and has several well-understood effects. By comparing what is known about norepinephrine in people to the theoretical simulation, the simulation could be tested for predictive capacity.

The Five Year Plan

The group resolved to create a list of goals that could be achieved within five years, should sufficient resources be applied to the work of a complete simulation of human biology—“Google Human.” Firstly, they wanted to create an inventory of all the data currently available and a preliminary inventory of all the missing data. Once the data have been created and compiled, a quality control check of all the data will be necessary to make sure that the data are correct and put into a format that is consistent for computer analysis. Along with creating standards, designing new computational tools will be an important early step in the program. For a detailed outline of the group’s thoughts, see Figure 1.